Methods for forecasting in the Danish National Transport model

Jeppe Rich[1]

Allan Steen Hansen

DTU Transport, Bygningstorvet 1, 2800 Kgl. Lyngby, Denmark

Abstract

The present paper is concerned with the forecasting methodology applied in the new Danish national model. The new national model will apply two forecast methods depending on the type of demand model considered. For models which can be estimated on the basis of TU data and is further covered by register data from Statistic Denmark, a prototypical sample enumeration approach will be used. For models, where this is not the case, a matrix model approach will be used. Typically, this will be the case for models where respondents include foreigners. In this case we do not have register data for the respondents and the TU data will only cover the Danish segment. The key to do forecasting based on a prototypical sample enumeration methodology is to apply a population synthesiser, which can forecast the population profile. By combining the population forecast with the micro-survey, it is possible to derive expansion factors which can be used to up-scale the demand model. The "expansion" is used to lifts the TU data base to a representative population level. The paper will first in brief terms discuss the choice of forecast methodology. Hereafter, we will consider the design of the population synthesiser in some details. Finally, we will test the proposed population synthesiser by back-casting.

[1] Corresponding author, email: jr@transport.dtu.dk , phone: +45 45251536, web: www.transport.dtu.dk

1 Introduction

Forecasting represents one of the greatest challenges in statistical modeling. The complexity arises because forecasting implies that not only, should we be able to build a proper model, but the model should also be representative for the population the forecast concern. Even in the baseline year in which the survey is collected, it may difficult to attain a completely representative model. Typically, the survey will be under- or overrepresented in various respects. This can be caused by a number of factors from different interview rates between different socio-groups to sampling errors and biased interview personal. The problem becomes even more difficult when we are to forecast demand. In this context, we need to up-scale the demand model to be representative for a future population, which at the time of the model simulation is unknown.

Historically, in transport modeling, two forecast methods have been applied[2].

- Prototypical sample enumeration (PSE)
- Matrix model forecasting (MM)

The idea of the PSE method is to stay with the micro-data underlying the demand model during the whole demand model process. This is done by expanding the micro-survey to the population level by a set of expansion factors (Daly, 1998). Typically, expansion factors are defined over a number of socio-groups in order to be able to up- and down weight different groups separately.

In the MM method, on the contrary, the micro-survey is only used to estimate the parameters of the demand model. In a second stage (the calculation or simulation stage) the model is re-formulated at a zone level (matrix level), with all inputs aggregated to the matrix level. The MM approach was applied in the OTM model (Vuc and Hansen, 2007) and in the recent TRANSTOOL II model (Rich et al., 2009) and is typically applied in situations where the data foundation is less detailed.

As stated above, the main idea of PSE is to decompose the respondents in the model into different socio-groups. An individual n is said to belong to socio-group $q(n)$ if he/she conforms to the characteristics of that particular socio-group. Say the TU survey has collected s_q individuals within a given socio-group q and that p_q is the number of people within q at the level of the population. The expansion factor e_q, which will bring the survey to a population level, is then given by

(1) $$e_q = p_q / s_q$$

It is noteworthy that the expansion factor is basically a by-product of the prototypical population profile. In other words, if we can forecast the population by mean of a population synthesiser, it will at the same time result in the expansion factors when combined with s_q. This have the immediate benefit that, as we roll the population backwards and forwards in time, we attain new expansion factors. Whereas expansion factors in future years are used for forecasting purposes, expansion factors in historical years can be used to make the survey in those years representative. The process is illustrated in Figure 1.1 below.

[2] Refer to Rich (2009) for an introduction.

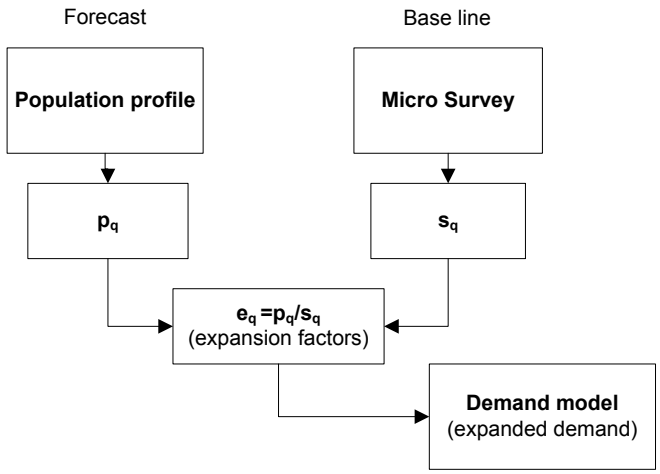

Figure 1.1: Illustration of prototypical sample enumeration forecast methodology.

An issue in the way the expansion factors are derived according to Figure 1.1 is that it requires an aggregation between the population profile and the definition of socio groups. As we will see later in the paper, the population profile will be represented with an overwhelming amount of details (e.g. 9 million cells corresponding to 0.6 individuals per cell), which cannot be used to the full extent as it would "overstretch" the TU data. Hence, the population profile will for the purpose of the expansion factors be aggregated considerably to serve this specific purpose. Still, a detailed population profile is preferable as it is used in a number of other contexts (e.g., construction of matrices). Moreover, it gives a great deal of flexibility in the aggregation.

To exemplify the PSE approach in more details, consider a tour demand matrix T_{idm} for transport mode m and destination d conditional on the residential zone i. In a PSE context, demand could be be derived as

(2) $$T_{idm} = \sum_n P_n(d, m|x_{ni}, z_{dmi}) T_{ni} e_{q(n)}$$

Where $P_n(d, m|x_{ni}, z_{dmi})$ represent the probability model for the demand function with x_{ni} representing exogenous variables related to individuals and the residential zone, z_{dmi} variables related to the zone-system and the choice of mode (typically level-of-service variables), T_{ni} a possible tour generation measure (most likely represented by a discrete choice model), and $e_{q(n)}$ the expansion factor related to individual n belonging to socio-group $q(n)$. By summing over n we attain at the one hand a demand measure that corresponds to the size of the population, and on the other hand a measure that reflects the structure of the demand model[3].

In a MM context it will be simpler in that we do not sum over individuals, but apply only variables that can be expressed at the zone level. That is,

(3) $$T_{idm} = P_i(d, m|x_i, z_{dmi}) T_i$$

[3] It could be that there were different expansion factors for the demand model and the tour generation module, however, for simplicity we assume only one expansion factor.

Here we have dropped the n index as we only consider variables that are aggregated at the zone level.

Theoretically, the PSE method is favourable because it rules out aggregation bias at the zone level. The MM on the other hand will lead to aggregation bias when aggregating to the matrix level. This is because discrete choice probability is non-linear in the input variables, e.g.

(4) $$\Pr\left(\frac{1}{N}\sum_{n} x_n\right) \neq \frac{1}{N}\sum_{n} \Pr\left(x_n\right)$$

The PSE on the other hand, require a micro-data foundation, which can be up-scaled properly, which is not always possible. As the MM only uses aggregated zone data, it need not be backed by micro data.

Although, the MM approach suffers from potential model bias, it generally works quite well as the only aggregation bias will result from bias in the socio-economic dimension of the model and not level-of-service variables. This is because level-of-service variables, which are produced by an external assignment model, are usually represented at the most detailed zone level in either of the two methods. Moreover, the MM method is well established and has been used in the OTM model as well as in the TRANSTOOL model.

2 Overview of forecasting methodology in the Danish National Model

In the case of the Danish National model, demand will be represented by several models as described in (Rich et al, 2010). For the set of models, which cover the travel behaviour of Danes, we can apply TU data and use register data to produce detailed expansion factors. However, for transport carried out by foreigners this is not so. We may well have a usable repeated preference data foundation, which is based on specific RP collections at various border crossings (The Copenhagen air port, Ferries, and foreigners intercepted at the Great Belt) however; we will not have a proper register data base to up-scale these respondents. In practise it means that the New Danish National model will apply a mixture of PSE and MM forecasting depending on the type of demand model as illustrated in Figure 2.1. This does not raise methodological problems as it is only a matter how the final tour matrices are calculated, either by expanding the micro sample or by simulating at the matrix level.

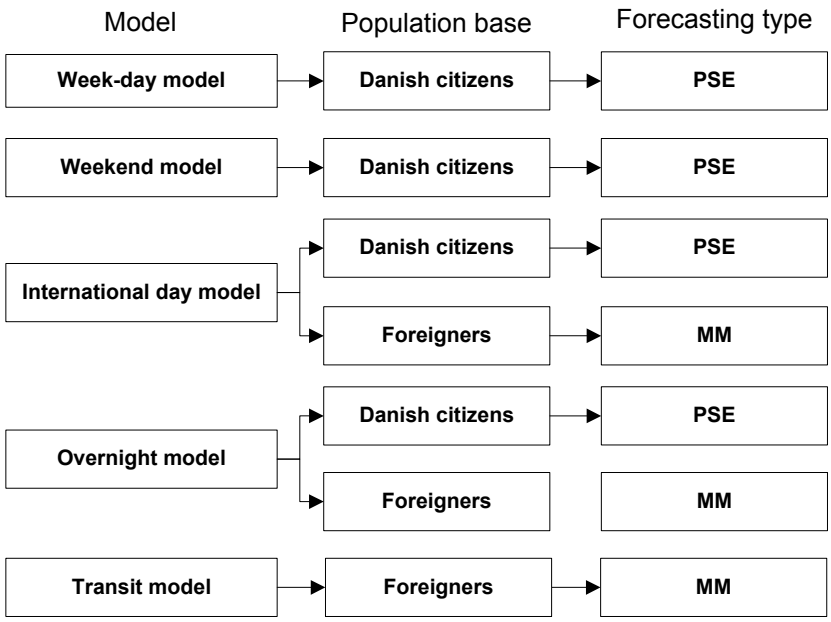

Figure 2.1: Forecasting methodology in the Danish National model.

An alternative to the structure presented in Figure 2.1 is to use only MM forecasting for the international day model and the overnight model.

As the MM approach is well-established in practise, the remainder of the paper will be concerned with the PSE forecast methodology. Moreover, as the PSE approach is basically a function of the population synthesiser the discussion will focus on the methodology of the population synthesiser.

The question is which synthesisers are needed? The answer to this question goes back to the model, which is decomposed into two parts; (i) a strategic model that operates at the household level, and (ii) a demand model that operates at the individual level (Rich et al., 2010) and Rich (2010a). As the two models are separated, answer different questions, and are based on different data, we need to have separate synthesisers for households and individuals. Moreover, it should be recognised that whereas these synthesisers cover the population structure, e.g. the "generation domain" of the model they do not cover the "attraction domain" of the model. In most demand models the attraction domain will be proportional to labour demand in one or more branches. Clearly, in a commuting model this will be the case, however, for shopping and leisure activities it will also be the case as these activities can be measured by the employment intensity in branches that are related to shopping and leisure activities.

If we look closer at the probability model in equation (2), $P_n(d, m|x_{ni}, z_{dmi})$, the exogenous information is divided into variables x_{ni} that define characteristics of model respondents by zone of residence i, and variables that relates to the zone structure z_{dmi}. However, the z_{dmi} variable can be decomposed further. Define $z_{dmi} = c_{dmi} + A_d$, where c_{dmi} is related to the level of level-of-service variables, and A_d is the zone "attraction" variables that relates to the destination zone d.

In most model applications the expansion of the data is carried out for the in the model generation part, e.g. x_{ni}, to ensure that the right number of people demand transport. However, we strongly advocate that the expansion of the employment demand should also be considered, e.g. expansion

or forecast of the A_d component. By also considering an employment demand synthesiser, we avoid that certain users of the model will apply over optimistic employment numbers, which does not conform to the supply base (individuals and households). In other word, the following synthesisers will be applied in the National Model;

- Population synthesiser
- Household synthesiser
- Labour demand synthesiser (firms and public institutions)

The difference between the employment demand synthesiser and the two other synthesisers is the way it is used. The employment synthesiser will be applied to generate forecasts of the A_d variable, whereas the population and household synthesiser will be used to calculate expansion factors $e_{q(n)}$. However, in each case, this is a matter of aggregating the results of the synthesisers to an appropriate expansion measure or labour demand measure.

2.1 Literature on population synthesising

Population synthesizing can be carryout in various ways depending on the data available and the methodology applied. In the Danish National model we will apply a Iterative Proportional Fitting (IPF) approach (Bishop et al.; 1975). The IPF algorithm is used in several different disciplines under different names. In statistics it is often referred to as “bi-proportional fitting”, in economy as the “RAS algorithm”, and in computer science as “matrix scaling”. In the transport community the well-known Furness method is essentially just a different representation of the IPF. Applications of the IPF algorithm in relation to transport have been presented by Beckman et al. (1996) and Arentze et al. (2007), even though these contributions concerned small-scale applications and marginal targets that are not cross-linked. Lee (2007) considered several issues relevant to this study by discussing the issue of making targets internally consistent and harmonized.

Another approach is the maximum entropy approach, in which the matrix estimation problem is formulated as a non-linear mathematical program (MP) with linear constraints. The IPF algorithm and the entropy maximisation approach are different representations of the same problem. Given that entries can be assumed to be Poisson distributed, which under constraints reduces to multinomial or product multinomial entries (Dobson, 1990), both the IPF and the entropy approach will render maximum likelihood estimates for the matrix entries. There is a strong incitement of using IPF because of its computational efficiency, since even very large problems can be solved fast. The downside of the IPF, however, is that it may be difficult to construct a consistent and feasible set of constraints. The entropy approach on the other hand provides an easy way of defining constraints as part of a non-linear MP. The downside is that it is computationally infeasible for large scale problems as considered in this paper.

The IPF have another very important feature, namely that different solutions are quite similar because the structure of the initial solution is preserved. This is especially important when considering population forecasts because the development of the population happens smoothly and with small changes from year to year. An issue, which has been raised in the literature (Daly, 1998) is the existence and preservation of structural zeros. In other words, if zeros exist in the initial solution then zeros will be preserved in the final solution. This however, may also be seen as a practical feature because usually there are entries which should be defined as zeros. For instance, the

ownership of cars may only apply to adults. However, zero's in the initial solution may not always be "strictly" structural, e.g. per definition zero. A relevant example is that an aging population will tend to populate socio-groups that were not represented in the initial solution. From an IPF perspective, there are only one solution to this challenge, namely to alter the initial solution to also represent possible new socio-groups. We will discuss this in Section 3.4.1.

Daly (1998) introduced a quadratic optimisation approach (QUAD) in which expansion factors were estimated based on a sum of squared deviations between targets and expansion factors multiplied by the survey proportion for a respective socio-class. Later in Fosgerau and Jørgen-Jordal (1998), Rich and Kveiborg (1998) and Rich (2002) a modified objective function were investigated. As the QUAD approach and the IPF serve somewhat similar purposes it is a question whether we should use one or the other. In our perspective this question should be answered by looking at the data. The fact that we in this particular situation have access to a very reliable initial matrix at the most detailed level, makes the IPF appealing as it replicates "structure" from the initial matrix. If however, as it is often the case in transport modelling, data are more uncertain, the QUAD approach could be better. In general terms, the QUAD seems to be strongest when the data foundation is weak, whereas, the IPF is the more natural approach when the underlying data is good.

3 The synthesiser methodology

The methodology of the synthesisers involves many mathematical details, which is beyond the scope of the present paper. As a result, the focus of the paper will be to outline the principles rather than going over all of the technical details.

3.1 The IPF algorithm

The basic idea of the IPF is to interpret the population matrix as a hypercube, which is fitted based on two inputs sources: (i) an initial matrix that defines structure or correlation between the different dimensions, and (ii) information about margins or combinations of margins. The strength of this partnership is that the "structure" of the starting matrix is preserved as are the restrictions provided by the margins. A simple 2-dimensional representation of the IPF algorithm is illustrated below.

Iterative proportional fitting algorithm

Step 1: Set $k = 0$ and set $t_{ij}^{k} = t_{ij}^{init}$ where t_{ij}^{init} represents the initial solution.

Step 2: Iterate equation (5) and (6).

(5) $$t_{ij}^{k+1} = \frac{t_{ij}^{k}}{\sum_j t_{ij}^{k}} O_i$$

(6) $$t_{ij}^{k+2} = \frac{t_{ij}^{k+1}}{\sum_i t_{ij}^{k+1}} D_j$$

Step 3: If $|t_{ij}^{K} - t_{ij}^{K+1}| > \varepsilon$ set $k = k + 1$ and go to Step 2. Otherwise stop.

In the illustrated 2-dimensional algorithm O_i define one set of target values (in an origin-destination matrix context this would be row totals) and D_j define another set of target values (column totals).

In this simple 2-dimensional example the IPF applies only first-order targets in the sense that the variable that is included in one target in not included in other targets. However, if more information is available in terms of higher-order interaction terms we need to utilities this information by allowing higher-order constraints. The introduction of higher-order constraints cause a problem for how we define targets, which we shall briefly consider in Section 3.3.

3.2 The master tables

The National Model will consider three synthesizers; a population synthesiser, a household synthesiser, and a labour demand synthesiser. The result of the synthesizers will be represented by a so-called "master table", which basically defines the population profile. The population master table is presented in Table 3.1 below. The dimensions of the table are made up of 2,640 socio-groups ($2 \times 10 \times 2 \times 6 \times 11$), which is then combined with different zones systems to give more or less detailed population matrices. As the model cover all zone systems, the most detailed syntheses of the population occurs when the 2,640 socio-groups are combined with the most detailed zone structure consisting of 3,640 zones. As a result, the synthesiser will generate a matrix with more than 9 million entries in the completely spanned master matrix with less than 0.6 individuals per cell.

Type	Categories	Comment	Index reference
Residential zone	98	L0 zone system	L_0
	176	L1 zone system	L_1
	907	L2 zone system	L_2
	3,640	L3 zone system	L_3
Children	2		c
Age group	10		a
Gender	2		g
Labour market association	6		l
Personal income	11		i
Cell combinations	2,640		

Table 3.1: Attributes and dimensionality of master table for individuals.

The grouping of the socio-economic matrix has been based on different criteria's. Firstly, it should only include information that is exogenous to the model. Although car ownership may be relevant for the travel demand, it is something that is endogenously determined in the model and not part of the population profiling. Secondly, it should capture as much demand variation as possible. Thirdly, it should resemble a grouping which enables us to use official forecasts from Denmark's Statistics. Fourthly, because of the potential size of the problem (primarily caused by the many zones), we should not use an overwhelming amount of dimensions.

The master tables for the household synthesiser and the employment synthesiser is shown below in Table 3.2 and Table 3.3.

Type	Categories	Comment	Index reference
Residential zone	98 176 907 3,640	L0 zone system L1 zone system L2 zone system L3 zone system	L_0 L_1 L_2 L_3
Number of adults Children Labour market association A Labour market association B Household income	3 3 6 6 11		n c l_A l_B i
Cell combinations	3,569		

Table 3.2: Attributes and dimensionality of master table for households.

Type	Categories	Comment	Index reference
Work zone	99 175 891 3,459	L0 zone system L1 zone system L2 zone system L3 zone system	L_0 L_1 L_2 L_3
Branch Highest education	111 9		b e
Cell combinations	999		

Table 3.3: Attributes and dimensionality of the employment demand table.

3.3 The target tables

The target tables represent the future restrictions imposed on the population. The first step when creating a target vector is to identify which targets and combinations of targets to include in the fitting. Obviously, if there are many targets the solution will be strongly restricted and be more precise given that the targets are correct. On the other hand, the more detailed the targets the more uncertain they are in a forecast perspective.

Another issue has to do with the users of the model. Some users will need to do regional forecasting and therefore need a relative detailed geographical classification. Other users will mainly be interested in aggregated nation-wide demand measures and will not need details at a regional level.

Target constraint ID	Variable combination	Notation	Dimensions
TP_{A1}	Age×Gender	$TP_{A1}(a,g)$	20 (10×2)
TP_{A2}	Age×Income	$TP_{A2}(a,i)$	110 (10×11)
TP_{A3}	Age×Lma	$TP_{A3}(a,l)$	60 (10×6)
TP_{A4}	Age×Children	$TP_{A4}(a,c)$	20 (10×2)
TP_{A5}	Income×Lma	$TP_{A5}(i,l)$	66 (11×6)
TP_{B1}	Age×L0	$TP_{B1}(a,L_0)$	980 (10×98)
TP_{B2}	Income×L0	$TP_{B2}(i,L_0)$	1078 (11×98)
TP_{B3}	Lma×L0	$TP_{B3}(l,L_0)$	588 (6×98)
TP_{B4}	Children×L0	$TP_{B4}(c,L_0)$	196 (2×98)
TP_{C1}	L1	$TP_{C1}(L_1)$	176
TP_{D1}	L2	$TP_{D1}(L_2)$	907
TP_{E1}	L3	$TP_{E1}(L_3)$	3670

Table 3.4: Targets applied in the population generator for individuals.

In Table 3.4 we firstly define aggregate socio-economic targets given by target TP_{A1}-TP_{A5}. These targets define the overall national socio-economic profile of the population. Target TP_{B1}-TP_{B4} combines various socio-economic attributes with a L_0 zone level (municipalities). These targets will benefit from a range of official forecasts at the municipality level. Target TP_{C1}, TP_{D1} and TP_{E1} represent only the population at the L_1, L_2 and L_3 level and does not include additional socio-economic information. The latter targets are relevant when considering regional projects.

A general problem is to ensure consistency between the many different targets, many of which may be cross-liked as for TP_{A1}-TP_{A5} where age and income enter two or more constraints. To deal with this consistency problem, an ordering of the different targets is required. The ordering will be used in a more general harmonisation process of the whole set of targets.

The harmonisation process (refer to Rich, 2010) is carried out by defining a ranking of the targets so that higher order targets define the absolute level of lower level targets. There are two objectives of the harmonisation process. Firstly, it is a tool to the users, which will ensure consistency according to the ranking scheme imposed. If users edit many different targets restrictions it can be quite a challenge to ensure that the final set of targets are all completely consistent. Secondly, it is needed as a pre-processing step to a linear-programming algorithm that solves a more general consistency problem in the target vector. If targets are not completely "harmonised" prior to the LP, the LP will fail to produce a feasible solution.

Below in Table 3.5 and Table 3.6, the target definitions for the household synthesiser and the employment synthesiser are shown.

Target constraint block	Variable combination	Notation	Dimensions
TH_{A1}	Income×Adults	$TH_{A1}(i,d)$	33
TH_{A2}	Income×Children	$TH_{A2}(i,c)$	33
TH_{A3}	Income×Lma(A)×Lma(B)	$TH_{A3}(i,l_A,l_B)$	396
TH_{B1}	Income×L0	$TH_{B1}(i,L_0)$	1078
TH_{B2}	Adults×L0	$TH_{B2}(d,L_0)$	294
TH_{B3}	Children×L0	$TH_{B3}(c,L_0)$	294
TH_{B4}	Lma(A)×Lma(B)×L0	$TH_{B4}(l_A,l_B,L_0)$	3528
TH_{C1}	L1	$TH_{C1}(L_1)$	176
TH_{D1}	L2	$TH_{D1}(L_2)$	907
TH_{E1}	L3	$TH_{E1}(L_3)$	3640

Table 3.5: Targets applied in the population generator for households.

Target constraint ID	Variable combination	Notation	Dimensions
TE_{A1}	Branch11	$TE_{A1}(b_1)$	11
TE_{A2}	Branch27	$TE_{A2}(b_2)$	27
TE_{A3}	Branch111	$TE_{A3}(b_3)$	111
TE_{B1}	Branch11×Education	$TE_{B2}(b_1,e)$	88
TE_{C1}	Branch11×L0	$TE_{C1}(b_1,L_0)$	1078
TE_{C2}	Branch27×L0	$TE_{C2}(b_2,L_0)$	2646
TE_{C3}	Branch111×L0	$TE_{C3}(b_3,L_0)$	10878
TE_{C4}	Education×L0	$TE_{C4}(e,L_0)$	784
TE_{D1}	L_1	$TE_{D1}(L_1)$	176
TE_{E1}	L_2	$TE_{E1}(L_2)$	907
TE_{F1}	L_3	$TE_{F1}(L_3)$	3640

Table 3.6: Targets applied in the labor demand generator.

As seen in Table 3.4 targets are cross-linked, e.g. the age variable enters several targets. This causes a problem for how we can obtain consistent targets. Consider a simple problem with three simple first-order targets represented by $T_1(a), T_2(i)$ and $T_3(l)$. A consistent target vector $T_q = T_{a,i,l}$ can then be derived as the product of marginal probabilities. The marginal probabilities is given by $Pr(a) = \frac{T_1(a)}{\sum_a T_1(a)}, Pr\ (i) = \frac{T_2(i)}{\sum_i T_2(i)}$ and $Pr\ (l) = \frac{T_3(l)}{\sum_l T_3(l)}$ and a consistent target vectors would be

(7) $$T(a,i,l) = \left(\sum\nolimits_a T_1(a)\right) Pr(a)\ Pr(i) Pr(l)$$

It is easily to see that the target vector in (7) fulfils all constraints if they are internally consistent (this will be ensured by the harmonization process). If however, the targets are cross-linked in the sense that one attribute enter several targets the target cannot be measured as a product of marginal probabilities. Consider instead a set the targets consisting of $T_1(a,g)$ and $T_2(a,i)$ and let $Pr\ (a,g,i)$ define the marginal probability of age, gender, and income, then

(8) $$Pr\ (a,g,i) \neq Pr(a,g) \times Pr(a,i)$$

In fact, the product $Pr(a,g) \times Pr(a,i)$ will not even be a probability.

It is therefore not simple to create a consistent target for the problem represented by Table 3.1 and Table 3.4. However, a general method has been proposed in Rich (2010). This involves running a linear mathematical program, including all constraints and an objective function that guide the target solution to the most likely representation of the targets initial solution.

3.4 The initial solution

The initial solution describes the correlation structure of the population matrix. If all dimensions are statistically independent, the initial matrix is simply the product of the marginal probabilities. However, this is clearly far from the case in that almost all dimensions of the problem represented by Table 3.1 are more or less correlated, e.g. age is strongly correlated with income.

A practical problem with the initial solution is that the complete span represents more than 9 million entries as we saw in Table 3.1. In other words, there is an average sample rate of 0.6 individuals per entry. This creates a confidentiality issue because single individuals and households can be identified from the cross between socio-economic attributes and the zone system. As the model are to be build and operated outside Denmark Statistics, we cannot rely on the exact initial vector.

To cope with this problem, we will define an initial solution, which is based on a random-sampled version of the true initial solution. The sampled version can then be generated in the protected DST environment and brought outside. The sampled initial solution will be fairly precise and in particular for large socio-groups where sampling is not needed.

3.4.1 Modifying the initial solution

When applying the IPF the normal premise is to stay with the initial solution and change the targets to conform to a future population. However, it could be argued that if we have additional information about a changing population structure this should be included in the IPF by changing the starting values. Two examples are particular relevant.

- When new cities emerge on locations where no people has been living before
- The aging effect

An example of the first example would be the Ørestad city expansion, however, the problem exist in many municipalities as well where certain parts are defines as “development areas” for firms as well as households.

The “aging effect” is a more general problem, which has to do with the fact that people in their seventies today is quite different from people in the seventies 20 years ago. If we assume this trend to continue, then people in 2030 will be different that they are today and this may cause us to under represent certain groups of individuals. However, in the present situation, we believe the problem is limited. As we represent the complete 5.4 million of individuals in Denmark, we should have a broad range of socio-groups represented. If certain groups are no represented, it is unlikely that these groups will have major impact on a medium range forecast horizon of 20-40 years.

Even so, we propose that it should be possible for users to alter the initial solution in order to investigate local issues that cannot be controlled in the target specification where the socio-economy is decoupled from the L2 and L3 zone level. On the other hand, only a rather limited and controlled editing should be allowed. More specifically, we suggest user can edit the structure represented by age, labour market association, and zone L3 $\{a, l, L_3\}$. The edited matrix could then be given by $t^{edit}_{\{a,l,L_3\}}$ and the final initial matrix $\hat{t}^{init}_{\tilde{q}}$ given by

(9)
$$\hat{t}_{\tilde{q}}^{init} = \frac{t_{\{a,l,L_3\}}^{edit}}{\sum_{a,l,L_3} t_{\{a,l,L_3\}}^{edit}} \sum_{g,i,f} t_{\tilde{q}}^{init}$$

The precise model for how $t_{\{a,l,L_3\}}^{edit}$ can be constructed is not discussed further here, however, it should take into account the current intial matrix as well as the expected development within $\{a, l, L_3\}$.

Another suggestion, which may reduce the confidentiality problem of the initial solution, is to work with an "average" starting solution rather than working with a specific baseline year. The problem of looking at only one year is that occasional deviations from the general trend will be introduce "noise" in the general forecast. A better idea could be to work with a moving average of solutions.

3.5 Description of the population generator

Although, we have left out many technical details, we will below describe the stepwise process for how the three generators are modelled.

Step 1: Carry out a harmonisation process of all socio-economic targets, e.g. only TP_{A1} through TP_{B4} for the population synthesiser represented by Table 3.1 and Table 3.4 (the most detailed zone targets are not included at this stage, e.g. TP_{C1} through TP_{E1}).

Step 2: Based on the harmonised targets from Step 1 calculate a consistent target vector based on a linear programming formulation (Refer to Rich, 2010a).

Step 3: Define the initial vector to be used.

Step 4: Run an IPF based on the target vector from Step 2 and the initial vector from Step 3.

Step 5: Based on the IPF solution from Step 4, calculate a new complete target vector for all dimensions including the detailed zone targets, e.g. TP_{C1} through TP_{E1} for the population synthesiser (refer to Rich, 2010a).

Step 6: Process the final IPF based on 5) and 3).

Although the stepwise process may seem complicated, it is rather efficient and will process a complete run of the population synthesiser in about 2 minutes. The whole process has been programmed in a SAS Environment and applies special functionality from the SAS/IML language and the Proc OPTMOD procedure.

4 Validation of the population synthesizer

To assess the forecasting ability of the population synthesizer, we have carried out a simple test in which year 2000 represents the target year and all other years represents baseline years. In other words, we try to predict the population structure in year 2000 based on initial solutions from 1994, 1996 and so forth. Clearly, as can be seen in Figure 4.1, the percent deviation is 0 in year 2000.

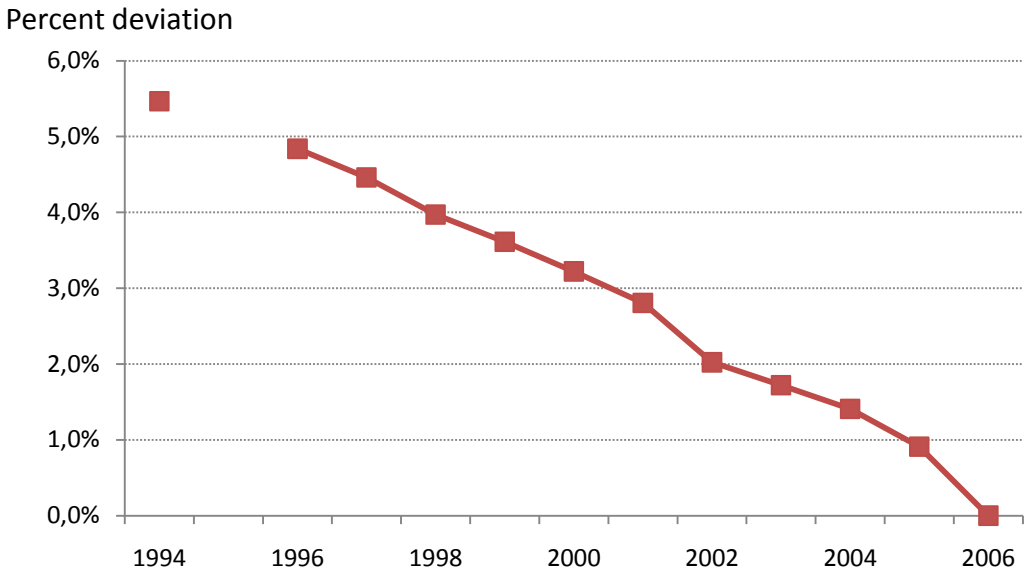

Figure 4.1: Forecast accuracy in the population synthesizer measured in terms of percent deviation. Year 2006 is target year and other years are baseline years.

The deviation is calculated by measuring the percentage deviation for each socio-group in the matrix and subsequently calculate a weighted sum (weighted by the size of the socio-group) of these.

It is important to stress that the experiment in Figure 4.1 is constructed with "correct targets" in the sense that the IPF is processed with the correct year 2006 targets and not just a forecast as would be necessary in reality. Hence, there is an uncertainty in the target specification, which is not included in the above deviation. The deviation therefore results only from divergence between the intial matrix in, say 1994, and the correct population matrix in 2006. Interestingly enough, is seems as if the deviation is a linear function of the length of the forecast period.

5 Conclusion

The paper presents the forecast methodology applied in the new Danish National Transport model.

The first issue we consider is the choice of forecasting methodology, which in turn depends on the available data foundation. In the National Transport model two forecasting methodologies will be applied. The primary forecast strategy will be to use a prototypical sample enumeration approach where the demand models can be based on TU data and register data information about the respondents. This applies to all segments where transport is carried out by Danish Citizens. In case the models cover demand of foreigners, as is the case for the international day model and the overnight model, we cannot apply this approach as we do not have proper register data. In this case we will apply a matrix modeling strategy, which has been used in OTM and Transtools II.

The paper then focus on the forecast methodology of the prototypical sample enumeration approach. It is described that this approach is essentially parallel to create a population synthesizer, as this will allow us to derive expansion factors that measure the profile of future populations. It is

pointed out that the Danish national model will rely on three synthesizers; a population synthesizer that represents individuals, a household synthesizer that represents households, and finally an employment demand synthesizer that synthesize the employment profile as represented by firms and public institutions. The latter is needed in order to avoid that various users of the model applies over-optimistic employment forecasts.

The structure of the population synthesizer is described in some details and the "master tables" of the three synthesizers are outline in order to describe the complete dimensionality of the population tables. The iterative proportional fitting methodology is briefly discussed, including a discussion of target generation, and the role of the initial solution.

In a final section, we provide a simple validation check of the precision of the population synthesizer, by using 2006 as forecast year, and prior years (from 1994 to 2005) as input years. Results indicate that the precision of the model is a linear function of the forecast period.

6 Literature

Arentze, T., Timmermans, H., Hofman, F. (2007), Creating Synthetic Household Populations - Problems and Approach, *Transportation Research Record*, No. 2014, pp.85-91, DOI:10.3141/2014-11.

Beckman, J.R., Baggerly, K.A., McKay, M.D. (1996), Creating Synthetic Baseline Populations, *Transportation Research Part A* **30**(6), pp.415-429.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975), *Discrete Multivariate Analysis – Theory and Practise*, MIT Press.

Daly, A. (1998), Prototypical Sample Enumeration as a basis for forecasting with disaggregate models. PTRC Proceedings (ed) *Transport Planning Methods*, Volume 1 (Seminar D), p.225-236.

Dobson, J.A. (1990), An Introduction to Generalized Linear Models, Chapman and Hall.

Fosgerau, M., Jordal-Jørgensen, J. (1998), PETRA: Weights, PETRA working paper no.3, COWI, 1998.

Lee, A. (2007), Generating Synthetic Unit-Record Data From Published Marginal Tables, Department of Statistics, University of Auckland, 103 pages.

Rich, J. (2010a), Population and Workplace Synthesiser, DTU Transport, Internal Report, 2010.

Rich, J. (2010b), The new Danish national passenger transport model, To be presented at Trafikdage, August 23-24 2010, Aalborg, Denmark.

Rich, J. (2002), Prototypical Sample Enumeration, Appendix C in PhD. thesis, Technical University of Denmark, 2002, Report 2002-1.

Rich J., Nielsen, O.A. (2001): A micro-economic model for car ownership, residential location and work location, PTRC proceedings 2001, Technical Innovations.

Rich J., Bröcker, J., Hansen, C.O., Korchenewych , A., Nielsen, O.A., Vuk, G. (2009): Report on Scenario, Traffic Forecast and Analysis of Traffic on the TEN-T, taking into Consideration the External Dimension of the Union – Trans-Tools Version 2; Model and Data Improvements, Funded by DG TREN, Copenhagen, Denmark.

Rich J. (2009): Introduction to Transport Models – Application with SAS Software, Lulu Press, Ed.5.05, 327 pages.

Rich, J., Aagaard, M. (2010), Modelling tourism in the new national transport model - a multi-day approach, To be presented at Trafikdage, August 23-24 2010, Aalborg, Denmark.

Rich, J., Nielsen, O.A., Brems, C. (2010), Overall Design of the Danish national transport model, To be presented at Trafikdage, August 23-24 2010, Aalborg, Denmark.

Rich, J., Prato, G.C., Daly, A. (2010) Activity-based demand modelling on a large scale: Experience from the new Danish National Model, To be presented at the European Transport Conference, October 9-11 2010, Glasgow, Scotland.

Van Ommeren, J.N, Rietveld, P., Nijkamp, P. (1998) Spatial moving behaviour of two-earner households. *Journal of regional Science* **38**(1), pp. 23-41.

Vuk, G., Hansen, C.O., Fox, J. (2009) The Copenhagen Traffic Model and its application in the Metro City Ring Project, *Transport Reviews*, **29**(2), pp.145-161.