Design and implementation of a method for the synthesis of travel diary data

Roland Martijn Kager

Dissertation committee:

prof. dr. ir. H. J. Grootenboer	Universiteit Twente, chairman/secretary
prof. dr. ir. M. F. A. M. van Maarseveen	Universiteit Twente, promotor
prof. dr. ir. E. C. van Berkum	Universiteit Twente
prof. dr. M. J. Dijst	Universiteit Utrecht
prof. dr. ir. A. Y. Hoekstra	Universiteit Twente
prof. ir. L. H. Immers	Katholieke Universiteit Leuven, Belgium
prof. dr. G. C. de Jong	University of Leeds, United Kingdom
prof. dr. M. J. Kraak	International Institute for Geo-Information Science and Earth Observation

TRAIL Thesis Series T2005/11, The Netherlands TRAIL Research School

TRAIL Research School
PO Box 5017
2600 GA Delft, The Netherlands
Telephone: +31 15 2786046
Telefax: +31 15 2784333
Email: info@rsTRAIL.nl

This thesis is the result of a Ph.D. study carried out between 2001 and 2004 at the University of Twente, faculty of Engineering Technology, department of Civil Engineering, Centre for Transport Studies.

Cover picture: 'Beste Katalysator der Welt' (Prinzipalmarkt, Münster)

Typeset in LaTeX
Printed by Febodruk BV, Enschede, The Netherlands.

ISBN 90-5584-076-9

DESIGN AND IMPLEMENTATION OF A METHOD FOR THE SYNTHESIS OF TRAVEL DIARY DATA

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de Rector Magnificus,
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 14 oktober 2005 om 13.15 uur

door

Roland Martijn Kager

geboren op 25 januari 1976
te Leiden

Dit proefschrift is goedgekeurd door de promotor:

prof. dr. ir. M. F. A. M. van Maarseveen

The challenge of transportation systems analysis is to intervene, delicately and deliberately, in the complex fabric of a society to use transport effectively, in coordination with other public and private actions, to achieve the goals of that society (Mannheim, 1979)

Thirty years of transforming a society and a space economy to high levels of mobility and energy intensity have failed to notice the decline of community, the loss of people from the streets, and the increasing isolation of children, the elderly and women as the built environment fails them and through them fails the idea of community itself (Whitelegg, 1997)

Contents

Appendices

References

Preface

-Se eu tivesse um carro	- If only I had a car,
havia de conhecer	I could get to know
toda a terra.	the entire world.
Se eu tivesse um barco	If only I had a ship,
havia de conhecer	I could get to know
todo o mar.	all of the sea.
Se eu tivesse um avião	If only I had a plane,
havia de conhecer	I could get to know
todo o céu.	all the skies.
-Tens duas pernas	- You have got two legs
e ainda não conheces	and you don't even know
a gente da tua rua.	the people in your street.

Child story,
Luísa Ducla Soares (1939-2004)

Cars, like ships and planes, have an important role to play in our societies. But, as the above poem and the front cover of this report illustrate, some reflection on this role is necessary. The pictures on the front cover were taken from an awareness campaign by the bus operating company of the historic city of Münster (Germany). Combined, the pictures vividly illustrate the implication of the land claim by cars in urban city centres.

The poem is on the vanity of daydreaming, or perhaps on its innocent and human nature. At the same time, the poem is positively enforcing: stop looking for barriers, look for your opportunities and use them to realise your dreams. Nonetheless, in real life, limited transport options are a barrier to many. At the same time, infrastructure extension programmes are ever more strongly confronted with high financial, environmental and (since recently) juridical barriers. This forces us to rethink of the people and the activities that we would like the transport system to support. This report aims to assist in such analysis

by delivering a tool for the synthesis of detailed data on the actual functioning of the transport system, basing itself on already existing data sources.

This report is based on the combining of large databases from various organisations. I am thankful to Wegener Direct Marketing for granting usage of their detailed socioeconomic database for an extensive study area. My gratitude also goes to Rijkswaterstaat Oost Nederland in Arnhem for allowing me to use a detailed trip matrix for the East of the Netherlands from the *NRM Oost-Nederland*. Next, I owe thanks to I&O Research in Enschede for granting usage of the *Koopstroomonderzoek Oost Nederland*, an extensive shop trip database for the same area. Finally, I would like to thank the Adviesdienst Verkeer en Vervoer (AVV) of the Dutch Ministry of Transport for having made available the *Onderzoek Verplaatsingsgedrag* at a low geographic aggregation level in 1995 and for their current, promising work on the successor of this survey, the *Mobiliteitsonderzoek Nederland*.

The writing of this report didn't always go with the ease that some accredited me with. Here, I would like to thank some of the people who assisted me and who were important during the writing. First is my supervisor, Martin van Maarseveen. Martin, you have taken the chance from me to find out how my life would have been without writing this thesis. I feel happy to have been given this chance and I want to thank you for your trust in me. At the university, it has been a pleasure to be part of the diverse group of people that together constitute the Centre for Transport Studies. In particular, Mark Zuidgeest, in awaiting of all our future travels and business projects, I indebt much thanks to you. Always full of energy when it comes to both helping and correcting others, I will badly miss having a teacher like you near. This also applies for Bas Tutert. Bas, you are a constant source of inspiration to many and so have you been invaluable for me. It was fun and always inspiring to have shared an office with you. Many good things in life, I learned from you and I am thankful for that.

Outside the university, yes often very far outside, Claudia, you always like to take the hard way. Therefore, it's no surprise you came into my life only for the second half of the writing of this report. A strong woman you are and I am happy to be with you; I hope the next book we write together. Very far in the other direction have been my parents. Jan and Carla, you have made me what I am now. It is your characters that I try to combine and it's a good mix. Thanks for always supporting me.

Roland Kager,
Deventer, September 2005

Chapter 1

Introduction

Inhabitants of the Dutch, medium-sized city of Almelo were asked to report three problems in their neighbourhood that they wished to be addressed with priority. A total of 4,224 randomly selected inhabitants answered this open question and reported over 12,000 priority neighbourhood problems during eight years of survey. Table 1.1 classifies these problems for six main categories. Remarkably, where 23.7% of the reported problems are related to the combination of crime, vandalism, social insecurity, waste disposal or noise pollution from industry, restaurants or bars, a much higher 39.4% of all problems are related to the local transport system alone.

"This neighbourhood problem should be addressed with priority" (open question, answers categorised)	Problems (N=12,504) (max. three problems per respondent)	Respondents (N=4,424)
Crime and non-transport-related nuisances	23.7%	43.1%
Level of urban facilities	17.2%	31.3%
Maintenance of housing and facilities	13.2%	24.0%
Contacts with local government	5.7%	10.3%
Transport system related	39.4%	66.8%
-Provision of transport infrastructure	8.4%	15.2%
-Transport system related nuisances	31.0%	56.3%
Other	1.0%	1.7%
Total responses	100.0%	177.2%

Table 1.1: Priority neighbourhood problems, Almelo, 1995-2003, (Gemeente Almelo, 2003)

The Almelo study is mentioned as an illustration of the large impact that the transport system has on our day-to-day neighbourhood life. Although Almelo is not considered atypical in this respect, many are perhaps inclined to search for special circumstances that caused these outcomes. Born and raised in societies that continuously communicate positive images of cars as representing individual freedom, prosperity and success, many of us have perhaps grown lenient to the problems associated with this means of transport. Probably, the above results are nonetheless obtained because this leniency is at its low-

est where the intrusion of cars is perceived at its highest: in someone's own neighbourhood. It is also for one's own neighbourhood that solutions are more rapidly perceived as easy and non-intrusive; the banning or diverting of cars or the imposing of lower speed limits. It is thus that these local transport system related problems are more actively perceived than those that manifest themselves at a higher level. Here, at the level of complete cities and where the above solutions generally seem less applicable, John Whitelegg offers an interesting analogy with other problems that for many centuries have been perceived as inseparable with urban life:

> Likewise, those in twentieth-century cities are now coming to realise that the car and the lorry are at least as big a problem as the polluted water and sewage systems of the nineteenth century. (Whitelegg, 1997)

This report aims to deliver a tool that decomposes traffic flows into all of its individual trips, relating these trips to the living situation of the persons who perform them, their motivations to make the trips and to give a full specification of the trip-pattern of which the trips are part. Relating traffic-induced problems to individual trip contexts and individual persons is considered an important weakness of current transport planning practice. Without such analysis, traffic flows remain intangible and anonymous, making it impossible to determine the net societal utility of traffic flows (the difference between the traffic-related benefits and traffic-induced problems for society). This report proposes the synthesis of travel diary data as a first step in the upheaval of this anonymity. For the Almelo case study, this could relate the neighbourhood problems to individual trips segments, the persons that perform them and the motivations and contexts of these trips.

Chapter outline The above discussion offers a first sketch of the background of this report and the contribution it intends to make to transportation science. The remainder of this chapter is devoted to a more detailed discussion of both aspects. First, section 1.1 reviews costs and benefits of transport systems and how these are evaluated by current transport planning practice. Next, section 1.2 discusses travel diary data, its potential usage for transport cost-benefit analysis and its collection by household travel surveys. Considering that no empirical data is required for some applications of travel diary data, section 1.3 proposes the synthesis of travel diary data as a cost-effective extension to such data collection. Next, section 1.4 discusses techniques that are available for such synthesis, based on a list of modelling classifications. Out of these techniques, section 1.5 selects a technique based on which this report designs and implements a method for the synthesis of travel diary data. The section also reviews two other research efforts that are considered the state of the art for the synthesis of travel diary data. The chapter concludes with the formulation of four research questions, a reading guide for the remainder of the report and a summary.

1.1 Costs and benefits of transport systems

This first section provides an overview of current research on the cost and benefit components of the transport system. The first paragraph distinguishes internal from external costs and benefits, and explains why only the latter are generally of interest for transport planning. Next, main findings are discussed for the external costs and the external benefits of transport systems. The section then reviews how both external costs and benefits are compared in current transportation practice. It is noted that because the current transport system is dominated by cars and road infrastructure, the discussion will mostly centre on this transport mode.

Internal and external costs and benefits The distinction between internal and external effects refers to whether or not these effects are experienced by the ones who cause these effects. For transport systems, internal costs are for example the money, time and effort invested by individuals for their travel. Internal benefits then are all personal benefits derived from this travel: the possibility to engage in socioeconomic and cultural activities out of home, the fulfilment of seeing distant cities and countries or the joy of visiting family and friends. Also the considerable economic activity associated with transport systems such as the automobile industry, gasoline industry or the car insurance, repair and maintenance sector, represent internal effects for the part that they are fully paid for by drivers themselves. In contrast, external costs are all costs not taken into account by the ones who cause these effects. Well-known examples are costs due to air pollution, noise pollution or traffic accidents. Less known costs are land claims, captivity and exclusion effects or loss of local control, discussed later in this section. External benefits are centred around the increased socioeconomic development of the area where car-users make their trips. Although these external benefits are often utilised as the strongest argument in favour of infrastructure provision, various studies (discussed later in this section) show that these effects are often just redistributions of socioeconomic development rather than true increases in overall socioeconomic development as it is argued later in this section. Ideally, a cost-benefit analysis for transport policy should address external costs and external benefits only. This is a consequence of the typical state provision of infrastructure whose role it is to defend societal interests rather than individual interests. Both external costs and benefits are now discussed in greater detail.

External costs of transport systems The external cost of the transport system -mainly the road system- is composed out of a long list of socioeconomic and environmental costs not paid for by the users of the system:

- *Accident costs* In the 1990s, an annual 420,000 global deaths and 9 million injuries have been reported as a result of road traffic accidents. These figures are estimated to rise to 2 million annual deaths and 50 million annual injuries in 2030 (Teufel, 1995). Although user-paid insurance covers some

of the costs associated with traffic accidents, many accident-related costs are rolled off to society. These accident-related external costs include emotional damage to victims and families, prevention costs (including for example the prohibiting of children to play outside or to walk to school unattended), traffic disruption costs, policing costs, court and procedural costs.

- *Air pollution* Air pollution is perhaps the most obvious and well-known external cost for transport. Remarkably, at the same time, popular knowledge and concern on the scope and size of these costs is limited. Apart from climate change costs discussed hereunder, the traffic-induced emission of carbon monoxide (CO), nitrogen oxides (NO_x), sulphur dioxides (SO_2), hydrocarbons (HC), formaldehyde, asbestos, diesel particulates, lead and, indirectly, ozone causes considerable environmental and health costs. Some believe that a decline of emissions per vehicle mile travelled (VMT) will eventually solve this problem. However, where engine and fuel improvements have indeed significantly reduced tailpipe emission rates under design conditions, a significant portion of driving occurs under non-design conditions. At the same time, non-tailpipe emissions such as tyre abrasion, road wear and road dust are not controlled by these technologies (Litmann, 2004d). In his comprehensive and well-readable book, Whitelegg (1993) reviews a long list of surveys on both the environmental costs and the health costs of air pollution. These studies vary from noting the overall 21.5% forest damage in Europe due to acidification in 1991 (WRI, 1994), the fivefold increase in hospital admissions for childhood asthma between 1979 and 1989 (Holgate, 1992), a minimal estimate of 10,000 to 24,000 premature US-deaths from air pollution in 1985 (American Lung Association, 1990), studies on transport-related air pollution and cancer incidence, the formal recognition by the WHO (1990) and the OECD (1991) that the scale of air pollution from motor vehicles is unacceptably high, as well as many urban studies that show how current air-pollution levels for many pollutants are already well above the official (sometimes not very demanding) standards. Litmann (2004d) reviews various research to quantify the costs of air pollution and arrives at a middle USD 0.112 estimate per VMT for the weighted fleet average in 1990 monetary levels and for health costs alone (Delucchi, 1996). Converted to 2005 euro estimates, assuming 1990 US pollution levels and an average 6 vehicle miles per liter of gasoline, this implies an approximate EUR 1.00 surcharge on European gasoline prices to account for health costs alone. This surcharge roughly equals the current average gasoline price in Europe, which is already heavily debated to be too high.

- *Climate change* Climate change or popularly, global warming, refers to the increase in global mean temperature as a consequence of the increased concentration of greenhouse gases in the atmosphere. Long disputed and ignored at first, the phenomenon is now accepted to be caused by mankind and lead to unprecedented changes in climate, causing loss of fauna and

flora, drastic changes in annual precipitation, increasing both the likelihood of droughts and floodings, the widespread retreat of glaciers and snow-cover in the Northern Hemisphere, rises in sea levels due to thermal expansion of the ocean and melting of glaciers, threatening extensive coastal regions, islands and atolls, increasing coastal erosion, flooding risk and the salinity of farming grounds, estuaries and aquifers (IPCC, 2001).

- *Equity costs* In case it is decided to invest in infrastructure for one transport mode, equity principles demand the same investments to be made for the infrastructure of alternative transport modes to the ratio in which these modes offer transport performance. In addition, where infrastructure investments for the one mode incur increased costs for other modes, these cost effects should be accounted for fully, again to the ratio in which these modes offer transport performance. Hence, equity costs are the rate to which such equity investments are left out, and thus quantify the degree to which certain transport modes and their associated user groups are disfavoured to the benefit of others. Both in the developed world and increasingly in developing countries (Vasconcellos, 2001), equity costs are tremendous with the most efficient transport modes (walking and biking) hardly receiving any funds, while at the same time suffering extensively from improvements made in road infrastructure, by the sacrificing of sidewalks, by increased exposure to accidents and ambient pollution levels, by increased road crossing barriers or reduced social status. Another aspect of equity costs is *intergenerational equity* expressing the degree in which future generations suffer from transport policies that mainly offer net benefits to current generations (see for example Litmann, 2004b).

- *Land use* Considering its transport performance of mostly just one person, the car consumes large and costly land resources both when moving or when parked. The total consumption of land for transport infrastructure, mainly roads and car parking space, accounts for 1.3% of the total available land area in the EU in 1990 (Whitelegg, 1993). Moreover, as Whitelegg reports for Switzerland in 1990, the land allocated to road transport purposes was five times higher than for all housing purposes (houses, gardens and yards), while the daily rate of land take for new roads in Germany equalled 23 hectares in the late 1980s (before the German unification and the consequent large-scale road investment schemes for the former German Democratic Republic). Apart from the land consumed directly, roads affect the quality of much larger areas of urban space, landscapes and farmlands due to the partitioning and depreciation of neighbourhoods, recreation areas, silence areas, farmlands and wildlife habitats adjoining the infrastructure (with several disturbance effects reaching up to several kilometres from the infrastructure itself). Finally, the land-use of roads is not evenly-balanced, but concentrated in urban areas where land is the most scarce, as noted by Burchell (1998).

- *Loss of local control* In his vivid and convincing plea on this topic, White-

legg (1997) introduces the loss of local control as one of the often overlooked but main socioeconomic cost components of large-scale transport investments. Under this cost component, he lists the giving away of local control on production and consumption, made possible by underpriced transport costs. This causes local economic welfare as well as local employment to become the responsibility of increasingly large, distant bodies that are less likely and less capable to defend local interests, while at the same time not being accountable for the local consequences of their policies. The same argument is made at the urban level by Knoflacher (1996) who in addition notes that although centralising is often defended as cost-saving for consumers, this generally overlooks transport costs that are rolled off to consumers.

- *Social alienation* The work of Appleyard (1981) has shown how traffic annoys and disturbs residents and disrupts social interaction at the level of the individual street. Significant effects were found of traffic density on the number of contacts and the sense of local responsibility. As one of the consequences, people on streets with heavy traffic were found to have three times as few local friends and twice as few acquaintances. Appleyard concludes (quoted from Whitelegg, 1993):

  > there was a marked difference in the way these streets were seen and used, especially by the young and elderly. [Streets with light traffic housed] a closely knit community whose residents made full use of their street. The street had been divided into different use zones by the residents. Front steps were used for sitting and chatting, sidewalks for children playing, and for adults to stand and pass the time of day ... [Streets with heavy traffic], on the other hand, had little or no sidewalk activity and were used solely as a corridor between the sanctuary of individual homes and the outside world. Residents kept very much to themselves. There was no feeling of community at all.

  Another report on this topic (here termed 'sociability' effects) is offered by Untermann and Vernez Moudon (1989). Yet another viewpoint is offered by Schaeffer and Sclar (2003) who offer an extensive and well-readable cultural analysis of the alienation and social crisis that stems from what they refer to as the automobile era.

- *Other costs* Other significant external costs by the transport system include: roadside pollution control and clean-up, fuel provision externalities (environmental damage from drilling and processing petroleum, spill-overs and evaporation, explosion risks, land-take of fuel stations and refineries), opportunity and interest costs (due to the considerable fund absorption by the transport sector), car industry life-cycle costs (air, land and water pollution by auto manufacturing, scrapped vehicles and tyres), increased energy dependency (leading to economic vulnerability and considerable military, strategic and political costs to safeguard supply) and

> the pollution and increased salinity levels from chemicals to keep roads free of ice. A good tabular overview of all environmental effects is offered in Banister (2005), including monetary estimates.

A large problem in dealing with the external costs of the transport system is that these costs are often not perceived as real. They are borne by the society at large and as such seem imaginative to many. In addition, each nation has their own lobby groups from car industry, freight transport sector, the road construction sector, aviation industry and motorist and touring clubs, which invest considerable resources to talk external costs down and use population dissatisfaction and frustration on low accessability, high transport costs, low safety and even environmental concerns, to promote their solutions: new road schemes, highway widening programmes, increased parking opportunities and increased (direct or indirect) subsidising of the road and air transport system. As examples of such lobbyist activity, see (Neumann, 1997) or (AASHTO, 1998) and (Jacoby, 1999) -commented in (Litmann, 2004c)- or six other articles individually commented in Littmann's article on this topic (Litmann, 2004a). An interesting historical account on the roots of the dealing with external costs from the car is offered by the excellent work of Mrazek (2002) on the role of transport technology in the colonial Dutch Indies.

External benefits of transport systems The external benefits of transport systems, thus your neighbour profiting from you travelling more often, are exclusively found in extended economic development. Increased transport options extend the number of people that have access to a specific service or supply of goods given a fixed amount of time or a fixed amount of cost. This can help the markets and facilities for these goods and services to pass threshold levels of minimal activity, and to become economically feasible as a consequence. This in turn offers jobs and new economic possibilities for your neighbour as a result of you driving more often. In its many variants, this is the fundamental economic rationale behind transport policies. However, important and far-ranging as this effect may be, considerable critique has accumulated on the linear extension of this principle, in particular for regions or nations with an already well-developed transport infrastructure. Many studies have revealed an historic, statistical relationship for the increase in transport volumes under a climate of economic growth. However, many have criticised the viewpoint that this statistical relationship should be turned into a causal relationship, in other words, pursuing economic growth by facilitating transport flows. As an example, Banister and Berechmann (2000) conclude their extensive research on this issue by stating:

> Our conclusion is that there is no reason why we should have transport growth in line with economic growth. Indeed, there are strong efficiency and environmental arguments for breaking that link (Banister and Berechman, 2000).

And yet, the historical, statistical link between transport growth and economic growth is used as exogenous input to many transport models used in practice.

Thus, many transport models simply *assume* that in line with a projected or desired economic growth, car ownership and transport demand also grows by a certain percentage (instead of predicting car ownership and transport demand under a given economic climate). This practice, in conjunction with the tendency in transport to provide for the transport that is predicted by the transport models -the so-called predict and provide paradigm- results in a self-perpetuating linkage between transport and economic growth. Instead of assuming historical correlations, a well-founded analysis of the external (economic) benefits of transport systems needs to explicitly model this relationship. Effects that need be included in such analysis, often ignored by current transport planning practice, include distributional effects, induced travel demand, changed production strategies and changed land-use patterns that result from transport investments:

- First, *distributional effects* are all economic effects that occur at a distance from where a transport policy is applied. Although it is undisputed that the opening or extension of an airport creates new job opportunities at and around the airport itself, how about job opportunities at other airports, how about jobs in competing transport sectors (buses, trains, car rental), how about jobs in the domestic tourism sector as the result of an increased net outflow of tourists, how about jobs of industry and offices that move from their original location to the new airport or would otherwise have opened elsewhere? Such distribution effects are typically difficult to quantify, as they are heterogeneous in nature and they spread over a very large area outside the study area itself. In addition, the effects are typically distributed over relatively isolated actors, each of whom is affected to some extent, but generally none of them heavily enough to generate sufficient economic and political backing to oppose the much more visible direct gains in the study area itself. Often, such peripherical economic downturn isn't perceived as somehow related to transport investments elsewhere, but ascribed to general factors such as the general economic climate, general politics, changing lifestyles or ironically, to the bad accessibility of these locations.

- A second effect, that of *induced travel demand*, directly threatens the above-discussed fundamental economic rationale of transport policy. If increased transport options make people change their activities such as making longer leisure trips, buying a house further from work, sending children to schools further away, the increased transport capacity is quickly consumed without contributing to what it was meant for: increasing accessibility of goods and services. Thus, although induced travel demand may be desirable for individual persons who choose to make use of the new transport options, it does not offer any external benefits.

- Finally, *changed production strategies and changed land-use patterns* refer to the reaction of economy and society on the new transport realities by new transport policy. First, for the economy, increased transport op-

> tions offer the opportunity to centralise production or allow individuals to move to a house further from work. Both processes increase transport flows, but in doing so, do not deliver external benefits to society: the fruits are enjoyed by the centralising company and the moving individual alone. In contrast, these processes actually cause external (equity) costs rather than benefits: companies which do not centralise, for example local groceries, loose customers to the new, large shopping malls made possible by the transport investments. The emergence of 'just-in-time' (JIT) production and distribution strategies is a clear example. The JIT-strategy offers benefits to firms by reducing the stocks of materials that need be stored. This production strategy however is accompanied by a larger demand on the transport system by requiring many small freight deliveries instead of several larger batches. In this example, what is sold as external economic benefits of the transport system, effectively is a subsidy to those industries that adopt JIT-strategies and other transport-extensive production methods at the expense of their competitors. Second, for society, Knoflacher (1987) falsifies the common belief that transport improvements lead to societal time gains in his extensive work on induced travel and changing activity patterns. Instead, he shows them to lead to urban sprawl and explains why this leads to cities that are less attractive for its inhabitants in the long run.

As a result of the systematic negligence of the above effects, strong critiques have been made on the current modelling of transport-related external benefits. A first example is offered by two expert reports from the Standing Advisory Committee on Trunk Road Assessment of the Department of Environment, Transport and Regions in the UK. Both reports represent some of the few serious government efforts to make an objective and high-level assessment of the true economic development effects of highway investment:

> These studies demonstrate convincingly that the economic value of a scheme can be overestimated by the omission of even a small amount of induced traffic. We consider that this matter is of profound importance to the value for money assessment of the road programme ... [Induced traffic is of importance in particular] where the road network is already congested; where travel behaviour has a high potential for change; where a scheme causes big changes in travel costs. In practice this means that traffic induction is most likely to occur on roads in and around urban areas, river crossings, and capacity-enhancing inter-urban schemes including motorway widening (SACTRA, 1994)

and in the second SACTRA-report:

> The available evidence does not support arguments that new transport investments in general has a major impact on economic growth in a country with an already well-developed infrastructure. ... We

> do not accept the results of macroeconomic studies which purport to identify very large returns from infrastructure investment. We are at present unpersuaded by the size of the impact of transport on jobs claimed by a number of European studies (SACTRA, 1999)

In addition, in his account on this topic, Litmann (2004c) reviews seven other critical reports on the external benefits delivered by highway investment schemes. One of these critiques notes how the annual rate of return from highway investments has dropped under that of private investments from the 1980s onwards and is expected to decrease further because the most cost-effective investments have already been made (Nadri and Mamuneas, 1996). Two other reports confirm the above-mentioned distributional effects of transport projects. First, the Dutch Central Planning Bureau concludes that

> much of the economic impacts that occur when highways are upgraded tend to be economic transfers, economic activity shifted from one location to another without overall gain. (CPB, 2002)

In another research, Boarnet and Haughwout (2002) conclude that increased highway capacity often provides economic benefits in one part of a region at the expense of other areas. Finally, Preston (2001) reviews several studies on the limited external economic benefits of transport investments, both at the national level (Fogel, 1964), as at the local level (Nelson et al., 1994). Summarising, the combination of the above quoted research indicates that the current modelling of the external benefits of transport systems is ill-founded. The next paragraph deals on the implications of this conclusion for contemporary cost-benefit analysis of transport systems.

Transport policy cost-benefit analysis As it is powerfully illustrated by the work of Hardin (1968), man cannot be expected to incorporate the external costs of a behaviour in his decisions as long as he is still profiting from it at an individual basis. This feature creates a fundamental role for the state to intervene in transport projects, even in case it would not be the prime funder of transport investments. Analysis frameworks to assess the net economic impacts of transport policy include integrated land-use transport models, benefit-cost analysis, input-output models, economic forecasting models, econometric models, case studies, real estate market analysis and fiscal impact analysis (Weisbrod, 2000). Particularly in the Netherlands, much effort has been made to assess all societal effects of transport infrastructure in a uniform, monetary framework by the *Overzicht Effecten Infrastructuur* method (OEI) (CPB/NEI, 2000). However, despite this array of analysis techniques available, their outcomes are as inclusive as their input data is. In particular, contemporary cost-benefit analysis for transport is criticised for structurally overestimating transport benefits while underestimating construction costs and for underestimating the size and scope of external costs. The influential critiques of the UK-governmental SACTRA committee quoted on page 9 offer first indications for this practice, further supported by the reports discussed in (Litmann, 2004c). As another example,

but on the application of cost-benefit appraisal in developing countries, Wright concludes

> Cost-benefit studies have as a rule focused on individual projects, neglecting the overall picture. They have often overestimated the benefits and underestimated the costs of bad projects and neglected to formulate better alternatives. Considerable mathematical skill has been employed to produce unrealistic predictions of the number of trips and the value of travel time that would supposedly be saved by some proposed boondoggle. The number and values give the impression of being sufficiently elastic to justify whatever is being proposed. (Wright, 1992)

In another critique on current transportation planning practice, addressing the limited attention for transport-incurred external costs, Banister states:

> ... if we are serious about achieving targets for reductions in emissions, improved safety, and a better quality of life in cities, then these issues must be fully reflected in the decision-making process. In many cases, only limited value seems to be placed on environment, accidents, and quality factors. Yet increasingly, these are the issues about which people claim to be most concerned. The transport planning process [however] is still driven by the desire to [primarily] reduce travel time and cost. (Banister, 2002)

It is concluded that although much research has been done on the external costs of transport systems, their external benefits are not well-known. It has been shown how transport-related benefits are often only predicted by extrapolating historical statistical relations between economic development and transport growth. However, it was shown that this relationship is not realistic for well-developed transport systems and in addition, the relationship was shown to be based in part on self-perpetuating transport planning practice. As a result of this failure to properly identify transport-related benefits and the consequent ease to overestimate these benefits or to underestimate competing costs, much of the work on the external costs of transport systems is effectively rendered of secondary importance in influencing the public and political decision making at the end of the day.

1.2 Travel diary data

The previous section argued that in contemporary transport cost-benefit analysis, the analysis of the (external) benefits of transport systems is much underdeveloped. It is argued that this is a direct consequence of the general absence of detailed data on the actual, socioeconomic performance of transportation systems. Conventional transport models that are used to assess transport system performance only generate aggregate data such as the amount of trips

facilitated by the transport system and their average length. However, for a well-founded benefit analysis of transport systems, it is indispensable to have data available on the trip contexts of the *individual* trips facilitated by the transport system. A well-founded benefit analysis would require knowledge on the population segments served, on trip purposes and departure times of each trip or on the larger trip chain where each trip is part of. Such detailed data is often collected by means of travel diaries. Let such detailed trip data be referred to as travel diary data (TDD), defined as:

> information on the consecutive trips performed by or expected for an individual person during a predefined, fixed time interval, in particular including the origin and destination locations, trip purpose and main transport mode of each trip

This section starts with a discussion on the typical usage of travel diary data. The remainder of this section mainly focuses on household travel surveys (HTS), being the principal method to obtain travel diary data. First, the survey type is introduced in brief. Next, its costs are discussed, followed by a discussion on the underreporting of trips in travel diary data by household travel surveys. Finally, various emerging data collection techniques are reviewed on their potential to collect travel diary data in extension to household travel surveys.

Usage of travel diary data Travel diary data is the principal data used for the calibration of travel demand models. Generally available for a small sample of the total population only, travel diary data is for example used to obtain insight in aggregate transport system performance indicators such as the number of trips made per person per trip purpose and VMT per person per day. However, with Mannheim (1979) noting that 'the best sources of information on current transportation issues are newspapers, news magazines, and transport-related journals', it is remarkable to note that most of this popular debate on transportation issues is hardly based on any travel diary data. In comparison; magazines, newspapers and news shows offer daily analyses of the number of people who would vote for each political party, who are unemployed, who visited a rock festival last weekend, who bought the new book, who have a certain opinion, often based on relatively small samples. Such analyses generally not just provide the number of people who show this behaviour, they also show what are the characteristics of all these people. The individuals voting for party X, how many previously voted for Y? Which are the income groups of the parents of the new first-year students for this year? What is the average family size of people who spend their holidays in their own country? Marketing agencies have good insight in the statistical inference of many personal characteristics on a consumer's interest in almost any perceivable product. Compared to these surveys, travel diary data seems much underrated when considering the key role of transport in our societies and the relatively large samples generally obtained.

Household travel surveys Travel diary data is generally collected by means of a household travel survey (HTS). For this report, household travel surveys

are defined as

> a survey of randomly selected respondents whose travel behaviour is queried by means of a travel diary

Although these surveys can be targeted at single individuals only, they are generally performed for all members of a household at once. The variants of the survey type mostly differ on their contact method. Contact can be made in the field or on-board, by fixed or rotating respondent panels, by mail or telephone or by derived forms such as Computer Assisted Telephone Interviewing (CATI). Reflecting the various backgrounds of household travel surveys, the respondent is either asked to report the main activity or time-use for subsequent time intervals (activity diary data) or to report trips only (trip diary data). New trends are the computer-assisted survey methods, an example of which is the virtual reality and GIS-based logging system (VIRGIL) discussed in Janssens et al. (2004). Two important aspects of travel surveys are respondent burden and respondent privacy. The high burden typically associated with travel surveys prevents many individuals from participating in the survey (Ampt, 2003), whereas privacy aspects prevent the publishing of detailed trip data and data on the respondent's residential location. This latter aspect requires survey agencies to publish their surveys at considerable data aggregations or to deliberately introduce distortion in their data-sets, for example discussed by Gouweleeuw et al. (1997) and Willenborg and van den Hout (2000). The considerable consequences of both practices for transport modelling are listed by Hague Consulting Group (1992), including in particular the impossibility to calibrate transport models for urban settings or small travel regions. Protected travel diary data can no longer be used to monitor transport at these lower geographic aggregation levels, although the vast majority of travel manifests itself at these levels.

Sample size The calibration of transport demand models requires very large samples of travel diary data. In the work of Bruton (1985), minimum values are proposed that range between a 1% coverage for populations over 1 million inhabitants and a 10% coverage for populations under 50,000. Although the work of Ortúzar and Willumsen (1995) proposed to use lower sample sizes by making use of intelligent sampling strategies, relatively large sample fractions seem nonetheless inevitable, especially for smaller populations (and thus for small study areas). In addition, the work of Stopher et al. (2001) and Kalfs and Evert (2003), identified several trends that are likely to reduce both response rates and the accuracy of responses. These trends are thus expected to require a higher number of individuals that need be contacted for obtaining an equivalent sample size:

- Increasing reluctance of people to participate in surveys in general, due to their subjection to many other surveys and marketing approaches by various organisations.

- Replacement of face-to-face interviews by telephone, CATI and mail-based surveys. These survey methods are known to result in significantly lower response rates and reduced response quality due to the lack of personal contact.
- Increased difficulties to reach individuals, driven by diverse factors such as the increasingly mobile personal lifestyles (making it harder to find respondents at home), less free time to participate in surveys, increased number of immigrants who speak different languages and the spread of telephone screening devices enabling people to avoid calls they do not wish to receive.

It is argued that the fundamental cause why travel surveys require large samples and are thus that expensive to perform is the intrinsic location-specific character of important travel behaviour aspects, in particular the departure and arrival locations of each trip. As a result, household travel surveys obtained for area A are of no use for area B. Thus, sharing a required minimum sample size over area A and B is no option, as each area needs a full sample for itself. It is noted that this location-specific characteristic of travel surveys makes them fundamentally different from other surveys such as opinion polls and consumer surveys. For these latter surveys, standard techniques such as extrapolation and population stratification can be adopted to use observations in one location for predicting behaviour in others.

Costs of travel surveys Travel surveys are expensive to perform, mostly because of the large sample sizes involved. As an indication, table 1.2 provides the reported costs on consultant services for various American household travel surveys in the mid 90s (Cambridge Systematics, Inc., 1996), (Greaves, 1998). Based on these costs, Stopher et al. (2001) estimates the full cost for travel surveys at $120 to $175 per completed household at the 1995 US price level.

Area and year of survey	Sample size (households)	Cost of consultant services	Contact method	Diary type
Albuquerque (1992)	2,000	$ 130,000	Phone/Mail	Travel
Detroit (1994)	7,400	$ 800,000	Phone/CATI	Activity
Dallas/Ft. Worth (1996)	6,000	$ 750,000	Phone/CATI	Time-Use
Honolulu (1996)	4,000	$ 500,000	Phone/CATI	Activity
Little Rock (1993)	856	$ 48,000	Phone/Mail	Travel
Salt Lake City (1993)	3,082	$ 300,000	Phone/Mail	Activity
Raleigh-Durham (1994)	2,000	$ 275,000	Phone/On-board/CATI	Time-Use (2-day)

Table 1.2: Costs for some US Household Travel Surveys, source: Cambridge Systematics, Inc. (1996) and Greaves (1998) in Stopher et al. (2001)

Underestimation of trips in travel surveys Apart from the reluctance of a respondent to cooperate with travel surveys in general (unit non-response), he or she can also fail to provide with complete travel information at the trip or

trip detail level (trip item non-response). Reasons for this problem are offered by Wolf et al. (2003), Brög et al. (1982) and Richardson et al. (1996):

- Lack of care in survey completion (e.g. forgotten trips)
- Skipping of trips due to false assumptions on which trips need be reported (e.g. elimination of trips that are considered too short, too unimportant, redundant or because they are made by a non-motorised travel means)
- Deliberate skipping of trips (e.g. unwillingness to report some trips)

The result of this non-response is an underestimation of the true number of trips. For car trips, André (1997) quantifies the underestimation by comparing a diary survey with an instrumented study that used vehicle equipment to report vehicle trips automatically. The number of instrumented vehicles in this study amounted to 56, trip data of which was recorded for 35 days on average. The results are shown in table 1.3.

Trip length segment (kilometres)	Diary survey % of trips	Instrumented study % of trips
$<$ 1 km.	9% - 12%	24%
1-3 km.	24% - 27%	28%
3-5 km.	14% - 16%	16%
5-10 km.	18% - 21%	18%
10-50 km.	21% - 26%	13%
$>$ 50 km.	5% - 7%	2%

Table 1.3: Underestimation of trip totals by travel surveys, (source: André, 1997)

It is thus shown that in particular for short trips, trip surveys make a significant underestimation of the number of trips. For long trips, André explains the difference by noting that whereas the diary survey defined a trip by trip purpose, the instrumented study defined a trip by vehicle start-up. A refuelling stop can thus cut one long trip into two shorter trips. Taking such effects into account, the underestimation by travel surveys of the number of trips under 1 km. was estimated at around 40%. In a second study by Wolf et al. (2003), similar comparisons were made based on 523 instrumented vehicles. This study similarly showed a 27.4% underreporting of trips. Although this percentage was not specified for various trip segments, it was shown that for households with more than 10 observed trips, the underreporting rate rose to 43.8% on average, compared to 20.4% for households with less observed trips. This again indicates that short trips are much more likely to be underreported. Various other research on this issue, summarised in Arentze, Timmermans, Hofman and Kalfs (2000), also indicates a general underreporting of trips by household travel surveys. This latter paper concluded that underreporting is most significant for off-peak trips, for non-homebased trips of short duration, for discretionary trips, for non-vehicular trips and for walking segments within trips. As a conclusion, predictions for these trip segments should be used with

care if they are based on household travel surveys, and preferably based on additional research.

Emerging technologies for obtaining travel diary data Automated survey methods that use mobile phones, Global Positioning Systems (GPS), Personal Digital Assistants (palm-top computers), electronic tolling systems and smart-cards offer innovative ways to both contact and query respondents for travel diary data collection. A review of such techniques is offered for example by the well-illustrated paper of Wermuth et al. (2003). Because the availability and cost-effectiveness of these techniques is rapidly improving and when considering the trends on page 13 that increase costs and reduce the reliability for traditional household travel surveys, such emerging techniques may seem likely to replace the large-scale collection of travel surveys in the coming decades. However, although an integrated usage of these survey methods is likely to counter the increase in sample costs per respondent, the technologies are not expected to replace traditional survey methods. This is because these new survey methods typically offer less possibilities to query respondents on information that cannot be recorded automatically (trip purpose etc.) and because not everyone is willing or capable to be surveyed via these new methods. As a conclusion, emerging technologies are expected to offer large increases of specific trip data for some trip segments in the short term, but are expected to remain complementary with traditional travel surveys. Furthermore, whereas these techniques may reduce the survey costs per respondent, practical reasons and privacy considerations are likely to prevent anything near of full coverage of all individuals in a given study area. The sampling nature and the location-specific character of travel surveys (see page 14) thus remain in place.

1.3 Synthesis of travel diary data

Not all applications of travel diary data require this data to be of an empiric nature. In studies where only an initial estimate of travel behaviour is required or where a full (and therefore unfeasible) coverage of inhabitants is required, it is often acceptable that the accuracy of travel diary data is lower than that of empirical data as long as no *systematic* biases are introduced. This leads to the idea of synthesising travel diary data. Here, the synthesis of travel diary data is defined as any method that generates (estimations of) travel diary data other than direct measurement. Section 1.4 reviews the two main techniques for this modelling activity. First, this section discusses the reasons for such data synthesis and the different complexity levels for which travel diary data can be synthesised.

Reasons for the synthesis of travel diary data The principal reason for the synthesis of travel diary data is its low cost. Despite its lower accuracy compared to empirical data, synthetic travel diary data offers a cost-effective solution for studies where no empirical data is required, examples of which

are given above. Other advantages of synthetic data, although of a secondary nature, include:

- *Elimination of privacy concerns.* Confidentiality of individual respondent data is of key importance in maintaining high response rates in the short and the long term. Because individual respondents are easily traced from detailed travel diary data, it is required that such data includes some form of data aggregation or deliberate data distortion (see page 13). For synthetic data, publishing could go without such data aggregation or deliberate distortion.

- *Elimination of high probability factor on local data observations.* Due to its high cost, travel diary data is generally obtained for a small sample of the population only. Whereas this observed data represents the true behaviour of the sample, this behaviour cannot be generalised for small trip segments or population segments due to the high chance factor in obtaining the sample. For synthetic data, travel behaviour can be generated for all individuals in a study area, thus eliminating the probability factor of small trip segments.

- *Analysis of travel survey data in GIS environments.* By the ability to generate travel behaviour where it is required instead of only having it available where it was sampled, travel diary data can be integrated directly into Geographic Information Systems (GIS). These analysis systems typically require information for all analysis units defined (e.g. small zones or individual persons).

- *Direct availability of aggregate output data.* Once synthetic travel diary data is generated for each individual in a study area, it allows for all imaginable selections and aggregations on this data without further concern of weighing or extrapolations as it is required for sample data.

Complexity of the synthesis of travel diary data Three levels of complexity can be distinguished for the synthesis of travel diary data. This level of complexity is related to the degree in which data elements to be synthesised are location-specific:

- *Non-location-specific travel diary data* Travel diary data that has been shown empirically *not* to correlate strongly with location-specific characteristics are the amount of trips made per day and the total travel time per day. This implies that the location where such data is observed or synthesised for, is not of relevance. This characteristic enables a simple generation of synthetic data for these aspects. However, it should be noted that synthetic data on these aspects is probably of low interest, as observed data already gives good insight in these aspects.

- *Low location-specific travel diary data* Most travel diary data is location-specific, at least to a limited degree. An example is the mode choice for a

trip, which is influenced by the local provision level of infrastructure. The train is unlikely to be selected from a location without a railway station nearby. The degree in which such travel diary data is location-specific is nonetheless termed as low, because in case two different locations can be found with roughly equal infrastructure provision levels, these different locations can be treated as equal. Depending on the rigidity by which conditions are defined as equal, this allows for the definition of a range of independent contexts where such travel diary data aspects can be observed or synthesised for in interchange.

- *High location-specific travel behaviour data* An aspect of travel diary data where the above approach cannot be followed is trip destination choice. Regardless of the similarity of two locations in terms of infrastructure provision, urbanisation level or geographic surroundings, trip destination choices obtained for two different locations can never be considered as comparable except if they are (very) near. Whereas a trip mode choice for a train can be explained rather well by noting the infrastructure provision of a location and the characteristics of a respondent, the interpretation of a trip destination choice for Amsterdam requires knowledge of the precise location where this choice was made (was Amsterdam selected from a zone nearby or from a zone far away?). This characteristic makes each location unique for trip destination choice.

The degree in which a travel diary data variable is location-specific, determines the complexity of the synthesis of this variable. The reason for this is that the higher this location-specificness, the higher the number of choice contexts that need be modelled. In addition, a higher number of independent choice contexts reduces the amount of observations that is available in a given database for each choice context. A low amount of observations per choice context results in limited possibilities for the design of data synthesis models. This report proposes as a solution to this fundamental problem by the *removal* of the location-specific nature of location-specific travel diary data prior to its modelling.

1.4 Techniques for the synthesis of travel diary data

This section reviews various techniques for the synthesis of travel diary data. For this purpose, the section first introduces some conceptual terminology according to which these techniques can be classified. Subsequently, specific techniques are reviewed on their existing or potential application for the synthesis of travel diary data.

Behavioural modelling versus data-driven algorithms In a broad sense, 'modelling' may be used to describe any computational procedure in which output data is generated from some available input data using a list of assumptions

on how this input data is related to the desired output data. This report will use the term in a more narrow sense, and reserve it for *behavioural* modelling. Behavioural modelling refers to a modelling effort in which not merely the exogenous (input) data is related to exogenous (output) data, but where this occurs with a detailed consideration of the underlying behavioural processes. This approach leads to an improved insight in the modelled phenomena and ideally results in models that are applicable under conditions that are fully different from those where the model was designed for. In contrast, this report adopts the term *data-driven* algorithms for the often simpler, yet less robust, models that link endogenous to exogenous variables based on observed data only. Because this approach often leaves out relevant variables or effects, it may only be put to use under conditions that are marginally different from those where a data-driven model was designed and calibrated for. As an illustration of the difference, consider stock market analysis. One school of analysts heavily relies on graph analysis of stock exchange prices and stock market indices, a pure data-driven technique. Based on such historical analyses, forecasts are made on future trading. In contrast, another school relies preferably on analysis of the economical performance of companies, markets and economies that underly stock trading, a behavioural approach. The difference between both approaches is clearly illustrated by observing the extrapolation school often making the better short-term predictions on market directions thanks to their straightforward models, sometimes to the frustration of technical analysts. However, it is generally the behavioural school that does much better in predicting long-term market conditions and major market changes. For current transportation practice, the focus is mostly on data-driven techniques, with some notable exceptions in research discussed later in this section. However, it should be noted that this model distinction characterises more the model development process as the nature or even the quality of the resultant models. This is easily proven by noting that in case a data-driven model maker is coincidentally offered precisely the same data as collected by a behavioural modeler, the outcomes are identical (assuming that both researchers deploy the same modelling skills).

Descriptive versus predictive models Another traditional model distinction is that between descriptive and predictive models. In his critical retrospective on the early development of transport models, Supernak (1982) labels the first as mostly useful, the second as the true core of interest. However, again it can be argued that although this model distinction is somewhat artificial: in case we would have a perfect description of the transport system, we would also understand how it changes under varying circumstances. The model distinction is therefore more one of degree: the degree in which a model captures all phenomena of relevance determines whether the model may be used for 'predictive' usage apart from a mere 'descriptive' usage.

Aggregate versus disaggregate modelling Yet another traditional modelling characterisation is that between aggregate and disaggregate models. Of-

ten strongly correlated with the behavioural and data-driven distinction, the aggregate-disaggregate distinction refers to the unit of analysis in modelling. First, aggregate modelling is a modelling technique that relies on aggregate (input) variables. Aggregate variables are not observed directly, but represent aggregations of individual observations instead. Examples of such aggregate variables for transport modelling include zonal population, zonal employment or zonal trip production and attraction. These variables do not represent true behavioural units as how they are observed. Instead, they are determined by our modelling assumptions only, like that of the zone scheme imposed. Apart from travel demand modelling, the technique has found application for price-elasticity models, value-of-time studies and travel-cost-based mode choice models, a general critique of which is offered by Preston (2001). In contrast, disaggregate modelling is primarily based on variables as they are observed for behavioural units in the field, without further aggregation. In the early years of transport modelling, an era with limited available computing power, disaggregate models were impossible to compute in practice. However, with the explosive increase of computing possibilities in the last decades, much of the rationale for aggregate models seems to be lost. Yet, in cases where limited observations are available for a specific process under varying conditions, aggregate figures may prove more robust as working with disaggregate input, because these are less prone to chance factors. Considering the identification of a low number of observations per location as the fundamental problem for the synthesis of trip destination choice data (see page 18), aggregate techniques can be justified here.

Activity based approach Using the classifications of modelling techniques discussed above, various modelling approaches are now discussed as they are currently used for transport modelling. In particular, all approaches are judged on their actual or potential ability to synthesise travel diary data. The first such approach is the activity based approach (ABA). The approach has been developed from the late 1960s onwards and stems from the disciplines of geography and urban planning. In retrospective, starting points of the approach are generally considered the work of Chapin (1974) on driving forces for individual activity participation and that of Hägerstrand (1970) on space-time constraints that limit the available activity space. A first comprehensive effort to use the activity based approach for travel behaviour modelling was made by Jones et al. (1983). A multitude of research efforts followed, for example summarised by Ettema (1996) and Timmermans (2000). The core of these ABA modelling applications has been formulated as

> the analysis of travel as daily or multi-day patterns of behaviour, related to and derived from differences in lifestyles and activity participation among the population (Jones et al., 1974)

The activity-based approach has been termed the only paradigm shift in the history of transport modelling, not only by providing a new framework for the analysis of travel behaviour, but also through the emphasis on understanding

travel behaviour (the why question) as opposed to the prior focus on forecasting (the who, what, where and how many of trips) (Pas, 1990), (Pas and Lee-Gosselin, 1997). Whereas the approach was successfully adopted for segments of travel behaviour, one of the few operational, comprehensive model for the synthesis of full activity patterns and derived travel behaviour today is ALBATROSS (Arentze and Timmermans, 2000). The model is discussed in more detail on page 23 as a possible technique for the synthesis of travel diary data.

Regression techniques Regression techniques attempt to correlate a set of exogenous variables directly to the endogenous variables of interest by means of standard statistical techniques. The so-called four-step modelling approach (FSM), the mainstream in travel demand modelling and indeed 'the entire institutional environment in transport practice' (McNally, 2000), can be classified as a regression technique. Key to the application of regression techniques is the availability of a large number of observations on what may be considered the same phenomenon. As it was discussed on page 18, it is precisely the unavailability of a large number of observations that constitutes the main problem for the synthesis of location-specific travel diary data. Thus, the application of regression techniques does not offer a useful framework for this modelling activity.

Discrete choice modelling Discrete choice modelling refers to the modelling of choice from a set of mutually exclusive and collectively exhaustive alternatives. The technique is of high importance in disaggregate transport modelling as it was extensively shown in the influential book of Ben-Akiva and Lerman (1985). More a computational modelling technique rather than a true modelling approach, it is easily integrated into other modelling approaches, including the behavioural activity-based approach as it was shown by the daily activity model of Bowman and Ben-Akiva (2000). Essentially, discrete choice modelling is based on the derivation of utility assignments as made by individual persons as they implicitly follow from observed choice behaviour for a large number of (identical) choice contexts. As such, the technique is basically a data-driven technique, although it can be applied within behavioural modelling frameworks as shown above. Most aspects of travel diary data representing the choice between a set of mutually exclusive and collectively exhaustive alternatives (e.g. trip destinations), discrete choice modelling is of interest for the synthesis of such data. In addition, discrete choice modelling offers ample possibilities for assuming two choice contexts that are different in location but equal in the nature of their choice problem to be defined as equal. For these reasons will the synthesis of travel diary data in this report be based on discrete choice modelling techniques.

Artificial Neural Networks Artificial Neural Networks (ANN) or simply neural networks have first been introduced as a computational simulation of biological neurons (McCulloch and Pitts, 1943). Similar to biological neutron

networks, the technique is capable of adjusting itself when fed with sufficient examples of cause and effect (Supervised Learning). In its modelling essence, ANN represents a special form of regression models that automatically searches for the optimal regression fit. The models have been found to perform particularly well in cases where many correlating exogenous variables exist. The adoption of the technique for transport modelling is for example proposed in Openshaw and Openshaw (1997) and extensively reviewed in Tillema (2005). However, for the synthesis of travel behaviour data, the technique is of little direct use, again in light of the fundamental sample size problem (see page 18). This modelling technique solely being based on sufficient training material, the low availability of observations per choice context rules out its application for the synthesis of location-specific travel diary data. Nonetheless, it is considered that the *output* of this report, synthetic databases of travel diary data, may well be used to feed neural networks in follow-up research.

Knowledge Discovery in Databases Knowledge Discovery in Databases (KDD), also known as data mining has been defined as

> The nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al., Fall 1992)

The approach is a general denominator for the usage of machine learning, statistics and visualisation techniques to discover and present knowledge in a form that is easily comprehensible to humans. The approach has been successfully adopted for the fast generation of new knowledge from phenomena on which much (cheap) data was readily available, such as for transaction data, internet usage and the weather. Observing the complexity of many transport models and the amount of data used to calibrate them, one could judge many transportation modelling efforts as data-mining techniques to some extent. However, the difference with a true KDD application is that the latter extracts new relationships from a large amount of (often previously unstructured) observations, rather than using the data to validate or calibrate model relations that are enforced on the data by the modeller. Using this definition, the application of KDD for transportation research has been limited.

Micro-simulation Like the term 'modelling', the term 'micro-simulation' both has a broad and a narrow definition. First, the broad definition merely states that micro-simulation attempts to describe economic and social events by modelling the behaviour of individual agents (Wiemers et al., 2003). Agents in this sense are typically defined as the smallest individual behavioural unit; for transport studies generally individual persons. However, this broad definition does not really distinguish micro-simulation from disaggregate modelling, rendering the term of little practical usage. Instead, this report prefers to use the term in its narrow sense. In this definition, micro-simulation refers to a process of first identifying homogeneous groups within a set of observations. Next, synthetic cases are generated, each of which are assigned to belong to one of these

homogenous groups. Considering that no observed case is more or less likely to occur as others within each homogeneous group (and correcting for this otherwise), it is now posed that for each additional case that would be observed, the chance of this case to be characterised by one of the already observed cases for the same homogeneous group, is equal. Similarly reasoning, each synthetic case may be assigned the observed behaviour of a random observed case of the same homogeneous group, as an unbiased estimate of what would have been observed for this synthetic case. Such a uniform sampling simulation is often referred to as a Monte Carlo simulation, for example discussed in Granger Morgan and Henron (1990). Examples of this modelling approach in transport are the RAMBLAS model on daily activity travel patterns (Veldhuisen et al., 2000) and the so-called HTS-simulation model (Greaves, 2003). Because this latter model had the synthesis of travel diary data as its explicit purpose, this model is considered in more detail in the subsequent section as a reference model. Because micro-simulation requires a relatively limited number of observations for each choice context only, this report chooses a micro-simulation framework for the synthesis of travel diary data.

1.5 Models for the synthesis of travel diary data

This section focuses on the two models that have been identified as the current state of the art for the synthesis of travel diary data. These are first the ALBATROSS model of Arentze and Timmermans (2000), representing the state of the art for behavioural modelling, and second the HTS-simulation model by Greaves, representing the state of the art in data-driven algorithms.

ALBATROSS The ALBATROSS model (A Learning-Based Transportation Oriented Simulation System) is a rule-based activity-based model, designed by the Urban Planning Group at the Eindhoven University of Technology and described in detail by Arentze and Timmermans (2000). The model was developed to explore possibilities of a rule-based activity-based modelling approach and to develop a travel demand model for policy impact analysis. The model structure is that of a production system, i.e. a set of decision rules of an if ... then ... else form (Newell and Simon, 1972). The model adopts the CHAID algorithm (Kass, 1980) to find such decision rules from a detailed, observed travel behaviour data set. In such manner, a model is constructed of how people select and prioritise activities and how these are subsequently scheduled into feasible activity patterns under differing conditions. Using the observations in a cleaned travel diary database collected from 1997 to 2000, the model was trained for generating travel behaviour for the city where activity-travel patterns were observed (Arentze, Hofman, van Mourik and Timmermans, 2000). Later research then compared this output to the performance of several utility-based activity-based models (Arentze et al., 2001). Next, the model was used to generate data for two different cities using the same input data (Arentze et al., 2002). Finally, the model was extended to generate patterns at the Dutch national scale as

well (Arentze et al., 2003). For all these studies, the authors concluded that ALBATROSS offered satisfactory results in comparison to existing transport models. In particular the two latter studies showed that ALBATROSS could be used to synthesise travel diary data. Chapter 8 returns to ALBATROSS and compares it with the synthesis method proposed by this report.

HTS simulation model In his Ph.D. thesis, Greaves (2003) proposed a model to simulate travel diary data for the US from existing household travel surveys data and census data using a micro-simulation approach. The aim was to find an alternative for the costly collection of travel diary data for Metropolitan Planning Organisations (MPO) in the US. For this purpose, Greaves proposed a Monte Carlo simulation from heterogeneous population clusters in the available household travel observations. This model thus adopts a micro-simulation approach as it was discussed on page 22. Next, Greaves demonstrated the model by simulating travel diary data for Baton Rouge, Louisiana, which was consequently compared with an actual HTS performed for Baton Rouge in 1997. Although significant differences were found between both datasets, some data items were replicated well. Because this data had been obtained at virtually no cost in comparison to the empirical data, it was concluded that the HTS-simulation approach is promising:

> The results presented here suggest that this concept of creating simulated HTS data is worth pursuing further. The procedure was able to provide trip rates that were generally comparable to actual data. The other salient trip characteristics (mode, departure time, and reported trip length) were less effectively replicated in the simulation procedure. These discrepancies were attributed to contextual differences between regions (e.g., population size, transit service), which intuitively affect these attributes to a greater extent than trip rates. Clearly, further research is needed to determine how local characteristics can be incorporated in the simulation process to capture these unexplained differences. (Stopher et al., 2001)

Later, the model approach was extended and applied for an Australian city as well, still using US household travel surveys, yielding comparable results (Stopher et al., 2002). However, travel diary data was only generated for non-location specific travel behaviour aspects (see page 17). This report is presented as a continuation of the research work initiated by Stopher and Greaves. As an extension to their HTS-simulation model, this report designs and implements a framework that allows for the simulation of location-specific travel diary data and in particular trip destination choice. This research aim is specified in greater detail in the subsequent section. Chapter 8 returns to the model of Stopher and Greaves in a comparison with the model presented by this report.

1.6 Research questions

In the previous sections, it has been argued that in contemporary transport cost-benefit analysis, the analysis of the (external) benefits of transport systems is hampered by the limited availability of travel diary data at the individual level and for detailed geographic levels. Transport system benefit analysis does not require highly accurate travel diary data for every single individual and thus allows the usage of synthetic data. However, it was shown that the synthesis of travel diary data is complicated by the spatial dimension of travel diary data, in particular trip destination choice. This spatial dimension makes trip destination choice observed in different locations incomparable and thus prohibits the development of a synthesis method for travel diary data on this or on derived aspects. This report provides a method that is able to transfer the spatial dimension of observed travel diary data and, as an application, is capable of generating synthetic travel diary data that can be used for the detailed analysis of the benefits of transport systems. Thus, the research questions to be answered by this report are:

1. How can trip destination choice be characterised independently of the area in which it has been observed and how can such a measure be implemented without making use of behavioural assumptions other than the information in an existing travel diary database?

2. Using such a location-independent characterisation of trip destination choice, how can travel diary data be synthesised that includes trip origin and destination locations, for each individual, at a detailed geographic level and without bias at any aggregation level?

3. What is the level of accuracy of this synthetic data at various aggregation levels?

The subsequent reading guide specifies where each research question is answered in this report and which chapters to read depending on a specific interest.

1.7 Reading guide

This report presents a method for the synthesis of travel diary data that includes trip destination choice. Hereunder, a thematic reading guide is offered for improved reference.

1. **Conceptual model design** Key to the conceptual model design are sections 2.1 and 2.2. Here, it is specified in detail what output data this report intends to generate from what input data. Of equal importance is section 2.5, that specifies a conceptual algorithm for the synthesis of travel diary data (summarised by figure 2.1). This conceptual algorithm is founded on a theoretical framework on the modelling of location-specific travel behaviour data, presented in section 2.4. Throughout the report,

use is made of a framework of auxiliary modelling concepts that is introduced mainly in sections 2.3 and 3.1. An extensive listing of these concepts and of the associated notational scheme is offered on page 221 and 227 respectively. These listings include references to the pages where these concepts or notations are first introduced.

2. **Model implementation** The actual model implementation is performed in chapters 3 and 4. These chapters answer the first research question by specifying a data-driven observation model for destination choice eccentricity. This measure is introduced as a location-independent measure of trip destination choice on page 40. The model specification is clarified by means of the much simplified and therefore illustrative case study Amsterdam-Rotterdam, introduced in section 3.3.

3. **Model demonstration** Chapter 5 offers a demonstration of the destination choice eccentricity observation model by means of the simplified, yet non-trivial case study Interregional Train Travel. The case study implements a full travel demand model for long distance train travel for the whole of the Netherlands based on a very limited set of observations. Implementing the case study step-by-step, this chapter may be a good starting point for those not interested in the detailed model specification in chapters 3 and 4. At the end of the chapter, section 5.6 includes a discussion on the possible applications of the destination choice eccentricity observation model, out of which the synthesis of travel diary data is one.

4. **Model validation** Model validation is performed in chapters 6 and 7. First, chapter 6 implements the conceptual synthesis algorithm of chapter 2 for the complex case study Southern Overijssel, thus answering the second research question. In this case study, travel diary data is synthesised for a total of 658,590 inhabitants, based on the observed behaviour of 111,484 observed individuals (mostly living outside the study area) and for a zone system that includes 3,766 zones. Chapter 7 compares the resulting database with two major external databases in order to answer the third research question. A second validation study is performed in section 5.5, here for the synthetic database generated for the above-discussed case study Interregional Train Travel.

5. **Summaries and conclusions** Each chapter is concluded by a short summary. Finally, chapter 8 draws conclusions, summarises the answers on all research questions, defines potential applications of the synthesis model and recommends future research.

6. **Appendices** As background information on the three case studies, various appendices are included at the end of this report. Three appendices specify algorithms used for the preparation of input data. The remaining appendices offer various excerpts of the used input data and the obtained output data.

1.8 Summary

This chapter has discussed the background of this report and defined the three research questions it intends to answer. First, a comprehensive review was made in section 1.1 of current practice in cost-benefit analysis for transport policy. It was shown that there exists widespread critique on such analysis for either systematically underestimating the varied external costs or for overestimating the external benefits. It was argued that both practices are mainly sustained by the ill-founded analysis of the external benefits of transport systems. This is considered a consequence of the limited availability of disaggregated transport performance indicators such as they could be offered by extensive travel diary data (TDD). Based on this conclusion, section 1.2 offered a discussion on the usage of this data, its collection by household travel surveys (HTS), its collection costs, the measurement failures by this survey type and on emerging trends for the collection of travel diary data. It was concluded that travel diary data is very costly to collect, which prohibits the sufficient buildup of travel diary databases required to de-anonymise existing traffic flows on the benefits they deliver to society. Considering that transport system benefit analysis need not be based on fully empirical travel diary databases, section 1.3 proposed the synthesis of travel diary data as an extension to its collection. Three levels of complexity were identified based on the degree in which travel behaviour variables are location-specific. For the most complex variable, that of trip destination choice, the low number of observations available for each destination choice context was defined the fundamental problem for the synthesis of such data. Considering this problem, section 1.4 reviewed a range of transport modelling techniques for their application on the synthesis of this data. It was concluded that within the data-driven techniques, micro-simulation is well-suited for the synthesis of non-location specific behaviour, and discrete choice modelling for location-specific behaviour. Finally, section 1.5 reviewed two operational models that are considered the current state of the art for the synthesis of travel diary data. These are first the ALBATROSS model by Arentze and Timmermans as a proponent of the behavioural modelling approach and second the HTS-simulation model by Stopher and Greaves for the data-driven approach. The chapter concluded with the formulation of three research questions in section 1.6 and a thematic reading guide in section 1.7.

Chapter 2

A conceptual algorithm for the synthesis of travel diary data

This chapter introduces a conceptual algorithm for the synthesis of travel diary data (TDD). First, section 2.1 specifies the intended output of the algorithm and section 2.2 specifies the input data that is considered available. Next, section 2.3 introduces several modelling concepts used in this report. The separate introduction of these concepts allows for a structured and more compact algorithm specification in this chapter and in subsequent chapters. Section 2.4 introduces a theoretic framework for the synthesis of travel diary data, focussing on the embedding of travel behaviour in the geographic setting in which it is observed. Based on this framework, section 2.5 introduces a conceptual algorithm for the synthesis of travel diary data, referred to as the TDD synthesis algorithm. The algorithm is summarised by a schematic overview on page 44. Chapters 3 and 4 continue on this conceptual algorithm by implementing its abstract components. The algorithm is applied in full for two case studies in chapters 5 and 6.

2.1 Algorithm output

The aim of this report is to design and implement a method for the synthesis of travel diary data. This section specifies what is aimed for: the precise travel diary data composition, its aggregation level and its level of accuracy. Two lists of specifications are offered for this purpose. The first two paragraphs of this section list the specifications where the synthesis method is theoretically designed for. A subset of these specifications, the subsequent two paragraphs list the specifications where the method is actually implemented and validated for in this report. A summary of these specifications is given by table 2.1.

Specification of theoretic travel diary data output On page 24, it was chosen to adopt a micro-simulation approach for the synthesis method proposed by this report. Micro-simulation, as specified on page 22, requires the availability of an existing travel diary database, whereas the technique generates output data equivalent in composition to that of the input database. Thus, as a consequence of the micro-simulation approach, the proposed data synthesis method is potentially capable of generating all travel behaviour variables that are present in an input travel diary database, using identical aggregation levels and classification schemes. Later in this section, a subset of variables is selected at a given classification scheme and aggregation level where the method will be implemented and validated for in practice.

Accuracy of theoretic travel diary data output For the overall accuracy of the above defined output data, it is clear that this can never exceed that of an underlying empirical database. Synthetic data, ultimately always based on empirical data, becomes less accurate *at the level of data observation* when more transformations are performed on the initial empirical data. However, at higher aggregation levels, it is possible for synthetic data to become more accurate if a modelling approach succeeds in joining previously unrelated empirical observations to describe behaviour. In case the resulting gain in accuracy is higher than the loss in accuracy due to the transforming of data, synthetic data on *groups* of observed units can contain more information than the empirical data observed for these groups. Also, it has been argued at the beginning of section 1.3 that for various transport studies, the high accuracy of empirical data is not required. For individual travel diary data, fully accurate travel diary data is even neither expected nor of high practical interest:

- The claim of having fully accurate travel diary data available at the individual person or trip level implies nothing less than an infeasible full observation of individuals, or complete knowledge of the living situation of all individuals, of the detailed socioeconomic structure of all locations, a full understanding of all human behaviour and so on. Even in case either of these were possible, then still should the availability of such highly accurate travel diary data at the individual person level be considered undesirable because it would constitute a serious violation into our private lives.
- Second, the practical interest of transport planning in travel diary data is almost exclusively on the behaviour of *groups* of individuals. Thus, whereas chapter 1 has argued it to be desirable to have an indication of travel behaviour for each individual and for each trip, a high accuracy is generally only of relevance when this individual data is aggregated for considerable groups of individuals or trips. For example; the accurate, detailed attribution of travel behaviour is typically not of high interest for transport planning purposes. Suppose that a group of interest is constituted by two similar individuals A and B, roughly residing in the same area. Now, consider two individual travel diaries being synthesised,

the one representing the behaviour of A, the other the behaviour of B. This suffices for practical transport planning applications. The behaviour need not be attributed to A or B specifically.

Specification of implemented travel diary data output The previous paragraphs have defined the type of output travel diary data to be theoretically equivalent to that of the input (observed) travel diary data. This means that the output specification is dependent on the input provided. It follows that the higher the complexity and level of detail of the input travel diary database used for a validation case study, the stronger this test on the potential application of the synthesis algorithm. This report has selected to work with one of the largest and most detailed travel diary databases available as the input travel diary database for one of its case studies. This database, the annual Dutch *Onderzoek Verplaatsings Gedrag* (OVG) or its successor the *Mobiliteitsonderzoek Nederland* (MON) includes between 60,000 to 140,000 respondents annually. In this database, departure and arrival locations of individual trips is available for zone schemes that represent an average population of a mere 4,500 individuals (the concept of zones is introduced on page 35). The OVG and the MON are introduced in detail in section 6.2. It is the structure of these databases that defines the output data that this report intends to generate and where it is actually validated for. In this output data, origin and destination locations are defined at the very low aggregation level of 3,776 zones, trip purpose by 6 purpose classes and travel mode by 7 main transport modes. Detailed specifications of these classification schemes are given in section 6.1. It was chosen to focus on these variables because they are the most important travel behaviour aspects for practical applications of transportation modelling, while the aspects of origin and destination location are considered the most complex to model (see page 17).

Accuracy of implemented travel diary output Above, the importance of the aggregation level of synthetic travel diary data on the desired level of accuracy has been discussed. In case this aggregation level is low, thus for very small trip segments, time intervals and/or population segments (and ultimately, for individual persons or trips), the required level of accuracy of output data is at its minimum. However, the higher this aggregation level, the higher the level of accuracy that is probably required for the aggregated data. Also, it has been noted that for all aggregation levels, it is important that synthetic data does not contain a structural over- or underestimation (bias) in order to be useful. Thus, the target level of accuracy for this report is specified twofold. First, for low aggregation levels, it is specified that the synthetic travel diary data should allow initial or *a-priori* estimates of travel behaviour that indicate a correct order of magnitude, without having a significant bias. For high aggregation levels, the travel diary data should not offer deviations from reality for a statistically significant number of cases, nor should it contain a significant bias. As an intuitive guess of what should be the threshold value between a low and a high aggregation level, a value of 500 persons or trips

is selected. In chapter 7 it will be approximated what the true value of this threshold value has been for the case study performed and the input data used.

Summary Table 2.1 offers a summary overview of the target output specifications, as well as the associated target levels of accuracy.

	Theoretical algorithm output	Implemented, validated algorithm output
Aggregation level	• One synthetic travel diary for each individual.	
Behavioural variables included	• All behavioural variables included in input travel diary database	• Origin and destination locations • Trip purpose (per trip) • Main mode of transport (per trip)
Aggregation level or classification scheme of output variables	• Equal to those of the input travel diary database	• For origin and destination locations: 3,766 zones • For trip purpose: 6 purposes • For mode of transport: 7 modes
Level of accuracy	• Unbiased estimates; no systematic over- or underestimations • For aggregations of less than 500 persons or trips: a-priori estimate; order of magnitude should be correct • For higher aggregations: no statistically significant deviations from reality	

Table 2.1: Output specification of the TDD synthesis algorithm

2.2 Algorithm input

This section specifies the input data that is assumed available for the generation of synthetic travel diary data with the above specifications. The first database should be an empirical travel diary database, although not necessarily for the study area where travel diary data is to be synthesised for. The second database should specify a synthetic population where travel diary data is to be synthesised for. Finally, a so-called generalised travel cost matrix is required. Because it is considered that the latter two databases are not really available for many study areas, this section proposes alternative databases to be used instead. In appendices A and F, this report includes two algorithms for the generation of the required input databases from these alternative databases. Each input database is now specified in greater detail.

Input travel diary database *TB* The first required input database should contain observed travel diary data, for example the result of an existing household travel survey. Let the symbol *TB* be used to refer to this input database (for travel behaviour) and let its respondents be indexed by r ($1 \leq r \leq N_{r,TB} =$ number of respondents in *TB*). Let the symbol ξ denote the population sample whose travel behaviour is stored in *TB*. This input database should meet the following specifications:

- *Travel behaviour variables* The database should include at least all travel behaviour variables where synthetic travel diary data is to be synthesised for. For the case studies in this report, these are specified by table 2.1. In addition, it is specified that the database should include a consecutive list of locations visited (no intermediate trips may be omitted) and that the data should be internally consistent (discussed separately hereunder).
- *Sample distribution* Although the population sample in *TB* is best randomly sampled, stratified samples or any other probability sample are acceptable. A probability sample is defined as a sample in which each individual has a known probability of being included in the sample (Blalock Jr., 1960). For simplicity, this report assumes that *TB* is based on a disproportional stratified population sample; a sample that is composed out of a list of random samples for each population cluster within a mutually exclusive and collectively exhaustive set of population clusters. The concept of population clusters is further introduced on page 37. Each population cluster that is defined need be represented by a significant number of respondents in the sample ξ for *TB*.
- *Respondent data* Apart from travel behaviour data, *TB* should also include the population cluster of each respondent (see page 37), or alternatively, all (probably socioeconomic) variables required to uniquely assign each respondent to such a cluster.
- *Geographic spread of respondents* The geographic spread of observed respondents over the area where travel diary data is to be synthesised for is not of high interest. The input travel diary data may even be collected outside of this area, as long as the observed locations do not differ significantly in their travel behaviour patterns from those where travel diary data is to be synthesised for.
- *Missing or inconsistent data* Finally, it is considered that empirical travel diary databases often contain missing or inconsistent data entries. Because such missing or inconsistent observations are more likely to occur for respondents who have stated complex travel behaviour, their exclusion would introduce bias in the remaining travel diary database. As a more preferable option, it should be tried to synthesise data for missing data entries and to correct for inconsistent data entries. Section 6.2 introduces a conceptual rule-based algorithm for this task, implemented in detail by appendix E.

Synthetic population *SP* or socioeconomic zonal database *SE* As a second input database, a synthetic population *SP* is required. This database should store a set of individual and/or household characteristics for all individuals in the geographic area for which travel diary data need be synthesised. Let all synthesised individuals in *SP* be indexed by s ($1 \leq s \leq N_{s,SP}$) where $N_{s,SP}$ = number of individuals in *SP*. This synthetic population should allow each individual s to be uniquely assigned to a population cluster. What

variables are thus required in *SP* depends on the population clustering scheme adopted. For the main case study implemented in this report, a population cluster is defined in section 6.2. However, almost independent of which characteristics should be included in *SP*, such data can generally not be obtained at the individual level. Instead, such data generally needs to be synthesised from other data sources. For this purpose, appendix A proposes an algorithm to generate synthetic populations. The algorithm is designed for the synthesis of all individual and household characteristics commonly considered as the most relevant for an individuals travel behaviour, as listed by table 2.2. The input of this algorithm is constituted by a database *SE* with detailed socioeconomic zonal data for the geographic area where travel diary data need be synthesised for. In contrast to synthetic populations, such databases are available for many countries.

Household variables	Individual variables
Household type Household level of prosperity	
Number of adults in household Number of children in household	Age category
Number of retired persons in household Number of (full-time) students in household Number of (un)employed persons in household	Social participation (retired, student, (un)employed, other)
Number of cars in household	Individual car availability

Table 2.2: Travel determinants synthesised by appendix A

Travel cost matrix *TC* or infrastructure database *NW* Finally, a matrix is required that provides the geographic separation between each combination of two zones. This separation is possibly expressed by the associated distance, but preferably by a more inclusive concept, such as generalised travel cost (see page 37). Let such a travel cost matrix be denoted by the symbol *TC*. Because section 2.1 proposed that this report should be capable of handling very fine zone systems, the number of matrix elements in *TC* can be very high. For many practical applications, it is unlikely that travel cost data is readily available for all the associated origin-destination pairs. For such cases, this report proposes an algorithm to generate entire travel cost matrices from a relatively limited infrastructure network *NW* in appendix F.

2.3 Modelling concepts and notational conventions

Before this report introduces two models that allow for the synthesis of travel diary data according to the above input-output specifications, this section first introduces several modelling concepts. These concepts are introduced separately, in order to allow for a more compact model specification and discussion later in this report. Along with the introduction of these concepts, a list

of notational conventions is introduced. This notational scheme is extended throughout this report. An index is included on page 227.

Zones and OD-pairs A basic concept in any geographic study, let a zone be defined as

> a relatively small geographic area that is modelled to have all of its socioeconomic contents located in its gravity centre

Often, not a single zone but the combination of the origin and destination zone of an (imagined) trip is of relevance. For this purpose, let an origin-destination pair (OD-pair) be defined as

> the pairwise combination of an origin and destination zone

Similarly, an OD-matrix is defined as a matrix stating the observed or predicted number of trips for each OD-pair.

Synthesise area As it has been discussed on page 30, it is assumed that practical interest is generally not in travel diary data for separate individuals, but rather in those of considerable groups of individuals. For convenience, this report assumes that synthetic travel diary data is always generated for all of the inhabitants of a given geographic area of interest. Let this area be termed the synthesise area, denoted by $\mathcal{Z}$. Also, let $\mathcal{Z}$ be split up into a number of residential zones. Let each of these residential zones be indexed by z ($1 \leq z \leq N_{z,\mathcal{Z}}$), where $N_{z,\mathcal{Z}}$ = number of zones within $\mathcal{Z}$.

Residence space Let the residential locations of two groups of individuals be aggregated into a number of zones. These two groups are first all observed respondents $r \in TB$ and second all synthetic individuals $s \in SP$ where travel diary data is to be synthesised for. Let such residential (home) zones be denoted by the symbol h and grouped into three zone sets. First, let the zones in which all observed respondents $r \in TB$ reside be termed the observed residence space, denoted by $\mathcal{H}_{\mathcal{TB}}$. Second, let the set of zones in which the synthetic population SP resides be referred to as the synthesised residence space, denoted by $\mathcal{H}_{\mathcal{SP}}$. Although this latter area is geographically equal to $\mathcal{Z}$, a different notation is introduced to allow for a customised zone scheme for both areas. For most case studies, $\mathcal{H}_{\mathcal{SP}} \subseteq \mathcal{H}_{\mathcal{TB}}$. Finally, let the union of $\mathcal{H}_{\mathcal{SP}}$ and $\mathcal{H}_{\mathcal{TB}}$ be referred to as the (total) residence space, denoted by the symbol $\mathcal{H}$.

Population sample Let the population sample for which travel diary data is available be denoted by the symbol ξ. This sample defines how the population composition of the observed respondents $r \in TB$ relates to that of the entire population residing in $\mathcal{H}_{\mathcal{TB}}$. As it was defined on page 33, ξ is expected a probabilistic population sample.

Trip origin and destination space Let the geographic space in which observed or synthesised individuals can travel, be modelled by two independent zone sets. First, let all zones from which these individuals can possibly start a trip be termed trip origin zones. Let such zones be denoted by the symbol i and let the group of all such possible origin zones be referred to as the trip origin space, denoted by the symbol $\mathcal{O}$. Second, let all zones where these individuals can possibly travel to be termed trip destination zones. Let these individual zones be denoted by the symbol j and let the set of all such destination zones be referred to as the trip destination space, denoted by the symbol $\mathcal{D}$. Let the number of zones in $\mathcal{O}$ be denoted by $N_{\mathcal{O}}$ and the number of zones in $\mathcal{D}$ by $N_{\mathcal{D}}$. Both areas $\mathcal{O}$ and $\mathcal{D}$ should basically represent the world at large in order to allow the modelling of all imaginable trips. However, for practical applications, it is only required to define a detailed zone system for $\mathcal{O}$ and $\mathcal{D}$ for the geographic area represented by $\mathcal{H}$ and its vicinity. All locations further away can be represented by a limited zone system of rough (macro) zones. The reason why a separation between $\mathcal{O}$ and $\mathcal{D}$ is introduced is to allow for efficient computation procedures. For the one zone set, a fine zone system may be required for specific sub-areas, whereas this represents a mere waste of computation capacity for the other zone set.

Trip segment This report allows the independent synthesis of trip destination choice data for different trip segments. For this purpose, let a trip segment be defined as

> a segment of trips that is relatively homogeneous in the type of destinations selected for them and the processes by which these are selected

The most apparent trip segmentation is based on trip characteristics themselves, such as trip purpose or travel mode. However, the above definition does not exclude non trip-based characteristics for the definition of a trip segment. Possible examples are characteristics of the trip-maker like age or social participation, time variables like time of the day or day of the week, or geographic characteristics like the urbanity level of the residential location of a trip-maker. Let trip segments be denoted by the symbol m. Also, let a trip segmentation scheme be introduced as an exclusive and exhaustive set of trip segments for a case study, denoted by the symbol $\mathcal{M}$. In fact, whether or not trip segments are indeed relatively homogeneous can only be determined after the processing of a large travel diary input database or the synthesis of an output travel diary database. In theory, this makes the definition of trip segments iterative with the synthesis of output data. In practice, common transport modelling knowledge will generally suffice to assume a trip segmentation scheme that is appropriate for a given case study. This assumption can later be validated by output data. In the definition of $\mathcal{M}$, homebound trips need generally not be included, because the destination choice of such trips is trivial.

Travel diary data notations Here, a list of notations is proposed to refer to elements of the travel diary data reported by respondent r in database TB. Let the set of travel behaviour data reported by r be denoted by vector $\mathbf{B}_r$ where the k-th element is denoted by B_{rk}. In addition, let various travel diary elements be assigned a dedicated notation. First, let the residential zone of respondent r be denoted by H_r. Second, let the number of trips reported by respondent r be denoted by $N_{t,r}$. and let these individual trips be indexed by t ($0 \leq t \leq N_{t,r}$). Also, let the origin zone of trip t of respondent r be denoted by O_{rt} and the destination zone of the same trip by D_{rt}. Finally, let the trip segment of trip t of respondent r be denoted by M_{rt}, the trip purpose by K_{rt} and the main mode of transport by L_{rt}.

Generalised travel cost Let the generalised travel cost of travelling from an origin zone i towards a destination zone j for a trip of trip segment m be denoted by ψ_{ijm}. Let generalised travel cost ψ_{ijm} be defined as

> the sum of all time, money, effort and inconvenience required for a trip from an origin zone i towards a destination zone j and of trip type m, measured in uniform (monetary) terms

Trip segment m can influence travel cost in case it is based on different transport modes. For ease of speech, this report will sometimes use the term *distance* in interchange with generalised travel cost. For the same reason, does this report sometimes prefer terms like *near* and *far* instead of the more accurate but elaborate *low generalised travel cost* and *high generalised travel cost.*

Travel behaviour determinant Let a travel behaviour determinant for this report be defined as

> an easily observable, objective variable that is assumed to have a large influence on the travel behaviour $\mathbf{B}_r$ of an individual person r

Examples of such variables as they are used in this report are age, social participation (employed, student, retired etc), mobility variables (car possession, possession of commutation ticket), status in household (wage earner, other adult household member, in-living child), household or personal income level, household type (single or multiperson household, with or without young or old children) and geography type of the residential location (urban, suburban, rural).

Population cluster Based on a set of travel behaviour determinants, let a population cluster be defined as

> a group of individuals who share one or more characteristics for a set of travel behaviour determinants, and who show relatively homogeneous travel behaviour $\mathbf{B}$

Let the population cluster of respondent r be denoted by P_r and let the mutually exclusive and collectively exhaustive set of population clusters for a case study be referred to as the population cluster scheme, denoted by $\mathcal{P}$. It is noted that a cluster scheme fully specifies what variables should be generated for the synthetic population *SP* (see discussion on page 33).

Geographic setting Travel behaviour **B** being a geographic phenomenon, many of its aspects are only meaningfully when interpreted for the spatial context where they take place. For example, the observation of a trip to a city 100 kilometres away is probably much less eccentric behaviour if it is the nearest city of considerable size, compared to the same distance in a heavily urbanised setting. For this purpose, let the concept of a geographic setting be introduced for an individual person r that plans a trip t. The concept is defined as

> a vector representing the generalised travel cost towards all possible destinations j for an individual person r and for trip t, weighted by the attractiveness A_{jm} of these destinations (where $m = M_{rt}$)

where the concept of attractiveness is defined later in this section. Let this geographic setting be denoted by vector symbol $\mathbf{\Gamma}_{rt}$. It follows from the definition that each imaginable geographic location has a unique geographic setting. This is because the generalised travel cost from the origin location O_{rt} towards the same zone within $\mathcal{D}$ equals 0. No other origin zone can have such a generalised travel cost towards this zone.

Destination choice A key concept in this report, let trip destination choice be denoted by the dichotomous variable y_{rtj}, defined as

> the choice of an individual person r to travel towards zone j for trip t

In case a respondent r is observed to travel to destination zone j for trip t, define $y_{rtj} = 1$. For all other destination zones, define $y_{rtj} = 0$. It might be argued that *choice* is not an appropriate term for describing how individuals generally select where to travel to. In many contexts, the location is a fixed location, and its selection doesn't represent what is normally considered a (free) choice. Here, however, choice is adopted in a descriptive modelling denotation and refers to the outcome of a behavioural process in which one alternative is selected from a finite set of alternatives. For this selection process, it is left undefined how long the selection was made before the actual trip and under what conditions it took place.

Marginal travel cost Let the marginal travel cost of trip t by respondent r be denoted by ω_{rtj}. This measure is defined as

> the total travel cost that is associated with destination choice y_{rtj} considering the entire trip chain of individual r starting from trip t onwards, compared to the expected trip chain when destination choice y_{rtj} would not have been made, but r would have stayed home or travelled back home instead

Thus, the marginal travel cost is defined as the difference between the total generalised travel cost of the entire trip chain in which r chooses to travel towards zone j for trip t, and how this trip chain would have looked like if r would *not* have made this choice for trip t, but just stayed at home or travelled back to home instead. As a simple approximation, let the marginal travel cost ω_{rtj} be approximated by twice the travel cost (back and forth) that an individual r is required to make when travelling from current origin zone O_{rt} towards $j = D_{rt}$. Thus, the marginal travel cost ω_{rtj} roughly equals $2\psi_{ijm}$ for destination choice y_{rtj} (where $i = O_{rt}$ and $m = M_{rt}$). Because marginal travel cost is a relative measure, the definition $\omega_{rtj} \approx \psi_{ijm}$ yields identical results. For simplicity, let this approximation be used until a more elaborate specification of ω_{rtj} is proposed in section 4.1.

Destination choice utility Let the utility of destination choice y_{rtj} be denoted by U_{rtj}. The measure is defined as

> the relative degree in which destination choice y_{rtj} is preferred by an average person in sample ξ in case he or she faces a choice situation like r faces at the onset of trip t

The concept of *utility* originates from economics. It has been widely embraced by the transport modelling community, in particular since the publication of the book of Ben-Akiva and Lehrmann in (1985). The concept however is given a slightly distinct interpretation in this report. Other than in the conventional interpretation, the above definition of U_{rtj} does not provide with the utility experienced by individual r, but rather that of an *average individual* from sample ξ who is faced with the same destination choice context y_{rtj} as r is. For notational simplicity, ξ is omitted in the notation U_{rtj} and is implicitly assumed instead. Chapter 3 is devoted completely to the concept of destination choice utility and its practical estimation for all possible destination choice contexts.

Destination choice selectiveness Let the selectiveness σ_{rtj} of a destination choice y_{rtj} be defined as

> the cumulative utility of all destination zones that require a lower marginal travel cost for an individual person r for trip t compared to a destination choice for zone j

The term selectiveness is adopted here as to indicate that for an observed destination choice $y_{rtj} = 1$, an individual r apparently wanted to perform an activity that could not be performed at any location nearer than j. The higher σ_{rtj}, the

more selective r is in his or her destination choice compared to average individuals in sample ξ, otherwise a more nearby destination would have sufficed. The reader might argue that this argument falsely assumes that individuals always choose to travel to the nearest option available. This assumption however is irrelevant as long as it is not defined what the individual is looking for when selecting j. Consider for example the observed selection of what is judged an 'unnecessarily' distant zone to perform a specific activity. Even for such (subjective) judgements, it can be argued that in case an unnecessarily distant option was looked for, the selected zone indeed represented the nearest option for r. The performing of unnecessarily distant activities cannot be performed at locations that are not unnecessarily distant.

Destination choice eccentricity The concept of destination choice selectiveness has been introduced as a first step to obtain a measure of destination choice that is independent of the geographic setting $\mathbf{\Gamma}_{rt}$ in which it is observed. For reasons discussed in section 2.4, a measure with such a property is desirable for the synthesis of trip destination choice data. However, in order to obtain such a measure, σ_{rtj} need be standardised for the total utility offered by all destination zones surrounding O_{rt}. This is because base locations O_{rt} located within urban surroundings are likely to offer higher total utility compared to more isolated base locations. Let this standardised variable be termed destination choice eccentricity, denoted by E_{rtj} and defined as

> the ratio of the destination choice selectiveness associated with destination zone j to the destination choice selectiveness of the destination zone that involves the highest marginal travel cost for an individual person r for trip t

Alternatively stated:

$$E_{rtj} = \frac{\sigma_{rtj}}{\sigma_{rtk}} \tag{2.1}$$

where k is the zone for which σ_{rtk} takes the highest value for trip segment M_{rt} and origin zone O_{rt}. Thus, the value of E_{rtj} can range between values of 0 and 1. Based on this definition, chapter 4 will show that the measure has two useful interpretations. First, E_{rtj} expresses the likeliness of an individual person r choosing j when starting trip t of trip segment M_{rt} from origin location O_{rt}. Second, the measure defines the chance by which r is expected to select a destination zone that requires less marginal travel cost compared to j. Both interpretations offer a measure of the associated destination choice y_{rtj} that is independent of the geographic location $\mathbf{\Gamma}_{rt}$ in which it is observed. In the following section, a general modelling framework on travel behaviour is introduced, based on this property.

2.4 A theoretic model on the synthesis of travel diary data

This section proposes a modelling framework for the synthesis of travel behaviour vector $\mathbf{B}_s$ for all synthetic individual $s \in SP$. Based on this theoretic model, section 2.5 derives a conceptual algorithm for the synthesis of travel diary data. Chapters 3 and 4 then implement this conceptual algorithm for practical application. In chapters 5 until 7, this implemented algorithm is applied for two different case studies, which are verified with external data sources.

1. Using the terminology and the notation scheme introduced in section 2.3, consider an individual person r whose travel behaviour $\mathbf{B}_r$ is observed within travel behaviour database TB. Furthermore, let vector $\mathbf{\Gamma}_r$ represent all geographic settings $\mathbf{\Gamma}_{rt}$ in which this individual is located at the beginning of each trip t (see page 38). Consider these geographic settings as the environment within which a more fundamental, *location-independent travel behaviour* manifests itself compared to the travel behaviour pattern that is observed in reality. Without further elaboration on its operational specification, let such location-independent travel behaviour be denoted by vector $\mathbf{B}_r^*$. In order to distinguish between both vectors, let vector $\mathbf{B}_r$ alternatively be referred to as *embedded travel behaviour*, referring to its embedding in geographic setting $\mathbf{\Gamma}_{rt}$. Finally, consider a function u that filters location-independent travel behaviour vector $\mathbf{B}_r^*$ from observed, embedded travel behaviour vector $\mathbf{B}_r$:

$$\mathbf{B}_r^* = u(\mathbf{B}_r, \mathbf{\Gamma}_r) \tag{2.2}$$

2. An individual's travel behaviour is partly bounded by opportunities and constraints that can be related to observable travel behaviour determinants (see page 37). Based on such determinants, consider a population clustering scheme $\mathcal{P}$ that uniquely assigns each individual person r to a population cluster P_r with relatively homogeneous travel behaviour (see page 37). Now, let vector $\mathbf{B}_{P_r,\text{avg}}^*$ denote the average, location-independent travel behaviour of all individuals who belong to population cluster P_r. Because $\mathbf{B}_{P_r,\text{avg}}^*$ is defined location-independent, the definition of such an average is straightforward, even in case individuals who have reported their travel behaviour did so in different geographic settings. Let the number of respondents within TB who belong to population cluster P_r be denoted by $N_{r,TB_{P_r}}$. Also, let the k^{th} element in $\mathbf{B}_r^*$ be denoted by B_{rk}^*. Similarly, let the k^{th} element in $\mathbf{B}_{P_r,\text{avg}}^*$ be denoted by $B_{P_rk,\text{avg}}^*$, defined as:

$$B_{P_rk,\text{avg}}^* = \frac{1}{N_{r,TB_{P_r}}} \sum_{r=1}^{N_{r,TB_{P_r}}} B_{rk}^* \tag{2.3}$$

3. Although the average, location-independent travel behaviour $\mathbf{B}^*_{P_r,\text{avg}}$ is assumed to be characteristic for respondents r who belong to population cluster P_r, a high variation is anticipated around $\mathbf{B}^*_{P_r,\text{avg}}$ due to a range of intangible factors, denoted by vector ε_r. Examples of such intangible travel behaviour determinants are individual, subjective perceptions, incomplete information, individual taste and preference or plain coincidence. No assumption is made on the distribution of ε_r. Now, assume an (imaginary) function v that provides with an individual's location-independent travel behaviour $\mathbf{B}^*_r$ based on $\mathbf{B}^*_{P_r,\text{avg}}$ and ε_r (where $P = P_r$). An individual, location-independent travel behaviour $\mathbf{B}^*_r$ is thus obtained, alternative to equation 2.2:

$$\mathbf{B}^*_r = v(\mathbf{B}^*_{P_r,\text{avg}}, \varepsilon_r) \tag{2.4}$$

4. Next, consider a synthesised individual s for whom the travel behaviour vector $\mathbf{B}^*_s$ need be known. For population cluster P_s, a characteristic travel behaviour $\mathbf{B}^*_{P_s,\text{avg}}$ is now available. The intangible vector ε_s however remains unknown. Thus, equation 2.4 can not be used directly to obtain an estimate of travel behaviour $\mathbf{B}^*_s$ for s. However, ε_s can be estimated based on the requirement for *TB* that each population cluster $P \in \mathcal{P}$ is represented by a significant number of observations, randomly sampled from all individuals who belong to this population cluster (see page 33). In case this requirement is fulfilled, it implies that all vectors ε_r from all respondents r for whom $P_r = P_s$ offer unbiased estimates of vector ε_s. The *random selection* of a respondent r for whom $P_r = P_s$ thus provides with an unbiased estimate for ε_s that is equal to ε_r. Based on this estimate, equation 2.4 provides with an unbiased estimate of $\mathbf{B}^*_s$ as follows:

$$\mathbf{B}^*_s = v(\mathbf{B}^*_{P_s,\text{avg}}, \varepsilon_r), \tag{2.5}$$

where r is randomly selected from *TB* where $P_r = P_s$. It is noted that because $\mathbf{B}^*_r = v(\mathbf{B}^*_{P_r,\text{avg}}, \varepsilon_r)$, the above equation can be simplified as $\mathbf{B}^*_s = \mathbf{B}^*_r$. This report uses the term *matching* to describe this microsimulation technique. It is defined as

> the random selection of an observed respondent r of the same population cluster as a synthesised individual s, in order to estimate the location-independent travel behaviour vector $\mathbf{B}^*_s$ by the observed behaviour $\mathbf{B}^*_r$

5. The now obtained travel behaviour $\mathbf{B}^*_s$ for s represents an abstract, location-independent travel behaviour. For practical purposes, it is required to know how this behaviour manifests itself in the specific geographic setting $\mathbf{\Gamma}_s$ where s is located during all trips t. For this purpose,

let function u' be introduced as the inverse of function u. Now, a travel behaviour pattern $\mathbf{B}_s^*$ that is embedded in the proper geographic setting $\mathbf{\Gamma}_s$ can be estimated as:

$$\mathbf{B}_s = u'(\mathbf{B}_s^*, \mathbf{\Gamma}_s) \tag{2.6}$$

6. Using the matching technique, the previous equation may be simplified as:

$$\mathbf{B}_s = u'(\mathbf{B}_r^*, \mathbf{\Gamma}_s) \tag{2.7}$$

where r is randomly selected from *TB* with $P_r = P_s$.

2.5 A conceptual algorithm for the synthesis of travel diary data

Based on the concepts introduced in section 2.3 and the theoretic matching model of section 2.4, this section proposes a conceptual algorithm for the synthesis of travel diary data, referred to as the TDD synthesis algorithm. Figure 2.1 presents the algorithm design in the form of a flowchart. Each of the computation steps shown in this figure are discussed hereunder.

1. The first step in the flowchart enhances the selection of a synthesise area $\mathcal{Z}$ (see page 35) and the generation of a synthetic population *SP* for this area (see page 33). An algorithm for the creation of such a synthetic population, in case not readily available, is offered by appendix A.

2. From the synthetic population *SP* thus created, the first individual s is selected, denoted in the flowchart by the assignment $s = 1$. The subsequent computational steps apply the general model on travel behaviour of section 2.4 for the synthesis of travel behaviour vector $\mathbf{B}_s$ (see page 37) for each individual $s \in SP$.

3. Next, a respondent r is randomly selected from the empirical travel diary database *TB* (see page 32). This respondent r should belong to the same population cluster P as s does, as denoted in the flowchart by the condition $P_r = P_s$. Under this constraint, r should be selected at random from *TB*, irrespective of the home zone of the respondent r and that of the synthesised individual s.

4. Based on the matching model of section 2.4, the location-independent travel behaviour $\mathbf{B}_r^*$ as reported by r may be assigned to s. In the flowchart, this is indicated by the assignment $\mathbf{B}_s^* = \mathbf{B}_r^*$. Section 2.1 has defined that this report aims to generate travel diary data on origin

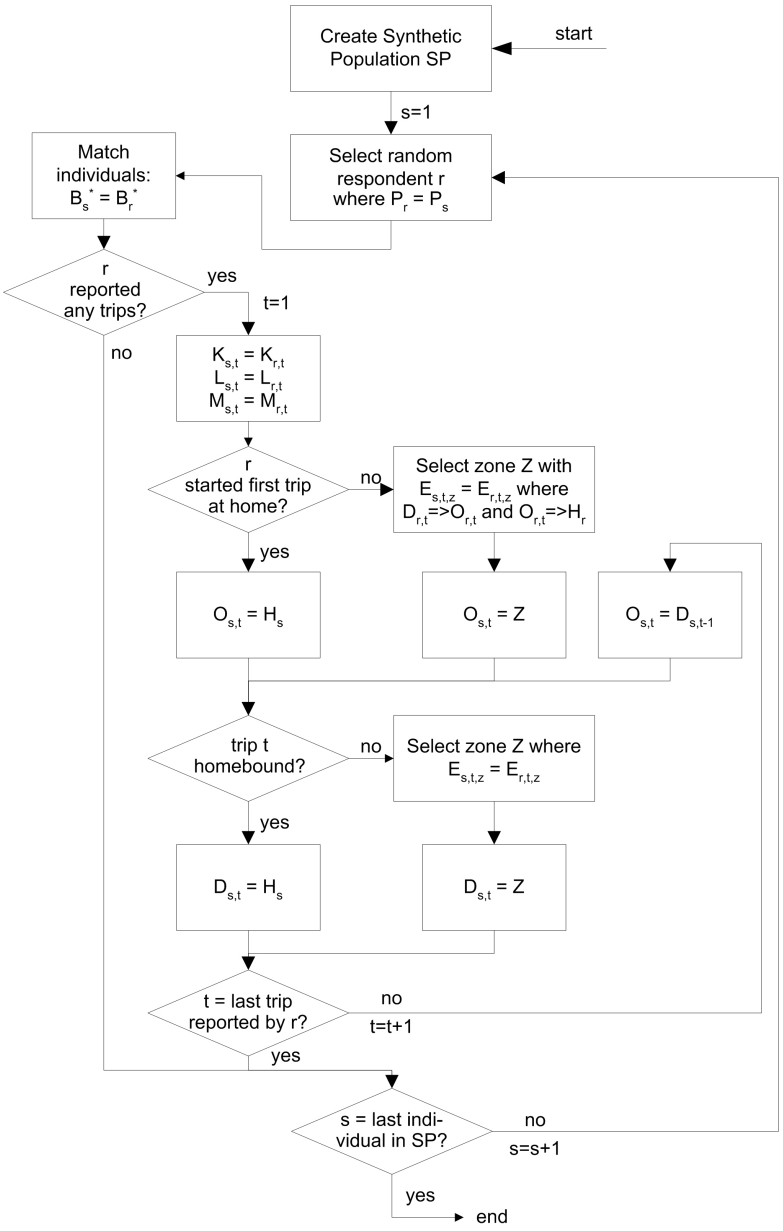

Figure 2.1: Conceptual algorithm for the synthesis of travel diary data

and destination locations, trip purpose and main mode of transport. This means that in the subsequent algorithm steps, the associated embedded travel behaviour aspects B_{sk} (see page 37) need be obtained from $\mathbf{B}_s^*$.

5. First, it is evaluated whether r reported any trips. In case r did not report any trips, no further modifications need be made to $\mathbf{B}_s^*$, simply because it does not contain any (trip) data on the above aspects. In this case, the algorithm continues with step 10.

6. From the above aspects, only origin locations O_{st} and destination locations D_{st} are determined using a non-trivial application of the general model in section 2.4. This choice is based on the discussion on page 17 that identified these aspects as the most location-specific of all travel behaviour aspects. Focussing modelling attention on these aspects, let trip purpose K_{st}, main mode of transport L_{st} and trip segment M_{st} be obtained trivially from $\mathbf{B}_r$ as is indicated in the flowchart by the assignments $K_{rt} = K_{st}$, $L_{rt} = L_{st}$ and $M_{rt} = M_{st}$. This is based on the assumption that these travel behaviour aspects are not influenced by different geographic settings, thus $u_k(x) = u'_k(x) = x$.

7. In many cases, the origin location of the first trip equals the home location of respondent r, but not necessarily so. Whether or not the first trip of r started from H_r is perhaps stated explicitly by $\mathbf{B}_r$. In case such a variable is not included in $\mathbf{B}_r$, it can alternatively be approximated by observing whether or not the first origin location O_{rt} equals H_r.

   (a) If it is thus determined that respondent r departs the first trip from home, the origin zone of the first trip for synthetic individual s is set equal to the home zone. This is denoted in the flowchart by the assignment $O_{st} = H_s$.

   (b) Alternatively, in case r departed from another location, a first origin location need be synthesised. The synthesis of a zone Z to represent this origin zone O_{rt} is analogous to the synthesis of a destination zone D_{rt}, discussed in step 8b. However, in order to apply this synthesis method for the former task, it is required to assume the home zone H_r as the origin zone of a fictive trip and the actually observed origin zone O_{rt} as the destination zone of this fictive trip. The eccentricity of this fictive trip can be used to synthesise an origin zone for s by means of a similar reversal of origin and destination zone for the synthetic trip.

8. For the determination of D_{st}, it is first evaluated whether or not r travelled home in trip t. Again, this is possibly denoted by a dedicated variable in TB. Alternatively, it can be approximated by observing whether the destination zone D_{rt} matches with H_r. Based on this evaluation, destination zone D_{st} is estimated as follows:

(a) First, in analogy with step 7a: if r returned home for trip t, the destination zone of trip t for synthesised individual s is set to the home zone H_s as well. This is shown in the flowchart by the assignment $D_{st} = H_s$.

(b) Alternatively, in case r did not travel towards home in trip t, destination zone D_{st} need be synthesised. For this synthesis, the model on travel behaviour of section 2.4 is used in conjunction with the concept of destination choice eccentricity (see page 40). This latter concept has been introduced as a location-independent measure for destination choice. As such, a destination choice $y_{rtj} = 1$ within geographic setting $\mathbf{\Gamma}_{rt}$ may be observed location-independent as E_{rtj} on a scale from 0 to 1. Using the terminology introduced in section 2.4, this can be formulated as:

$$E_{rtj} = u(D_{rt}, \mathbf{\Gamma}_{rt}) \quad \begin{array}{rl} \text{where} & j = D_{rt} \\ \text{and} & y_{rtj} = 1 \end{array} \tag{2.8}$$

Because of the independency of a geographic setting, E_{rtj} may be used to estimate the (location-independent) destination choice eccentricity of s in trip t. In order to determine an actual zone D_{st} that corresponds to this measure, it need be embedded in the geographic setting in which s is located:

$$D_{st} = u'(E_{rtj}, \mathbf{\Gamma}_{st}) \tag{2.9}$$

where r is randomly selected from *TB* with $P_r = P_s$ and function u' embeds destination choice eccentricity E_{rtj} into geographic setting $\mathbf{\Gamma}_{st}$ as it is experienced by s in trip t. In figure 2.1, equation 2.8 is denoted alternatively by the condition $E_{stZ} = E_{rtj}$ (where $j = D_{rt}$, $y_{rtj} = 1$ and $D_{st} = Z$).

9. At this point, the origin zone O_{st}, the destination zone D_{st}, trip purpose K_{st}, main mode of transport L_{st} and trip segment M_{st} are all determined for trip t by synthesised individual s. Next, it is determined whether r has performed additional trips.

   (a) In case r did report at least one additional trip, the trip counter t is raised by one, as denoted in figure 2.1 by the assignment $t = t + 1$. First, the synthesis of the *origin zone* of this subsequent trip is trivial. Because the travel behaviour database *TB* should store a list of all consecutive zones visited (see page 33), the origin zone of r, and thus that of s, is equal to the destination zone of the previous trip, as denoted in the flowchart by the assignment $O_{st} = D_{s,t-1}$. Second, for determining the destination zone of this subsequent trip, the algorithm returns to step 8.

(b) In case r did not report any additional trips, the algorithm continues with the final step.

10. Finally, it is evaluated whether or not s was the last individual within SP. In case this is true, the algorithm ends here, and a synthetic travel diary database has been generated for the entire synthetic population SP. Otherwise, the next individual is selected within SP, as denoted in the flowchart by the assignment $s = s+1$ and the algorithm returns to step 3.

2.6 Summary

This chapter introduced a conceptual algorithm for the synthesis of travel diary data (TDD) based on the matching of synthesised individuals and observed individuals. Table 2.1 specified the travel diary data output that this report aims to generate: origin and destination locations, trip purpose and transport mode for all trips made by all individuals residing in a randomly defined study area. Innovative to this report is that it aims to synthesise origin and destination data for a highly detailed zone set and that it allows the usage of travel diary data obtained from outside the study area. Next, section 2.2 specified the input data that is assumed available for the generation of this output data. This input was grouped over three databases. First, a travel behaviour database is required that contains observed travel behaviour and socioeconomic characteristics of the respondents who reported their behaviour. Second, a synthetic population is required that allows each of its members to be classified according to a separately defined population clustering scheme. In case no synthetic population is readily available, a zonal database was proposed as an alternative. Appendix A proposes an algorithm to convert this database into a synthetic population. Finally, a travel cost matrix is required that provides with the generalised travel cost for travelling between each pair of zones. Again, in case such a database is not readily available, an alternative database was proposed in the form of a basic infrastructure network, where appendix F proposes an algorithm to derive the required matrix from such a network. Next, section 2.3 introduced a range of auxiliary modelling concepts, the most important being geographic setting, destination choice and destination choice eccentricity. Based on these concepts, section 2.4 introduced a general model on travel behaviour, focussing on the relation between travel behaviour and the geographic setting in which it is embedded. It also formally introduced the key concept of matching. Finally, based on this general model and the modelling concepts of section 2.3, a conceptual algorithm for the synthesis of individual travel diary data was presented, referred to as the TDD synthesis algorithm. The algorithm is presented in the form of a flowchart by figure 2.1 and has been discussed step by step in section 2.5. It is termed a conceptual algorithm because it does not fully implement the concept of destination choice eccentricity yet. This location-independent measure of destination choice was introduced on page 40 and is further implemented in chapters 3 and 4.

Chapter 3

Destination choice utility

This chapter implements the observation of destination choice utility from an existing travel diary database. Destination choice utility is the main determinant of destination choice eccentricity (see on page 40). In turn, destination choice eccentricity is the single measure that needs further implementation in the TDD synthesis algorithm proposed in section 2.5. This implementation is subdivided over two chapters. First, this chapter designs and implements a model for the observation of destination choice utility. Second, chapter 4 is devoted to all remaining determinants of destination choice eccentricity. In the current chapter, section 3.1 starts by introducing several modelling concepts and an associated notation systems that extends section 2.3. Next, section 3.2 provides with a conceptual model for the estimation of destination choice utility. In this model, destination choice utility is defined from two base components; zone-based *attractiveness* and travel cost-induced *attractiveness decay*. The second part of the chapter is devoted to the implementation of both concepts in sections 3.4 and 3.5 respectively. This model specification is facilitated by a stepwise application for the illustrative case study Amsterdam-Rotterdam, introduced in section 3.3.

3.1 Modelling concepts and notational conventions

This section introduces various modelling concepts and an associated notation system for the modelling of destination choice utility. It thus extends the concepts and notations already introduced in section 2.3.

Dual zone system As it has been defined in section 2.1, it is the aim of this study to synthesise travel behaviour patterns at a low geographic aggregation level. This implies that destination choice eccentricity and thus destination choice utility need be modelled for a high number of OD-pairs (see page 35). The quadratic rise of independent OD-pairs with an increase in the number

of zones quickly leads to model systems that cannot be computed in practice. An associated problem for such fine zone systems is the occurrence of *sparse matrix problems*. For example, the number of trips for an OD-pair may be expected either zero or a low number for most OD-pairs that are far apart. Neither of such entries, if observed, are statistically significant. Thus, for all OD-pairs that are far apart, a fine zone system is not just impractical, but also not of much (statistical) interest. As to illustrate, for an individual living within or near a town, it is of interest whether he or she chooses to visit the centre, the north or the south of the town. However, in case someone covers 200 km to visit another town, it is mostly of interest that this particular town is chosen, not as much whether it was the centre, the north or the south. This observation leads to the idea of a two-layered or *dual* zone system. Let such a zone system be defined as

> a zone system in which each geographic area is modelled by two layers of zonal representation, with one layer representing an aggregation of the other layer

Let such a dual zone system be used to model the trip destination space $\mathcal{D}$. In such a zone system, the destination zone of an OD-pair is selected from either layer based on the generalised travel cost between origin and destination location. First, in case the generalised travel cost between a modelled origin zone and a physical destination exceeds a certain break travel cost, the zonal representation in the rough, aggregated layer is selected to represent the physical destination zone. Otherwise, the zone in the fine, non-aggregated layer is selected for the OD-pair. Effectively, this approach reduces the total number of OD-pairs that need be considered, because many OD-pairs are grouped into one. The precise reduction is determined by the level of aggregation in the rough zonal layer and the break travel cost. For most settings, this approach eliminates the quadratic increase of OD-pairs for an increase of the number of zones. The main disadvantage of this approach is that zones in the fine and rough layers are not independent of each other. Thus, in adopting a dual zone system, it need always be considered that predictions for specific zones in either layer only represent part of the modelled reality. A final prediction should always be based on the combination of both partial predictions. A visual representation of a dual zone system is offered by figure 6.3.

Notations for destination zones in a dual zone system In order to identify between the two zonal representations of each location, this paragraph proposes an extension to the notation system for trip destination space $\mathcal{D}$ introduced in section 2.3. Let destination zones that are included in the fine, non-aggregated layer be referred to as *single* destination zones and let those in the rough, aggregated layer be referred to as *composite* destination zones. Let single zones be indexed by $\tilde{\jmath}$ and let the set of all such single zones be denoted by $\mathcal{D}_S$. For composite destination zones, the notations J and $\mathcal{D}_C$ are used respectively. Next, the trip destination space $\mathcal{D}$ for dual zone systems is redefined for dual zone systems to represent the set of all possible destination

zones in *both* layers ($\mathcal{D} = \mathcal{D}_S + \mathcal{D}_C$). Similarly, the notation of a destination zone j (see page 36) is redefined for dual zone systems to refer to either a zone in $\mathcal{D}_S$ or in $\mathcal{D}_C$. Finally, let the subset of all single destination zones $\tilde{j} \in \mathcal{D}_S$ that are aggregated into composite zone $J \in \mathcal{D}_C$ be denoted by $\mathcal{D}_{S,J}$. Thus, $\mathcal{D}_{S,J} \subseteq \mathcal{D}_S$.

Other notations related to a dual zone system In the above discussion, a break travel cost was introduced for which a destination zone is selected from either the fine, non-aggregated or from the rough, aggregated destination zone layer. Let this break travel cost be denoted by the symbol ψ^*. In case $\psi^* = 0$ or $\psi^* = \infty$, the dual zone system collapses to a normal zone system, consisting only of the aggregated or the non-aggregated layer respectively. Assuming that $0 < \psi^* < \infty$, it is useful to possess of a standard notation to denote which origin zones are connected to which of the two possible destination zone layers. For this purpose, let a derivation of the generalised travel cost variable ψ_{ijm} be introduced, denoted by the symbol $\tilde{\psi}_{ijm}$. It is defined separately for single and for composite destination zones as:

$$\tilde{\psi}_{i\tilde{j}m} = \begin{cases} \psi_{i\tilde{j}m} & \text{where} \quad \psi_{iJm} \leq \psi^* \\ \infty & \text{where} \quad \psi_{iJm} > \psi^* \end{cases} \quad \text{and} \quad \tilde{j} \in \mathcal{D}_{S,J}$$

$$\tilde{\psi}_{iJm} = \begin{cases} \infty & \text{where} \quad \psi_{iJm} \leq \psi^* \\ \psi_{iJm} & \text{where} \quad \psi_{iJm} > \psi^* \end{cases} \tag{3.1}$$

Finally, let the set of all destination zones that can be selected by a trip-maker who starts a trip from origin zone i be denoted by $\mathcal{D}_i$. This set thus includes all destination zones j that can be selected from i as expressed by the condition $\tilde{\psi}_{ijm} < \infty$.

Cost circle Let a cost circle be defined as

> a range of marginal travel cost values that is modelled as one discrete cost category

Thus, a cost circle, or simply circle, constitutes a discrete representation of the continuous variable of generalised or marginal travel cost (see page 38). In their meta-study on travel demand models, Jong et al. (2004) used a similar modelling concept, that they labelled (distance) band. Both labels refer to the visual representation on a map by circular bands in case travel cost increases linearly with distance from a fixed origin zone. Let a circle be denoted by the symbol c, and let the lower and upper travel cost boundaries of a cost circle be denoted by $\omega_{c,\min}$ and $\omega_{c,\max}$ respectively. Thus, all values ω for which $\omega_{c,\min} \leq \omega < \omega_{c,\max}$, are represented by cost circle c and are modelled as having a fixed travel cost $\omega_{c,\text{avg}}$ that can be defined at wish ($\omega_{c,\min} \leq \omega_{c,\text{avg}} \leq \omega_{c,\max}$).

Let a cost circle scheme be defined as the set of all circles defined for a case study, denoted by $\mathcal{C}$. In the selection of $\mathcal{C}$, care should be taken that each of its circles represents a significant number of observations. Furthermore, it has no use to define circles that are smaller in their travel cost range as the average travel cost *within* zones, because the zone system itself prohibits the discerning of such travel costs. In such cases, the cost circle scheme would claim an accuracy that is not supported by the zone system. Assuming that attractiveness is evenly distributed over a zone and that zones are circle-shaped in terms of travel costs from the central point, the standard deviation σ_z by which the actual travel cost from address to address differs from the cost from centroid to centroid equals:

$$\sigma_z = \sqrt{2}(\frac{4r}{3\pi}) \tag{3.2}$$

where r = travel cost from the centre of an average zone to the edge. The equation is based on the formula for the gravity centre of a half circle, which equals $\frac{4r}{3\pi}$. The additional factor $\sqrt{2}$ represents that this error is made both for the origin zone as well as the destination zone. Because the above assumptions are mostly not true, this estimate can be considered a conservative estimate. However, the tempering effect from these conservative assumptions is countered by the fact that several zones may be much larger than the average zone. At the same time, trips from near the border of one zone to just over this border of the neighbouring zone are more likely to occur as longer trips. Thus, the standard deviation σ_z serves as the minimum travel cost range of individual cost circles.

Observed destination choice Later in this chapter, destination choice utility is defined from the number of trips observed in *TB* for all OD-pairs in a case study. For this purpose, let the notation $T^O_{ijm\xi}$ be introduced to denote the number of trips of trip segment m observed between origin zone i and destination zone j by the population ξ sampled for an empirical travel diary database *TB*. In order to interpret the value of $T^O_{ijm\xi}$, let two weigh factors be introduced. First, let $W_{im\xi}$ denote the number of observed trips that depart from i for trip segment m by population sample ξ:

$$W_{im\xi} = \sum_{j \in \mathcal{D}} T^O_{ijm\xi} \tag{3.3}$$

As a second weigh factor, let $W_{jm\xi}$ be defined as the number of observed trips that possibly could have headed towards destination zone j. It is defined as:

$$W_{jm\xi} = \sum_{i \in \mathcal{D}} W_{im\xi} \quad \text{where} \quad \tilde{\psi}_{ijm} < \infty \quad \text{(compare eq. 3.1)} \tag{3.4}$$

Aggregated home zones For the implementation of part of the destination choice utility model, an aggregation over home zones is desired. Page 35 introduced the notation h for home zones, as well as the symbol $\mathcal{H}$ to refer to the set of all home zones; the residence space. Let all these home zones alternatively be referred to as *base* home zones. Next, let an *aggregated* home zones be introduced as the aggregation of a group of base home zones, indexed by H (in analogy to the notation J for composite destination zones). Also, let the set of all aggregated zones be denoted by $\mathcal{H}_C$. Let $\mathcal{H}_H$ denote the set of all base home zones $h \in \mathcal{H}$ that are grouped into aggregated home zone H. Finally, let the part of population sample ξ that resides in H be denoted by ξ_H. Despite its analogy in notation, it is noted that the aggregation of home zones does not represent a dual zone system as it has been defined for destination zones. Zone set $\mathcal{H}_C$ just represents a second zone system that can be used in case a rougher zone system is preferred over $\mathcal{H}$. Because both zone sets are not used simultaneously, properties of H need not be combined with those of its substituent zones $h \in \mathcal{H}_H$ in order to obtain a complete estimate as required for a destination zone $\tilde{j}$ or J.

3.2 A model for the observation of destination choice utility

This section sets up a conceptual model for the observation of destination choice utility U_{rtj} for each destination zone j that could have been selected by individual r for trip t from a travel behaviour database *TB*. In this model, the utility U_{rtj} of destination choice y_{rtj} is assumed to be influenced by two main components. The first component, attractiveness, is the amount of activity opportunities offered by each alternative destination zone $j \in \mathcal{D}$, as these are perceived and evaluated by the population sample ξ within *TB* (Note that population sample ξ is implicitly assumed in the notation U_{rtj}, see page 39). The second component, attractiveness decay, represents the marginal travel cost ω_{rtj} as it is perceived and evaluated by population sample ξ. This section first formalises the above concepts in greater detail, before deriving the concept of destination choice utility from them.

Attractiveness For the definition of attractiveness, first let the concept of an activity opportunity be introduced as

> the possibility of performing an activity at some location

In this abstract definition, one single location can offer many activity opportunities. Next, let the attractiveness $A_{jm\xi}$ of a destination zone j for trips of trip segment m and for population sample ξ be defined as

> the relative amount of activity opportunities offered by destination zone j for a given trip segment m as it is perceived on average by an individual who belongs to population sample ξ

Attractiveness is thus defined independent of the (marginal) travel cost that need be invested to travel to j. Instead, it represents a characteristic of zone j only. In a basic interpretation, the measure of attractiveness $A_{jm\xi}$ expresses how many trips zone j is expected to attract from population sample ξ, if travel cost would not be an issue. Being a relative parameter, the total number of trips by ξ towards j can (in first order) be estimated as the total number of trips reported by ξ, times the ratio of $A_{jm\xi}$ to the cumulative attractiveness for all feasible, alternative destination zones. Section 3.4 proposes a model for the observation of attractiveness from an existing travel diary database.

Attractiveness decay For the definition of attractiveness decay, the concept of marginal travel cost is recalled from page 38. Based on this marginal travel cost, let an attractiveness decay function $\Phi_{m\xi_H}(\omega)$ be defined as

> a function providing the relative evaluation of an activity opportunity for trip segment m at a marginal travel cost ω as it is evaluated on average by population sub-sample ξ_H

The function is expected to decrease with increasing travel cost, because for equal activity opportunities, the one that is located further away is generally less visited. However, whether or not the function indeed decreases with increasing travel costs is left open by the above definition. Based on $\Phi_{m\xi_H}(\omega)$, let the attractiveness decay factor $\phi_{rtj\xi_H}$ be defined as

> the value of attractiveness decay function $\Phi_{m\xi_H}(\omega)$ at travel cost $\omega = \omega_{rtj}$ for trip segment m as it applies on average for population sub-sample ξ_H

This factor expresses the relative attractiveness of zone j as a possible destination zone for person r and for trip t as a result of travel cost ω_{rtj} that need be invested to reach j. Attractiveness decay functions and the derived attractiveness decay factors are defined irrespective of the destination zone j for a destination choice y_{rtj}. Instead, they only represent the effect of travel cost ω_{rtj} that need be invested to reach j. Attractiveness decay factors thus reveal the distribution of trips in case all destination zones would have equal attractiveness. Section 3.5 proposes a model for the observation of attractiveness decay factors from an existing travel diary database.

Destination choice utility The above components can be combined to estimate destination choice utility U_{rtj} for an average person residing in $\mathcal{Z}$:

$$U_{rtj} = A_{jm\xi}\ \Phi_{m\xi_H}(\omega_{rtj}) \ = \ A_{jm\xi}\ \phi_{rtj\xi_H} \tag{3.5}$$

where $H_r \in H$ and $m = M_{rt}$ and where U_{rtj} implicitly applies for population sample ξ only (see page 39)

An implemented model for the observation of attractiveness and attractiveness decay for population sample ξ is presented in section 3.4 and 3.5 respectively. In order to facilitate this model specification, first, an illustrative case study is introduced that will be implemented parallel to the model specification.

3.3 Case study Amsterdam-Rotterdam

This section introduces the case study Amsterdam-Rotterdam. The case study is introduced to clarify the elaborate and rather abstract model specification for attractiveness and attractiveness decay in sections 3.4 and 3.5. The case study is set up with very limited complexity in order to be as illustrative as possible. As a consequence, the outcomes of the case study are trivial. Two subsequent case studies in chapter 5 and 6 entail much more complexity so the reader is referred to these chapters for a less trivial application.

Circle	Description	Euclidean Distance (km)		
c		$\omega_{c,\min}$	$\omega_{c,\text{avg}}$	$\omega_{c,\max}$
1	Nearby cities	13	19	25
2	More distant cities	25	38	50
3	Distant cities	50	75	100

Table 3.1: Case study Amsterdam-Rotterdam, cost circle scheme

Case study specification The aim of the case study Amsterdam-Rotterdam is to obtain destination choice utilities for residents of large cities who intend to travel to a medium-sized or large city located at around 50 km of distance on a regular workday. In order to simplify, only destination choices are considered where an individual starts travelling from the home zone. This implies that for this case study, $\mathcal{H} = \mathcal{O}$, and that home zones h and trip origin zones i can be used interchangeably. In addition, only residents from Amsterdam and Rotterdam, the two largest cities of the Netherlands are studied and only a limited number of destination zones is considered. As such, the case study models destination choice utilities, *provided* that an individual travels to one of these cities and starts from the home zone. For this modelling, let a basic circle scheme $\mathcal{C}$ (see page 52) be defined as shown by table 3.1. Because the case study only models trips starting from the home zone, the marginal travel cost ω_{rtj} for each trip t is directly proportional to ψ_{ijm} (where $i = H_r$ and $m = M_{rt}$). For both Amsterdam and Rotterdam and for each circle $c \in \mathcal{C}$ around them, one medium-sized Dutch city has been selected at random. In addition, three additional cities are selected for the most distant circle. These include the cities of Amsterdam and Rotterdam themselves for destination choices from the other city. The trip destination space $\mathcal{D}$ is thus defined as the set of cities of Amersfoort, Amsterdam, Arnhem, Breda, Dordrecht, Haarlem and Rotterdam. The location of these cities is shown in figure 3.1. The figure shows that the city of Arnhem is distant for both Amsterdam and Rotterdam, where other cities are nearby for either Amsterdam or Rotterdam but distant

for the other one. Finally, all trips are considered to form one homogeneous trip segment. As a consequence, trip segment m takes a uniform value in all notations for this case study.

Figure 3.1: Case study Amsterdam-Rotterdam, trip destination space $\mathcal{D}$

Basic zone system This paragraph specifies the zone list and input data for the case study Amsterdam-Rotterdam by means of a basic zone system. The subsequent paragraph does the same for a dual zone system (see page 49). The comparison of both paragraphs illustrates the impact of the dual zone system on model complexity. Starting with the basic zone system, table 3.2 shows the population data for all individual zones showing that the selected destination cities are comparable in size. The population totals have been rounded to the nearest 5.000 inhabitants from data available for 1995. Section 3.4 will show that although population data is not strictly required for destination choice utility modelling, it can provide with an initial estimate for the iterative attractiveness estimation.

Next, table 3.3 specifies all origin and destination zones for the case study Amsterdam-Rotterdam as well as the euclidian distances in between. Because

Zone type	City	Population	# Respondents (=$W_{im\xi}$)
Origin i	1.Amsterdam	725,000	110
($i \in \mathcal{H} = \mathcal{O}$)	2.Rotterdam	600,000	98
Destination j	1.Amersfoort	125,000	-
($j \in \mathcal{D}$)	2.Amsterdam	725,000	-
	3.Arnhem	140,000	-
	4.Breda	160,000	-
	5.Dordrecht	120,000	-
	6.Haarlem	150,000	-
	7.Rotterdam	600,000	-

Table 3.2: Case study Amsterdam-Rotterdam, basic zone system

all these cities are connected to each other by direct highway and railway connections of comparable quality and congestion levels, the euclidian distances represent a valid estimate of the relative marginal travel cost ω on these routes. Table 3.3 also shows the circle to which all OD-pairs belong.

Origin $i \in \mathcal{O}$	1.Amsterdam		2.Rotterdam	
Destination $j \in \mathcal{D}$	ω_{rtj}	c	ω_{rtj}	c
1.Amersfoort	41	2	70	3
2.Amsterdam	-	-	58	3
3.Arnhem	82	3	99	3
4.Breda	87	3	43	2
5.Dordrecht	65	3	21	1
6.Haarlem	17	1	54	3
7.Rotterdam	58	3	-	-

Table 3.3: Case study Amsterdam-Rotterdam, travel costs in basic zone system

As the input travel diary database for the case study, the Dutch National Travel Survey (OVG) of 1995 has been used. The database is further introduced on page 124. Because only respondents are included that reported just one trip, the number of respondents per origin zone shown in table 3.2 equals the weigh factor $W_{im\xi}$ (see page 52). Finally, table 3.4 shows the observed number of trips $T^{O}_{ijm\xi}$ (see page 52) towards destination zone j from origin zone i by the population sample ξ in this database.

Destination $j \in \mathcal{D}$	i =Amsterdam	i =Rotterdam	total
1.Amersfoort	15	5	20
2.Amsterdam	-	29	29
3.Arnhem	6	6	12
4.Breda	3	14	17
5.Dordrecht	5	38	43
6.Haarlem	73	6	79
7.Rotterdam	8	-	8
total	110	98	208

Table 3.4: Case study Amsterdam-Rotterdam, observed destination choices for basic zone system

Dual zone system This paragraph adds a layer of *composite* zones to the basic zone system introduced in the previous paragraph. Two composite zones are defined, each of which aggregates the two most distant destination zones from either Amsterdam or Rotterdam. Table 3.5 presents the resulting dual zone system using the notation conventions introduced on page 50 and onwards.

Zone type	City	Population	# Respondents ($=W_{im\xi}$)
Origin i	1.Amsterdam	725,000	110
($i \in \mathcal{O}$)	2.Rotterdam	600,000	98
Destination $\tilde{j}$	1.Amersfoort	125,000	-
($\tilde{j} \in \mathcal{D}_S$)	2.Amsterdam	725,000	-
	3.Arnhem	140,000	-
	4.Breda	160,000	-
	5.Dordrecht	120,000	-
	6.Haarlem	150,000	-
	7.Rotterdam	600,000	-
Destination J	91.Arnhem/Breda	300,000	-
($J \in \mathcal{D}_C$)	92.Amersfoort/Arnhem	265,000	-

Table 3.5: Case study Amsterdam-Rotterdam, dual zone system

The seven physical destination cities are now represented by nine modelled zones. However, not each of these nine zones may be selected from each of the origin zones. For example, in the dual zone system, individuals from Amsterdam can no longer choose to travel to either Arnhem or Breda. Instead, they can only choose to travel to the composite zone which represents *both* Arnhem and Breda. Thus, although the zone system includes more zones, the number of OD-pairs has been reduced from the initial $2*(7-1) = 12$ to $2*(6-1) = 10$. This is illustrated by the revised travel cost table 3.6. In this table, the travel costs towards the newly created composite zones are determined by weighing the travel cost towards the individual zones with the associated population. For example the distance from Rotterdam to the combined zone of Arnhem and Amersfoort is determined as : (99 km * 140.000 + 70 km * 125.000) / (140.000 + 125.000) = 85,5 km. Finally, table 3.7 shows the observed trip frequencies T^O_{ijm} as a revision of table 3.3. The table also shows the weigh factors $W_{jm\xi}$ that vary per destination zone as a consequence of the dual zone system.

Origin $i \in \mathcal{O}$	1.Amsterdam		2.Rotterdam	
Destination $j \in \mathcal{D}$	cost ω_{rtj}	circle c	cost ω_{rtj}	circle c
1.Amersfoort	41	2	-	-
2.Amsterdam	-	-	58	3
3.Arnhem	-	-	-	-
4.Breda	-	-	43	2
5.Dordrecht	65	3	21	1
6.Haarlem	17	1	54	3
7.Rotterdam	58	3	-	-
91.Arnhem + Breda	85	3	-	-
92.Amersfoort + Arnhem	-	-	86	3

Table 3.6: Case study Amsterdam-Rotterdam, travel costs for dual zone system

Destination $j \in \mathcal{D}$	$T^O_{ijm\xi}$, $i =$ A'dam	R'dam	Weigh factors $\sum_i T^O_{ijm\xi} W_{jm\xi}$		Possible origin zones
1.Amersfoort	15	-	15	110	{Amsterdam}
2.Amsterdam	-	29	29	98	{Rotterdam}
3.Arnhem	-	-	0	0	∅
4.Breda	-	14	14	98	{Rotterdam}
5.Dordrecht	5	38	43	208	{Amsterdam,Rotterdam}
6.Haarlem	73	6	79	208	{Amsterdam,Rotterdam}
7.Rotterdam	8	-	8	110	{Amsterdam}
91.Arnhem + Breda	9	-	9	110	{Amsterdam}
92.Amersfoort + Arnhem	-	11	11	98	{Rotterdam}
total	110	98	208	1,040	(= 5 * 208)

Table 3.7: Case study Amsterdam-Rotterdam, observed destination choices for dual zone system

3.4 A model for the observation of attractiveness

This section presents a model for the (data-driven) observation of the attractiveness $A_{jm\xi}$ of each destination zone $j \in \mathcal{D}$, for each trip segment $m \in \mathcal{M}$ and as it is perceived on average by population sample ξ.

Conceptual observation model Attractiveness $A_{jm\xi}$ has been defined on page 53 to measure the relative provision of activity opportunities by destination zone j for trip segment m and for population sample ξ. In theory, such a measure can be observed by counting how often zone j is visited for trips of trip segment m and by population sample ξ in input travel diary database *TB* (see page 32). The more often a zone is visited by ξ, the more activity opportunities it apparently offers. However, it need be considered that each individual destination choice is made in a different geographic setting (see page 38). For example, the selection of a destination zone in the one geographic setting may reveal more activity opportunities as the selection of the same zone from another geographic setting. Consider a destination zone that is located far from most respondents in ξ. If such a zone receives relatively few trips, this need not necessarily indicate a low attractiveness. Second, consider a zone that is surrounded by several highly attractive zones. Such a zone can be considered in competition with these zones in attracting trips. As a consequence, it is more likely to receive a relatively low amount of trips from its vicinity. In case the same zone would not have such attractive neighbouring zones, it would probably attract more trips. Thus, each destination choice in *TB* need be considered in conjunction with the geographic setting in which the destination choice is made. Because page 38 has defined geographic setting based on the concept of attractiveness itself, it follows that both concepts need be estimated iteratively. Thus, let an initial guess of the attractiveness of each destination zone and for each trip segment be assumed, possibly by uniform value. Under the assumption of this distribution of attractiveness, it can be

determined how trips are most likely to be distributed based on average destination choice patterns in *TB*. Comparing this expected trip distribution with the actual distribution of trips in *TB*, it can now be determined how the initial attractiveness estimates need be adjusted. Before this iterative observation model is implemented, consider the following set of notational conventions.

Notational conventions For notational simplicity, let the population sample ξ be implicitly assumed for the remainder of this section and omitted in the notations $A_{jm\xi}$, $T^O_{ijm\xi}$, $W_{im\xi}$ and $W_{jm\xi}$ like it was proposeed for $U_{rtj\xi}$ on page 39. In addition, let the symbol n be used to index iteration steps in the iterative estimation of attractiveness. As an example, let A_{jmn} denote the estimate of $A_{jm\xi}$ in iteration step n, with implicit reference to population sample ξ. Also, let the notation $\tilde{\mathbf{\Gamma}}_{mn}$ be introduced to denote the geographic setting of an average origin zone for trips of trip segment m in iteration step n. Finally, let A^0_{jmn} denote the initial estimate of attractiveness in iteration step n. For the first iteration step, A^0_{jmn} can be set to a uniform value larger than 0 for all zones $j \in \mathcal{D}$. However, a reduced number of iteration cycles is probably required for the estimation of attractiveness, in case initial estimates are provided that correlate with the final attractiveness estimate A_{jm}. For example, for m = all trips, the choice for common population data seems appropriate, whereas for m = commuting trips, the choice for employment data seems more appropriate (if available). Initial attractiveness estimates and convergence speed are further discussed for a real-world case study on page 92.

Model implementation The above-proposed conceptual model for the observation of attractiveness is summarised as an iterative process that starts with the assumption of a distribution of attractiveness A^0_{jmn} for all destination zones $j \in \mathcal{D}$. Based on average destination choice behaviour observed for an average geographic setting $\tilde{\mathbf{\Gamma}}_{mn}$ in *TB*, it can be determined what would be the most likely corresponding distribution of trips. Differences between the actual trip distribution in *TB* and expected trip distribution under the assumption of a given attractiveness distribution, can now be used to improve this attractiveness distribution. An implementation of this approach is proposed by figure 3.2. The flowchart is organised into the various aggregation levels of the observed or estimated variables. In the flowchart, variables are shown by squared boxes with variable names shown in bold. Transitions from the one variable to the other are shown by rectangles with triangled sides, linking the associated variables with arrows. For each iteration step n, an estimation cycle starts at the upper right corner with an initial estimate for attractiveness A^0_{jmn} and ends at the lower right corner with an improved estimate of attractiveness A_{jmn}. The main variables that define this estimation process are F^O_{ijm} in the upper left corner and F^E_{ijmn} in the lower left corner. First, F^O_{ijm} is a convenient way of expressing the observed distribution of trips from origin zone i towards destination zone j for a standardised number of 100 respondents. These and all remaining variables are now discussed in individual paragraphs for the remainder of this section. In order to facilitate model dissemination, each paragraph

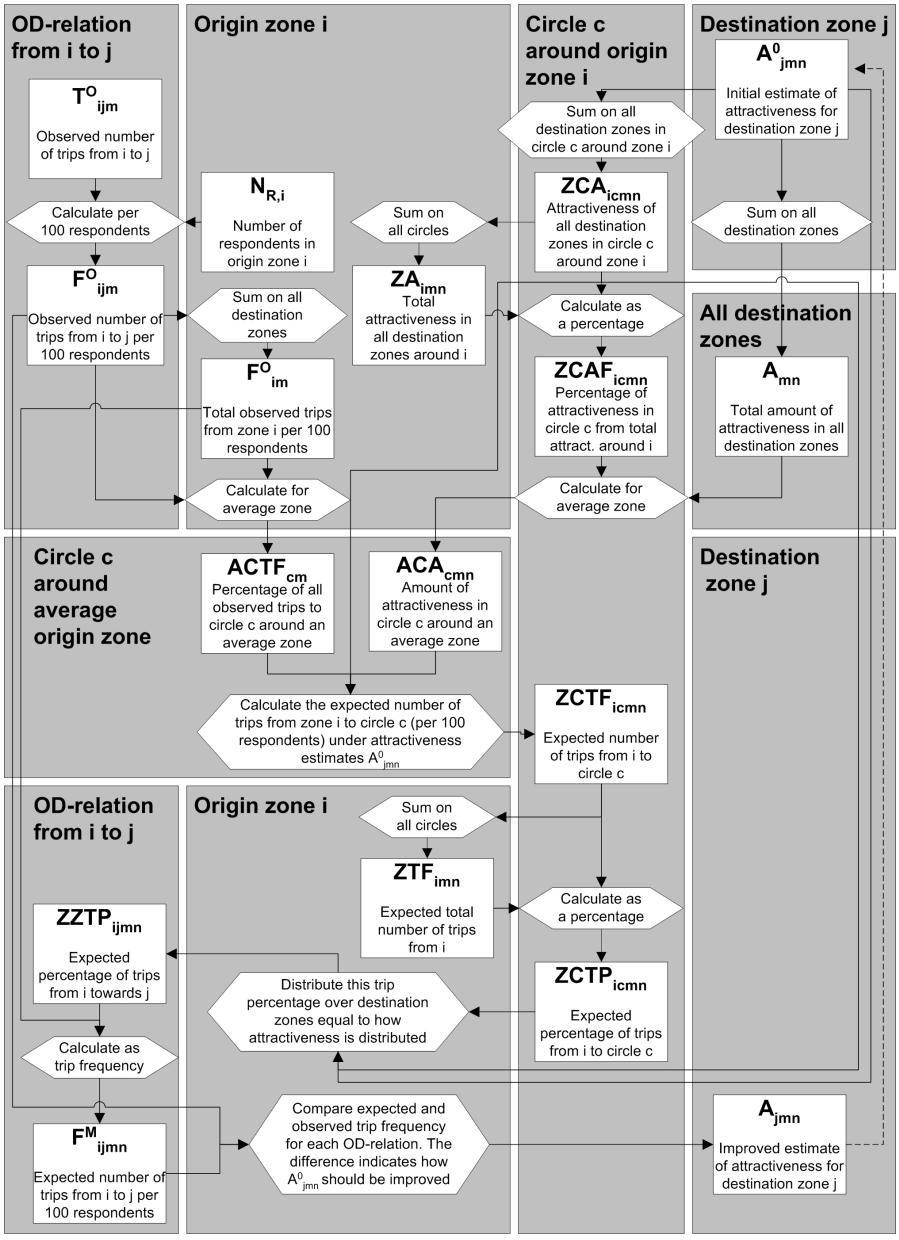

Figure 3.2: Attractiveness observation model

is accompanied by a sample application for the case study Amsterdam-Rotterdam, as it was introduced in section 3.3 for this purpose.

Initial attractiveness estimates A^0_{jmn} The initial attractiveness estimate A^0_{jmn} is assumed for the first iteration step $n = 0$. For the case study Amsterdam-Rotterdam, these initial values are assumed the population totals of all destination zones, as shown in table 3.5. For subsequent iteration steps, A^0_{jmn} is defined equal to the improved attractiveness estimate of the previous iteration step:

$$A^0_{jmn} = A_{jm(n-1)} \quad \textit{where} \quad n > 1 \tag{3.6}$$

Observed trip frequency F^O_{ijm} Let the observed number of trips $T^O_{ijm\xi}$ (see page 52 and table 3.7) be transformed to a standardised frequency. Let this frequency be denoted by F^O_{ijm}, defined as the observed number of trips of trip segment m from zone i towards zone j per 100 respondents residing in a home-zone $h \in i$:

$$F^O_{ijm} = \frac{T^O_{ijm\xi}}{N_{r,i}} * 100 \tag{3.7}$$

As it was proposed on page 60, the index for population sample ξ is omitted for notational simplicity. Reference to ξ is thus implicit for this and all its derived variables later in this section. Table 3.8 displays all values for F^O_{ijm} for the case study Amsterdam-Rotterdam. In addition, figure 3.3 offers a graphical illustration. In the case study, the observed number of respondents happens to be near 100 for both Amsterdam and Rotterdam (98 and 110 respectively). This explains why the values for F^O_{ijm} in table 3.8 do not differ much from those of $T^O_{ijm\xi}$ in table 3.7.

	Origin zone i	
Destination zone j	Amsterdam	Rotterdam
1.Amersfoort	13.6	-
2.Amsterdam	-	29.6
3.Arnhem	-	-
4.Breda	-	14.3
5.Dordrecht	4.5	38.8
6.Haarlem	66.4	6.1
7.Rotterdam	7.3	-
91.Arnhem + Breda	8.2	-
92.Amersfoort + Arnhem	-	11.2
total (F^O_{im})	100	100

Table 3.8: Case study Amsterdam-Rotterdam, observed destination choice frequencies

Figure 3.3: Case study Amsterdam-Rotterdam, observed trip frequencies per OD-pair

Total observed trip frequency F_{im}^{O} A second variable is introduced to represent observed trip frequencies, but aggregated for origin zones, instead of OD-pairs. Let this total observed trip frequency be denoted by F_{im}^{O} :

$$F_{im}^{O} = \sum_{j \in \mathcal{D}} F_{ijm}^{O} \quad \text{where} \quad \psi_{ijm} \geq \omega_{c_0,\min} \tag{3.8}$$

For case study Amsterdam-Rotterdam, the values for F_{im}^{O} amount to 100 precisely, as shown by table 3.8. This however is not the general case. Instead, these estimates are a result of two special characteristics of the case study. First, only respondents have been selected who reported *at least* one trip to any of the destination zones and second, none of them reported *more* than one trip to any of these destination zones for the same travel diary day. For real-world case studies where all possible destinations are covered by the zone system, where all respondents are included and where all trip segments are considered (apart from homebound trips), the value of F_{im}^{O} is expected to be in

the range from 120 to 250 per 100 respondents for a regular workday. This estimate is based on an average of 1.2 to 2.5 non-homebound trips per individual for a regular workday.

	Destination zones where $i =$	
Circle c	Amsterdam	Rotterdam
1.(12.5 to 25 km)	6.Haarlem	5.Dordrecht
2.(25 to 50 km)	1.Amersfoort	4.Breda
3.(50 to 100 km)	5.Dordrecht 7.Rotterdam 91.Arnhem + Breda	2.Amsterdam 6.Haarlem 92.Amersfoort + Arnhem

Table 3.9: Case study Amsterdam-Rotterdam, destination zones per combination of origin zone and circle

Zone to circle attractiveness ZCA_{icmn} Let ZCA_{icmn} sum the attractiveness A_{jmn} of all destination zones $j \in \mathcal{D}$ that are located in circle c around origin zone i:

$$ZCA_{icmn} = \sum_{j \in \mathcal{D}} A^0_{jmn} \quad \text{where} \quad \omega_{c,\min} \leq \psi_{ijm} < \omega_{c,\max} \tag{3.9}$$

The variable denotes the total attractiveness represented by circle c for a tripmaker who starts a trip of trip segment m from zone i for iteration step n. In case ZCA_{icmn} takes a relatively high value, one would expect a relatively high number of trips towards this circle. For the case study Amsterdam-Rotterdam, the computation of the values for ZCA_{icmn} is illustrated by two tables. First, table 3.9 shows which destination zones j are located within each circle around both origin zones. Now, the values for ZCA_{icmn} can be obtained by replacing the city names in this table by the corresponding attractiveness estimate A_{jmn} and summing them for each combination of origin zone i and circle c. Table 3.10 shows all resulting estimates for ZCA_{icmn} for case study Amsterdam-Rotterdam in iteration step $n = 0$. .

	$i =$Amsterdam		$i =$Rotterdam	
Circle c	ZCA_{icmn}	$ZCAF_{icmn}$	ZCA_{icmn}	$ZCAF_{icmn}$
1.(12.5 to 25 km)	150,000	0.116	120,000	0.084
2.(25 to 50 km)	125,000	0.096	160,000	0.113
3.(50 to 100 km)	1,020,000	0.788	1,140,000	0.803
total $=ZA_{imn}$	1,295,000	(1.000)	1,420,000	(1.000)
total $=A_{mn}$		1,353,894		

Table 3.10: Case study Amsterdam-Rotterdam, zone to circle attractiveness totals

Total origin zone attractiveness ZA_{imn} Let ZA_{imn} be defined as an aggregation of ZCA_{icmn} over all circles $c \in \mathcal{C}$, defined as:

$$ZA_{imn} = \sum_{c \in \mathcal{C}} ZCA_{icmn} \tag{3.10}$$

The values for case study Amsterdam-Rotterdamare presented in table 3.10. The table illustrates that variations in ZA_{imn} are caused by the exclusion of certain zones from the destination spectrum $\mathcal{D}_i$ for both of the origin zones. For example, Amsterdam cannot be chosen as the destination zone of a trip by a respondent from Amsterdam, while respondents in Rotterdam cannot choose to travel to Rotterdam. Note that the dual zone system does not influence ZA_{imn}. In real-world case studies where $\mathcal{O}$ represents the world at large, ZA_{imn} takes a constant value for all origin zones.

Average total attractiveness A_{mn} Let A_{mn} be introduced as the average total attractiveness from which a trip-maker in population sample ξ can choose for trips of trip segment m and for iteration step n:

$$A_{mn} = \frac{1}{\sum_{i \in \mathcal{O}} W_{im}} \sum_{i \in \mathcal{O}} W_{im}\, ZA_{imn} \tag{3.11}$$

For case study Amsterdam-Rotterdam, A_{mn} is calculated as 1 / 208 (110 * 1,295,000 + 98 * 1,420,000) = 1,353,894 for $n = 0$ and $m =$ all trips (see table 3.10).

Zone to circle attractiveness fraction $ZCAF_{icmn}$ By means of the aggregate variable ZA_{imn}, it is possible to express ZCA_{icmn} as a standardised fraction. Let this standardised fraction be denoted by $ZCAF_{icmn}$ and defined by equation 3.12. Table 3.10 shows all values for $n = 0$ for the case study Amsterdam-Rotterdam.

$$ZCAF_{icmn} = \frac{ZCA_{icmn}}{ZA_{imn}} \quad (0 \leq ZCAF_{icmn} \leq 1) \tag{3.12}$$

Average circle attractiveness ACA_{cmn} The above variables can be combined to obtain the zone to circle attractiveness for an average origin zone. Let this measure be termed the average circle attractiveness, denoted by ACA_{cmn}. :

$$ACA_{cmn} = \frac{1}{\sum_{i \in \mathcal{O}} W_{im}} \sum_{i \in \mathcal{O}} W_{im}\, ZCAF_{icmn}\, A_{mn} \tag{3.13}$$

For a circle scheme $\mathcal{C}$ that is composed out of circles with a fixed incremental increase in radius $\Delta\psi = \omega_{c,\max} - \omega_{c,\min}$, the variable ACA_{cmn} increases approximately linearly with $\omega_{c,\text{avg}}$ in case attractiveness is distributed evenly

over the geographic surface and travel cost is lineair with geographic distance. For example, in case travel cost is modelled as distance only and subdivided in discrete circles of a 10 km range, the circle of 50-60 km (on average 55 km) is roughly expected to contain $(10\text{km} * 2\pi * 55\text{km}/10\text{km} * 2\pi * 25\text{km}) = 2.2$ times as much attractiveness as the circle at 20-30 km (on average 25 km) from an average zone. Note that for circles describing a very large travel cost, ACA_{cmn} is expected to show a decreasing trend, because an increasing part of the circle area covers a geographic area located outside the trip destination space $\mathcal{D}$ that is modelled in detail (for example seas and foreign countries). For the Amsterdam-Rotterdam study, all values for ACA_{cmn} for the first iteration step are shown in table 3.11. The pattern in this table differs much from the above described pattern. This is caused by the specific selection of destination zones as well as the varying increase in radius for each circle.

Circle c	ACA_{cmn}	% of total	$ACTF_{cm}$	% of total
1.(12.5 to 25 km)	136,639	10.1%	53.4	53.4%
2.(25 to 50 km)	140,818	10.4%	13.9	13.9%
3.(50 to 100 km)	1,076,438	79.5%	32.7	32.7%
total	1,353,895	100.0%	100.0	100.0%

Table 3.11: Case study Amsterdam-Rotterdam, average circle attractiveness

Average circle trip frequency $ACTF_{cm}$ Analogous to the average circle attractiveness measure ACA_{cmn}, this paragraph introduces an average circle trip fraction. This average percentage, denoted by $ACTF_{cm}$, states the total percentage of trips of trip segment m that is expected to head to a circle c surrounding an average origin zone:

$$ACTF_{cm} = \frac{100\%}{\sum_{i\in\mathcal{O}} W_{im}} \sum_{i\in\mathcal{O}} W_{im} \frac{F^O_{ijm}}{F^O_{im}} \quad \text{where} \quad \omega_{c,\min} \leq \psi_{ijm} < \omega_{c,\max} \tag{3.14}$$

Table 3.11 shows the values for $ACTF_{cm}$ for each circle c for case study Amsterdam-Rotterdam. The table contrasts the values of $ACTF_{cm}$ to those of ACA_{cmn}. For example, the table shows that although circle 1 contains a mere 10% of all attractiveness surrounding an average origin zone, it attracts an average of 53% of all trips. The table also explains the values for $ACTF_{cm}$ for circle $c = 2$ and $c = 3$ respectively. It might seem strange that circle 2 attracts a mere 14% of all trips where the more distant circle 3 attracts 32.7% on average. However, the table shows that this is due to circle 3 containing an average of nearly eight times as much attractiveness as circle 2.

Zone to circle trip factor $ZCTF_{icmn}$ The two average variables ACA_{cmn} and $ACTF_{cm}$ define an average relationship between attractiveness and travel cost per circle c for an average zone. Thus, in case a certain amount of attractiveness ZCA_{icmn} is located in circle c surrounding zone i, it can be determined which

proportion of trips expected to this circle on average. Let this expected trip fraction be introduced as the zone to circle trip factor, denoted by $ZCTF_{icmn}$ and defined as:

$$ZCTF_{icmn} = ACTF_{cm} \frac{ZCA_{icmn}}{ACA_{cmn}} \tag{3.15}$$

All values for $ZCTF_{icmn}$ for the Amsterdam-Rotterdam study are shown in table 3.12. These factors are a first expectation of how trips would be distributed over the trip destination space $\mathcal{D}$ in case all attractiveness estimates A^0_{jmn} would be correct. Subsequent variables translate this aggregate expectation down to individual OD-pairs so that these can be compared with the actually observed trip fractions F^O_{ijm}. Based on these comparisons for all OD-pairs all estimates of attractiveness A^0_{jmn} can be improved.

Origin i	1.Amsterdam		2.Rotterdam	
Circle c	$ZCTF_{icmn}$	$ZCTP_{icmn}$	$ZCTF_{icmn}$	$ZCTP_{icmn}$
1.(12.5 to 25 km)	58.6	57.5	46.9	48.2
2.(25 to 50 km)	12.3	12.1	15.8	16.2
3.(50 to 100 km)	31.0	30.4	34.6	35.6
total	101.9	100.0	97.3	100.0

Table 3.12: Case study Amsterdam-Rotterdam, zone to circle trip factors

Total zone to circle trip factor $ZCTP_{icmn}$ As illustrated by 3.12, the zone to circle trip factors do not add up to a constant total. Because this would be a useful property, let trip factor $ZCTP_{icmn}$ be introduced as a standardised variant of $ZCTP_{icmn}$:

$$ZCTP_{icmn} = \frac{ZCTF_{icmn}}{\sum_{c \in \mathcal{C}} ZCTF_{icmn}} \quad (0 \leq ZCTP_{icmn} \leq 1) \tag{3.16}$$

Table 3.12 shows all values for $ZCTF_{icmn}$ for the case study Amsterdam-Rotterdam. The fact that the values for $ZCTF_{icmn}$ and $ZCTP_{icmn}$ are rather similar in this table, is again mainly caused by the selection of only those respondents who reported precisely one trip to one of the destination zones (see page 63).

Expected zone to zone trip fraction $ZZTP_{ijmn}$ The relationship between attractiveness and trip attraction has been defined as a linear relationship in the definition of attractiveness (see page 53). Because of this linear relationship, the zone to circle trip percentage $ZCTP_{icmn}$ may be distributed over the individual zones of a circle, proportional to the associated distribution of attractiveness. Let such an expected trip factor be termed the zone to zone trip fraction, denoted by $ZZTP_{ijmn}$ and defined as:

$$ZZTP_{ijmn} = ZCTP_{icmn} \frac{A_{jmn}}{ZCA_{icmn}} \tag{3.17}$$

Table 3.13 shows all values for $ZZTP_{ijmn}$ for the Amsterdam-Rotterdam study. These values may be interpreted as the most likely distribution of trips over the destination zones from all origin zones, in case all attractiveness estimates A^0_{jmn} would be correct.

Origin i	1.Amsterdam			2.Rotterdam		
Destination j	$ZZTP_{ijmn}$	F^E_{ijmn}	F^O_{ijm}	$ZZTP_{ijmn}$	F^E_{ijmn}	F^O_{ijm}
1.Amersfoort	0.121	13.3	13.6	-	-	-
2.Amsterdam	-	-	-	0.226	22.2	29.6
3.Arnhem	-	-	-	-	-	-
4.Breda	-	-	-	0.162	15.9	14.3
5.Dordrecht	0.036	4.0	4.5	0.483	47.3	38.8
6.Haarlem	0.575	63.2	66.4	0.046	4.5	6.1
7.Rotterdam	0.179	19.7	7.3	-	-	-
91.Arnhem + Breda	0.089	9.8	8.2	-	-	-
92.Amersfoort + Arnhem	-	-	-	0.083	8.1	11.2
total	1.000	100.0	100.0	1.000	100.0	100.0

Table 3.13: Case study Amsterdam-Rotterdam, trip estimates per OD-pair

Expected number of trips F^E_{ijmn} From $ZZTP_{ijmn}$, an expected number of trips can be determined for each OD-pair. Let this expected number of trips be denoted by the symbol F^E_{ijmn}, defined as:

$$F^E_{ijmn} = \frac{ZZTP_{ijmn}}{100} F^O_{im} \tag{3.18}$$

This variable states the expected number of trips for all OD-pairs again in case all attractiveness estimates A^0_{jmn} would be correct. Table 3.13 shows all values for $ZZTP_{ijmn}$ for the case study Amsterdam-Rotterdam. The table also shows the actually observed trip totals F^O_{ijm} from table 3.8. The comparison of all values for F^O_{ijm} and F^E_{ijmn} can now be used to improve the current attractiveness estimates.

Attractiveness estimation The subsequent paragraphs draw conclusions from the differences between F^O_{ijm} and F^E_{ijmn} in order to improve the current estimates of attractiveness for all destination zones. Note that these differences for the Amsterdam-Rotterdam study are relatively small for most OD-pairs (see table 3.13). This is a direct result of the mere two origin zones included in the case study. Because F^E_{ijmn} represents some form of expected, average behaviour, it cannot deviate much from the observed behaviour F^O_{ijm} that in itself constitutes around 50% of this average behaviour. However, for real-world case studies, much more origin zones will be defined, and individual observations of F^O_{ijm} may be expected highly influenced by chance for many OD-pairs. Despite this high chance factor, in case F^O_{ijm} is structurally higher compared to F^E_{ijmn} for a given destination zone j, it indicates that the attractiveness estimate of j should be increased. Unfortunately, the observation of such differences is much

complicated by the dual zone system. Because of this dual zone system, the following discussion requires several additional computational steps to arrive at an improved attractiveness estimate for each destination zone. Let all intermediate variables be referred to as attractiveness estimates, for simplicity denoted by A^{*}_{jmn}, A^{**}_{jmn}, A^{***}_{jmn} and A^{****}_{jmn} and discussed separately hereunder.

First attractiveness estimate A^{*}_{jmn} Because the attractiveness of a zone is defined linearly dependent with the amount of trips it attracts (see page 53), a first improved attractiveness estimate may be obtained as:

$$A^{*}_{jmn} = \frac{1}{\sum_{i\in\mathcal{O}} W_{im}} \sum_{i\in\mathcal{O}} W_{im} \frac{F^{O}_{ijm}}{F^{E}_{ijmn}} A^{0}_{jmn} \tag{3.19}$$

The equation states that the factor by which A^{0}_{jmn} need be adjusted equals the average factor between F^{O}_{ijm} and F^{E}_{ijmn} for all origin zones, weighted by the number of observed destination choices from these origin zones. Table 3.14 shows the resultant attractiveness estimates A^{*}_{jmn} for all zones j.

Destination j	A^{*}_{jmn}	A^{**}_{jmn}	A^{***}_{jmn}	A^{****}_{jmn}
1.Amersfoort	141,000	169,750	154,546	177,235
2.Amsterdam	751,100	-	751,100	861,368
3.Arnhem	-	157,543	157,543	180,672
4.Breda	140,960	146,880	144,091	165,245
5.Dordrecht	124,680	-	124,680	142,984
6.Haarlem	185,850	-	185,850	213,134
7.Rotterdam	243,600	-	243,600	279,362
91.Arnhem + Breda	275,400	-	-	-
92.Amersfoort + Arnhem	359,870	-	-	-
total (1-7)	2,222,460	-	1,761,410	2,020,000

Table 3.14: Case study Amsterdam-Rotterdam, intermediate attractiveness estimates for destination zones

Second attractiveness estimate $A^{**}_{\tilde{j}mn}$ Attractiveness estimate A^{*}_{jmn} incorrectly treats the two zonal representations $\tilde{j}$ and J of each physical geographic destination as independent of each other. In order to harmonise the attractiveness estimates $A^{*}_{\tilde{j}mn}$ and A^{*}_{Jmn} where $\tilde{j} \in \mathcal{D}_{S,J}$, let a second attractiveness estimate be introduced that distributes A^{*}_{Jmn} over the single zones $\tilde{j} \in \mathcal{D}_{S,J}$ of composite destination zone J. Thus, the second attractiveness estimate is defined for single destination zones $\tilde{j}$ only and defined as:

$$A^{**}_{\tilde{j}mn} = \frac{A^{0}_{\tilde{j}mn}}{A^{0}_{Jmn}} A^{*}_{Jmn} \quad \text{where} \quad \tilde{j} \in \mathcal{D}_{S,J} \tag{3.20}$$

This distributes the first attractiveness estimate for J over its constituent single destination zones $\tilde{j} \in \mathcal{D}_{S,J}$. For this distribution, the best available estimate

of attractiveness is selected which is A^0_{jmn}. For the very uncommon zone system adopted for case study Amsterdam-Rotterdam, equation 3.20 cannot be applied directly because the single destination zone of Arnhem is represented by two composite destination zones ($J = 91$ as well as $J = 92$). Therefore, it was necessary to compute the attractiveness for Arnhem based on estimates of two macro-zones. In order to do so, the two estimates A^*_{Jmn} have been combined and weighted by the number of observed destination choice contexts for which either composite zone could have been selected. Thus, the attractiveness estimate $A^{**}_{\tilde{j}mn}$ for Arnhem in case study Amsterdam-Rotterdam has been computed as:

$$A^{**}_{\tilde{j}mn} = \frac{1}{110 + 98}(98A^*_{91mn} + 110A^*_{92mn}) \quad \text{where} \quad \tilde{j} = \text{Arnhem}$$

The resulting estimates for $A^{**}_{\tilde{j}mn}$ are shown in table 3.14.

Third attractiveness estimate $A^{***}_{\tilde{j}mn}$ Attractiveness estimates $A^*_{\tilde{j}mn}$ and $A^{**}_{\tilde{j}mn}$ denote the improved attractiveness estimate of destination zone $\tilde{j}$ based on the number of expected trips towards the *single* zone $\tilde{j}$ and the associated *composite* zone J respectively. Both estimates are now combined by the third attractiveness estimate $A^{***}_{\tilde{j}mn}$, defined as:

$$A^{***}_{\tilde{j}mn} = \frac{1}{\gamma W_{\tilde{j}m} + W_{Jm}}(\gamma W_{\tilde{j}m} A^*_{\tilde{j}mn}) + (W_{Jm} A^{**}_{\tilde{j}mn}) \quad \text{where} \quad \tilde{j} \in \mathcal{D}_{S,J} \quad (3.21)$$

Attractiveness estimate $A^{***}_{\tilde{j}mn}$ is thus obtained by combining $A^*_{\tilde{j}mn}$ and $A^{**}_{\tilde{j}mn}$, weighted for the number of respondents on which they are based. The resulting estimates for $A^{***}_{\tilde{j}mn}$ for the Amsterdam-Rotterdam study are presented in table 3.14. By means of scale factor γ, it is possible to assign a higher weight to the attractiveness estimate for the single zone. This can be justified because estimate $A^*_{\tilde{j}mn}$ is targeted precisely at zone $\tilde{j}$ whereas $A^{**}_{\tilde{j}mn}$ is only roughly so. Choosing $\gamma > 1$ could thus lead to a higher convergence speed. However, the disadvantage of such an approach is that it could also lead to instabilities in the iteration process. For example, consider a single destination zone $\tilde{j}$ to act as a regional centre. In such case, the single zone $\tilde{j}$ is probably observed to attract a high number of trips from the $W_{\tilde{j}m}$ trips that are observed from the surroundings of $\tilde{j}$. In contrast, the W_{Jm} observed destination choices that can only select to travel towards J, probably includes much less trips towards $\tilde{j} \in J$. Consider that this composite zone J would *not* attract any trips at all from any of these W_{Jm} observed destination choices. In case $\gamma = 1$, both observations equally determine the improved attractiveness estimates for $\tilde{j}$. However, in case $\gamma >> 1$, a very high attractiveness estimate for $\tilde{j}$ could be the result, even though it is only based on the behaviour of a small group of respondents. Subsequent iterations cannot sustain this high estimate and will correct it downwards again. This is explained as follows: because $\tilde{j} \subset J$; a high

attractiveness estimate for $\tilde{j}$ also results in a high estimate of J. Because no-one travels towards J, such estimates for J cannot be sustained for subsequent iterations. This leads to instabilities in the attractiveness estimation that might also affect other zones than $\tilde{j}$ and J. Based on this discussion, this study proposes a modest scale factor for single zones of $\gamma = 3$, assuming around 50 single zones per composite zone.

Fourth attractiveness estimate $A^{****}_{\tilde{j}mn}$ Attractiveness estimate $A^{***}_{\tilde{j}mn}$ provides with an improved estimate of the attractiveness of all single destination zones. However, if these attractiveness estimates are aggregated for all single destination zones, a different total might be obtained compared to an aggregation of the initial A^{0}_{jmn}. Because attractiveness is a relative measure, this is no problem. However, in case the improved estimate should be compared to the initial estimate, this property is undesirable. Therefore, let another, standardised attractiveness estimate $A^{****}_{\tilde{j}mn}$ be defined that is comparable to A^{0}_{jmn}:

$$A^{****}_{\tilde{j}mn} = \frac{\sum_{k\in\mathcal{D}_S} A^{***}_{kmn}}{\sum_{k\in\mathcal{D}_S} A^{0}_{kmn}} A^{***}_{\tilde{j}mn} \quad \text{where} \quad \tilde{j} \in \mathcal{D}_S \tag{3.22}$$

Final attractiveness estimate A_{jmn} Attractiveness estimate $A^{****}_{\tilde{j}mn}$ has been defined for single destination zones $\tilde{j}$ only. Based on these estimates, a final attractiveness estimate is obtained for composite destination zones J by aggregating all single destination zones $\tilde{j} \in \mathcal{D}_{S,J}$. The final attractiveness estimate for a destination zone j is thus defined as:

$$A_{jmn} = \begin{cases} A^{****}_{jmn} & (\text{where } j \in \mathcal{D}_S) \\ \sum_{\tilde{j}\in\mathcal{D}_{S,j}} A^{****}_{\tilde{j}mn} & (\text{where } j \in \mathcal{D}_C) \end{cases} \tag{3.23}$$

Table 3.15 shows the final attractiveness estimates A_{jmn} of the first iteration for case study Amsterdam-Rotterdam and compares these with the initial attractiveness estimates A^{0}_{jmn} (equalling population totals). The table also shows the percentages by which all attractiveness estimates have been adjusted in this first iteration step. The table reveals that Rotterdam is much less attractive as to what might be estimated based on its considerable population. However, note that this estimation only applies for inhabitants of Amsterdam, who might behave distinctly in their destination choice behaviour for Rotterdam compared to inhabitants of other cities. Cities that appear significantly more attractive in relation to their population are Amersfoort and Haarlem. Both appear more attractive for inhabitants of Rotterdam and Amsterdam compared to Breda, even though they both have a smaller population.

Model input/output specification The above attractiveness observation model is summarised by table 3.16, listing all input and output variables of the model, organised for varying data objects.

Destination j	A^0_{jmn}	A_{jmn}	
1.Amersfoort	125,000	177,235	+42%
2.Amsterdam	725,000	861,368	+19%
3.Arnhem	140,000	180,672	+29%
4.Breda	160,000	165,245	+3%
5.Dordrecht	120,000	142,984	+19%
6.Haarlem	150,000	213,134	+42%
7.Rotterdam	600,000	279,362	-53%
total	2,020,000	2,020,000	0%
91.Arnhem + Breda	300,000	345,917	+15%
92.Amersfoort + Arnhem	265,000	357,907	+35%

Table 3.15: Case study Amsterdam-Rotterdam, final attractiveness estimates for destination zones

Data object	Input data	Output data
Origin zones $i \in \mathcal{O}$	Zone i # observed trips W_{im}	
Single destination zones $\tilde{\jmath} \in \mathcal{D}_S$	Zone $\tilde{\jmath}$ Zone J for which $\tilde{\jmath} \in \mathcal{D}_{S,J}$ Initial attractiveness $A^0_{\tilde{\jmath}mn}$ # possible trips $W_{\tilde{\jmath}m}$	Attr. $A_{\tilde{\jmath}mn}$
Composite destination zones $J \in \mathcal{D}_C$	Zone J Initial attractiveness A^0_{Jmn} # possible trips W_{Jm}	Attr. A_{Jmn}
OD-pairs $ij \in \mathcal{O} \times \mathcal{D}$	Zone i, Zone j # observed trips T^O_{ijm} Travel cost ψ_{ijm}	
Circles $c \in \mathcal{C}$	Circle c Average travel cost $\omega_{c,\text{avg}}$ Minimal travel cost $\omega_{c,\text{min}}$ Maximum travel cost $\omega_{c,\text{max}}$	

Table 3.16: Input/output for attractiveness estimation model

3.5 A model for the observation of attractiveness decay

This section introduces a model for the observation of attractiveness decay functions $\Phi_{m\xi_H}(\omega)$ from an existing travel diary database. For this observation model, it is assumed that the best available attractiveness estimates are used as they result from the final iteration step n. For notational simplicity, this section and later chapters omit the n index, and simply refer to the attractiveness of a destination zone by A_{jm}.

Conceptual model Because both the marginal travel costs ω_{rtj} for all possible destination choices y_{rtj} (see page 38) and the attractiveness A_{jm} of each destination zone j are known (see section 3.4), the observation of attractiveness decay is relatively straightforward. Apart from trip segment m, attractiveness decay functions are considered location-specific. For this purpose, let attractiveness decay functions be estimated independently for parts of the population

sample ξ according to their aggregated home zone H (see page 53), denoted by ξ_H. In order to estimate $\Phi_{m\xi_H}(\omega)$, let a circle scheme $\tilde{\mathcal{C}}$ be assumed that is independent from $\mathcal{C}$ (see page 52). For each circle $\tilde{c} \in \tilde{\mathcal{C}}$ around origin zone O_{rt} and for all trips made by persons who reside in $H_r \subseteq H$, it is now observed how much attractiveness could be reached in this circle ($\omega_{\tilde{c},\min} \leq \omega_{rtj} < \omega_{\tilde{c},\max}$). Also, it is observed how many trips have been reported to this circle for trip segment m. The ratio of both measures offers the average value of the decay function for circle $\tilde{c}$ and can thus be regarded an estimate of $\Phi_{m\xi_H}(\omega)$ where $\omega = \omega_{\tilde{c},\text{avg}}$. Remaining values of ω may be obtained by interpolation between $\omega_{\tilde{c},\text{avg}}$ for adjoining circles $\tilde{c}$. This conceptual model is now implemented in detail. Like for section 3.4, this model specification is illustrated by a sample application for case study Amsterdam-Rotterdam. For this case study, circle scheme $\tilde{\mathcal{C}}$ is taken equal to $\mathcal{C}$ and aggregate home zones H are defined equal to all home zones $h \in \mathcal{H}$.

Aggregated zone to circle attractiveness $ZCA_{H\tilde{c}m}$ Let $ZCA_{H\tilde{c}m}$ sum the attractiveness A_{jm} of all destination zones j that are located in circle $\tilde{c}$ around all trip origin zones O_{rt} visited by respondents r who reside in $H_r \subseteq H$ for trips t with trip segment $M_{rt} = m$. The measure is thus defined as:

$$ZCA_{H\tilde{c}m} = \sum_{r \in \xi_H} \sum_{t \in \mathbf{B}_r} \sum_{j \in \mathcal{D}} A_{jm} \tag{3.24}$$

$$\text{where } H_r \in H,\ m = M_{rt} \text{ and } \omega_{\tilde{c},\min} \leq \omega_{rtj} < \omega_{\tilde{c},\max}$$

The variable is a variant of ZCA_{icmn} (see page 64), but where ZCA_{icmn} provides an *average* zone to circle attractiveness per respondent, $ZCA_{H\tilde{c}m}$ provides the *total* attractiveness for all respondents. For case study Amsterdam-Rotterdam, table 3.17 shows all values for $ZCA_{H\tilde{c}m}$ for Amsterdam and Rotterdam respectively. For a better display of results, all values for $ZCA_{H\tilde{c}m}$ are divided by a factor 1,000,000 in these tables. Because attractiveness is a relative variable, this does not affect any of the subsequent estimates derived from $ZCA_{H\tilde{c}m}$.

	H = Amsterdam			H = Rotterdam		
Circle $\tilde{c}$	$ZCA_{H\tilde{c}m}$	$ZCT^O_{H\tilde{c}m}$	$\phi^O_{H\tilde{c}m}$	$ZCA_{H\tilde{c}m}$	$ZCT^O_{H\tilde{c}m}$	$\phi^O_{H\tilde{c}m}$
1.(12.5 to 25 km)	23.45	73	3.11	14.01	38	2.71
2.(25 to 50 km)	19.50	15	0.77	16.19	14	0.86
3.(50 to 100 km)	84.51	22	0.26	140.38	46	0.33
total / weighted avarage	127.46	110	0.86	170.58	98	0.57

Table 3.17: Case study Amsterdam-Rotterdam, attractiveness decay functions

Zone to circle trip total $ZCT^O_{H\tilde{c}m}$ Analogous to $ZCA_{H\tilde{c}m}$, let $ZCT^O_{H\tilde{c}m}$ denote the total number of destination choices reported in *TB* that involve a marginal travel cost ω for each of the travel cost circles $\tilde{c} \in \tilde{\mathcal{C}}$ ($\omega_{\tilde{c},\min} \leq \omega < \omega_{\tilde{c},\max}$).

Table 3.17 shows all values for $ZCT^{O}_{H\tilde{c}m}$ for case study Amsterdam-Rotterdam, calculated as:

$$ZCT^{O}_{H\tilde{c}m} = \sum_{r \in TB} \sum_{t \in \mathbf{B}_r} \sum_{j \in \mathcal{D}} y_{rtj} \tag{3.25}$$

$$\text{where } H_r \in H,\ m = M_{rt} \text{ and } \omega_{\tilde{c},\min} \leq \omega_{rtj} < \omega_{\tilde{c},\max}$$

Circle trip to attractiveness ratio $\phi^{O}_{H\tilde{c}m}$ The ratio of both previous variables provides a first estimate for the attractiveness decay function $\Phi_{m\xi_H}(\omega)$ where $\omega = \omega_{\tilde{c},\text{avg}}$. Let this ratio be denoted by $\phi^{O}_{H\tilde{c}m}$:

$$\phi^{O}_{H\tilde{c}m} = \frac{ZCT^{O}_{H\tilde{c}m}}{ZCA_{H\tilde{c}m}} \tag{3.26}$$

Table 3.17 shows all values for $\phi^{O}_{H\tilde{c}m}$ for case study Amsterdam-Rotterdam. In this table, the decreasing value observed for $\phi^{O}_{H\tilde{c}m}$ for increasing circles $\tilde{c}$ indicates that attractiveness of destination zones is devaluated in case they require a higher marginal travel cost (as it could be expected). However, the values for $\phi^{O}_{H\tilde{c}m}$ for Amsterdam and Rotterdam are not directly comparable with each other. In order to allow for such a comparison, the observed ratios $\phi^{O}_{H\tilde{c}m}$ should first be standardised for the (weighted) average ratio per aggregated home zone H. These weighted averages are also shown in table 3.17.

Circle trip to attractiveness ratio derivative $\frac{\delta\phi^{O}_{H\tilde{c}m}}{\delta\omega}(\omega)$ The estimates $\phi^{O}_{H\tilde{c}m}$ can be improved for marginal travel costs $\omega \neq \omega_{\tilde{c},\text{avg}}$ by interpolation. For this purpose, let a derivative of $\phi^{O}_{H\tilde{c}m}$ be defined for marginal travel cost ω. Let this derivative be defined as:

$$\frac{\delta\phi^{O}_{H\tilde{c}m}}{\delta\omega}(\omega) = \begin{cases} 0 & \text{where} \quad \omega \leq \omega_{\tilde{c},\text{avg(min)}} \\ \dfrac{\phi^{O}_{H\tilde{c}m} - \phi^{O}_{H(\tilde{c}-1)m}}{\omega_{\tilde{c},\text{avg}} - \omega_{(\tilde{c}-1),\text{avg}}} & \text{where} \quad \omega_{\tilde{c},\min} \leq \omega < \omega_{(\tilde{c}+1),\text{avg}} \\ \dfrac{\phi^{O}_{H(\tilde{c}+1)m} - \phi^{O}_{H\tilde{c}m}}{\omega_{(\tilde{c}+1),\text{avg}} - \omega_{\tilde{c},\text{avg}}} & \text{where} \quad \omega_{(\tilde{c}-1),\text{avg}} \leq \omega < \omega_{\tilde{c},\max} \\ 0 & \text{where} \quad \omega \geq \omega_{\tilde{c},\text{avg(max)}} \end{cases} \tag{3.27}$$

Attractiveness decay function $\Phi_{m\xi_H}(\omega)$ Based on the derivative defined above, the attractiveness decay function $\Phi_{m\xi_H}(\omega)$ is now defined formally:

$$\Phi_{m\xi_H}(\omega) = \phi^O_{H\tilde{c}m} + (\omega - \omega_{\tilde{c},\text{avg}}) \frac{\delta \phi^O_{H\tilde{c}m}}{\delta \omega}(\omega) \quad \text{where} \quad \omega_{\tilde{c},\min} \leq \omega < \omega_{\tilde{c},\max} \tag{3.28}$$

Attractiveness decay factors $\phi_{rtj\xi_H}$ The attractiveness decay factors $\phi_{rtj\xi_H}$ (see page 54) can now be obtained for every person r, trip t and destination zone j:

$$\phi_{rtj\xi_H} = \Phi_{m\xi_H}(\omega_{rtj}) \quad \text{where } H_r \in H \text{ and } m = M_{rt} \tag{3.29}$$

3.6 Summary

This chapter proposed a model for the observation of destination choice utility. First, section 3.1 introduced a list of auxiliary modelling concepts in addition to the list of section 2.3. The most important new concepts were the dual zone system for destination zones and cost circles. A dual zone system entails the flexible modelling of destination zones and results in a computationally more efficient representation of OD-pairs. The concept of cost circles offers a discrete representation of the continuous variable of marginal travel cost. Section 3.2 introduced a conceptual model for the observation of destination choice utility from an existing travel diary database. Here, it was proposed to define destination choice utility U_{rtj} from two main components. The first component, attractiveness, should measure the utility of a destination zone, independent of the marginal travel cost required to get there. Attractiveness can thus be interpreted as the utility offered by destination zones in case no travel cost would be required for travelling. The second component, attractiveness decay, should express the influence of marginal travel cost on the utility of a destination choice. Section 3.3 introduced a case study for illustrating the implementation of both components. Section 3.4 then proposed an implemented model for the calibration of the attractiveness $A_{jm\xi}$ of each destination zone j, trip segment m and for population sample ξ. The model states that the number of times in which a zone is observed to be visited by an observed population sample ξ reveals the associated number of activity opportunities and thus its attractiveness for ξ. However, this total number of zone visits is blurred by the geographic setting in which individual destination choices have occurred. For this reason, an observation model is constructed that controls for geographic settings. Finally, section 3.5 proposed a model for the observation of attractiveness decay functions for marginal travel cost. This observation is based on the ratio of the aggregated attractiveness of all observed destination choices and the associated number of trips per cost circle per aggregated home zone of the trip-maker.

Chapter 4

Destination choice eccentricity

The previous chapter has provided a model to observe the destination choice utility $U_{rtj\xi}$ for each destination choice y_{rtj}, as it is revealed for population sample ξ. This chapter discusses the derivation of destination choice eccentricity (see page 40) from destination choice utility. First, section 4.1 implements marginal travel cost ω_{rtj} for practical application. This implementation intends, in a computationally efficient manner, to eliminate part of the trip-based nature of observing just generalised travel costs between OD-pairs as it was initially proposed on page 38. Next, section 4.2 discusses the actual derivation of destination choice eccentricity from destination choice utility. Section 4.3 concludes with a discussion on the comparison of destination choice eccentricities obtained for different geographic settings as it is required for practical implementation of the TDD synthesis algorithm of section 2.5.

4.1 Marginal travel cost

The marginal travel cost ω_{rtj} was initially implemented on page 38 as the generalised travel cost ψ_{ijm} where $i = O_{rt}$ and $j = D_{rt}$. However, for non-home-based trips, the actual marginal travel cost can be very different. This section provides an extended implementation of marginal travel cost by including the residential home zone H_r. In addition, a framework is proposed by which the resulting marginal travel costs, despite their dependency on three zones, can be stored in a computationally efficient matrix of two dimensions.

Conceptual implementation method The simplest implementation of marginal travel cost ω_{rtj} for a destination choice y_{rtj} defines it proportional to the generalised travel cost ψ_{ijm} where $i = O_{rt}$ and $m = M_{rt}$. This was proposed on page 38 and this definition has been used for case study Amsterdam-Rotterdam in chapter 3 and will also be used for case study Interregional

Train Travel in chapter 5. The definition represents a fully trip-based implementation, appropriate for simplified case studies. For real-world case studies, a second implementation is proposed that incorporates the location of the home zone $h = H_r$ in addition to that of the trip origin zone $i = O_{rt}$. The home location influences marginal travel cost, because it is likely that the travel cost ψ_{jhm} from the selected destination zone j back to the home zone h need be covered later in the travel chain, as a consequence of destination choice y_{rtj}. For destination choices in the direction of the home zone, this leads to a lower marginal travel cost compared to destination choices in the opposite direction. Correcting for the sunk travel cost ψ_{ihm}, an improved definition of ω_{rtj} is thus proposed:

$$\omega_{rtj} = \psi_{ijm} + \psi_{jhm} - \psi_{ihm} \qquad (4.1)$$

where $i = O_{rt}$, $h = H_r$ and $m = M_{rt}$

For many destination choice contexts, this implementation offers an improvement to the initial definition on page 38. However, destination choice contexts can be thought of where equation 4.1 actually worsens the estimate of ω_{rtj} in comparison to a definition of ω_{rtj} proportional to ψ_{ijm}. An example would be a business trip or a lunch trip that starts from the workplace and returns there directly after. In this case, $\omega_{rtj} = \psi_{ijm}$ is to be preferred over equation 4.1. However, if the visited zone is relatively perpendicular to the route from home to the work location, the distortion is minimal, like for the general case where the travel cost is small in comparison to the commuting travel cost (if this cost would be high, it probably means that the destination zone is not visited from the workplace or not followed by a trip back to the workplace). Consequently, the number of cases where equation 4.1 worsens the estimate of ω_{rtj} are limited and largely cancel each other out. In case these small distortions are nonetheless expected to be of relevance for a specific research question, follow-up research is needed to implement a more elaborate definition of marginal travel cost.

Framework for a computationally efficient processing of destination choice For a computationally efficient processing of observed destination choice, it would be an advantage if all destination choices could be analysed by a simple matrix of just origin zones i and destination zones j, rather than by a cube of home zones h, origin zones i and destination zones j (as required by equation 4.1). The remainder of this section introduces a modelling framework that allows for such computationally efficient matrix storage of destination choice despite the three zones of relevance for each destination choice. Key to this framework is the definition of an eccentricity origin zone on page 79; a fictive trip origin zone that replaces both the actual trip origin zone and the residential zone of the trip-maker. This eccentricity origin zone is defined for each destination choice context using the concepts of intermediate zones and home elongation factors, which are introduced hereunder.

Intermediate zone $\tilde{k}_{ijmq}$ Let a constant fraction between 0 and 1 be denoted by the symbol q. Furthermore, let a set of such fractions be denoted by $\mathcal{Q}$. Based on such fractions, let an intermediate zone $\tilde{k}_{ijmq} \in \mathcal{O}$ be defined as a zone between zone i and zone j that is located at a fraction q of the generalised travel cost ψ_{ijm} from i:

$$\tilde{k}_{ijmq} = k \in \mathcal{O} \quad \text{where} \quad \min\Big(\left(q\psi_{ijm} - \psi_{ikm}\right)^2 + \left((1-q)\psi_{ijm} - \psi_{jkm}\right)^2\Big) \quad \text{and} \quad 0 \leq q \leq 1, \quad q \in \mathcal{Q} \tag{4.2}$$

An intermediate zone $\tilde{k}_{ijmq}$ is thus defined as the zone $k \in \mathcal{O}$ that is the most approximately located from i at a fraction q of the travel cost ψ_{ijm} and at a fraction $(1-q)$ of this travel cost from j.

Home elongation factor HEF_{rtj} As a second concept for the matrix storage of marginal travel cost, let the home elongation factor HEF_{rtj} be introduced. This factor is defined as the degree in which the selected destination zone j is further located from the home-zone $h = H_r$ as is the origin zone O_{rt} of trip t. It is defined at the $[0, 1]$ interval as:

$$HEF_{rtj} = \begin{cases} 0 & \text{where} \quad \psi_{him} = 0 \\ \left(\dfrac{\psi_{hjm}}{\psi_{him}}\right) & \text{where} \quad \psi_{him} \neq 0, \quad \psi_{hjm} < \psi_{him} \\ 1 & \text{where} \quad \psi_{him} \neq 0, \quad \psi_{hjm} \geq \psi_{him} \end{cases}, \tag{4.3}$$

where $h = H_r$, $i = O_{rt}$, $j = D_{rt}$ and $m = M_{rt}$

In case HEF_{rtj} is near to 0, it indicates that the destination choice y_{rtj} from i for j appears mostly influenced by the location of the home zone h. In contrast, if HEF_{rtj} is near to 1, the destination choice appears mostly influenced by the location of the origin location i.

Eccentricity origin i^*_{rtj} Based on the two concepts above, it is now possible to select a zone $k \in \mathcal{O}$ that appears to represent *both* the origin zone O_{rt} and the residential zone H_r for a destination choice y_{rtj}. Based on equation 4.1, the generalised travel cost from this zone k towards zone j should match (as closely as possible):

$$\psi_{kjm} = \psi_{ijm} + \psi_{jhm} - \psi_{ihm} \tag{4.4}$$

where $h = H_r$, $i = O_{rt}$, $j = D_{rt}$ and $m = M_{rt}$

Let such a zone k be labelled the eccentricity origin zone of destination choice y_{rtj}, denoted by i^*_{rtj}. This naming convention is adopted because this zone

indicates the location of the destination zone that, if selected, incorporates a minimal marginal travel cost ω_{rtj} and consequently a minimal destination choice eccentricity E_{rtj}. Based on the concepts of intermediate zones and the home elongation factor, it is defined as:

$$i^*_{rtj} = \begin{cases} H_r \quad \text{where} & \mathit{HEF}_{rtj} = 0 \\ \tilde{k}_{ijmq} \quad \text{where} & \begin{cases} 0 < \mathit{HEF}_{rtj} < 1 \\ \min \left| \mathit{HEF}_{rtj} - q \right| ,\ q \in \mathcal{Q} \\ i = O_{rt} \ ,\ m = M_{rt} \end{cases} \\ O_{rt} \quad \text{where} & \mathit{HEF}_{rtj} = 1 \end{cases} \tag{4.5}$$

The equation states that in case $\mathit{HEF}_{rtj} = 0$, the home zone H_r is selected as i^*_{rtj}. Alternatively, when $\mathit{HEF}_{rtj} = 1$, i^*_{rtj} is defined selected the origin zone O_{rt}. For intermediate cases, the value $q \in \mathcal{Q}$ is selected that is nearest to HEF_{rtj}. For this nearest value of q, intermediate zone $\tilde{k}_{ijmq}$ is selected as the best estimate of eccentricity origin zone i^*_{rtj}. Using eccentricity origin zones, the marginal travel cost ω_{rtj} may be estimated as the generalised travel cost $\psi_{i^*_{rtj}jm}$. After the eccentricity origin zone i^*_{rtj} is determined for each observed destination choice y_{rtj}, marginal travel cost can be retrieved from a simple travel cost matrix instead of a travel cost cube.

4.2 Utility to eccentricity

This section defines destination choice eccentricity estimates E_{rtj} for all possible destination choice contexts y_{rtj}, based on the associated destination choice utility U_{rtj}. Because chapter 3 estimated $U_{rtj\xi}$ as it applies for population sample ξ, the resulting estimates for E_{rtj} also apply for population sample ξ only. For notational simplicity, this chapter omits the ξ index in analogy with chapter 3.

Destination choice selectiveness σ_{rtj} The destination choice selectiveness σ_{rtj} has been introduced on page 39 as the sum of the utility of all destination zones that involve less marginal travel cost than the selected destination zone. Thus σ_{rtj} is defined as:

$$\sigma_{rtj} = \sum_{k \in \mathcal{D}} U_{rtk} \quad \text{where} \quad \omega_{rtk} < \omega_{rtj} \tag{4.6}$$

Destination choice eccentricity E_{rtj} Next, destination choice eccentricity itself is defined on page 40 as a standardised form of σ_{rtj}:

$$E_{rtj} = \frac{\sigma_{rtj}}{\sigma_{rtk} + U_{rtk}}, \tag{4.7}$$

where k represents the destination zone that takes the highest value for σ_{rtk} for any value $k \in \mathcal{D}$. This is the zone that requires the highest marginal travel cost for destination choice context y_{rtj}.

Destination choice eccentricity range The estimate in 4.6 and 4.7 can be considered a lower boundary estimate of σ_{rtj} and E_{rtj} respectively. This is because, as a result of the aggregation by the zone system, it cannot be distinguished which destination location is selected *within* zone j. In case an individual heads to a location within j that is somewhat further off compared to other locations within j, a somewhat more selective and thus somewhat more eccentric destination choice is made. All ignored locations within j that are nearer than the selected location within j offer part of the utility offered by the zone as a whole. Ignoring this more nearby utility means one is more selective in destination choice, and thus more eccentric. Based on this reasoning, the upper boundary estimate for σ_{rtj} can be defined as a slight modification of equation 4.6:

$$\sigma^*_{rtj} = \sum_{k \in \mathcal{D}} U_{rtk} \quad \text{where} \quad \omega_{rtk} \leq \omega_{rtj} \tag{4.8}$$

Equation 4.7 can now be extended to include both a lower and an upper estimate of the possible destination choice eccentricities associated with destination choice y_{rtj}:

$$E_{rtj,\min} = \frac{\sigma_{rtj}}{\max(\sigma^*_{rtk})} \tag{4.9}$$

$$E_{rtj,\max} = \frac{\sigma^*_{rtj}}{\max(\sigma^*_{rtk})} \tag{4.10}$$

Based on these measures, let the destination choice eccentricity range $\mathbf{E_{rtj}}$ be introduced as

> the range of destination choice eccentricity values that describe the eccentricity of a destination choice y_{rtj}

According to the above discussion, $\mathbf{E_{rtj}}$ is bounded by a lower limit $E_{rtj,\min}$ and an upper limit $E_{rtj,\min}$ as defined by equation 4.9 and 4.10 respectively.

Necessity of considering eccentricity ranges The necessity of considering eccentricity ranges instead of destination eccentricity points is illustrated as follows. Consider a destination choice for j by an individual r for trip t. Using

equation 4.7, eccentricity E_{rtj} can be observed for this destination choice as a point. Now, consider that zone j is split into two parts $j_{(A)}$ and $j_{(B)}$. A random part of the original socioeconomic contents of j is assigned to $j_{(B)}$, the other part remains in $j_{(A)}$. In addition, the location of $j_{(A)}$ is unchanged to j, but the location of $j_{(B)}$ is defined at some further travel cost from most destination choice eccentricity origin zones i^*_{rtj}. Now, because $j_{(A)}$ is smaller than the original zone j, and no new zone has been defined that requires a lower marginal travel cost ω_{rtj}, destination choice eccentricity $E_{rtj_{(A)}}$ should at maximum equal E_{rtj}, but probably takes a lower value. This is because the same percentage of trips still head to $j_{(A)}$ or some nearer zone, minus the percentage of trips that head to $j_{(B)}$. Similarly, the lowest possible value for $E_{rtj_{(A)}}$ is reached in case all socioeconomic contents are assigned to $j_{(B)}$. In this case, $E_{rtj_{(A)}}$ takes the value of E_{rtj} minus the percentage of trips attracted by the original j. These two extreme values represent the definition of the upper and the lower boundary of the destination choice eccentricity range. The range $\mathbf{E_{rtj}}$ defines all values that the destination choice eccentricity E_{rtj} could have taken in case (irrelevant) parts of zone j would have been assigned to other zones.

Interpretation of destination choice eccentricity The destination choice eccentricity range $\mathbf{E_{rtj}}$ offers three useful measures for the expected behaviour in destination choice context y_{rtj}. First, let P_{rtj} be introduced as the chance by which zone j is expected to be selected for destination choice context y_{rtj}. Using $\mathbf{E_{rtj}}$, this frequency is simply defined as:

$$P_{rtj} = E_{rtj,\max} - E_{rtj,\min} \tag{4.11}$$

Second, the likeliness P^-_{rtj} of destination choices that head towards destination zones that require a *lower* marginal travel cost than ω_{rtj} for destination choice context y_{rtj}, is obtained from $\mathbf{E_{rtj}}$ as:

$$P^-_{rtj} = E_{rtj,\min} \tag{4.12}$$

Alternatively, the likeliness P^+_{rtj} of destination choices that head towards destination zones that require a *higher* marginal travel cost than ω_{rtj} for destination choice context y_{rtj}, is obtained from $\mathbf{E_{rtj}}$ as:

$$P^+_{rtj} = 1 - E_{rtj,\max} \tag{4.13}$$

Destination choice eccentricity observation model Equation 4.9 has finalised a model for the observation of destination choice eccentricities for each possible destination choice y_{rtj} by means of a travel diary database *TB* and a travel cost matrix *TC* only. In the remainder of this report, let this model be referred to as the destination choice eccentricity observation model, defined as

the model presented in chapters 3 and 4 for the data-driven observation of destination choice eccentricity as the product of attractiveness and attractiveness decay as they follow with maximum likelihood from an existing travel diary database *TB* and an associated travel cost matrix *TC*

As it has been noted before, this observation model actually offers eccentricity estimates that apply for population sample ξ only (but this index is omitted for notational simplicity). In case it is applied for a randomly sampled person from a population with another composition as ξ, erroneous estimates may occur. However, for the TDD synthesis algorithm of section 2.5, where the destination choice eccentricity observation model has been designed for, this is not of relevance. This is because all synthesised individuals are exclusively assigned a travel behaviour as it is observed for ξ.

4.3 Destination choice eccentricity comparison

The TDD synthesis algorithm presented in section 2.5 compares two values of destination choice eccentricity with each other in step 8b. However, in the above implementation, it has been shown that destination choice eccentricity can only be estimated as a *range* of possible values $\mathbf{E_{rtj}}$ instead of point estimates. This section proposes a method to compare such ranges obtained for different destination choice contexts. Consider a destination choice y_{rtj} is represented by eccentricity range $\mathbf{E_{rtj}}$. Now consider a second individual $\hat{r}$ who starts trip $\hat{t}$. The following algorithm defines which zones $k \in \mathcal{D}$ have equal eccentricity $E_{\hat{r}\hat{t}k} = E_{rtj}$ and with which probability.

1. First it is checked if a zone k exists, for which the following two conditions hold: $E_{\hat{r}\hat{t}k,\min} \leq E_{rtj,\min}$ and $E_{\hat{r}\hat{t}k,\max} \geq E_{rtj,\max}$. In this special case, *only* a choice for k by $\hat{r}$ in trip $\hat{t}$ represents destination choice y_{rtj} in its associated eccentricity. This means that a destination choice for zone k resembles as no other destination choice the observed destination choice y_{rtj}. Note that the reverse is not necessarily true.

2. In all other cases, two or more destination zones k for $\hat{r}$ can represent destination choice y_{rtj} in its eccentricity. In this case, one of these alternative zones need be selected at random. However, not all these zones are equally likely to represent a match. Therefore, each alternative zone is assigned a chance mass in which it represents a match with y_{rtj}.

3. First, all possible destination zones need be selected. These are all zones $k \in \mathcal{D}$ which satisfy at least one of the following two conditions:

$$E_{rtj,\min} \leq E_{\hat{r}\hat{t}k,\min} < E_{rtj,\max}$$

or

$$E_{rtj,\min} < E_{\hat{r}\hat{t}k,\max} \leq E_{rtj,\max}$$

4. Next, for each of these alternative zones k, the overlap of $\mathbf{E_{\hat{r}\hat{t}k}}$ with $\mathbf{E_{rtj}}$ is computed. Let this overlap be denoted by ζ_k and defined as:

$$\zeta_k = \min(E_{rtj,\max}, E_{\hat{r}\hat{t}k,\max}) - \max(E_{rtj,\min}, E_{\hat{r}\hat{t}k,\min}) \tag{4.14}$$

5. Subsequently, each alternative zone k is assigned a chance mass P_k for which E_{rtk} is expected to resemble E_{rtj}:

$$P_k = \frac{\zeta_k}{\Delta E_{rtj}} \quad \text{where} \quad \Delta E_{rtj} = E_{rtj,\max} - E_{rtj,\min} \tag{4.15}$$

6. Finally, a zone is randomly selected from all alternative zones k, weighted by the associated chances P_k of occurring. The choice for the thus selected destination choice by person $\hat{r}$ for trip $\hat{t}$ is an unbiased representation of the observed destination choice for zone y_{rtj} in its eccentricity.

4.4 Summary

This chapter discussed the observation of destination choice eccentricity based on the implementation of destination choice utility in chapter 3. First, the concept of marginal travel cost was extended from the initial definition on page 38 to also consider the residential location of trip-makers. The section also offered a computationally efficient method to store marginal travel cost that is based on three zone dimensions. Next, the derivation of destination choice eccentricity from destination choice utility was discussed in section 4.2. In this discussion, it was explained that destination choice eccentricity need be estimated as a range rather than a point value, as a consequence of the zone system. Section 4.3 concluded the chapter in offering a method for the comparison of two destination choice eccentricity ranges. Such a method is required because the TDD synthesis algorithm proposed by section 2.5 is based on such comparisons.

Chapter 5

Case study Interregional Train Travel

This chapter provides an example application of the destination choice eccentricity observation model for a real-world case study. Although chapter 3 has already provided extensive illustration for all modelling steps, this has been done for the trivial case study Amsterdam-Rotterdam only. This served well for the illustration of all model variables and all computation steps, but it is admitted that the final model outcomes hardly added information to the original observations in table 3.7. By selecting a more complex case study in this chapter, the initial focus on model illustration is shifted to model demonstration. The case study set up for this aim models interregional train trips in the Netherlands. Section 5.1 starts with a discussion of all modelling choices and input data for this case study Interregional Train Travel. Sections 5.2 until 5.4 apply the destination choice eccentricity observation model for an existing travel diary database of interregional train travel. First, section 5.2 observes the attractiveness of each destination zone in the case study. Next, section 5.3 observes three distinct attractiveness decay functions. Section 5.4 finalises model implementation by estimating destination choice eccentricity ranges for all possible destination choices in the case study. The results of this exercise are extensively validated in section 5.5. This validation also argues to what degree the original input data has been enriched by application of the destination choice eccentricity observation model. Section 5.6 concludes the chapter with a discussion on possible practical applications.

5.1 Case study specification

Let an interregional train trip be defined as

> a trip starting in the home zone of the trip-maker which heads to any other zone and for which the train is the main mode of transport

The case study models the destination choice eccentricity for each possible interregional train trip. This first section discusses all input data and modelling choices that have been used for this task. Note that the above definition of an interregional train trip excludes homebound train trips (see also page 36).

Zone system For the case study, let the Netherlands be divided into 16 regions around the major railway stations. The resulting zone system is shown in table 5.1 and in figure 5.1. First, table 5.1 shows that many zones actually represent a set of important railway stations that are in each other's vicinity or related to each other network-wise. In such cases, the most dominant or most central station has been selected. Second, figure 5.1 shows how each municipality has been allotted to each of these railway stations by the fastest railway connection. Figure 5.1 and table 5.1 reveal that the thus defined zones are relatively comparable in their geographic size, but not in their population. Note that the larger number of zones is the main difference with the trivial Amsterdam-Rotterdam study in chapter 4. Where case study Amsterdam-Rotterdam contains just 5 destination choice options from both of the 2 origin zones, case study Interregional Train Travel contains 15 possible destination choices from 16 origin zones. The resulting total of OD-pairs (see page 35) is a good measure of the complexity of a case study, because each OD-pair represents an independent choice option that needs to be modelled. Where the complexity of case study Amsterdam-Rotterdam is 10 OD-pairs, case study Interregional Train Travel includes 240 OD-pairs. However, even this much increased complexity is still modest compared to the 757,662 OD-pairs in the final application of the destination choice eccentricity model in chapters 6 and 7.

ID	Railway Station	Other major stations served	Population (1995)
1	Leeuwarden	Heereveen	573,940
2	Groningen	Assen	731,540
3	Zwolle	Emmen	921,020
4	Enschede	Hengelo, Almelo	552,260
5	Deventer	Apeldoorn, Zutphen	511,990
6	Alkmaar	Hoorn, Enkhuizen, Den Helder	745,720
7	Amsterdam	Haarlem, Schiphol, Almere	1,703,860
8	Den Haag	Leiden, Gouda	1,386,940
9	Rotterdam	Dordrecht	1,671,360
10	Utrecht	Amersfoort, Ede-Wageningen	1,691,050
11	Arnhem	Nijmegen	943,450
12	Middelburg	Vlissingen	347,060
13	Breda	Roosendaal	595,810
14	Den Bosch	Tilburg	769,190
15	Eindhoven	Venlo, Roermond	1,220,140
16	Maastricht	Heerlen, Sittard	633,680
total			14,999,010

Table 5.1: Case study Interregional Train Travel, zone system

Zone reference system For the case study Amsterdam-Rotterdam, the zone reference system has been made artificially complex, but this is not re-

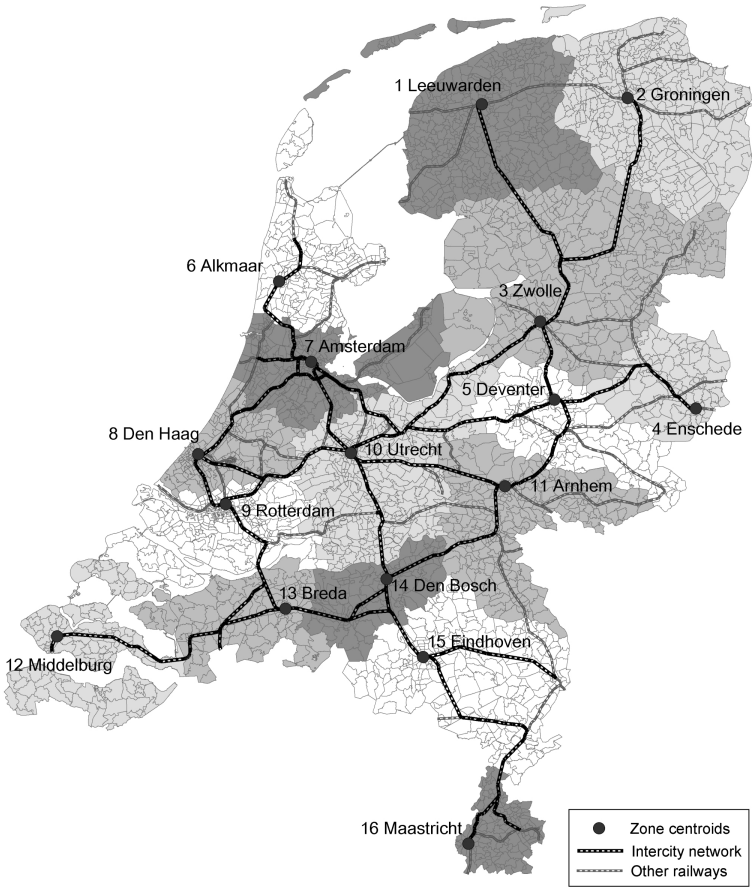

Figure 5.1: Case study Interregional Train Travel, zone system, rail network

quired for this chapter. The high complexity for the case study Amsterdam-Rotterdam was mainly caused by the introduction of a dual zone system so that the model implications of this zone system could be illustrated. This illustration being made, there is no need to adopt such a zone system for the case study Interregional Train Travel because the number of OD-pairs is manageable. The zone reference system may also be kept very simple, because for the case study all zone sets of interest $\mathcal{H}, \mathcal{O}$ and $\mathcal{D}$ can be defined equal (the case study Amsterdam-Rotterdam and the final case study Southern Overijssel in chapter 6 and 7 offer examples where this is not the case). Only for the (base) home zones $h \in \mathcal{H}$, it seems useful to introduce a layer of aggregated home zones $H \in \mathcal{H}_C$ in order to allow for the independent estimation of attractiveness decay functions (see section 3.5). Let this aggregated zone system $\mathcal{H}_C$ distinguish for three categories in which a zone is geographically separated from the Randstad, as indicated by table 5.2. The Randstad is the large ag-

glomeration of cities in the West of the Netherlands, including the four largest cities of the country Amsterdam, Rotterdam, The Hague and Utrecht. The total resulting zone reference system is shown in table 5.3.

H	Separation from the Randstad
91	Distant from the Randstad
92	Adjoining the Randstad
93	Within the Randstad

Table 5.2: Case study Interregional Train Travel, aggregated home zones

Major railway station	h, i, j	H	Major railway station	h, i, j	H
Leeuwarden	1	91	Rotterdam	9	93
Groningen	2	91	Utrecht	10	93
Zwolle	3	92	Arnhem	11	92
Enschede	4	91	Middelburg	12	91
Deventer	5	91	Breda	13	92
Alkmaar	6	92	Den Bosch	14	92
Amsterdam	7	93	Eindhoven	15	91
Den Haag	8	93	Maastricht	16	91

Table 5.3: Case study Interregional Train Travel, zone identification

Network The intercity rail network of the Netherlands is shown in figure 5.1. Nearly all links of this network are served on at least a half-an-hour basis by intercity trains. Apart from this high frequency, Dutch intercity trains are faster, more comfortable, have better connections with other services, but are not more expensive than non-intercity trains. The result is that virtually all interregional train travel in the Netherlands is handled by the intercity network. A single exception is the Leeuwarden-Groningen local service, because using an intercity service for this trip would require a large detour over Zwolle. Thus, all interregional train travel may be considered to be handled by the intercity network alone and by the Leeuwarden-Groningen local service. Table 5.4 displays a list of the travel times on all individual links thus defined. All other railway links shown in figure 5.1 have merely been incorporated in this case study for the assignment of municipalities to each of the 16 major railway stations.

Trip segmentation scheme $\mathcal{M}$ For case study Interregional Train Travel, let a uniform trip segment represent all possible interregional train trips: $\mathcal{M} = \{\text{interregional train trips}\}$. The inclusion of trip segment m in notations such as A_{jm} is thus obsolete, but is maintained for a constant notational style in this report. The assumption of one uniform trip segment is reasonable to make because the case study already limits itself to one specific trip segment only.

Link	Description	Time (min.)	Link	Description	Time (min.)
1-2	Leeuwarden-Groningen	51	7-10	Amsterdam-Utrecht	30
1-3	Leeuwarden-Zwolle	58	8-9	Den Haag-Rotterdam	22
2-3	Groningen-Zwolle	61	8-10	Den Haag-Utrecht	38
3-5	Zwolle-Deventer	30	9-10	Rotterdam-Utrecht	36
3-7	Zwolle-Amsterdam	69	9-13	Rotterdam-Breda	34
3-10	Zwolle-Utrecht	54	10-11	Utrecht-Arnhem	34
4-5	Enschede-Deventer	42	10-14	Utrecht-Den Bosch	28
5-7	Deventer-Amsterdam	68	11-14	Arnhem-Den Bosch	45
5-10	Deventer-Utrecht	58	12-13	Middelburg-Breda	59
5-11	Deventer-Arnhem	36	13-14	Breda-Den Bosch	33
6-7	Alkmaar-Amsterdam	32	14-15	Den Bosch-Eindhoven	19
7-8	Amsterdam-Den Haag	40	15-16	Eindhoven-Maastricht	64

Table 5.4: Case study Interregional Train Travel, network links

Input travel diary database *TB* The Dutch National Travel Survey (OVG) for the year 2001 has been used for obtaining a trip database on interregional train travel. (the OVG is introduced in more detail on page 123). The database includes a total of 130,145 respondents for the whole of the Netherlands out of which 1,808 reported an interregional train trip. For demonstration purposes, only the 1,565 independent interregional train trips are considered, defined as the first train trip observed per household. It follows that each independently observed train trip in fact represents (1,808/1,565) = 1.16 interregional train trips on average. No selection has been made on the day for which this trip has been reported. The observed behaviour thus applies for an average day of the year. Table 5.5 shows the total number of trips T^O_{ijm} observed for each of the 240 OD-pairs of the case study. In addition, table 5.6 shows various aggregate measures of the observed trips in the OVG of 2001 for each of the 16 zones defined.

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	0	18	0	0	1	0	1	0	1	1	0	0	0	0	0	0	22
2	9	0	10	3	0	0	4	3	0	7	1	1	0	1	1	1	41
3	4	12	0	12	2	0	8	6	2	11	1	0	0	0	1	0	59
4	2	1	6	0	11	0	5	3	2	5	0	1	0	0	0	0	36
5	0	1	13	1	0	3	10	3	2	15	24	0	0	3	0	0	75
6	0	0	1	0	1	0	118	4	3	8	2	0	0	0	0	0	137
7	2	1	3	0	7	19	0	72	16	80	4	0	3	2	2	3	214
8	1	3	3	0	1	2	89	0	68	46	4	1	1	2	4	0	225
9	0	1	1	0	0	2	23	73	0	20	4	3	13	2	5	0	147
10	2	3	6	3	6	1	95	37	22	0	14	2	2	17	5	2	217
11	1	0	3	0	7	0	17	6	2	44	0	1	2	18	6	4	111
12	0	0	1	0	0	0	1	5	4	1	1	0	2	2	0	0	17
13	0	0	0	0	0	0	3	12	15	3	0	3	0	14	4	0	54
14	0	0	0	2	0	0	12	3	5	17	13	2	9	0	29	1	93
15	1	0	0	0	1	0	14	5	4	11	6	1	2	21	0	18	84
16	0	0	0	0	0	0	2	2	4	5	1	0	0	4	15	0	33
Σ_i	22	40	47	21	37	27	402	234	150	274	75	15	34	86	72	29	1565

Table 5.5: Case study Interregional Train Travel, observed destination choices

Zone	# respondents in OVG	# total departures	# independent departures	# independent arrivals	max. arrivals ($= W_{jm}$)
1. Leeuwarden	5,134	26	22	22	1,543
2. Groningen	6,206	46	41	40	1,524
3. Zwolle	8,038	68	59	47	1,506
4. Enschede	4,755	43	36	21	1,529
5. Deventer	4,674	86	75	37	1,490
6. Alkmaar	6,461	158	137	27	1,428
7. Amsterdam	14,236	244	214	402	1,351
8. Den Haag	12,680	248	225	234	1,340
9. Rotterdam	12,182	174	147	150	1,418
10. Utrecht	13,951	253	217	274	1,348
11. Arnhem	9,859	135	111	75	1,454
12. Middelburg	3,256	21	17	15	1,548
13. Breda	5,384	63	54	34	1,511
14. Den Bosch	6,377	109	93	86	1,472
15. Eindhoven	11,213	96	84	72	1,481
16. Maastricht	5,739	38	33	24	1,532
Total	130,145	1,808	1,565	1,565	23,475

Table 5.6: Case study Interregional Train Travel, observed trip totals

Population sample ξ It is assumed that the sample ξ of the 1,565 respondents who performed the above-mentioned trips, represents a good cross-section of the total population that is expected to make an interregional train trip for an average day. In addition, it is assumed that persons who have reported an interregional train trip in the OVG of 2001 are not more or less likely to be included in the OVG as other persons. Thus, it follows that in case 1% of all respondents in the OVG of 2001 have reported an interregional train trip from their residential zone h, an approximate 1% of the total population of h is expected to perform such a trip for an average day of the year (in case h is sufficiently large).

Generalised travel costs ψ_{ijm} Table 5.4 shows the intercity travel times for all OD-pairs that are connected to each other by a direct intercity rail link. Based on these link times, let the generalised travel cost ψ_{ijm} for all other OD-pairs be defined from the following rules:

1. Let generalised travel cost be modelled by three components: monetary cost, travel time and discomfort. For simplicity, monetary cost and discomfort are assumed linearly dependent on *intercity* travel time. This assumption is supported by the nearly linear tariff structure of the Dutch Railways for interregional train travel. Consequently, let generalised travel cost be measured in terms of intercity travel time equivalents.

2. For a real-world intercity trip between a physical origin and destination location, the intercity service only constitutes part of the total generalised travel cost. Three additional travel cost components are considered in this study: access time, egress time and transfer time. First, access time and

	Destination															
Origin	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1
2	0	-	0	1	1	1	0	0	0	0	1	1	1	1	1	1
3	0	0	-	0	0	1	0	0	0	0	0	1	0	0	1	1
4	1	1	0	-	0	1	0	0	0	0	1	1	1	1	2	2
5	1	1	0	0	-	1	0	0	0	0	0	1	0	0	1	1
6	1	1	0	1	1	-	0	1	1	0	0	1	1	1	1	1
7	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	-	0	0	1	0	0	0	0	1
9	0	0	0	0	0	1	0	0	-	0	1	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0	-	0	1	0	0	0	0
11	1	1	0	1	0	0	0	1	1	0	-	1	0	0	1	1
12	1	1	1	1	1	1	0	0	0	1	1	-	0	1	1	2
13	1	1	0	1	0	1	0	0	0	0	0	0	-	0	0	1
14	1	1	0	1	0	1	0	0	0	0	0	1	0	-	0	0
15	1	1	1	2	1	1	0	0	0	0	1	1	0	0	-	0
16	1	1	1	2	1	1	0	1	1	0	1	2	1	0	0	-

Table 5.7: Case study Interregional Train Travel, number of train transfers required

egress time relate to the time required to reach the origin major railway station from the true origin location, or the time required to reach the final destination from the destination major railway station. Second, transfer time is the time required for possible changes of trains while travelling from the origin major railway station to the destination major railway station. Both components are now defined by intercity travel time equivalents.

3. For access and egress time, an additional fixed 20 minutes of intercity travel time equivalents is assumed for every OD-pair. This additional travel time represents the travel time from the true departure point to the origin (major) railway station (10 min.) and that from the destination railway station to the true destination point (10 min.). Whereas these travel times may be assumed higher for many cases, the associated monetary travel costs are probably less in comparison to intercity travel time.

4. Second, for transfer time, let a fixed 10 minutes of intercity travel time be assumed for each change of trains. This estimate represents the half hour intercity service frequency on the entire network. This results in an average 15 minutes of transfer time. However, due to the design of the time table, in most cases a less-then-average transfer time is required, hence the 10 minutes estimate. For this estimation, the absence of monetary costs for transfer time is assumed to cancel out the high discomfort cost component associated with transfer time (in respect to intercity travel time). This high discomfort level is caused by uncertainty and the chance of a missed train connection, apart from the physical change of trains and the waiting time itself. The number of changes required is shown in table 5.7

As an example, the travel cost ψ from Zwolle to Alkmaar for this study is computed as follows. Table 5.4 reveals that the shortest path is formed by the combinations of the link from Zwolle to Amsterdam and the link from Amsterdam to Alkmaar. The associated travel costs are 69 and 32 intercity travel minute equivalents respectively. To this travel cost, 20 minutes are added for access and egress time. Finally, table 5.7 shows that for this link, one change of trains is required. Thus the total travel cost for this origin-destination pair amounts to $69 + 32 + 20 + 10 = 131$ intercity travel minute equivalents.

Circle schemes $\mathcal{C}, \tilde{\mathcal{C}}$ The rather limited number of 240 possible origin-destination pairs calls for a rather rough circle scheme $\mathcal{C}$ for the attractiveness estimation. A fine circle scheme would result in few observations for each travel cost circle and thus lead to insignificant observations. Table 5.8 shows the circle scheme $\mathcal{C}$ adopted for the case study. For the estimation of attractiveness decay functions, an alternative circle scheme $\tilde{\mathcal{C}}$ is proposed on page 96 (see table 5.11).

c	$\omega_{c,\min}$	$\omega_{c,\max}$	Intercity travel time equivalent
45	30	60	up to 1 hour
75	60	90	1 hour to 1.5 hours
105	90	120	1.5 hour to 2 hours
135	120	150	2 hours to 2.5 hours
165	150	180	2.5 hours to 3 hours
195	180	300	3 hours and more

Table 5.8: Case study Interregional Train Travel, circle scheme

5.2 Attractiveness estimation

This section discusses the observation of attractiveness A_{jm} based on the input data specified by section 5.1. First, table 5.9 shows the results of five iterations by the attractiveness observation model of section 3.4 to observe the attractiveness A_{jm} of each zone $j \in \mathcal{D}$ from the OVG of 2001. In this table, the attractiveness estimates A_{jmn} for iteration number $n = \{1, 2, 3, 4, 5\}$ are expressed as indexed values of the initial attractiveness estimate (where A^0_{jm1}=1.00). For this case study the population data of a zone (see table 5.1) has been taken as the initial attractiveness estimate A^0_{jm1}. The section discusses the influence of the initial attractiveness estimation, convergence speed and finally, the interpretation of attractiveness for practical applications.

Influence of initial attractiveness estimates On page 62, it was claimed that initial attractiveness estimates can increase convergence speed, but that no initial attractiveness estimation A^0_{jmn} is required. Table 5.9 has taken the first approach by using population data as the initial attractiveness (compare table 5.1). In order to study the claim that no initial attractiveness estimate is required, the same exercise has been performed, but based on uniform initial

Destination zone j	A^0_{jm1} n=1	A_{jmn} (indexed; $A^0_{jm1} = 1.00$) n=1	n=2	n=3	n=4	n=5	A_{jmn} n=5
1.Leeuwarden	573,940	0.80	0.70	0.67	0.65	0.65	372,421
2.Groningen	731,540	0.80	0.76	0.74	0.74	0.73	535,082
3.Zwolle	921,020	0.49	0.49	0.49	0.49	0.49	453,195
4.Enschede	552,260	1.05	0.97	0.93	0.92	0.92	506,014
5.Deventer	511,990	0.61	0.63	0.64	0.64	0.64	328,074
6.Alkmaar	745,720	0.50	0.47	0.45	0.44	0.44	330,454
7.Amsterdam	1,703,860	1.93	2.10	2.15	2.16	2.17	3,690,638
8.Den Haag	1,386,940	1.70	1.70	1.70	1.71	1.71	2,366,885
9.Rotterdam	1,671,360	0.73	0.72	0.72	0.72	0.72	1,209,119
10.Utrecht	1,691,050	0.99	0.99	1.00	1.00	1.01	1,699,957
11.Arnhem	943,450	0.86	0.86	0.86	0.86	0.86	813,347
12.Middelburg	347,060	1.53	1.40	1.35	1.33	1.31	455,905
13.Breda	595,810	0.50	0.49	0.49	0.49	0.49	291,434
14.Den Bosch	769,190	0.77	0.79	0.80	0.81	0.81	621,851
15.Eindhoven	1,220,140	0.71	0.68	0.67	0.67	0.67	,817,507
16.Maastricht	633,680	0.93	0.85	0.82	0.80	0.80	507,122
Total/weighted average	14,999,010	1.00	1.00	1.00	1.00	1.00	14,999,005
91.Far from Randstad	2,838,480	0.97	0.89	0.86	0.84	0.84	2,376,544
92.Adjoining Randstad	5,707,320	0.65	0.64	0.64	0.64	0.64	3,655,862
93.Within Randstad	6,453,210	1.32	1.37	1.38	1.39	1.39	8,966,599
Total/weighted average	14,999,010	1.00	1.00	1.00	1.00	1.00	14,999,005

Table 5.9: Case study Interregional Train Travel, attractiveness estimation per iteration step based on initial attractiveness estimates by population data

attractiveness estimates A^0_{jmn} of $937,438$ or one-sixteenth of the total population. The results of this second attractiveness observation are shown in table 5.10. Note that because the initial attractiveness estimates differ for both tables, all attractiveness estimates expressed as indexes cannot be compared between both tables. However, it can be observed that the final attractiveness estimate for iteration number $n = 5$ in 5.10 is just a few percent different from table 5.9 at maximum. Thus, it can be concluded that the quality of the initial attractiveness estimate indeed seems of little effect on attractiveness estimates in case several iterations cycles are performed.

Convergence speed From both tables 5.9 and 5.10 it can be observed that the attractiveness estimates for most zones radically differ from the initial estimate for the first iteration cycle. In contrast, subsequent iterations hardly change the estimates to a comparable degree. This high convergence speed is a desirable property as it indicates an appropriate modelling of the underlying (first-order) processes that define attractiveness. However, zones which differ from this desirable pattern are zones 1 (Friesland), 2 (Groningen), 4 (Twente), 6 (Noord Holland), 12 (Zeeland) and 16 (Zuid-Limburg). Although this is a considerable number of zones, it can be argued that this slow convergence speed is not as much caused by the model itself, but because of specific characteristics of the case study. First, consider that the slow convergence speed only occurs for locations in the periphery of the Netherlands, far away from the Randstad. This causes the attractiveness estimates of these peripherical zones to be

Destination zone j	A^0_{jm1} n=1	A_{jmn} (indexed; A^0_{jm1} = 1.00) n=1	n=2	n=3	n=4	n=5	A_{jmn} n=5
1.Leeuwarden	937,438	0.70	0.50	0.44	0.41	0.40	377,071
2.Groningen	937,438	0.80	0.67	0.61	0.58	0.57	538,189
3.Zwolle	937,438	0.47	0.47	0.48	0.48	0.48	451,888
4.Enschede	937,438	0.79	0.64	0.58	0.56	0.54	510,693
5.Deventer	937,438	0.33	0.34	0.35	0.35	0.35	328,252
6.Alkmaar	937,438	0.47	0.41	0.38	0.36	0.35	331,730
7.Amsterdam	937,438	3.07	3.62	3.82	3.90	3.92	3,678,337
8.Den Haag	937,438	2.40	2.46	2.50	2.51	2.52	2,363,484
9.Rotterdam	937,438	1.16	1.25	1.28	1.28	1.29	1,207,490
10.Utrecht	937,438	1.55	1.70	1.77	1.80	1.81	1,695,154
11.Arnhem	937,438	0.82	0.85	0.86	0.86	0.87	812,284
12.Middelburg	937,438	0.77	0.61	0.53	0.50	0.49	462,431
13.Breda	937,438	0.30	0.30	0.31	0.31	0.31	291,188
14.Den Bosch	937,438	0.58	0.62	0.65	0.66	0.66	620,016
15.Eindhoven	937,438	0.93	0.88	0.87	0.87	0.87	816,874
16.Maastricht	937,438	0.87	0.67	0.59	0.56	0.55	513,924
Total/weighted average	14,999,008	1.00	1.00	1.00	1.00	1.00	14,999,005
91.Far from Randstad	5,624,628	0.81	0.66	0.60	0.58	0.57	3,219,182
92.Adjoining Randstad	5,624,628	0.49	0.50	0.50	0.50	0.50	2,835,358
93.Within Randstad	3,749,752	2.04	2.26	2.34	2.37	2.39	8,944,465
Total/weighted average	14,999,008	1.00	1.00	1.00	1.00	1.00	14,999,005

Table 5.10: Case study Interregional Train Travel, attractiveness estimation per iteration step based on uniform initial attractiveness estimate

highly influenced by the observations of the number of trips T^O_{ijm} from origin zones that are far away. Because of the rather limited number of observations for case study Interregional Train Travel (1,565 respondents in total but less than 40 respondents for some zones), the number of trips observed on these relations is highly influenced by chance. Only integer trip volumes T^O_{ijm} can be observed such as no trip, one trip, two trips etc. In case a low number of trips is observed to depart from a zone, the associated trip frequencies F^O_{ijm} can only be estimated very roughly. As a consequence, for OD-pairs over a long distance, mostly no trips are observed. However, for the rare cases where a trip is observed towards the peripherical zones from very distant origin zones, the attractiveness estimation model interprets this as the peripherical zone attracting a trip from a distance over which even zones with high attractiveness such as Amsterdam or Utrecht are not attracting trips. This, however, is not caused by the limited attractiveness of Amsterdam and Utrecht, but simply because nor Amsterdam nor Utrecht is connected to any origin zone for such a distance. For these cases, the model rewards the selected peripherical zones by the assignment of a large share of total attractiveness. In table 5.5, it can be observed that this is the case for zone 1 (Friesland attracting one trip from Eindhoven), zone 12 (Zeeland attracting one trip from Twente) and zone 16 (Limburg attracting one trip from Groningen). All three trips are within the largest cost circle of more than three hours of intercity travel time equivalents. No other city attracts a trip from such a distance, because most cities are not connected to cities that far away. Precisely these three zones experience the

slowest convergence speed of the six remote zones mentioned before. However, because these high attractiveness estimates are not sustained by the observed number of trips from all remaining origin zones, later iterations correct for these unrealistic high estimates. Related to these corrections, the attractiveness estimates of the three other peripherical zones are modified upwards.

Practical interpretation of attractiveness This case study measures attractiveness in population equivalents. This is achieved by distributing a number of attractiveness units over all zones that is equal to the total population. Also the subsequent case study in chapter 6 will do so. Despite this linking of attractiveness to population, note that population and attractiveness are theoretically unrelated to each other. The only reason why this study prefers to express the attractiveness A_{jmn} of a zone j in population equivalents is because of an anticipated improved interpretation. Of all possible size variables, the population total of a zone, be it a single village, a town, a region or an entire country, is the one most commonly used. By expressing attractiveness in terms of population equivalents, it is possible to hook on with these common size dimensions. This facilitates interpretation and quickly pinpoints attention to zones where attractiveness estimates differ much from population size. Also, note that attractiveness expresses the attraction force for trip type m. For the case study Interregional Train Travel this excludes all homebound trips. (see page 36). Consider a zone which has a very high population, but with low attractiveness. For case study Interregional Train Travel, this should be interpreted as a zone attracting very little trips from residents of other zones. However, this does *not* necessarily imply that people rarely travel to this specific zone. A typical example of such a zone could be a suburb. Although many people might have their home here, relatively few persons from outside travel towards it. Instead, the observed travel towards it is mostly formed by the residents themselves when they return home (but these trips are not included in trip type segment m). An example of the opposite case; a zone with very low population, but with very high attractiveness might be an inner-city, or any other zone which offers a high amount of facilities in relation to its population total. Other examples are zones with large shopping malls, industrial estates, educational institutes or leisure facilities.

Attractiveness estimates for interregional train travel For the case study Interregional Train Travel, all zones are of considerable size. Therefore, localised effects as described above are expected of minor importance: no zone is fully characterisable as a suburb or an inner-city. As a consequence, attractiveness estimates are expected to represent the population distribution. However, table 5.9 indicates relatively large differences in the ratio attractiveness per inhabitant. While all zones outside the Randstad ($H = 91, 92$) show low attractiveness per capita, all zones within the Randstad ($H = 93$) show much higher attractiveness ratios. For example, each inhabitant in zone Amsterdam represents (3,690,638/1,703,860=) 2.17 units of attractiveness compared to a figure of around 0.80 for inhabitants in 'peripherical' Leeuwarden,

Zwolle, Breda or Alkmaar. Taking into account the aggregation level of these zones, this is a very significant difference. Most likely, the difference cannot be explained by stating that each inhabitant in Amsterdam is surrounded by three times as much attractiveness as the inhabitants of the other cities are. This could be the case if just the city of Amsterdam is compared to just the city of Leeuwarden. However, for this case study, this effect is tempered because 'Amsterdam' and 'Leeuwarden' are understood as large regions around the cities themselves, including regional areas and suburbs. Thus, it is proposed that the high differences in the attractiveness per capita ratio are caused by specific characteristics of the case study. Consider that interregional train travel is in competition with car travel over the highway network. Thus, interregional train travel represents a more favourable option where car travel is relatively less favourable: over often congested routes and/or towards congested cities. Both phenomena are most likely to occur in the Randstad. Although these phenomena indeed also occur for the way back in the afternoon, these are mostly returning residents, ignored in the above attractiveness modelling. Furthermore, consider that the attractiveness of many interregional train trips is influenced by the quality of the public transport system at the destination zone. Again, these systems are better developed for the Randstad. Both factors are likely to explain the high differences in the attractiveness per capita ratio for zones in the Randstad and zones outside the Rnadstad. Yet another factor probably causes the striking difference between attractiveness per capita estimates for Den Haag and Rotterdam. Where the former city appears almost twice the size of Rotterdam in attractiveness units, actually it is Rotterdam that is larger in population. This seems only partly explained by the more densely populated area in the Den Haag zone compared to the one for Rotterdam. The main reason is likely to be that the typical governmental offices for Den Haag attract more interregional person train trips (and more distant ones) as the typical harbours and industry do for Rotterdam.

5.3 Attractiveness decay function estimation

This section demonstrates the observation of attractiveness decay functions $\Phi_{m\xi_H}(\omega)$ for the case study Interregional Train Travel, according to the model specified by section 3.5. The required input data has been specified by section 5.1 and by section 5.2 for attractiveness. This section first specifies a circle scheme required for the observation model. Next, the various steps are discussed for the observation of $\Phi_{m\xi_H}(\omega)$ from the above input data. Finally, a general discussion is given on the interpretation of attractiveness decay functions and the results obtained for the case study.

Circle scheme $\tilde{\mathcal{C}}$ The attractiveness decay observation model of section 3.5 allows for the specification of an independent circle scheme $\tilde{\mathcal{C}}$ compared to that used for the observation of attractiveness. For the case study Interregional Train Travel, let circle scheme $\tilde{\mathcal{C}}$ be composed out of 20-minutes intervals of

intercity travel time cost equivalents, as shown by table 5.11. Although these intervals are rather rough, they are less so in comparison to the circle scheme $\mathcal{C}$ used for the attractiveness observation. This finer circle scheme is possible because the attractiveness decay function is estimated for aggregated home zones H (see table 5.2), instead of for base home zones h. This results in a higher number of observations per home zone, that consequently allows for smaller discrete intervals without decrease in significance. A second argument for a finer zone scheme is that the attractiveness estimates A_{jm} are likely to be more significant than the number of trips T^O_{ijm} which has been used for the attractiveness observation model.

	circle scheme $\tilde{\mathcal{C}}$			attractiveness ratio $\phi^O_{H\tilde{c}m}$		
$\tilde{c}$	$\omega_{\tilde{c},\min}$	$\omega_{\tilde{c},\text{avg}}$	$\omega_{\tilde{c},\max}$	$H = 91$	$H = 92$	$H = 93$
1	40	50	60	-	2.489	1.393
2	60	70	80	7.438	0.930	0.961
3	80	90	100	5.010	0.554	0.368
4	100	110	120	1.160	0.201	0.272
5	120	130	140	0.712	0.107	0.192
6	140	150	160	0.253	0.302	0.104
7	160	170	180	0.320	0.049	0.109
8	180	190	300	0.133	0.017	-

Table 5.11: Case study Interregional Train Travel, attractiveness decay factors (daily trips per 100.000 attractiveness units by 100 individuals performing an interregional train trip)

Circle trip to attractiveness ratios $\phi^O_{H\tilde{c}m}$ Based on the above defined circle scheme $\tilde{\mathcal{C}}$, the aggregated home zones $\mathcal{H}_C$ and the attractiveness estimates in table 5.9, equations 3.24 and 3.26 provide the trip to attractiveness ratios $\phi^O_{H\tilde{c}m}$. These ratios are shown in table 5.11 for all three composite origin zones $H \in \mathcal{H}_C$ and all circles $\tilde{c} \in \tilde{\mathcal{C}}$ in the case study.

Attractiveness decay function $\Phi_{m\xi_H}(\omega)$ By means of equations 3.27 and 3.28, the above attractiveness-to-trip ratios $\phi^O_{H\tilde{c}m}$ are interpolated to obtain estimates of decay function $\Phi_{m\xi_H}(\omega)$ for $\omega = \omega_{\tilde{c},\text{avg}}$ where $\tilde{c} \in \tilde{\mathcal{C}}$ and ξ_H is the sample of respondents that resides within H. As an example, figure 5.2 provides with a graphical illustration of the trip to attractiveness ratios and the interpolated decay function for $H = 92$. The bend in this graph at $\omega = 150$ minutes is discussed at the end of this section.

Attractiveness decay factors $\phi_{rtj\xi_H}$ Based on the decay function for $H = 92$ shown in figure 5.2 and those for $H = 91$ and $H = 92$, all decay factors ϕ_{rtj} for any trip t by all individuals r and for any possible destination zone j can be obtained by entering the associated marginal travel cost ω_{rtj} into $\Phi_{hm}(\omega)$. As an example, the values of ϕ_{rtj} for all destination zones are presented in table 5.12 for an individual r residing in home zone $h = 15$ (Eindhoven) and starting

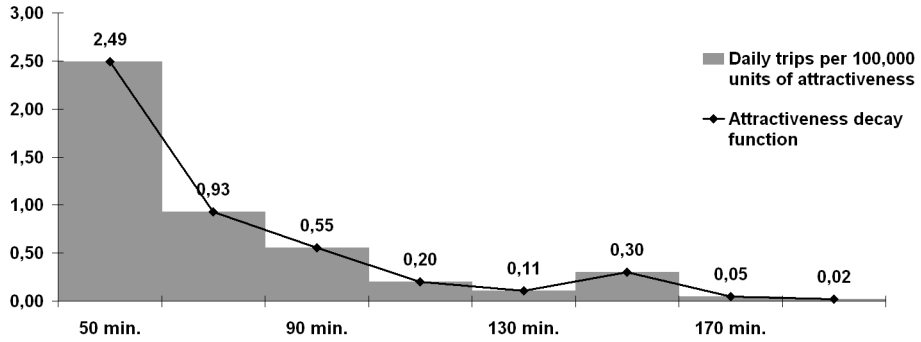

Figure 5.2: Case study Interregional Train Travel, attractiveness decay function $\Phi_{m\xi_H}(\omega)$ for zones adjoining the Randstad (H= 92)

a trip t from there (assuming that $\omega_{rtj} = \psi_{hjm}$). The remaining columns of table 5.12 are discussed in section 5.4.

Zone j	ϕ_{rtj}	A_{jm}	U_{rtj}	$E_{rtj,\min}$	$E_{rtj,\max}$	P_{rtj}	T^E_{ijm}	T^O_{ijm}
Leeuwarden	0.0188	372,421	7,008	0.998	0.999	0.001	0.1	1
Groningen	0.0172	535,082	9,219	0.999	1.000	0.001	0.1	0
Zwolle	0.1167	453,195	52,872	0.967	0.974	0.007	0.6	0
Enschede	0.0299	506,014	15,139	0.996	0.998	0.002	0.2	0
Deventer	0.1069	328,074	35,073	0.962	0.967	0.005	0.4	1
Alkmaar	0.1947	330,454	64,349	0.974	0.983	0.009	0.7	0
Amsterdam	0.4308	3,690,638	1,589,872	0.605	0.818	0.213	17.9	14
Den Haag	0.2896	2,366,885	685,464	0.871	0.962	0.092	7.7	5
Rotterdam	0.3249	1,209,119	392,845	0.818	0.871	0.053	4.4	4
Utrecht	1.1641	1,699,957	1,978,930	0.207	0.472	0.265	22.3	11
Arnhem	0.4837	813,347	393,475	0.552	0.605	0.053	4.4	6
Middelburg	0.2142	455,905	97,675	0.983	0.996	0.013	1.1	1
Breda	0.8927	291,434	260,153	0.472	0.507	0.035	2.9	2
Den Bosch	2.4892	621,851	1,547,937	0.000	0.207	0.207	17.4	21
Eindhoven	-	817,507	-	-	-	-	-	-
Maastricht	0.6671	507,122	338,303	0.507	0.552	0.045	3.8	18
Total	-	14,999,005	7,468,312	-	-	1.000	84.0	84

Table 5.12: Case study Interregional Train Travel, destination choice eccentricities from Eindhoven

Attractiveness decay factor interpretation The attractiveness decay factor ϕ_{rtj} should be interpreted as a discount factor to the attractiveness A_{jm} of zone j, as a consequence of the marginal travel cost ω_{rtj} associated with the choice to travel to j by individual r in trip t (see page 54). Attractiveness decay factors thus measure that a zone j with travel cost independent attractiveness A_{jm} is probably less visited by an individual who needs to make high travel costs to get there. Two general patterns should be expected. First, attractiveness decay factors will generally decrease with marginal travel costs

increasing, all else equal. Second, relatively high values for decay factors can be expected for residents of isolated regions, at least for destination zones in their vicinity. This effect is caused by the high relative attractiveness of the sparse attractiveness located near these zones. For the case study Interregional Train Travel, both patterns can be observed. First, for all three zones, a general decrease of the decay function can be observed for an increase of ω (see table 5.11). An exception to this general pattern is found for $H = 91$ and $H = 92$ for $\omega_{\tilde{c},\text{avg}} = 170$ and $\omega_{\tilde{c},\text{avg}} = 150$ minutes of intercity travel time equivalents. For this cost circle, *more* trips per unit of attractiveness are observed as for zones at a smaller marginal travel cost. In order to explain this pattern, note that it manifests itself precisely for the marginal travel cost range in which the large cities in the Randstad can be reached from most of the base home zones $h \subset H \in \{91, 92\}$. Second, the effect is observed for travel costs higher than two and a half hours of intercity travel time equivalent. Train travel is a relatively favourable option for these ranges of travel cost. The second expected effect, that attractiveness decay factors are high for the vicinity of isolated zones can also be observed in table 5.11. In this table, aggregated home zones in the periphery ($H \in \{91, 92\}$), show much higher ratios for the smallest cost circle compared to that of $H = 93$. This illustrates that all attractiveness that is located in the vicinity of these more isolated regions is much better utilised per unit of attractiveness as it would be if it was located in the urbanised Randstad.

5.4 Destination choice eccentricity estimation

This section combines the attractiveness estimates A_{jm} and the attractiveness decay factors $\phi_{rtj\xi_H}$ obtained in section 5.2 and 5.3. By combining these components, first the destination choice utilities U_{rtj} and ultimately destination choice eccentricities E_{rtj} are determined for each possible destination choice y_{rtj}. The section concludes with a brief discussion on the interpretation of the resulting outcomes.

Destination choice utility U_{rtj} Equation 3.5 defined the utility U_{rtj} of a destination choice as the product of the attractiveness A_{jm} of the selected destination zone and the attractiveness decay factor $\phi_{rtj\xi_H}$. Table 5.12 showed both the attractiveness estimates of all possible destination zones and all attractiveness decay factors for residents of Eindhoven. In the third column, the table also shows destination choice utility estimates U_{rtj} for these residents, obtained by multiplying the first two columns. Note that for residents in Eindhoven, the highest utility for an interregional train trip is offered by Utrecht, followed by Amsterdam and Den Bosch. The high utilities of Amsterdam and Utrecht are explained by their high attractiveness while Den Bosch is included in this list because of its vicinity to Eindhoven.

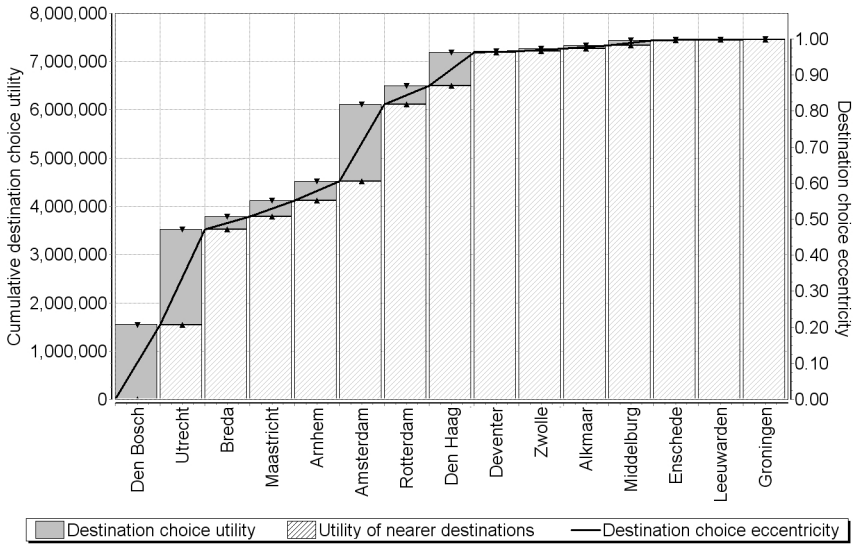

Figure 5.3: Case study Interregional Train Travel, destination choice eccentricity ranges observed for Eindhoven

Destination choice eccentricity range $\mathbf{E_{rtj}}$ From these utility estimates, the destination choice eccentricity range of each destination choice is obtained by equations 4.9 and 4.10. Table 5.12 indicates the resulting ranges by specifying the minimum and maximum eccentricity values $E_{rtj,\min}$ and $E_{rtj,\max}$. The same results are graphically displayed in figure 5.3 for residents departing from Eindhoven. The figure shows the values for $E_{rtj,\min}$ by light gray bars and $E_{rtj,\max}$ by the sum of the light gray and the dark grey bars. Finally, the line drawn through the dark gray bars represents the possible destination choice eccentricity values for all zones j. The left and the right axes of the figure illustrate that destination choice eccentricity represents a standardisation of cumulative utility. In addition, the figure illustrates the need of considering destination choice eccentricity *ranges* instead of point eccentricity values. This is because only eccentricity ranges reflect the indiscriminating effect of the zone system in targeting precise destination choice (see page 81).

Destination choice eccentricity interpretation The interpretation of destination choice eccentricity is two-fold. First, the difference between $E_{rtj,\min}$ and $E_{rtj,\max}$ provides with the expected chance P_{rtj} of destination zone j to be selected by an average individual in destination choice context y_{rtj}. Second, the difference between 0 and either $E_{rtj,\min}$ or $E_{rtj,\max}$ expresses the degree in which such a choice for j is a selective or eccentric choice (see page 40). For example, table 5.12 reveals that an interregional train trip from Eindhoven to Amsterdam is a trip that is expected to occur relatively often: 21.3% of all

trips that start from Eindhoven. However, it is also expected that no less than 60.5% of all trips head to a location that requires less marginal travel cost, while a mere 18.2% of all trips is expected to head to a location with a higher cost. Thus, although the above destination choice is expected to be a rather common one, it also represents a rather eccentric destination choice.

5.5 Destination choice eccentricity validation

The question arises how accurate the above obtained destination choice eccentricities should be judged (and hence the destination choice eccentricity model itself). To answer this question, this section compares the above obtained destination choice eccentricities with external empirical data.

Synthesised destination choice probability P^s_{ij} The measure of destination choice eccentricity is defined jointly by $E_{ij,\min}$ and $E_{ij,\max}$. Section 4.2 has discussed two usages of these measures. First, the (average) height of $E_{ij,\min}$ and $E_{ij,\max}$ is of relevance only in the *ordering* of destination choices for the purpose of the TDD synthesis algorithm of section 2.5. Second, the difference between $E_{ij,\min}$ and $E_{ij,\max}$ is proportional to the chance P_{rtj} that j is selected as the destination of a trip that departs from i. Whereas the first usage of destination choice eccentricity only has meaning within the modelling framework, the second usage also has meaning for the modelled reality and can thus be validated by empirical data of this reality. For this purpose, let the measure of destination choice probability be introduced, denoted by P_{ij} and defined as

> the chance that a person who starts a trip from origin zone i, selects destination zone j as the destination of a trip

Using the available destination choice eccentricity estimates, this destination choice probability can be estimated directly. Let this estimate be denoted by P^s_{ij} and defined as:

$$P^s_{ij} = P_{rtj} = E_{ij,\max} - E_{ij,\min} \qquad (5.1)$$

where $i = O_{rt}$ and $j = D_{rt}$

The thus defined P^s_{ij} is referred to as the synthetic destination choice probability. It represents an estimate of the true but unknown chance P_{ij}. The values obtained for P^s_{ij} for case study Interregional Train Travel are shown by table 5.13. As for subsequent tables, the table shows these chance values by percentile values (e.g. a value of 5.2 indicating a chance of 5.2% or 0.052). Furthermore, in order to obtain a compact overview, percentile values have been rounded to integer values. Because this eliminates much of the precision, especially for low chance estimates, estimates under 1% are shown as one-decimal values (e.g. 0.8% is shown as .8).

	Destination j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	34	26	1	3	.5	11	6	3	10	2	.5	.5	2	.9	.5
2	28	-	28	2	3	.5	10	7	4	1	2	.7	.5	2	1	.7
3	4	6	-	.7	12	.7	30	7	4	3	7	.2	.7	3	1	.2
4	1	1	19	-	13	.8	24	6	3	15	12	.6	.6	1	2	.6
5	.7	.9	14	10	-	1	26	5	3	16	20	.1	.5	3	1	.1
6	.1	.1	.6	.2	.9	-	75	7	2	10	1	.1	.7	.8	1	.1
7	.5	.7	2	1	2	5	-	34	9	29	5	.6	.9	6	3	.7
8	.4	.5	1	.5	.8	1	45	-	17	21	3	.8	2	3	3	.3
9	.4	.6	1	.5	.9	.7	24	35	-	23	3	1	4	4	3	.4
10	.5	.6	3	.9	2	2	37	21	11	-	8	.3	1	6	6	.7
11	.9	1	3	1	7	.5	24	8	5	36	-	.4	1	8	4	1
12	.6	1	.6	.6	.6	.6	15	18	17	5	4	-	24	7	5	.6
13	.0	.2	.7	.2	.6	.9	7	20	27	13	4	4	-	14	8	1
14	.1	.2	1	1	.9	.3	21	11	6	30	8	.4	5	-	15	1
15	.1	.1	.7	.2	.5	.8	21	9	5	27	5	1	4	21	-	5
16	.6	.6	.6	.6	.3	.3	11	6	3	12	2	.6	.9	16	47	-

Table 5.13: Case study Interregional Train Travel, synthetic destination choice probabilities per OD-pair

Observed destination choice probability P^o_{ij} The OVG database of the year 2001 that underlies the destination choice eccentricity estimates and hence the synthetic estimates P^s_{ij}, can also be used to obtain an estimate of P_{ij} directly. Let this second estimate of P_{ij} be denoted by P^o_{ij}, defined as:

$$P^o_{ij} = \frac{T^o_{ij}}{T^o_i} \quad \text{where} \quad T^o_i = \Sigma_j T^o_{ij} \tag{5.2}$$

where T^o_{ij} = the number of interregional train trips observed from zone i towards j in the OVG of 2001. Let this second estimate be referred to as the observed destination choice probability. All values for T^o_{ij} for the OVG of 2001 can be derived from table 5.5. The resulting values are shown by table 5.14, that can be compared to table 5.13.

Improved observed destination choice probability P^{o*}_{ij} The question arises how often both of the above defined estimates P^s_{ij} and P^o_{ij} are correct estimates of P_{ij}. In case both frequencies could be measured, this would quantify the enrichment offered by the destination choice eccentricity model. Even without such a method, in a first analysis of the differences between table 5.13 and 5.14, it should be noted that whereas the first table contains a high number of chances under 1%, the second table contains just two chance in this range. Instead, the second table contains a high number of 0% chance values. Even although these values occur for OD-pairs that are far apart or heading towards destinations of low attractiveness, a destination choice probability of exactly 0 is very unlikely to occur. Rather than representing reality, these zero observations are caused by the high chance factor in observing low destination choice probabilitys for interregional train trips from the OVG of 2001 alone. For

	Destination j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	40	20	0	5	0	10	0	10	10	0	0	0	0	5	0
2	38	-	29	4	0	0	4	4	0	17	0	4	0	0	0	0
3	11	11	-	16	4	0	29	0	2	22	4	0	0	0	0	0
4	5	5	21	-	37	0	0	0	0	11	21	0	0	0	0	0
5	8	8	8	8	-	0	14	0	5	14	35	0	0	0	0	0
6	0	0	0	0	0	-	82	5	0	10	0	0	0	2	2	0
7	1	0	2	1	0	8	-	20	7	49	4	0	0	2	6	1
8	0	1	0	0	0	4	36	-	28	21	3	0	2	0	1	4
9	2	2	0	0	2	0	11	46	-	25	2	2	2	4	5	0
10	.8	0	3	.8	2	0	47	17	10	-	12	0	0	2	3	2
11	0	0	3	0	3	0	19	3	11	35	-	0	0	16	8	3
12	0	5	0	0	0	0	5	10	50	5	10	-	15	0	0	0
13	4	0	0	0	0	4	4	4	30	0	0	13	-	17	22	0
14	3	0	0	0	0	0	6	6	0	24	24	0	9	-	27	0
15	0	0	0	0	0	0	14	7	0	14	3	0	0	41	-	21
16	0	0	0	0	0	0	6	6	0	25	6	0	0	6	50	-

Table 5.14: Case study Interregional Train Travel, observed destination choice chances

example, table 5.5 shows that some departure zones have less than 20 independent interregional train trips departing from them. In such cases, an observed destination choice probability to any of the destination zones can merely be 0%, 5%, 10% etc. This rough observation of destination choice probability for the observed destination choice probabilities P^o_{ij} is a first indication that the associated synthetic estimates P^s_{ij} are probably more accurate. However, in order to confirm this indication, it is required to possess of an accurate observation of destination choice probability for interregional train trips. For this purpose, a number of years of the OVG has been analysed jointly. The OVG was succeeded by the similar *Mobiliteitsonderzoek Nederland* in 2004 (MON), which has been used as well. Assuming that the distribution of interregional train trips has not changed significantly for most 240 OD-pairs during the different years of these databases, such a combined database increases the amount of observations per departure zone and hence the statistical significance of the observed destination choice probabilities. Let this improved, observed estimate of P_{ij} be denoted by P^{o*}_{ij}, defined by:

$$P^{o*}_{ij} = \frac{T^{o*}_{ij}}{T^{o*}_{i}} \quad \text{where} \quad T^{o*}_{i} = \Sigma_j T^{o*}_{ij} \tag{5.3}$$

where T^{o*}_{ij} is the number of interregional train trips from zone i towards j in the combined database. For this report, a combined travel diary database has been created from the OVG databases for 1995, 1998, 2001 and from the MON of 2004. Together, they contain 6,147 observations of independent interregional train trips, compared to the mere 1,565 observations in the OVG of 2001 alone. Appendix C offers the observed values T^o_{ij} for each of these databases individually.

True destination choice probability P_{ij} The combined database contains a minimum of 68 and an average of 384 trips per departure zone. Because of this high number of trips per departure zone, the law of large numbers may be applied which states that

> if repeated random samples of size N are drawn from any population (of whatever form) having a mean μ and a variance σ^2, then as N becomes large the sampling distribution of sample means approaches normality, with mean μ and variance σ^2/N (Blalock Jr., 1960).

The combined database can thus be used to estimate the true but unknown destination choice probabilities P_{ij}. Applying the law of large numbers, the estimation of this chance P_{ij} is normally distributed with a mean $\bar{P_{ij}}$ and a standard variance of σ_{ij}^2 (hereafter denoted in abbreviated form as $N(\bar{P_{ij}}, \sigma_{ij}^2)$). Both distribution parameters can be estimated as

$$\bar{P_{ij}} \approx P_{ij}^{o*} \quad \text{where} \quad P_{ij}^{o*} = \frac{T_{ij}^{o*}}{T_i^{o*}} \tag{5.4}$$

and

$$\sigma_{ij} \approx s_{ij}^{o*}\sqrt{N_i^{o*}} \quad \text{where} \quad s_{ij}^{o*}\sqrt{N_i^{o*}} = \sqrt{\frac{\sum_i \sum_j (T_{ij}^{o*} - \bar{T^{o*}})^2}{N_i^{o*}}} \quad \text{and} \quad \bar{T^{o*}} = \frac{\sum_i \sum_j T_{ij}^{o*}}{N_i^{o*}} \tag{5.5}$$

where N_i^{o*} is the number of trips observed that depart from zone i. For a dichotomised binomial distribution like the destination choice probability P_{ij}, equation 5.5 can be simplified as (Blalock Jr., 1960)

$$\sigma_{ij} \approx s_{ij}^{o*}\sqrt{N_i^{o*}} = \sqrt{\frac{P_{ij}^{o*}(1 - P_{ij}^{o*})}{N_i^{o*}}} \quad \text{where} \quad P_{ij}^{o*} = \frac{T_{ij}^{o*}}{T_i^{o*}} \tag{5.6}$$

Tables 5.15 and 5.16 offer the thus obtained estimates of $\bar{P_{ij}}$ and σ_{ij} respectively. Again, values under 1.0 are rounded to one extra decimal, as for table 5.13 and 5.14.

As an example, the chance P_{ij} of an interregional train trip that departs from Leeuwarden ($i = 1$) to head for Groningen ($j = 2$) is estimated by $\bar{P_{ij}} = 46\%$ and a standard deviation σ_{ij} of a (high) 4.9%. The latter percentage is not to be confused with the relative standard error which for this case amounts to

	Estimate of $\bar{P_{ij}}$ in percentile values per destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	46	15	1	4	1	6	5	5	9	4	1	1	0	3	0	100
2	20	-	26	4	1	.6	15	6	3	14	3	1	2	1	1	1	100
3	9	16	-	14	7	.4	17	6	4	20	4	0	.4	1	2	0	100
4	4	4	20	-	27	2	10	4	2	14	9	.7	1	1	.7	0	100
5	2	2	15	10	-	2	13	4	2	19	27	.4	0	3	.4	0	100
6	.2	0	.5	.2	.5	-	84	5	.9	6	.5	0	.4	.9	.4	.2	100
7	.7	.4	2	.7	2	12	-	29	8	37	3	.4	.8	2	2	1	100
8	.7	.7	.8	.2	.5	.9	34	-	30	23	3	.6	1	1	2	1	100
9	.5	1	.4	.5	1	.9	11	49	-	19	3	2	6	2	2	.5	100
10	.6	.5	5	1	2	.9	43	19	10	-	8	.4	.4	5	2	1	100
11	.9	.9	3	2	6	1	16	8	4	35	-	.4	2	13	7	2	100
12	0	1	1	0	0	1	10	16	31	10	4	-	19	3	1	0	100
13	.4	0	.4	0	0	.4	6	14	30	5	3	7	-	23	10	1	100
14	1	.6	1	1	.6	1	13	6	5	19	12	.6	12	-	26	1	100
15	.6	.6	.9	0	.9	.3	12	7	3	16	8	.3	5	28	-	18	100
16	0	0	0	0	0	0	9	6	7	17	8	0	1	10	42	-	100

Table 5.15: Case study Interregional Train Travel, mean of true destination choice probability (Source: OVG 1995, 1998, 2001 and MON 2004)

4.9%/46% = 10.6%. In contrast, due to the much higher number of observations from a large city like Amsterdam ($i = 7$), the mean chance of an interregional train trip starting from there to head to the city of Utrecht ($j = 10$) is estimated to be 37%, but with a much lower standard deviation of 1.8% (relative standard error of 4.8%). Because for 25 OD-pairs no trips were observed at all in the combined database, the standard deviation of these OD-pairs could not be computed. These OD-pairs are indicated in table 5.17 by an asterisk ($*$). For these OD-pairs, no test can be performed of the accuracy of P^s_{ij} and P^o_{ij} compared to P_{ij}. Consequently, these OD-pairs are excluded from further analysis. Hence, the number of OD-pairs with an estimated $N(\bar{P_{ij}}, \sigma^2_{ij})$ distribution for the destination choice probability P_{ij} of trips starting from i equals $16 * (16 - 1) - 25 = 215$.

Statistical test for the accuracy of synthetic estimates P^s_{ij} Now that the true destination choice probability P_{ij}, distributed by the unknown distribution $N(\bar{P_{ij}}, \sigma^2_{ij})$, is approximated by the observable $N(P^{o*}_{ij}, (s^{o*}_{ij}\sqrt{N^{o*}_i})^2)$, it can be determined how many of the original synthetic estimates P^s_{ij} have been correct (with a given level of certainty). By comparing this number with that for the directly observable estimates P^o_{ij}, it can be quantified how the destination choice eccentricity algorithm has succeeded in enriching the original data. Because the true destination choice probability is given by a stochastic distribution rather than a fixed value, a statistical test has to be used for this purpose. In this test, a balance must be found between two types of errors that should be minimised but which are inversely related. Both types of errors are related to the width of the range around the estimated mean $\bar{P_{ij}}$ of the destination choice probability that is accepted to include the true but unknown destination choice probability P_{ij} for the associated OD-pair. Let the size of this range be

	Estimate of σ_{ij} in percentile values per destination j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-	4.9	3.5	1.0	1.9	1.0	2.3	2.1	2.1	2.8	1.9	1.0	1.0	$*$	1.6	$*$
2	3.0	-	3.3	1.6	.79	.56	2.6	1.8	1.4	2.6	1.4	.79	1.0	.79	.79	.79
3	1.7	2.3	-	2.1	1.6	.38	2.3	1.5	1.2	2.5	1.2	$*$	.38	.65	.75	$*$
4	1.6	1.6	3.4	-	3.7	1.2	2.6	1.7	1.2	3.0	2.4	.72	1.0	1.0	.72	$*$
5	.82	.91	2.3	1.9	-	.82	2.2	1.3	1.0	2.5	2.9	.41	$*$	1.1	.41	$*$
6	.18	$*$	.31	.18	.31	-	1.5	.91	.39	1.0	.31	$*$	.25	.39	.25	.18
7	.31	.24	.52	.31	.52	1.2	-	1.7	1.0	1.8	.61	.24	.34	.52	.53	.44
8	.29	.29	.31	.17	.23	.33	1.6	-	1.6	1.4	.58	.26	.39	.40	.42	.37
9	.31	.43	.25	.31	.50	.39	1.3	2.1	-	1.7	.68	.53	1.0	.65	.65	.31
10	.26	.24	.68	.37	.51	.30	1.6	1.3	1.0	-	.89	.21	.21	.73	.51	.35
11	.45	.45	.77	.63	1.1	.55	1.7	1.3	.93	2.3	-	.32	.59	1.6	1.2	.70
12	$*$	1.5	1.5	$*$	$*$	1.5	3.7	4.5	5.6	3.7	2.5	-	4.8	2.0	1.5	$*$
13	.43	$*$	.43	$*$	$*$	.43	1.5	2.3	3.0	1.4	1.1	1.6	-	2.8	2.0	.75
14	.55	.39	.55	.55	.39	.55	1.8	1.3	1.1	2.1	1.7	.39	1.7	-	2.3	.61
15	.41	.41	.51	$*$	.51	.29	1.7	1.4	1.0	2.0	1.5	.29	1.2	2.4	-	2.1
16	$*$	$*$	$*$	$*$	$*$	$*$	2.4	1.9	2.1	3.2	2.2	$*$	1.0	2.5	4.1	-

Table 5.16: Case study Interregional Train Travel, destination choice probability standard deviation (Source: OVG 1995, 1998, 2001 and MON 2004)

measured by an amount of Z times the estimated standard deviation σ_{ij} under or above the mean $\bar{P}_{ij}$. In case a low Z value is selected, it means that a high chance exists that the true value P_{ij} is located outside this narrow range of values. Thus, the lower the value of Z, the higher the risk that estimates of P^s_{ij} or P^o_{ij} are classified as incorrect, while they are actually correct. This type of error is generally denoted as a type I error. The height of this risk is given by the standard $N(0,1)$ distribution, representing the area under the normal curve for X-values smaller than $-Z$ or larger than Z. In contrast, in case a high value of Z is selected, the range of accepted values becomes that large that ever more estimates of P^s_{ij} or P^o_{ij} are classified as correct, while in fact they should be rejected as such (a type II error). Table 5.17 gives the number of OD-pairs for which an incorrect estimation of destination choice probability is offered by P^s_{ij} and P^o_{ij} for different Z values and with an indication of the associated risk of type I and type II errors. The total number of OD-pairs for which both methods are tested has been computed above as 215. The number of (probably) correct predictions is thus determined as 215 - the amount of incorrect predictions. In addition, figure 5.4, has plotted the Z deviation of the synthetic estimate P^s_{ij} from P_{ij} against the associated Z deviation of the observed estimate P^o_{ij} from P_{ij}.

Where the risk of a type I error is easily quantified, this is not possible for the risk of a type II error. The reason is that such quantification requires knowledge of the distributions of P^s_{ij} and P^o_{ij}, which are unknown. However, the error is known to follow a strictly increasing S-shaped curve with increasing Z value (Blalock Jr., 1960). Using this knowledge, table 5.17 states the relative risk of a type II error for varying values of Z. Because the risk of a type I error is known, using the $N(0,1)$ distribution as discussed above, it is possible to correct

	Risk of error (per OD-pair)		Number of OD-pairs estimated incorrectly (number in brackets corrected for type I error)			
Z	Type I	Type II	synthetic estimate P^s_{ij}		observed value P^o_{ij}	
0.25	80%	negligible	193	(21)	201	(29)
0.38	70%	very low	180	(29)	196	(45)
0.52	60%	low	163	(34)	189	(60)
0.67	50%	low	145	(37)	171	(63)
0.84	40%	average	131	(45)	164	(78)
1.03	30%	average	118	(53)	133	(68)
1.28	20%	high	99	(56)	119	(76)
1.65	10%	high	81	(59)	91	(69)
1.98	5.0%	very high	64	(53)	61	(50)
2.24	2.5%	very high	60	(55)	43	(38)
2.58	1.0%	near certain	45	(43)	30	(28)
3.09	0.2%	near certain	35	(35)	17	(17)

Table 5.17: Case study Interregional Train Travel, number of incorrectly estimated OD-pairs, N = 215

the amount of incorrect predictions for each Z. This correction is possible by noting that a total of 215 OD-pairs is estimated. Thus, for each percent risk of a type I error, the number of correctly estimated OD-pairs classified as incorrect equals 2.15 OD-pairs on average. Estimates in brackets in table 5.17 indicate the number of incorrectly estimated destination choice probabilities after such correction for the risk of type I errors. However, it should be noted that this correction is in itself based on chance. As a consequence, this correction method should be used with care, especially for low values of Z.

Analysis of incorrectly estimated destination choice probabilities Table 5.17 as well as figure 5.4 reveal a differentiated conclusion on the accuracy of P^s_{ij} compared to that of P^o_{ij}. In case only a narrow Z range is accepted around $\bar{P_{ij}}$ then the synthetic estimates P^s_{ij} clearly outperform the directly observed values P^o_{ij}. On average, the directly observed value is incorrect for 20 to 25 OD-pairs more compared to the number of incorrect synthetic estimates. The exact amount is dependent on the chosen analysis value of Z and the associated risk of a type I and type II error that is accepted. Assuming an average value for both errors by allowing a narrow range of $Z = \pm 0.84$ standard errors around P^{o*}_{ij}, the synthetic destination choice eccentricity model offers (215-45)=170 correct estimates out of 215 estimates in total (=79.1%), whereas direct observation of the OVG only gives (215-78)=137 correct estimates for the same total (=63.7%). Thus, at this value of Z, the synthetic estimate is correct for 25% more OD-pairs compared to direct observation. However, table 5.17 also shows that in case Z is increased to (very) high values, the directly observable estimates improve themselves more rapidly compared to the synthetic estimates. In order to analyse the reason why, two tables are presented with the individual OD-pairs with the highest under- and over-estimations of $\bar{P_{ij}}$ by the synthetic estimate P^s_{ij}. The tables are ordered by the amount of Z standard errors $s^{o*}_{ij}\sqrt{N^{o*}_i}$ that the synthetic estimate is under or above $\bar{P_{ij}}$.

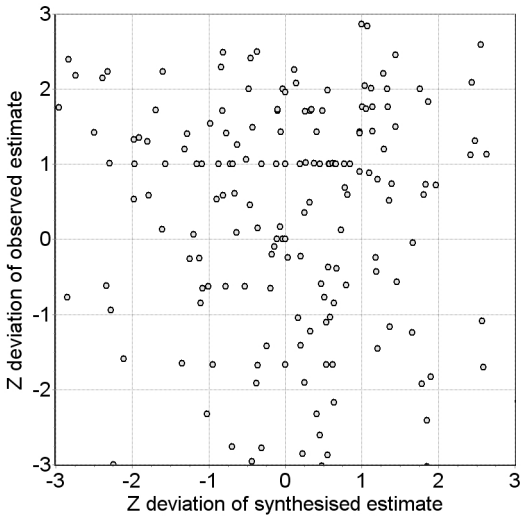

Figure 5.4: Case study Interregional Train Travel, Z deviations from P_{ij} for synthetic and observed estimates

Correction for significantly observed OD-pairs For most OD-pairs included in table 5.18 and 5.19, a significant number of trips T^O_{ijm} has been observed (see table 5.5). It might be wondered how it is possible that for OD-pairs with a significant number of trips T^O_{ijm} and hence a significant observation of P^o_{ij}, the destination choice eccentricity observation model generates estimates P^s_{ij} that are incorrect, even although the model is correct most of the times. And not just are these estimates incorrect, they are even incorrect to such degree that they end up in a top-ten of incorrect estimates, long before all OD-pairs for which an insignificant number of trips has been observed and that may consequently be expected to be much more difficult to estimate correctly. The next section will argue that this is caused by strong OD-pair specific factors not included in the destination choice eccentricity observation model. Although at first sight this may seem a failure of the model, the next section will argue that it actually is a feature of the model because it allows for a quantification of location-specific factors that are intangible otherwise. And because factors that can be quantified can also be incorporated into subsequent modelling, the eccentricity observation model can be extended by a second modelling cycle to eliminate most entries in table 5.18 and 5.19. However, even if this were not the case, it is obvious that for observations that in themselves are significant, any modelled estimate that is deviating may be discarded. Thus, by either approach, the synthetic estimates P^s_{ij} for all of the OD-pairs in table 5.18 and 5.19 can be improved without requiring additional input data. Because the (absolute) Z deviations for all 20 OD-pairs in table 5.18 and 5.19 are all larger than the highest Z in table 5.17, it follows that

	Origin (i)	Destination (j)	P_{ij}^s	P_{ij}	σ_{ij}	Z
1.	Den Haag (8)	Rotterdam (9)	17.39%	29.86%	1.57%	- 7.94
2.	Rotterdam (9)	Den Haag (8)	34.73%	48.85%	2.10%	- 6.73
3.	Eindhoven (15)	Maastricht (16)	4.53%	17.60%	2.06%	- 6.34
4.	Alkmaar (6)	Amsterdam (7)	75.02%	83.92%	1.54%	- 5.78
5.	Amsterdam (7)	Alkmaar (6)	5.44%	11.48%	1.19%	- 5.08
6.	Den Bosch (14)	Eindhoven (15)	14.64%	25.62%	2.29%	- 4.80
7.	Zwolle (3)	Enschede (4)	3.78%	13.69%	2.12%	- 4.68
8.	Zwolle (3)	Groningen (2)	5.57%	15.97%	1.79%	- 4.60
9.	Amsterdam (7)	Utrecht (10)	28.88%	36.51%	2.26%	- 4.26
10.	Den Bosch (14)	Breda (13)	4.73%	11.85%	1.70%	- 4.19

Table 5.18: Case study Interregional Train Travel, OD-pairs with most significant under-estimations of interregional train trips by the destination choice eccentricity observation model

	Origin (i)	Destination (j)	P_{ij}^s	P_{ij}	σ_{ij}	Z
1.	Rotterdam (9)	Amsterdam (7)	23.54%	11.33%	1.33%	+ 9.18
2.	Utrecht (10)	Eindhoven (15)	6.10%	2.47%	0.51%	+ 7.12
3.	Amsterdam (7)	Den Bosch (13)	5.49%	1.94%	0.51%	+ 6.96
4.	Den Haag (8)	Amsterdam (7)	44.86%	34.31%	1.62%	+ 6.51
5.	Deventer (5)	Amsterdam (7)	26.29%	13.17%	2.17%	+ 6.04
6.	Den Bosch (14)	Utrecht (10)	30.44%	18.73%	2.05%	+ 5.71
7.	Breda (13)	Utrecht (10)	12.67%	4.78%	1.41%	+ 5.60
8.	Zwolle (3)	Amsterdam (7)	30.41%	17.49%	2.34%	+ 5.52
9.	Enschede (4)	Amsterdam (7)	24.09%	10.07%	2.55%	+ 5.50
10.	Eindhoven (15)	Amsterdam (7)	21.29%	11.73%	1.74%	+ 5.49

Table 5.19: Case study Interregional Train Travel, OD-pairs with most significant over-estimations of interregional train trips by the destination choice eccentricity observation model

by the above improvement, the number of incorrectly estimated OD-pairs by the synthetic estimate P_{ij}^s are corrected downwards by at least 20 OD-pairs (both for uncorrected and for corrected estimates). This increases the number of correctly estimated destination choice probabilities by the eccentricity observation model up to at least 190 out of 215 pairs (=88.4%). This implies an increase of correctly estimated OD-pairs of at least 39% in comparison to direct observation, almost double the enrichment offered by the initial application of the destination choice eccentricity model. It is concluded that although the destination choice eccentricity model in itself offers significant enrichment of the input travel diary data, this enrichment can potentially be increased by correcting for significant, OD-pair specific attraction. This topic is further discussed in the subsequent section.

5.6 Practical application of output data

The output data for the case study Interregional Train Travel are the attractiveness estimates for all destination zones (see table 5.9), the attractiveness decay functions for three aggregated zones (see table 5.11 for $H = 92$) and

destination choice eccentricity ranges for individual destination choices (see table 5.12 for destination choices from Eindhoven). In comparison with the outcomes for case study Amsterdam-Rotterdam in chapter 4, the outcomes for case study Interregional Train Travel are no longer trivial. This section discusses five practical applications of the output obtained. These are first an enrichment of observed travel diary sample data, second, the generation of trip OD-matrices, third, the quantification of location-specific factors that influence travel behaviour, fourth, extended link analysis and finally, the synthesis of travel diary data. Each application is discussed in the separate paragraphs of this section.

Enrichment of observed travel diary data The validation exercise in section 5.5 has illustrated how the destination choice eccentricity observation model is capable of enriching local samples from a given travel diary database. This is achieved for trip relations for which no significant number of observations is available in the travel diary database. An example is the observation of trip volumes per OD-pair. The enrichment is made possible because the eccentricity observation model observes destination choice location-independent. In doing so, observations from different contexts can be used to estimate destination choice for other locations, trip segments or population groups. This technique allows for large increases of the number of observations available for local trip samples. Data enrichment occurs when the resulting increase in statistical significance outweighs the loss of data due to the transforming of data for areas other than where it was observed for. The validation test in section 5.5 showed that this is the case for a significant number of OD-pairs.

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	8	6	.3	.6	.1	2	1	1	2	.4	.1	.1	.4	.2	.1	22
2	12	-	12	.6	1	.2	4	3	2	4	1	.3	.2	.8	.5	.3	41
3	3	3	-	2	7	.4	18	4	2	12	4	.1	.4	2	.8	.1	59
4	.4	.5	7	-	5	.3	9	2	1	5	4	.2	.2	.5	.6	.2	36
5	.5	.7	10	7	-	.9	20	4	2	12	15	.1	.4	2	.8	.1	75
6	.1	.1	.8	.3	1	-	103	10	2	14	2	.1	.9	1	2	.1	137
7	1	1	5	3	4	13	-	73	19	62	12	1	2	12	7	1	214
8	.9	1	3	1	2	2	101	-	39	48	6	2	5	7	6	.7	225
9	.6	.9	2	.8	1	1	35	51	-	33	4	2	6	5	4	.6	147
10	1	1	6	2	4	3	81	46	24	-	17	1	3	14	13	2	217
11	1	1	3	1	7	.6	27	9	5	40	-	.4	1	9	4	1	111
12	.1	.2	.1	.1	.1	.1	3	3	3	1	.6	-	4	1	.8	.1	17
13	0	.1	.4	.1	.3	.5	4	11	15	7	2	2	-	8	4	.7	54
14	.1	.2	1	.9	.8	.3	19	10	5	28	7	.4	4	-	14	1	93
15	.1	.1	.6	.2	.4	.7	18	8	4	22	4	1	3	17	-	4	84
16	.2	.2	.2	.2	.1	.1	4	2	1	4	.7	.2	.3	5	15	-	33
Σ_i	20	19	56	20	34	23	446	235	125	296	80	11	31	85	73	12	1565

Table 5.20: Case study Interregional Train Travel, expected trip frequencies T^E_{ijm}

Trip OD-matrix estimation Section 5.5 has shown how the destination choice eccentricity observation model can be used to obtain improved estimates of the destination choice probability P_{ij} for each OD-pair compared to direct observation in an existing travel diary database. This results in an improved estimate of the number of trips T^E_{ijm} made by all respondents in the population sample ξ of the input travel diary database. Using knowledge of the representation of the total population by this sample, also the total trip volumes for the entire population can be estimated. On page 90, it was argued that the population sample ξ obtained for case study Interregional Train Travel may be considered as random. Second, it was argued that the fraction of the total population that is expected to make an independent interregional train trip for an average day and from a given origin zone may be estimated by the fraction of respondents in the OVG who reported such a trip from this zone. Next, it need be considered that each independently observed train trip actually represents 1.16 actual interregional train trips (see page 5.1). Taking this scale factor into account, the total amount of interregional train trips T^{OD}_{ij} per OD-pair may be estimated from observed eccentricities:

$$T^{OD}_{ij} = 1.16 * (E_{rtj,\max} - E_{rtj,\min}) * \frac{N^{ITT}_{r,i}}{N_{r,i}} P_i \qquad (5.7)$$

where $i = O_{rt} = H_r, j = D_{rt}$

where $N_{r,i}$ = total number of respondents in the OVG who reside in zone i, $N^{ITT}_{r,i}$ = total number of respondents in the OVG who reside in zone i and for whom an independent interregional train trip has been observed, P_i = population of i. All values for $N_{r,i}$ and $N_{r,i}$ are shown as the first and third column in table 5.6. Population totals P_i are shown in table 5.1 and the eccentricity intervals are shown in table 5.13. As an example, the total amount of daily train trips from Eindhoven to Breda is estimated as $T^{OD}_{ij} = 1.16 * (0.507 - 0.472) * (84/11,213) * 1,220,140 = 367$ trips. Table 5.21 shows all estimates for the Netherlands. Note that these estimates apply for an average day in 2001 because the database T^E_{ijm} is observed for average days in this year. Also, it should be noted that these estimates predict the number of residents in the origin zone who travel to the destination zone. Hence, the figures neither represent trip flows on individual railway links nor trip flows between each pair of zones. Nonetheless are both easily obtained from the figures in table 5.21. First, total trip flows per OD-pair are estimated by adding the number of interregional train trips between the destination and the origin to the figure between origin and destination. This represents that all residents of zone A who travel towards B will just as often travel from B back to A. The total daily number of trips between Eindhoven and Breda is thus determined as $367 + 525 = 892$. Second, trip flows on railway links are easily computed by assigning all of these trip flows to the network. Finally, note that the figures in table 5.21 apply for an average day. For *workdays* these figures

may be expected to be somewhat higher, while they may be lower for holidays. An adjusted observation of T^E_{ijm} could offer estimates for specific day types if desired.

Quantification of location-specific factors that influence destination choice At the end of section 5.5 it was noted that the eccentricity observation model generates unlikely estimates for several OD-pairs. Remarkably, whereas the eccentricity observation model has been shown to increase the accuracy of the observed number of trips for OD-pairs with relatively little trips in between, it is mostly for a subset of the OD-pairs with high trip volumes that the eccentricity model makes the most significant errors. As an example, compare the amount of trips observed in the OVG of 2001 with that expected by the eccentricity model for the OD-pair from Eindhoven to Maastricht. Table 5.12 shows that in the OVG of 2001, 18 trips out of the 84 trips starting from Eindhoven are observed to head to Maastricht (=21%). However, the eccentricity model predicts just 3.8 trips for this OD-pair (=4.5%). That the latter estimate is wrong is indicated by table 5.18. A more significant observation of the number of trips between both cities shows that the actual percentage is distributed around a mean of 17.6% with a standard variance of 2.1% only. Although the observed value of 18 out of 84 trips has probably been too high, the estimated value of 3.8 trips is even worse (and even represents one of the worst estimates for all 215 OD-pairs). In order to explain this large error, note that the destination choice eccentricity range for the Eindhoven-Maastricht OD-pair is estimated under the assumption that Maastricht offers the same attractiveness A_{jm} to inhabitants from Eindhoven as to inhabitants of all other zones. Suppose now that strong socioeconomic or cultural ties would exist between both cities and that it would be the only pair of cities with such strong ties. In observing destination choice behaviour from all other cities, an average behaviour is assessed that does not incorporate these OD-pair specific ties. Clearly, when this average behaviour is transferred to the OD-pair, an estimate of behaviour is obtained of how the behaviour would be in case these OD-pair specific ties would not exist. When compared with reality (where these ties do exist), a large error is the result. Such OD-pair specific ties can both be positive or negative and be the result of historical relationships, administrative ties, strong cultural similarities or of the opposite, cultural animosity. But regardless of the reason of these OD-pair specific ties, the destination choice eccentricity model cannot reproduce such ties as long as they manifest themselves for one specific OD-pair only. This is because the model uses behaviour observed for other OD-pairs to make predictions of the former OD-pair, whereas these other OD-pairs are not influenced by the above ties. This might seem disappointing at first but at a closer look, it can also be used as a feature. Note that where the above mentioned OD-pair specific ties are often intangible, the destination choice eccentricity observation model is able to quantify the strength of these ties. For such quantification, it is necessary that a significant observation of destination choice probability for such an OD-pair is available (that this is relatively likely is illustrated by tables 5.18 and 5.19). Now, the ratio between the observed destination choice

	Destination j							
Origin i	1	2	3	4	5	6	7	8
1. Leeuwarden	-	976	729	39	78	13	300	182
2. Groningen	1,571	-	1,598	82	177	27	546	396
3. Zwolle	330	436	-	291	912	53	2,366	515
4. Enschede	54	66	921	-	641	40	1,161	294
5. Deventer	63	89	1,293	900	-	114	2,496	469
6. Alkmaar	14	13	107	40	161	-	13,776	1,298
7. Amsterdam	202	208	651	346	513	1,607	-	10,072
8. Den Haag	114	152	354	152	240	304	12,770	-
9. Rotterdam	96	144	303	128	208	176	5,522	8,139
10. Utrecht	155	198	846	268	522	451	11,436	6,430
11. Arnhem	112	134	346	145	802	67	2,975	981
12. Middelburg	12	25	13	12	12	12	320	381
13. Breda	-	13	51	13	39	64	512	1,382
14. Den Bosch	14	28	140	126	112	42	2,704	1,401
15. Eindhoven	13	12	76	25	50	88	2,262	973
16. Maastricht	26	26	26	26	13	13	452	233
total	2,726	2,519	7,452	2,594	4,480	3,072	59,582	33,146

	Destination j								total
i	9	10	11	12	13	14	15	16	Σ_j
1	91	286	52	13	13	52	26	13	2,862
2	219	588	136	41	27	110	68	41	5,629
3	278	1,626	569	13	53	238	106	13	7,799
4	147	708	560	26	27	67	80	27	4,817
5	254	1520	1,875	13	50	253	102	12	9,502
6	267	1,914	228	13	121	147	254	14	18,352
7	2,646	8,562	1,607	180	277	1,635	984	194	29,633
8	4,948	6,100	746	228	671	886	709	89	28,465
9	-	5,314	639	288	941	846	622	96	23,460
10	3,399	-	2,369	99	409	1,932	1,862	212	30,587
11	557	4,467	-	44	145	992	479	122	12,366
12	357	111	74	-	492	198	98	13	2,082
13	1,868	871	243	243	-	999	525	89	6,912
14	757	3,964	1,008	56	617	-	1,905	154	13,027
15	556	2,818	556	139	366	2,199	-	480	10,615
16	129	491	91	26	39	659	1,990	-	4,239
total	16,472	39,340	10,753	1,422	4,249	11,162	9,810	1,568	210,347

Table 5.21: Case study Interregional Train Travel, expected daily number of interregional train trips for entire population for 2001

probability and the corresponding synthetic estimate quantifies the strength of such OD-pair specific ties. As an example, for the Eindhoven-Maastricht relationship, local factors are estimated to attract $18/3.8 = 4.7$ times as much interregional train trips as would be expected on average for comparable cities in comparable geographic settings. Now that the strength of this tie is quantified, it can also be corrected for. For example, the utilities of Maastricht for a destination choice from Eindhoven can be multiplied by 4.7. In doing so, more accurate predictions will be made by the destination choice eccentricity model, without requiring any additional data. For example table 5.21 could be improved further by such corrections. In addition, the quantification can be used in the evaluation of such ties and to monitor changes.

Extended link analysis Above, it has been shown how the destination choice eccentricity observation model can be used to estimate trip volumes per OD-pair with an improved accuracy. These trip estimates are the (aggregated) outcome of a modelling at the individual trip level. Although the case study Interregional Train Travel has not yet used its outcomes to synthesise full travel diary data for train-travelling persons in the Netherlands, all of its outcomes (for example the OD-matrix in table 5.21) can be related directly to observations in the source travel diary database. For example, consider that it need be analysed which type of respondents travel from Eindhoven to Zwolle. For this OD-pair, 152 trips are predicted, based on an observed destination choice eccentricity range from 0.967 to 0.974 (see table 5.12). Note that this eccentricity range is not just constituted by persons who travel from Eindhoven to Zwolle but also for OD-pairs that start from other zones. The trip-makers observed for these OD-pairs behave similarly as the ones observed travelling from Eindhoven to Zwolle and thus have contributed to the prediction of 152 trips for this OD-pair. For example, OD-pairs with an overlapping eccentricity range from Utrecht are Utrecht to Alkmaar (0.955 to 0.969) and Utrecht to Enschede (0.969 to 0.978). Because these ranges are only partly overlapping with the above 0.967 to 0.974 range, only part of the associated respondents for these OD-pairs should be selected. From all respondents who reported a trip from Utrecht to Alkmaar, a random selection of $(0.969 - 0.967)/(0.969 - 0.955) = 14\%$ need be made and from all respondents from Utrecht to Enschede, a random selection of $(0.974 - 0.969)/(0.978 - 0.969) = 56\%$ need be made. If this procedure is repeated for each origin zone, the respondents thus selected are the respondents who 'caused' the estimate of 152 daily train trips between Eindhoven and Zwolle. For the whole of the Netherlands, a total of $(0.974 - 0.967) * 1565 = 11$ respondents within the OVG of 2001 can thus be identified. These respondents can be analysed to get an initial insight in the type of person likely to travel from Eindhoven to Zwolle or in the trip context for which such trips are likely to occur. The destination choice eccentricity observation model allows for such analysis even although not a single trip between Eindhoven and Zwolle has been observed in the OVG of 2001. For OD-pairs with a higher eccentricity range, the number of individuals that can be identified by this method is proportionally higher. For example, a total of

333 observed OVG-respondents can be identified to behave similarly as people who travel from Eindhoven to Amsterdam and 415 from Eindhoven to Utrecht. These figures are much higher than the 14 and 11 respondents actually observed for both travel relations.

Generation of synthetic travel diary data The previous paragraphs have provided with four practical applications of the destination choice eccentricity observation model. However, the actual reason why the model has been designed is the generation of synthetic travel diary data, as proposed by chapters 1 and 2. The destination choice eccentricity observation model allows for the location-independent observation of destination choice and consequently represents an implementation of modelling step 8b in the TDD synthesis algorithm of section 2.5. The remainder of this report is devoted to this application. For this application, chapter 6 introduces a third case study where it is tried to generate individual travel diary data for six trip purposes, seven transport modes and for a detailed zone system of 3,766 zones.

5.7 Summary

After the specification of the destination choice eccentricity observation model in chapters 3 and 4, this chapter demonstrated the model for a first real-world application. For this purpose, section 5.1 proposed a case study on interregional train trips, with an interregional train trip defined as a long distance train trip that starts from the home zone. Model specification was kept as simple as possible to enable a fully transparent model application while at the same time as complex as to obtain model outcomes that are not easily deducted otherwise from the input data. Sections 5.2 until 5.4 then estimated the main variables of the destination choice eccentricity observation model: first destination zone attractiveness, then attractiveness decay functions and finally destination choice eccentricity ranges themselves. It was shown that zones in the densely-populated Randstad in the West of the Netherlands show exceptionally high attractiveness, probably because the train is a relatively favourable mode of transport for travelling to these cities. Also it was demonstrated that the convergence speed of the attractiveness estimate is high, which eliminates the need to make initial estimates for attractiveness. For attractiveness decay functions, it was shown that these functions generally decrease with increasing marginal travel cost as expected, but with the exception of those travel costs that represent visits to the large cities in the West from the more peripherical regions in the Netherlands. At the same time, attractiveness decay functions showed high estimates for the immediate vicinity of these peripherical regions, indicating that the relatively sparse attractiveness around these zones is much better utilised than elsewhere. Section 5.5 then made a validation of the output obtained by means of an external database. The section concluded that the destination choice eccentricity observation model has increased the accuracy of the initial data considerably. Some 25% additional OD-pairs were estimated

correctly compared to the observed data for a narrow range of 0.84 times the standard deviation around the true trip volume. Also, it was argued how this percentage could be raised further under a simple extension of the destination choice eccentricity observation model that requires no additional data. Finally, section 5.6 discussed five practical applications of the destination choice eccentricity observation model. Practical applications include the above identified data enrichment, the estimation of OD-matrices, the quantification of otherwise intangible location-specific factors, the ability to perform extended link analyses and the synthesis of travel diary data. This latter application is further demonstrated in the subsequent chapters of this report.

Chapter 6

Case study Southern Overijssel

This chapter generates a synthetic travel diary database for all inhabitants of a large study area in the Netherlands. The selected area is located around a string of cities in the southern part of the Dutch province of Overijssel, housing over 650,000 inhabitants in total. Section 6.1 starts with a list of specifications such as the geographic areas $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{O}$, $\mathcal{D}$ and their zone systems. Sections 6.1 until 6.4 prepare the various input data required for the generation of synthetic travel diary data. Section 6.2 prepares a behavioural database *TB* from the large Dutch National Travel Survey of 1995. Section 6.3 prepares a synthetic population *SP* and a population cluster scheme $\mathcal{P}$ for $\mathcal{Z}$. Section 6.4 prepares a travel cost matrix *TC* for all 757,662 OD-pairs that are defined for the case study. Based on these input databases, section 6.5 concludes with the generation of the synthetic travel diary database. The database is validated in chapter 7 by comparison with two external trip databases.

6.1 Case study specification

This section lists the specifications for which the case study Southern Overijssel is implemented. These include the geographic areas $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{O}$, $\mathcal{D}$ and their constituent zone systems, the trip type segmentation scheme $\mathcal{M}$ and both cost circle schemes $\mathcal{C}$ and $\tilde{\mathcal{C}}$.

Synthesise area $\mathcal{Z}$ The synthesise area $\mathcal{Z}$ defines the area where individual travel behaviour is synthesised for. For case study Southern Overijssel a large geographic area is selected with 658,590 inhabitants for the year 2001. The location and geography of the area is shown in figure 7.1. Apart from its size, the area is considered useful for validation purposes because of its very heterogeneous urban structure. It includes some of the most rural areas in the Netherlands next to a large polycentric, urbanised area in its east (Almelo,

Hengelo, Enschede) and a mono-centric urbanised area in its west (Deventer). In addition, the area is enclosed by strong boundaries with concentrated and relatively limited traffic flows crossing these borders: the international border with Germany in the east, north-east and south-east, the large river IJssel in the west and low-populated rural areas in the north and south.

Residence space $\mathcal{H}$ The residence space $\mathcal{H}$ includes all zones for which individuals are observed in the travel diary input database. Using the algorithm for the synthesis of travel diary data of section 2.5, individuals who reported their travel behaviour and who reside in $\mathcal{H}$ offer 'examples' for the travel behaviour to be synthesised for the synthetic individuals in $\mathcal{Z}$. For case study Southern Overijssel, the residence space $\mathcal{H}$ is defined to include the whole of the Netherlands, including all 114,721 respondents in the travel behaviour database *TB* used for this study (see section 6.2).

Trip origin space $\mathcal{O}$ and trip destination space $\mathcal{D}$ Finally, the trip origin space $\mathcal{O}$ and trip destination space $\mathcal{D}$ model the location from which a synthesised individual s living in $\mathcal{Z}$ or an observed individual r living in $\mathcal{H}$ can start or end a trip. Both areas are defined as the world at large, although with only a basic consideration of locations outside the Netherlands. The precise zone systems of these and the previous zone sets are discussed hereunder.

Zone systems for $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{O}$, $\mathcal{D}$ The most frequently used zone system in the Netherlands is the national postal code system. For this reason, all four zone sets $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{O}$, $\mathcal{D}$ are based on this zone system. The Dutch postal code system is composed out of six digits; four numbers followed by two letters, ranging from 1000AA to 9999ZZ. This study uses three aggregation levels: the fully specified, six-digit postal code level (e.g. 1000AA), the four-digit level (e.g. 1000 = 1000AA until 1000ZZ) and the two-digit level (e.g. 10 = 1000AA until 1099ZZ). The average population is a mere 35 for six-digit zones, 4,500 for four-digit zones and 180,000 for two-digit zones. Table 6.1 specifies how each zone set is specified for the case study Southern Overijssel in terms of precision, zone range and number of zones. For a rough location of each postal code, figure 6.1 shows the two-digit zone system for the Netherlands. An overview map of the synthesise area $\mathcal{Z}$ for case study Southern Overijssel is given by figure 6.2. This latter map includes most of the four-digit postal code zones in $\mathcal{Z}$.

Table 6.1 shows that an identical zone set is selected for $\mathcal{O}$ and $\mathcal{D}_S$ (page 36 showed that this is not necessarily the case). Furthermore, the table shows the inclusion of three foreign zones in addition to the detailed national zone system. Table 6.2 shows the foreign zones modelled in this study. All foreign zones are arbitrarily modelled to have their gravity centre at a fixed travel cost from the Dutch border. The total travel cost from an origin zone to one of these foreign zones is calculated as the sum of the travel cost to the nearest boundary crossing plus an additional (foreign) travel cost as indicated by table

Figure 6.1: Case study Southern Overijssel, road network and two-digit zone system $\mathcal{D}_C$

Zone set	precision	zone range national (# zones)	foreign (# zones)
synthesise area $\mathcal{Z}$	six-digit	7411AA - 7679ZZ (19,269)	-(0)
residence space $\mathcal{H}$			
-base zones $\mathcal{H}$	four-digit	1000-9999 (3,766)	901-903 (3)
-aggregated zones $\mathcal{H}_C$	two-digit	10 - 99 (90)	901-903 (3)
trip origin space $\mathcal{O}$	four-digit	1000-9999 (3,766)	901-903 (3)
trip destination space $\mathcal{D}$			
-single zones $\mathcal{D}_S$	four-digit	1000-9999 (3,766)	901-903 (3)
-composite zones $\mathcal{D}_C$	two-digit	10-99 (90)	901-903 (3)

Table 6.1: Case study Southern Overijssel, zone sets

6.2. The travel cost from a foreign zone back to a national zone is calculated similarly.

ID	Zone	relevant boundary	additional travel cost
702	Germany	nearest German-Dutch boundary	30 min. equivalent
707	Belgium	nearest Belgian-Dutch boundary	30 min. equivalent
734	Other countries	nearest Belgian-Dutch boundary	2 hr. 30 min. equivalent
	(or:)	nearest German-Dutch boundary	4 hr. equivalent

Table 6.2: Case study Southern Overijssel, foreign zones

Trip type segmentation scheme $\mathcal{M}$ A segmentation on trip type allows for the independent modelling of attractiveness, attractiveness decay and thus destination choice utility and destination choice eccentricity for trip segments that are considered heterogeneous in these aspects. However, because the independent modelling of many trip type segments quickly reduces the number of observations per segment and consequently decreases the significance of output, the number of such segments should be kept as low as possible. For this reason, case study Southern Overijssel distinguishes trip type segments according to trip purpose and day type only. Both factors are considered the strongest single determinants of the above trip destination choice aspects. First, for trip purpose, let a special trip segment be reserved for home-going trips. Because the destination choice for such trips does not need to be modelled, trips of this trip segment need not be considered further (see page 36). The resulting trip purpose segmentation is shown by table 6.3.

Second, for day type, let average workdays be considered independent from all other days. An average workday is defined as a Mondays to a Friday, except for all national holidays and all days in July, August and the Christmas holidays. For this report, only average workdays are modelled because all validation data in chapter 7 also applies to this day type. Table 6.3 includes the number of trips observed in the travel diary input database *TB* (see section 6.2) that meet both selection criteria.

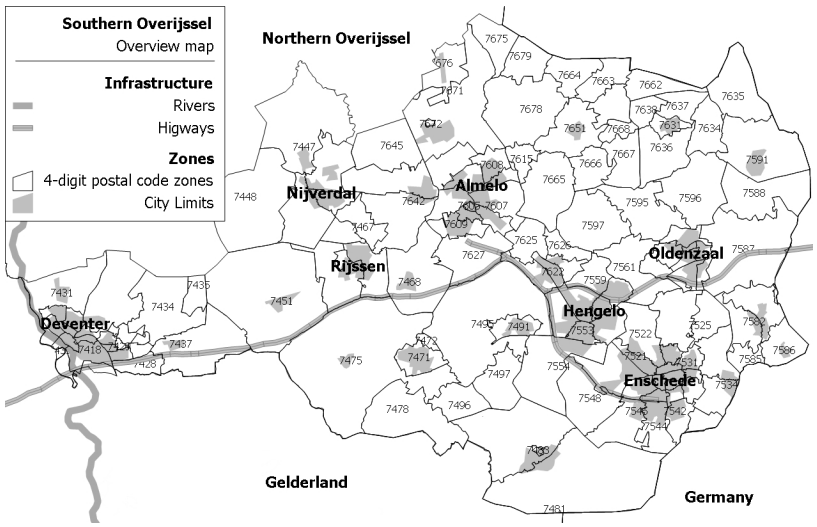

Figure 6.2: Case study Southern Overijssel, overview map of $\mathcal{Z}$

Trip type m (average workday)	Typical destinations	Trip purposes in TB	# trips in TB
0. Home	Private homes	Going home	177,858
1. Work	Offices, industry	Commuting	48,310
2. Education	Schools, universities	Education trips	30,104
3. Shopping	Shops	Shopping	56,988
4. Social	Private homes	Visiting friends and family	33,515
5. Other	Other	Leisure, holiday, touring, escort trips business, religious, medical visits	101,752
Total			448,527

Table 6.3: Case study Southern Overijssel, trip segmentation scheme

Marginal travel cost ω_{rtj} For the case study Southern Overijssel, the marginal travel cost for an individual r travelling to j for trip t is assumed dependent on the location of destination zone j in respect to the origin zone $i = O_{rt}$ and the home zone $h = H_r$ as defined by equation 4.1. This equation estimates ω_{rtj} from generalised travel costs ψ_{ijm}, ψ_{ihm} and ψ_{jhm}. For this case study, generalised travel cost for all trip segments m is assumed linearly dependent on travel time by car over the road network. Thus, travel cost is measured in car travel time minute-equivalents. A matrix TC for travel costs for all OD-pairs is prepared in section 6.4.

Break travel cost ψ^* The break travel cost ψ^* regulates the number of origin zones that are modelled to be connected to composite destination zones instead of single destination zones (see page 51). For case study Southern Overijssel, let this break travel cost be selected at 20 minutes. The model was also prepared with settings for 15 and 25 minutes in order to find an optimal

setting for ψ^*. The result of this analysis is shown in table 6.4.

	# OD-pairs with $T^O_{ijm} \geq 1$ and $\tilde{j} \in \mathcal{D}_S$			# trips T^O_{ijm} and $\tilde{j} \in \mathcal{D}_S$		
trip segment	ψ^*=15m.	ψ^*=20m.	ψ^*=25m.	ψ^*=15m.	ψ^*=20m.	ψ^*=25m.
$m = 1$	15,256	18,202	20,174	34,493	37,730	40,342
$m = 2$	8,020	8,747	9,175	26,980	27,712	28,258
$m = 3$	7,640	10,645	10,972	52,336	53,418	54,169
$m = 4$	8,489	9,170	9,573	27,445	28,555	29,427
$m = 5$	17,128	18,633	19,607	88,026	90,691	92,788
$1 \leq m \leq 5$	33,445	39,843	43,539	229,280	238,106	244,984
	# OD-pairs with $T^O_{ijm} \geq 1$			# trips in T^O_{ijm}		
$\tilde{j} \in \mathcal{D}_S$	33,445	39,843	43,539	229,280	238,106	244,984
$\tilde{j} \in \mathcal{D}$	46,985	50,079	52,519	262,899	263,091	263,251
percentage	71.2%	79.6%	82.9%	87.2%	90.5%	93.1%
	# OD-pairs					
with $T^O_{ijm} \geq 1$	46,985	50,079	52,519			
with $T^O_{ijm} \geq 0$	616,176	766,032	973,526			
percentage	7.63%	6.54%	5.39%			

Table 6.4: Case study Southern Overijssel, influence of break travel cost on model precision and model efficiency

For the selection of an optimal value for ψ^*, two aspects have to be weighted against each other. These are first the resulting model precision and second the resulting computational model efficiency. First, model precision is associated with the percentage of OD-pairs and trips that is modelled within the detailed zone layer $\mathcal{D}_S$. Table 6.4 shows that model precision increases with higher settings for ψ^*, but less so between $\psi^* = 20$ and $\psi^* = 25$ compared to the change from $\psi^* = 15$ to $\psi^* = 20$. Second, computational efficiency is first determined by the total number of OD-pairs that need be modelled and second by the percentage of these OD-pairs that actually attract one or more trips of the 263,000 trips in total. Where the total number of OD-pairs increases steeply to 973,526 for $\psi^* = 25$ compared to 616,176 pairs for $\psi^* = 15$, the percentage of OD-pairs that actually attract at least one trip decreases from 7.63% down to 5.39%. Considering the increase in model accuracy with higher values for ψ^*, the reduced incremental increase from 20 to 25 minutes, and the rapidly increasing total number of OD-pairs, especially from 20 to 25 minutes, the break time value of 20 minutes is selected as an optimal value for this case study.

Circle schemes $\mathcal{C}, \tilde{\mathcal{C}}$ On page 52, two guidelines are offered for the design of a circle scheme. For case study Southern Overijssel, the average zone size is estimated at 7.5 km of diameter. Using equation 3.2 and an average speed of 45 km/hour (see table 6.10), this results in a minimum standard error by the zone system of around 3 minutes for travel costs. Considering this error, a circle interval of at least 5 travel minutes is proposed for the case study. Furthermore, in reflection of the reduced number of observations for high travel costs, this interval size is increased for high travel costs. Table 6.5 presents the

resulting circle scheme. It is proposed both as the circle scheme $\mathcal{C}$ for the attractiveness estimation as for circle scheme $\tilde{\mathcal{C}}$ for the attractiveness decay function calibration.

Circle c, $\tilde{c}$	$\omega_{c,\text{avg}}$, $\omega_{\tilde{c},\text{avg}}$	$\omega_{c,\text{min}}$, $\omega_{\tilde{c},\text{min}}$	$\omega_{c,\text{max}}$, $\omega_{c,\text{max}}$
	(car travel time equivalents)		
1	2.5 min.	0 min.	5 min.
2	7.5 min.	5 min.	10 min.
3	12.5 min.	10 min.	15 min.
4	17.5 min.	15 min.	20 min.
5	22.5 min.	20 min.	25 min.
6	27.5 min.	25 min.	30 min.
7	32.5 min.	30 min.	35 min.
8	37.5 min.	35 min.	40 min.
9	42.5 min.	40 min.	45 min.
10	47.5 min.	45 min.	50 min.
11	52.5 min.	50 min.	55 min.
12	57.5 min.	55 min.	60 min.
13	1 hr. 05 min.	1 hr. 00 min.	1 hr. 10 min.
14	1 hr. 15 min.	1 hr. 10 min.	1 hr. 20 min.
15	1 hr. 25 min.	1 hr. 20 min.	1 hr. 30 min.
16	1 hr. 35 min.	1 hr. 30 min.	1 hr. 40 min.
17	1 hr. 45 min.	1 hr. 40 min.	1 hr. 50 min.
18	1 hr. 55 min.	1 hr. 50 min.	2 hr. 00 min.
19	2 hr. 15 min.	2 hr. 00 min.	2 hr. 30 min.
20	2 hr. 45 min.	2 hr. 30 min.	3 hr. 00 min.
21	3 hr. 30 min.	3 hr. 00 min.	∞

Table 6.5: Case study Southern Overijssel, circle schemes

6.2 Input travel diary database *TB*

This section prepares the input travel diary database *TB* for case study Southern Overijssel. For the synthesis of travel diary data, this database serves three purposes. First, it provides the observed number of trips T^O_{ijm} for each OD-pair. This is an important input variable for both the estimation of attractiveness A_{jmn} as for the calibration of attractiveness decay functions $\Phi_{m\xi_H}(\omega)$. Second, the database should contain socioeconomic data on the individuals whose travel behaviour it stores. Based on these variables, the population cluster of each respondent can be determined, which is used for the matching of respondents r ($r \in TB$) with synthesised individuals s ($s \in SP$, see page 33). Finally, the travel behaviour database should provide the observed travel behaviour vector $\mathbf{B}_r$ that is transferred from its geographic setting $\mathbf{\Gamma}_r$ to become new, synthetic behaviour $\mathbf{B}_s$ within geographic setting $\mathbf{\Gamma}_s$ (see section 2.4). This section first introduces the database used for the case study Southern Overijssel. Typical for such large, raw databases is the high amount of missing or inconsistent data entries. For this purpose, an algorithm is introduced to correct for many of these raw data problems in the remainder of the section.

Raw data source The travel behaviour database used for this study is the Dutch National Travel Survey (*Onderzoek Verplaatsingsgedrag*, OVG) of the year 1995. This database is collected at an annual basis by Statistics Netherlands (*Centraal Bureau voor de Statistiek*, CBS) for the Dutch Ministry of Transport. The survey has been running since the 1970s and targets a 1% response rate of the entire Dutch population per year. The trip origin and destination data in the OVG is aggregated for the municipal level, except for the 1995 edition that was also published at the four-digit postal code level. This low aggregation level makes it attractive for the application of the TDD synthesis algorithm in this report because it enables output that is defined at this low aggregation level. The survey was succeeded in 2004 by the similar *Mobiliteitsonderzoek Nederland* (MON) which publishes all of its origin and destination data at the level of four-digit postal codes.

Raw data specification The OVG encompasses a wide range of over 150 variables. This paragraph lists all variables used for the case study and introduces an associated notation scheme. Appendix A specifies the possible values for each of these variables. Two data groups can be distinguished; respondent data and trip data. First, respondent data should provide socioeconomic information of each respondent r, so that it can be assigned a population cluster $P_r \in \mathcal{P}$ (see section 6.3). Furthermore, it should include the residential zone of each respondent, which is a key variable in the analysis of trip destination choice behaviour.

- r = Respondent identifier (see page 32)
- H_r = Home zone of r (H_r expected $\in \mathcal{H}$, see page 37)
- NT_r = Flag indicating that r did not report any trips
- DH_r = Flag indicating that r started the first trip from home
- AC_r = Age category of r
- FS_r = Family status of r
- SP_r = Societal participation of r
- CP_r = Car possession of r
- HC_r = Household composition of r
- HS_r = Household size of r
- W_r = Household income of r

Second, trip data should provide the actual travel behaviour at a trip level. For case study Southern Overijssel, these variables are proposed for each trip t:

- t = Trip counter (see page 37)
- DH_{rt} = Flag indicating that trip t started at the home H_r of r
- O_{rt} = Origin zone of trip t by r (O_{rt} expected $\in \mathcal{O}$, see page 37)
- D_{rt} = Destination zone of trip t by r (D_{rt} expected $\in \mathcal{D}$, see page 37)
- $T_{dep,rt}$ = Departure time of trip t by r
- $T_{arr,rt}$ = Arrival time of trip t by r
- M_{rt} = Trip purpose of trip t by r (or activity undertaken at D_{rt})
- MOT_{rt} = Main mode of transport of trip t by r

Raw data omissions and inconsistencies Most respondents of the OVG have been asked to complete a travel diary by themselves. This being a relatively low-cost and thus efficient survey method, it has the disadvantage of resulting in many data-omissions and inconsistencies. This effect is fortified by the many variables sampled from each respondent, including several that should be answered for each trip. Also the zones reported by the respondent are not always included in $\mathcal{H}$, $\mathcal{O}$ and $\mathcal{D}$, due to the reporting of non-existing zones or an inaccurate zone specification. These factors result in only around 60% of the available respondents in the raw data to have provided valid entries for *all* variables listed above. The remaining 40% contains at least one missing, ambiguous or inconsistent data-entry. The simple solution of ignoring these 40% of observations for case study Southern Overijssel is undesirable for two reasons. First, it would result in a much smaller behavioural database available. This in turn would lead to synthetic travel diary data that is of a lower statistical significance. Second, it would introduce a bias for respondents who report few trips. This is because respondents who report more trips are more likely to skip some questions or provide with incomplete entries. As an alternative procedure, appendix E proposes a rule-based algorithm to reduce the amount of incomplete or inconsistent records in the OVG. This algorithm is based on inspection of the respondents with partly missing or inconsistent data. Application of these rules increased the number of respondents with fully specified and fully consistent behaviour up to 111,484 respondents out of the available total of 114,718 respondents (=97.2%).

Number of respondents The total observed number of respondents with fully specified and fully consistent travel behaviour is shown per two-digit postal code zone by table 6.6. The location of these zones is shown in figure 6.1 earlier in this chapter.

6.3 Synthetic population *SP*

No synthetic population *SP* was readily available for the synthesise area $\mathcal{Z}$. Therefore, the algorithm of appendix A has been used to estimate *SP* from a zonal socioeconomic database *SE* for $\mathcal{Z}$. This database was kindly made available by Wegener Direct Marketing BV, who offers comparable databases for the whole of the Netherlands. The average size of the six-digit postal code zones in *SE* equals a mere 17 households or 35 individuals for 17,706 zones in total. Because of this high level of detail, *SE* offers a sound basis for the estimation of a synthetic population. As an excerpt of this database, table 6.7 shows the number of six-digit postal code zones in *SE* per income level.

Creation of synthetic population From the above specified socioeconomic database *SE*, the algorithm of appendix A has been used to generate a synthetic population *SP* for the entire synthesise area $\mathcal{Z}$. The resulting synthetic population comprised of 282,988 households and 658,590 individuals. Table 6.7

Zone	# Resp.	Zone	# Resp.	Zone	# Resp.	Zone	# Resp.	Zone	# Resp.
10	1,782	28	925	46	956	64	1,272	82	1,042
11	1,562	29	1,867	47	1,242	65	1,549	83	934
12	1,182	30	2,128	48	1,639	66	1,400	84	904
13	1,012	31	1,695	49	687	67	1,106	85	375
14	1,468	32	1,963	50	1,952	68	1,350	86	373
15	1,021	33	1,802	51	1,188	69	1,276	87	270
16	1,352	34	1,482	52	1,676	70	1,046	88	348
17	1,731	35	1,272	53	1,374	71	960	89	559
18	1,070	36	652	54	1,734	72	986	90	598
19	1,307	37	1,802	55	1,378	73	1,391	91	385
20	1,224	38	2,147	56	2,336	74	1,875	92	1,225
21	1,441	39	1,850	57	1,509	75	2,225	93	490
22	2,290	40	471	58	707	76	1,470	94	917
23	1,481	41	843	59	1,612	77	1,016	95	468
24	980	42	799	60	1,678	78	950	96	954
25	1,864	43	1,151	61	1,405	79	1,160	97	1,360
26	1,862	44	815	62	1,380	80	1,636	98	313
27	1,299	45	741	63	800	81	945	99	770
							total respondents:		111,484

Table 6.6: Case study Southern Overijssel, number of respondents per two-digit postal code area

shows some example characteristics of these synthetic households $f \in \mathcal{F}_z$ and synthetic individuals $s \in SP$. For example, it can be seen that the population of $\mathcal{Z}$ is somewhat less wealthy as the average population of the Netherlands is, although the region shows an above-average share of high incomes. A specification of income-levels and household types is offered by tables A.2 and A.1 in appendix A.

Population cluster scheme $\mathcal{P}$ The population cluster scheme controls which respondents in *TB* can be matched to synthesised individuals in *SP*. A population cluster scheme can be based on any variable included in both *TB* and *SP*. For case study Southern Overijssel, a population cluster scheme $\mathcal{P}$ is defined with 98 independent population segments, based on five socioeconomic variables. These variables are age, social participation, household type, car availability and household income. The clustering is selected such that each cluster includes a minimum number of around 500 respondents in *TB*. Because of its size, the resulting cluster scheme is included separately as appendix D.

Validation of synthetic population generating algorithm The population cluster scheme in appendix D includes three frequency statistics for each population cluster. These include the number of respondents in *TB*, a projection of this number of respondents for all inhabitants in $\mathcal{Z}$ and finally the number of synthetic individuals in *SP* that belong to each population cluster. These statistics are further explained and discussed in appendix D. The discussion shows that significant differences between the latter two statistics are mostly explained by special characteristics of the synthesise area $\mathcal{Z}$. The

Zones $z \in \mathcal{Z}$ (N = 17,706)		Households $f \in \mathcal{F}_z$ (N = 282,988)		Individuals $s \in SP$ (N = 658,590)	
Characteristic	%	Characteristic	%	Characteristic	%
Zonal income level W_z		Household type HT_f		Age category G_s	
-high	5.9%	-household type 1	2.3%	-child, $\leq$ 9 yr	13.1%
-above-average	23.3%	-household type 2	2.5%	-child, 10-17 yr	9.4%
-average	38.2%	-household type 3	3.4%	-child, $\geq$ 18 yr	3.6%
-under-average	15.1%	-household type 4	15.8%	-main, $\leq$ 35 yr	29.1%
-low	4.6%	-household type 5	12.2%	-main, 35-59 yr	30.9%
mixed	13.0%	-household type 6	10.1%	-main, $\geq$ 60 yr	13.9%
		-household type 7	21.7%	Social participation	
		-household type 8	21.1%	-employed	40.0%
		-household type 9	10.8%	-student	2.4%
		Household income level W_f		-retired	8.7%
		-high	8.9%	-other	48.9%
		-above-average	22.7%	Car possession $S_{\text{car},s}$	
		-average	33.2%	-yes	67.9%
		-under-average	23.4%	-no	32.1%
		-low	11.8%		

Table 6.7: Case study Southern Overijssel, summary of synthetic population generation

appendix concludes that the algorithm for the generation of synthetic populations in appendix A performs well for the Wegener database and the population cluster scheme $\mathcal{P}$.

Excerpts of synthetic population As an illustration, two excerpts are given of the synthetic population generated for case study Southern Overijssel. First, table 6.8 shows the amount of households per four-digit postal code zone in the city of Enschede for all of the nine household types distinguished (see table A.1 in appendix A). As a second example, table 6.9 shows the number of synthetic individuals synthesised per population cluster in the municipality of Enschede (see appendix D).

6.4 Travel cost matrix TC

For the observation of attractiveness and for the calibration of attractiveness decay functions, the generalised travel cost ψ_{ijm} need be known for each OD-pair. For studies with a detailed zone system, such a travel cost matrix is typically not readily available. For example, the current case study includes 3,776 x 3,776 = 14,258,176 OD-pairs. At the same time, as a consequence of the dual zone system and the discrete observation of travel cost by the circle scheme, a high accuracy of this matrix is only of secondary importance. Therefore, even in case detailed travel cost matrices would be available, they would probably not justify their cost. As an alternative to such matrices, appendix F proposes a rough travel cost estimation algorithm, based on the input of a basic highway network only.

	Household type (see table A.1)									
Zone	1	2	3	4	5	6	7	8	9	total
7511	971	419	652	421	657	633	395	396	82	4,626
7512	260	178	191	417	432	346	401	429	116	2,770
7513	227	170	198	437	358	349	424	284	62	2,509
7514	173	107	153	337	305	241	303	368	109	2,096
7521	301	285	446	766	678	417	788	592	162	4,435
7522	1,706	160	203	419	508	425	444	412	58	4,335
7523	433	384	370	671	680	532	637	524	192	4,423
7524	5	65	69	129	187	190	143	191	53	1,032
7525	13	2	15	15	14	7	21	26	15	128
7531	151	259	237	641	562	505	597	680	199	3,831
7532	38	73	136	387	217	203	377	353	128	1,912
7533	124	102	193	360	325	193	347	352	105	2,101
7534	120	169	146	813	555	355	782	601	209	3,750
7535	58	65	91	303	222	231	282	258	83	1,593
7541	81	78	134	216	149	162	231	256	88	1,395
7542	116	282	295	829	673	346	806	821	315	4,483
7543	316	197	230	608	358	329	591	450	159	3,238
7544	259	412	748	1,202	1,039	817	1,153	1,058	433	7,121
7545	411	402	535	1,296	948	649	1,245	1,144	347	6,977
7546	67	113	87	777	452	321	777	602	245	3,441
7547	7	8	11	32	47	37	35	56	16	249
7548	15	36	110	181	212	193	148	218	44	1,157
total	5,852	3,966	5,250	11,257	9,578	7,481	10,927	10,071	3,220	67,602

Table 6.8: Case study Southern Overijssel, number of households per household type for Enschede

Road network The road network used for the travel cost estimation algorithm of appendix F is shown in figure 6.1. The network was obtained from the *Nationaal WegenBestand* (National Road Inventory, NWB) for the year 1995, as published annually by the Dutch Ministry of Transport. The NWB database was also used to specify the locations where this network could be accessed (highway ramps etc.). Figure 6.1 also shows several boundary roads as discussed in appendix F. The thicknesses of links represent different road classes. Table 6.10 lists these road classes and the associated travel speeds (see also appendix F).

Exclusion of OD-pairs for the dual zone system The dual zone system for destination zones requires the selection of a set of destination zones that may be chosen from each independent origin zone. For the case study Southern Overijssel, the following OD-pairs are considered valid choice options:

1. All OD-pairs from origin zones i towards *single* destination zones $\tilde{\jmath}$ for which the travel cost from i to the associated composite destination zone J ($\tilde{\jmath} \in \mathcal{D}_{S,J}$) is less than break travel cost ψ^* (see page 6.1).

2. All OD-pairs from origin zones i towards *composite* destination zones J for which the travel cost from i to J equals or exceeds ψ^*.

ID	# Indiv.	ID	# Indiv.	ID	# Indiv.	ID	# Indiv.	ID	# Indiv.
1	5,486	21	224	41	1,962	61	1,739	81	941
2	6,051	22	295	42	803	62	1,287	82	1,298
3	4,752	23	308	43	2,775	63	1,582	83	433
4	2,830	24	197	44	522	64	4,555	84	824
5	454	25	1,022	45	529	65	3,449	85	779
6	3,571	26	2,394	46	488	66	1,513	86	651
7	4,025	27	2,360	47	519	67	872	87	989
8	4,898	28	1,615	48	711	68	862	88	1,182
9	304	29	2,302	49	1,184	69	712	89	1,973
10	237	30	2,877	50	286	70	1,201	90	960
11	173	31	3,651	51	350	71	1,494	91	931
12	443	32	1,741	52	352	72	2,279	92	242
13	1,149	33	1,558	53	1,010	73	1,085	93	1,106
14	88	34	1,261	54	1,609	74	2,105	94	1,106
15	170	35	1,497	55	1,484	75	1,094	95	900
16	331	36	1,726	56	1,172	76	1,482	96	1,395
17	539	37	2,557	57	1,703	77	2,499	97	1,779
18	439	38	2,843	58	2,091	78	273	98	2,612
19	490	39	2,895	59	2,823	79	368	total:	
20	624	40	963	60	1,913	80	316		146,494

Table 6.9: Case study Southern Overijssel, number of individuals per population cluster for Enschede

Link type	within Randstad	outside Randstad
Highways	90 km/h	110 km/h
Major regional roads	60 km/h	
Connecting links	40 km/h	
Ferry routes	20 km/h	
	Single zone	Composite zone
Access link	40 km/h	50 km/h
Direct link	45 km/h	

Table 6.10: Case study Southern Overijssel, link speeds

For the above OD-pairs, the generalised travel cost is determined by appendix F. The travel cost for all other OD-pairs is set to ∞ to indicate that it is not a valid choice option. As an example, figure 6.3 illustrates which destination zones can be selected from the city centre of Deventer (origin zone 7411). The figure shows that persons who start a trip from 7411, can either choose from fine-mazed, single destination zones that are near, or from composite destination zones that require a higher travel cost. In addition, the figure uses shades to indicate travel cost with lighter shades representing higher travel cost. The above assignment method uses the travel cost from zone 7411 towards the centre of adjoining composite destination zones in order to determine if this composite destination zone is to be modelled as a whole or by its substituent single destination zones. This is the main reason why figure 6.3 does not show a neat circle surrounding surrounding zone 7411 in fine-mazed zones. The other reason is the different qualities of the road network in different directions.

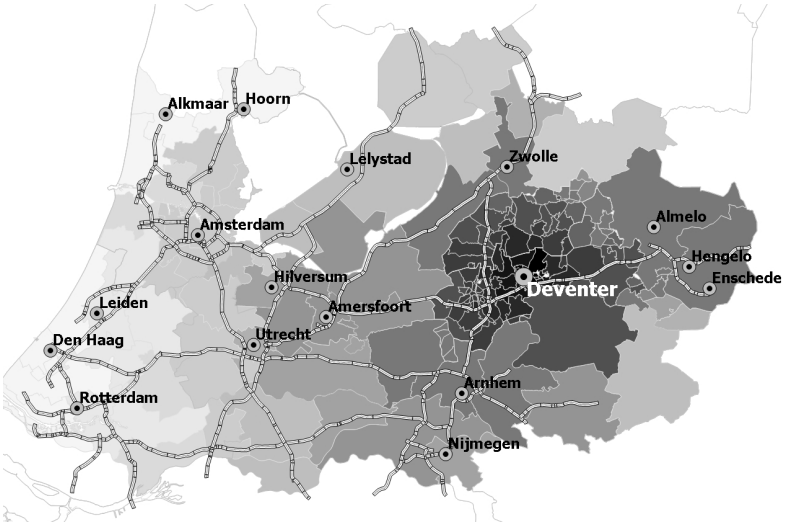

Figure 6.3: Case study Southern Overijssel, dual zone system example

Trip types In case a trip type segmentation scheme has been adopted that distinguishes for different transport modes, the above algorithm can be run for different networks or associated link speeds. For case study Southern Overijssel, all trip type segments $m \in \mathcal{M}$ have been defined by trip purpose only, so there is no need to specify travel costs ψ_{ijm} for different values of m.

6.5 Generation of synthetic trip database

Based on the databases prepared in sections 6.2 until 6.4, this final section generates a synthetic travel diary database for all inhabitants of the study area of southern Overijssel. For this task, the first paragraph determines the attractiveness for all five trip segments (see table 6.3) and for all 3,766 four-digit postal code zones in the Netherlands (zone set $\mathcal{H}$, see table 6.1). The next paragraph determines the attractiveness decay functions for all 90 two-digit postal code zones (zone set $\mathcal{H}_C$, see table 6.1). Both results are then combined to generate the complete spectrum of destination choice eccentricity and associated destination choices from each possible origin zone. Examples of such a spectrum are given for two zones in figure 6.9 on page 137.

Observation of zonal attractiveness For the TDD synthesis algorithm proposed by section 2.5 it is required to know the eccentricity of each destination choice from $\mathcal{O}$ to $\mathcal{D}$, made by a resident of $\mathcal{H}$ (see figure 2.1). Chapters 3 and 4 defined this measure by means of the concepts of attractiveness and attractiveness decay. Using the attractiveness observation model of section 3.4, the modelling specifications and databases prepared earlier in this chapter, the attractiveness is observed for all 3,766 four-digit postal code areas in the

Netherlands. For this observation, a total of 15,758,880 attractiveness units is distributed over all 3,766 zones. Because this total equals the total population of the Netherlands in 1995, the resulting attractiveness estimates represent population equivalents. A high number of 10 iterations is performed for each of the five trip segments (see discussion on convergence speed on page 92). For a first overview, table 6.11 aggregates these 3,766 attractiveness estimates to the 12 provinces of the Netherlands for all five trip segments. In addition, figures 6.4 until 6.6 offer graphical illustrations for three of the five trip segments and for three arbitrarily selected regions along the Dutch coast. These examples are now discussed separately.

		Attractiveness (in population equivalents)				
Province	Population	Work	Education	Shopping	Social	Other
Groningen	560,040	530,411	358,249	380,090	806,775	655,497
Friesland	620,940	474,482	304,989	187,227	1,174,322	433,187
Drenthe	467,120	240,480	644,727	177,817	571,433	509,937
Overijssel	1,070,660	800,359	1,396,774	513,475	1,541,094	1,177,066
Gelderland	1,906,720	1,536,365	1,640,540	3,400,543	2,083,903	2,362,869
Noord Holland	2,511,290	3,030,797	2,453,837	3,301,495	2,123,252	3,175,618
Zuid Holland	3,358,550	3,546,322	3,979,330	2,803,708	2,599,179	2,649,648
Utrecht	1,109,460	1,476,165	2,242,454	1,885,093	894,873	1,384,500
Flevoland	306,410	308,371	226,618	44,773	305,610	256,781
Zeeland	370,640	463,869	120,523	323,774	1,055,675	542,822
Noord Brabant	2,337,710	1,789,017	1,656,592	2,057,992	1,595,786	1,814,011
Limburg	1,139,340	1,562,002	734,401	682,603	1,006,878	796,745
Netherlands	15,758,880	15,758,640	15,759,034	15,758,590	15,758,780	15,758,681
Other countries	-	510,407	6,755	59,131,524	1,050,278	112,401,123

Table 6.11: Case study Southern Overijssel, attractiveness estimates for provinces per trip segment

Table 6.11 shows large differences in attractiveness per province. For comparison purposes, the population totals per provinces are included but note that this data, nor any other socioeconomic data, has been used to obtain the various attractiveness estimates (see section 3.4). Yet, table 6.11 reflects much common knowledge (for Dutch inhabitants) on the socioeconomic, educational and recreational importance of the twelve provinces. For example, the provinces of Noord Holland, Zuid Holland and Utrecht show profoundly higher shares in employment, education and shopping-related attractiveness compared to their share in population. In contrast, attractiveness for the remaining trip segments (mainly social and recreational traffic) is distributed in relatively high shares over the remaining provinces.

Figure 6.4 offers a graphical illustration of the distribution of work attractiveness for four-digit postal code zones within the Randstad (see page 88). Darker shades indicate higher ratios of work attractiveness to the zonal population. Two factors can cause a high ratio. First, in case the population of a zone is low, already a relatively low amount of work

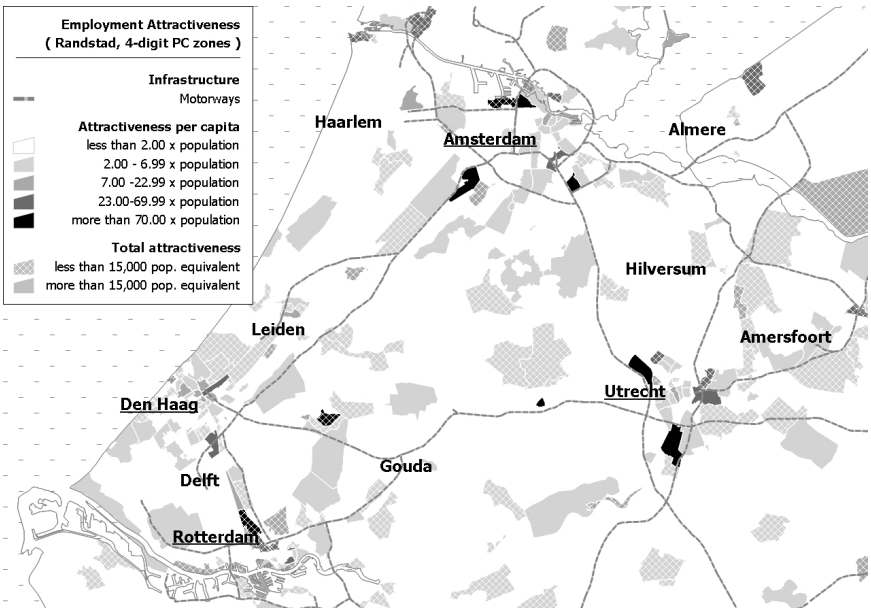

Figure 6.4: Distribution of employment attractiveness in the Randstad

attractiveness leads to a high ratio. In case the population is high, an exceptionally high amount of attractiveness is required to arrive at the same ratio. In order to differentiate, the figure uses a gray pattern-fill for zones with an observed work attractiveness less than 15,000 population equivalents. Apart from these zones, the figure correctly identifies inner-cities, large industrial estates, harbours and the national airport as the main centres of employment.

Figure 6.5 offers another illustration of the observed attractiveness values, now for the distribution of attractiveness for trip segment 5 over the four-digit postal code zones of the province of Zeeland. This trip segment is composed out of a multitude of trip purposes, but mainly out of leisure, holiday and touring trips (see table 6.3). Different from figure 6.4, darker shades in this figure indicate a higher total amount of attractiveness per zone. Gray patterns indicate that a zone has a ratio of less than two attractiveness units per inhabitant. Zones without such a pattern of crosses are therefore zones that are apparently very attractive for this trip segment, an attractiveness that cannot be explained by their population. In the figure, these zones are mostly located at the coastlines, where beaches and nature attract people from far.

Figure 6.6 zooms in at the city of The Hague. The figure shows the observed distribution of shopping attractiveness for the entire city, using a mapping legend similar to the figure for Zeeland. Note that the algorithm correctly

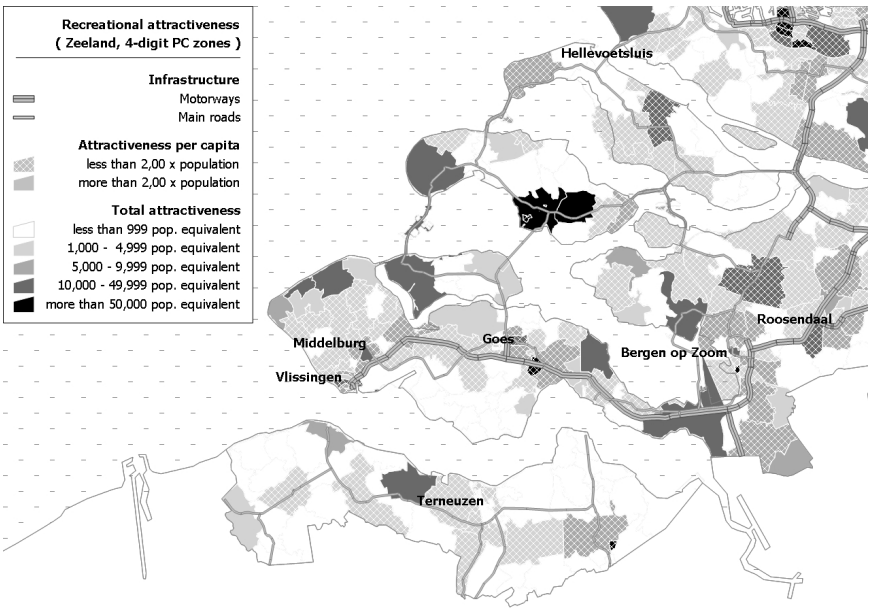

Figure 6.5: Distribution of recreational attractiveness in Zeeland

identifies very high shopping attractiveness for the inner-city, but also for the popular (seashore) locations of Scheveningen and Kijkduin.

Appendix G offers a list of all 90 two-digit attractiveness estimates for all five trip segments. A list of all 3,766 four-digit zones is not included, as this would cover some 35 pages of very small printing. Instead, a sample of four digit attractiveness estimates is included for all zones between 2500 and 2599, located within and around The Hague.

Observation of attractiveness decay Attractiveness decay is observed by the observation model introduced in section 3.5. For case study Southern Overijssel it was determined in section 6.1 to observe independent attractiveness decay functions for all 90 two-digit postal code zones (see table 6.1) and for all five trip segments (see table 6.3). As an example, figure 6.7 shows the attractiveness decay functions observed for the two-digit postal code zone 75. This zone includes all four-digit postal code zones from 7500 up to 7599, including the cities of Enschede and Hengelo which are an important part of the synthesise area $\mathcal{Z}$ for case study Southern Overijssel. For this area, the figure shows the amount of daily trips for an average respondent in population sample ξ per trip segment and per 10,000,000 units of cumulative attractiveness per travel cost circle $\tilde{c} \in \tilde{\mathcal{C}}$. This arbitrary scale factor results in convenient trip frequencies per respondent per trip segment. Note that because decay factors vary to a high degree, a logarithmic scale has been used. Also consider that the highest travel cost circles are of a larger 'bandwidth' than the lower ones but nonethe-

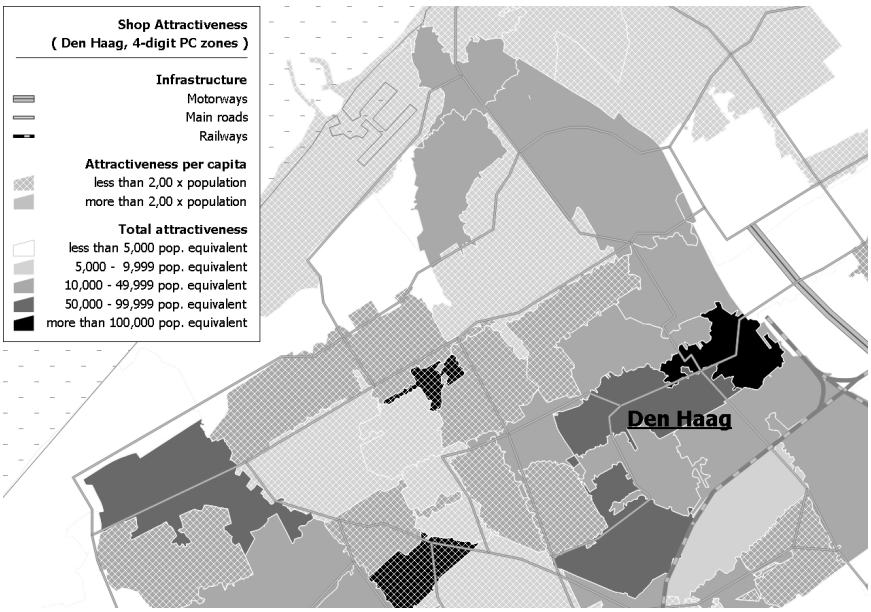

Figure 6.6: Distribution of shop attractiveness in Den Haag

less shown on a linear scale. In addition, observed values of precisely 0 trips for a cost circle have mostly been interpolated for improved readability. This happens if not a single trip is observed for the combination of a trip segment and a cost circle from any of the observed origin zones. In case the cost circle before and after did not have such a zero observation, as a consequence of the logarithmic scale, this results in a large sawtooth, making the figure harder to read. In such cases, the observed line is simply interpolated in the figure, but without drawing an observation point for the zero observation. This practice is only performed in the display of results in figure 6.7. For the actual computation of the case study, zero observations have been left as they are. Although the logarithmic scale can seem to blur out the differences between the trip segments at first sight, closer inspection reveals considerable differences. Whereas mainly shop trips seem to take place over the shortest travel cost circles, they quickly disappear for intermediate distances, only to come back at the highest travel cost circles from Enschede and Hengelo (this is the cost circle that includes the travel cost towards all of the four large cities in the Randstad from Enschede and Hengelo). Finally, note that commuting trips are more resilient to travel cost. The associated attractiveness decay function takes the lowest value for the lowest cost circles (nearly 10 times less trips as for shopping on the same distance), but with relatively high values for most intermediate cost circles. For the cost circles representing the most distant destination zones, it is virtually only social and other (mainly recreational) travel that is of relevance.

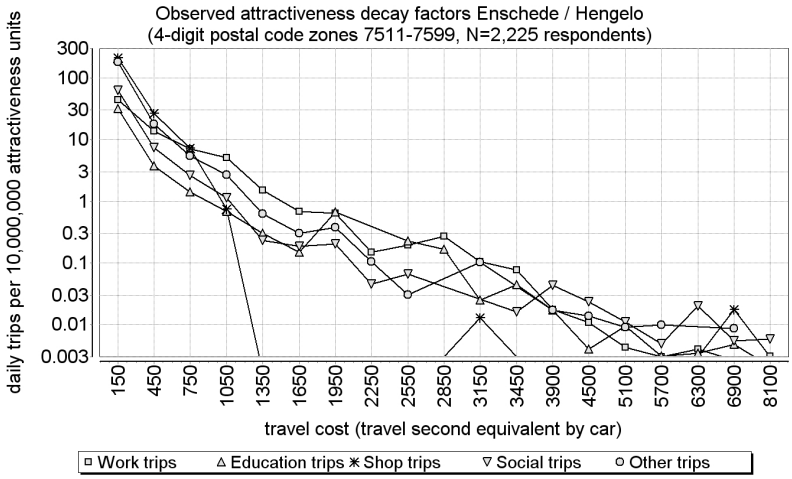

Figure 6.7: Attractiveness decay factors Enschede/Hengelo, all trip segments

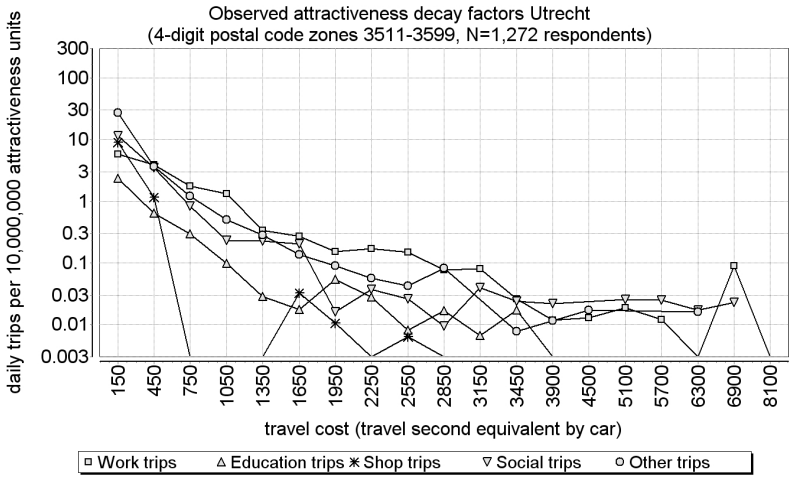

Figure 6.8: Attractiveness decay factors Utrecht, all trip segments

As a contrasting example, figure 6.8 shows the observed attractiveness decay factors for $H = 35$, representing mainly the city of Utrecht. The figure shows a remarkable different decay pattern from the pattern observed in figure 6.7. In general, all decay factors seem much more resilient to travel cost for Utrecht. First, for attractiveness located in the vicinity it can be observed that this attractiveness is less utilised compared to the observation for Enschede and Hengelo. In contrast, attractiveness located at a high travel cost is travelled to much more often. Various factors can be thought of to explain the observed pattern. Although the centrality of Utrecht is cancelled out as a disturbing factor in the observation of attractiveness decay compared to the situation for Enschede, as a second-order effect, this centrality is likely to attract persons with a highly mobile lifestyle to settle in Utrecht and hence result in more mobile destination choice to be observed from Utrecht. Also varying income levels and culture are likely to have their part in explaining this difference.

Generation of destination choice eccentricity ranges The above determined distribution and decay of attractiveness are now combined to obtain the eccentricity of each possible destination choice for any trip context. Figure 6.9 offers an example of the destination choice eccentricity ranges from the inner-cities of Enschede and Utrecht for work trips. Enschede is a major but isolated city in the Netherlands. It functions as a strong regional centre in an area that is mainly rural. It also is the main city located within the synthesise area $\mathcal{Z}$ for case study Southern Overijssel. In contrast, Utrecht, twice the size of Enschede is the most central city of the Netherlands and part of the heavily urbanised Randstad (see page 88). Figure 6.9 shows the spectrum of destination choice eccentricities from both inner-cities at varying scales. Left, the entire spectrum is shown, from the minimum eccentricity value of 0 (the home zone) towards the maximum eccentricity value of 1 (the most distant zone). In the middle, the range from 0.85 to 1.00 is zoomed in. Finally, at the right, the 1% of most eccentric destination choices (0.990-1.000) is shown in detail. In all three graphs, eccentricities for work trips starting from Enschede are shown at the left, and trips from Utrecht at the right. In the graphic, each rectangle represents a zone that can be selected as the destination of a work trip. The higher the rectangle is located in the graph, the more eccentric the associated destination choice is. The larger the height of the rectangle, the more trips are expected to head to this zone. In case an eccentricity range is large enough, the postal code or zone identifier is shown at the bottom of this rectangle (see table 6.1 and 6.2). As an example, the middle graph shows that apparently, the choice from the inner-city of Enschede towards postal code area 7551 is as eccentric as the choice from the inner-city of Utrecht for postal code areas 27 or 52 for a commuting trip. Whereas the first zone represents the inner-city of Hengelo, a mere 8 km from Enschede, the latter zones represent Zoetermeer and Den Bosch respectively, at 50 and 60 km distance from Utrecht. Despite this difference in distance, for both cases some 88% of commuting trips is observed to head to nearer locations, and a mere 11% to locations further away. The frequency in which the inner-city of Hengelo is selected as the destination

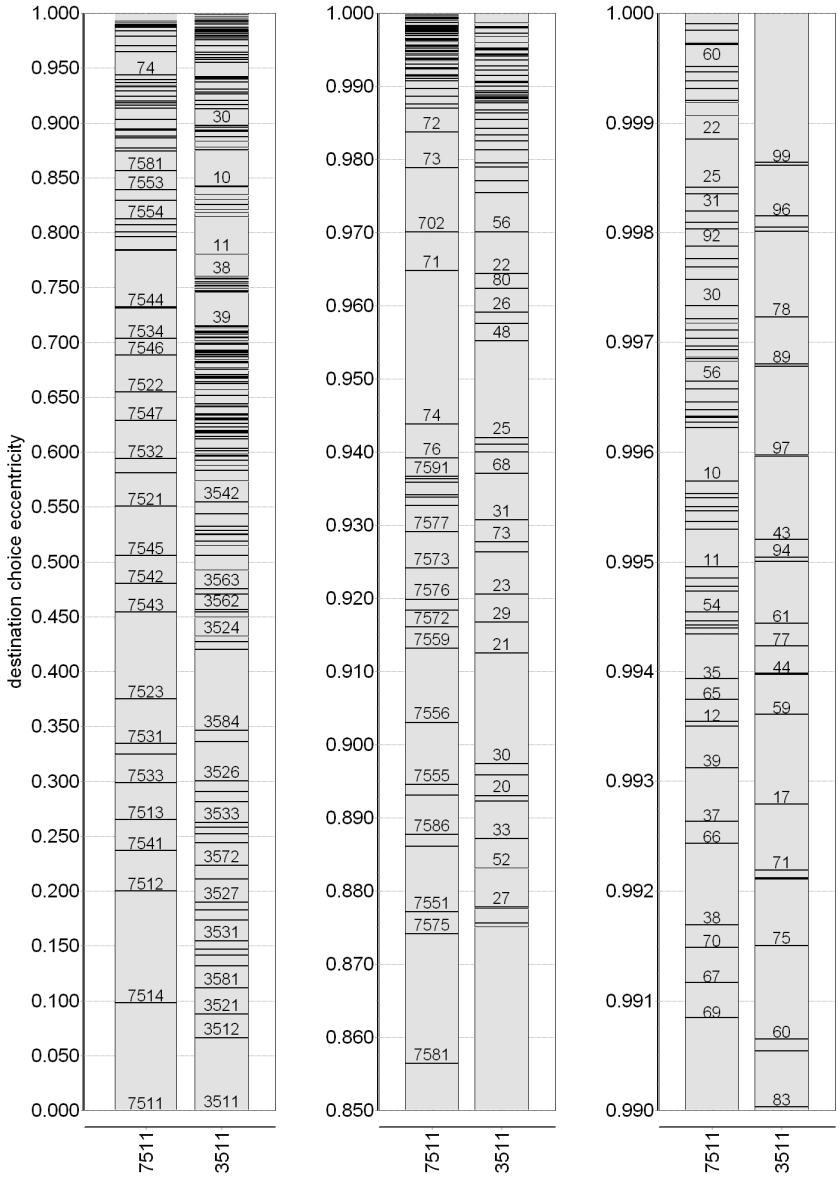

Figure 6.9: Destination choice eccentricities from inner-cities of Enschede (7511) and Utrecht (3511) for work trips

of a work trip starting from Enschede is a bit less than 1%. This frequency is approximately 0.5% for work trips starting from Utrecht for Zoetermeer and 0.3% for Den Bosch. A map of the four-digit postal code areas around Enschede is shown in figure 6.10. The above mentioned two-digit postal can be looked up in figure 6.1.

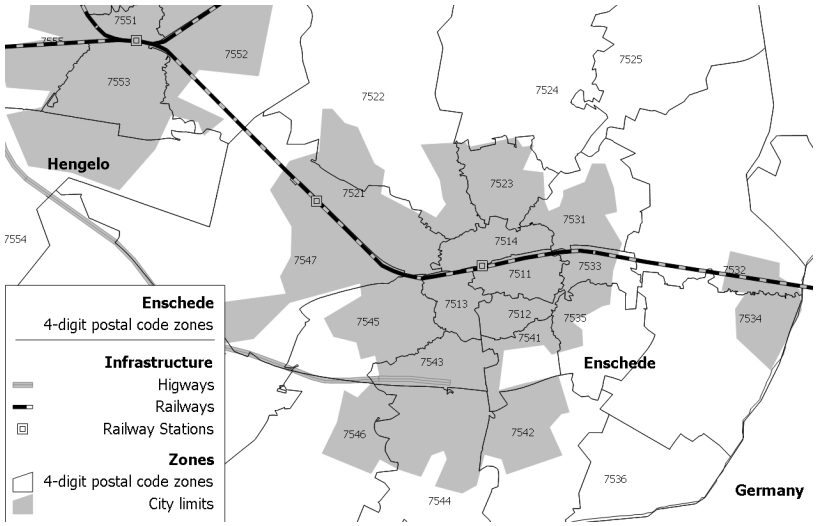

Figure 6.10: Map of four-digit postal code zones in Enschede

Generation of synthetic travel diary database All data is now prepared to run the TDD synthesis algorithm of section 2.5. First, a random respondent in *TB* is selected who has the same population cluster as a synthetic individual in *SP*. The travel behaviour reported by this respondent is now transferred trip by trip to the synthesised individual. For destination choice of non-homebound trips, a zone is looked for that has the same destination choice eccentricity for the synthetic individual as that of the observed destination choice. Using this method, a total of 2,495,726 trips is synthesised for all 658,590 synthetic individuals in the synthesise area $\mathcal{Z}$ of case study Southern Overijssel. The resulting trip database represents the detailed travel behaviour of each individual person living in Southern Overijssel. The question arises how accurate this database should be judged. In an attempt to answer this question, the next chapter is devoted to the validation of the above obtained synthetic trip database.

6.6 Summary

This chapter discussed the collection and preparation of various input data for a large scale application of the TDD synthesis algorithm of section 2.5. The case study Southern Overijssel set up for this purpose intends to synthesise a

travel diary database with trip origin and destination locations at four-digit postal code zones (home to 4,500 individuals on average), for six trip segments (homebound, work, education, shop, social and other trips), for seven transport modes (car driver, car passenger, train, bus/tram/metro, bicycle, walking, other), for an average working-day from midnight to midnight and for the year of 1995. The area for which household travel surveys are synthesised was selected a string of cities including the surrounding countryside of 658,590 individuals in total. Sections 6.2 until 6.4 prepared a set of input databases according to the above specifications. Section 6.5 used these input databases for the actual synthesis of travel diary data. Throughout the chapter, various illustrated examples have been discussed of the input and output of this modelling exercise. All output followed expectation and hence offered an initial test of the proper functioning of the various algorithms. A more extensive validation of the now available synthetic travel diary database is performed in chapter 7.

Chapter 7

Validation for case study Southern Overijssel

This chapter validates the synthetic trip database generated in section 6.5 by two external trip databases. The first external database stems from a large shopping survey in the East of the Netherlands. The second trip database is an excerpt of the official Dutch transport model. Whereas the first database is based on data that is fully exogenous to the data used for the generation of the synthetic trip database, the second database is derived from the same data sources as the synthetic database is. In this manner, the synthetic trip database is validated for two aspects: how the synthetic estimates compare to an independently observed reality and how these estimates compare to estimates obtained by an established method that is based on the same input data. The chapter is outlined as follows. Section 7.1 introduces both external databases and specifies how they have been converted to a trip matrix with identical properties as the synthetic trip database. Section 7.2 compares the zonal trip production and attraction totals for both pairs of trip matrices. Next, section 7.3 compares synthetic estimates with external figures at the level of OD-pairs (see page 35). Section 7.4 draws some conclusions from these validation exercises.

7.1 Introduction of validation databases

This section introduces the two external databases used to validate the synthetic trip database that was generated in chapter 6. These databases are a shopping survey for the East of the Netherlands and the official transport model for the same area. Both databases are introduced in the first two paragraphs of this section. Each of these paragraphs also derives a so-called trip OD-matrix (see page 35) from these external databases that can be used to validate the synthetic trip database. The section concludes with a paragraph on how the synthetic trip database itself has been transformed into two trip OD-matrices

that are of identical properties as the above defined reference matrices.

Shopping survey for the East of the Netherlands The *Koopstroomonderzoek Oost Nederland* (KSO) is a survey on shop visits in the East of the Netherlands that is performed every five years. The KSO database has been kindly made available for this report by I&O Research in Enschede, who conducted the survey. The survey in the year 2000 included 21,361 respondents, out of whom 12,139 reside within the synthesise area $\mathcal{Z}$ for the case study Southern Overijssel. This area was defined on page 118 and a first map was drawn as figure 6.2, showing the location of the 137 constituent four-digit postal code zones. Figure 7.1 draws a second map that shows the location of the 23 municipalities in the area.

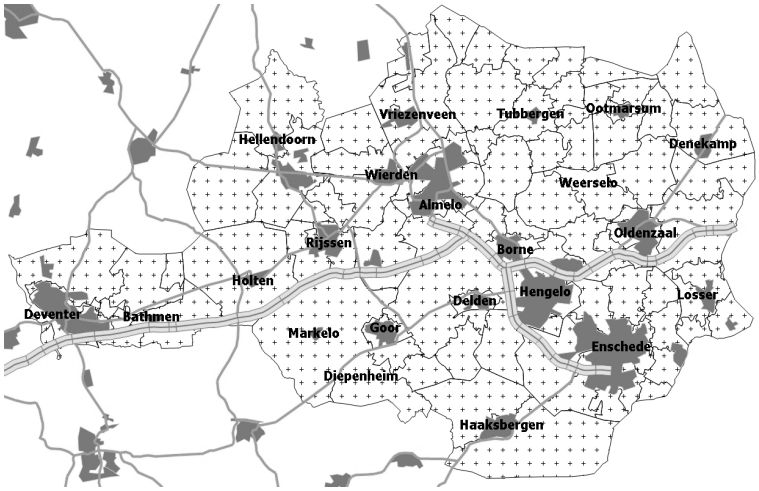

Figure 7.1: KSO municipalities within synthesise area $\mathcal{Z}$

In the survey, a random sample of respondents in each four-digit postal code zone has been asked which shop locations they visit for ten different product categories. In addition, respondents have been asked for the frequency in which they visit these locations per week, month or year. Under the assumption that on average, all trips start from the home zone, the survey is transformed to an OD-matrix (see page 35) for shop trips with their origin and destination within $\mathcal{Z}$. This matrix provides with the monthly number of shop trips from one of the 137 four-digit residential zones towards one of the 4,618 individual shop locations. In order to obtain a statistically significant OD-matrix that is comparable with the synthetic trip database, two further data conversions are performed. First, the total of 4,618 individual shop locations is aggregated by their 23 municipalities. This high aggregation is justified by the high chance factor that would otherwise exist for any of the $137 * 4,618 = 632,666$ cells in the OD-matrix. Based on this aggregation, the total is reduced to $137 * 23 = 3,151$

cells, which is more in line with the total of 12,139 respondents sampled. Table 7.1 presents a list of all municipalities in $\mathcal{Z}$ and their constituent four-digit postal code zones. The table also shows the number of respondents that have been queried for the survey.

Municipality	4-digit zone range	N_{resp}	Municipality	4-digit zone range	N_{resp}
Almelo	7601-7609	796	Holten	7451-7451	200
Ambt Delden	7495-7497	200	Losser	7581-7588	400
Bathmen	7437-7437	200	Markelo	7475-7475	196
Borne	7621-7627	400	Oldenzaal	7571-7577	400
Denekamp	7591-7591/7634-7638	201	Ootmarsum	7631-7631	200
Deventer	7411-7435	1,549	Rijssen	7461-7463	400
Diepenheim	7478-7478	196	Stad Delden	7491-7491	200
Enschede	7511-7547	3,000	Tubbergen	7614-7615/7651-7679	405
Goor	7471-7472	400	Vriezenveen	7611-7611/7671-7676	196
Haaksbergen	7481-7483	400	Weerselo	7561-7561/7595-7597	400
Hellendoorn	7441-7448	600	Wierden	7466-7468/7641-7645	400
Hengelo	7551-7559	800			
			total	7411-7679	12,139

Table 7.1: Shopping survey KSO, number of respondents

Second, the trip frequency in the KSO database is converted from a monthly estimate to that of a workday (Monday till Friday). The factor used for this conversion is calculated using an average of 365 days/12 months = 30.42 days/month, out of which 5/7 are workdays. This amounts to an average of 21.73 workdays per month, equalling 0.0460 month per workday.

New Regional Model for the East of the Netherlands The official Dutch transport model for four-digit postal code zones in the late 1990s is titled the *Nieuw Regionaal Model* (NRM). The model was developed for the Dutch Ministry of Transport and implemented by a consortium of consultancy firms during the mid-90s. The model is split up into several regional parts, out of which the one for the East of the Netherlands is used for this report (*NRM-Oost Nederland 1995*), as it was kindly made available for this report by Rijkswaterstaat Oost-Nederland in Arnhem. The first step of the NRM modelling technique is the derivation of a base trip OD-matrix for the initial mobility situation. Here, this base matrix is used as a validation of the synthetic trip database. For such a validation test, it is of interest that both the synthetic trip database as well as the base trip matrix of the NRM are based on the same travel behaviour database, the OVG of 1995 (see page 124). However, other than the synthetic trip database, the NRM base matrix is also based on traffic counts, detailed socioeconomic zonal data and a detailed road network (used to obtain a detailed travel cost estimate for each OD-pair). As its output for the base situation, the NRM delivers trip matrices for fixed combinations of trip purpose, transport mode and time frames. Out of these trip matrices, this report uses the matrix for car trips (drivers only), for all trip purposes and for a 24-hour workday period. This matrix is selected for being the matrix with the highest number of trips in the NRM and thus probably of the highest statistical

significance for both the NRM as the corresponding synthetic estimates. In order to make the reader more familiar with this trip matrix, table 7.2 shows a summary OD-matrix that has been obtained by aggregating all of the NRM's original 121 zones within $\mathcal{Z}$ into 6 arbitrary macro-zones. The geographic locations of these macro-zones are shown in figure 7.2.

Macro-zone (origin)	4-digit zone range	(destination) 1	2	3	4	5	6
1.Deventer	7411-7439	60,063	2,430	1,223	587	612	1,023
2.NW-Twente	7441-7468	2,232	46,894	4,132	1,465	1,841	9,966
3.SW-Twente	7471-7497	1,199	4,259	81,632	8,626	8,422	4,998
4.Enschede	7511-7547	577	1,492	8,304	120,790	31,307	7,029
5.E-Twente	7551-7597	628	1,959	8,516	31,406	106,573	16,445
6.NE-Twente	7601-7679	945	10,058	4,907	6,956	16,544	139,220

Table 7.2: NRM-Oost Nederland, daily car trips for all trip purposes

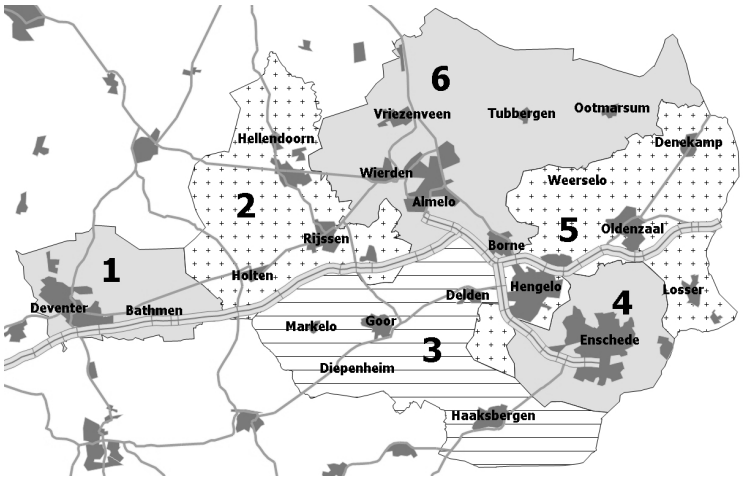

Figure 7.2: NRM macro zone system for synthesise area $\mathcal{Z}$

Because of a slight divergence of the zone system used in the NRM and the four-digit postal code system used for the generation of the synthetic trip database in chapter 6, five out of the 121 zones have been excluded from the validation exercise (NRM zones 439, 473, 477, 491 and 538 respectively). Another difference between the NRM and the synthetic trip database is that the NRM matrix includes foreign passenger trips, here defined as

> trips with both their origin and destination zone within the synthesise area but made by individuals residing outside of the synthesise area

Because the synthetic trip matrix only includes passenger trips made by inhabitants of $\mathcal{Z}$, it need be anticipated that for this reason, the NRM matrix can

show trip volumes that are somewhat higher. The degree in which the NRM includes foreign trips is expected to vary with the size of $\mathcal{Z}$ and the average distance of an OD-pair, origin or destination zone to all boundaries of $\mathcal{Z}$. As a rough estimate, the share of foreign trips is expected to be around 5% of the total NRM trip estimate. This issue is returned to in greater depth on page 150.

Generation of synthetic trip matrices The trip database generated in chapter 6 represents a list of all trips synthesised for the population of $\mathcal{Z}$, and as such, is not directly comparable with the trip OD-matrices delivered by the KSO and the NRM. Of course, such OD-matrices are easily obtained once a synthetic trip database is available. For this purpose, two synthetic trip OD-matrices are made from the trip database of chapter 6. First, the synthetic equivalent of the KSO trip matrix is obtained by selecting all trips with a shop motive from the synthetic trip matrix, after which their numbers are counted for all OD-pairs. Similarly, the synthetic equivalent of the NRM base matrix for car trips is obtained by counting the amount of car trips per OD-pair in the synthetic trip database. In comparing both synthetic OD-matrices with the KSO and the NRM matrices, it need be considered that the synthetic trip database was generated by using input data of other base years as that of the KSO and the NRM. For this purpose, table 7.3 offers an overview of the varying base years used in the creation of the synthetic trip database in chapter 6 and those applying for the KSO and the NRM.

Trip matrix	Travel behaviour database	Synthetic population	Travel cost matrix
1a. Synthetic shop trips	1995	2001	1995
1b. KSO shop trips	(2000)	(2000)	(2000)
2a. Synthetic car trips	1995	2001	1995
2b. NRM car trips	(1995)	(1995)	(1995)

Table 7.3: Case study Southern Overijssel, base year of synthetic database and validation databases

As it can be seen from table 7.3, the (average) base years of the synthetic trip matrices and the reference trip matrices vary to some extent. However, neither the travel behaviour patterns, nor the population, nor the infrastructure changed dramatically from 1995 to 2001 for the synthesise area $\mathcal{Z}$ of southern Overijssel. Thus, where changes occurred locally, these could lead to differences in the synthetic and the reference matrix. However, for the general case, both matrices are considered comparable despite the differences in their base years. Only minor changes in the zone systems of the various base years, as a consequence of a regrouping of municipalities, have been corrected for.

7.2 Explorative validation of synthetic trip matrix

Now that two synthetic trip OD-matrices have been obtained, together with a corresponding set of validation matrices, the next two sections explore their differences. This first section starts with an introduction of the comparison technique that will be used for this purpose. Subsequent paragraphs apply this comparison technique for various levels of aggregation and for both sets of matrices. Section 7.3 then continues to compare these matrices directly at the (non-aggregated) level of individual OD-pairs. With these two sections focussing mainly on the pure observation of differences between all OD-matrices, section 7.4 aims to interpret the differences. By doing so, the section draws four main conclusions on the quality of the synthetic trip matrices and hence of the synthesis method.

Matrix comparison technique Statistically speaking, the trip OD-matrices that have been obtained in section 7.1 can be perceived as a long list of experiments on how geographically separated zones produce and attract trips. When two such matrices need be compared with each other, conventional statistics offer a variety of comparison techniques. For example, it can be tested whether or not the difference between a pair of elements is significant or it can be estimated what is the likelihood that both data-sets represent the same underlying phenomena and should thus be considered as their equals. Unfortunately, the application of these conventional techniques for trip OD-matrices is unreliable because this comparison problem typically does not satisfy the (data) assumptions on which these techniques are based. For example, it is typically necessary to accommodate for high but varying interdependencies for all OD-matrix elements, and for high, unknown and heterogenous variances of individual matrix elements. Neither is acceptable for conventional statistical techniques. Also the application of alternative, highly complex (and disputed) statistical techniques is not considered justified here, considering the mainly explorative nature of this report. Instead, it has been chosen to perform trip OD-matrix validation only by elementary, direct comparison. First, this section compares the grand trip total and the zonal trip totals for all OD-pairs. Section 7.3 then compares all OD-matrix trip volumes for individual OD-pairs. For all of these direct analyses, a goodness-of-fit parameter r^2 has been calculated, using the following equation

$$r^2 = 1 - \frac{\sum (X_i - Y_i)^2}{\sum (X_i - \bar{X})^2} \quad \text{where} \quad \bar{X} = \frac{\sum X_i}{N} \qquad (\infty \leq r^2 \leq 1) \qquad (7.1)$$

where X_i is the validation value for case i ($i = 1 \ldots N$), N is the number of cases, $\bar{X}$ is the average of all validation values and Y_i is the synthetic estimate for case i.

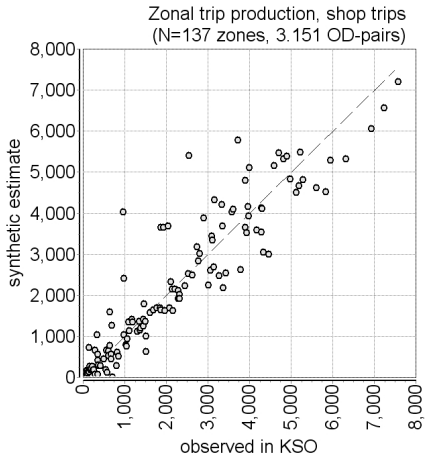

Figure 7.3: Validation of zonal trip production of shop trips, ($r^2 = 0.813$)

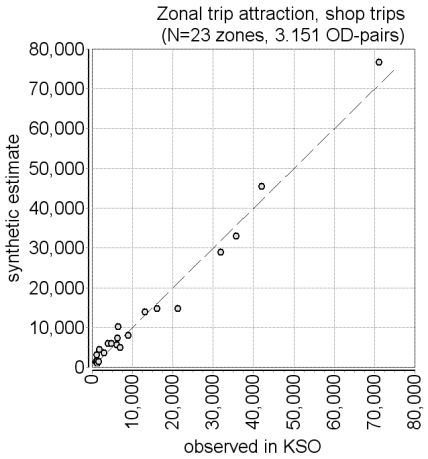

Figure 7.4: Validation of zonal trip attraction of shop trips, ($r^2 = 0.980$)

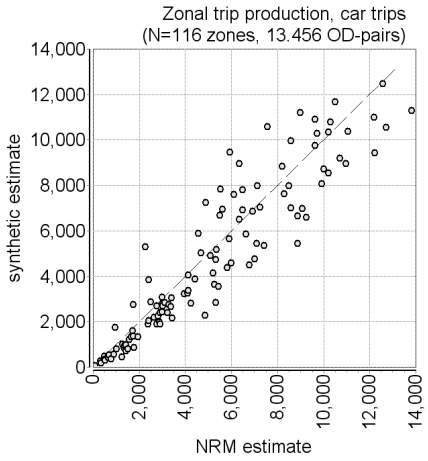

Figure 7.5: Validation of zonal trip production of car trips, ($r^2 = 0.840$)

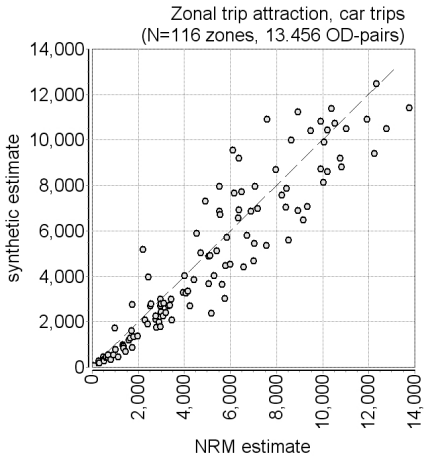

Figure 7.6: Validation of zonal trip attraction of car trips, ($r^2 = 0.842$)

Comparison of grand trip totals As a first explorative comparison, table 7.4 shows the aggregated, grand trip totals for both the synthetic trip matrices as well as the corresponding validation matrices. This comparison shows whether or not the trip estimates made by the synthetic matrices are systematically lower or higher compared to the validation values.

	Daily shop trips	Daily car trips
Synthetic matrix trip total	297,071	560,002
Validation matrix trip total	291,570	595,173
Synthetic matrix trip index	101.0%	94.1%
Validation matrix trip index	100.0%	100.0%

Table 7.4: Case study Southern Overijssel, validation of grand trip totals for synthetic trip matrices

The table shows that neither of the synthetic trip matrices contain a large bias for trip estimates on average. For the daily shop trip matrix, the synthetic overestimation is a negligible 1.0% whereas for daily car trips the synthetic trip matrix makes an average 5.9% underestimation. In the interpretation of this latter difference it is relevant that the NRM includes foreign passenger trips but the synthetic trip matrix does not. The share of these foreign trips was estimated at around 5% of the NRM trip total (see page 145).

Comparison of zonal trip totals As a second explorative, aggregate analysis, the zonal trip totals of both pairs of matrices are compared. A zonal trip total is the total amount of trips that either depart or arrive in a specific zone, and is consequently referred to as zonal trip production and attraction respectively. Statistically, these trip totals represent the row and column totals of the trip OD-matrix. Figures 7.3 until 7.6 explore these trip totals for both sets of trip matrices and for zonal trip production and attraction respectively. The numbers of origin and destination zones in the shop trip matrices differ as a consequence of the OD-matrix aggregation performed in section 7.1: they contain 137 origin zones at the four-digit postal code level against 23 destination zones at the municipal level. For the car trip matrices, the number of both origin and destination zones equals 116 zones at the four-digit postal code level. At a first glance, figures 7.3 until 7.6 all suggest a reasonable fit between the synthetic estimate for zonal trip production and attraction compared to what has been observed in the KSO or estimated by the NRM. The differences are further explored in the subsequent paragraphs.

Comparison of zonal trip totals for OD-pairs with low and high shop trip volumes In order to further investigate the differences between the synthetic estimates and observed values for shop trips (as they are shown in figures 7.3 and 7.4), the zonal trip production and attraction totals in the KSO are split into two components. Let the first component be defined as the total trip volume of all OD-pairs that have a low associated trip volume in the KSOand that depart or arrive in the corresponding zone. The second component is

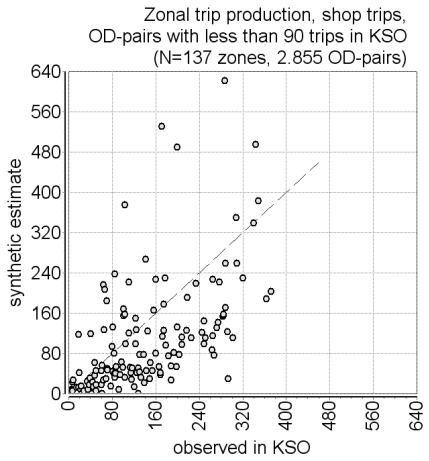

Figure 7.7: Validation of zonal trip production for OD-pairs with low shop trip volumes, ($r^2 = 0.210$)

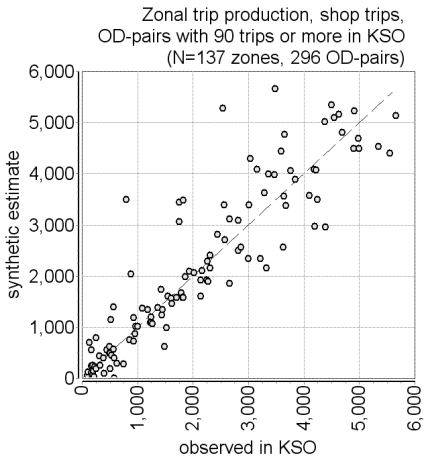

Figure 7.8: Validation of zonal trip production for OD-pairs with high shop trip volumes, ($r^2 = 0.794$)

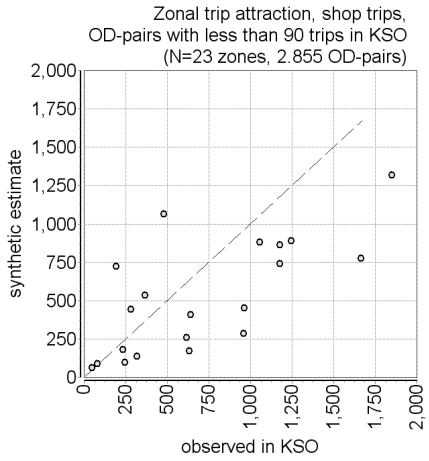

Figure 7.9: Validation of zonal trip attraction for OD-pairs with low shop trip volumes, ($r^2 = -0.038$)

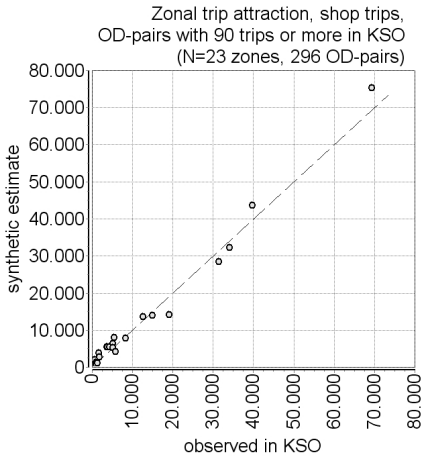

Figure 7.10: Validation of zonal trip attraction for OD-pairs with high shop trip volumes, ($r^2 = 0.983$)

then defined as the total trip volume of all OD-pairs with high trip volumes. For example, let a zonal trip production of 100 trips that is divided over 4 destination zones in a 2:3:45:50 ratio be split into a zonal trip total of 5 trips from OD-pairs with low trip volumes and a complementary zonal trip total of 95 trips from OD-pairs with high trip volumes. The distinction between both components is of interest, because it is likely that the second trip total is of much higher significance in both the validation trip matrix as the synthetic trip matrix. For what constitutes a small and what constitutes a high trip volume, let the average trip volume per OD-pair be taken as the break value. For the KSO survey, this average trip volume equals 291,570 trips / 3,151 OD-pairs = 92.5 trips, rounded downwards to 90 trips. First, figures 7.7 and 7.8 show the zonal trip production of shop trips for OD-pairs with a low and a high number of trips respectively. Figures 7.9 and 7.10 do the same for zonal trip attraction. Both sets of figures, as well as their associated r^2 values, clearly illustrate that the synthetic estimates match better with the KSO observations for OD-pairs with high volumes of shop trips. This could be ascribed to prediction errors in the synthetic estimates for OD-pairs with low trip volumes, but could also be caused by the lower significance that should be attributed to corresponding KSO observations. Section 7.4 returns on this issue.

Comparison of zonal trip totals for OD-pairs with low and high car trip volumes Similarly as for shop trips in figures 7.7 until 7.10, also zonal trip production and attraction of car trips is split into two zonal trip total components that aggregate OD-pairs with either low or high trip volumes only. Again, let the break value for low and high trip volumes be defined as the average car trip volume per OD-pair. Here, this average value equals 595,173 trips / 13,456 OD-pairs = 44.2 trips, rounded upwards to 45 trips. First, figure 7.11 and 7.12 show the zonal trip production of car trips aggregated over OD-pairs with a low and a high number of trips respectively. Figures 7.13 and 7.14 do the same for zonal trip attraction. Again, both pairs of graphs, as well as their associated r^2 values suggest that the synthetic estimates match better with the NRM estimates for OD-pairs with high volumes of car trips. At the same time, this improved match for higher trip volumes is less outspoken as it was for shop trips in the KSO observations. Yet again, considering the clear difference between OD-pairs for low and high trip volumes, the larger differences for low trip volumes can both be ascribed to prediction errors in the synthetic trip matrix as well as to chance elements in the NRM trip matrix, a topic returned to in section 7.4. For now, it is noted that the fit nonetheless is reasonable for both sets of graphs. Also the fact that the fit is better for high trip volumes can be considered of interest on its own, as these OD-pairs are generally the focal point of interest in practical applications. As another example of this latter observation, the comparison of zonal trip totals on all OD-pairs in figures 7.5 and 7.6 resemble more those in 7.12 and 7.14 on the few OD-pairs with high trip volumes as they resemble figures 7.11 and 7.13 on the much larger number of OD-pairs with low trip volumes.

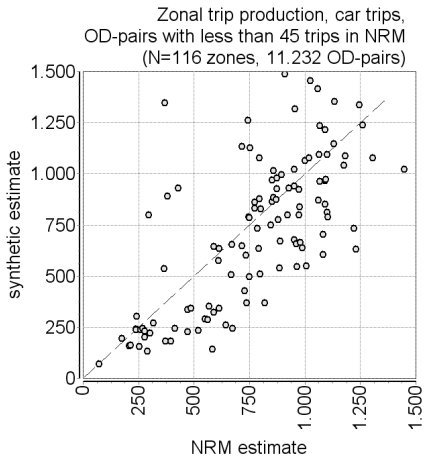

Figure 7.11: Validation of zonal trip production for OD-pairs with low car trip volumes, ($r^2 = 0.470$)

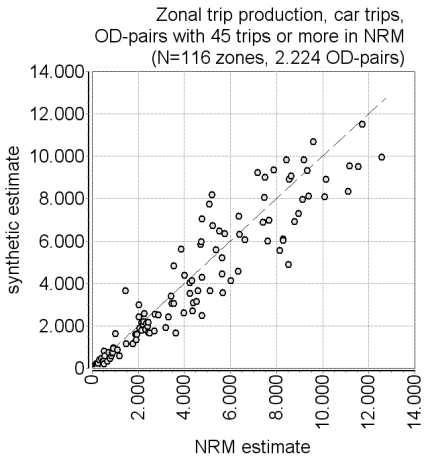

Figure 7.12: Validation of zonal trip production for OD-pairs with high car trip volumes, ($r^2 = 0.836$)

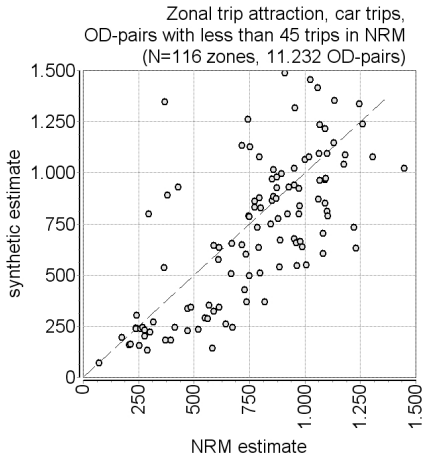

Figure 7.13: Validation of zonal trip attraction for OD-pairs with low car trip volumes, ($r^2 = 0.470$)

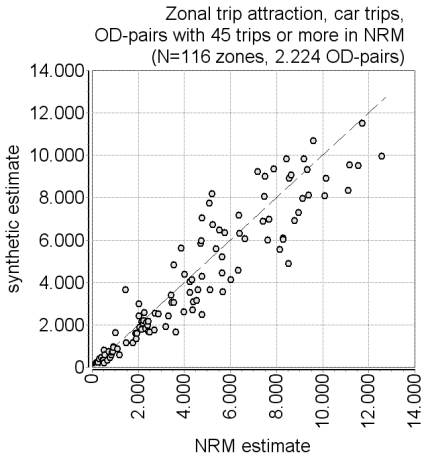

Figure 7.14: Validation of zonal trip attraction for OD-pairs with high car trip volumes, ($r^2 = 0.836$)

Comparison of car trip zonal trip totals for zones near synthesise area boundary In the introduction of the NRM database in section 7.1, it was remarked that the excerpt taken from this database is not fully comparable with the synthetic estimates for the synthesise area $\mathcal{Z}$. This is because the NRM estimate includes foreign trips where the synthetic estimate does not (see page 145). In order to assess the influence of this mismatch, it is considered that the NRM OD-matrix is likely to include more foreign trips with decreasing travel cost to the boundaries of $\mathcal{Z}$ of the associated OD-pairs. Thus, an extra analysis has been performed where the zonal trip totals of zones near the boundaries of $\mathcal{Z}$ are analysed separately from the remaining zones that are further from these boundaries. Thus, 7.15 is identical with 7.5, but with the estimates of zonal trip production for zones near boundaries drawn by crosses. Similarly, figure 7.6 has been regenerated as figure 7.16 for the zonal attraction of car trips but with the boundary zones drawn by crosses. For this analysis, the following procedure has been followed for determining which zones are boundary zones. First, the travel cost of each zone to the nearest zone outside of $\mathcal{Z}$ is determined. Next, the average boundary travel cost for an OD-pair is defined as the average of this boundary travel cost for both its origin and destination zone. Finally, the average boundary travel cost for an origin or destination zone is taken the average of all average minimal boundary travel costs of all OD-pairs originating or destining in this zone. For figures 7.15 and 7.16, a break value of 15 travel minute equivalents has been selected for distinguishing boundary zones from other zones. Whereas the above procedure might seem unnecessarily complicated compared to a straightforward measurement of the travel cost of each zone to the nearest boundary of $\mathcal{Z}$, the method should be preferred in order to give a fair consideration of the varying effects caused by zones enclosed in peninsular-shaped extensions of $\mathcal{Z}$, zones located next to international boundaries of $\mathcal{Z}$ or various localised effects caused by the dual zone system. Figures 7.15 and 7.16 both indicate that a large part of relatively strong mismatches between the NRM trip estimate and the synthetic estimate can be ascribed to the small set of the thus defined boundary zones. This observation is further discussed in section 7.4.

7.3 Detailed validation of synthetic trip matrix

Now that the previous section has made some explorative analyses at aggregated levels, this section compares synthetic estimates with validation values directly at the level of OD-pairs. In order to give a first overview, figures 7.17 and 7.18 have plotted synthetic estimates against corresponding NRM estimates for low trip volumes (< 200 trips) and high trip volumes (< 1200 trips) respectively. However, both figures do not offer much insight in the specific nature of the fit between the synthetic estimates and the NRM estimates. Confronted with the dense cloud of all 13,456 OD-pairs, many overlapping each other, it is hard to filter the associated high r^2 value ($r^2 = 0.856$) from both graphs. Therefore, an alternative comparison method is proposed that uses

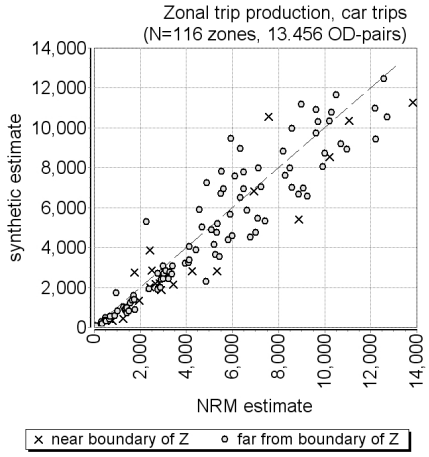

Figure 7.15: Zonal car trip attraction, influence of synthesise area boundary, ($r^2 = 0.840$)

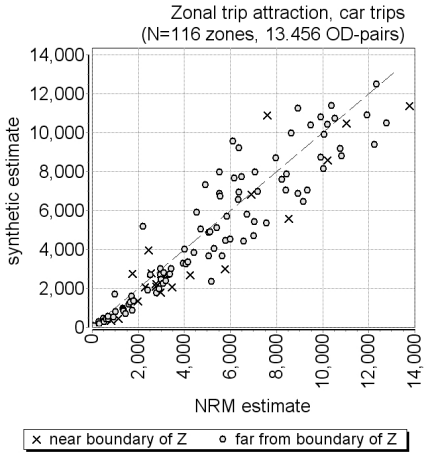

Figure 7.16: Zonal car trip production, influence of synthesise area boundary, ($r^2 = 0.842$)

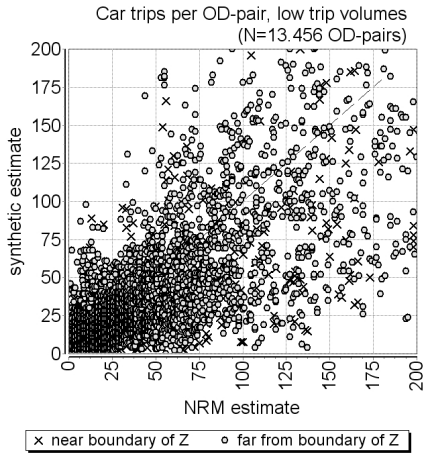

Figure 7.17: Validation of OD-pairs with low car trip volumes, influence of synthesise area boundary, ($r^2 = 0.856$)

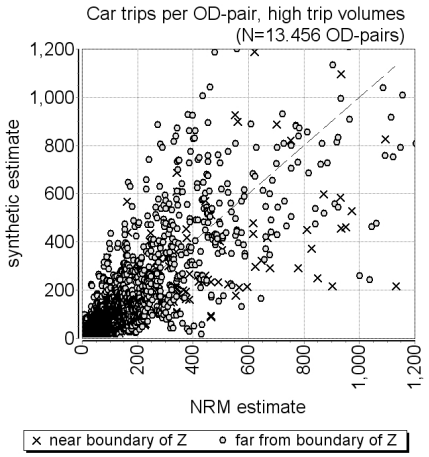

Figure 7.18: Validation of OD-pairs with high car trip volumes, influence of synthesise area boundary, ($r^2 = 0.856$)

indexed values for a range of trip segments.

Introduction of a method for OD-matrix comparison using indexed values Let all synthetic estimates be classified into five categories. In the first category, include all synthetic estimates that are less than a third of the corresponding validation value. In a rough interpretation, these synthetic estimates should be classified as unacceptably low. In the second category, include all synthetic estimates that range from one third up to two thirds of the reference value. Similar to the first category, these estimates should be classified as too low, but to a degree that is easier to accept. Next, let the third category include all synthetic estimates that range from one third under the reference value up to one third above it. For general transport modelling purposes, all synthetic estimates in this category may be considered as equal to the validation value. Finally, let the fourth and the fifth category be defined in analogy to the first and second category, but for synthetic estimates that are higher than the validation value. Unfortunately, it need be considered that the direct adoption of this categorisation for all OD-pairs of entire trip OD-matrices generally does not make much sense. This is because it generally makes an important difference whether a synthetic estimate of 2 trips is made where the reference value is just 1 trip compared to the case where a synthetic estimate is made of 2000 trips, where the reference value is 1000 trips. Both synthetic estimates represent a 100% overestimation from the reference value, but the latter difference is likely to be of high interest whereas the first is of no interest. Therefore, let the above categorisation scheme be adopted for specific trip segments only, rather than for the entire OD-matrix at once. Furthermore, a special case occurs where the reference value equals 0 trips for an OD-pair. In case the corresponding synthetic estimate for such an OD-pair equals 0 trips as well, clearly there is no over- or underestimation and the OD-pair should be counted in the third category of good fits. However, in case the synthetic estimate is 1 trip or more, this estimate represents an infinite overestimation and the pair is included in the fifth category of significant overestimation, whereas no percentile value can be defined that expresses the magnitude of this overestimation. Because of this special property, the category of 0 trips in the reference matrix is best defined a trip segment on its own. The subsequent paragraphs now adopt this comparison technique for the KSO and NRM reference matrices, the results of which are shown in figures 7.19 and 7.20.

Comparison of OD-pair trip volumes for daily shop trips Figure 7.19 shows the percentage of synthetic estimates for shop trips per OD-pair in $\mathcal{Z}$ that is within a certain range of the corresponding KSO observation. Using the categorisation scheme introduced above, the middle category represents synthetic estimates that are near the observed estimate; from one third too low up to one third too high. As it can be seen from the figure, the percentage of synthetic estimates in this category increases in case the number of trips in the corresponding KSO observation increases. First, the percentage of synthetic estimates that represent a near match equals 76% for the special case where the

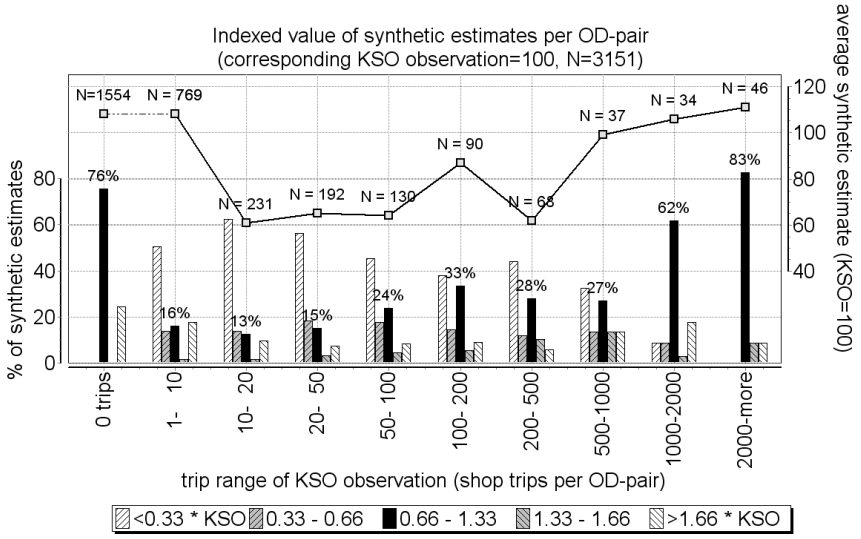

Figure 7.19: Comparison synthetic estimates with KSO observation per OD-pair

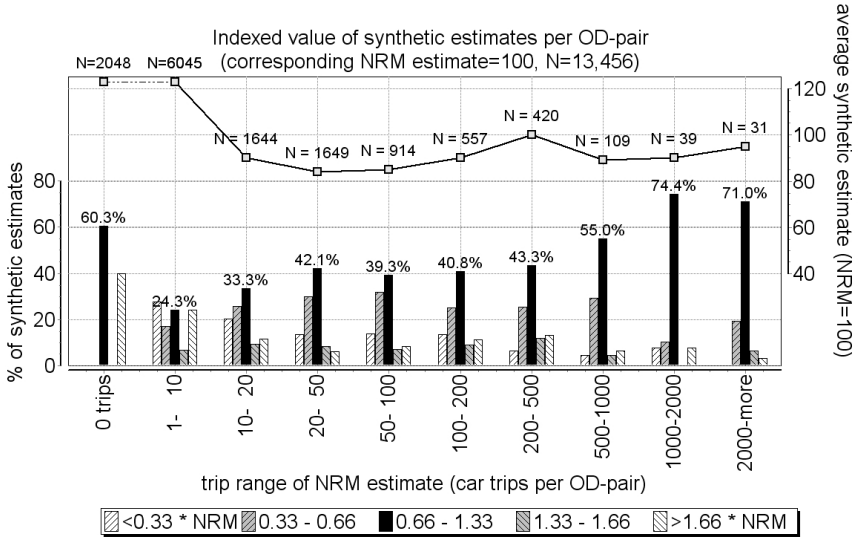

Figure 7.20: Comparison synthetic estimates with NRM observation per OD-pair

observed trip volume is 0 trips, then down to an average 17% for observed trip volumes from 1 to 100 trips, up to 30% for observed trip volumes from 100 to 1000 trips and finally 70% for trip volumes higher than 1000 shop trips per day. In addition, the figure shows how the percentage of medium to high synthetic underestimations decreases from up to 60% for low observed trip volumes to 0% for very high observed trip volumes. As an aggregate measure of the fit between the synthetic estimates and the observed values, the figure also draws a line with the average synthetic estimate if expressed in KSO indexed units. This line shows that for OD-pairs where a low shop trip volume has been observed by the KSO, the synthetic estimate on average reproduces a mere 60% to 70% of the observed trip volume. In contrast, for high volumes, this matching percentage rises to 100% and somewhat above. In the interpretation of this line, a value of 100% indicates that for all OD-pairs in the KSO with a trip volume in the associated trip range, the corresponding synthetic estimates do not produce structural over- or underestimations of this trip volume. Finally, the figure includes the number of OD-pairs for each of the trip range segments defined. Thus, it can be seen that the first trip segment of 0 trips already represents nearly half of the 3151 OD-pairs in total, whereas just 80 OD-pairs had more than 1000 shop trips per day observed.

Comparison of OD-pair trip volumes for daily car trips Figure 7.20 analyses the fit between the synthetic estimates and the corresponding NRM estimates at the OD-pair level. Including more than four times the amount of OD-pairs for the same geographic area as the previous analysis, but with comparable trip volumes per OD-pair, the same trip segmentation scheme is adopted. This makes both graphs well comparable, showing that neither the bias for underestimations for low KSO trip volumes, nor the very high percentage of synthetic estimates in the medium to high underestimation category in figure 7.19, is repeated in figure 7.20. Instead, the average synthetic estimate, expressed in the corresponding NRM indexed values, hovers between 90% and 100% for most trip segments, indicating only a slight structural underestimation compared to the NRM. Also the percentage of OD-pairs in the third category of good fits is significantly higher for most trip segments compared to the values obtained in figure 7.19. For the comparison with the NRM OD-matrix, these values range from 25% for the trip segment of 1 to 10 trips, up to 40% for most other OD-pairs with intermediate trip volumes and finally a high 70% for high trip volumes.

7.4 Conclusions from synthetic trip matrix validation

The number of cases for both the zonal trip total analysis in section 7.2 and the OD-pair analysis in section 7.3 is high to very high. In addition, the synthesise area $\mathcal{Z}$ for the synthetic trip matrix represents a large geographic area with a very heterogeneous urban structure (see page 117). Finally, the reference trip

matrices directly stem from established trip databases. For these reasons, the validation exercise of the synthetic trip matrices with these reference matrices is considered a detailed test of the synthetic matrices themselves as for the method used to generate these matrices. This section presents some conclusions that can be drawn from the analyses in section 7.2 and 7.3 on the quality of these synthetic matrices and hence on the quality of the synthesis method.

Conclusion 1: Synthesis method reduces the need to obtain local transport data The synthesis method proposed by this report can generate synthetic travel diary data for local study areas without requiring much data from this area itself. However, these lower local data demands are offset by higher data demands outside the local study area. It is for this reason that the synthesis algorithm does not permit the reduction of sample sizes for travel diary data in general; it is always dependent on the availability of a sufficiently large empirical travel diary database. However, once such a database is available, the synthesis method allows parts of its data to be used for areas other than where it was observed for, effectively reducing the need to obtain additional local transport data for local transport studies.

Conclusion 2: Synthesis method does not structurally over- or underestimate trips The grand trip total analysis as well as the zonal trip total analysis in section 7.2 both show that the synthetic trip matrices do not significantly over- or underestimate trips. In contrast, figure 7.19 seems to indicate a significant underestimation of trips for shop trips for intermediate trip volume segments. This underestimation however is compensated by over-estimations for the zero trip segment (by mostly only a few trips, but for a high number of OD-pairs) and for the very high trip volume segment (which includes a few OD-pairs only, but which count heavily in absolute trip numbers). Thus, this underestimation of trips represents a local effect that is compensated for by other trip segments. The fourth conclusion returns on this observation.

Conclusion 3: Synthesis method generates close approximations of trip totals for high trip volumes All validation exercises, both at the zonal trip total level in section 7.2 as well as on the OD-pair level in section 7.3 have in common that the synthetic estimates resemble the reference trip matrix better when the associated trip volume increases. This leads to a third conclusion, stating that the synthesis method offers close approximations of trip estimates for high trip volumes. As a break value for what constitutes a high trip volume, figures 7.7 until 7.14 suggest a value of 500 to 1,000 trips. This estimate is based on noting the high percentile differences (and associated low r^2 values) for all odd-numbered figures compared to the much better fits for all even-numbered figures between figures 7.7 and 7.14. A similar estimate can be made from the analysis at the level of OD-pairs in figures 7.19 and 7.20.

Conclusion 4: Synthesis method generates first estimations of trip totals for low trip volumes In analogy to the third conclusion, all syn-

thetic estimates for low trip volumes, be it at the zonal level or at the OD-pair level, show an increasing percentile spread around the corresponding reference values with decreasing trip volumes. However, other than for high trip volumes, both reference databases cannot be considered a reliable source of trip frequency prediction at low aggregation levels. An example is given for the KSO database. Despite being based on a high number of respondents, the trip volume prediction is based on either zero or a handful of respondents for rarely travelled OD-pairs. In case only one inhabitant of zone X with 100 inhabitants performs a weekly shop trip in zone Y (the actual trip volume for this OD-pair thus equals 0.2 daily trips), it matters whether or not this respondent happens to be included in the KSO sample or not. In case he or she is, the KSO observation for the entire OD-pair is extrapolated to an unrealistic high value of, say, $0.2 * 25 = 5$ when a 4% sample was taken from X. However, if the respondent is not included, the extrapolated daily trip volume would have equalled 0 trips. This effect causes the significant underestimation of daily shop trip volumes in figure 7.19 for low (but non-zero) trip volumes. Nonetheless, this effect doesn't imply that the synthetic estimates for these low trip aggregation levels cannot be validated neither as it doesn't render the reference matrices useless at these aggregation levels. Instead, trip totals for low trip volumes should be interpreted with a higher-than-usual spread. Because the synthetic estimates are of a comparable magnitude and whilst noting conclusions 2 and 3, it is concluded that the synthesis method may be used for obtaining a first estimate (a prediction indicating the order of magnitude rather than a precise trip total) of trip volumes for low trip aggregation levels, similar to how both reference databases should be used.

Conclusion 5: Synthesis method generates enhanced trip data For both the KSO and the NRM database, the level of OD-pairs is the lowest meaningful aggregation level. In contrast, the synthetic trip database is generated for individual persons. Note that both synthetic trip matrices validated in this chapter, the one on daily shop trips and the one on daily car trips, actually stem from the same synthetic database generated in chapter 6. By coupling the synthetic trip matrices to the original trip database, a precise specification is obtained of all trips that together constitute the estimate for a single OD-pair. Thus, it can be queried what are the person characteristics of all individuals who travel over each OD-pair. Alternatively, a list can be made of the timing of all these trips or of any other variable included in the synthetic trip database and somehow related to an origin zone, destination zone or combination thereof. Although these enhanced analyses cannot be validated by the travel databases currently available, this feature is likely to be subject to conclusions 3 and 4. This means that for any trip segment of whatever composition, the synthetic prediction is likely to be accurate in case it exceeds 1,000 trips and a good first estimate otherwise. It is believed that this property bears important potential for the improvement of the monitoring and modelling of the benefits offered by transport systems, as it was identified a weakness of current transport modelling practice in chapter 1.

Chapter 8

Conclusions and recommendations

Previous chapters have implemented the travel diary data (TDD) synthesis algorithm of section 2.5 for three case studies. The synthetic travel diary databases thus obtained have been validated either informally by comparison with intuitive outcomes or formally by comparison with external data sources. An important component of the TDD synthesis algorithm is the destination choice eccentricity observation model as it has been specified by chapters 3 and 4. In addition to its relevance for the synthesis of travel diary data, the output of this latter model allows for other practical application as well (see section 5.6). This chapter summarises the main conclusions from the validation of the above case studies, formulates answers on the research questions of section 1.6, proposes possible applications of the synthesis method and finally proposes directions of future research.

8.1 Conclusions from case study validation

Three formal validation exercises have been performed for two of the three case studies implemented in this report. This section starts with a summary of the results obtained from these validation exercises. At the end of the section, several earlier conclusions from the outcomes of this validation exercises are summarised.

Validation of case study Interregional Train Travel In chapter 5, a matrix was estimated with the number of long-distance train trips for a total of 215 OD-pairs. First, this matrix was directly observed from the OVG of 2001 that included 1,565 independent long-distance train trips. For many OD-pairs, this estimate is highly influenced by chance because of the low number of trips starting from the associated origin zone. This results in large standard errors that need be considered when using these estimates. The same data

was then used for the destination choice eccentricity observation model that resulted in a synthetic prediction of the number of train trips for all these OD-pairs. In a comparison of both estimates with a larger empirical database of 6,147 independent long-distance train trips, it turned out that the synthetic estimates were correct for 170 out of 215 OD-pairs (=79%) whereas direct observation was correct for only 137 out of the 215 OD-pairs (=64%). These percentages were obtained for a maximum deviation of 0.84 times the standard deviation under or above the estimated mean of trips in the larger empirical database. This corresponds to a 40% chance that the actual trip volume of an OD-pair is located outside the thus defined range. Further analysis showed that the synthetic estimates can be improved by means of an extension of the destination choice eccentricity observation model, requiring no additional data. In doing so, correct trip volumes could be estimated for at least 190 out of 215 OD-pairs (88%). Stated differently, analysis of the OVG data of 2001 by means of the destination choice eccentricity observation model resulted in the correct prediction of traffic flows for 53 additional OD-pairs out of a total of 215 intercity travel relations. This reduced the number of OD-pairs with incorrectly predicted traffic flows from 78 down to 25, a result obtained without making use of additional data.

Validation of case study Southern Overijssel In chapter 6, a synthetic travel diary database was generated for all of the individual 658,590 inhabitants registered in Southern Overijssel for the year 2001 and for a regular workday. This resulting database includes a total of 2,495,726 individual trips, specified for a total of 14,182,756 OD-pairs, 6 trip purposes and 7 transport modes. For two of the possible aggregations, external data was available so that this output could be validated.

- *Daily shop trips* First, the KSO database offers significant observations of the number of daily shopping trips made on a regular workday for a total of 3,151 OD-pairs within the synthesise area of the case study. Compared to this empirically observed database, the synthetic travel diary database made an over-prediction of trips of a mere 1.0%. The goodness-of-fit between the observed number of trips per zone and the associated synthetic estimates amounted to a r^2 between 0.813 and 0.980 for trip production and attraction respectively. These high r^2 values were maintained at the level of individual OD-pairs with more than 90 daily shop trips, the average trip volume per OD-pair in the KSO database.

- *Daily car trips* Second, the NRM database offers estimates of the number of daily car trips made on a regular workday for a total of 13,456 OD-pairs within the synthesise area of the case study. Here, the synthetic travel diary database offered a systematic underestimation of trips of 5.9% which was explained in part by foreign trips. The goodness-of-fit for zonal trip production and attraction amounted to a high 0.842, which was sustained for individual OD-pairs of more than 45 daily car trips in

the NRM. In addition, the synthetic estimate was shown to improve with increasing travel cost to the border of the synthesise area.

Conclusions from validation exercises Based on the above validation exercises, several conclusions were drawn in section 5.5 and 7.4 that are summarised hereunder.

- *Synthesis method reduces the need to obtain local transport data.* Considering that the output generated by the TDD synthesis algorithm is comparable to that of databases that are based on considerably higher amounts of *locally obtained* transport data, the method was proven to permit reduced local data collection. These lower local data demands however are offset by higher data demands outside the study area. It is for this reason that the synthesis algorithm does not permit the reduction of sample sizes for travel diary data in general; it is always dependent on the availability of a sufficiently large empirical travel diary database. However, once such a database is available, the TDD synthesis algorithm allows parts of its data to be used for areas other than where they were observed for, effectively reducing the need to obtain additional local transport data for local transport studies.

- *Synthesis method does not structurally over- or underestimate trips.* Although all synthetic travel diary databases are based on considerably less local data as the three above-discussed validation databases were, they showed no significant systematic over- or underestimations of trips in all three validation exercises.

- *Synthesis method generates close approximations of trip totals for high trip volumes.* For the latter two validation exercises, the TDD synthesis method produced trip data that did not deviate significantly from validation data for aggregate trip volumes of around 500 to 1,000 trips and higher. In the first validation exercise for case study Interregional Train Travel, it was shown that most errors for high trip volumes could be removed based on an additional analysis that does not require any additional data.

- *Synthesis method generates first estimations of trip totals for low trip volumes.* For low aggregate trip volumes, the precise accuracy of the synthetic data could not be assessed because for these aggregation levels, also the validation data cannot be considered accurate. Nonetheless, noting that trip totals were generated with a comparable order of magnitude, the conclusion was drawn that the synthesis method offers an initial, a-priori estimate for aggregation levels under 500 to 1,000 trips.

- *Synthesis method generates enhanced trip data.* In addition to the generation of the trip OD-matrices that have been validated by external trip matrices, the TDD synthesis algorithm also allows for the detailed analysis of trip flows. Briefly illustrated in section 5.6, it is the generation

of this detailed data on the usage of transport systems that constituted the background of this report, as argued in chapter 1. In addition, the TDD synthesis algorithm resulted in various quantitative output on the geographic functioning of cities and regions, as it was also illustrated in section 5.6. Such data could be used for varied applications such as the evaluation of spatial policy measures, the assessment of unexploited potentials in urban planning or the identification of under- or over-dimensioned infrastructure.

Positioning with reference models The conclusions presented above allow for a positioning of the TDD synthesis method with the two reference models identified in chapter 1 for the synthesis of travel diary data. First, compared to the ALBATROSS model (see page 23), the TDD synthesis method cannot claim a similar robustness for small trip aggregation levels or for drastically changed exogenous conditions (compared to the conditions where an input travel diary database was collected for). This is a consequence of the lacking of an explicit behavioural model on travel behaviour in the TDD synthesis algorithm. At the same time, this property makes the TDD synthesis method require less input data, makes it more transparent and allows faster computation. Compared to the HTS synthesis method by Greaves (see page 24), the method proposed by this report offers a substantial extension by allowing the synthesis of origin and destination location data. This ability, allowed by the inclusion of the destination choice eccentricity observation model to the TDD synthesis framework, increases the practical application of the synthesis method considerably without requiring additional input data.

8.2 Answers on research questions

This section summarises the answers offered by this report on the three research questions posed in section 1.6.

1. **How can trip destination choice be characterised independently of the area in which it has been observed and how can such a measure be implemented without making use of behavioural assumptions other than the information in an existing travel diary database?**

   This report has introduced the measure of destination choice eccentricity for this purpose. The measure defines how eccentric a given trip destination choice is, compared to trip destination choices as they are observed on average from the same location. The measure is based on 1) the perceived attractiveness of the selected and all other possible destination zones and 2) on the perceived discount factor on attractiveness due to the (marginal) travel cost required to travel to a zone. The measure is implemented by the destination choice eccentricity observation model, specified in chapters 3 and 4. This implementation is based on an initial,

uniform estimate of the attractiveness of each possible destination choice. By comparing these estimates with the observations in an existing travel diary database, the estimates are improved in an iterative process.

2. **Using such a location-independent characterisation of trip destination choice, how can travel diary data be synthesised that includes trip origin and destination locations, for each individual, at a detailed geographic level and without bias at any aggregation level?**

   This report implements a micro-simulation approach for the synthesis of travel diary data. The approach first requires the generation of a synthetic population where synthetic travel diary data is to be synthesised for. Next, it requires a population clustering scheme that distinguishes for relatively homogeneous travel behaviour. Finally, it requires an empirical travel diary database of sufficient sample size for all population clusters. By a theoretic model introduced in section 2.4, it is shown that the random selection of a respondent in the empirical travel diary database offers an unbiased estimate of the travel behaviour of a synthetic individual that belongs to the same population cluster. In case the synthetic individual resides in another location as the observed individual, the destination choice eccentricity model can be used to transfer the observed trip destination choices for the different spatial context. The report offers a step-by-step example on the synthesis of travel diary data for each individual in the southern part of the Dutch province of Overijssel.

3. **What is the level of accuracy of this synthetic data at various aggregation levels?**

   The report includes various validation studies on the accuracy of two synthetic databases, the results of which are summarised in the previous section. It was shown that for many cases, the accuracy of local aggregations of synthetic travel diary data is significantly higher than that of the underlying, empirical travel diary data alone. As it could be expected, accuracy was also shown to increase with higher aggregation levels. For the remainder, the accuracy of synthetic travel diary data seems dependent on a complex interaction of factors such as the sample size in the input travel diary database, the geographic spread of its respondent sample, the variability in behaviour, the number of choice options, the zonal level of detail, the degree to which strong location-specific effects exist and various modelling decisions in the observation of destination choice eccentricity. These and other factors can be grouped in factors related to the input data of a case study, the level of detail for this case study, the aggregation level of output data and finally to the study area where output data is required for. Despite this long list of factors that is expected to influence the accuracy of synthetic data to an unknown degree, the report offered two case studies of high complexity for which an acceptable level of accuracy was achieved. Based on these results, the next section

proposes possible applications of synthetic travel diary data generated by the TDD synthesis algorithm.

8.3 Possible applications

This section lists possible practical applications of the TDD synthesis algorithm and the constituent destination choice eccentricity observation model. Part of these possible applications are discussed in greater detail in section 5.6.

- *Assessment of zonal trip attraction factors for conventional transport models* The destination choice eccentricity observation model is easily integrated in the so-called four step transport model, as it is widely used in current transportation practice. The attractiveness estimates obtained for each trip segment and for each destination zone are a good replacement of socioeconomic zonal data as it is used to estimate zonal trip production and attraction.

- *Evaluation of long term spatial policy and transport policy* The quantification of location-specific deviations from average destination choice behaviour can be used to evaluate the net impact of spatial policy and transport policy. Also, in processing input travel diary databases for various years, changes in this behaviour can be monitored and possibly related to such policies.

- *Prioritising of alternative location choices for investment schemes* The above discussed differences between local destination choice behaviour and average behaviour can alternatively be used to prioritise alternative location choices for investment schemes. In case several destination zones are relatively little visited from a base location compared to what would be expected on average, these differences are perhaps easily overcome by focussing attention on the (too negative) perception of the accessibility of these destination zones. Such relatively easy accomplished changes in attitude can be anticipated in the location choice of new city extensions or industrial estates.

- *A priori estimates of travel behaviour* For case study Southern Overijssel, it has been demonstrated how the synthetic travel diary databases may be used to make a-priori estimates of travel behaviour. This is possible for all combinations of trip-maker characteristics, trip origin and destination locations, trip characteristics or trip contexts included in the synthetic travel diary database. It is such insight in the functioning of transport that is considered fundamental to an improved analysis of the societal benefits from transport systems as it was pleaded for in chapter 1.

- *Eliminating privacy concerns for existing travel diary databases* In case empirical travel diary databases are complemented by synthetic travel diary data for all non-observed inhabitants, it becomes impossible to

distinguish actual respondents from synthesised persons. This effectively guards the privacy of the respondents even in case their behaviour is published without any aggregation: it has become impossible to assess whether this behaviour has actually been reported or is synthesised for this person.

- *Adoption of innovative analysis methods on travel behaviour data* Because the TDD synthesis algorithm can produce an estimate of travel behaviour for each individual respondent in a given study area, this output is easily integrated in GIS environments (which require estimates of the modelled phenomenon for each entity defined). This allows for innovative spatial analysis of the data, improved data storage and increased presentation possibilities. Synthetic travel diary databases also open new possibilities for the adoption of data-mining techniques and neural networks in the analysis of travel behaviour data.

8.4 Future research

Concluding this report, several research areas are proposed to further advance knowledge and application of the synthesis method proposed by this report.

- *Accuracy of synthetic travel diary databases* The synthetic travel diary databases generated for the case studies of this report were shown to be comparable (from a transport modeler's point of view) with very detailed external data sources. It is thus indicated that the TDD synthesis algorithm is capable of generating travel diary data that is of sufficient accuracy to be used for conventional transport studies. However, it was concluded in section 8.2 that the conditions under which data of such quality has been generated, remain largely unspecified. In follow-up research, a case study could be set up for which detailed travel diary data is available. For this case study, synthetic travel diary data could then be generated using varying input data sets and adopting various modelling choices. Also, it would be of relevance to assess in what statistical distributions the synthetic data relates to the empirical data at varying aggregation levels.

- *Influence of socioeconomic database on model outcomes* A factor that seems of particularly high relevance for the accuracy of synthetic data is the quality of the initial synthetic population. For the case study Southern Overijssel, a synthetic population has been generated using a very detailed zonal socioeconomic database for an exceptionally low aggregation level of just 35 individuals per zone. Because such highly detailed databases are not available for many countries or too expensive where available, it need be researched what is the influence of aggregation errors in this database.

- *Implementation of marginal travel cost* Equation 4.1 provided an implementation of marginal travel cost that is based on three zone dimensions. Further research could assess how a more advanced implementation of marginal travel cost would lead to more accurate output. However, significant changes in accuracy are only expected for the limited trip segments where the actual marginal travel cost is not well represented by equation 4.1.

- *Quality of input travel diary database* The quality of the input travel diary database directly determines the quality of the output travel diary database. A more elaborate research on this issue could assess how errors in the input database (e.g. underreporting of trips) affect output data, and hence what errors need be anticipated.

- *Influence of other modelling choices* Other important modelling choices whose influence on output data accuracy need be assessed are the break travel cost for the dual zone system, the specification of both circle schemes (for attractiveness and attractiveness decay observation) and the population clustering scheme.

- *Inclusion of route choice model* By inclusion of a route choice model, the synthetic travel diary output can be assigned to infrastructure networks. After such assignment, output data can be validated by observations of actual traffic volumes. Also, such a model extension enables the detailed analysis of traffic flows as it was proposed in chapter 1.

- *Application areas of synthetic data* The synthetic data generated by this report represents a new type of travel behaviour data available for transport research. In its data structure, these synthetic databases look similar to existing, empirical travel diary databases. However, they are distinct in three important aspects that have implications on potential application. First, rather than representing a fixed reality, each data element represents a stochastic quantity. Based on the law of large numbers, this data was shown to give accurate predictions at high aggregation levels, but to be inaccurate at the individual level. Second, the synthetic data is a product of all modelling choices that generated it. As a consequence, synthetic data can never be used to research behaviour that is affected by these modelling choices. Third, large databases are now available for every individual in the study area instead of a limited database for a population sample only. This allows much simplified analysis methods (although it may require different software tools), but could also support innovative modelling techniques such as GIS-based analysis, data-mining techniques, neural network training and the development of complex behavioural models. Whereas the adoption of synthetic data for such research need explicitly consider the previous two aspects, such research seems promising because of the ability of the TDD-synthesis algorithm to concentrate information from an entire sample into a small study area.

Appendices

Appendix A

An algorithm for the generation of synthetic populations

Chapter 2 has introduced a conceptual algorithm for the synthesis of travel diary data. The first step of this algorithm enhances the generation of a synthetic population *SP* for a synthesise area $\mathcal{Z}$. This first step is included in the algorithm because a synthetic population is typically not readily available (see page 33). This appendix proposes an algorithm to generate such synthetic populations for transport studies. The algorithm is capable of the disaggregation of zonal data to individual data for a list of travel behaviour determinants (see page 37). Section A.1 first specifies the precise output generated by the algorithm. The assumed structure of the input zonal database *SE* is specified by section A.2. Finally, section A.3 until A.5 specify the algorithm for the generation of a synthetic population.

A.1 Specification of desired algorithm output

The flowchart in figure 2.1 (see page 44) reveals that a synthetic population *SP* serves one single purpose within the TDD synthesis algorithm. This purpose is that it should allow each member s of a synthetic population to be assigned to a population cluster P_s (see page 38). Based on these assignments, each individual s is matched to a random respondent $r \in TB$, for which $P_r = P_s$. As it has been defined on page 37, a population cluster should be based on a set of shared travel behaviour determinants. This implies that what need be synthesised for *SP* is limited to all travel behaviour determinants used for the population clustering scheme $\mathcal{P}$. Furthermore, because population clusters have been defined based on *similar* characteristics of individuals only (see page 37), no very precise synthesis of each of these determinants is required. Instead, the synthesis of rather rough, categorical estimates for each of these determinants

suffices. This considerably facilitates algorithm design as well as input data requirements.

Household output parameters The next two paragraphs specify which travel behaviour determinants are considered for this report. Also, for each of these determinants, a notation is introduced and it is specified which categories are distinguished (where appropriate). Much of the determinant selection as well as their categorisation has been co-determined by the input database available for this study (see section A.2). The selected travel behaviour determinants are subdivided into two components; determinants defined at the household level and those defined at the individual level. This paragraph first presents all determinants at the household level. For this specification, let a synthetic household within *SP* be indexed by f (for 'family'). Also, let the set of all synthetic households within a zone $z \in \mathcal{Z}$ be denoted as $\mathcal{F}_z$:

- *Household type* HT_f This parameter should provide a rough classification of the type of the household an individual belongs to. Table A.1 provides an overview of the household types that are distinguished.

HT_f	Description
1	one-person-households younger than 35 yr.
2	one-person-households 35 up to 59 yr.
3	one-person-households 60 yr. and older
4	multi-person-households, no children, younger than 35 yr.
5	multi-person-households, no children, 35 up to 59 yr.
6	multi-person-households, no children, 60 yr. and older
7	multi-person-households, 1 or 2 children younger than 10 yr.
8	multi-person-households, 1 or 2 children 10 yr. and older
9	multi-person-households, 3 or more children.

Table A.1: Overview of household types

- *Number of children in household* $N_{c,f}$ For this report, children are considered all in-living children in a household, irrespective of their age.
- *Number of adults in household* $N_{a,f}$ Adults are considered all members of the household who are not in-living children. In order to distinguish them from in-living children, they are sometimes referred to as *main household members*.
- *Household prosperity level* W_f Five household prosperity categories are proposed, presented in table A.2. The symbol W_f stands for 'wealth'.

W_f	Description	W_f	Description
1	High prosperity level	4	Under-average prosperity level
2	Over-average prosperity level	5	Minimal prosperity level
3	Average prosperity level		

Table A.2: Overview of household prosperity categories W_f

- *Number of cars in household* $N_{car,f}$ The number of cars possessed by all members of the household.
- *Number of retired individuals in household* $N_{ret,f}$ The number of individuals in the household who are retired from work and who receive a pension.
- *Number of students in household* $N_{stud,f}$ The number of full-time students in higher education who live in the household and who are not retired.
- *Number of unemployed individuals in household* $N_{unempl,f}$ The number of individuals in the household who are unemployed and who receive an unemployment allowance for this.

Individual output parameters Second, the group of travel behaviour determinants is listed that is defined at the individual level. In this specification, a synthesised individual is denoted by s (see page 33). Let the set of all synthesised individuals within a household f be denoted by $\mathcal{S}_f$.

- *Age* G_s The age of a synthesized individual need only be expressed roughly. Table A.3 presents the six age categories used.

G_s	In-living children	G_s	Main household members
1	Young child ($<$ 10 years)	4	Young ($<$ 35 years)
2	Teenager (10-17 years)	5	Mid-age (35-59 years)
3	Grown-up child ($\geq$ 18 years)	6	Old ($\geq$ 60 years)

Table A.3: Overview of age categories G_s

- *Individual car availability* $S_{car,s}$ Indicates whether or not individual s generally possesses over a car.
- *Individual retired status* $S_{ret,s}$ Indicates whether or not individual s is retired from work as in the definition above.
- *Individual student status* $S_{stud,s}$ Indicates whether or not individual s is a full-time student, as in the definition above.
- *Individual employment status* $S_{empl,s}$ Indicates whether or not individual s is generally employed for 12 hours per week or more.

A.2 Specification of required algorithm input

For the writing of this report, a large socioeconomic database has been kindly made available by Wegener Direct Marketing BV, a commercial data-supplier operating in the Netherlands and in other countries. As it has been noted in section A.1, the choice of output parameters as well as their categorisation has been partly determined by the structure of this specific database. This

section selects all data elements from this database used for the generation of a synthetic population. Subsequently, a short discussion is provided on the level of aggregation of input data.

Input data variables Table A.4 presents a list of all input variables required by the algorithm and their notations. In order to provide with a complete input/output overview, the table also repeats all algorithm output from section A.1. The precision level of all percentage estimates in the Wegener database is 10%. The selected variables are standard data, expected available in many countries.

Input data (for zone $z \in \mathcal{Z}$)	Output data (for household $f \in \mathcal{F}_z$)	(for indiv. $s \in \mathcal{S}_f$)
$HT_{1,z}$ (% household type 1/zone) ... $HT_{9,z}$ (% household type 9/zone)	HT_f (household type)	G_s (age category)
$N_{f,z}$ (# households/zone) $N_{s,z}$ (# individuals/zone)	$N_{a,f}$ (# adults) $N_{c,f}$ (# children)	
W_z (average household prosperity level)	W_f (prosperity level)	
$P_{\text{ret},z}$ (% retired of $N_{s,z}$) $P_{\text{stud},z}$ (% students of $N_{s,z}$) $P_{\text{unempl},z}$ (% unemployed of $N_{s,z}$) $P_{\text{cars},z}$ (% car penetration/household)	$N_{\text{ret},f}$ (# retired) $N_{\text{stud},f}$ (# students) $N_{\text{unempl},f}$ (# unempl) $N_{\text{car},f}$ (# car poss.)	$S_{\text{ret},s}$ (retired) $S_{\text{stud},s}$ (student) $S_{\text{empl},s}$ (employed) $S_{\text{car},s}$ (car avail.)

Table A.4: Synthetic population algorithm, input and output

Input data aggregation level The creation of a synthetic population of individuals out of a zonal database is a disaggregation procedure. This means that out of a single zonal estimate, it is tried to make predictions on all individuals who make up this zone. Synthesised results are thus more accurate for small zones as for large zones, as less prediction is required. Although the algorithm in this section does not place any formal restrictions to zone size in the zonal database, increasing zone size is expected to lead to rapidly decreasing output accuracy. For the Wegener database used in this report, this is not considered a problem, because it is based on very small zones. It is based on the Dutch six-digit postal code level, which houses only 30 to 70 individuals on average. The accuracy of the output generated by means of this database can thus be expected to be good, as it is checked in appendix 205. For databases based on larger zone sizes, additional tests are required before the algorithm can be applied here.

A.3 Conceptual algorithm design

The proposed algorithm should disaggregate data observed at a zonal level to that of individual persons. This section lists two main criteria that are put to the algorithm, introduces a disaggregation hierarchy and then presents a conceptual design.

Algorithm output criteria Two criteria should be met by the disaggregation algorithm. These are summarised as the consistency of the disaggregation and second the generation of most likely, heterogenous distributions. The first criterium states that the number of individuals synthesised for an area and the characteristics assigned to them should aggregate back to the original zonal figures. The second criterium states that correlations between individually assigned characteristics should correspond to correlations observed in the zonal source database.

Algorithm hierarchy As specified in section A.1, output data is required at two aggregation levels. Some of the desired output characteristics are defined at the individual level, but others are defined at the household level. Thus, the disaggregation procedure is split in two separate steps. First, the population of each zone is disaggregated to synthetic households. Second, these households are disaggregated further to synthetic individuals.

Conceptual algorithm design Based on the above criteria and the above algorithm hierarchy, a conceptual algorithm is proposed. In this specification, output databases for households and individuals are referred to as $\mathcal{F}_z$ and $\mathcal{S}_\mathcal{Z}$ respectively (see page 171). Sections A.4 and A.5 provide with a detailed specification.

1. First, a set of synthetic households $\mathcal{F}_z$ is generated, based on available zonal data for all zones $z \in \mathcal{Z}$:

   (a) For each zone z, a number of household records is inserted into $\mathcal{F}_z$ equal to the observed number of households $N_{f,z}$.

   (b) Each of these households is assigned a random household type HT_f, by means of weighted chances. The chance weight of each household type is determined by the observed frequencies of household types $HT_{1,z}$ to $HT_{9,z}$.

   (c) The average distribution in household-size is observed for each household-type in the entire input database or obtained from external statistics. From these distributions, probabilistic chance weights are derived for each household-type and for each household-size. The number of adults $N_{a,f}$ and the number of children $N_{c,f}$ is now randomly selected for each household f, based on HT_f and the above determined chance weights.

   (d) In case the sum of household-sizes $N_{a,f} + N_{c,f}$ assigned to all households in a zone differs from the observed $N_{s,z}$, individuals are added or subtracted from random household where this is feasible (respecting minimum or maximum household sizes).

   (e) A normal distribution $N(\mu, \sigma^2)$ is assumed for the wealth level of individual households in zone z. The mean μ of this distribution is selected the observed zonal wealth level W_z, while the variance

σ^2 is assumed for different values of W_z. For each household f, the wealth level W_f is randomly selected from this distribution.

(f) For all other household characteristics in $\mathcal{F}_z$, the correlation of their (categorical) values with either household-type or household wealth is observed in the input database or obtained from external statistics. This defines a chance weight by which each household-type or wealth level 'causes' a specific categorical value. Based on these chance weights, a value is randomly selected for each of these household characteristics and for all households in $\mathcal{F}_z$.

2. Second, each synthetic household generated above, is disaggregated further to a list of synthetic individuals.

   (a) First, for each household f in a zone, a number of records is inserted in $\mathcal{S}_\mathcal{Z}$ equal to $N_{a,f} + N_{c,f}$.

   (b) The age category is modelled for each individual member of a household in analogy with the procedure in step 1f. If the combination of the resulting individual age categories for a household does not match with its household-type HT_f, it is tried to correct individual age categories already assigned.

   (c) All remaining individual output variables are yes/no variables, expressing whether or not an individual belongs to a certain sub-group or not (students, car-owners, employees an retired individuals). At the household level, it has already been determined how many of its members belong to each of these sub-groups. Considering observed correlations of these sub-groups with age class, employment status or retirement status in the entire input data set, this total number of members is randomly distributed over specific individuals in $\mathcal{S}_\mathcal{Z}$.

A.4 Synthetic households

The next two sections provide a detailed implementation on each of the conceptual steps proposed above. This first section specifies the disaggregation from zonal data to household data. Next, section A.5 specifies the disaggregation from household to individual data. Both sections are organised into paragraphs, each devoted to the generation of a specific output parameter. Many of these steps depend on output generated in previous paragraphs. The sequence in which the disaggregation algorithm is presented is based on these dependencies.

Household-type HT_f The following steps are proposed to generate the household type HT_f for each household $f \in \mathcal{F}_z$ ($f = 1 \dots N_{f,z}$). For the algorithm, let ht index a household-type ($ht = 1 \dots ht_{\max} = 9$). Input data required are the relative frequencies $HT_{ht,z}$ ($ht = 1 \dots HT_{\max}$) by which each household type ht occurs in zone z.

1. The sum $HT_{\text{tot},z}$ is taken of all relative frequency estimates values $HT_{ht,z}$.

$$HT_{\text{tot},z} = \sum_{ht=1}^{ht_{\max}} HT_{ht,z} \tag{A.1}$$

   In case $HT_{\text{tot},z}$ equals 0 for a zone z (for example due to missing data for this area), all estimates of $HT_{ht,z}$ are assigned a uniform value of 10 ($\approx 10\%$), and $HT_{\text{tot},z}$ of 90.

2. The chance that the household type of a household within z equals ht can now be determined as:

$$P(HT_z = ht) = \frac{HT_{ht,z}}{HT_{\text{tot},z}} \tag{A.2}$$

3. A distribution function $F_{HT_z}(ht)$ of stochastic variable HT_z is defined based on these chances:

$$F_{HT_z}(ht) = P(HT_z \leq ht) \tag{A.3}$$

   which is the same as:

$$F_{HT_z}(ht) = \begin{cases} 0 & ht \leq 1 \\ \sum_{k=1}^{ht} P(HT_f = k) & 1 < ht < ht_{\max} \\ 1 & ht \geq ht_{\max}) \end{cases} \tag{A.4}$$

4. For each household f an equidistant z-value z_f is defined as:

$$z_f = \begin{cases} \frac{1}{N_{f,z}}(f-1) \quad \text{where} \quad N_{f,z} \geq 20 \\ \frac{1}{N_{f,z}+1}(f-1) + \frac{0.5}{N_{f,z}} \quad \text{where} \quad N_{f,z} < 20 \end{cases} \tag{A.5}$$

   As it can be seen, a more complex assignment is preferred for zones with a small number of households. The reason is that the simple assignment always assigns a z-value of 0 to the first household. For zones with a small number of households, this could lead to an over-representation of the household type $ht = 1$ as it becomes clear from the next step. However, in zones with a high number of households, the more complex assignment could lead to an over-representation of all other household-types. The break-point of 20 households per zone in equation A.5 reflects the 10% precision level in the input data $HT_{1,z} \ldots HT_{9,z}$ (see page 172). This implies an average 5% under- or overestimation by the input data (one twentieth of the number of households), allowing a similar accuracy level for the disaggregation algorithm.

5. Household f is now assigned household-type ht based on z-value z_f and distribution function $F_{HT_z}(ht)$:

$$HT_f = ht \quad \text{where} \quad \begin{cases} z_f \geq F_{HT_z}(ht-1) \\ z_f < F_{HT_z}(ht) \end{cases} \tag{A.6}$$

Number of individuals per household $N_{s,f}$ Let the number of individuals living in each household f be denoted by $N_{s,f}$. ($N_{s,f} = N_{a,f} + N_{c,f}$). Although this variable is not directly required as an output variable by table A.4, it is required to synthesise $N_{a,f}$ and $N_{c,f}$ later on. The variable can be synthesised by means of the total number of individuals $N_{s,z}$ in zone z and the household type HT_f of f. For this, assume a correlation matrix $\mathbf{D_n}$ for household-size and household-type. Let each element $D_n(ht, n)$ of this matrix define the relative occurrence of a household of household-type ht ($1 \leq ht \leq ht_{\max} = 9$) containing n members ($1 \leq n \leq n_{\max} = 8$). As an example, the correlation matrix $\mathbf{D_n}$ calibrated for this report on page 189:

$$\mathbf{D_n} = \left| \begin{array}{cccccccc} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 40 & 43 & 0 & 0 & 0 & 0 \\ 0 & 10 & 40 & 43 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 51 & 562 & 148 & 39 & 10 \end{array} \right| \tag{A.7}$$

The above matrix thus states that the maximum number of individuals $n_{\max} = 8$ can only be assigned to households of type $ht = 9$ (multi-person households with many children) and only for $10/(51+562+148+39+10) = 1.24\%$ of these households. Based on this matrix, let household-size $N_{s,f}$ for each household f be synthesised by the following algorithm:

1. Define the chance that a household f has a household-size $N_{s,f}$ equal to n as:

$$P(N_{s,f} = n \mid HT_f = ht) = \frac{D_n(ht, n)}{\sum_{k=1}^{n_{\max}} D_n(ht, k))} \tag{A.8}$$

2. Distribution function $F_{N_{s,f}}(n)$ is now defined analogous to equation A.3:

$$F_{N_{s,f}}(n) = P(N_{s,f} \leq n) \tag{A.9}$$

3. For each household f, a random percentile z-value z_f is selected ($0 \leq z_f < 1$). The initial estimate of the number of individuals $N_{s,f}$ living in household f is determined as:

$$N_{s,f} = n \quad \text{where} \quad \begin{cases} z_f \geq F_{N_{s,f}}(n-1) \\ z_f < F_{N_{s,f}}(n) \end{cases} \tag{A.10}$$

4. The sum of all household sizes is now compared with the total population $N_{s,z}$ living in z. Subsequent steps try to fine-tune the initial assignments of step 3 in case both totals are not equal. For this, let ΔN be introduced to denote the number of individuals assigned too much by the initial estimate of step 3. In case $\Delta N = 0$, the assignment procedure ends here.

$$\Delta N = \left(\sum_{f=1}^{N_{f,z}} N_{s,f} \right) - N_{s,z} \tag{A.11}$$

5. Select a random household f ($1 \leq f \leq N_{f,z}$). Now, in case:
   (a) ($\Delta N \geq 1$) $\rightarrow$ on average, too many trips have been assigned in the initial assignment. Therefore, it is tried to decrease the household-size $N_{s,f}$ of household f by one individual. However, this adjustment is only acceptable in case the resulting household-size $N_{s,f} - 1$ has been assigned a positive chance mass in matrix $\mathbf{D_n}$; thus if $\mathbf{D_n}(ht, N_{s,f} - 1) > 0$. Under this condition, both $N_{s,f}$ and ΔN are decreased by 1.
   (b) ($\Delta N \leq -1$) $\rightarrow$ on average, too little trips have been assigned in the initial assignment. Therefore, it is tried to increase the household-size $N_{s,f}$ of household f by one individual. Analogous to the above, this is only acceptable in case $\mathbf{D_n}(ht, N_{s,f} + 1) > 0$. Under this condition, both $N_{s,f}$ and ΔN are increased by 1.
6. The previous step is repeated until $\Delta N = 0$. In practical applications, it is advisable to set a maximum number of iterations. This prevents the algorithm from entering infinite loops if is provided with insolvable input data.

Number of children per household $N_{c,f}$ This paragraph introduces an algorithm to synthesise the number of children $N_{c,f}$ ($0 \leq N_{c,f} < N_{s,f}$) of household f. As input to this algorithm, let correlation matrix $\mathbf{D_c}$ be introduced, in which each element $D_c(ht, c)$ expresses the relative frequency in which a household of type ht contains a number of children c. This matrix is defined 0 for all household-types without children. For household-types with children, the matrix can be estimated by statistical analysis of the entire input database *SE* or obtained from external statistics. An example calibration for $D_c(ht, c)$ is made on page 191, based on a publication of Statistics Netherlands (CBS, 2000):

$$D_c(ht, c) = \left| \begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 103 & 100 & 0 & 0 & 0 & 0 \\ 103 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 549 & 144 & 38 & 10 \end{array} \right| \tag{A.12}$$

1. In case HT_f represents a household-type without children ($HT_f \in \{1, 2, 3, 4, 5, 6\}$), $N_{c,f}$ is defined 0. For other cases, the following steps are performed:

2. First, let the chance of household f having c children ($c \in 1, \cdots, c_{\max}$) be defined as:

$$P(N_{c,f} = c \mid HT_f = ht) = \frac{D_c(ht, c)}{\sum_{k=1}^{c_{\max}} D_c(ht, k)} \tag{A.13}$$

3. In analogy to equation A.3, let distribution function $F_{N_{c,f}}(c)$ be defined as:

$$F_{N_{c,f}}(c) = P(N_{c,f} \leq c) \tag{A.14}$$

4. A random z-value z_f is selected for each household f ($0 \leq z_f < 1$). Based on z_f, an initial estimate of the number of children $\hat{N}_{c,f}$ is defined as follows:

$$\hat{N}_{c,f} = c \quad \text{where} \quad \begin{cases} z_f \geq F_{N_{c,f}}(c-1) \\ z_f < F_{N_{c,f}}(c) \end{cases} \tag{A.15}$$

5. Now, the number of children $\hat{N}_{c,f}$ assigned to each household f is compared with the total household-size $N_{s,f}$ as it has been synthesised above. At least one non-child should live in each household. Thus the constraint $N_{c,f} \leq N_{s,f} - 1$ is put. Based on this constraint, the initial estimates $\hat{N}_{c,f}$ are now corrected in order to obtain $N_{c,f}$:

$$N_{c,f} = \begin{cases} \hat{N}_{c,f} & \text{if} \quad \hat{N}_{c,f} \leq (N_{s,f} - 1) \\ N_{s,f} - 1 & \text{if} \quad \hat{N}_{c,f} > (N_{s,f} - 1) \end{cases} \tag{A.16}$$

   The children which are -synthetically spoken- removed from their families are stored in ΔC:

$$\Delta C = \sum_{f=1}^{N_{f,z}} (\hat{N}_{c,f} - N_{c,f}) \tag{A.17}$$

   In case $\Delta C = 0$, the assignment algorithm ends here.

6. As a final step, the children that were 'removed' from households should be assigned to other households in order to prevent a structural over-representation of households with a small number of children. In order to do so, a random household f is selected ($1 \leq f \leq N_{f,z}$).

   (a) $\mathbf{D_c}(HT_f, N_{c,f}+1) = 0 \rightarrow$ It is not feasible to add a child to household f, because no chance mass has been assigned in association matrix $\mathbf{D_c}$ for the resulting number of children $N_{c,f} + 1$ for the household type of f. No child is added to this household.

   (b) $\mathbf{D_c}(HT_f, N_{c,f}+1) > 0 \rightarrow$ It is feasible to add a child to household f. As a consequence, $N_{c,f}$ is increased by one, where ΔN is decreased by one.

7. The previous step is repeated until $\Delta C = 0$. Again it is advisable to put a maximum number of iterations for practical applications (see page 177).

Number of adults per household $N_{a,f}$ The number of adults in each household f is based on the estimates for $N_{s,f}$ and $N_{c,f}$ which have been obtained in the previous paragraphs:

$$N_{a,f} = N_{s,f} - N_{c,f} \tag{A.18}$$

Level of prosperity per household W_f The average household prosperity level in zone z is provided by input parameter W_z. In the conceptual design in section A.3, it was proposed to assume the prosperity level of individual households by a normal distribution $N(\mu, \sigma^2)$ in which $\mu = W_z$. The standard deviation σ^2 is assumed fixed at 0.6 prosperity categories. However, for the 'mixed' category $W_z = 6$, μ is assumed the average prosperity category $W_z = 3$, where $\sigma^2 = 1.0$ prosperity categories. Now, for each household f, a random number z_f is selected from normally distributed $N(\mu, \sigma^2)$. Now, a prosperity level W_f is assigned based on z_f:

$$W_f = \begin{cases} 1 & \text{if} \quad z_f < 1 \quad \text{(highest prosperity level)} \\ z_f & \text{if} \quad 1 \leq z_f \leq 5 \\ 5 & \text{if} \quad z_f > 5 \quad \text{(lowest prosperity level)} \end{cases} \tag{A.19}$$

Number of retired individuals per household $N_{\mathbf{ret},f}$ The input parameter $P_{\text{ret},z}$ provides with the percentage of retired individuals in zone z. Let the total number of retired individuals in z be denoted $N_{\text{ret},z}$, estimated as $N_{\text{ret},z} = P_{\text{ret},z} N_{s,z}$. This paragraph proposes an algorithm to distribute this zonal total over the individual households in z. Out of the parameters which have been synthesised for households at this moment, household type is considered the most strongly correlated with the occurrence of retired individuals within a household. Therefore, it is proposed to control the distribution of retired individuals by the household type HT_f of each household f. For this purpose, let correlation vector $\mathbf{D}_{\text{ret}}$ be introduced with each element $D_{\text{ret}}(ht)$ providing the relative chance in which a retired individual lives in a household of type ht. This vector can be observed from the entire input database *SE* or be obtained from external statistics. For this study, an example vector is calibrated on page 191:

$$\mathbf{D}_{\text{ret}} = \left| \begin{array}{cccccccccc} 0 & 14 & 446 & 0 & 20 & 632 & 20 & 62 & 18 \end{array} \right| \tag{A.20}$$

Based on this vector, let $N_{\text{ret},f}$ be synthesised for each household f by the following algorithm:

1. Assume an initial estimate of $\hat{N}_{\text{ret},f} = 0$ for all households in z ($1 \leq f \leq N_{f,z}$)

2. Suppose that one retired individual resides in z. The chance that the household F_{ret} in which this individual resides, equals household f can then be expressed as:

$$P(F_{\text{ret}} = f \mid HT_f = ht) = N_{a,f} \frac{D_{\text{ret}}(ht)}{\sum_{k=1}^{ht_{\max}} D_{\text{ret}}(k)} \tag{A.21}$$

   This chance is co-determined by the number of adults $N_{a,f}$ in household f, because a higher number of adults in f increases the chance that it includes a retired individual. Now, in analogy to equation A.3, the chance by which F_{ret} equals f can be defined by distribution function $F_{F_{\text{ret}}}(f)$:

$$F_{F_{\text{ret}}}(f) = P(F_{\text{ret}} \leq f) \tag{A.22}$$

3. Instead of denoting the chance that *one* retired individual in z makes part of household f, equation A.22 could also be used for denoting the chance that one *additional* retired individual resides in f. In such an assignment of retired individuals to a household, it should however additionally be checked that the total amount of retired individuals assigned never exceeds the total number of adults in this household. Under this condition, let a random z-value z_{ret} be selected ($0 \leq z_{\text{ret}} < 1$) and let an additional retired individual be synthesised for household f accordingly:

$$N_{\text{ret},f} = \hat{N}_{\text{ret},f} + 1 \quad \text{where} \quad \begin{cases} z_{\text{ret}} \geq F_{F_{\text{ret}}}(f-1) \\ z_{\text{ret}} < F_{F_{\text{ret}}}(f) \\ \hat{N}_{\text{ret},f} \leq N_{a,f} - 1 \end{cases} \tag{A.23}$$

4. The previous step is repeated until a total of $N_{\text{ret},z} = P_{\text{ret},z} N_{s,z}$ retired individuals have been successfully assigned to all households in z. In analogy with the comments on page 177, it is advisable to put a maximum number of iterations for practical applications.

Number of students per household $N_{\mathbf{stud},f}$ In order to assign the number of students $N_{\text{stud},f}$ per household f, based on a zonal estimate of $N_{\text{stud},z} = P_{\text{stud},z} N_{s,z}$, the same algorithm can be used as for the assignment of retired individuals (see page 179). In this assignment procedure, only equation A.23 need be modified to represent that a retired individual cannot be a full-time student at the same time. Thus, let the third condition in this equation be replaced by:

$$\hat{N}_{\text{stud},f} \leq N_{a,f} - N_{\text{ret},f} - 1 \tag{A.24}$$

Also, instead of correlation vector $\mathbf{D}_{\text{ret}}$, let vector $\mathbf{D}_{\text{stud}}$ be used. Let this vector be introduced to denote the correlation of household-type and the number of students in such households. For this purpose, let each element $D_{\text{stud}}(ht)$ provide with the relative chance in which a person belonging to a household of type ht is a full-time student. The vector used for this study is calibrated on page 192:

$$\mathbf{D}_{\text{stud}} = \left| \begin{array}{cccccccccc} 138 & 6 & 3 & 50 & 3 & 2 & 6 & 137 & 116 \end{array} \right| \tag{A.25}$$

Number of unemployed individuals per household $N_{\mathbf{unempl},f}$ Let the number of unemployed persons in zone z be estimated as $N_{\text{unempl},z} = P_{\text{unempl},z} N_{s,z}$. As with students, let this total number of persons be subdivided over individual households in analogy to the algorithm used for the distribution of retired individuals (see page 179). However, for this distribution algorithm, let the third condition in equation A.23 be replaced by:

$$\hat{N}_{\text{unempl},f} \leq N_{a,f} - N_{\text{ret},f} - 1 \tag{A.26}$$

Also, let a correlation vector on the number of unemployed persons not be based on household type, but on the level of household prosperity. For this purpose, let a level of household prosperity be denoted by w ($1 \leq w \leq w_{\max}$). Now, let correlation vector $\mathbf{D}_{\text{unempl}}$ be introduced of which each element $D_{\text{unempl}}(w)$ provides the relative chance in which a household member is unemployed in case he or she belongs to a household of prosperity level w. The vector used for this study is calibrated on page 194:

$$\mathbf{D}_{\text{unempl}} = \left| \begin{array}{ccccc} 0 & 0 & 10 & 19 & 42 \end{array} \right| \tag{A.27}$$

Number of cars per household $N_{\mathbf{car},f}$ Finally, let the total number of cars in zone z be defined as $N_{\text{car},z} = P_{\text{cars},z} N_{f,z}$. Let vector $\mathbf{D}_{\text{car}}$ be introduced to denote the correlation between car possession and household prosperity. Let the elements $D_{\text{car}}(w)$ of this vector denote the relative chance in which a household possesses over a car in case it has a household of prosperity level w. The vector used in this report is derived on page 194:

$$\mathbf{D}_{\text{car}} = \left| \begin{array}{ccccc} 127 & 94 & 73 & 42 & 20 \end{array} \right| \tag{A.28}$$

Based on this vector, the total number of cars $N_{\text{car},z}$ can be distributed over the households in z in order to obtain estimates of $N_{\text{car},f}$, analogous to the algorithm for the distribution of retired individuals (see page 179). In this algorithm, the third condition in equation A.23 may be discarded, because no maximum number of cars can be put to each household.

A.5 Synthetic individuals

The above section has disaggregated zonal data to household data, thus implementing the first half of the disaggregation algorithm discussed in section A.3. This section now implements the second half; from synthetic households to synthetic individuals. For this implementation, let synthetic individuals be indexed by s (see page 33) where $1 \leq s \leq N_{s,f}$ and where f represents the household where s belongs to. The data elements which need be synthesised for each such individual has been specified in table A.4. As in section A.4, the generation of each of these output variables is discussed in separate paragraphs.

Individual age category G_s In table A.3, a categorisation scheme has been proposed which distinguishes for two types of age categories: in-living children and adult members of a household. In-living children are subdivided in three age categories with 10 and 18 years of age as breakpoints. Similarly, adults are subdivided in three age categories around 35 and 60 years of age. Because these age categories correspond with the definition of household types in table A.1, individual age categories G_s can mostly be assigned by simple deterministic rules for all household members s. The list below provides assignment rules for all household types ht.

1. In case $HT_f = 1$ for a household f (young one-person households), then its single member s is assigned $G_s = 4$ (young adult).

2. Similarly, in case $HT_f = 2$ or 3 for a household f (mid-age and old one-person households respectively), then its single member s is assigned $G_s = 5$ or 6 respectively (mid-age vs. old adult).

3. The same rules can be applied for pairs without children, if it is assumed that on average both of its members belong to the same age category on average. Thus, in case $HT_f = 4$, 5 or 6 for a household f (young, mid-age or old multi-person households without children respectively), then its members are both assigned $G_s = 4$, 5 or 6 (young, mid-age, old adults respectively).

4. In case $HT_f = 7$ for a household f (households with younger children), then all children are assigned $G_s = 1$ (young child). These children are defined as the first $N_{c,f}$ individuals in $s = \{1, \ldots, N_{s,f}\}$. The remaining (main) members of this household need be assigned an age category stochastically. For this stochastic assignment, let g denote age category ($g = 1, \ldots, g_{\max}$) Furthermore, let correlation vector $\mathbf{D_{A,ht_7}}$ be introduced where each element $D_{A,ht_7}(g)$ provides the relative chance of an adult member of a household with young children (household type 7), to belong to age category g. By definition, this chance equals 0 for the age categories 1 to 3 (in-living children). For the remaining age categories, an example calibration of $D_{A,ht_7}(g)$ is made on page 195:

$$\mathbf{D_{A,ht_7}} = \begin{vmatrix} 0 & 0 & 0 & 23 & 10 & 0 \end{vmatrix} \tag{A.29}$$

Now, for all adult members in a household of household type 7, the chance that their age category G_{ht_7} equals g can be expressed as:

$$P(G_{ht_7} = g) = \frac{D_{A,ht_7}(g)}{\sum_{k=1}^{g_{\max}} D_{A,ht_7}(k)} \tag{A.30}$$

In analogy to equation A.3, the chance by which stochastic variable G_{ht_7} equals g can also be described by a distribution function $F_{G_{ht_7}}(g)$:

$$F_{G_{ht_7}}(g) = P(G_{ht_7} \leq g) \tag{A.31}$$

Now, let a random z-value z_s be selected ($0 \leq z_s < 1$) for each adult member s who lives in a household of type 7. This adult is assigned an age category G_s according to equation A.32.

$$G_s = g \quad \text{where} \quad \begin{cases} z_s \geq F_{G_{ht_7}}(g-1) \\ z_s < F_{G_{ht_7}}(g) \end{cases} \tag{A.32}$$

5. Finally, for households of type 8 or 9 (households with 1 or 2 in-living children of 10 years and older and households with three or more in-living children respectively), both the age categories of the children and their parents should be synthesised. For this stochastic assignment, an algorithm can be applied which is analogous to the one introduced above for the main household members of household type 7. For such an algorithm, four new association vectors are required. First, let vector $\mathbf{D_{C,ht_8}}$ be introduced where each element $D_{C,ht_8}(g)$ expresses the relative chance of a *child* member to belong to age category g if he or she belongs to a household of household type 8. An example estimate of this vector is calibrated on page 196:

$$\mathbf{D_{C,ht_8}} = \begin{vmatrix} 0 & 20 & 10 & 0 & 0 & 0 \end{vmatrix} \tag{A.33}$$

A similar vector, but then for the children who live in a household of type 9 is calibrated on page 197.

$$\mathbf{D_{C,ht_9}} = \begin{vmatrix} 76 & 68 & 10 & 0 & 0 & 0 \end{vmatrix} \tag{A.34}$$

Second, let vector $\mathbf{D_{A,ht_8}}$ be introduced where each element $D_{A,ht_8}(g)$ expresses the relative chance of a *main member* in a household of household type 8, to belong to age category g. This vector is calibrated on page 196:

$$\mathbf{D_{A,ht_8}} = \begin{vmatrix} 0 & 0 & 0 & 28 & 253 & 10 \end{vmatrix} \tag{A.35}$$

A similar vector, but for the main members of a household of type 9 is calibrated on page 198:

$$\mathbf{D_{A,ht_9}} = \begin{vmatrix} 0 & 0 & 0 & 35 & 55 & 0 \end{vmatrix} \tag{A.36}$$

Individual retired-status $S_{\mathbf{ret},s}$ Consider all cases where at least one individual within a household is a retired individual ($N_{\text{ret},f} \geq 1$). Out of the variables already synthesised, the state of being retired is considered to be the most strongly correlated with age. Therefore, let vector $\mathbf{D}^*_{\text{ret}}$ be introduced, where each element $D^*_{\text{ret}}(g)$ expresses the relative chance of an individual s to be retired, provided his or her age category G_s. An example estimate of this vector is calibrated on page 198:

$$\mathbf{D}^*_{\text{ret}} = \begin{vmatrix} 0 & 0 & 0 & 0 & 10 & 618 \end{vmatrix} \tag{A.37}$$

Based on this vector, let the number of retired individuals within a household $N_{\text{ret},f}$ be assigned to specific household members by means of the following algorithm. In this discussion, let $S_{\text{ret},s}$ denote the retired status of each household member s ($S_{\text{ret},s} = 1 \rightarrow s$ is retired, $S_{\text{ret},s} = 0 \rightarrow s$ is not retired)

1. First, let the initial estimate of the retired status $\hat{S}_{\text{ret},s}$ of each individual s be assumed 0.
2. Suppose that $N_{\text{ret},f} = 1$. The chance that individual s^*_{ret} ($1 \leq s^*_{\text{ret}} \leq N_{s,f}$) represents this one retired individual in f can be expressed as:

$$P(s^*_{\text{ret}} = s) = \frac{\mathbf{D}^*_{\text{ret}}(G_s)}{\sum_{k=1}^{g_{\max}} \mathbf{D}^*_{\text{ret}}(G_k)} \tag{A.38}$$

3. The above chance can also be expressed as a chance distribution, in analogy with equation A.3:

$$F_{s^*_{\text{ret}}}(s) = P(s^*_{\text{ret}} \leq s) \tag{A.39}$$

4. Now, let a random z-value z_s be selected ($0 \leq z_s < 1$) for each adult member s in household f. According to this z-value, the retired status of an individual s is set to 1 according to the conditions in equation A.40. The same procedure applies in case more than one retired individual should be distributed over the members of a household ($N_{\text{ret},f} > 1$). In such cases however, it should be controlled that individuals who have been assigned as retired, cannot be assigned as such for a second time. In the definition of $S_{\text{ret},s}$ hereunder, this is controlled by the third condition.

$$S_{\text{ret},s} = 1 \quad \text{where} \quad \begin{cases} z_s \geq F_{s^*_{\text{ret}}}(s-1) \\ z_s < F_{s^*_{\text{ret}}}(s) \\ \hat{S}_{\text{ret},s} = 0 \end{cases} \tag{A.40}$$

5. The above procedure should be repeated until a total number of $N_{\text{ret},f}$ retired individuals has been successfully distributed over the household. In practical applications, a maximum number of iterations should be set (see page 177).

Individual student-status of $S_{\mathbf{stud},s}$ Let vector $\mathbf{D}^*_{\text{stud}}$ be introduced as the correlation between student status and age category, where each element $D^*_{\text{stud}}(g)$ provides with the relative chance of an individual s to be a full-time student, provided his or her age category G_s. On page 199, an estimate for this vector is calibrated:

$$\mathbf{D}^*_{\text{stud}} = \left| \begin{array}{cccccc} 0 & 0 & 357 & 51 & 4 & 2 \end{array} \right| \tag{A.41}$$

Based on this vector, the student status of each individual can be determined by the algorithm introduced for retired individuals at page 184. In this algorithm, vector $\mathbf{D}^*_{\text{ret}}$ should be replaced by vector $\mathbf{D}^*_{\text{stud}}$. In addition, it is required to add a new constraint to equation A.40. Next to the -adjusted- condition $\hat{S}_{\text{stud},s} = 0$ (an individual can only be assigned to be a student in case he or she hasn't been assigned so already), it should also be checked that $S_{\text{ret},s} = 0$, because of the definition of a student on page 171.

Individual employment-status $S_{\mathbf{empl},s}$ For employed individuals, a different approach need be followed as for retired individuals and students. This is because, other than for students and retired individuals, no aggregate parameter is yet available which provides the total number of employed individuals in household f. Instead, each individual household member s who is neither retired, nor a full-time student ($S_{\text{ret},s} = 0$ and $S_{\text{stud},s} = 0$) is *potentially* employed (see page 171). It is assumed that such individuals are employed under two conditions. First, the person should belong to the labour force. A person belongs to the labour force in case he or she is employed for 12 hours per week or more, or both capable to and willing to do so (see page 171). Let

this status be denoted as $S_{\text{lab},s} = 1$. Whether or not a person belongs to the labour force can stochastically be determined by means of the age category of the individual and a correlation between average participation in the labour force and age. Second, the person should indeed have found a job of 12 hours per week or more. Because the number of officially unemployed individuals in a household is known by parameter $N_{\text{unempl},f}$, it can now be synthesised which of the household members are employed.

1. Denote $\lambda_{\text{lab},g}$ as the absolute chance by which an individual of age category g belongs to the labour force as defined above. For this report, these chances are taken from national statistics for the Netherlands for 1995 (see page 199):

g	Description	$\lambda_{\text{lab},g}$	g	Description	$\lambda_{\text{lab},g}$
1	Young child	0.0%	4	Young adult	76.1%
2	Teenager	3.3%	5	Mid-age	62.6%
3	Grown-up child	39.5%	6	Old	19.2%

Table A.5: Labour force ratio $\lambda_{\text{lab},g}$ per age category

2. Let a random z-value z_s be selected for each household member s in household f ($0 \leq z_s < 1$). Now, each of these individuals is assigned to belong to the labour force ($S_{\text{lab},s} = 1$) in case the following three conditions are met.

$$S_{\text{lab},s} = 1 \quad \text{where} \quad \begin{cases} S_{\text{ret},s} = 0 \\ S_{\text{stud},s} = 0 \\ 0 \leq z_s < \lambda_{\text{lab},G_s} \end{cases} \tag{A.42}$$

3. Let the total number of employed individuals in household f be denoted by $N_{\text{empl},f}$, determined as:

$$N_{\text{empl},f} = \sum_{s=1}^{N_{s,f}} (S_{\text{lab},s}) - N_{\text{unempl},f} \tag{A.43}$$

4. Define an initial estimate $\hat{S}_{\text{empl},s}$ for each individual s within household f that he or she is not employed ($\hat{S}_{\text{empl},s} = 0$). In case $N_{\text{empl},f} \leq 0$, the assignment procedure ends here.

5. Let the ratio of employed individuals for each age category g be denoted as $\lambda_{\text{empl},g}$, in case these individuals belong to the labour force. Table A.6 shows the estimates used in this report, calibrated on page 200.

g	Description	$\lambda_{\text{empl},g}$	g	Description	$\lambda_{\text{empl},g}$
1	Young child	0.0%	4	Young adult	95.2%
2	Teenager	79.0%	5	Mid-age	94.6%
3	Grown-up child	92.2%	6	Old	85.9%

Table A.6: Employment ratio $\lambda_{\text{empl},g}$ per age category for individuals who belong to the labour force

6. Based on the employment ratio $\lambda_{\text{empl},g}$, the chance for each individual within household f to be employed is now expressed as:

$$P(s^*_{\text{empl}} = s) = \frac{S_{\text{lab},s}\lambda_{\text{empl},G_s}}{\sum_{k=1}^{N_{s,f}} S_{\text{lab},k}\lambda_{\text{empl},G_k}} \tag{A.44}$$

7. Let the above chance be rewritten as a chance distribution (compare equation A.3):

$$F_{s^*_{\text{empl}}}(s) = P(s^*_{\text{empl}} \leq s) \tag{A.45}$$

8. Finally, let a second random z-value z_f be selected ($0 \leq z_f < 1$). Individual s is assigned to be employed under the following conditions:

$$S_{\text{empl},s} = 1 \quad \text{where} \quad \begin{cases} z_f \geq F_{s^*_{\text{empl}}}(s-1) \\ z_f < F_{s^*_{\text{empl}}}(s) \\ \hat{S}_{\text{empl},s} = 0 \end{cases} \tag{A.46}$$

9. The previous step is repeated until a number of $N_{\text{empl},f}$ individuals within f has successfully been assigned to be employed. (compare final step on page 177)

Individual car availability $S_{\mathbf{car},s}$ Like individual employment, neither can individual car availability $S_{\text{car},s}$, be synthesised as straightforward as individual retired status $S_{\text{ret},s}$ or student status $S_{\text{empl},s}$ have been. This is because *two* individual characteristics are assumed to be of major influence on the car availability of an individual. These are first the age category of the individual G_s and second his or her employment status $S_{\text{empl},s}$. In order to incorporate both dependencies, the following algorithm is proposed:

1. First, let correlation vector $\mathbf{D}^*_{\text{car}}$ be introduced as the correlation between car possession and age category. Let each of the elements $D^*_{\text{car}}(g)$ of this vector express the relative chance of an individual s to generally have a car available for him- or herself, in case its household possesses one car

precisely, and the age category of this individual is G_s. On page 200, an example estimate for this vector is calibrated as:

$$\mathbf{D}^*_{\text{car}} = \left| \begin{array}{cccccc} 0 & 0 & 268 & 490 & 579 & 395 \end{array} \right| \tag{A.47}$$

2. Second, let factor α_{empl} denote the average factor by which in individual is more likely to have a car available, in case the individual is employed. The factor is calibrated on page 201:

$$\alpha_{\text{empl}} = 1.87 \tag{A.48}$$

3. Based on these definitions, the chance for individual s^*_{car} to have a car available mainly for him- or herself equals (in case the household possesses over precisely one car):

$$P(s^*_{\text{car}} = s) = \frac{\alpha_{\text{empl}}\mathbf{D}^*_{\text{car}}(G_s)}{\sum_{k=1}^{N_{s,f}} \alpha_{\text{empl}}\mathbf{D}^*_{\text{car}}(G_k)} \tag{A.49}$$

4. Let the above chance be expressed as a chance distribution as in equation A.50 (compare equation A.3). This equation does not only apply where a household possesses of one car only, but also for each additional car possessed by the household.

$$F_{s^*_{\text{car}}}(s) = P(s^*_{\text{car}} \leq s) \tag{A.50}$$

5. Define an initial estimate $\hat{S}_{\text{car},s}$ for each individual s within household f that no car is generally available ($\hat{S}_{\text{car},s} = 0$).

6. Finally, let a second random z-value z_s be selected ($0 \leq z_s < 1$). Individual s is assigned to have a car available under the following conditions:

$$S_{\text{car},s} = 1 \quad \text{where} \quad \begin{cases} z_s \geq F_{s^*_{\text{car}}}(s-1) \\ z_s < F_{s^*_{\text{car}}}(s) \end{cases} \tag{A.51}$$

7. The previous step is repeated until a total number of $N_{\text{car},f}$ cars have been assigned successfully over the members of household f (compare final step on page 177). Note the absence of a $\hat{S}_{\text{car},s} = 0$ condition, as multiple cars may be available to only one person.

Appendix B

Calibration of synthetic population generation vectors

This appendix gives example calibrations of the correlation matrices and vectors introduced in appendix A for the generation of synthetic populations. The appendix is subdivided in two sections. First, section B.1 calibrates all correlation matrices used in the disaggregation of zonal socioeconomic data to household characteristics. Next, section B.2 calibrates all correlation matrices used in the disaggregation of synthetic household data further down to individual characteristics. For the calibration, use has been made of the Wegener database (see page 171), national statistics of Statistics Netherlands (CBS) and the OVG database (see page 124).

B.1 From zones to households

Individuals per household-type $D_n(ht, n)$ The household type segmentation in table A.1 defines the household-size for all households without children. These household sizes equal 1 for all single households (type 1-3) and 2 for all pairs without children (type 4-6). For this assignment, it is noted that both the Wegener database and the OVG consider addresses with three or more adult household members as multiple separate households at the same address. This directly defines the correlation matrix between household size and household type for these six household types. Thus, should only the correlation between household size and household types with children need be analysed further (type 7-9). Because in the Wegener database, household size in itself is not included, Dutch national statistics have been used to calibrate this correlation. First, table B.1 shows the number of children for one-parent households and two-parent households in 1996.

Number of children $N_{c,f}$	one-parent households	two-parent households
1 child	222,766	778,422
2 children	108,547	920,469
3 children or more	37,462	406,746
	368,775	2,105,637

Table B.1: Number of in-living children in multi-person household-types for the Netherlands in 1996 (source: CBS, 1997c)

For households with 3 or more children, another publication of the CBS on youth statistics specifies that in 1999, households with 3 children occurred 2.87 times as much as households with 4 children or more (CBS, 2000). Let this ratio also be assumed for the occurrence of households with 4 children in relation to households with 5 children or more. Considering a maximum number of 6 children per household, also assume the same ratio for the occurrence of households with 5 children in relation to households with 6 children. Based on these assumptions, table B.1 is transformed into table B.2.

Number of children $N_{c,f}$	one-parent households	two-parent households
1 child	222,766	778,422
2 children	108,547	920,469
3 children	27,737	301,155
4 children	7,298	79,234
5 children	1,922	20,866
6 children	505	5,491
	368,775	2,105,637

Table B.2: Number of in-living children in multi-person household-types for the Netherlands in 1996 (continued)

Based on table B.2, the correlation between household size and household type can now be estimated for households with children. First, for households with 1 or 2 children (type 7 or 8) the relative occurrence of household sizes with 2,3 and 4 members equals 222,766 : (778,442 + 108,547) : (920,469 + 37,462) or 10:40:43 respectively. Similarly, for households with 3 or more children (type 9), the relative occurrence of 4,5,6,7 and 8 members equals 27,737 : (301,155 + 7,298) : (79,234 + 1,922) : (20,866 + 505) : 5,491 or 51:562:148:39:10 respectively. Based on the above discussion, matrix $\mathbf{D_n}$ is calibrated as:

$$\mathbf{D_n} = \begin{vmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 40 & 43 & 0 & 0 & 0 & 0 \\ 0 & 10 & 40 & 43 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 51 & 562 & 148 & 39 & 10 \end{vmatrix} \tag{B.1}$$

Children per household-type $D_c(ht, c)$ Table B.2 can also be used to correlate the number of children in a household to household type. First, because household-types 1 until 6 should not contain any children, the matrix is zero for these household types. For household types 7 and 8 (one or two children), the occurrence ratio is (222,766 + 778,422) : (108,547 + 920,469) = 100:103 for one and two children respectively. For household type 9 (three or more children), the occurrence ratio can similarly be determined as 549:144:38:10. The entire matrix $\mathbf{D_c}$ thus becomes:

$$\mathbf{D_c} = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 103 & 100 & 0 & 0 & 0 & 0 \\ 103 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & 549 & 144 & 38 & 10 \end{vmatrix} \tag{B.2}$$

Retired individuals per household-type $D_{\mathbf{ret}}(ht)$ In order to calibrate this vector, two tables are considered. First, table B.3 denotes the number of pensioned individuals per household category. In order to receive a pension, one has to be 65 or older in the Netherlands. However, for younger individuals, two other types of retirement allowance are possible as well. Table B.3 shows the number of individuals who receive either type of retirement allowance for the year 1997, for both relevant age categories distinguished in this report.

	# households with n pensioned individuals (×1000)						
Household category	$n = 0$		$n = 1$		$n = 2$		Total
Single household	1,588	(66.4%)	803	(33.6%)	-	-	2,391
-private	1,485	(68.8%)	673	(31.2%)	-	-	2,158
-institutional	103	(44.2%)	130	(55.8%)	-	-	233
Pair	1,429	(72.5%)	146	(7.4%)	396	(20.1%)	1,971
-private	1,422	(72.6%)	146	(7.5%)	391	(19.9%)	1,959
-institutional	7	(58.3%)	0	(0%)	5	(41.7%)	12
Family	2,328	(94.4%)	96	(3.9%)	41	(1.7%)	2,465
Total	5,345	(78.3%)	1,045	(15.3%)	437	(6.4%)	6,827

Table B.3: Number of retired individuals per household category for the Netherlands in 1997, source: CBS (1998b), CBS (1998a)

From tables B.3 and B.4, the relative frequencies in which each household-type contains one (additional) retired individual is determined in two steps. First, this relative frequency for all single households in relation to pairs and families is determined as 33.6 : (7.4 + 2*20.1) : (3.9 + 2*1.7) = 460:652:100. Second, these ratios are equally subdivided over mid-age and elder household types in the ratio 63,000:2,041,620 = 10:324. Finally, retired individuals are spread equally over the three family household types as the relative frequencies of these household types. These relative frequencies are determined by means of table

Household type	#individuals in age category 4 (35-60 years)	#individuals in age category 5 (> 60 years)
Early pension (pre-pensioen)	34,510	106,790
Early Retirement Scheme (VUT)	28,510	15,830
Pension	-	1,919,000
Total retired	63,020	2,041,620
Other	5,206,240	711,462
Total population	5,269,260	2,753,082

Table B.4: Number of retired individuals per age category, source: CBS (1998c), CBS (1997a)

B.1 and the assumption that retired individuals are three times as often head of a family with elder children (household type 8) as of a family with young children (household type 7). The resulting distribution key for household type 7:type 8:type 9 thus equals 20:62:18. This fully calibrates vector $\mathbf{D}_{\text{ret}}$:

$$\mathbf{D}_{\text{ret}} = \left| \begin{array}{ccccccccccc} 0 & 14 & 446 & 0 & 20 & 632 & 20 & 62 & 18 \end{array} \right| \tag{B.3}$$

Students per household-type $D_{\mathbf{stud}}(ht)$ No general statistical data could be found to relate the occurrence of students in a household directly to its household-type HT_f. Instead, the OVG has been used for this purpose. This database includes all required socioeconomic characteristics of around 1% of the Dutch population. In addition, it specifies to what degree each individual respondent is under- or overrepresented in the sample compared to the total population. This is specified by scale factors per respondent, defining for how many individuals in the Dutch population the observation is valid. Because the sample represents a rough 1% of the total Dutch population, the scale factors hover around 100. In order to calibrate $\mathbf{D}_{\text{stud}}$, it is first required to specify household type HT_f for the household f of each respondent r in the OVG. This variable can be derived from a combination of variables in the OVG such as household characteristic, household size, respondent age and position in the household. The latter two categories can be joined to represent age category AC_r (see table A.3). This data operation is straightforward because the specification of these OVG-variables precisely match with the specification of age category in table A.3. Thus, the following algorithm for the derivation of household type HT_f bases itself on household characteristic and household size as specified in the OVG, in combination with respondent age AC_r.

1. In case the household characteristic = single household, and respondent age category = 4,5 or 6, the individual is assigned to live in a household of type 1,2 or 3 respectively.

2. In case the household characteristic = pair without young children, and respondent age category = 4,5 or 6, the individual is assigned to live in a household of type 4,5 or 6 respectively.

3. In case the household characteristic = pair with young children (up to 10 year) and household size = 3 or 4, the individual is assigned to live in a household of type 7. This assignment is also made in case the household characteristic = single parent with children, and household size = 3.
4. In case the household characteristic = pair with older children (10 and older) and household size = 3 or 4, the individual is assigned to live in a household of type 8. This assignment is also made in case the household characteristic = single parent with children, and household size = 4.
5. All remaining individuals are assigned to belong to a household of type 9.

type	Age category						total
ht	1	2	3	4	5	6	population
1	0	0	0	758,009	0	0	758,009
2	0	0	0	0	434,238	0	434,238
3	0	0	0	0	0	826,062	826,062
4	0	0	0	1,367,035	0	0	1,367,035
5	0	0	0	0	1,261,270	0	1,261,270
6	0	0	0	0	0	1,558,208	1,558,208
7	1,261,740	328,165	9,355	1,453,247	400,895	3,290	3,456,692
8	1,990	510,842	860,339	110,489	1,341,588	197,249	3,022,497
9	671,593	601,163	342,774	436,926	464,173	14,310	2,530,939
total	1,935,323	1,440,170	1,212,468	4,125,706	3,902,164	2,599,119	15,214,950

Table B.5: Population per household type and age category in OVG 1995

Table B.5 shows the expected number of individuals per household type and age category for the whole of the Netherlands, based on the scaled number of observations in the OVG of 1995. Second, table B.6 shows the number of students per age category and household type. Children up to 17 year are excluded from this table, because the OVG labels high school students as students as well and vector $\mathbf{D}_{\text{stud}}$ should refer to higher education students.

type	Age category				total	total pop.	
ht	3	4	5	6	students	population	$D_{\text{stud}}(ht)$
1	0	104,275	0	0	104,275	758,009	13.8%
2	0	0	2,805	0	2,805	434,238	0.6%
3	0	0	0	2,420	2,420	826,062	0.3%
4	0	68,719	0	0	68,719	1,367,035	5.0%
5	0	0	3,981	0	3,981	1,261,270	0.3%
6	0	0	0	2,570	2,570	1,558,208	0.2%
7	4,243	6,129	1,733	0	12,105	1,866,787	0.6%
8	317,766	20,466	5,351	390	343,973	2,509,665	13.7%
9	132,106	10,999	3,453	0	146,558	1,258,183	11.6%
total	454,115	210,588	17,323	5,380	687,406	11,839,457	5.8%

Table B.6: Students per household type and age category in OVG 1995 (Source: CBS, 1995)

The last column in table B.6 reveals the percentage of individuals in each household type which are higher education students. Because a person younger

than 18 years cannot be such a student because of the above assumption, this percentage is based on the total number of individuals in the household which are 18 years or older (as indicated in the table). The resulting percentages define vector $\mathbf{D}_{\text{stud}}$ as:

$$\mathbf{D}_{\text{stud}} = \left| \begin{array}{cccccccccc} 138 & 6 & 3 & 50 & 3 & 2 & 4 & 114 & 58 \end{array} \right| \tag{B.4}$$

Unemployed individuals per wealth category $D_{\mathbf{unempl}}(w)$ From this vector, each element $D_{\text{car}}(w)$ should provide with the relative frequency in which a household of wealth category w contains an unemployed individual. An individual has been defined as unemployed in case he or she receives an unemployment allowance (see page 171). For the calibration of this vector, no official statistical data could be found. Instead, the correlation between both variables has been estimated by observation of the entire Wegener database as a whole. Let variables $W_z(1) \cdots W_z(5)$ be defined as binary indicators equal to 1 in case a zone is characterized by wealth category $1 \cdots 5$ respectively, or 0 otherwise. Also, define the percentage of unemployed individuals in a zone by P_z. Performing a regression analysis through the origin on the entire Wegener database, the amount of unemployed individuals could be estimated as:

$$P_z = 0W_z(1) + 0W_z(2) + 4.5W_z(3) + 8.7W_z(4) + 15.7W_z(5) \tag{B.5}$$

Based on equation B.5, vector $\mathbf{D}_{\text{unempl}}$ is calibrated as:

$$\mathbf{D}_{\text{unempl}} = \left| \begin{array}{ccccc} 0 & 0 & 10 & 19 & 42 \end{array} \right| \tag{B.6}$$

Cars in household per wealth category $D_{\mathbf{car}}(w)$ Let vector $\mathbf{D}_{\text{car}}$ consist out of $w_{\max}$ elements. Each element $D_{\text{car}}(w)$ should provide with the relative frequency in which a household of wealth category w is likely to possess an (additional) car out of a total amount of cars to be distributed for a zone. In order to calibrate this vector, consider table B.7, which reveals the number of cars possessed by households of various income groups for the Netherlands in 1996.

	Percentage of households with n cars			
Income category	$n = 0$	$n = 1$	$n = 2$	$n \geq 3$
1 (high)	5.7%	64.2%	28.0%	2.1%
2 (above-average)	14.9%	76.4%	8.5%	0.2%
3 (average)	29.9%	67.1%	3.0%	0.0%
4 (under-average)	58.9%	40.4%	0.7%	0.0%
5 (low)	80.2%	19.6%	0.2%	0.0%

Table B.7: Car possession per income category in 1996, (source: CBS, 1997e)

If the average amount of cars for households with 3 or more cars is assumed to equal 3.2 cars (considering the steep decrease for two cars to three or more cars), vector $\mathbf{D}_{\text{car}}$ is calibrated from table B.7 as:

$$\mathbf{D}_{\text{car}} = \begin{vmatrix} 127 & 94 & 73 & 42 & 20 \end{vmatrix} \tag{B.7}$$

B.2 From households to individuals

Age category of adults in families with young children $D_{A,ht_7}(g)$ Correlation vector $\mathbf{D_{A,ht_7}}$ should define the relative frequency in which the parents of a household belong to an age category g, if this household contains young children (household type 7). First, table B.8 shows historical data on the age of the mother at the birth of a child for the Netherlands.

# years	Amount of births per age category of the mother						
before 1996	$\leq$ 19 yr	20-24 yr	25-29 yr	30-34 yr	35-39 yr	40-44 yr	$\geq$ 45 yr
0-4 yr	11,202	100,892	326,932	378,357	131,908	17,808	1,028
5-9 yr	15,729	133,416	380,195	321,910	93,386	13,105	1,182
10-14 yr	16,478	170,256	383,409	236,984	62,430	8,851	994
15-17 yr	12,493	121,456	239,735	128,127	27,590	4,914	527
18-19 yr	8,631	86,153	157,953	76,245	16,140	3,383	341
20-24 yr	32,146	263,318	421,893	164,666	52,553	14,310	1,188
25-29 yr	45,439	349,913	419,464	229,271	107,295	34,596	3,492
30-34 yr	41,987	271,200	409,429	288,735	157,046	57,879	5,494
35-39 yr	25,797	206,492	394,785	311,179	186,606	67,721	7,224
40-44 yr	19,380	176,409	358,386	317,953	188,175	80,092	8,344

Table B.8: Historical births per age category of the mother, (source: CBS, 1997b)

Second, table B.9 estimates the percentage of all individuals which were born in a specific year, who still live with at least one of their parents. This percentage is calculated as the product of two other percentages. First, the individual should still be alive, as expressed by the average survival rate for his or her age category. Second, the individual should still live with one or both of his or her parents (this also incorporates a third factor; the combined survival rate of the parents). Both percentages are estimated by statistics for 1996.
This historical birth data in table B.8 and table B.9 can be transformed to a third table which estimates the absolute number of in-living children for each age category who live together with parents of each age-category for 1996. These estimates are shown in table B.10. The table represents the age category of the mother if all age intervals are subtracted with 2.5 years (half of the five-year age intervals per age category). However, this effect is largely compensated for in case it is assumed that on average, fathers are older than the mothers.

Finally, the age categories of the parents of table B.10 are aggregated for the three adult age categories distinguished in this study. Table B.11 presents the resulting association matrix between the age category of an in-living child and

Age category	% average survival rate	% individuals still living with parents	% all births still living with parents
0-4 yr	99.3%	99.3%	98.6%
5-9 yr	99.2%	99.3%	98.5%
10-14 yr	99.1%	98.9%	98.0%
15-17 yr	99.0%	96.7%	95.7%
18-19 yr	98.9%	84.8%	83.9%
20-24 yr	98.7%	26.3%	26.0%
25-29 yr	98.4%	5.9%	5.8%
30-34 yr	98.1%	2.7%	2.5%
35-39 yr	97.6%	2.6%	2.5%
40-44 yr	96.8%	1.8%	1.7%

Table B.9: Percentage of historical births still living with parents per age category for 1996, (source: CBS, 1997d)

	Age of children		
Age of parents	$\leq$ 10 years	10-17 years	$\geq$ 18 years
$<$ 20 years	11,046	0	0
20 - 24 years	114,978	0	0
25 - 29 years	453,793	16,150	0
30 - 34 years	747,592	178,828	7,239
35 - 39 years	447,167	492,052	80,599
40 - 44 years	109,550	461,773	203,461
45 - 49 years	13,923	183,847	194,887
50 - 54 years	1,164	35,088	88,471
55 - 59 years	0	5,679	46,212
60 - 64 years	0	505	30,969
65 - 69 years	0	0	20,618
$\geq$ 70years	0	0	18,878
total	1,899,214	1,373,921	691,332

Table B.10: Number of in-living children per parent and child age category for 1996

his or her parents:

Based on table B.11, correlation vector $\mathbf{D_{A,ht_7}}$ is be estimated as:

$$\mathbf{D_{A,ht_7}} = \begin{vmatrix} 0 & 0 & 0 & 23 & 10 & 0 \end{vmatrix} \tag{B.8}$$

Age category of children in families with older children $D_{C,ht_8}(g)$ Analogous to vector $\mathbf{D_{A,ht_7}}$, let vector $\mathbf{D_{C,ht_8}}$ provide with the relative frequency in which the in-living children of a household belong to an age category g, if this household contains old children (household type 8). This vector is estimated from table B.11 as:

$$\mathbf{D_{C,ht_8}} = \begin{vmatrix} 0 & 20 & 10 & 0 & 0 & 0 \end{vmatrix} \tag{B.9}$$

Age of parents	Age of children $\leq$ 10 years ($g=1$)	10-17 years ($g=2$)	$\geq$ 18 years ($g=3$)	total
$<$ 35 years ($g=4$)	1,327,409	194,978	7,239	1,529,626
35 - 59 years ($g=5$)	571,804	1,178,439	613,629	2,363,872
$\geq$ 60years ($g=6$)	0	505	70,465	70,969
total	1,899,214	1,373,921	691,332	3,964,467

Table B.11: Number of in-living children per parent and child age category for 1996

Age category of adults in families with older children $D_{A,ht_8}(g)$ Similar to the calibration of $\mathbf{D_{C,ht_8}}$ above, let vector $\mathbf{D_{A,ht_8}}$ be estimated from table B.11, where children of age category 2 and 3 together represent the 'old' children. This calibrates the vector:

$$\mathbf{D_{A,ht_8}} = \left| \begin{array}{cccccc} 0 & 0 & 0 & 28 & 253 & 10 \end{array} \right| \tag{B.10}$$

Age category of children in families with many children $D_{C,ht_9}(g)$ Vector $\mathbf{D_{C,ht_9}}$ is defined out of $g_{\max}$ elements, where each element $D_{C,ht_9}(g)$ represents the relative frequency of an individual having age category g provided he or she is an in-living child in households with three or more children (household type 9). This paragraph calibrates this vector for the Netherlands in 1996 for families with three children. The same vector is then assumed for households with three or more children. Furthermore, it is considered that the children in such households differ by two years each on average. Also, let all children be modelled to have the same age as the middle child. If all children are still to be alive, and still to live with their parents, the chance percentages in table B.9 need be put to the third power. This is to incorporate that where it is not uncommon for one child to remain with its parents after the age of 25, it would be much more uncommon if all three children out of three would still do so. As a second modification of table B.9 it is considered that children under the age of two cannot belong to a three-child household in the above model setup: their younger brother or sister is not yet born. This reflects that the average age of the children in households with three children cannot be as low as for households with less children, because most often they will differ in age for some years. Because these first two years represent 40% of the age interval from 0 up to 4 years, only 60% of the children in this age category are eligible to belong to a three-child household. Table B.12 reveals the consequent relative index in which children of various age categories belong to a household of three children. Second, table B.13 states the number of children who live in a household with a specified number of children. This table is obtained directly from table B.2 on page 190. Finally, tables B.12 and B.13 can be joined to distribute the children living in households with three children over their respective age categories, as shown in table B.14. From table B.14, vector $\mathbf{D_{C,ht_9}}$ is calibrated as:

$$\mathbf{D_{C,ht_9}} = \left| \begin{array}{cccccc} 76 & 68 & 10 & 0 & 0 & 0 \end{array} \right| \tag{B.11}$$

Age category	relative index of children eligible to belong to a three-child household	% chance of all three children still living with parents	relative index of children which are part of three-child household
$\leq$ 4 yr	60	95.9 %	57.5
5-9 yr	100	95.6 %	95.6
10-14 yr	100	94.1 %	94.1
15-17 yr	100	87.7 %	87.7
18-19 yr	100	59.0 %	59.0
20-24 yr	100	1.7 %	1.7
$\geq$ 25 yr	100	0.0 %	0.0

Table B.12: Relative index of children which are part of a three-child household per age category, 1996

number of in-living children per household	number of households	number of in-living children
1 child	1,001,188	1,001,188
2 children	1,029,016	2,058,032
3 children	328,892	986,676
4 children	86,532	346,128
5 children	22,788	113,940
6 children	5,996	35,976
	2,474,412	4,541,94

Table B.13: Number of in-living children in multi-person household-types for the Netherlands in 1996 (continued)

Age category of adults in families with many children $D_{A,ht_9}(g)$ The last age correlation vector $D_{A,ht_9}(g)$ can be calibrated in analogy with $\mathbf{D_{A,ht_7}}$ (see page 195). However, instead of using the relative chance percentages of table B.12, the chances in table B.9 are used as they apply for in-living children in three-children households. Now, a similar table as B.11 can be obtained from the historical birth table B.8. Next, the outcomes are fitted to the total number of children which lived in three-children households in 1996. The resulting correlations are shown in table B.15. Vector $\mathbf{D_{A,ht_9}}$ can now be calibrated as:

$$\mathbf{D_{C,ht_9}} = \left| \begin{array}{cccccc} 0 & 0 & 0 & 35 & 55 & 0 \end{array} \right| \tag{B.12}$$

Retired status per age category $D^*_{\mathbf{ret}}(g)$ Let vector $\mathbf{D}^*_{\text{ret}}$ consist of $g_{\max}$ categories, where each element $D^*_{\text{ret}}(g)$ provides with the relative chance of an individual of age-category g to be retired. This vector is calibrated by means of table B.4. Whereas no people in the age categories below 35 years are retired,

	number of in-living children	relative index of children which are part of three-child household	number of children in three-child household
$\leq$ 4 years	980,013	57.5	185,553
5 - 9 years	961,591	95.6	302,702
10 - 14 years	898,547	94.1	278,418
15 - 17 years	540,194	87.7	155,997
18 - 19 years	313,705	59.0	60,945
20 - 24 years	546,772	1.7	3,061
25 - 29 years	337,272	0.0	0
total	4,557,089	395.6	986,676

Table B.14: Number of children per age category in households with three children for 1996 (source: CBS, 1997a)

	Age of children			
Age of parents	$\leq$ 10 years ($g = 1$)	10-17 years ($g = 2$)	$\geq$ 18 years ($g = 3$)	total
$<$ 35 years ($g = 4$)	321,722	61,567	1,678	384,968
35 - 59 years ($g = 5$)	163,930	365,897	71,553	601,381
$\geq$ 60 years ($g = 6$)	0	152	175	328
total	485,652	427,617	73,407	986,676

Table B.15: Age category of parents if all children would live in a three-child family for 1996

the percentages of retired individuals for the age category 5 (between 35 and 60 years) and age category 6 (60 years or older) are $63,020/5,269,260 = 1.2\%$ and $2,041,620/2,753,082 = 74.2\%$ respectively. From these percentages, vector $\mathbf{D}^*_{\text{ret}}$ is calibrated as:

$$\mathbf{D}^*_{\text{ret}} = \left| \begin{array}{cccccc} 0 & 0 & 0 & 0 & 10 & 618 \end{array} \right| \tag{B.13}$$

Student status per age category $D^*_{\mathbf{stud}}(g)$ Vector $\mathbf{D}^*_{\text{stud}}$ should consist of $g_{\max}$ categories, of which each element $D^*_{\text{stud}}(g)$ provides with the relative chance of an individual of age-category g to be a full-time student. This vector is calibrated by combining table B.5 and table B.6. Table B.16 groups the total population and number of students for all relevant age categories from both tables. The ratio of both totals defines $D^*_{\text{stud}}(g)$:

$$\mathbf{D}^*_{\text{stud}} = \left| \begin{array}{cccccc} 0 & 0 & 357 & 51 & 4 & 2 \end{array} \right| \tag{B.14}$$

Labour force ratios per age category $\lambda_{\mathbf{lab},g}$ Let $\lambda_{\text{lab},g}$ ($g = 1 \cdots g_{\max}$) denote the absolute chance of an individual of age-category g to belong to the national labour force. No general statistics could be found for the calibration of this vector. Instead, the OVG is used, as it was done for the calibration of

Age g	total population	total students	$D^*_{\text{stud}}(g)$
3	1,212,468	454,115	35.7%
4	4,125,706	210,588	5.1%
5	3,902,164	17,323	0.4%
6	2,599,119	5,380	0.2%
total	11,839,457	687,406	5.8%

Table B.16: Total students per household type and age category in OVG 1995 (Source: CBS, 1995)

$D_{\text{stud}}(ht)$ on page 192. Table B.17 shows the total population, the number of employed individuals, and the number of individuals who is either employed or looking for work (the labour force) for each age category. The table represents the number of observed respondents in the OVG database, extended for the entire population by means of the appropriate scale factors (see page 192).

g	total population	labour force	ratio $\lambda_{\text{lab},g}$	population employed	employment ratio	$\lambda_{\text{empl},g}$
1	1,935,324	0	0.0%	0	0.0%	-
2	1,440,170	57,935	4.0%	45,772	31.8%	79.0%
3	1,212,468	647,453	53.4%	596,634	49.2%	92.2%
4	4,125,705	3,135,706	76.0%	2,984,295	72.3%	95.2%
5	3,902,164	2,477,388	63.5%	2,342,567	60.0%	94.6%
6	2,599,118	149,442	5.8%	128,336	4.9%	85.9%
total	15,214,949	6,467,924	42.5%	6,097,604	40.1%	94.3%

Table B.17: Labour force 1996 and employed population 1995 CBS (1995), CBS (2003)

Table B.17 includes the labour force ratios per age group. As it is expected, this ratio is very low for children under 18 years, as well as for individuals of over 60 years of age, and high for the age categories in between.

Employment ratios per age category $\lambda_{\mathbf{empl},g}$ Factor $\lambda_{\text{empl},g}$ $(g = 1 \cdots g_{\max})$ is defined as the chance of an individual of age-category g to be employed in case he or she belongs to the labour force. These factors can be observed directly in table B.17. The ratio takes a value of 79% for age category 2 up to 95.2% for age category 4. The estimate for age category 2 may seem high, after all, these are in-living children between 10 and 17 years. However, it should be considered that for this ratio to apply, the child should first belong to the labour force. Only four percent of these children belong to this group.

Individual car-availability per age category $D^*_{\mathbf{car}}(g)$ Let vector $\mathbf{D}^*_{\text{car}}$ consist of $g_{\max}$ categories, where each element $D^*_{\text{car}}(g)$ provides with the relative chance of an individual of age-category g to have a car available. In order to calibrate this vector, consider table B.18. This table reveals the ratio by which an individual of each age category has at least one car available.

Age	total population	population with car available	car availability ratio
$\leq$ 17 yr	3,401,860	0	0.0%
18-19 yr	369,749	1,849	5.0%
20-24 yr	1,079,959	303,468	28.1%
25-34 yr	2,619,979	1,215,670	46.4%
35-39 yr	1,238,102	694,575	56.1%
40-49 yr	2,355,818	1,392,288	59.1%
50-59 yr	1,675,340	988,451	59.0%
60-64 yr	692,210	364,102	52.6%
65-74 yr	1,180,553	481,666	40.8%
$\geq$ 75 yr	880,319	231,524	26.3%
total	15,493,889	5,673,594	36.6%

Table B.18: Car availability ratio per age group, 1996, (source: CBS, 1997e)

Because this data source does not distinguish between in-living children and independently living adults, table B.18 cannot be used to assess the car availability ratio for the age categories defined in this study. Therefore, the OVG is once again used to estimate car availability for these age categories. The estimates are shown in table B.19, and can be compared to those of table B.18.

Age	#respondents		population equivalents		
g	no car available	car available	no car available	car available	$D^*_{\text{car}}(g)$
1	0	23,822	0	1,935,324	0.0%
2	0	17,565	0	1,440,170	0.0%
3	4,127	8,421	325,433	887,035	26.8%
4	23,667	19,318	2,021,962	2,103,743	49.0%
5	26,289	19,765	2,260,033	1,642,132	57.9%
6	11,185	13,762	1,025,369	1,573,749	39.5%
total	65,268	102,653	5,632,797	9,582,153	37.0%

Table B.19: Car availability ratio per age group, 1996, (source: CBS, 1997e)

The resulting vector $\mathbf{D}^*_{\text{car}}$ equals:

$$\mathbf{D}^*_{\text{car}} = \left| \begin{array}{cccccc} 0 & 0 & 268 & 490 & 579 & 395 \end{array} \right| \tag{B.15}$$

Car-availability for employed individuals α_{empl} Finally, let α_{empl} denote the relative frequency in which employed individuals of all ages have availability over a car, compared to non-employed individuals. It is first tried to calibrate this constant by means of table B.20. This table provides insight in the car availability ratio per social group, but no data was available for the relative sizes of each segment. Without such data, table B.20 could not be aggregated in order to estimate α_{empl}.
Therefore, let a second table be considered. Table B.21 provides the number of respondents in the OVG of 1995 which are either employed or have the availability over a car. The total number of respondents in this database totals 167,921. However, by means of scale factors included in the database, this number could be projected for the entire Dutch population in 1995 (see page

social group	car availability ratio
Employed $<$ 30 hours	44.4 %
Employed $\geq$ 30 hours	66.2 %
Work in household	35.0 %
Student	1.6 %
Retired	44.0 %
Unemployed	
Receiving disability benefit (WAO)	57.2 %
Others	24.2 %

Table B.20: Car availability ratio per social group 1996, (source: CBS, 1997e)

200). In this way, table B.21 includes the number of Dutch inhabitants which belong to either classification. The weighted totals included in the table denote the relative frequency in which employed and non-employed individuals of 18 years or older have a car available.

Employment Status	Age	N	% without car	% with car
Not employed	$\leq$ 17 yr	3,362,491	100.0%	0.0%
	18 – 29 yr	1,351,842	85.2%	14.8%
	30 – 39 yr	1,080,307	60.4%	39.6%
	40 – 59 yr	2,120,011	57.4%	42.6%
	$\geq$ 60 yr	2,523,931	61.8%	38.2%
weighted average (excluding individuals $\leq$ 17 yr)			64.7%	35.3%
Employed	$\leq$ 18 yr	19,977	100.0%	0.0%
	18 – 29 yr	1,412,027	48.7%	51.3%
	30 – 39 yr	1,455,569	33.1%	66.9%
	40 – 59 yr	1,811,069	24.0%	76.0%
	$\geq$ 60 yr	77,725	20.6%	79.4%
weighted average (excluding individuals $\leq$ 17 yr)			34.1%	65.9%

Table B.21: Relationship between employment status and age category and car possession in the Netherlands in 1995, (source: CBS, 1995)

Based on table B.21, the constant α_{empl} is determined as 65.9% to 35.3% = 187 to 100:

$$\alpha_{\text{empl}} = 1.87 \tag{B.16}$$

Appendix C

Observed interregional train trips 1995-2004

This appendix gives the number of interregional train trips observed for four different years. Section 5.5 compares these observations with the synthetic estimates presented in table 5.20. Tables C.1 until C.4 show the number of train trips observed from an origin zone i towards destination zone j where the train was the main mode of transport (for a list of zones, see table 5.1). In case more interregional train trips have been reported within the same household, only the first such trip is included, in order to preserve the independency of the observed trips. Note that no direct comparison can be made from differences for individual cells or trip totals between the tables, because the data in each table is based on different sample sizes.

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	11	6	1	2	1	1	3	0	3	0	1	0	0	2	0	31
2	7	-	18	2	2	1	10	4	4	6	3	0	2	1	0	1	61
3	7	15	-	8	9	1	14	6	3	12	2	0	0	2	1	0	80
4	2	3	15	-	13	2	4	2	1	6	2	0	2	1	1	0	54
5	0	0	15	13	-	0	13	1	1	14	17	1	0	3	1	0	79
6	1	0	1	1	2	-	162	13	1	8	1	0	1	1	0	1	193
7	1	0	6	2	4	22	-	59	12	74	4	1	3	4	2	2	196
8	4	2	3	1	1	0	80	-	80	64	7	2	2	5	3	1	255
9	1	2	0	3	2	3	19	111	-	38	3	2	8	6	6	3	207
10	2	1	19	4	10	4	121	61	31	-	24	1	1	14	8	3	304
11	2	3	3	4	10	4	24	17	10	50	-	1	3	19	10	2	162
12	0	0	0	0	0	0	2	2	5	2	0	-	5	0	0	0	16
13	0	0	1	0	0	0	1	11	25	4	2	5	-	22	11	3	85
14	2	1	2	0	2	2	17	9	3	20	9	0	16	-	28	2	113
15	0	0	3	0	1	0	8	8	5	25	10	0	8	35	-	17	120
16	0	0	0	0	0	0	6	3	2	7	4	0	2	5	22	-	51
Σ_i	29	38	92	39	58	40	482	310	183	333	88	14	53	118	95	35	2007

Table C.1: Interregional train trips observed for 1995 (Source: OVG 1995)

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	11	6	0	0	0	2	2	2	3	4	0	1	0	0	0	31
2	10	-	12	2	0	0	11	3	2	8	2	0	1	0	1	0	52
3	7	10	-	9	6	0	11	4	4	19	5	0	1	1	2	0	79
4	0	0	3	-	6	1	5	1	0	7	6	0	0	1	0	0	30
5	1	1	5	7	-	1	4	6	1	12	12	0	0	1	0	0	51
6	0	0	1	0	0	-	146	8	1	14	0	0	1	3	1	0	175
7	1	2	3	2	3	34	-	54	23	61	8	2	0	6	5	4	208
8	1	0	1	1	2	2	87	-	78	65	11	2	6	5	5	5	271
9	1	2	1	0	5	0	16	66	-	38	7	3	11	4	0	0	154
10	1	1	13	4	5	3	129	55	27	-	21	1	1	14	6	3	284
11	1	1	5	4	7	2	21	12	2	50	-	0	2	14	11	3	135
12	0	0	0	0	0	1	3	2	2	3	0	-	3	0	1	0	15
13	0	0	0	0	0	0	8	9	21	4	4	4	-	14	4	0	68
14	1	1	2	2	0	2	16	9	9	23	14	0	15	-	27	2	123
15	1	2	0	0	1	1	14	8	2	14	11	0	7	28	-	19	108
16	0	0	0	0	0	0	4	2	4	9	5	0	0	4	15	-	43
Σ_i	25	31	52	31	35	47	477	241	178	330	110	12	49	95	78	36	1827

Table C.2: Interregional train trips observed for 1998 (Source: OVG 1998)

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	18	0	0	1	0	1	0	1	1	0	0	0	0	0	0	22
2	9	-	10	3	0	0	4	3	0	7	1	1	0	1	1	1	41
3	4	12	-	12	2	0	8	6	2	11	1	0	0	0	1	0	59
4	2	1	6	-	11	0	5	3	2	5	0	1	0	0	0	0	36
5	0	1	13	1	-	3	10	3	2	15	24	0	0	3	0	0	75
6	0	0	1	0	1	-	118	4	3	8	2	0	0	0	0	0	137
7	2	1	3	0	7	19	-	72	16	80	4	0	3	2	2	3	214
8	1	3	3	0	1	2	89	-	68	46	4	1	1	2	4	0	225
9	0	1	1	0	0	2	23	73	-	20	4	3	13	2	5	0	147
10	2	3	6	3	6	1	95	37	22	-	14	2	2	17	5	2	217
11	1	0	3	0	7	0	17	6	2	44	-	1	2	18	6	4	111
12	0	0	1	0	0	0	1	5	4	1	1	-	2	2	0	0	17
13	0	0	0	0	0	0	3	12	15	3	0	3	-	14	4	0	54
14	0	0	0	2	0	0	12	3	5	17	13	2	9	-	29	1	93
15	1	0	0	0	1	0	14	5	4	11	6	1	2	21	-	18	84
16	0	0	0	0	0	0	2	2	4	5	1	0	0	4	15	-	33
Σ_i	22	40	47	21	37	27	402	234	150	274	75	15	34	86	72	29	1565

Table C.3: Interregional train trips observed for 2001 (Source: OVG 2001)

	Destination j																
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Σ_j
1	-	8	4	0	1	0	2	0	2	2	0	0	0	0	1	0	20
2	9	-	7	1	0	0	1	1	0	4	0	1	0	0	0	0	24
3	5	5	-	7	2	0	13	0	1	10	2	0	0	0	0	0	45
4	1	1	4	-	7	0	0	0	0	2	4	0	0	0	0	0	19
5	3	3	3	3	-	0	5	0	2	5	13	0	0	0	0	0	37
6	0	0	0	0	0	-	49	3	0	6	0	0	0	1	1	0	60
7	1	0	2	1	0	8	-	20	7	49	4	0	0	2	6	1	101
8	0	1	0	0	0	4	37	-	29	21	3	0	2	0	1	4	102
9	1	1	0	0	1	0	6	26	-	14	1	1	1	2	3	0	57
10	1	0	4	1	2	0	59	21	12	-	15	0	0	3	4	3	125
11	0	0	1	0	1	0	7	1	4	13	-	0	0	6	3	1	37
12	0	1	0	0	0	0	1	2	10	1	2	-	3	0	0	0	20
13	1	0	0	0	0	1	1	1	7	0	0	3	-	4	5	0	23
14	1	0	0	0	0	0	2	2	0	8	8	0	3	-	9	0	33
15	0	0	0	0	0	0	4	2	0	4	1	0	0	12	-	6	29
16	0	0	0	0	0	0	1	1	0	4	1	0	0	1	8	-	16
Σ_i	23	20	25	13	14	13	188	80	74	143	54	5	9	31	41	15	748

Table C.4: Interregional train trips observed for 2004 (Source: MON 2004)

Appendix D

Population cluster scheme of case study Southern Overijssel

Table D.1 shows the specification of all 98 population groups defined for case study Southern Overijssel. Several of the classification variables require further clarification. First, social participation distinguishes between employed individuals, students, retired individuals and other individuals. For each age category, this latter social participation category refers to all individuals not specifically addressed by other population segments in the respective age category. Second, household-type distinguishes between single households, pairs, and families. The latter two of these categories actually refer to any multi-person household, the first without children, the second with one or more children. Finally, income is denoted by symbols: $++$ denotes a very high income, $+$ a high income, 0 an average income, $-$ a low income and $--$ a very low income, as they are represented in the OVG database.

Frequency statistics For each population group, three frequency statistics are provided. First, N_{TB}, specifies the total number of respondents that was available within TB (equalling the complete OVG, but filtered for average workdays, see section 6.2). These respondents can live anywhere in the Netherlands. Second, N^*_{TB} specifies the expected number of individuals per population group in the synthesise area $\mathcal{Z}$ for case study Southern Overijssel. This expected number of individuals is determined by multiplication of N_{TB} with scale factors included in the OVG. These scale factors correct for population groups that are under- or over-represented in the sample. By means of these scale factors, the observed sample N_{TB} is extended to an expected number of individuals of the same population cluster who reside in $\mathcal{Z}$ in case $\mathcal{Z}$ would be a perfectly normal and average sub-area of the Netherlands. Finally, N_{SE} specifies the number of synthesised individuals in SE. This database is a result of the synthetic pop-

ulation algorithm of appendix A, applied to the socioeconomic database *SE* (see section 6.3). In order to interpret the differences between N^*_{TB} and N_{SE}, it need be considered that the estimates for N^*_{TB} are projections of N_{TB} for the study area of the case study. These estimates are obtained by multiplying the observed frequencies N_{TB} by scale factors included in the OVG that define the relative occurrence of groups of respondents in the OVG and comparable groups of respondents in the whole of the Netherlands. Possible causes of the differences between N^*_{TB} and N_{SE} are:

- The scale factors in the OVG are national scale factors, whereas the region for which it is applied may well differ from the national average.
- The scale factors are not appropriate for some population groups, as they do not control for some of its variables.
- The socioeconomic database *SE* is not accurate for some of its zones or on some of its variables.
- The various simplifications in the synthetic population generation algorithm lead to incorrect assignments for some of the synthesised zones or for some of the synthesised variables.

Only the first of these causes can be estimated. The study area is mostly rural, with some medium-sized cities. In addition, it is located at a large distance from the socioeconomic centre of the Netherlands, the four large cities in the West. This characterisation causes the study area to be distinct from the national average in the following aspects:

- Because main offices and industry is not located within or near to the area, the number of highly-paid jobs, and thus the number of very rich households is expected relatively low.
- At the same time, also the number of very poor households is expected to be relatively low, as the area does not contain any major socioeconomic problem areas, as these are typically found in the large cities.
- Because of relatively low house prices and relatively good availability of rented houses, people can be expected to live on their own sooner than on average in the Netherlands.
- Because of the limited and rather small higher education institutions in the study area, a lower number of students is expected

The above list provides explanations for most (but not all) of the observed large differences between N^*_{TB} and N_{SE}. Thus, it is concluded that the generation of a synthetic population by the algorithm specified in appendix A, by the calibrated matrices and vectors from appendix B and by the Wegener database at the six-digit postal code level, has performed well for the subsequent population cluster scheme $\mathcal{P}$.

cluster P_r	age category	in-living child	social particip.	household type	car poss.	household income	N_{TB}	N^*_{TB}	N_{SE}
1	$\leq$ 9 yr	yes	-	-	-	++ / +	8582	38110	28187
2	$\leq$ 9 yr	yes	-	-	-	0	6446	27799	28759
3	$\leq$ 9 yr	yes	-	-	-	-	2306	9877	19508
4	$\leq$ 9 yr	yes	-	-	-	- -	1850	7979	9864

cluster P_r	age category	in-living child	social particip.	household type	car poss.	household income	N_{TB}	N^*_{TB}	N_{SE}
5	10-17 yr	yes	empl.	-	-	all	349	1981	2379
6	10-17 yr	yes	other	-	-	++ / +	4712	32499	19374
7	10-17 yr	yes	other	-	-	0	2624	17648	19736
8	10-17 yr	yes	other	-	-	- / - -	1370	10206	20224
9	$\geq$ 18 yr	yes	empl.	-	yes	++ / +	1615	8059	1524
10	$\geq$ 18 yr	yes	empl.	-	yes	0	622	3328	1367
11	$\geq$ 18 yr	yes	empl.	-	yes	- / - -	63	358	841
12	$\geq$ 18 yr	yes	empl.	-	no	++ / +	1419	9808	2444
13	$\geq$ 18 yr	yes	empl.	-	no	0 / - / - -	647	4271	5282
14	$\geq$ 18 yr	yes	stud.	-	-	++ / +	1938	12778	305
15	$\geq$ 18 yr	yes	stud.	-	-	0	768	4872	442
16	$\geq$ 18 yr	yes	stud.	-	-	- / - -	296	2005	703
17	$\geq$ 18 yr	yes	other	-	yes	all	291	1384	2901
18	$\geq$ 18 yr	yes	other	-	no	++ / +	478	3008	2330
19	$\geq$ 18 yr	yes	other	-	no	0	280	1809	2445
20	$\geq$ 18 yr	yes	other	-	no	- / - -	136	799	2945
21	$\leq$ 34 yr	no	empl.	single	yes	++ / +	179	1545	906
22	$\leq$ 34 yr	no	empl.	single	yes	0	430	3837	1233
23	$\leq$ 34 yr	no	empl.	single	yes	- / - -	924	8496	1362
24	$\leq$ 34 yr	no	empl.	single	no	++ / +	36	458	656
25	$\leq$ 34 yr	no	empl.	single	no	0 / - / - -	821	9770	3261
26	$\leq$ 34 yr	no	empl.	pair	yes	++ / +	3516	19045	12905
27	$\leq$ 34 yr	no	empl.	pair	yes	0	1308	7156	11740
28	$\leq$ 34 yr	no	empl.	pair	yes	- / - -	263	1577	7297
29	$\leq$ 34 yr	no	empl.	pair	no	++ / +	2045	14575	12405
30	$\leq$ 34 yr	no	empl.	pair	no	0	992	7122	13801
31	$\leq$ 34 yr	no	empl.	pair	no	- / - -	317	2664	15810
32	$\leq$ 34 yr	no	empl.	family	yes	++ / +	3499	17048	9602
33	$\leq$ 34 yr	no	empl.	family	yes	0	1932	9615	8567
34	$\leq$ 34 yr	no	empl.	family	yes	- / - -	1114	5684	5662
35	$\leq$ 34 yr	no	empl.	family	no	++ / +	1745	10455	8402
36	$\leq$ 34 yr	no	empl.	family	no	0	1076	6488	9097
37	$\leq$ 34 yr	no	empl.	family	no	- / - -	586	3634	10956
38	$\leq$ 34 yr	no	stud.	single	-	all	388	4513	4931
39	$\leq$ 34 yr	no	stud.	pair/fam.	-	all	595	4601	5532
40	$\leq$ 34 yr	no	other	single	-	all	417	4190	3113
41	$\leq$ 34 yr	no	other	pair	yes	all	341	1555	9400
42	$\leq$ 34 yr	no	other	pair	no	++ / +	134	758	4398
43	$\leq$ 34 yr	no	other	pair	no	0 / - / - -	420	2935	12186
44	$\leq$ 34 yr	no	other	family	yes	++ / +	831	3531	2702
45	$\leq$ 34 yr	no	other	family	yes	0	877	3777	2793
46	$\leq$ 34 yr	no	other	family	yes	- / - -	621	2781	2069
47	$\leq$ 34 yr	no	other	family	no	++ / +	1084	5427	2907
48	$\leq$ 34 yr	no	other	family	no	0	1656	8430	3612
49	$\leq$ 34 yr	no	other	family	no	- / - -	1295	6905	4653
50	35-59 yr	no	empl.	single	yes	++ / +	266	2267	1320
51	35-59 yr	no	empl.	single	yes	0	278	2542	1459
52	35-59 yr	no	empl.	single	yes	- / - -	395	3479	1467
53	35-59 yr	no	empl.	single	no	all	318	3211	3941
54	35-59 yr	no	empl.	pair	yes	++ / +	4470	23589	7826
55	35-59 yr	no	empl.	pair	yes	0	1268	7025	7087
56	35-59 yr	no	empl.	pair	yes	- / - -	429	2327	4632
57	35-59 yr	no	empl.	pair	no	++ / +	1761	9922	7745
58	35-59 yr	no	empl.	pair	no	0	617	3346	8429
59	35-59 yr	no	empl.	pair	no	- / - -	215	1152	10533
60	35-59 yr	no	empl.	family	yes	++ / +	4014	19789	11003
61	35-59 yr	no	empl.	family	yes	0	1415	7132	9760
62	35-59 yr	no	empl.	family	yes	- / - -	602	3213	6368
63	35-59 yr	no	empl.	family	no	++ / +	1577	8024	9337
64	35-59 yr	no	empl.	family	no	0 / - / - -	908	4373	22393
65	35-59 yr	no	ret.	-	-	all	960	6070	15501
66	35-59 yr	no	other	single	-	all	715	6456	6021
67	35-59 yr	no	other	pair	yes	++ / +	1117	5626	4177
68	35-59 yr	no	other	pair	yes	0	650	3514	3915
69	35-59 yr	no	other	pair	yes	- / - -	388	2310	2673
70	35-59 yr	no	other	pair	no	++ / +	1982	10030	5032

cluster P_r	age category	in-living child	social particip.	household type	car poss.	household income	N_{TB}	N^*_{TB}	N_{SE}
71	35-59 yr	no	other	pair	no	0	1567	8230	5907
72	35-59 yr	no	other	pair	no	- / - -	1150	6134	7735
73	35-59 yr	no	other	family	yes	++ / +	795	3694	5945
74	35-59 yr	no	other	family	yes	0 / - / - -	759	3659	9827
75	35-59 yr	no	other	family	no	++ / +	1168	5140	6339
76	35-59 yr	no	other	family	no	0	892	4090	7531
77	35-59 yr	no	other	family	no	- / - -	537	2552	9559
78	$\geq$ 60 yr	no	empl.	-	yes	all	590	4107	1231
79	$\geq$ 60 yr	no	empl.	-	no	all	147	1447	1522
80	$\geq$ 60 yr	no	ret.	single	yes	++ / +	135	1042	1776
81	$\geq$ 60 yr	no	ret.	single	yes	0 / - / - -	635	5561	4261
82	$\geq$ 60 yr	no	ret.	single	no	++ / + / 0 / -	291	3310	5307
83	$\geq$ 60 yr	no	ret.	single	no	- -	495	5591	1702
84	$\geq$ 60 yr	no	ret.	pair/fam.	yes	++ / +	1676	8837	4122
85	$\geq$ 60 yr	no	ret.	pair/fam.	yes	0	1123	6078	3690
86	$\geq$ 60 yr	no	ret.	pair/fam.	yes	- / - -	1346	7365	2859
87	$\geq$ 60 yr	no	ret.	pair/fam.	no	++ / +	638	4559	4652
88	$\geq$ 60 yr	no	ret.	pair/fam.	no	0	623	4277	5308
89	$\geq$ 60 yr	no	ret.	pair/fam.	no	- / - -	887	6051	7799
90	$\geq$ 60 yr	no	other	single	yes	all	619	4709	4340
91	$\geq$ 60 yr	no	other	single	no	++ / + / 0 / -	553	5812	3629
92	$\geq$ 60 yr	no	other	single	no	- -	866	8708	949
93	$\geq$ 60 yr	no	other	pair/fam.	yes	++ / +	495	2173	5463
94	$\geq$ 60 yr	no	other	pair/fam.	yes	0	471	2145	5006
95	$\geq$ 60 yr	no	other	pair/fam.	yes	- / - -	504	2364	3866
96	$\geq$ 60 yr	no	other	pair/fam.	no	++ / +	1462	8699	6464
97	$\geq$ 60 yr	no	other	pair/fam.	no	0	1538	9137	7398
98	$\geq$ 60 yr	no	other	pair/fam.	no	- / - -	1823	10525	10266
total							114364	658543	658590

Table D.1: Case study Southern Overijssel, population cluster scheme (ret. = retired; receiving pension or other retirement benefit, empl. = employed; part-time or full-time, stud. = full-time student; excluding high school)

Appendix E

An algorithm for cleaning up travel diary databases

This appendix presents a rule-based algorithm for filling data-omissions and correcting inconsistencies in observed travel diary databases. Trivial as some of its rules may seem, each of them increased the number of fully consistent respondents when applied to the raw OVG of 1995 in combination with zone sets $\mathcal{H}$, $\mathcal{O}$ and $\mathcal{D}$ for case study Southern Overijssel (see page 118). The algorithm also assures that the origin and destination zones of raw trips obey to the conventions of the dual zone system. This means that a trip reported from a single origin zone i to a distant single destination zone $\tilde{j}$ is recorded as a trip between i and composite destination zone J, where $\tilde{j} \in \mathcal{D}_{S,J}$.

1. In case the home-zone H_r of a respondent r does not appear in $\mathcal{H}$, an alternative home-zone is looked for. This is tried by adding a small number (1, -1, 2, -2 up to a maximum of 5 and -5) to the raw home-zone as stated by the respondent. In case this new home-zone still shares its first two digits with the raw home-zone, and the new home-zone is included in $\mathcal{H}$, it replaces the home-zone as stated by the respondent. This rule is based on most four-digit zones that are nearly identical in their ID to be located in each others vicinity, as long as their associated two-digit zone is identical. For example, zone range 8097-8099 is not located in the vicinity of 8102, but 8100-8107 does.

2. In case $DH_r = 1$ (the respondent states that the first trip of the day left from home), the first zone of departure O_{rt} is set equal to H_r, irrespective of what the respondent has stated for O_{rt}.

3. In case no trips are reported by a respondent, flag indicator NT_r is set equal to 1, irrespective of what has been stated by the respondent for NT_r.

4. Otherwise, NT_r is set equal to 0. In this case, for each of the individual

trips t reported by the respondent, the following additional rules are performed:

(a) Unless it is the first trip, the origin zone O_{rt} is set equal to the destination zone D_{rt} of the previous trip $(t-1)$.

(b) In case the respondent states that the activity undertaken at the destination M_{rt} is 'be at home', the destination zone D_{rt} is set equal to the home-zone H_r.

(c) In case trip $O_{rt} = H_r$, the flag DH_{rt} (trip left from home) is set to 1, unless the previous trip $(t-1)$ already had this flag set to 1.

(d) In case flag $DH_{rt} = 1$, it is determined whether or not the destination zone D_{rt} occurs in $\mathcal{D}$. In case it does, but no travel cost is defined for the origin-destination pair O_{rt} towards D_{rt} (see section 6.4), D_{rt} is set to composite destination zone J where $D_{rt} \in \mathcal{D}_{S,J}$.

(e) In case D_{rt} does *not* occur in $\mathcal{D}$, first it is tried whether the subsequent origin zone $Z_{dep,t+1}$ specifies a valid destination zone of trip t. This of course is not possible if t is the last trip reported by r. If this procedure does not lead to a valid destination zone D_{rt}, then it is tried to synthesise a substituent destination zone as follows:

- In case *no* valid departure time $T_{dep,rt}$ or destination time $T_{arr,rt}$ has been reported for trip t, it is tried to look for a substituent destination zone $T_{arr,rt}$ analogous as to how a substituent home-zone has been looked for in step 1.
- Otherwise, trip travel cost $\hat{\psi}$ is determined as $T_{arr,rt} - T_{dep,rt}$. Subsequently, the following steps are performed:
- All zones j similar to D_{rt} are selected from $\mathcal{D}$. Two zones are similar in case the first two digits of their ID are the same (see step 1). Next, for all these zones, travel cost ψ_{ijm} is determined (see section 6.4, where $i = O_{rt}$). Only zones j are kept for which ψ_{ijm} can be defined.
- Chance factor F_{ij} is determined as the third power of the factor between $\hat{\psi}$ and ψ_{ijm} $(0 \leq F_{ij} \leq 1)$. F_{ij} is raised to the third power, in order to give a disproportional weight to zones with a near match in travel cost from i. In addition, let F_{ij} be weighted by a weigh factor W_j in order to assign more chance mass to larger zones. For case study Southern Overijssel, W_j = number of inhabitants of j.

$$F_{ij} = \begin{cases} W_j \left(\dfrac{\psi_{ijm}}{\hat{\psi}} \right)^3 & \text{where} \quad \hat{\psi} \geq \psi_{ijm} \\ W_j \left(\dfrac{\hat{\psi}}{\psi_{ijm}} \right)^3 & \text{where} \quad \hat{\psi} < \psi_{ijm} \end{cases} \tag{E.1}$$

- A random number is drawn between 0 and $\sum_j F_{ij}$. Now, O_{rt} is assigned the destination zone associated with this draw.

(f) In case the reported departure time $T_{dep,rt}$ is invalid or omitted, but arrival time $T_{arr,rt}$ is not, $T_{dep,rt}$ is estimated as:

$$T_{dep,rt} = T_{arr,rt} - \psi_{ijm} \quad \text{where} \quad i = O_{rt} \quad \text{and} \quad j = D_{rt}. \tag{E.2}$$

(g) In case the reported departure time $T_{dep,rt}$ is equal to or earlier than the arrival time $T_{arr,t-1}$ of the *previous* trip, $T_{dep,rt}$ is set to $T_{arr,t-1}$ + 2 minutes, in order to represent a minimum time for which the respondent stayed at a location.

(h) Similarly as for (4f), in case the arrival time $T_{arr,rt}$ is invalid or omitted, but departure time $T_{dep,rt}$ is not, $T_{arr,rt}$ is estimated as:

$$T_{arr,rt} = T_{dep,rt} + \psi_{ijm} \quad \text{where} \quad i = O_{rt} \quad \text{and} \quad j = D_{rt}. \tag{E.3}$$

(i) In case the arrival time $T_{arr,rt}$ is equal to or earlier than the departure time $T_{dep,rt}$, it is considered invalid, and the above approach is followed.

Appendix F

An algorithm for creating travel cost matrices

This appendix proposes an algorithm to roughly estimate the travel costs associated with an OD-pair. As its input, the algorithm merely relies on the locations of all origin and destination zones, as well as the specification of a possibly basic (highway) network.

Road network First, define the highway or main road network for the area for which travel cost need be estimated from a set of straight links. Define node points where this network can be accessed and left, as well as where junctions or significant curves exist. Apart from this highway network, the area is assumed to be covered by an implicit, uniform, basic road network that connects all physical locations. Where this is explicitly not the case for larger areas, such as mountain ranges, large lakes or sea-arms, the boundaries of these areas should also be modelled as tertiary roads where they exist. Tunnels, bridges or ferry links can be defined where they cross such natural barriers.

Link speeds and travel costs of network links Second, define a road categorisation in which each road is assigned a fixed speed or generalised cost factor per unit of geographic distance. The travel cost for covering each link is now simply computed by dividing its length with the speed or cost factor attributed to its road category.

Access links Apart from the various network links, let two other link types be introduced. First, let access links be defined as synthetic links that connect zone centroids to the road network. For these network connections, let it be stored for each node whether the road network can be accessed or left from this node. As an example, nodes representing a highway crossing generally do not offer the possibility to access or leave the network. Let each origin zone i as well as each destination zone j be attached to a minimum of one and a

maximum of three of such access nodes in different directions. The following algorithm is proposed for this task.

1. First, search for the nearest access node of the road network in any direction around a zone centroid. This node thus found will be the first road network access node of the respective zone.

2. Second, look for the nearest road network access nodes in two opposite directions. For this task, search for the nearest access nodes in angles of 60 to 180 degrees and 180 to 270 degrees, relative to the direction of the first access node.

3. In case an alternative road access node is found in either direction at a maximum of three times the euclidian distance towards the nearest access node, this alternative access node is assumed to constitute a relevant alternative next to the first access node. As such, it is added to the list of alternative access nodes of the associated zone. Each zone is thus assigned a minimum of one, and a maximum of three road network access nodes.

As an example of the above algorithm, figure F.1 illustrates how all zones in and around Amsterdam are possibly connected to the surrounding road network. The figure shows the road network in bold lines, where the thickness of the lines represents link speed. Also, the figure shows road network access nodes as large bright dots, and zone centroids as small black dots. Finally, the thin lines represent the access links as they were found using the above search algorithm. The travel cost of each of these links is now computed as the euclidian distance divided by the speed or cost factor assigned to the access link.

Access link speeds For dual zone systems, it is advised to assume different speeds for access links connecting single destination zones to the road network and for access links connecting composite destination zones. This is because the gravity centre of a composite zone can be located relatively far from a network access point, simply because it is the average of all individual zones. In order to compensate for this unrealistic elongation from the network, a higher access link speed is assumed here.

Direct links Different from access links, let a direct link be defined as the synthetic link connecting each OD-pair as the crow flies, thus, completely ignoring the road network. The links are based on the assumption that at any location some basic form of infrastructure has been realized. However, because the attainable speeds on these links are generally relatively low, such synthetic links should be assigned a relatively low link speed. In the final assignment of travel cost between an OD-pair, the travel cost over the direct link is compared to the shortest network alternative.

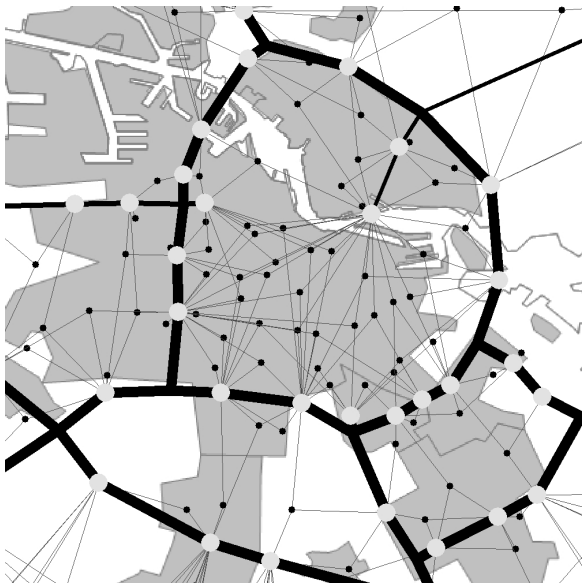

Figure F.1: Access links Amsterdam (example)

Network routes According to the above definitions, a possible travel route over the network is formed by a chain of three separate links. The first link is constituted by the access link from the origin zone i to any of the zone's network access nodes. Analogously, the last link is the egress link between the destination zone j and one of its network access nodes. The intermediate, second link is formed by the shortest network route in between both access nodes. In practical applications, this shortest network route can be computed by means of the Floyd-Warshall algorithm. Because origin zones as well as destination zone each have a maximum of three access nodes, a total of nine network routes is possible for each OD-pair. Out of these nine routes the shortest route is selected as the shortest network route for an origin-destination relation.

Direct travel route Besides the network route alternatives, the direct route follows the direct link and is also considered for the determination of ψ_{ijm}. This is required because of the low detail level of the network compared to that of the zone system. At the same time, because of the relatively low link speed assigned to such direct links, do they hardly compete with the network for OD-pairs that are far apart.

Travel cost assignment Based on the above determined link costs, the generalised travel cost ψ_{ijm} of each OD-pair is determined as the cheapest of all possible network routes and the direct route. This travel cost assignment method is further illustrated by figure F.2. The figure shows a small road

network, two zones and their six access nodes on a road network. Network links are shown as thick lines, access links as thin lines, and the direct link between both zones as a dashed line. Whether or not the connection between an OD-pair is modelled as a network route or a direct route, is determined by the speeds assigned for access links, network links and direct links as well as their geographic lengths.

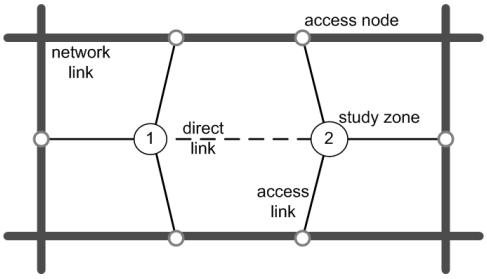

Figure F.2: Shortest route determination (example)

Appendix G

Attractiveness estimates by case study Southern Overijssel

This appendix gives a selection of the attractiveness estimates obtained for case study Southern Overijssel. For this case study, a total of five trip segments was defined (see table 6.3) as well as 90 two-digit postal code zones and 3,766 four-digit postal code zones in $\mathcal{D}$ (see table 6.1). Thus, a total of 5 * (90 + 3,766) = 19,280 attractiveness estimates has been obtained for this case study. Two excerpts of this output data are presented in this appendix. First, the estimates for all two-digit zones are listed by table G.1.

		Attractiveness estimate				
Zone	Population	Work	Education	Shopping	Social	Other
10	641970	973365	1266156	1828825	339199	818203
11	273660	674323	353883	168388	248257	605900
12	163260	108849	87789	162207	158344	140611
13	176270	108708	8011	105100	111337	153284
14	197020	197123	170960	83971	179280	232945
15	157320	56431	27737	36333	44400	35129
16	173740	121994	45244	282167	136504	173477
17	239180	247129	103069	318497	417669	546480
18	143930	183053	60790	84848	171852	130586
19	184030	206570	163370	198783	117324	122563
20	185800	101183	78509	51794	212049	211418
21	182760	162974	139583	68338	146200	168347
22	298350	405178	214881	267000	270896	514985
23	207550	239009	464250	84039	212082	147297
24	124160	72574	30154	37923	90481	82769
25	440710	842469	1334936	1140023	389394	718695
26	254640	232544	722217	224214	187741	165839
27	164420	186552	86803	65913	211034	78706
28	122160	95729	72104	24329	108648	100089
29	255980	148398	32367	64384	137428	49083
30	541760	556162	631808	583881	344460	297436

		Attractiveness estimate				
Zone	Population	Work	Education	Shopping	Social	Other
31	269050	300715	139960	65623	162312	199864
32	280530	158533	69073	131752	228614	152212
33	257930	181218	113804	46230	114881	73792
34	194810	196690	140230	107199	100227	123449
35	232780	564472	1292222	752220	196420	387095
36	96760	78860	38741	42738	73305	124135
37	244480	270718	240396	622720	215967	310855
38	295090	405148	747363	251098	249367	343318
39	218860	207380	269419	291901	202719	334864
40	75080	19932	8753	15626	10431	46076
41	114430	87389	42326	263221	86108	80950
42	108170	76670	31770	22434	72792	21909
43	144300	358479	10429	316446	585225	331366
44	95670	76151	107733	2018	168986	111881
45	107090	9134	1058	4493	288906	68596
46	129990	113517	5788	4864	84031	136623
47	164670	125706	8791	121708	83729	81927
48	231790	150354	212470	167579	207357	149400
49	88340	27170	34027	12132	56774	42268
50	266900	167356	461021	189282	280050	233943
51	148320	68047	171807	440803	39868	304034
52	239520	145935	212131	71992	128878	123354
53	176190	57727	89601	153131	129374	105593
54	202320	262689	24647	341110	151365	176079
55	166060	85690	31861	27564	77397	114738
56	303010	406312	392651	463642	245385	324620
57	191630	123016	20683	74518	73515	74910
58	92810	155014	10153	2659	81638	53078
59	180830	108430	42935	25191	67070	153890
60	218480	373703	93114	11808	129728	158945
61	179220	264571	35541	96295	23370	129007
62	190960	225277	505264	488728	191889	168690
63	106700	158627	2961	14231	187661	5593
64	193840	319494	26615	40168	236710	61059
65	222500	261837	168401	142713	295480	413115
66	169590	107856	76425	183782	62486	112586
67	134760	94286	202431	54456	114845	153874
68	203880	166598	305089	554439	390163	452073
69	162750	58634	5491	408031	114154	87907
70	131580	72617	46432	195572	158776	161896
71	102870	180031	4223	396520	44696	18829
72	115570	86813	31699	351667	310040	144506
73	174480	131256	89092	580888	341703	323200
74	222920	184108	166522	80259	316352	355199
75	299990	266170	510544	99160	185957	236769
76	177740	29585	113643	24682	327606	151362
77	113290	101157	15625	298919	346808	133443
78	119840	73897	12217	3290	255328	147940
79	136820	46720	25279	35006	122243	29522
80	198590	145992	582222	74394	377358	314757
81	109720	33050	36198	60757	66286	34371
82	152280	160439	188588	19901	200926	228001
83	109660	224200	197569	10479	163748	91913
84	106440	16361	12269	42046	237250	238785
85	45460	5211	364	5937	140681	5318
86	45270	2394	1441	1856	19724	862

Zone	Population	Attractiveness estimate				
		Work	Education	Shopping	Social	Other
87	34660	6421	1477	79	36781	16485
88	46300	26175	10940	44312	267226	79900
89	83720	86067	245008	19559	84348	16863
90	63980	1345	341	35	58441	13905
91	43970	5340	4389	788	89411	4834
92	140530	312329	16878	72648	221722	41115
93	56920	7379	14995	16297	11685	8648
94	128980	74790	589501	28095	84204	246665
95	68670	5006	4542	2775	33456	53293
96	151160	36614	2021	11172	66629	22633
97	219160	271891	344865	359298	623368	555794
98	36050	3302	467	550	51155	1712
99	93500	222607	2201	14156	71295	58877
total	15758880	15758640	15759034	15758590	15758780	15758681

Table G.1: Case study Southern Overijssel, attractiveness distribution for two-digit postal code zones

Second, table G.2 lists a sample of attractiveness data for all four-digit postal code zones that together constitute the two-digit zone 25. These zones are all located in and around the Hague. The attractiveness estimates for shopping are graphically illustrated in figure 6.6. The selection is limited to these four-digit zones only, as a full listing would require some ninety densely printed pages.

Zone	Population	Attractiveness estimate				
		Work	Education	Shopping	Social	Other
2511	2410	32499	6183	325970	1900	20513
2512	6210	13434	4530	70037	14230	22018
2513	6860	9943	4980	41937	6131	8035
2514	3530	38324	33800	14340	1377	11867
2515	8420	50301	19266	20612	5121	25343
2516	8680	28600	89975	14437	3305	8739
2517	7280	51512	54310	5809	14730	19241
2518	10580	19254	7284	14016	8690	12372
2521	6240	13352	16682	9686	2492	8879
2522	10440	4092	7229	13645	4591	7130
2523	7630	2365	5270	34944	4179	4134
2524	4780	1760	3245	6509	2441	2170
2525	17970	14120	66885	60484	7417	13745
2526	19170	7252	13734	33313	7931	10574
2531	9150	4662	6423	15308	12190	5274
2532	4150	9546	2958	5620	4439	7050
2533	6070	7340	4435	5188	2509	5804
2541	10670	5920	7486	10619	11255	6279
2542	9210	2789	10938	11279	11721	7089
2543	6990	3270	4648	5447	2702	20067
2544	9570	42735	22754	12460	8783	37891
2545	7040	19446	78696	19205	6280	16403
2547	15510	12448	20871	17149	9498	8350
2548	2440	1560	1544	8626	2473	4056
2551	12670	4418	8104	17184	15690	10045
2552	18710	17728	19954	37213	12986	16235
2553	5190	17616	3251	3750	19112	17021

		Attractiveness estimate				
Zone	Population	Work	Education	Shopping	Social	Other
2554	1400	8230	864	2583	2702	26749
2555	8640	18559	104036	54149	9612	26595
2562	25000	13886	17452	21246	10007	12067
2563	6600	4087	11759	5508	10828	4178
2564	10100	5882	7968	8161	3951	8465
2565	14680	11596	129321	15569	13633	12814
2566	4750	7747	28261	3772	2575	12429
2571	4520	15803	10348	24143	5246	8112
2572	13250	5460	14415	11606	8620	10793
2573	11880	4373	8333	10241	7482	9787
2574	6130	1437	4212	5203	2464	2926
2581	4940	1739	3311	3924	3534	2296
2582	9360	18926	41346	15725	6567	9055
2583	8910	5995	78113	8615	8777	15412
2584	7110	10852	4518	16923	3466	16118
2585	6310	28865	10312	23309	7009	5516
2586	7080	20540	22612	9846	2351	71891
2587	8080	5734	4849	6588	3556	42034
2591	5010	9676	19479	14483	29207	6006
2592	8430	8395	24529	15332	3148	23297
2593	8650	5478	20625	11738	7466	4482
2594	1220	37724	45673	884	5436	7324
2595	6150	64931	157069	4817	5142	9590
2596	5210	48480	23227	3981	6956	9778
2597	9730	41788	16869	6890	15486	24657
total	440710	842469	1334936	1140023	389394	718695

Table G.2: Case study Southern Overijssel, attractiveness distribution for four-digit postal code zones in The Hague

Glossary

activity opportunity = the possibility of performing an activity at some location (see page 53)

attractiveness $A_{jm\xi}$ = the relative amount of activity opportunities offered by destination zone j for a given trip segment m as it is perceived on average by an individual who belongs to population sample ξ (see page 53)

attractiveness decay factor $\phi_{rtj\xi_H}$ = the value of attractiveness decay function $\Phi_{m\xi_H}(\omega)$ at travel cost $\omega = \omega_{rtj}$ for trip segment m as it applies on average for population sub-sample ξ_H (see page 54)

attractiveness decay function $\Phi_{m\xi_H}(\omega)$ = a function providing the relative evaluation of an activity opportunity for trip segment m at a marginal travel cost ω as it is evaluated on average by population sub-sample ξ_H (see page 54)

circle = see cost circle

composite destination zone = destination zone that is part of the rough, aggregated zone layer $\mathcal{D}_C$ in a dual zone system (see page 50)

cost circle = a range of marginal travel cost values that is modelled as one discrete cost category (see page 51)

cost circle scheme = the set of all circles defined for a case study (see page 52)

destination choice y_{rtj} = the choice of an individual person r to travel towards zone j for trip t (see page 38)

destination choice probability P_{ij} = the chance that a person who starts a trip from origin zone i, selects destination zone j as the destination of a trip (see page 101)

destination choice eccentricity E_{rtj} = the ratio of the destination choice selectiveness associated with destination zone j to the destination choice selectiveness of the destination zone that involves the highest marginal travel cost for an individual person r for trip t (see page 40)

destination choice eccentricity observation model = the model presented in chapters 3 and 4 for the data-driven observation of destination choice eccentricity as the product of attractiveness and attractiveness decay as they follow with maximum likelihood from an existing travel diary database *TB* and an associated travel cost matrix *TC* (see page 82)

destination choice eccentricity range $\mathbf{E_{rtj}}$ = the range of destination choice eccentricity values that describe the eccentricity of a destination choice y_{rtj} (see page 81)

destination choice selectiveness σ_{rtj} = the cumulative utility of all destination zones that require a lower marginal travel cost for an individual person r for trip t compared to a destination choice for zone j (see page 39)

destination choice utility U_{rtj} = the relative degree in which destination choice y_{rtj} is preferred by an average person in sample ξ in case he or she faces a choice situation like r faces at the onset of trip t (see page 39)

dual zone system = a zone system in which each geographic area is modelled by two layers of zonal representation, with one layer representing an aggregation of the other layer (see page 49)

eccentricity origin zone = fictitious origin zone of a trip that combines the true origin zone and the residential zone of a trip maker (see page 79)

embedded travel behaviour = travel behaviour pattern that is embedded in a specific geographic setting (see page 41)

foreign trips = trips with both their origin and destination zone within the synthesise area but made by individuals residing outside of the synthesise area (see page 144)

geographic setting $\mathbf{\Gamma}_{rt}$ = a vector representing the generalised travel cost towards all possible destinations j for an individual person r and for trip t, weighted by the attractiveness A_{jm} of these destinations (where $m = M_{rt}$) (see page 38)

generalised travel cost ψ_{ijm} = the sum of all time, money, effort and inconvenience required for a trip from an origin zone i towards a destination zone j and of trip type m, measured in uniform (monetary) terms (see page 37)

household travel survey (HTS) = a survey of randomly selected respondents whose travel behaviour is queried by means of a travel diary (see page 13)

intermediate zone = zone that is located at a given fraction on the shortest route from an origin zone towards a destination zone (see page 79)

interregional train trip = a trip starting in the home zone of the trip-maker which heads to any other zone and for which the train is the main mode of transport (see page 85)

location-independent travel behaviour = travel behaviour pattern that is measured independent of the geographic setting in which it is rooted (see page 41)

marginal travel cost ω_{rtj} = the total travel cost that is associated with destination choice y_{rtj} considering the entire trip chain of individual r starting from trip t onwards, compared to the expected trip chain when destination choice y_{rtj} would not have been made, but r would have stayed home or travelled back home instead (see page 38)

matching = the random selection of an observed respondent r of the same population cluster as a synthesised individual s, in order to estimate the location-independent travel behaviour vector $\mathbf{B}_s^*$ by the observed behaviour $\mathbf{B}_r^*$ (see page 42)

residence space = set of all home zones where respondents or synthesised individual reside (see page 35)

observed residence space = set of all home zones where respondents reside (see page 35)

OD-matrix = a matrix stating the number of trips for each OD-pair(see page 35)

OD-pair = the pairwise combination of an origin and destination zone (see page 35)

population cluster = a group of individuals who share one or more characteristics for a set of travel behaviour determinants, and who show relatively homogeneous travel behaviour $\mathbf{B}$ (see page 37)

Randstad = an highly populated urban congolomeration in the West of the Netherlands that includes the four largest cities of the country Amsterdam, Rotterdam, The Hague and Utrecht (see page 88)

single destination zone = destination zone that is part of the fine, non-aggregated zone layer $\mathcal{D}_S$ in a dual zone system (see page 50)

synthetic population = database of all individuals who reside in a specific area, in which each separate individual is attributed a set of characteristics (see page 33)

synthesise area = geographic area where synthetic travel diary data is to be generated for (for all of its inhabitants, see page 35)

synthesised residence space = set of all home zones of synthetic population (see page 35)

TDD synthesis algorithm = the microsimulation algorithm presented in section 1.5 for the synthesis of travel diary data (see page 24)

travel behaviour determinant = an easily observable, objective variable that is assumed to have a large influence on the travel behaviour $\mathbf{B}_r$ of an individual person r (see page 37)

travel diary data (TDD) = information on the consecutive trips performed by or expected for an individual person during a predefined, fixed time interval, in particular including the origin and destination locations, trip purpose and main transport mode of each trip (see page 12)

trip destination choice = see destination choice

trip destination space = set of all zones in which respondents or synthesised individuals can end a trip (see page 36)

trip origin space = set of all zones in which respondents or synthesised individuals can start a trip (see page 36)

trip type = a segment of trips that is relatively homogeneous in the type of destinations selected for them and the processes by which these are selected (see page 36)

zone = a relatively small geographic area that is modelled to have all of its socioeconomic contents located in its gravity centre (see page 35)

List of abbreviations

ABA = Activity Based Approach (see page 20)

ANN = Artificial Neural Network (see page 21)

CATI = Computer Aided Telephone Interview (see page 13)

FSM = Four-step modelling approach (see page 21)

GIS = Geographic Information System (see page 17)

GPS = Global Positioning System (see page 16)

HTS = Household Travel Survey (see page 12)

KDD = Knowledge Discovery in Databases (see page 22)

KSO = *Koopstroom Onderzoek Oost Nederland* (see page 142)

MON = *Mobiliteitsonderzoek Nederland* (see page 103)

MPO = *Metropolitan Planning Organisations* (see page 24)

NWB = *Nationaal WegenBestand* (see page 128)

NRM = *Nieuw Regionaal Model* (see page 143)

OD = Origin-Destination (-pair/matrix) (see page 35 and 35)

OEI = *Overzicht Effecten Infrastructuur* (see page 10)

OVG = *Onderzoek Verplaatsingsgedrag* (see page 124)

TDD = Travel Diary Data (see page 12)

VMT = Vehicle Miles Travelled (see page 4)

Index of symbols

This list contains all symbols used in this study. For each symbol, the page number is provided where it was first introduced. Expressions printed in italics are separately referenced in the glossary. Greek symbols are included separately at the end of the list.

A_{jm} = *attractiveness* of destination zone j for *trip segment* m (see page 53 and 53)

A_{jmn} = final estimate for the *attractiveness* of destination zone j for *trip segment* m in iteration step n (see page 60)

A^0_{jmn} = initial estimate for the *attractiveness* of destination zone j for *trip segment* m in iteration step n (see page 60)

A^*_{jmn}, A^{**}_{jmn}, A^{***}_{jmn}, A^{****}_{jmn} = first to fourth *attractiveness* estimate (see page 69)

$A_{jm\xi}$ = *attractiveness* of destination zone j for *trip segment* m as observed for population sample ξ (see page 53 and 53)

A_{mn} = average total *attractiveness* (see page 65)

ACA_{cmn} = average circle *attractiveness* (see page 65)

$ACTF_{cm}$ = average *circle* trip frequency (see page 66)

$\mathbf{B}_r$ = *embedded travel behaviour* as stated by the household travel survey of person r (see page 37)

$\mathbf{B}^*_r$ = *location independent travel behaviour* as stated in the household travel survey of person r (see page 41)

B_{rk} = k^{th} aspect of *embedded travel behaviour* as stated in the household travel survey of person r (see page 37)

B^*_{rk} = k^{th} aspect of *location independent travel behaviour* as stated in the household travel survey of person r (see page 41)

$\mathbf{B}^*_{P_r,\mathrm{avg}}$ = *location independent travel behaviour* as stated by the household travel survey of person r (see page 41)

$B^*_{P_r k,\text{avg}}$ = k^{th} aspect of *location independent travel behaviour* as stated by the household travel survey of person r (see page 41)

$c, \tilde{c}$ = travel cost *circle* ($c \in \mathcal{C}$ or $\tilde{c} \in \tilde{\mathcal{C}}$ see page 51)

$\mathcal{C}, \tilde{\mathcal{C}}$ = *circle scheme* (see page 52)

$\mathcal{D}$ = *trip destination space* = set of all *single* and *composite* destination zones ($\mathcal{D} = \mathcal{D}_S + \mathcal{D}_C$, see page 36 and 51)

$D_{A,ht_7}(g)$ = relative chance vector for adults to have age category g, provided they live in household type 7 (see page 182)

$D_{A,ht_8}(g)$ = relative chance vector for adults to have age category g, provided they live in household type 8 (see page 184)

$D_{A,ht_9}(g)$ = relative chance vector for adults to have age category g, provided they live in household type 9 (see page 184)

$D_c(ht, c)$ = relative chance matrix for a household to have c children, provided household type ht (see page 177)

$\mathcal{D}_C$ = set of all *composite destination zones* ($\mathcal{D}_C \subseteq \mathcal{D}$, see page 50)

$\mathbf{D_{C,ht_8}}$ = relative chance vector for children to have age category g, provided they live in household type 8 (see page 183)

$D_{C,ht_9}(g)$ = relative chance vector for children to have age category g, provided they live in household type 9 (see page 183)

$D_{\mathbf{car}}(w)$ = relative chance matrix for a household to possess a car, provided prosperity level w (see page 181)

$D^*_{\mathbf{car}}(g)$ = relative chance matrix for an individual to have a car available, provided age category g (see page 188)

$\mathcal{D}_i$ = set of all destination zones which can be selected from *origin zone* i ($\mathcal{D}_i \subseteq \mathcal{D}$, see page 51)

$D_n(ht, n)$ = relative chance matrix for a household to exist out of n individuals, provided household type ht (see page 176)

D_{rt} = *destination zone* of trip t for individual r (see page 37)

$D_{\mathbf{ret}}(ht)$ = relative chance vector for a household member to be retired, provided household type ht (see page 179)

$D^*_{\mathbf{ret}}(g)$ = relative chance vector for a household member to be retired, provided age category g (see page 184)

$\mathcal{D}_S$ = set of all *single destination zones* ($\mathcal{D}_S \subseteq \mathcal{D}$, see page 50)

$\mathcal{D}_{S,J}$ = set of all *single destination zones* which are included in *composite destination zone* J ($\mathcal{D}_{S,J} \subseteq \mathcal{D}_S$, see page 51).

$D_{\mathbf{stud}}(ht)$ = relative chance vector for a household member to be a student, provided household type ht (see page 181)

$D^*_{\mathbf{stud}}(g)$ = relative chance vector for a household member to be retired, provided age category g (see page 185)

$D_{\mathbf{unempl}}(w)$ = relative chance vector for a household member to be unemployed, provided prosperity level w (see page 181)

E_{rtj} = *destination choice eccentricity* of a destination choice for j by individual person r in trip t (see page 40)

$\mathbf{E_{rtj}}$ = *destination choice eccentricity range* $[E_{rtj,\min}, E_{rtj,\max}>$ of a destination choice for j by individual person r in trip t (see page 81)

$E_{rtj,\min}$ = upper boundary of *destination choice eccentricity range* $\mathbf{E_{rtj}}$ of a destination choice for j by individual person r in trip t (see page 81)

$E_{rtj,\max}$ = upper boundary of *destination choice eccentricity range* $\mathbf{E_{rtj}}$ of a destination choice for j by individual person r in trip t (see page 81)

f = synthetic household (family) ($f \in \mathcal{F}_z$, see page 170)

$\mathcal{F}_z$ = set of all synthetic households which reside in z (see page 170)

F^E_{ijmn} = expected number of trips from zone i towards zone j of *trip segment* m, per 100 respondents residing in i, for iteration step n (see page 68)

F^O_{ijm} = observed number of trips from zone i towards zone j of *trip segment* m, per 100 respondents residing in i (see page 62)

F^O_{im} = total observed number of trips departing from zone i of *trip segment* m (see page 63)

g = age category ($1 \leq g \leq g_{\max}$, see page 182)

G_s = age category of synthetic individual s (see page 171)

h = (base) home zone of a trip-maker ($h \in \mathcal{H}$, see page 35 and 53)

H = aggregated home zone ($H \in \mathcal{H}_C$, see page 53)

$\mathcal{H}$ = *residence space* ($\mathcal{H} = \mathcal{H}_{\mathcal{SP}} + \mathcal{H}_{\mathcal{TB}}$, see page 35)

$\mathcal{H}_C$ = set of all aggregated home zones (see page 53)

$\mathcal{H}_H$ = set of all base home zones that constitute aggregated home zone H (see page 53)

H_r = home zone of respondent r (see page 37)

$\mathcal{H}_{\mathcal{SP}}$ = *synthesised residence space* (see page 35)

$\mathcal{H}_{\mathcal{TB}}$ = *observed residence space* (see page 35)

HEF_{rtj} = *home elongation factor* of destination choice y_{rtj} (see page 79)

ht = household type ($1 \leq ht \leq ht_{\max}$, see page 174)

HT_f = household type of synthetic household f (see page 170)

$HT_{1,z}...HT_{9,z}$ = percentage of households in zone z that is of household type 1 ... 9 (see page 172)

i = origin zone ($i \in \mathcal{O}$, see page 36)

i^*_{rtj} = *eccentricity origin zone* for destination choice context y_{rtj} ($i^*_{rtj} \in \mathcal{O}$, see page 79)

NW = infrastructure network database (see page 34)

j = *single* or *composite* destination zone ($j \in \mathcal{D}$, see page 36 and 51)

$\tilde{j}$ = *single* destination zone ($\tilde{j} \in \mathcal{D}_S$, see page 50)

J = *composite* destination zone ($J \in \mathcal{D}_C$, see page 50)

$\mathbf{k}$ = zone or travel behaviour aspect

K_{rt} = trip purpose of trip t for individual r (see page 37)

L_{rt} = main mode of transport of trip t for individual r (see page 37)

m = *trip segment* ($m \in \mathcal{M}$) see page 36)

$\mathcal{M}$ = trip segmentation scheme (see page 36)

M_{rt} = *trip segment* of trip t for individual r (see page 37)

n = iteration step (see page 60)

$N_{a,f}$ = number of adults in synthetic household f (see page 170)

$N_{c,f}$ = number of children in synthetic household f (see page 170)

$N_{\mathbf{car},f}$ = number of cars in synthetic household f (see page 171)

$N_{\mathcal{D}}$ = number of zones in *trip destination space* (see page 36)

$N_{\mathbf{empl},f}$ = number of employed individuals in synthetic household f (see page 186)

$N_{f,z}$ = number of households who reside in zone z (see page 172)

$N_{\mathcal{O}}$ = number of zones in *trip origin space* (see page 36)

$N_{\mathbf{ret},f}$ = number of retired individuals in synthetic household f (see page 171)

$N_{s,f}$ = number of individuals in synthetic household f ($N_{s,f} = N_{a,f} + N_{c,f}$, see page 176)

$N_{s,z}$ = number of individuals that reside in zone z (see page 172)

$N_{\mathbf{stud},f}$ = number of students in synthetic household f (see page 171)

$N_{t,r}$ = number of trips reported by respondent r (see page 37)

$N_{r,TB}$ = total number of respondents in TB (see page 32).

N_{r,TB_P} = total number of respondents in TB who belong to population cluster P (see page 41)

$N_{\mathbf{unempl},f}$ = number of unemployed individuals in synthetic household f (see page 171)

$N_{z,\mathcal{Z}}$ = number of study zones z in synthesise area $\mathcal{Z}$ (see page 35).

O_{rt} = origin zone of trip t for individual r (see page 37)

$\mathcal{O}$ = *trip origin space* (see page 36)

P_r = *population cluster* of individual r (see page 38)

$\mathcal{P}$ = population cluster scheme (see page 38)

$P_{\mathbf{cars},z}$ = car penetration in zone z per household (see page 172)

P_{ij} = true, unknown destination choice probability (see page 101)

$\bar{P}_{ij}$ = estimated mean of the true, unknown destination choice probability (see page 104)

P^s_{ij} = synthetic destination choice probability (see page 101)

P^o_{ij} = observed destination choice probability (see page 102)

P^{o*}_{ij} = improved observed destination choice probability (see page 102)

P_k = chance by which *destination choice eccentricity* E_{rtk} equals E_{rtj} (see page 84)

$P_{\mathbf{ret},z}$ = percentage of individuals in zone z who are retired (see page 172)

P_{rtj} = chance of an average individual choosing to travel to zone j when confronted with destination choice context y_{rtj} (see page 82)

P^-_{rtj} = chance of an average individual choosing to travel to zones that require a lower marginal travel cost as j when confronted with destination choice context y_{rtj} (see page 82)

P^+_{rtj} = chance of an average individual choosing to travel to zones that require a higher marginal travel cost as j when confronted with destination choice context y_{rtj} (see page 82)

$P_{\mathbf{stud},z}$ = percentage of individuals in zone z who are students (see page 172)

$P_{\mathbf{unempl},z}$ = percentage of individuals in zone z who are unemployed (see page 172)

q = constant percentage ($0\% \leq q \leq 100\%$, see page 79)

$\mathcal{Q}$ = set of fixed percentages q (see page 79)

r = respondent ($1 \leq r \leq N_{r,TB}$, see page 32).

r^2 = goodness-of-fit parameter ($\infty \leq r^2 \leq 1$, see page 146).

s = synthesised individual (depending on the context; $s \in \mathcal{S}_f$ or $1 \leq s \leq N_{s,SP}$, see page 33)

$S_{\mathbf{car},s}$ = car availability of synthesised individual s (see page 171)

$S_{\mathbf{empl},s}$ = employment status for synthesised individual s (see page 171)

$\mathcal{S}_f$ = set of all synthesised individuals who are part of synthetic household f (see page 171)

$S_{\mathbf{ret},s}$ = retired status for synthesised individual s (see page 171)

$S_{\mathbf{stud},s}$ = student status for synthesised individual s (see page 171)

SE = socioeconomic database (see page 34)

SP = *synthetic population* database (see page 33)

t = trip ($0 \leq t \leq N_{t,r}$, see page 37)

$T_{arr,rt}$ = arrival time of trip t by respondent r (see page 124)

$T_{dep,rt}$ = departure time of trip t by respondent r (see page 124)

T^E_{ijm} = expected number of trips of *trip segment* m between origin zone i and destination zone j, based on related *destination choice eccentricity range* estimates (see page 111)

$T^O_{ijm\xi}$ = observed number of trips of *trip segment* m between origin zone i and destination zone j for population sample ξ (see page 52)

TB = travel behaviour database (see page 32)

TC = travel cost matrix (see page 34)

u = function that filters a location-independent travel behaviour vectors from a travel behaviour vector embedded in a geographic setting (see page 41)

u' = inverse function of u(see page 43)

U_{rtj} = *destination choice utility* of selecting destination zone j for an average individual who is confronted with a destination choice context as was respondent r for trip t (see page 39)

v = function that synthesised location-independent travel behaviour (see page 42)

w = prosperity level ($1 \leq w \leq w_{\max}$, see page 181)

W_f = prosperity level of household f ('wealth', see page 170)

$W_{im\xi}$ = observed number of trips of *trip segment* m for population sample ξ that depart from origin zone i (ξ ca be omitted for notational simplicity, see page 52)

$W_{jm\xi}$ = observed number of trips of *trip segment* m for population sample ξ that can possibly head towards destination zone j (ξ ca be omitted for notational simplicity, see page 52)

W_z = average household prosperity level in zone z ('wealth', see page 172)

z = residential zone, the inhabitants of which travel behaviour is synthesised ($z \in \mathcal{Z}$, see page 35)

z_f, z_s = z-value for household f or individual s respectively (see page 175)

$\mathcal{Z}$ = *synthesise area* (see page 35)

ZA_{imn} = total origin zone attractiveness (see page 64)

ZCA_{icmn} = average zone to circle attractiveness (see page 64)

$ZCA_{H\bar{c}m}$ = aggregated zone to circle attractiveness (see page 73)

$ZCAF_{icmn}$ = zone to circle attractiveness fraction (see page 65)

$ZCT^{O}_{H\bar{c}m}$ = zone to circle observed trip total (see page 74)

$ZCTF_{icmn}$ = zone to circle trip factor (see page 67)

$ZCTP_{icmn}$ = zone to circle trip percentage (see page 67)

Greek symbols

$\alpha_{\mathbf{empl}}$ = relative chance increase factor for an individual to have a car available, provided he or she is employed (see page 188)

γ = scale factor in which attractiveness estimates for single destination zones are attributed more weight as such estimates for composite zones (see page 70)

$\mathbf{\Gamma}_{rt}$ = *geographic setting* in which individual r is located at the start of trip t (see page 38)

$\tilde{\mathbf{\Gamma}}_{mn}$ = average *geographic setting* for *trip segment* m in iteration step n (see page 60)

ε_r = vector of intangible travel behaviour determinants for individual person r (see page 42)

ζ_k = overlap of *destination choice eccentricity range* $\mathbf{E_{\hat{r}\hat{t}k}}$ with $\mathbf{E_{rtj}}$ (see page 84)

ξ = population sample (see page 35)

ξ_H = subset of population sample ξ that resides in aggregated home zone H (see page 53)

$\lambda_{\mathbf{empl},g}$ = absolute chance of an individual with age category g to be employed ($0 \leq \lambda_{\mathrm{empl},g} \leq 1$, see page 186)

$\lambda_{\mathbf{lab},g}$ = absolute chance of an individual with age category g belonging to the labour force ($0 \leq \lambda_{\mathrm{lab},g} \leq 1$, see page 186)

σ_{rtj} = *destination choice selectiveness* of zone j in case selected by an average individual who is confronted with a destination choice context as is individual r for trip t (see page 39)

$\Phi_{m\xi_H}(\omega)$ = *attractiveness decay function* for a trip of trip segment m by an individual residing in zone H, as it is observed for population sub-sample ξ_H (see page 54)

$\phi^{O}_{H\tilde{c}m}$ = observed circle trip to attractiveness ratio (see page 74)

$\phi_{rtj\xi_H}$ = *attractiveness decay factor* for trip t by individual person r in case he or she travels to j, as it is observed on average for population sub-sample ξ_H ($H_r \in H$, $\phi_{rtj} = \Phi_{hm}(\omega_{rtj})$, see page 54)

ψ^* = break travel cost for which individual, non-aggregated destination zones are considered (see page 51)

ψ_{ijm} = *generalised travel cost* required for travelling from i to j for trips of *trip segment* m (see page 37)

$\tilde{\psi}_{ijm}$ = *generalised travel cost* required for travelling from i to j for trips of *trip segment* m, where $\tilde{\psi}_{ijm} = \infty$ expresses that j cannot be selected from i as a cause of the dual zone system (see page 51)

ω = *marginal travel cost* (see page 38)

ω_{rtj} = *marginal travel cost* for trip t of individual person r in case he or she would travel to destination zone j (see page 38).

$\omega_{c,\mathrm{avg}}$ = fixed *generalised travel cost* for all trips in *cost circle* c (see page 52)

$\omega_{c,\mathrm{min}}$ = minimum *generalised travel cost* in *cost circle* c (see page 51)

$\omega_{c,\mathrm{max}}$ = maximum *generalised travel cost* in *cost circle* c (see page 51)

Bibliography

AASHTO (1998). Transportation and the economy, American Association of State Highway and Transportation Officials, Washington DC, USA. *http://www.aashto.org/

American Lung Association (1990). The health costs of air pollution: a survey of studies published 1984-1989, Washington DC, USA.

Ampt, E. S. (2003). Respondent burden, *in* P. Stopher and M. L. Lee-Gosselin (eds), *Understanding travel behaviour in an era of change*, Pergamon.

André, M. (1997). Vehicle uses and operating conditions: on-board measurements, *in* P. Stopher and M. L. Lee-Gosselin (eds), *Understanding travel behaviour in an era of change*, Pergamon.

Appleyard, D. (1981). *Livable streets*, University of California Press, Berkeley, USA.

Arentze, T. A., Borgers, A., Hofman, F., Fujii, S., Joh, C., Kikuchi, A., Kitamura, R., Timmermans, H. J. P. and van der Waerden, P. (2001). Rule-based versus utility-maximizing models of activity-travel patterns: A comparison of empirical performance, *in* D. A. Hensher (ed.), *Travel Behaviour Research: The leading edge*, Pergamon, pp. 569–584.

Arentze, T. A., Hofman, F. and Timmermans, H. J. P. (2003). Reinduction of ALBATROSS decision rules with pooled activity-travel diary data and an extended set of land use and cost-related condition states, *Transportation Research Record* **1831**: 230–239.

Arentze, T. A., Hofman, F., van Mourik, F. and Timmermans, H. J. P. (2000). ALBATROSS: multi-agent, rule-based model of activity pattern decisions, *Transportation Research Record* **1706**: 136–144.

Arentze, T. A., Hofman, F., van Mourik, F. and Timmermans, H. J. P. (2002). Spatial transferability of the ALBATROSS model system: Empirical evidence from two case studies, *Transportation Research Record* **1805**: 1–7.

Arentze, T. A. and Timmermans, H. J. P. (2000). *ALBATROSS: A Learning-Based Transportation Oriented Simulation System*, European Institute of Retailing and Services Studies, Eindhoven, the Netherlands.

Arentze, T. A., Timmermans, H. J. P., Hofman, F. and Kalfs, N. (2000). Data needs, data collection, and data quality requirements of activity-based transport demand models, *Transport Surveys: Raising the Standard, Grainau, Germany*, Transportation Research Circular E-C008, Transportation Research Board.

Banister, D. (2002). *Transport planning, second edition*, Spon Press, New York, USA.

Banister, D. (2005). *Unsustainable Transport*, Routledge, Oxfordshire, UK.

Banister, D. and Berechman, J. (2000). *Transport investment and economic development*, Spon Press, New York, USA.

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete choice analysis*, The MIT Press, Cambridge, MA, USA.

Blalock Jr., H. M. (1960). *Social statistics, International student edition*, McGraw-Hill, Ljubljana, Slovenia.

Boarnet, M. G. and Haughwout, A. F. (2002). Do highways matter? Evidence and policy implications of highways' influence on metropolitan development, Brookings Institute. *http://www.brookings.edu/es/urban/boarnetexsum.htm

Bowman, J. L. and Ben-Akiva, M. E. (2000). Activity-based disaggregate travel demand model system with activity schedules, *Transportion Research Part A Policy and Practice* **35**: 1–28.

Brög, W., Erl, E., Meyburg, A. H. and Wermuth, M. J. (1982). Problems of nonreported trips in survey of nonhome activity patterns, *Transportation Research Record* **891**: 1–5.

Bruton, M. J. (1985). *Introduction to Transport Planning*, UCL Press, Hutchinson, UK.

Burchell, R. (1998). The costs of sprawl - Revisited, TCRP Report 39, Transportation Research Board. *http://www.trb.org/

Cambridge Systematics, Inc. (1996). Travel survey manual, *Technical report*, U.S. DOT and U.S. EPA, TMIP Program, Washington DC.

CBS (1995). *Onderzoek Verplaatsingsgedrag 1995 (PRAM-uitgave)*, Centraal Bureau van de Statistiek (CBS), Voorburg, The Netherlands.

CBS (1997a). *Bevolking, Kencijfers 1996*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1997b). *Geboorte, Kerncijfers 1996*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1997c). *Huishoudens, Kerncijfers 1996*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1997d). *Sterfte, Kerncijfers 1996*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1997e). *Voertuigenbezit huishoudens 1996*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1998a). *Huishoudens, Kerncijfers 1997*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1998b). *Samenstelling inkomen van AOW-huishoudens 1997*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (1998c). *VUT; aantal personen 1997*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (2000). *Jeugd 1999, cijfers en feiten*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

CBS (2003). *Historie Arbeid 1899-2002*, Centraal Bureau voor de Statistiek, Voorburg, The Netherlands (in Dutch).

Chapin, F. S. (1974). *Human activity patterns in the city*, John Wiley & Sons, New York.

CPB (2002). Strategic competition with public infrastructure: ineffective and unwelcome?, Dutch Bureau of Economic Policy Analysis. *http://ideas.repec.org/p/cpb/discus/8.html

CPB/NEI (2000). *Evaluatie van infrastructuurprojecten, leidraad voor kosten-batenanalyse*, Centraal Planbureau/Nederlands Economisch Instituut.

Delucchi, M. (1996). Annualized social cost of motor vehicle use in the United States, based on 1990-1991 data, Institute of Transportation Studies, University of California at Davis. summarized in 'Total Cost of Motor-vehicle use', Access (www.uctc.net), 1996, pp. 7-13.

Ettema, D. (1996). *Activity-based travel demand modeling*, PhD thesis, Technische Universiteit Eindhoven, Eindhoven, the Netherlands.

Fogel, R. W. (1964). *Railroads and American Economic Growth: Essays in Economic History*, John Hopkins University Press, Baltimore.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C. (Fall 1992). Knowledge discovery in databases: An overview, *AI Magazine* pp. 213–228.

Gemeente Almelo (2003). Leefbaarheid & Veiligheid Almelo 1997 t/m 2003, *Technical report*, Almelo. (in Dutch).

Gouweleeuw, J., Willenborg, L. and de Wolf, P. P. (1997). Een kennismaking met PRAM voor het OVG, Centraal Bureau voor de Statistiek, Voorburg, the Netherlands. (in Dutch).

Granger Morgan, M. and Henron, M. (1990). *Uncertainty, A guide to dealing with uncertainty in quantative risk and policy analysis*, Cambridge University Press, Cambridde, UK.

Greaves, S. P. (1998). Applications of GIS technology in recent household travel survey methodologies, *Technical report*, Travel Model Improvement Program, Federal Highway Administration.

Greaves, S. P. (2003). *Simulating Household Travel Survey Data in Metropolitan*, PhD thesis, Louisiana State University, Baton Rouge, USA.

Hägerstrand, T. (1970). What about people in regional science?, *Papers of the regional science association* **23**: 7–21.

Hague Consulting Group (1992). Gevolgen van beveiligingsvoorwaarden OVG voor verkeersonderzoek, Leiden, the Netherlands. Report 252-1 (in Dutch).

Hardin, G. (1968). The tragedy of the commons, *Science* **162**(3859): 1243–1248.

Holgate, S. T. (1992). Asthma and our environment, *Medical research council news, December* .

IPCC (2001). *Third Assessment Report*, International Panel on Climate Change, Geneva, Switzerland.

Jacoby, A. (1999). Recent advances in understanding the effects of highway investments on the us economy, *Transportation Quarterly* **53**(3): 27–34.

Janssens, D., Wets, G., De Beuckeleer, E. and Vanhoof, K. (2004). Collecting activity-travel diary data by means of a new computer-assisted data collection tool. Transportation Research Institute, Diepenbeek, Belgium. *http://www.uhasselt.be/dam/Papers/2004_collecting.pdf

Jones, P., Koppelman, F. and Orfeuil, J. P. (1974). *Developments in dynamic and activity-based approaches to travel analysis*, Gower, Aldershot, UK, chapter Activity analysis: State-of-the-art and future directions.

Jones, P. M., Dix, M. C., Clarke, M. I. and Heggie, I. G. (1983). *Understanding travel behaviour*, Gower, Aldershot, UK.

Jong, G., Gunn, H. and Ben-Akiva, M. (2004). A meta-model for passenger and freight transport in europe, *Transport Policy* **11**: 329–344.

Kalfs, N. and Evert, H. (2003). Nonresponse and travel surveys, *in* P. Stopher and P. Jones (eds), *Transport Survey Quality and Innovation*, Pergamon.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied statistics* **29**: 119–127.

Knoflacher, H. (1987). *Verkehrsplanung für den Menschen, Band 1 - Grundstrukturen*, ORAC Verlag.

Knoflacher, H. (1996). *Zur Harmonie von Stadt und Verkehr. Freiheit vom Zwang zum Autofahren*, Böhlau-Verlag.

Litmann, T. (2004a). Evaluating criticism of TDM, TDM Encyclopedia, Victoria Transport Policy Institute. *http://www.vtpi.org/

Litmann, T. (2004b). Evaluating TDM equity, TDM Encyclopedia, Victoria Transport Policy Institute. *http://www.vtpi.org/

Litmann, T. (2004c). Transport and economic development: Impacts on productivity, employment, business activity and investment, TDM Encyclopedia, Victoria Transport Policy Institute. *http://www.vtpi.org/

Litmann, T. (2004d). Transportation costs and benefits, resources for measuring transportation costs and benefits, TDM Encyclopedia, Victoria Transport Policy Institute. *http://www.vtpi.org/

Mannheim, M. L. (1979). *Fundamentals of Transportation System Analysis*, The MIT Press, Cambridge, USA.

McCulloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics* **5**: 115–133.

McNally, M. G. (2000). The activity-based approach, *in* D. A. Hensher and K. J. Button (eds), *Handbook of transport modelling*, Elsevier Science, pp. 53–69.

Mrazek, R. (2002). *Engineers of happy land: Technology and nationalism in a colony*, Princeton University Press, Princeton, USA.

Nadri, M. I. and Mamuneas, T. P. (1996). Contribution of highway capital to industry and national productivity growth, FHWA, USDOT.

Nelson, J., Leitham, S. and McQuaid, R. (1994). Transport and commercial location decisiona: some recent evidence, *Transportation Planning Systems* **2**(4): 41–58.

Neumann, W. (1997). Mobilität und Umwelt. Wieviel Mobilität verkraften wir?, *Mobilitäet oder Freiheit: Gleichung oder Widerspruch?*, Friedrich-Ebert Stiftung, Akademie der Politischen Bildung, Bonn, Germany.

Newell, A. and Simon, H. (1972). *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, USA.

OECD (1991). The state of the environment, Organisation for Economic Co-operation and Development, Paris, France.

Openshaw, S. and Openshaw, C. (1997). *Artificial intelligence in geography*, John Wiley & Sons, Chichester, UK.

Ortúzar, J. and Willumsen, L. G. (1995). *Modelling transport (second edition)*, John Wiley & Sons, Chichester, UK.

Pas, E. I. (1990). Is travel demand analysis and modelling in the doldrums?, *in* P. M. Jones (ed.), *Developments in dynamic and activity-based approaches to travel analysis*, Avebury, Aldershot, UK.

Pas, E. I. and Lee-Gosselin, M. E. H. (1997). The implications of emerging contexts for travel-behaviour research, *in* P. Stopher and M. E. H. Lee-Gosselin (eds), *Understanding travel behaviour in an era of change*, Pergamon, Oxford, UK.

Preston, J. (2001). Integrating transport with socio-economic activity, a research agenda for the new millennium, *Journal of transport geography* **9**: 13–24.

Richardson, A. J., Ampt, E. S. and Meyburg, A. H. (1996). Nonresponse issues in household travel surveys, *Conference on Household travel surveys: new concepts and research needs*, National Academy Press, Washington D.C.

SACTRA (1994). Trunk roads and the generation of traffic, Standing Advisory Committee on Trunk Road Assessment, Department of Environment, Transport and Regions (DETR), London, UK. *http://www.roads.detr.gov.uk/

SACTRA (1999). Transport and the economy, Standing Advisory Committee on Trunk Road Assessment, Department of Environment, Transport and Regions (DETR), London, UK. *http://www.roads.detr.gov.uk/

Schaeffer, K. H. and Sclar, E. (2003). The automobile era: a cultural analysis, *in* A. Root (ed.), *Delivering sustainable transport: a social science perspective*, Pergamon, Amsterdam, the Netherlands, pp. 117–126.

Stopher, P. R., Bullock, P. and Rose, J. (2002). Simulating household travel survey data in australia: Adelaide case study, *25th Australasian Transport Research Forum*, Canberra, Australia.

Stopher, P. R., Greaves, S. P., Kothuri, S. and Bullock, P. (2001). Synthesizing household travel survey data, application to two urban areas, *23rd Conference of Australian Institutes of Transportation Research*, Monash University, Australia.

Supernak, J. (1982). Transportation modelling: lessons from the past and tasks for the future, *PTRC Summer Annual Meeting*.

Teufel, D. (1995). Folgen einer globalen Motorisierung, Heidelberg Umwelt und Prognose Institute, Heidelberg, Germany. (in German).

Tillema, F. (2005). *Development of a data driven Land Use Transport Interaction model*, PhD thesis, University of Twente, Enschede, the Netherlands.

Timmermans, H. J. P. (2000). Models of activity scheduling behaviour, *Stadt, Region, Land*, Vol. Heft 71 of *AMUS; Aachener Kolloquium: Mobilität und Stadt*, RWTH Aachen, pp. 33–47.

Untermann, R. and Vernez Moudon, A. (1989). Street design: Reassessing the safety, sociability and economics of streets, University of Washington, Seattle, USA.

Vasconcellos, E. A. (2001). *Urban transport, environment and equity: the case for developing countries*, Earthscan, London, UK.

Veldhuisen, K. J., Timmermans, H. J. P. and Kapoen, L. L. (2000). RAMBLAS: a regional planning model based on the microsimulation of daily activity travel patterns, *Environment and planning A* **32**: 427–443.

Weisbrod, G. (2000). *Current practices for assessing economic development impacts from transportation*, National Cooperative Highway Research Program Synthesis Report # 290, TRB, Washington, USA.

Wermuth, M. J., Sommer, C. and Kreitz, M. (2003). Impact of new technologies in travel surveys, *in* P. Stopher and P. Jones (eds), *Transport Survey Quality and Innovation*, Pergamon.

Whitelegg, J. (1993). *Transport for a sustainable future: the case for Europe*, John Wiley & Sons, Chicester, UK.

Whitelegg, J. (1997). *Critical Mass. Transport, Environment and Society in the Twenty-first Century*, Pluto Press, London, UK.

WHO (1990). Environment and health - the European charter and commentary, World Health Organisation, Regional Office for Europe, Copenhagen.

Wiemers, E., Ballas, D. and Clarke, G. (2003). A spatial microsimulation model for rural Ireland, evidence from the 2002 Irish Census of Population. Rural Economy Research Centre, Ireland and University of Leeds, UK. *http://www.teagasc.ie/research

Willenborg, L. and van den Hout, A. (2000). Possibilities of PRAM to protect statistical data, *Of Significance...* **2**(1): 65–69.

Wolf, J., Loechl, M., Thompson, M. and Arce, C. (2003). Trip rate analysis in gps-enhanced personal travel surveys, *in* P. Stopher and P. Jones (eds), *Transport Survey Quality and Innovation*, Pergamon.

WRI (1994). World resources 1994-1995, World Resource Institute, Oxford University Press, Oxford, UK.

Wright, E. O. (1992). *Fast wheels, slow traffic - urban transport choices*, Temple University Press, USA.

Summary

Transport has been called the maker and breaker of cities. A maker of cities because of its vital role in the bringing together of people, goods and services. A breaker of cities because of its effects on the quality of the living environment. In balancing both aspects, a detailed cost-benefit analysis of transport policy is imperative. However, compared to the analysis of (external) costs, the analysis of (external) benefits is hampered by the low availability of travel diary data.

Travel diary data reveals the usage that individuals make of the transport system. It shows which people travel from where to where, at what times, for what trip purposes and by what transport modes. If such data were available for a large proportion of the population, a detailed benefit analysis could be implemented. However, due to its high collection cost, travel diary data is only available for small to very small sample fractions of the population. This means that although a household travel survey may well include 100,000 individuals in total, when zooming in on the functioning of specific transport infrastructure, only a handful of relevant observations remains. This report proposes a method to use elements of all unused observations to estimate travel diary data as it is the most likely to have been reported for a selection of interest.

The idea is illustrated by figure S.1. In the left part of the figure, individuals with observed travel diary data are indicated by black crosses. Some of these observations are located within a study area (visualised by a rectangular box) but most observations are located outside. All individuals living inside this study area with unknown behaviour are indicated by a grey cross. In the right part of the picture, it is visualised what this report intends to deliver; a method to synthesise travel diary data for individuals with unknown behaviour.

This data synthesis is pursued by a micro-simulation approach that synthesises travel diary data for each individual, based on the observations from a sufficiently large empirical travel diary database. Innovative to the method proposed is that it allows observations external to a study area to be used for the synthesis of trip destination choice for the study area. This requires a detailed, quantitative modelling of the spatial context where travel diary data is observed in or synthesised for.

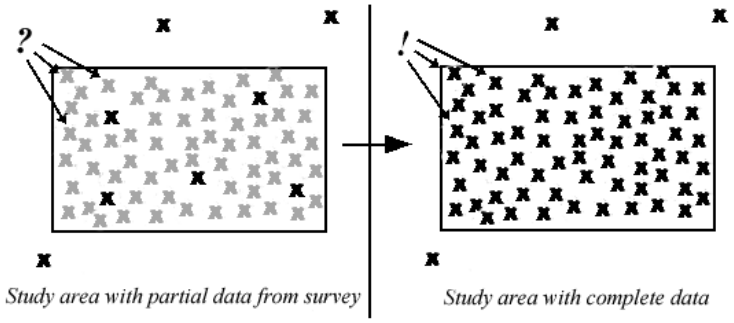

Figure S.1: Synthesis of travel diary data

The accuracy level of synthetic data can never exceed that of empirical data *for individual persons*. This is because all data transformations required to generate such data are inevitably based on assumptions external to the original empirical data. However, for *aggregations* of individual behaviour for sub-selections of the total sample, it is possible for synthetic data to be more accurate than the original empirical data for the same sub-selection. Such data enrichment occurs under three conditions. First, empirical data external to the sub-selection should contain information on behaviour that takes place within the sub-selection. Second, the synthesis method should be successful in the filtering of this information from the external observations and the transferring of this information into the sub-selection. Third, the amount of information thus introduced into the study area should outweigh the natural information loss due to the required data transformations.

This report researches this potential for the synthesis of travel diary data, in particular for the key aspect of trip destination choice. Trip destination choice is judged a discrete choice process where one zone is selected from a set of possible destination zones. In order to obtain transferrable information on this choice process, a measure is looked for that describes destination choice independent of the spatial context in which it is observed. This measure is obtained by an explicit modelling of this spatial context. First, the perceived attractiveness of each possible destination is deducted from the distribution of trips over all destination zones in a sufficiently large, empirical travel diary database. In its operational complexity, this attractiveness observation model is one of the main contributions made by this report. Second, it is observed for various sub-areas how the (marginal) travel costs required to reach all these destinations are perceived on average by local inhabitants as a discount factor on the attractiveness of a destination. Combining both effects, each possible destination choice in each possible spatial context is represented by an amount of utility. Ordering all destination choice utilities by their travel cost from a trip origin location, it is determined how eccentric each destination choice is when observed for each possible spatial context. This degree of eccentricity,

determined relative to the average of locally observed behaviour, is independent of the spatial context in which it is observed and may thus be transferred to other spatial contexts. This allows a micro-simulation approach to be implemented that includes the synthesis of trip destination choice. Travel diary data can now be synthesised for individuals who are expected to behave similar with other, observed individuals, but who have reported their behaviour in a different spatial context.

The report implements three case studies of varying scale and level of complexity as applications of this micro-simulation approach. First, the case study Interregional Train Travel shows how the synthesis method can be applied to estimate the amount of individuals who make long distance train trips in the Netherlands between all 16 major railway stations. Comparing these estimates with the number of trips observed in the input travel diary database and then relating both to an external database that contained the actual figures, it is shown that the synthesis method offered correct estimates for a significant higher percentage of travel relations. Second, in the case study Southern Overijssel, individual travel diary data is synthesised for an area of 658,590 inhabitants. From this database, the number of daily shop trips and the number of daily car trips is determined for any of the 13,456 travel relations within the study area. Comparing this data with two external databases on daily shop trips and daily car trips respectively, it is shown that the synthetic database offered a good reproduction of these external databases, but without requiring local data for the study area.

Based on these results, the case studies show that indeed it is possible to enrich travel diary data for a study area of interest by travel diary data obtained elsewhere. As it was noted above, this data enrichment only occurs for travel behaviour aspects that are not strongly influenced by local factors. Such strong factors can be caused by strong cultural or historical ties between pairs of cities or any other strong local deviation from average conditions in the empirical travel diary database used as input to the data synthesis. However, it was also shown that the strongest and therefore the most relevant of these local factors manifest themselves at a high aggregation level. For such aggregation levels, a significant number of observations is often already available in the input database. Under this condition, the case studies have identified a second main application of the synthesis method: the quantification of intangible, location-specific factors. The net strength of these factors is obtained by comparing the predictions of a synthetic travel diary database (which does not include these location-specific factors) with an empirical travel diary database (which, due to its empirical nature, includes all location-specific factors). Apart from a monitoring of these factors in time, this quantification also enables the generation of more accurate synthetic travel diary data by performing a second synthesis cycle.

It is concluded that the synthesis method proposed by this report serves as

a useful extension to the collection of travel diary data. Provided that travel diary data is collected for a significant sample fraction, the method allows the synthesis of additional travel diary data for small constituent study areas that would otherwise be represented by an insignificant number of observations. Due to its independence of intangible, location-specific factors, the method can also be used to quantitatively assess such location-specific factors. This creates new applications of existing travel diary databases in the evaluation of spatial policy and transport policy. It is hoped that both applications can assist in the further advancement of operational cost-benefit analysis of transport systems and hence, as Mannheim was quoted in the beginning of this report, in the coming to an effective usage of transport to achieve the goals of a society.

Samenvatting

Dutch summary

De huidige verkeerspraktijk besteedt steeds meer aandacht aan het modelleren en monitoren van de uiteenlopende (externe) effecten van het verkeerssysteem. In vergelijking hiermee is de gedetailleerde analyse van het maatschappelijk nut van het functioneren van het verkeerssysteem achtergebeleven. Een belangrijke oorzaak hiervan is de geringe beschikbaarheid van gedetailleerde, individuele verplaatsingsgedragdata voor lokale studiegebieden.

Verplaatsingsgedragdata geeft inzicht in het gebruik van het verkeerssysteem door individuele personen. Het laat onder andere zien welke bevolkingsgroepen gebruik maken van het verkeerssysteem, op welke tijden, met welke vervoerwijzen en voor welke activiteiten. Idealiter zou dergelijke data leidraad moeten bieden bij het opstellen, monitoren en evalueren van verkeersbeleid. Helaas zijn er hoge kosten verbonden aan het verzamelen van zulke data en wordt in Nederland slechts op nationaal niveau een significante waarnemening gemaakt door het Mobiliteitsonderzoek Nederland (MON) en voorheen door het Onderzoek Verplaatsingsgedrag (OVG). Dit rapport stelt een methode voor waarmee zoveel mogelijk informatie uit deze nationale database gebruikt kan worden om uitspraken te doen over het functioneren van onderdelen van het verkeerssysteem op lokaal niveau.

De grondgedachte achter de methode is geïllustreerd in figuur S.1 (zie pagina 244). In de linkerhelft staat een studiegebied gesymboliseerd door een rechthoek. Kruisjes symboliseren de personen woonachtig in dit gebied waarvoor inzicht in het verplaatsingsgedrag gewenst is. Echter, alleen voor de personen met een zwart kruisje is dit gedrag daadwerkelijk geobserveerd, bijvoorbeeld in het kader van het MON. Zoals te zien is betreft het maar een klein aantal personen binnen het studiegebied, maar daartegenover staat dat er ook veel waarnemingen buiten het studiegebied beschikbaar zijn. Ter indicatie zijn drie kruisjes buiten het studiegebied afgebeeld, maar voor praktijksituaties zullen dit meerdere tienduizenden waarnemingen zijn. Dit rapport stelt een methode voor waarmee dergelijke waarnemingen getransformeerd kunnen worden naar vergelijkbaar gedrag binnen het studiegebied. Enerzijds vindt hierdoor een vervuiling van de oorspronkelijke data plaats; de data moet im-

mers bewerkt worden om deze transformatie te kunnen maken. Anderzijds maakt de methode het mogelijk observaties te gebruiken die anders buiten beschouwing waren gebleven, waardoor de statistische significantie van de voor het studiegebied beschikbare waarnemingen wordt vergroot. Voor alle situaties waarin het tweede effect groter is dan het eerste, vindt dataverrijking plaats voor het studiegebied. Het eindresultaat is een (synthetische) schatting van het verplaatsingsgedrag van elk individu in het studiegebied, zoals gevisualiseerd in de rechterhelft van figuur S.1.

De synthetisatie van verplaatsingsgedrag op basis van extern waargenomen gedrag wordt in dit rapport uitgevoerd door een micro-simulatie benadering. Innovatief aan de voorgestelde methode is dat deze expliciet is ontwikkeld voor het genereren van bestemmingskeuzegedrag met mogelijk een gedetailleerd zoneringssysteem. Hiertoe stelt het rapport een methode voor om de geografische omgeving van elke bestemmingskeuze kwantitatief te karakteriseren. Aan de hand hiervan wordt voor elke bestemmingskeuze bepaald hoe excentriek deze is ten opzichte van andere bestemmingskeuzes in een vergelijkbare context. Voor deze kwantitatieve modellering van de geografische context waarbinnen bestemmingskeuzes plaats vinden, wordt gebruik gemaakt van de attractiviteit van elke mogelijke bestemmingszone per trip segment. Dit rapport stelt hiertoe onder andere een uitgebreid, *data-driven* model voor waarmee deze attractiviteiten kunnen worden geschat op basis van het MON en een reiskostenmatrix.

Om de bovengeschetste methode in de praktijk te testen zijn in dit rapport drie voorbeeldstudies uitgewerkt. Als eerste biedt de studie Amsterdam-Rotterdam een triviale toepassing, vooral bedoeld om de complexe modelspecificatie inzichtelijk te maken. Als tweede voorbeeld is een studie over interregionaal treinverkeer opgenomen. In deze studie wordt onder andere aangetoond dat de 1,565 waarnemingen in 2001 van langeafstand-treintrips door geheel Nederland kunnen worden gebruikt om lokale voorspellingen van treingebruik significant te verbeteren. Tot slot, in de studie Zuid-Overijssel wordt het individueel verplaatsingsgedrag voor elk van de 658,590 inwoners in dit gebied gesynthetiseerd, waarbij bestemmingskeuze op het vier-cijferig postcodeniveau wordt bepaald. Middels een vergelijking met twee gedetailleerde externe trip databases wordt hierbij aangetoond dat een significante verrijking heeft plaatsgevonden in de synthetische data ten opzichte van de oorspronkelijke MON-data voor het studiegebied. Hiermee blijkt de gegenereerde synthetische data voor veel verkeerskundige toepassingen van vergelijkbare waarde als in de praktijk gebruikte verkeerskundige databases, zonder dat in de synthetisatie van deze data gebruik is gemaakt van lokaal verkregen data, zoals verkeerstellingen, een gedetailleerd verkeersnetwerk of sociaal-economische data van het voorzieningsniveau in elke zone.

Hiermee is aangetoond dat het voor veel praktijktoepassingen inderdaad mogelijk is lokaal geobserveerd verplaatsingsgedrag te verrijken middels extern

geobserveerd verplaatsingsgedrag. Tevens bleek de in dit rapport ontwikkelde methode diverse betekenisvolle, kwantitatieve output te leveren over de geografische structuur van het gebied waarvoor verplaatsingsgedragdata beschikbaar is. Op verschillende plaatsen in het rapport wordt aangegeven hoe dergelijke output te gebruiken is voor het monitoren van ruimtelijk ordenings- en verkeersbeleid, het identificeren van veelbelovende ontwikkelingslocaties, het kwantificeren van ongrijpbare historische en culturele relaties of de monetarisering van (eigenschappen) van locaties of verbindingen.

Samenvattend wordt geconcludeerd dat de in dit rapport voorgestelde methode een bruikbare aanvulling is op de huidige empirische observatie van verplaatsingsgedrag. Zolang dergelijke data verzameld blijft worden voor een significante steekproef voor geheel Nederland, biedt de methode innovatieve mogelijkheden om deze data additioneel te benutten. Hieronder vallen in elk geval de generatie van gedetailleerd, individueel verplaatsingsgedrag voor lokale studiegebieden en de kwantificatie van de geografische structuur van deze studiegebieden. Deze toepassingen lijken goed inpasbaar binnen een gedetailleerde analyse van het nut dat individuele onderdelen van het verkeerssysteem de samenleving bieden en zouden daarmee kunnen worden ingezet voor het verder ontwikkelen van een integraal ruitemlijke ordenings- en verkeersbeleid.

About the author

Roland Kager was born in Leiden, the Netherlands, on 25 January 1976. He grew up in the village of De Zilk near an extensive park of coastal dunes. When a highway was extended through the area, this made a deep impression. After attending the 'Fioretticollege' in Lisse for secondary education, he studied Civil Engineering and Management at the University of Twente in Enschede, specialising in transport modelling. During his studies, he got interested in international development issues by lectures of the Technology Development Group (TDG). What followed were a 4-month visit to various East-African countries and two half-years in Bandung (Indonesia) for his traineeship and Masters project. For this latter project he proposed a hybrid travel demand model that was adapted for usage in developing cities by giving special modelling attention to the consequences of rapid urban sprawl, high population densities, extreme income inequalities and a low availability of data. He won a grant to present the research at the 9th World Conference on Transport Research (WCTR) in Seoul in 2001. Afterwards, the research constituted the basis of this Ph.D. work, performed from 2001 until 2004 at the University of Twente. Besides his research work, Roland set up various education modules on data usage for transport studies and he computerised a multi-actor simulation game on transport planning processes. Having resided for three years in the German cities of Münster and Mainz, Roland returned to his home country in 2005 to work as a project leader in local transport policy projects for the consulting firm Witteveen+Bos in Deventer. In his free time, Roland enjoys to learn foreign languages and to go outside.

TRAIL Thesis Series

A series of The Netherlands TRAIL Research School for theses on transport, infrastructure and logistics.

Nat, C.G.J.M., van der, *A Knowledge-based Concept Exploration Model for Submarine Design*, T99/1, March 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Westrenen, F.C., van, *The Maritime Pilot at Work: Evaluation and Use of a Time-to-boundary Model of Mental Workload in Human-machine Systems*, T99/2, May 1999, TRAIL Thesis Series, Eburon, The Netherlands

Veenstra, A.W., *Quantitative Analysis of Shipping Markets*, T99/3, April 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Minderhoud, M.M., *Supported Driving: Impacts on Motorway Traffic Flow*, T99/4, July 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Hoogendoorn, S.P., *Multiclass Continuum Modelling of Multilane Traffic Flow*, T99/5, September 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Hoedemaeker, M., *Driving with Intelligent Vehicles: Driving Behaviour with Adaptive Cruise Control and the Acceptance by Individual Drivers*, T99/6, November 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Marchau, V.A.W.J., *Technology Assessment of Automated Vehicle Guidance - Prospects for Automated Driving Implementation*, T2000/1, January 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Subiono, *On Classes of Min-Max-Plus Systems and Their Applications*, T2000/2, June 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Meer, J.R., van, *Operational Control of Internal Transport*, T2000/5, September 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Bliemer, M.C.J., *Analytical Dynamic Traffic Assignment with Interacting User-Classes: Theoretical Advances and Applications using a Variational Inequality Approach*, T2001/1, January 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Muilerman, G.J., *Time-based Logistics: An Analysis of the Relevance, Causes and Impacts*, T2001/2, April 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Roodbergen, K.J., *Layout and Routing Methods for Warehouses*, T2001/3, May 2001, TRAIL Thesis Series, The Netherlands

Willems, J.K.C.A.S., *Bundeling van Infrastructuur, Theoretische en Praktische Waarde van een Ruimtelijk Inrichtingsconcept*, T2001/4, June 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Binsbergen, A.J., van, J.G.S.N. Visser, *Innovation Steps towards Efficient Goods Distribution Systems for Urban Areas*, T2001/5, May 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Rosmuller, N., *Safety Analysis of Transport Corridors*, T2001/6, June 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Schaafsma, A., *Dynamisch Railverkeersmanagement, Besturingsconcept voor Railverkeer op basis van het Lagenmodel Verkeer en Vervoer*, T2001/7, October 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Bockstael-Blok, W., *Chains and Networks in Multimodal Passenger Transport. Exploring a design approach*, T2001/8, December 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Wolters, M.J.J., *The Business of Modularity and the Modularity of Business*, T2002/1, February 2002, TRAIL Thesis Series, The Netherlands

Vis, F.A., *Planning and Control Concepts for Material Handling Systems*, T2002/2, May 2002, TRAIL Thesis Series, The Netherlands

Koppius, O.R., *Information Architecture and Electronic Market Performance*, T2002/3, May 2002, TRAIL Thesis Series, The Netherlands

Veeneman, W.W., *Mind the Gap; Bridging Theories and Practice for the Organisation of Metropolitan Public Transport*, T2002/4, June 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Van Nes, R., *Design of Multimodal Transport Networks: A Hierarchical Approach*, T2002/5, September 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Pol, P.M.J., *A Renaissance of Stations, Railways and Cities, Economic Effects, Development Strategies and Organisational Issues of European High-Speed-Train Stations*, T2002/6, October 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Runhaar, H., *Freight Transport: At Any Price? Effects of Transport Costs on Book and Newspaper Supply Chains in The Netherlands*, T2002/7, December 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Spek, S.C., van der, *Connectors. The Way beyond Transferring*, T2003/1, February 2003, TRAIL Thesis Series, Delft University Press, The Netherlands

Lindeijer, D.G., *Controlling Automated Traffic Agents*, T2003/2, February 2003, TRAIL Thesis Series, Eburon, The Netherlands

Riet, O.A.W.T., van de, *Policy Analysis in Multi-Actor Policy Settings. Navigating Between Negotiated Nonsense and Useless Knowledge*, T2003/3, March 2003, TRAIL Thesis Series, Eburon, The Netherlands

Reeven, P.A., van, *Competition in Scheduled Transport*, T2003/4, April 2003, TRAIL Thesis Series, Eburon, The Netherlands

Peeters, L.W.P., *Cyclic Railway Timetable Optimisation*, T2003/5, June 2003, TRAIL Thesis Series, The Netherlands

Soto Y Koelemeijer, G., *On the Behaviour of Classes of Min-Max-Plus Systems*, T2003/6, September 2003, TRAIL Thesis Series, The Netherlands

Lindveld, Ch.D.R., *Dynamic O-D Matrix Estimation: A Behavioural Approach*, T2003/7, September 2003, TRAIL Thesis Series, Eburon, The Netherlands

Weerdt, de M.M., *Plan Merging in Multi-Agent Systems*, T2003/8, December 2003, TRAIL Thesis Series, The Netherlands

Langen, de P.W, *The Performance of Seaport Clusters*, T2004/1, January 2004, TRAIL Thesis Series, The Netherlands

Hegyi, A., *Model Predictive Control for Integrating Traffic Control Measures*, T2004/2, February 2004, TRAIL Thesis Series, The Netherlands

Lint, van, J.W.C., *Reliable Travel Time Prediction for Freeways*, T2004/3, June 2004, TRAIL Thesis Series, The Netherlands

Tabibi, M., *Design and Control of Automated Truck Traffic at Motorway Ramps*, T2004/4, July 2004, TRAIL Thesis Series, The Netherlands

Verduijn, T. M., *Dynamism in Supply Networks: Actor Switching in a Turbulent Business Environment*, T2004/5, September 2004, TRAIL Thesis Series, The Netherlands

Daamen, W.,*Modelling Passenger Flows in Public Transport Facilities*,T2004/6, September 2004, TRAIL Thesis Series, The Netherlands

Zoeteman, A., *Railway Design and Maintenance from a Life-Cycle Cost Perspective: A Decision-Support Approach*, T2004/7, November 2004, TRAIL Thesis Series, The Netherlands

Bos, D.M., *Changing Seats: A Behavioural Analysis of P&R Use*, T2004/8, November 2004, TRAIL Thesis Series, The Netherlands

Versteegt, C., *Holonic Control For Large Scale Automated Logistic Systems*, T2004/9, December 2004, TRAIL Thesis Series, The Netherlands

Wees, K.A.P.C. van, *Intelligente Voertuigen, Veiligheidsregulering en Aansprakelijkheid. Een Onderzoek naar Juridische Aspecten van Advanced Driver Assistance Systems in het Wegverkeer*, T2004/10, December 2004, TRAIL Thesis Series, The Netherlands

Tampère, C.M.J., *Human-Kinetic Multiclass Traffic Flow Theory and Modelling: With Application to Advanced Driver Assistance Systems in Congestion*, T2004/11, December 2004, TRAIL Thesis Series, The Netherlands

Rooij, R.M., *The Mobile City. The Planning and Design of the Network City from a Mobility Point of View*, T2005/1, February 2005, TRAIL Thesis Series, The Netherlands

Le-Anh, T., *Intelligent Control of Vehicle-Based Internal Transport Systems*, T2005/2, April 2005, TRAIL Thesis Series, The Netherlands

Zuidgeest, M.H.P., *Sustainable Urban Transport Development: A Dynamic Optimisation Approach*, T2005/3, April 2005, TRAIL Thesis Series, The Netherlands

Hoogendoorn-Lanser, S., *Modelling Travel Behaviour in Multimodal Networks*, T2005/4, May 2005, TRAIL Thesis Series, The Netherlands

Dekker, S., *Port Investment - Towards an integrated planning of port capacity*, T2005/5, June 2005, TRAIL Thesis Series, The Netherlands

Koolstra, K., *Transport Infrastructure Slot Allocation*, T2005/6, June 2005, TRAIL Thesis Series, The Netherlands

Vromans, M., *Reliability of Railway Systems*, T2005/7, July 2005, TRAIL Thesis Series, The Netherlands

Oosten, W., *Ruimte voor een democratische rechtsstaat. Geschakelde sturing bij ruimtelijke investeringen*, T2005/8, September 2005, TRAIL Thesis Series, Sociotext, The Netherlands

Le-Duc, T., *Design and control of efficient order picking*, T2005/9, September 2005, TRAIL Thesis Series, The Netherlands

Goverde, R., *Punctuality of Railway Operations and Timetable Stability Analysis*, T2005/10, October 2005, TRAIL Thesis Series, The Netherlands

Kager, R.M., *Design and Implementation of a Method for the Synthesis of Travel Diary Data*, T2005/11, October 2005, TRAIL Thesis Series, The Netherlands