TOWARDS MICROSIMULATION OF PASSENGER AND FREIGHT TRANSPORT COMPETITION: ADVANCES IN SYNTHETIC POPULATION GENERATION AND SIMULATION OF THE BEHAVIOUR OF FREIGHT ACTORS

Johan Barthélemy[1], johan.barthelemy@math.fundp.ac.be

Eric Cornelis[1], eric.cornelis@math.fundp.ac.be

Bart Jourquin[2], bart.jourquin@fucam.ac.be

Jérémy Piotte[2], jeremy.piotte@fucam.ac.be

Philippe Toint[1], philippe.toint@math.fundp.ac.be

[1] Department of Mathematics - Transportation Research Group
University of Namur - FUNDP
Rue de Bruxelles, 61
B-5000 Namur (Belgium)

[2] Group Transport & Mobility
Facultés Universitaires Catholique de Mons - FUCaM
Chaussée de Binche, 151
B-7000 Mons (Belgium)

ABSTRACT

Competition of freight transport and passenger traffic on the road infrastructure is an increasingly important problem. The approach considered in this paper is that of disaggregate modelling using multi-agents systems. We focus on two aspects of this methodology: synthetic population generation for passengers' agents and behaviour' simulation of the freight transportation actors. Each of these challenges is covered as follow: firstly the existing approaches are discussed, then the proposed technique is detailed and finally validation and computational results are presented.

Keywords: Microsimulation, Multi-agents systems, Synthetic population, Time dependant origin-destination matrices, Freight and passengers transports

INTRODUCTION

Competition of private vehicles and truck traffic on the road infrastructure is an increasingly important problem, especially when expanding the road network is difficult or impossible, this last situation being common in developed countries (*e.g.* Hicks, 1977 and Thompson and Taniguchi, 2001). Indeed, joint use of the road by individual travellers and freight firms has a significant impact on road congestion, and thereby on economy, environment and life quality. Its analysis and strategic planning is therefore of growing significance. This competitive use of public infrastructure has been studied in the past in the framework of traffic assignment: Dafermos (1971 and 1972) and Potts and Oliver (1972) extended the Wardrop user equilibrium condition to handle the multiclass traffic assignment. This condition was restated by Fernandez and Friez (1983) (see also Van Vliet *et al.*, 1986). Toint and Wynter (1995) study the impact of competition for space on the use of multiclass equilibrium assignment techniques.

The approach considered in this paper is that of disaggregate modelling, using microsimulation. It is the core of the DIDAM[1] research project from which results discussed here are extracted. The general philosophy of this project is to construct, for the country of Belgium, a set of agents representing the individual car passengers and the actors of the freight transport system, each associated with its own set of behavioural rules and localized at the municipality (NUTS-5) level. These two groups then generate a simulated dynamic demand for network use, which are then combined in a joint microsimulation assignment model.

The number of agents (autonomous entities interacting with their environment by their decisions and actions) involved in disaggregate travel demand models, such as activity-based travel techniques, usually involve a large number of agents (*e.g.* Bhat *et al.*, 2004, Hensher *et al.*, 2004, Los Alamos National Laboratory, 2005, Mathis 2009 and Chaker *et al.*, 2009). In the DIDAM framework, the first challenge is the generation of the set of potential car passengers, which we identify with the complete country's population. It represents a total of approximately ten millions individuals in approximately four millions households. Even if collecting detailed individual data for each of these agents were possible (which is not the case for obvious cost considerations), its use would be very problematic in view of the stringent privacy laws of Belgium. The construction of an artificial (synthetic) population for the country therefore appears to be the necessary. Fortunately, the use of existing databases (such as "*Enterprise"* from the Service public de Wallonie, 2009) for the construction of the population of freight firms is easier and not subject to legal restrictions, the difficulty being then to model the behaviour of these actors. The present paper focuses on these two challenges in the DIDAM project: synthetic population generation for passenger agents and behaviour of the freight system actors.

[1] Disaggregated Demand And Assignment Models for Combined Freight and Passenger Transport

The paper is organized as follows. In Section 2, we first review existing techniques for the generation of synthetic populations of individuals and households and indicate their limitations in our context. We pursue our state-of-the-art presentation by considering, in Section 3, previous proposals and techniques for multi-agents behavioural modelling. Building on the conclusions of these sections, we next present in Section 4 a new synthetic population generator whose objective is to circumvent the exposed difficulties and analyze its performance in terms of accuracy. Section 5 then presents the authors' views on the construction of a time-dependant demand matrix for freight transport based on multi-agent technology. Some conclusions and perspectives are finally discussed in Section 6.

EXISTING APPROACHES FOR GENERATING SYNTHETIC INDIVIDUALS AND HOUSEHOLDS

Iterative proportional fitting

To date, the conventional approach for building synthetic population is based on the method developed by Beckman et al. (1996). The main idea behind a population synthesizer consists of merging aggregate data from one source covering the complete population with disaggregated data from a sample in order to get a complete and detailed disaggregated data set for the population of interest. Typically the aggregate data set is an aggregate outcome from an existing census and the disaggregated data set is drawn from a survey over a sample of the population. The aggregate data consist of a set of marginal distributions for some of the relevant characteristics of the true population. We refer to these distributions and variables as target and control variables. Furthermore, the disaggregate data provides full information about the attributes of interest, but only for a sample of agents and is referred to as the seed.

The population synthesis procedure usually starts with identifying the relevant (categorical) socio-demographic variables of the agents. Assuming that the seed presents n attributes of interest and denoting by $\boldsymbol{V} = \{v_1, v_2, ..., v_n\}$ the vector of variables representing these attributes, each combination of values of v_i's therefore defines a socio-demographic group. The synthetic population is then generated by a two steps procedure:

1. Starting from the seed, estimate the *k*-way joint-distribution of the true population, where k ($\leq n$) is the number of control variables, such that the resulting distribution is consistent with the marginal distributions (margins) of the target and preserves the correlation structure of the seed.

2. Select agents from the sample and copy them in the synthetic population in a proportion derived from the distribution computed in the previous step.

The most popular way for estimating a *k*-way joint distribution table based on some known marginal distributions and a sample is the well known iterative proportional fitting procedure (IPFP) originally described by Deming and Stephan (1940). The procedure implies an initial representative sample of the true population being available. This requirement is important since Mosteller (1968) pointed out that the procedure preserves the interaction structure of the sample as defined by the odd ratios. According to Ireland and Kullback (1968), the IPFP also produces the estimated contingency table that minimizes the discrimination information, *i.e.* it yields the constrained maximum entropy estimator. Moreover Little and Wu (1991) demonstrated that IPFP results in a maximum likelihood estimator of the true contingency table. The IPFP uses an initial contingency table of the control variables build from the seed as a starting point. The procedure then iteratively updates the cells depending on the marginal distributions of the target until the margins of the table match the target's ones.

Once the expected number of agents in all the socio-demographic groups is estimated, each sampled agent is associated with a probability of being selected in the synthetic population. This probability typically depends on the agent's sampling weight and the expected number of similar agents in the true population. Based on these probabilities, agents are randomly drawn from the seed using a Monte Carlo procedure until the expected number of agents is reached for each socio-demographic group. When a sampled agent is drawn, then all its attributes, including the uncontrolled ones, are pasted in a new synthetic agent who is added to the synthetic population.

Limitation of IPFP in the Belgian case

As expected, the IPFP results largely rely on the quality of the data. In particular, it is important to notice that the method requires consistency of the margins across the targets and representativeness of the initial sample of the true population. For example if a class of agents is not represented in the seed then this particular class will remain unpopulated in the final synthetic population. These two requirements limit the applicability of the IPFP in real situations.

In addition, recent mobility surveys such as EGT (Direction Régionale de l'Equipement d'Ile-de-France, 2005), MOBEL (Hubert and Toint, 2001) or NTS (Office of UK National Statistics, 2010) suggest that the travel behaviour of an individual is significantly influenced by the type and composition of his/her household. This illustrates another limitation of the conventional approach: it is very unlikely for analysts to have access to a single dataset detailing the joint-distribution of individuals' and households' attributes simultaneously. Since the estimation step of the algorithm is designed to deal with a single contingency table, the conventional approach can consequently account either for individual-level or for household-level control variables. In other words this process results in a synthetic population where either the households or individuals joint-distributions match the desired ones but not both. Note that historically, households' distributions accuracy has been preferred (Ye et al, 2007).

Guo and Bhat (2007) proposed an algorithm to overcome this last issue by controlling simultaneously the individual- and household-level variables. Their algorithm generates a population where the household-level distributions are closed to those estimated using the IPFP, while simultaneously improving the fit of person-level distributions. Arentze *et al.*

(2007) proposed another method using relation matrices to convert distributions of individuals to distributions of households such that marginal distributions can be controlled at the person level as well. Ye et al. (2009) further built on previous work and proposed a practical heuristic approach called Iterative Proportional Updating (IPU), based on adjusting households' weights such that both household- and individual-level distributions can be matched as closely as possible.

However these new approaches remain based on the conventional IPFP and thus rely on the same assumptions, i.e. that a significant sample of the population of interest is available and that the available data is consistent in the sense that margins extracted from available but different joint distributions are equal. The last requirement is critical for convergence of the IPFP algorithm.

Unfortunately both these requirements could not be met in the case of the Belgian population for the DIDAM project. In the first place, a representative sample at the municipality level (which is the needed spatial disaggregation level) is indeed not available. A second problem is that all necessary information (*i.e.* distributions of variables of interest) is not available from a single database (which would hopefully guarantee consistency), but has to be extracted from different datasets, typically produced by different institutions and/or using different protocols or data cleaning mechanisms. This results in significant differences between margins, as is illustrated in Table 1 (Cornelis, Legrain and Toint, 2005) for the Charleroi and Nivelles districts.

Table 1 - Examples of inconsistencies between margins extracted from different data sources

Joint-distribution	Data source	Charleroi		Nivelles	
		Margins	Prop.	Margins	Prop.
municipality **x** gender **x** age	GéDAP, 2001	**405,491**	1	**342,028**	1
municipality **x** household type	GéDAP, 2001	**380,653**	0.94	**321,777**	0.94
municipality **x** diploma	GéDAP, 2001	**426,372**	1.05	**321,144**	0.94
municipality **x** activity status	GéDAP, 2001	**396,594**	0.97	**299,145**	0.87
district **x** household type **x** age	INS, 2001	**357,884**	0.88	**302,240**	0.88
district **x** diploma **x** age	INS, 2001	**398,582**	0.98	**299,443**	0.88
district **x** activity status **x** age	INS, 2001	**385,024**	0.95	**289,989**	0.85

In this table, the total number of inhabitants in these districts using the most reliable of our data sources (first row of the table) is compared with the same statistic extracted from other data files to be used in our population synthesis (rows 2 to 7). Whether or not these data files have the same origin does not prevent the differences between numbers of inhabitants to reach around 10%. Differences accruing from surveys at different periods are even more substantial. It is therefore desirable to develop an alternative population synthesis tool which would not suffer from the lack of a representative sample at the most disaggregate level and/or from (moderate) inconsistencies between different data sources. This is the object of Section 4.

MULTI-AGENTS SYSTEMS

DIDAM is based on state-of-the-art concepts and methods. Central to the disaggregate approach is the “activity chain” concept, that is a chain of actions/movements whose scheduling is performed by all the agents who are involved, facing the constraints imposed by infrastructure, time organization and prevailing regulations (such as land-use planning and fiscal (dis)incentives). Activity based models for passenger transport demand has indeed long advocated that people do not move for the sake of it, but rather because of their desire to participate in successive activities taking place at different locations. The chain of the activities, with their locations and timings, is then considered as the determinant matrix from which individual trips result. The chain of successive handling operations and transportation modes in a (multi-modal) trip for freight follows the same paradigm. It is also the case of the (multi-modal) logistic chain in the same transportation sector.

The modelling approaches which are developed exploit this common paradigm of a chain or sequence of "activities" (taken in a broad sense) subject to constrained decisions regarding their actual content (how many activities, of which nature?), their intrinsic characteristics (such as activity nature, mode of transport or localization) and their timing/scheduling. In such an approach, the various choice models are typically applied to an artificial population of actors, which also need further specific research in how such synthetic populations can be generated.

Working at a fully disaggregate level is however not easy. If such models are already available for passenger demand and traffic, disaggregate tools are still largely missing for freight transport. This is not only true at the operational level, but, more crucially, at the conceptual level. This is why the research program adopts, in this domain, a progressive approach that introduces disaggregation gradually into existing methods and models. This entails research in a full spectrum of issues, ranging from concepts definition (who are the actors, how can they be characterized ...) to validation exercises using available data sources.

Multi-agents systems become also more and more used in microsimulation, *i.e.* focused on individual entities. In transport models, microsimulations try to represent the behaviour of every single vehicle; at the opposite macro-simulations study aggregated traffic flows. There is not a worse or better type of simulation, each one having its advantages: microsimulations are adapted to study interactions between actors within the traffic, but become too expensive for large-scale networks while macro-simulations, well adapted for large-scale networks, do not take care about the interactions between vehicles.

As stated earlier, multi-agents systems are used to model freight transportation. There are many different definitions of what an agent is. Therefore it is common to define an agent by its properties, *i.e.* an agent is an entity

- able to act in its environment which it has a limited awareness;

- able to communicate with other agents;
- able to take its own decisions;
- and autonomous.

All these characteristics make multi-agents systems an interesting way to simulate interactions and negotiations between actors in a dynamic system.

There is not a lot of scientific literature about the use of multi-agents in transport models. Multi-agents systems have however been used in a few studies. For example, El Hmam, *et al.* (2006) proposed a hybrid model of traffic flow using multi-agents systems. The authors were able to combine a macroscopic and a microscopic model in a single integrated model. They used the Payne's model (1971), a second order macroscopic model derived on the basis of car-following considerations taking into account the driver's reaction time which leads to a dynamic mean speed equation, to describe the sections of highways and multi-agents systems to simulate the behaviour of vehicles at discontinuities like roundabouts, crossroads… Another example is the work of Sirikijpanichkul *et al.* (2007), who used multi-agents to optimize the location of intermodal freight hubs. They used agents (hub owners, transport network infrastructure providers, hub users, and communities) with conflicting objectives that initiate negotiations that result in a location choice for a new intermodal freight hub. In their paper, Wisinee *et al.* (2007) proposed a microsimulation model for urban freight movement incorporating the relationship between freight agents in supply chains where each of them tries to minimize the cost of each activity. Using their model, they analyzed the urban freight movements in the Tokyo Metropolitan Area. In the work of Van Katwijk *et al.* (2004) a test bed for agent-based road traffic management is presented. It consists of three combined models: interaction model, intelligence model and world model, respectively used to model interactions between the agents, the artificial intelligence of the agents (with or without experience notion), and the traffic using Paramics, a microscopic traffic simulation package. By means of two scenarios, they show their test-bed offers a valuable aid for in-depth analysis in the field of traffic management. Another interesting source is the OVID research project (2002-2005), launched by the German Ministry of Research and Education. The goal of OVID is to analyze the impact of intelligent information systems in road transport. In this project the interaction between human agents, transport models, transport data sources and software agents to assess the benefits of better information and of de-central decision-making is modelled.

Multi-agents systems have been tackled from a more theoretical point of view by J. M. Vidal (2007). Although not focused on transport but more on games and AI it covers all the aspects that a multi-agent practitioner should be familiar with.

A NEW POPULATION SYNTHESIS TECHNIQUE

We start the description of our work on the DIDAM challenges by describing our new tool for synthetic population generation which is applicable in the context of Belgium. The general philosophy of our proposal is construct individuals and households by drawing their characteristics or members at random within the relevant distribution at the most disaggregate level available, while maintaining known correlations as well as possible.

We now outline the main steps of our procedure and discuss its application to the 11,060,573 individuals in 4,333,769 households located in 589 municipalities (NUTS-5 level), situated themselves in 43 districts (NUTS-3 level) containing between 2 and 35 municipalities each. In order to preserve the scope of the current paper, our description attempts to highlight the main ideas but is not fully detailed: a complete and formal algorithm description is given and discussed in Barthélemy and Toint (2010).

Proposed generator

The algorithm consists of a 3-steps procedure applied to each municipality:

1. a pool of available individuals is generated for the current municipality, namely the individuals' attribute joint-distribution denoted by *Ind*;

2. the households' joint-distribution is estimated and stored in the contingency table *Hh*;

3. the synthetic households are constructed by randomly drawing individuals from the individual's pool *Ind*. This is achieved while preserving the distribution computed in the second step. Once a household has been built, it is added in the synthetic population.

We now provide more information on each of these successive steps.

Step 1: Estimating the individuals' distribution

The first step aims at building the *Ind* pool of available synthetic individuals. This pool is built individual by individual and the contingency table updated accordingly. Because disaggregated data is unavailable at the municipality level for some attributes (while margins are), we have to use a more aggregate level to obtain (approximate) information on the missing attributes. In our case, this level turned out to be the district level.

Assume that the individuals are characterized by a vector of attributes $\boldsymbol{V} = \{V_1,...,V_n\}$ and denote by v_i the value taken by the i^{th} attribute V_i for a particular individual. Denote by $Mun = V_{m1} \times V_{m2} \times ... \times V_{mi}$ and $Dis = V_{d1} \times V_{d2} \times ... \times V_{dj}$ the joint-distribution of variables respectively at the municipality and district level, where V_{mk} and V_{dl} correspond to a component of V. The cells of the *Mun* contingency table thus correspond to the numbers individuals with the attributes (v_{m1}, v_{m2}, ..., v_{mi}) and are considered sequentially. For each of the individuals in the considered cell, the generator randomly draws the remaining missing attributes accordingly to the *Dis* joint-distribution. Once the individual is completely defined, the *Ind* contingency table is updated.

Since draws from the district-level joint-distributions were used to assign some characteristics, the margins of *Ind* for these particular variables can be inconsistent with the known true one. A correction is then made to *Ind* to make it consistent with the margins at municipality level. This correction is computed by suitably shifting some of the attributes' values of certain individuals. Only shifts between two contiguous modalities are allowed, *e.g.* if an individual's age class is 5, then the shift allowed are either 4 or 6.

Step 2: Estimating the households' repartition

Denote the vector of the households related attributes by $\boldsymbol{W} = \{W_1,...,W_m\}$ and the value taken by a household for the j^{th} attribute by w_j. Now that a pool of individuals has been built, the next step is to find an estimator of the households' type contingency table denoted by *Hh* given some known data provided by different sources. Each cell of *Hh* correspond thus to a number of a particular household type given by a combination of w_j's.

In our algorithm, the estimation of *Hh's* cells given known data is obtained as the solution of an optimization problem, where the entropy is maximized under the (linear) constraints implied by the known margins on household types. This approach has the advantage of producing a more reasonably spread-out distribution amongst all household's type with respect to the constraints than would be produced by a least-squares formulation. Denoting by x the vector of the unknown cells of *Hh*, the corresponding objective function of this problem can then be written as

$$\min_x \sum_i x_i \ln(x_i) - x_i$$

such that the constraints on household types are satisfied.

Unfortunately, due to the corrupted and inconsistent nature of the available data, the constraints of this optimization problem are often formally incompatible. Our approach is then to impose only a subset Ω of them (those corresponding to higher quality data) as strict constraints, the other constraints being then incorporated in the objective function in a form penalizing their violation. Each of these penalized constraints is affected with a penalization weight *pen* depending on the quality of the associated data source and the numbers of households. Denoting by *A* and *b* the matrix and the vector derived from the subset of the scaled inconsistent constraints, the new objective function can now be formulated as

$$(EN) \qquad \min_x \ \|Ax - b\|_2^2 + \sum_i x_i \ln(x_i) - x_i$$

and the minimization is then carried out under the constraints in the set Ω only. In general, the solution of this optimization problem yields non-integer solution, which is unsuitable for representing household numbers. The solution's components of this optimization problem is then rounded and we finally compute

$$f_{EN}(\hat{x}) = \sum_i w_i \times |b_i - \hat{b}_i| + \sum_i \hat{x}_i \ln(\hat{x}_i) - \hat{x}_i$$

where $\hat{x}$, $\hat{b}$ and w respectively denote the rounded solution of (*EN*), the vector and the weight associated to the i^{th} constraint. This function can be seen as a performance measure describing how well the rounded integer solution fits the whole set of the initial constraints. We then loop over a set of values for the penalization parameter *pen*, picking the best rounded solution associated with the lowest f_{EN}. This solution is finally used as a starting point of a combinatorial optimization problem minimizing f_{EN} using a tabu-search algorithm (see Cvijovic *et al.* 1995, Glover 1989, Glover 1990 and Glover *et al.* 1997), in order to get a final estimation of *Hh*. More details on this (admittedly involved) process are provided in Barthélemy and Toint (2010), but it is enough to note here that the household-type distribution is essentially computed, for each municipality, as the approximate solution of a maximum entropy problem in discrete variables under constraints given by statistics known for the municipality.

Step 3: Households' generation

Individuals' and households' distributions being estimated, the last step consists of gathering individuals into households by randomly drawing households' constituent members. Households' types are considered sequentially and household's members generated as follow: a household head is first drawn from the pool of individuals, and then, depending on the household's type, a mate, children and additional adults are also drawn from the pool if relevant. Figure 1 gives a schematic description of the procedure.

We provide some more detailed comments on this process. The construction of a household starts with the selection of its head. The corresponding attributes' values are either randomly drawn according to known joint-distributions or directly derived from the current household type considered. Once the individual's class from which the household head is to be extracted is determined, the pool of individual is updated, *i.e.* the associated cell in *Ind* is decremented. More formally, this selection procedure is organized in 3 steps:

1. determine the desired attributes values (*i.e.* the v_i's): either derived from the current household type, or randomly drawn according to known distributions;

2. add a household head to the household being generated:

   - if the corresponding individual's class is still populated in the pool of individuals, extract an individual from this class and make it the household's head;

- else the generator tries to find a suitable household head in the constituents members of previously generated households. This last individual is then replaced with an appropriate one randomly drawn in the pool of the remaining individuals.

3. *Ind* is updated accordingly to the actions performed in step 2.

Depending on the household type, the generator may pursue the construction of the current household by selecting a head's mate, children and additional adults. The corresponding selection procedures are similar to the head's one. The generated household is finally added to the synthetic population with its constituent members and the *Hh* cell corresponding to the current household type is decremented Depending on the value of $Hh[w_1][w_2]…[w_m]$, the generator performs one of the following actions:

- if $Hh[w_1][w_2]…[w_m] = 0$ then the next household class is examined;

- else the generator builds a new household of the same class.

When there is no more household class remaining to examine, then the synthetic municipality is declared complete and the procedure moves on to the next municipality, if any.

When the procedure stops for a municipality after exhausting either the pool of individuals or the pool of households, data inconsistencies may result in two types of inaccuracies. In the first case, *i.e.* when the pool of individuals is exhausted first, the final number of households is smaller than anticipated, while the final number of individuals is smaller than estimated in the second case. We show below that the second case is much more likely in our experience.

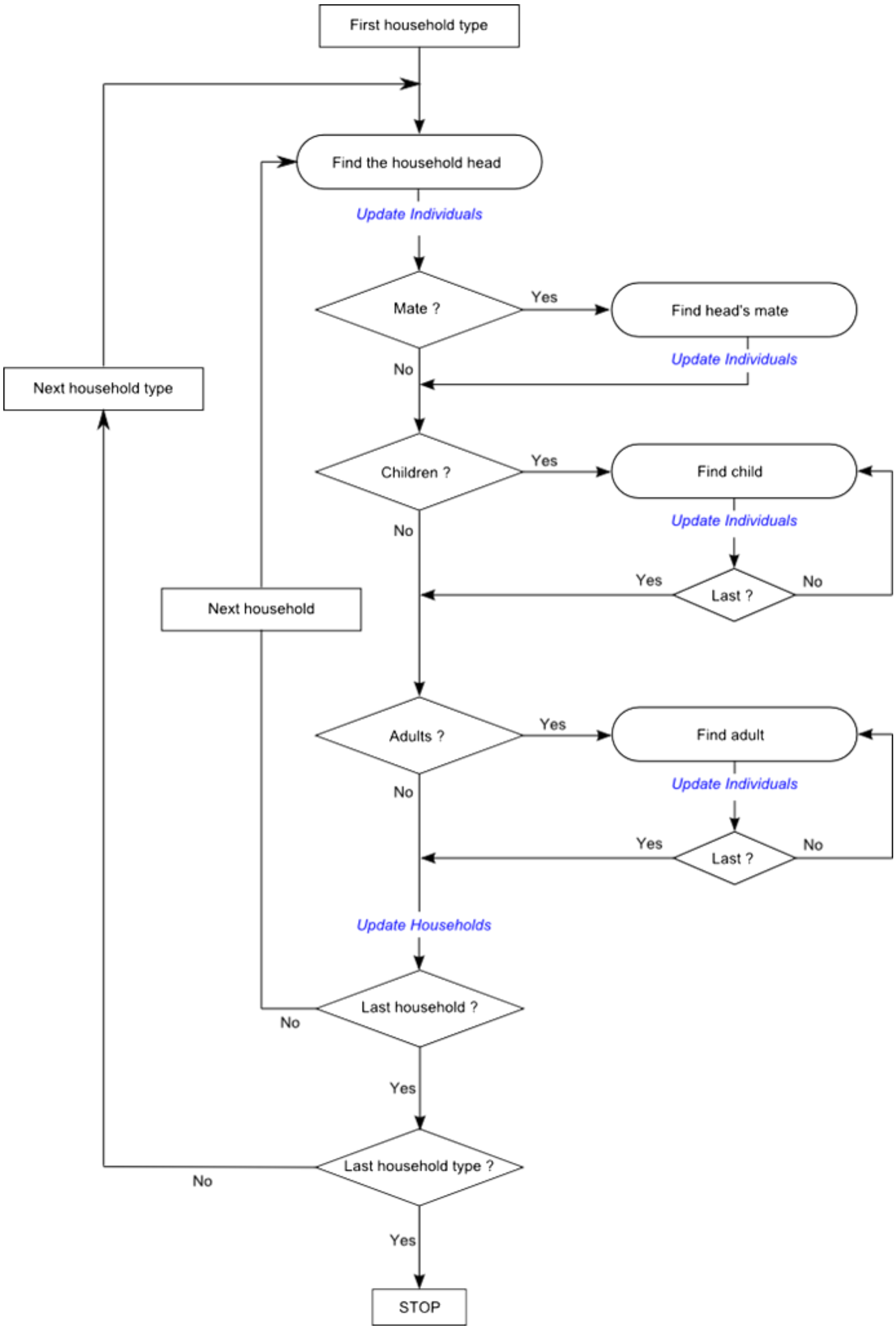

Figure 1 - households' generation

Results

The procedure outlined in the previous subsection has been applied to generate a synthetic population for Belgium over the 589 municipalities which is estimated to be 11,060,573 individuals gathered in 4,333,769 households. The individuals and households attributes are respectively described in Tables 2 and 3.

Table 2 - individuals' characteristics

Attribute	Values
Gender	male; female
Age class	0-5; 6-17; 18-39; 40-59; 60+ years old
Activity	student; active; inactive
Diploma	primary; high school; higher education; no diploma
Driving license ownership	yes; no

Table 3 - households' characteristics

Attribute	Values
Type	single man alone single woman alone single man with children and possibly additional adults single woman with children and possibly additional adults couple (without children and additional adults) couple with children and possibly additional adults
Number of children	0 to 5
Number of additional adults	0 to 2 (mate not included)

One is then faced to the question of estimating the quality of the generated population. As in Bhat (2007), one possible performance measure to assess the generator accuracy is the absolute percentage difference (APD) between the estimated contingency tables computed in the first step of the generator and the corresponding ones resulting from the household generation step. This measure is calculated for a particular cell (u_1, …, u_p) as follow:

$$APD_{T,T'}(u_1, u_2, \dots, u_p) = \left| \frac{T'[u_1] \dots [u_p] - T[u_1] \dots [u_p]}{T[u_1] \dots [u_p]} \right|$$

where *T* and *T'* denotes respectively the estimated and the generated table. The lower the APD is, the better the generated table fits the estimated one. Results are reported in Table 4.

Table 4 - Generated agents

	Estimated	Generated	Difference	APD
Individuals	11,060,573	10,638,112	422,461	0.039
Households	4,333,769	4,333,762	7	< 0.001

First note that the procedure was able to generate 10,608,112 individuals gathered in 4,333,762 households, meaning that it could build a synthetic population where the number of households is very close to the estimated one and differs less than 4% for the number of individuals. This is highly encouraging.

At a more disaggregate level, Figure 4 gives a representation of the mean and the standard errors of the APD of each individual type over the 589 municipalities. This figure suggests that the generator produces relatively small APD on average. Moreover, these APDs are associated with small standard errors, meaning that the APDs' values are relatively stable across the municipalities.

Table 5 presents some basic statistics on the average APDs value (AAPD) of the cells of the distributions computed across all the municipalities.

Table 5 - AAPD statistics

Distribution	Min	Max	Std deviation	Mean
Ind'	0.000	0.065	0.013	0,020
Hh'	0.000	0.006	< 0.001	< 0.001

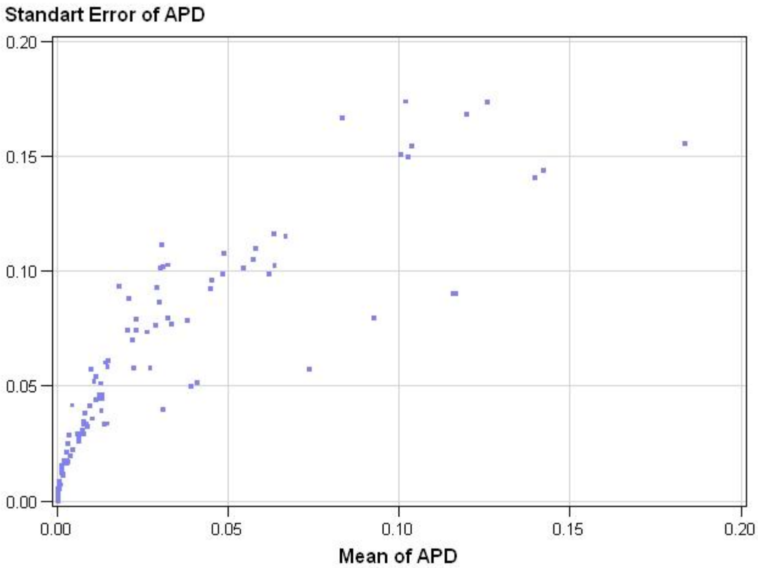

Figure 2 - APDs' mean and standard deviation for each individual type

The APDs related to *Hh* are not represented graphically but are very small. This property is obtained by construction of the algorithm, because households' generation terminates either when all the households are built, *i.e. Hh' = Hh* or when the pool of individual is empty such that no more households can be generated. Fortunately this last situation occurs very rarely: in the Belgian case, it happens only for a single municipality (Borgloon), where the difference between the total numbers of desired (4881) and generated (4874) households is 7, resulting in a very small AAPD (0,006). Because of this property ensuring highly accurate Household structures, the remainder of this subsection focuses on individual-related distributions.

The synthetic municipality associated with the worst AAPD value for the individuals is Tubize. The details of this municipality are described in Table 6. It indicates that even if this entity is the least accurate, the distribution of generated individuals is still reasonably close to the estimated one: in average, the APD between the estimated and generated distribution for a given individual class is less than 7%. Moreover the generator produces a population having 9.3% less individuals than the estimated one. These results are illustrated on Figure 5 representing the numbers of individuals generated against the number of estimated ones for each class of individuals. If all the class had been perfectly matched, then all points would be aligned on a diagonal. Nevertheless one can easily observe that the differences between the generated number of agents and the desired one are generally low for each class, *i.e.* the contingency tables produced by the generator fits the initial one quite accurately, given the initial level of data inconsistencies.

Table 6 - Tubize

	Value
Estimated	24,441
Generated	22,170
Difference	2,271
APD(Estimated, Generated)	0.093
AAPD	0.065

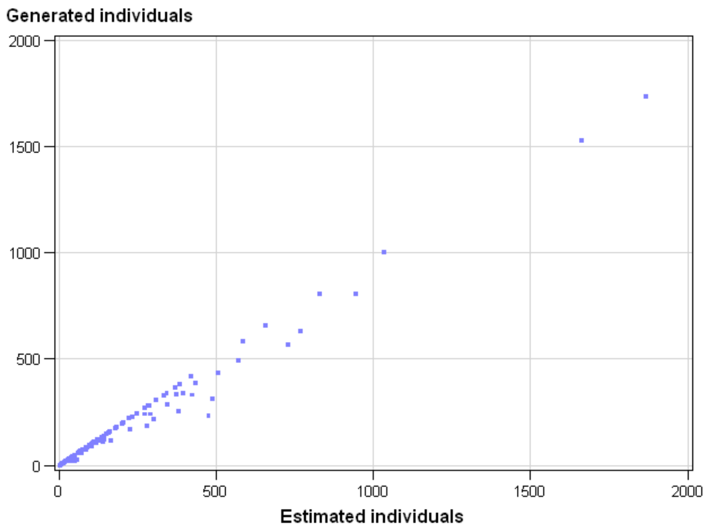

Figure 3 - Estimated x Generated agents for each individual type

More tests on the quality of fit will be reported in Barthélemy and Toint (2010), but the preliminary results presented here indicate that the generator has a real potential in practice.

TIME-DEPENDENT FREIGHT OD MATRICES GENERATION USING MULTI-AGENTS METHODOLOGY

In this section, we shall address the time dependant origin-destination (OD) matrices generation involving freight transport. As said earlier, the methodology in this part of the project is to create a multi-agents environment where the entities representing the miscellaneous actors of freight transport interact. The generation of these agents, the simulation of the negotiation process and the time dependant OD matrix generation will be looked over in this section.

This research is still in progress and the simulations carried out are still in continuous evolution. We started our research in this field doing a simple simulation representing an enterprise surrounded by some transport societies. This enterprise had some goods to move, so it made a transport demand to the transporters. These ones proposed a price to fulfil the contract in function of the distance they had to cover to reach the origin of the transport from the location of their trucks, the distance to cover to make the transport itself and the distance to cover to go back home from the destination of the transport. The enterprise selected the best offer (the cheapest) and accepted the contract proposition of the transport society associated.

As said earlier, in order to simulate the interactions in a realistic way, we used a database: "*Entreprises*", from "*La Direction des Réseaux d'Entreprises de la Région wallonne*". It contains more than 4800 enterprises located in Wallonia including 433 transport societies. It is based on a written survey validated by phone calls. The information is updated every 6 months.

Thank to this database, we've been able to generate the Walloon enterprises and transport societies taking into account their location, number of employees and activity sector. This information gives us an idea of the available fleets for the transporters in quantity and type of vehicle. So each transporter agent is fitted with its own trucks fleet characterized by a number and kind of trucks allowing it to transport freight of one or another NSTR chapter. Furthermore, the number of employees of an enterprise is used to associate it with a proper demand frequency. In the present state, this frequency is only a function of the number of employees. However, this assumption is not realistic and we are working to improve the model to take account other characteristic. Regarding the location of the agents, we decided work at NUTS-5 level of aggregation which represents the municipality level in Belgium.

Simulation: OD(t) generation

As previously stated, this simulation consists in the negotiation between the actors (agents) of the freight transport. We will consider direct interactions between demand and offer. We won't take into account potential intermediaries which could play a role in the market. We have on one side the enterprises agents and on the other side the transporter agents.

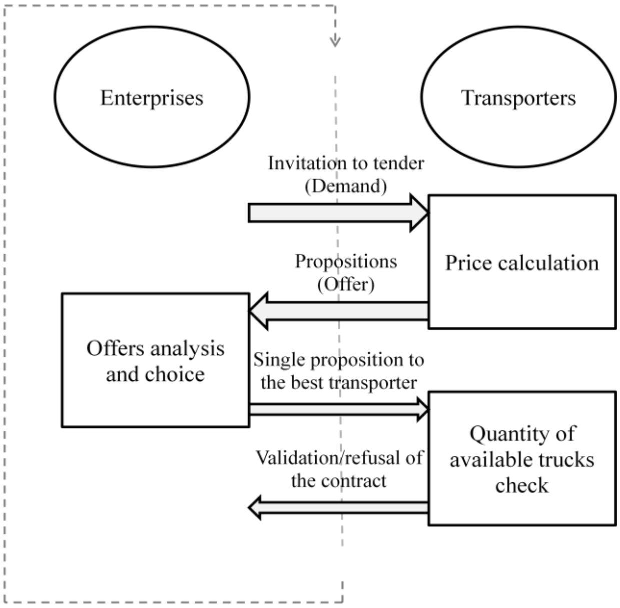

Figure 4 - General scheme of negotiation process between enterprises and transporters

The enterprises first make a public call for transport service and the transport agents able to propose a contract answer by proposing a price. Each enterprise agent will typically receive several answers in which case it will choose the cheapest one and will sign a contract with its sender. In this exploratory phase, every enterprise agent makes a public demand at a given frequency which is proper to it and defined as told in previous section. The general scheme of the simulation is showed in Figure 6.

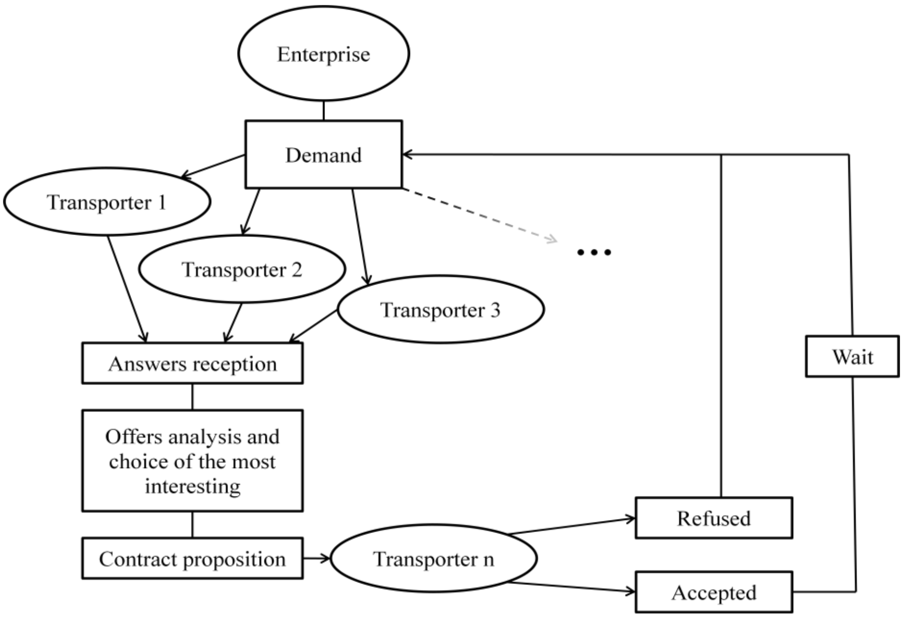

Figure 5 - Behaviour of an enterprise agent

Let's give a special attention to the behaviour model of each kind of agent. As we said earlier, every enterprise (Figure 7) will make a transport demand at its own defined frequency. At the moment of the generation of this demand, it will create a need of transport of a random quantity of freight of a given type related to the sector of activity of the enterprise. This attribution of a realistic freight quantity in the demand is in progress at the same time as the demand frequency problem we already mentioned. We'll talk about it in more details in the chapter about the perspectives of our simulation. A message holding this demand is sent to a set of transporter geographically close from the origin, the destination or the axis between these two points. When all the answers are received, the enterprise agent chooses the best offer (the cheapest, until now) and sends back a contract proposition to the transporter agent that had made the offer. If it is accepted, the enterprise agent goes to sleep until next time it has to move freight, if refused, the demand is dropped but saved as dropped demand for further analysis. Notice that the number of transporters included in the set of neighbours of an enterprise is a model parameter that directly influence the computing time and the size of the output OD(t) matrix. So as this parameter is low, the computing time decreases but more demands are dropped because the enterprises don't consider enough neighbours and these neighbours can't fulfil the demand.

Concerning the transporter agents (Figure 8) it is suitable to notice that they are made up of two behaviours: "the waiting for demands" and "the processing of demands". At the beginning, a transporter is waiting (behaviour 1). When it receives a message holding a demand, it processes it calculating the cost of the demanded transport and then sends back a message with the price proposition and goes back to waiting state. When it receives a message holding a contract proposition (behaviour 2), it checks if there is still enough trucks available to fulfil the contracts and accept or refuse it in function. Once this is done, it goes back to a waiting state.

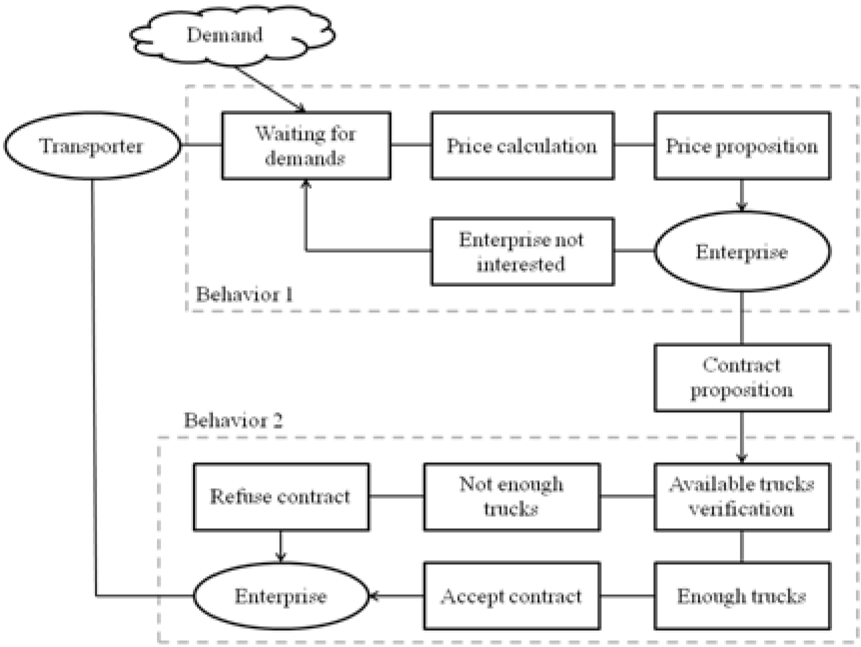

Figure 6 - behaviour of an agent representing a freight carrier

Results

During the simulation, each time a contract is sealed by a transport agent, the details of the transport are saved in a file to create the time dependent OD matrix. Each line contains an origin point, a destination point, a quantity transporter and the starting time of the transport. In a further version, we might also save information about the number and kind of vehicles used to make the transport. It is important to notice that a sealed contract is generally associated with three lines in the matrix: One representing the trucks moving from the transport society parking to the transport origin, one for the transport itself and one for the way back to the transport society parking. So our OD matrix contains the empty trips.

Depending on the number of neighbours considered by the enterprises (and the period one tries to simulate), the simulation computing time varies from 1 hour to 48 or more hours. A period of one week has been simulated to analyze the impact of some simulation parameters as the number of neighbours. The program builds a time dependent OD matrix for the Wallonia at aggregation level NUTS-5. The impact of the number of neighbours on computing time, dropped demands and OD size has been analyzed in Table 7. In this case, the simulation represents some 23,688 demands. Note that this numbers are subject to changes in the next months. Indeed, one can imagine that the fact of taking into account more realistic shipment frequencies can change the size of the matrix. Furthermore, the possibility for the transport societies to make better use of their trucks by the use of a flow consolidation strategy, trying to combine several transports, will also hugely change this matrix. This modification about the transporters politic will be discussed in Section 6.

Table 7 - Number of neighbours analysis

Neighbours	Computing time (min)	OD size (lines)	Dropped demands	Dropped demands (%)	Time(s) / demand
1	62	37,335	8,990	37.95	0.2531
2	66	47,892	4,821	20.35	0.2100
3	104	51,633	3,249	13.72	0.3053
4	119	54,143	2,018	8.52	0.3295
5	126	55,000	153	4.87	0.3355
10	140	56,985	209	0.88	0.3578
20	146	55,890	4	0.02	0.3699
50	153	55,154	0	0	0.3875
100	171	55,164	0	0	0.4331
200	173	55,000	0	0	0.4382
300	193	55,051	0	0	0.4889

Consider both the computing time and the portion of neglected demand in function of the number of neighbours considered by the enterprises (see Figure 7). One can see that as the neglected part of the demand falls from 38% to 0%, the computing time increases from 60 minutes to 150 minutes. And then even when more and more neighbours are considered by the enterprises, as the dropped demand doesn't vary, the computing time slowly increase. This short study leads to the conclusion that neglected part of the freight transport demand can be lowered to 0. Besides, the neighbours' parameter increases the computing time mostly when it permits to reduce the dropped demand.

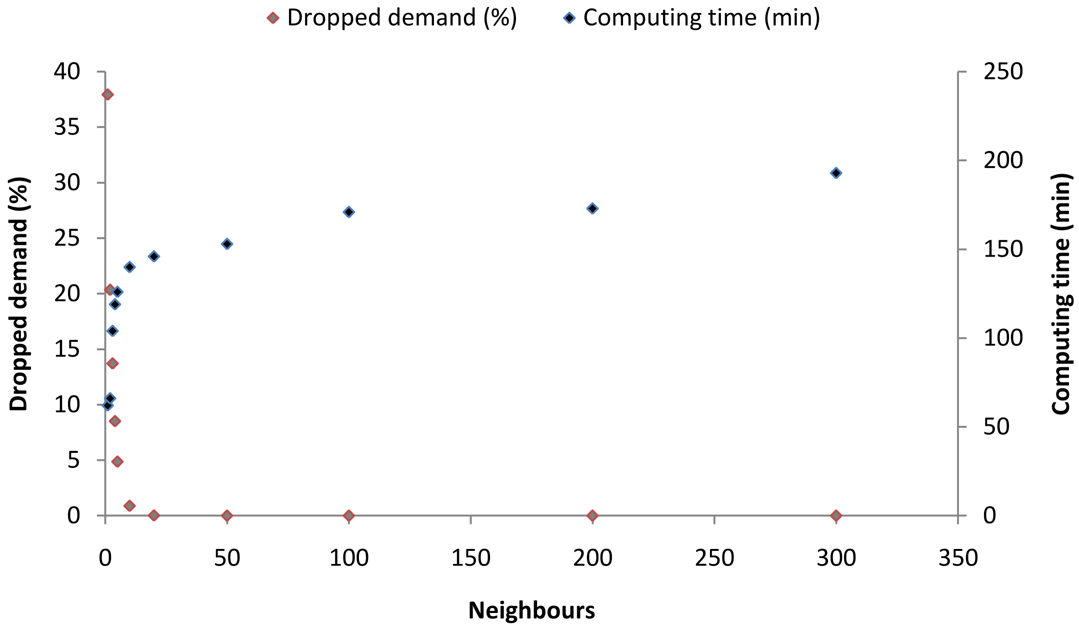

Figure 7 - Dropped demand and computing time in function of the number of neighbours considered by the enterprises

Finally, one can discuss the ratio between the size of the OD matrix and the number of demands that it represents. In fact this ratio is an estimator of the effectiveness of the transporters strategy aiming at maximizing the number of combined demands. For example, consider one transporter facing *N* demands. Assuming that these demands all concern different OD pairs and different from the transporter location. If this transporter doesn't use any strategy to combine some demands, the OD(t) matrix resulting will contain (3 *x N*) lines. For each demand, one to go from the transporter location to the origin of the demand, one for the transport and one to come back. On the contrary, if the transporter applies a strategy that permits to combine all N demands, his trucks will be source of (2 *x N*) + 1 lines in the matrix. Notice that for 100,000 demands, one would expect that the OD(t) matrix would contain 300,000 lines. But in fact a lot of transporters are located at the origin or the destination of a demand and consequently only produce 2 lines in the matrix instead of 3. By now the ratio is about 2.33 (with neighbours-value for which no demand is disregarded). One can then conclude that only 33% of the demands are fulfilled by a transporter that is located neither at the origin nor the destination of the transport.

PERSPECTIVES, CONCLUSIVE REMARKS AND FUTURE RESEARCH

Synthetic population

We have presented a new synthetic population generator which obviates the need for a significant sample of households and individuals. As in IPFP-based approaches, the proposed procedure attempts to maximize the entropy of the unknown contingency tables. It also has the advantages of allowing the merging of several data sources and of handling reasonable inconsistencies between them. Even though preliminary results indicates that the new methodology has potential, more tests and validation procedures have to be conducted in order to assess its efficiency, accuracy and robustness. Evolution of the databases is also considered, which will undoubtedly test the stability and practicality of the new algorithm further.

Temporal OD matrix for freight transports

This project is still in progress, implying that some assumptions need to be modified or dropped, *e.g.* the one about the demand frequency and shipment size by the enterprises. This work is underway using the ECHO (Guilbaut M. *et al*, 2008) survey conducted by the *Institut National de Recherche sur les Transports et leur Sécurité* (INRETS, France). It intends to have a better understanding of transport practices. The originality of this survey lies in the study of the shipment size and covers 3,000 establishments of more than 10 employees in the industry, wholesale trade and warehousing sectors. A shipment is generally associated with a transport service and a vehicle movement. As a result it is directly linked to the number of vehicles-kilometres on our networks.

Furthermore, the ECHO survey shows an important rate of small-size shipments (50% of them have a weight lower than 30kg). So in order to be realistic, we will have to insert a flow consolidation strategy of the transporters. This leads to another necessary improvement: in the current model every enterprise “wakes up and goes back to sleep” one by one. Once an enterprise wakes up to make its demand, it sends a demand message to several transporters and waits for answers. During this time, nothing happens on the side of enterprises. In this approach, the transporters don’t have the opportunity to set up a strategy maximizing their benefits. Because they can only handle one demand at a time, they do not know when and where the next demand will occur. In other words they are unable to wait for multiple demands at the current development stage.

A first proposal to overcome this issue is to wake up earlier the enterprise agents, *i.e.* when an agent is waiting answers others wake up and make other demands. Consequently the transporters could handle several demands at the same time and try to make better offers to enterprises that would require less empty trips taking account of this flow consolidation strategy.

Other factors are still not taken into account as seasonality effects and own account transport. These phenomena are to be implemented in the next months.

Finally, the enterprises always choose the cheapest offer. Indeed, only the transport costs are taken into account in the calculations and negotiations. These costs depend on the localization of the transporter in regard to the origin and destination of the transport demand. It would be interesting to take into account time slots. For example an enterprise could ask its freight to arrive at the destination at a defined time of the day. In this case, transporters able to propose a service at the desired time would be privileged. Besides, experience rules upon punctuality could be introduced in the model to make it more realistic.

When the simulation will be considered realistic and consistent enough, we will first try to validate the time dependent OD matrices built using common OD matrices holding relation between different NUTS-5 regions for the NSTR chapters, themselves built on the basis of the last information in the frame of TRANSTOOLS European project at the NUTS-2 aggregation level.

ACKNOWLEDGEMENTS

The authors wish to thank the *Groupe d'étude de démographie appliquée* (GéDAP, Université Catholique de Louvain, Belgium) for providing the data, derived from the most recent available datasets collected for the 2001 Belgian census, used as inputs for the synthetic population generator. Helpful corrections from Xavier Pauly, Amélie Schatteman, Laurent Van Malderen and Fabien Walle are also gratefully acknowledged.

REFERENCES

Arentze T., H.J.P. Timmermans, and F. Hofman (2007), 'Creating Synthetic Household Populations: Problem an Approach' – Transportation Research Record: Journal of the Transportation Research Board, 2014, 2007, pp. 85-91.

Barthélemy J., Cornélis E., Jourquin B., Limbourg S., Piotte J. (2009), 'The DIDAM framework Disaggregated demand and assignment models for combined passengers and freight transport', Proceedings of the BIVEC-GIBET Research Day, http://hdl.handle.net/2268/40417.

Barthélemy J., Toint Ph. (2010), 'Synthetic population: a generator obviating the conventional approach's limitations', in preparation.

Beckman R.J., Baggerly K.A., and McKay M.D. (1996), 'Creating synthetic baseline populations', Transportation Research Part A, 30(6), 415-429.

Bellifemine F., Caire G. and Greenwood D. (2007), 'Developing multi-agent systems with JADE', *Wiley Series in Agent Technology*, Michael Wooldridge eds, Liverpool University, UK.

Ben-Akiva M. and Lerman S.R. (1985), 'Discrete Choice Analysis: Theory and Application to Travel Demand', *The Mit Press*, Cambridge, Massachusetts.

Casey H.J. (1955), 'Applications to traffic engineering of the law of retail gravitation', *Traffic Quarterly,* IX(1), 23-35.

Chaker W., Proulx M-J., Moulin B. and Bédart Y. (2009), 'Modélisation, simulation et analyse d'environnements urbain peuplés – Approche multi-agent pour l'étude des déplacements multi-modaux', *Revue internationale de Géomatique*, 19(4), 413-442.

Combes F. (2009), 'Sur le choix de la taille d'envoi en transport de fret', Thèse de doctorat spécialité transport Université Paris Est - Ecole doctorale Ville Environnement*. Laboratoire Ville Modalité Transport (LVMT).*

Cornélis E., Legrain L. and Toint Ph. (2005), 'Synthetic populations: a tool for estimating travel demand', Proceedings of the BIVEC-GIBET Research Day 2005 Part I, 217-235.

Cvijovic D. and Klinowski J. (1995), 'Taboo search - an approach to the multiple minima problem'. Science, 267, 664-666.

Dafermos S.C. (1971), 'An Extended Traffic Assignment Model with Applications to Two-way Traffic', *Transportation Science*, 5, 366-389.

Dafermos S.C. (1972), 'The Traffic Assignment Problem for Multiclass-User Transportation Networks', *Transportation Science*, 6, 73-87.

Deming W.E., and Stephan F.F (1940), 'On a least-squares adjustment of a sampled frequency table when the expected marginal totals are known', Annals of Mathematical Statistics, 11, 428-444.

Dial R.B. (1971), 'A probabilistic multipath traffic assignment model which obviates path enumeration' *Transportation Research* B, Vol 5, pp. 83-111.

Direction Régionale de l'Equipement d'Ile-de France (2005), 'Les cahiers de l'Enquête Globale de Transport', http://www.ile-de-france.equipement.gouv.fr/.

Domencich T. & McFadden D. (1975), 'Urban Travel Demand: A Behavioural Analysis', *North-Holland,* Amsterdam*.*

El Hmam M.S., Abouaïssa H., Jolly D. and Benasser A. (2006), 'Simulation hybride de flux de trafic basée sur les systèmes multi-agents' *6e conférence Francophone de MOdélisation et SIMulation: MOSIM'06.*

Fernandez J.E.L. and Friesz, T.L. (1983), 'Equilibrium Predictions in Transport Markets: The State of the Art', *Transportation Research*, 17B(2), 155-172.

Fienberg S.E. (1970), 'An iterative procedure for estimation in contingency tables', *Annals of Mathematical Statistics*, 41, 907-917.

Frejinger E. (2007), 'Random sampling of alternatives in a route choice context' *Proceedings of the European Transport Conference*, http://infoscience.epfl.ch/record/117157.

Furness K.P. (1965) 'Time function iteration' *Traffic Engineering and Control,* ***7****(7), 458-60.*

Glover F. (1989), 'Tabu Search — Part I', *ORSA Journal on Computing*, 1: 3, 190-206.

Glover F. (1990), 'Tabu Search — Part II', *ORSA Journal on Computing*, 2: 1, 4-32.

Glover F. and M. Laguna. (1997), 'Tabu Search', *Kluwer*, Norwell, MA.

Guilbault M., Costa G., Franc P., Gouvernal E. Hémery C. and Rizet C. (2008), 'ECHO: Envois – Chargeurs – Opérateurs de transport' *Synthèse INRETS n°56.*

Guo J.Y. and Bhat C.R. (2007), 'Population synthesis for microsimulating travel behaviour', *Transportation Research Record: Journal of the Transportation Research Board*, 2014, pp. 92-101.

Hicks S. (1977), 'Urban freight' in *Urban Transport economics* by D.A. Densher, Cambridge University press, Cambridre (UK), Ch. 7:100-130.

Henser D. A. and Kenneth J. B. (2000), 'Handbook of transport modeling', *Pergamon, Elsevier*, Amsterdam (NL)

Ireland C.T., and Kullback S. (1968), 'Contingency tables with given marginals', *Biometrika*, 55(1), pp. 179-188.

Jourquin B. and Beuthe, M. (1996), 'Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe' *Transportation Research C*, Vol 4, N° 6, pp. 359-371.

Liedtke G. (2006), 'An actor-based approach to commodity transport modeling', *PhD thesis, Institute for Economic Policy Research (IWW), University of Karlsruhe (TH).*

Little R.J.A. and Wu M-M. (1991), 'Models for contingency tables with known margins when target and sampled populations differ', *Journal of the American Statistical Association*, 86, pp. 87-95.

Mathis P. (2009), 'An urban dynamic micro simulation based on activities program: use of the cellular graph for a multi-agents simulation', *Recherche Transports Sécurité*, 102, pp. 23-45.

Mosteller F. (1968), 'Association and estimation in contingency tables', Journal of the American Statistical Association, 63, 1-28.

Ortùzar J. De D. and Willumsen L.G. (1990), 'Modeling transport', John *Wiley & Sons*, Chichester, England.

Payne H.J. (1971), 'Models of Traffic and Control' *Simulation Council Proceedings*, Vol. 1. Mathematical Models of Public System, Ch. 6:51-61.

Potts R.B. and Oliver R.M. (1972), 'Flows in Transportation Networks', *Academic Pres*, New York.

Ryan J., Maoh H., Kanaroglou P. (2007), 'Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms' Center for Spatial Analysis – Working Paper Series, McMaster University, Hamilton (Canada).

Service public de Wallonie: Direction générale opérationnelle, économie, emploi et recherche (2008), 'Entreprises' *Région Wallone*: http://economie.wallonie.be.

Sirikijpanichkul A., Van Dam K.H., Ferreira L. and Lukszo S. (2007), 'Optimizing the location of intermodal freight hubs: an overview of agent based modelling approach', *Journal of transportation systems engineering and information*, pp. 71-81.

Thompson R.G. and Taniguchi E. (2001), 'City Logistics and Freight Transport', *Handbook of logistics and supply-chain management – Volume 2*, pp. 393-405, Pergamon, Elsevier, Amsterdam (NL).

Toint Ph.L. and Wynter L. (1996), 'Asymmetric Multiclass Traffic Assignment: A coherent formulation', *Transportation and Traffic Theory*, pp. 237-360.

Train K.E. (2003), 'Discrete Choice Methods with Simulation', *Cambridge University Press*, Cambridge (UK).

Office of UK National Statistics (2010), 'National Travel Survey', *Department of Transport.*

Van Katwijk R.T., Van Koningsbruggen P., De Schutter B. and Hellendoorn J. (2005), 'A test bed for multi-agent control systems in road traffic management', *Proceedings of the 84th Annual Meeting of the Transportation Research Board*, Washington (DC), Paper 05-0774.

Van Vliet D., Bergman T. And Scheltes W.H. (1986), 'Equilibrium Traffic Assignment with Multiple User Classes', *Proceedings of the PTRC Summer Annual Meeting*

Vidal J.M. (2007), 'Fundamentals of multiagent systems', http://www.multiagent.com.

Williams H.C.W.L. (1977), 'On the formation of travel demand models and economic evaluation measures of user benefit', *Environment and Planning,* 9A(3), 285-344.

Williams H.C.W.L. (1981), 'Travel demand forecasting: an overview of theoretical developments', *Transport and Public Policy Planning,* Mansell, London.

Wilson A.G. (1974), 'Urban and Regional Models in Geography and Planning', *John Wiley & Sons,* London.

Wisinee W., Kazushi S., Shoji M. and Pairoj R. (2007), 'Microsimulation model for modeling freight agents interactions in urban freight movement' - *86th annual meeting of the Transportation Research Board*, Washington (DC), Paper 07-2224.

Wong D.W.S. (1992), 'The reliability of using the iterative proportional fitting procedure', Professional Geographer, 44(3), pp. 340-348.

Ye X., Konduri K., Pendyala R. M., Sana B., Waddell P. (2009), 'A methodology to match distributions of both household and person attributes in the generation of synthetic populations' presented at the 88th annual meeting of the Transportation Research Board, Washington (DC).