Methods for forecasting in the Danish National Transport model

Jeppe Rich

DTU Transport

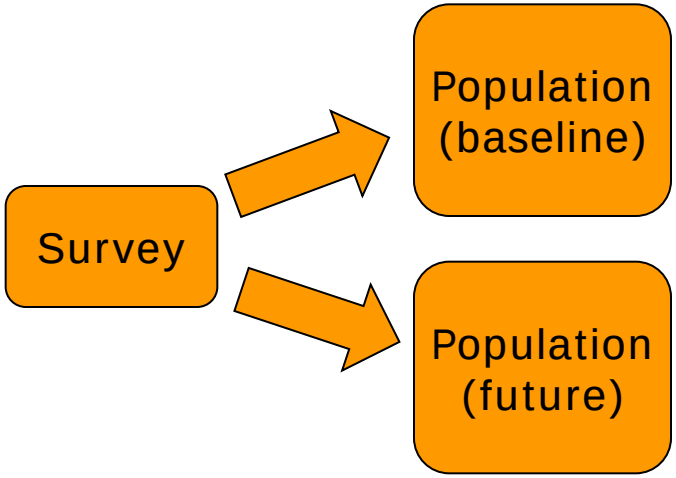
DTU Transport
Department of Transport

Outline

- Introduction – forecasting is difficutl!
- Overall model structure
- The general forecast approach
- Structure of the population syntheziser
  - Definition of master table
  - Targets
  - Initial solution
- Test of precision
- Summary and conclusion

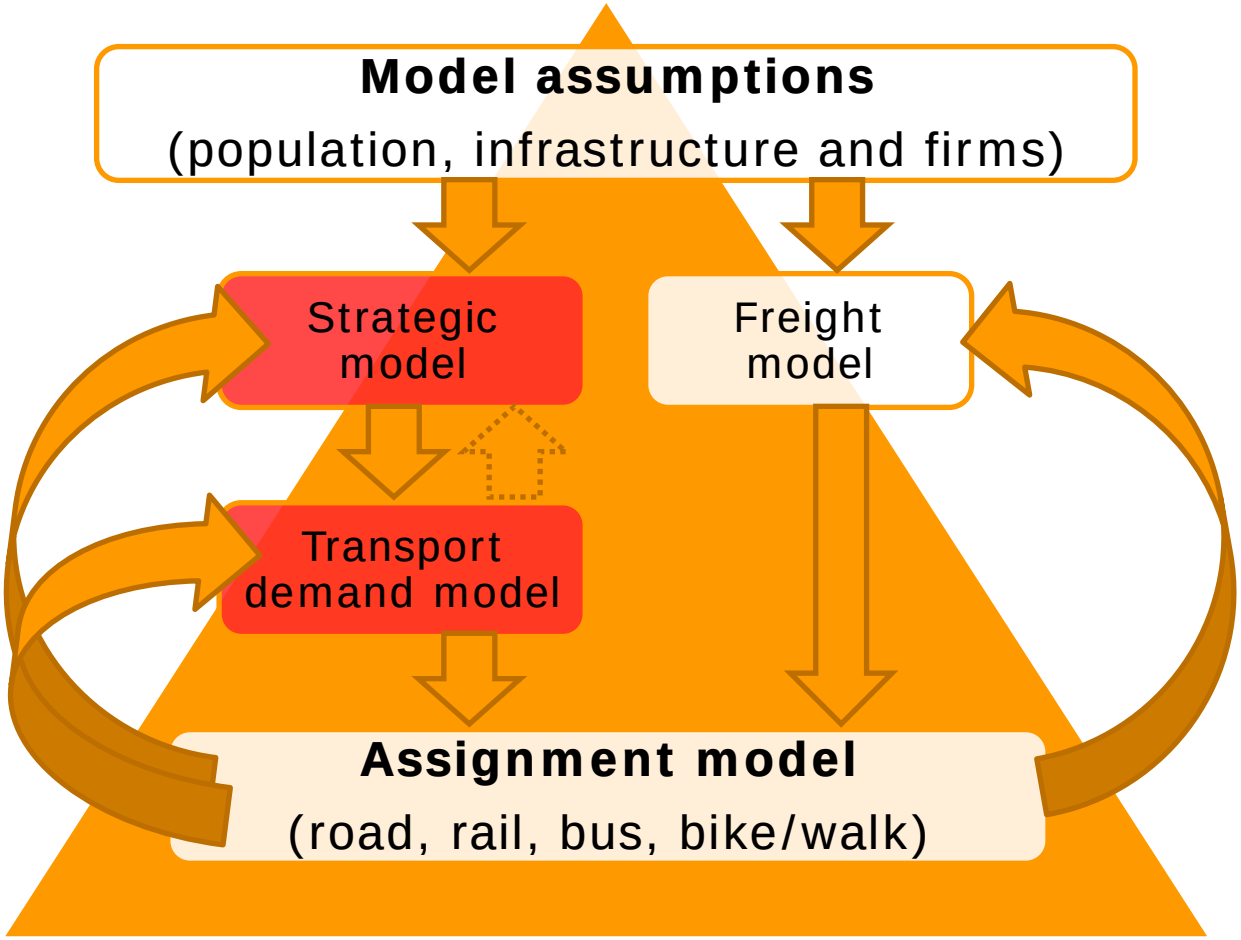
Introduction

- Forecasting of transport demand is difficult
- It require that we are able to explain the demand of the population on the basis of a survey
  - Even in the baseline it may be difficult to replicate demand (the survey may not be representative for the population)
  - More difficult when forecating as the future population is unkonwn



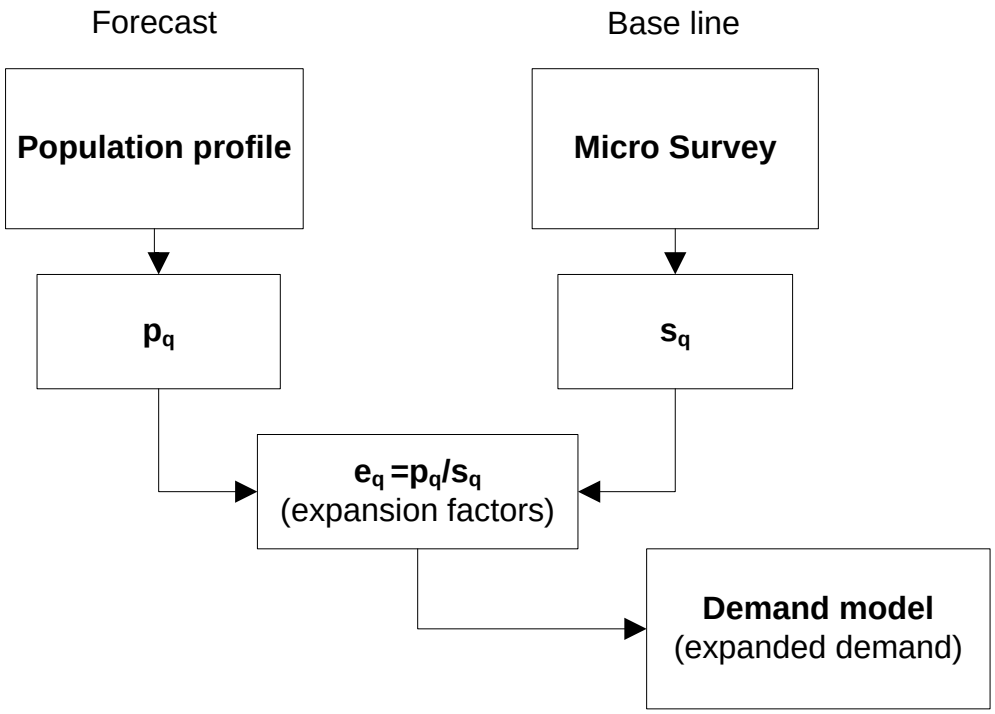
Overall model structure

- The framework will consist of the several componts



The general approach

- The standard approach will be “sample enumeration”
- We divide the population in different socio-groups *q*
  - s_q *represent the number of respondents in socio-group in the survey*
  - p_q *represent the number of respondents in socio-group in the population*
  - $e_q = p_q / s_q$ *is the expansion factor that “lift” the survey to the national level*



Prototypical sample enumeration (PSE)

- Matrices are then represented by a possible probability model, a frequency matrixs, and scaled with expansion factors

$$T_{idm} = \sum_n P_n(d,m \mid x_{ni}, z_{dmi}) T_{ni}\, e_q(n)$$

- The up-weighting is applied directly to the survey model
  - Summing over *n* replicate the entire population
- PSE is only possible if we have a solid RP data foundation and can generate $e_q(n)$
  - *E.g. require TU and register data*

A matrix approach

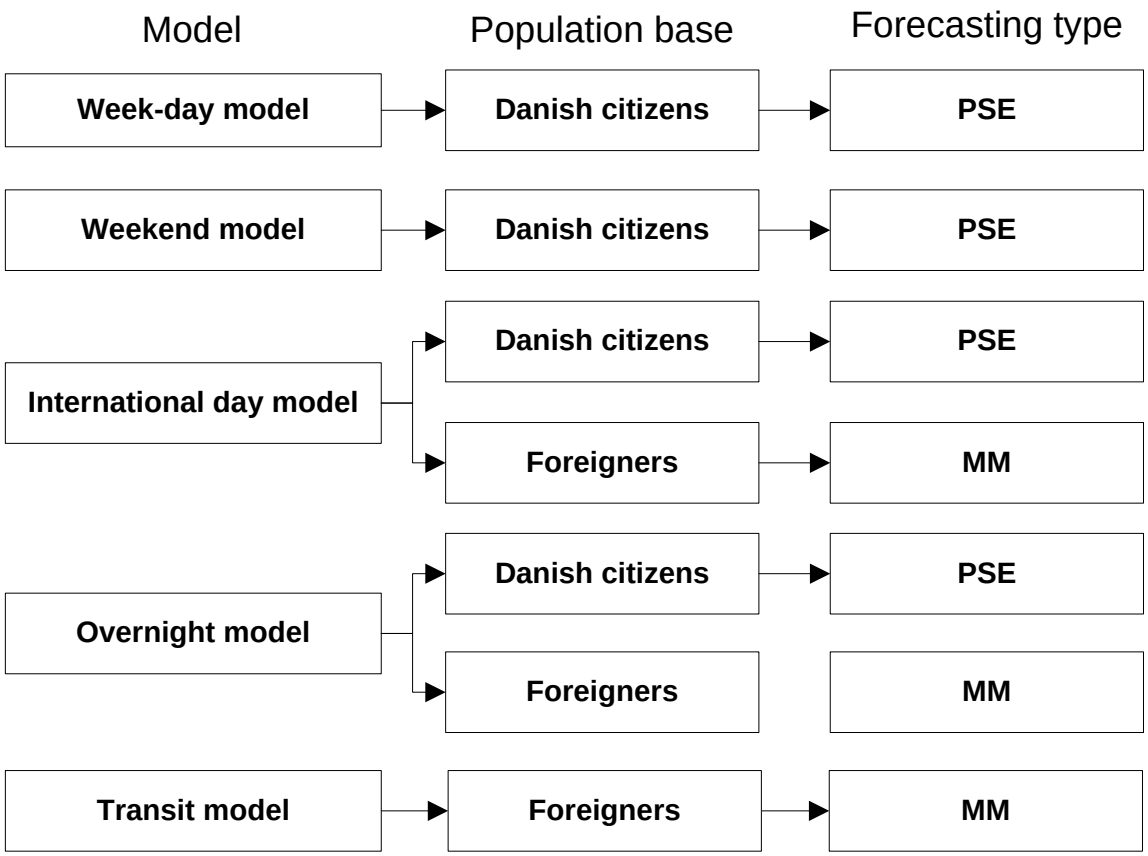
- The model is formulated at the matrix level

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

- Index *n* has been skipped and we only consider matrices
- If the model is calibrated (at the matrix level) to replicate the baseline matrix, the model will replicate the population demand
- Fewer data is required as the modelling entity is zones
- However, can lead to aggregation bias as

$$Pr([\Sigma_n x_n / N]) \neq [\Sigma_n Pr(x_n)] / N$$

PSE and MM in the National model



The PSE synthesizers

- The key to do forecasting is to calculate expansion factors to represent the structure of the future population
  - Expansion factors are essentially derived from the formula $e_q = p_q/s_q$
- As a result, the key to do forecasting is therefore to derive a population table p_q at any point in time
- In the national model, three synthesisers are developed;

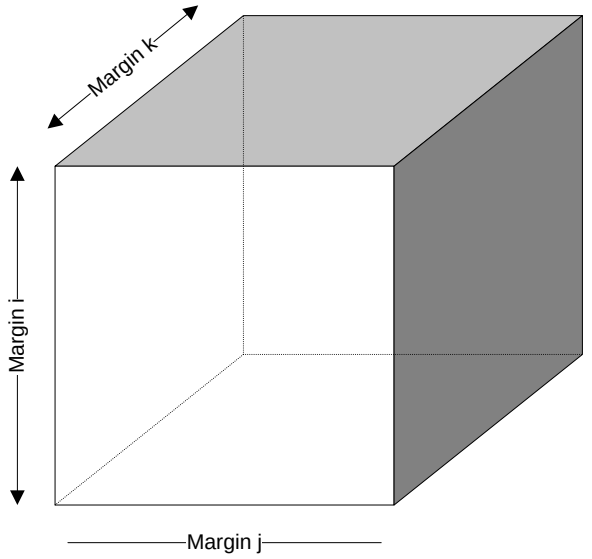
(i) Population synthesiser

(ii) Household synthesiser

(iii) Labour demand synthesiser (firms and public institutions)

Synthesiser methodology

- The synthesisers will be based on an iterative proportional fitting (IPF) algorithm
- The population tables are defines as a ”hyper-cube”
  - The objective is to estimate the ”interior” of the cube
  - This is done on the basic of (i) data on the margins, (ii) and an initial solution
- Forecasts are then developed by changing ”margins” or ”targets” according to, e.g. official forecasts



Simple ”two-target” example

- Consider two targets; Income and Age
  - Income is defined for three income groups 0-200.000, 200.000-500.000, and 500.000 – DKK.
  - Age is defined for three age groups 0-25, 26-59, 60- years
- Gray area define ”initial slution” from survey
- The ”master table” is the age×Income (3 by 3 table)

	0-25	26-59	60-	**Income target**
0-200	*43*	*25*	*17*	**3.000**
200-500	*39*	*55*	*23*	**5.000**
500-	*9*	*27*	*19*	**2.000**
Age target	**4.000**	**4.500**	**1.500**	**10.000**

Master tables for the population synthesizer

- The design of the socio-grouping should be relevant from a transport perspective
  - More group will in principle enable a more precise synthesizer, however, only if we can forecast these
  - The most detailed master table represent 9 million entries

Type	Categories	Comment
Residential zone	98	L0 zone system
	176	L1 zone system
	907	L2 zone system
	3,640	L3 zone system
Children	2	
Age group	10	
Gender	2	
Labour market association	6	
Personal income	11	
Cell combinations	2,640	

Household master table

- The household table include information about two workers
  - Income is defined as household income

Type	Categories	Comment
Residential zone	98	L0 zone system
	176	L1 zone system
	907	L2 zone system
	3,670	L3 zone system
Number of adults	3	
Children	3	
Labour market association A	6	
Labour market association B	6	
Household income	11	
Cell combinations	3,569	

Employment demand

- The table is aggregated from register data by simply counting people in the register database
- It represent the only the satiated demand (unemployment or excess demand not considered)
  - Branches is combined with highest education of the employed people
  - Will give further information about the structure of the workplaces
  - Make it possible to develop a ”attraction profile” that is specific to individuals

Type	Categories	Comment
Work zone	98	L0 zone system
	176	L1 zone system
	907	L2 zone system
	3,670	L3 zone system
Branch	111	
Highest education	9	
Cell combinations	999	

Defining targets

- The definition of targets is important because it defines the dimensions (margins on the ”hyper-cube”) that are going to be forecasted
  - Relevant to select targets that can be backed by official statistics and are relevant for transport
  - All to many targets may in principle give detailed output, however, if they cannot be forecasted it is of less value
- Another issue is to ensure consistency between targets
  - In the synthesiser we have embedded a ”harmoniser” which will make all targets consistent according to a ranking scheme of the targets
  - For users it means that targets will be ”harmonised” after they have been changed

Targets for the population synthesiser

- We first consider targets an aggregate socio-economic level (TP_{A1} – TP_{A5})
- A second set of targets represent links between the municipality level and socio-economy (TP_{B1} – TP_{B4})
- Finally, we set targets for the more detailed zone systems
- The ranking in the ”harmoniser” is based on the order of the rows

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

Targets for the household synthesiser

- Aggregate socio-economic targets (TH_{A1} – TH_{A3})
- Links between the municipality level and socio-economy (TH_{B1} – TH_{B4})
- Finally, we set targets for the more detailed zone systems
- The ranking in the ”harmoniser” is based on the order of the rows

Target constraint block	Variable combination	Dimensions
TH_{A1}	Income×Adults	33
TH_{A2}	Income×Children	33
TH_{A3}	Income×Lma(A)×Lma(B)	396
TH_{B1}	Income×L0	1078
TH_{B2}	Adults×L0	294
TH_{B3}	Children×L0	294
TH_{B4}	Lma(A)×Lma(B)×L0	3528
TH_{C1}	L1	176
TH_{D1}	L2	907
TH_{E1}	L3	3670

Targets for employment synthesizer

Target constraint ID	Variable combination	Dimensions
TE_{A1}	Branch11	11
TE_{A2}	Branch27	27
TE_{A3}	Branch111	111
TE_{B1}	Branch11×Education	88
TE_{C1}	Branch11×L0	1078
TE_{C2}	Branch27×L0	2646
TE_{C3}	Branch111×L0	10878
TE_{C4}	Education×L0	784
TE_{D1}	L_1	176
TE_{E1}	L_2	907
TE_{F1}	L_3	3670

The ”harmoniser” making targets consistent

- The harmonisation ensures that the level is defined at the highest ranking target
  - Lower ranking targets are then defined by using the relative distribution of these, but scaled with the correct absolute level
- Consider a simple example age = {3500, 4000, 3500} and income = (3000, 4000, 3700)
- If age dominate income, we would ”harmonise” income as

  Income = (3000/10700, 4000/10700, 3700/10700)*11000

	0-25	26-59	60-	**Income target**
0-200	*43*	*25*	*17*	**3.084**
200-500	*39*	*55*	*23*	**4.011**
500-	*9*	*27*	*19*	**3.803**
Age target	**3.500**	**4.000**	**3.500**	**11.000**

Consistency when targets are cross-linked

- A more serious problem occurs when targets are cross-linked
  - One target variable are represented in more than one target

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

Consistent targets

- Consider a simple example
- Three targets that are not cross-linked, e.g. $T_1(a)$, $T_2(i)$, and $T_3(l)$ with marginal probabilities given by

  $Pr(a) = T_1(a) / \sum_a T_1(a)$

  $Pr(i) = T_2(i) / \sum_i T_2(i)$

  $Pr(l) = T_3(l) / \sum_l T_3(l)$
- A consistent target vector T(a,i,l) is given by

  $T(a,i,l) = [\sum_a T_1(a)]^* Pr(a)^* Pr(i)^*Pr(l)$
- However, if targets are cross-linked, e.g. T1(a,i) and T2(a,l) then

  $Pr(a,i,l) \neq Pr(a,i)^*Pr(a,l)$
- A solution can be found by solving a special LP problem

Initial solution

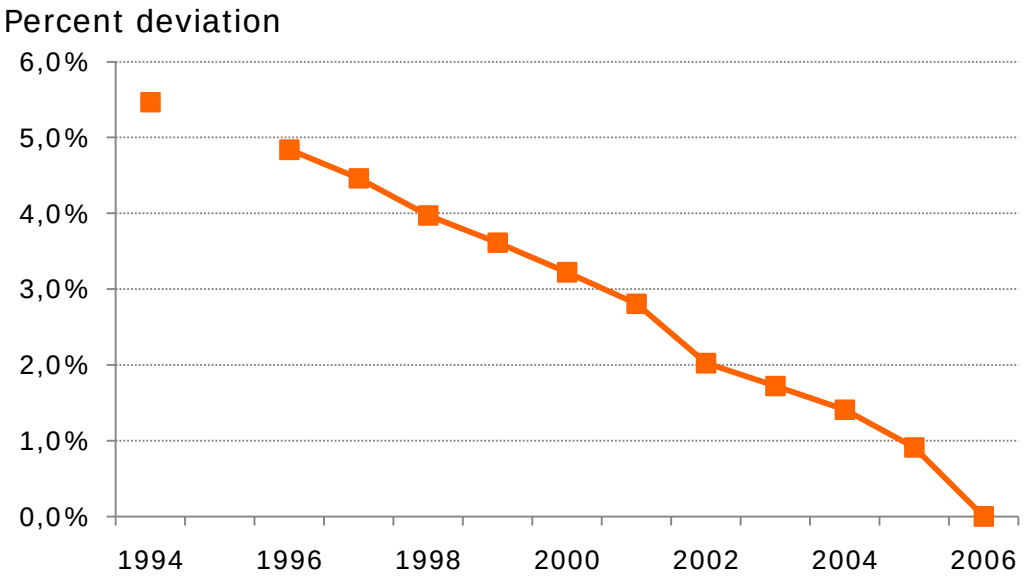
- We will allow editing of the initial solution as well
- If the initial solution have a zero in an entry, the solution will return a zero
- This is not always reasonable
  - People are becomming older and there could be an ”aging” effect that needs to be considered
  - Development areas, that are ”empty” in the baseline, but ”filled” in the future (Ørestad region is one example) is also a potential problem

Running the syntheziser

- **Step 1:** Carry out a harmonisation process of all socio-economic targets, e.g. only TP_{A1} through TP_{B4} for the population synthesiser
- **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).

Forecast example

- To test the forecast accuracy we have defined 2006 as ”target year”
- All other years are applied as ”initial years”
- The premise is that the ”targets” are correct
  - An almost linear decline in the precision
  - A 5.5% overall perecent deviation on a 12 year period



Summary and conclusions

- Two frorecast strategies are applied; a prototypical sample enumeration approach and a matrix approach
  - The PSE approach is based on the calculation of expansion factors
  - The calculation of expansion factors are based on a population synthesiser
- Three synthesiser are considered
  - Population, household, and employment demand
- An IPF algorithm is applied
- Definition of consistent targets is an issue
  - A harmoniser is used
  - Cross-linked targets are dealt with in a prior LP program
- A test of an ”ideal” forecast is considered and results are promising