Modelling hospital visitors for the city of Leuven as input for a FEATHERS-MATSim simulation

Jan Vuurstaek[1]
Luk Knapen[1]
Bruno Kochan[1]
Tom Bellemans[1]
Davy Janssens[1]
Geert Wets[1]

Abstract: *Investigation of the local effects of travel demand and the study of thin flows (low volume travel demand) require micro-modelling and simulation. The ongoing Smart-PT project focuses on demand responsive public transport and other collective transportation services; the Leuven region is used as a research case. This paper reports on the first steps in the project. Those steps aim to model the large number of trips generated by the academic hospital in the moderately sized city. The hospital attracts patients and their visitors from all over Flanders. The method used to sample them from a synthetic population is described and first results are reported. The daily agendas for the synthetic individuals are generated by the FEATHERS activity-based model. The next steps consisting of schedule adaptation for hospital personnel and visitors are briefly discussed. Those are non-trivial but essential to support the micro modelling required to study local public and collective travel demand.*

Keywords*:* Activity-based modelling, FEATHERS, MATSim, schedule adaptation, attraction sites, hospitals

1. Introduction

Activity-based transport micro-simulations are commonly used to determine network loads. When these simulations are capable of simulating public transport, they can even be used to determine the capacity utilization of public transport. This is particularly interesting for public transport companies who want to optimize public transport in city centres. However, in order to accurately determine public transport occupation at city level, large attraction sites need to be taken into account. This paper investigates the capacity utilization for buses in the city of Leuven. This region is heavily influenced by hospital patients/personnel/visitors going to the University Hospitals Leuven, college and university students, and large company sites such as Interbrew. In a first attempt we will account for the hospital patients/personnel/visitors. The simulation will be executed for a Tuesday during July or August, to cancel out the effects of the students.

FEATHERS will be used to generate a daily agenda for each member of the synthetic population. In order to account for the large number of trips generated by the University Hospitals Leuven, hospital patients/visitors will be simulated. Thereto individuals will be sampled from the population to become a patient/visitor. Their schedule will be adapted by insertion of a hospital appointment/visit. This paper reports on the patient/visitor sampling.

This paper is organized in the following way. In Section 2 we describe the relationship between FEATHERS and MATSim. Section 3 focusses the techniques used to account for a large hospital attraction site, whilst Section 4 presents the preliminary results of these techniques.

[1] Hasselt University, Transportation Research Institute (IMOB), Wetenschapspark 5 bus 6, 3590 Diepenbeek, Belgium

The validation of the results is given in Section 5. Section 6 discusses the justification of this research. Finally, Section 7 concludes the paper and provides an expose of planned future work.

2. FEATHERS-MATSim simulation

Simulation of the city of Leuven will be done using the open-source software MATSim (Multi-Agent Transport Simulation). To simulate public transport in MATSim (Rieser and Nagel, 2009), three types of input data are required: network, plans and public transport. The network is extracted from OpenStreetMap (OSM). Public transport data of the public bus operator servicing the region of Leuven (De Lijn), was found on the GTFS (General Transit Feed Specification) Data Exchange website. Last but not least, MATSim needs an initial schedule for each agent in the simulation. These schedules where generated by FEATHERS, an activity-based schedule generator for mutually independent individuals. FEATHERS, like other activity-based schedule generators, is a TAZ (traffic analysis zone) based predictor, trained using travel surveys (totalling to nearly 10,000 respondents) over the complete Flemish region. MATSim on the other hand is a coordinate based micro-simulator. Hence, every location in a FEATHERS schedule needs to be replaced by the coordinates of a street address in the corresponding TAZ. This is done using the Flemish CRAB database (Centraal ReferentieAdressenBestand; Central Reference Address Database). Another problem due to FEATHERS being a TAZ based predictor is that it is not able to accurately capture local effects caused by the presence of a large attraction site, unless the site constitutes a TAZ by itself. Furthermore, FEATHERS does not correctly model night workers since it assumes everyone to be at home at 3.00 a.m. This results in the absence of time-shift work in the generated schedules. This means we need to adapt the schedules predicted by FEATHERS in order to account for the local effects of the hospital.

3. Hospital attraction site

A large hospital, such as the University Hospitals Leuven, attracts many people who arrive and leave at different times over the course of the day. Based on the behaviour of large groups of people arriving at /leaving from the hospital (e.g. visitors need to abide visiting hours), we can divide these people into three main categories: patients, visitors and personnel. For the first two categories (patients and visitors) we discuss the home location determination of each person as well as the schedule adaptation procedure. The personnel category is not discussed here, since this paper covers work in progress.

3.1. Hospital patients

Patients are divided into two subclasses: critical (intensive care, palliative care, small children, etc.) and non-critical. The latter subclass also contains consultations. This is again based on the different behaviour between these two groups of patients. The details of these differences are discussed in the sections below.

3.1.1. Location sampling

Patients for each class are sampled uniformly from the complete population. This is based on the following assumption. The probability for an individual to be a hospital patient, given his/her home TAZ, is given by

$$Prob(p(i)|h(i) = TAZ) = Prob(p(i)|f(i)) \cdot Prob(f(i)|h(i) = TAZ)$$

where i denotes the individual, $p(i)$ is a boolean predicate stating that i is a patient, $h(i)$ is the home location for i and $f(i)$ is a characteristic of individual i that is a vector of predictors to be a patient. Each element in f is the value for a specific element in F which is the set of predictors for being a patient. This set contains age, gender, education level, etc.

In the current version we assume that $Prob(p(i)|f(i)) = Prob(p(i)) = C$; this makes the probability to become a hospital patient independent of the individual's characteristics. As a consequence we assume that the probability distribution is uniform. Hence, we do not need to assume that $Prob(f(i)|h(i) = TAZ) = Prob(f(i))$ which means that the probability for a vector of influencing factors to occur is the same in each TAZ. Those considerations enable easy extension of the method if required. For now they lead to

$$Prob(p(i)|h(i) = TAZ) = C$$

In order to determine a patient's home location, we assume that a deterrence function similar to the one for social contacts holds, i.e. we assume that the attractivity is proportional to the distance (see (Lambiotte et al., 2008), (Wang et al., 2011) and (Krings et al., 2009)) to the hospital. Furthermore, we assume that the larger a hospital, the higher the probability to attract patients; we do not assume that this relationship is linear. Finally, an academic hospital is assumed to be more attractive to critical patients than a non-academic one. All these assumptions lead to

$$a_{h,p} \propto c_{h,p} \cdot (S_h)^{\alpha} \cdot \left(d_{p,h}\right)^{-\beta}$$

where $a_{h,p}$ is the attraction of hospital h to a patient p, $c_{h,p}$ is the weight coefficient for the hospital to attract the patient, S_h is the hospital size (number of beds), α is the coefficient that specifies the effect of the size on the attraction and $d_{p,h}$ is the distance [km] between the patient's home TAZ and the hospital TAZ; β is a coefficient modulating the contribution of the distance. The weight coefficient $c_{h,p}$ depends on the hospital type (academic, non-academic) and the patient class (critical, non-critical). The reason for making the weight dependent on the patient class is the fact that critical patients often are sent to university hospitals by their advising medical doctors. For non-academic hospitals (critical and non-critical patients) and academic hospitals (non-critical patients), $c_{h,p} = 1$, for academic hospitals (critical patients) $c_{h,p} =$ hospitalWeightCritAcad.

The patient class $c_p$is sampled using the critPatientsFraction configuration setting.

For a given sampled individual i, the TAZ $h(i)$ is determined. Then the attractivity for each hospital is determined and the probability for the hospital to be selected is given by

$$p(h,i) = \frac{a_{h,i}}{\sum_{h \in H} a_{h,i}}$$

where H is the set of hospitals. A dataset listing the Flemish hospitals, their addresses and size (in terms of beds) is used. Note that this implies that each hospital is known by each patient.

The patient population is determined by sampling individuals uniformly from the complete population and determining at which hospital the patient resides until a sufficient number of individuals are collected for the University Hospitals Leuven. This number can be configured through the referenceHospitalPatients setting. The amount of patients for each class is determined using the critPatientsFraction.

3.1.2. Schedule adaptation

Building a schedule for a hospital patient, is easier than building a schedule for a person from one of the other two categories. This due to the fact that we do not need to keep the original activities, i.e. we add the hospitalisation to the schedule and drop all other out-of-home activities. This is based on the assumption that people clear their schedule if they are hospitalized. Arrivals of critical patients are distributed uniformly over the day; arrivals/departures of non-critical patients and departures of critical patients are uniformly distributed over the patient intake periods.

3.2. Hospital visitors

Based on the division of the patients, visitors are also divided into two subclasses: critical patient visitors and non-critical patient visitors. This is based on the behaviour of the visitors of these two classes, e.g. critical patients are accompanied by relatives around the clock, whilst non-critical patients are visited only during official visit time periods. Our current sampling method does not yet take this kind of behaviour in account, since this paper is based on work in progress.

3.2.1. Location sampling

Once we know who the patients are and where they live, we can determine where the visitors of a patient live and who they are. Papers (Lambiotte et al., 2008), (Wang et al., 2011) and (Krings et al., 2009) deliver evidence for a gravity model using the square of distance to correlate spatial distance to social closeness. In this project we use a simple distribution similar to the one found in (Krings et al., 2009) and defined by

$$f(d(a,b)) \propto \frac{k}{d^2(a,b)}$$

where $f(d(a,b))$ is the probability density for the distance between the homes of individuals a and b where $d(a,b) \geq 1$. Based on (Krings et al., 2009) it is assumed that for every individual i a fraction of 0.99 of i's acquaintances is living at a distance less than 100[km] from i's home TAZ. For $a = 100$, this leads to

$$\int_1^a \frac{k}{d^2} dx = k\left(1 - \frac{1}{a}\right) = 0.99$$

$$\Rightarrow k = 1$$

The resulting cumulative distribution is given by

$$F(x) = \int f(x)dx = 1 - x^{-1}$$

and with a uniformly sampled value F_0 the distance between a patient and his/her visitor becomes $d_{p,v} = \frac{1}{1-F_0}$.

The home location for a visitor is sampled as follows. Let $\bar{a}$ be the average area of a TAZ, then $r = \sqrt{\bar{a}}$ is used as a tolerance value to determine a set of TAZ locations $\mathcal{L} = \{L_i | d_{p,v} - r \leq d(L_p, L_i) \leq d_{p,v} + r\}$, where L_p is the patient's home location and $d_{p,v}$ is the distance between the patient's location and the visitor's home. If $\mathcal{L}$ is empty due to the way TAZs that lie within the tolerance zone at the required distance are shaped, then the tolerance value is doubled and $\mathcal{L}$ is recalculated. This process is repeated until $\mathcal{L}$ is non-empty. Each location in $\mathcal{L}$ is a possible home location for the visitor and has a probability proportional to its population size to be selected by sampling:

$$p(L_i) = \frac{pop(i)}{\sum_{L_i \in \mathcal{L}} pop(L_i)}$$

Visitors are sampled until a sufficient number of individuals are visiting a patient at the University Hospitals Leuven. This number can be configured through the referenceHospitalVisitors setting.

3.2.2. Schedule adaptation

In order to build a schedule for a hospital visitor of a non-critical patient, a schedule predicted by FEATHERS is taken and decomposed into individual activities. The purpose is to add a hospital visit while minimally decreasing the utility of the given schedule (Knapen et al., 2014). We begin with trying to replace an existing social visit activity by a hospital visit. If that fails, a new social visit activity is simply added. Next, a new schedule is constructed using the new activity set. During this reconstruction alternative locations can be chosen (e.g. for daily shopping), start and end times can be modified and trip modes can be altered. The utility for the resulting schedule is calculated, which can be lower than the utility for the original schedule due to duration compression. Therefore, in a final stage activities will be dropped if that leads to higher utility.

Building a schedule for a visitor of a critical patient is easier, since these visitors will adapt their schedule thoroughly.

4. Preliminary Results

Essential input data have been summarized in the Table 1. Data have been collected from annual reports published by the academic hospital board and from literature (e.g. the average daily visitors for a patient) (Duncan and Heady, 1976). The value for critPatientsFraction is derived from the ratio intensive care beds to regular beds and the value for hospitalWeightCritAcad is assumed (educated guess) since no relevant data have been found yet. The determination of the α and β are discussed in Section 5.

Variable	Value
critPatientsFraction	0.03
hospitalWeightCritAcad	4
referenceHospitalPatients	3356
referenceHospitalVisitors	3213
α	1
β	2

Table 1 : Configuration settings.

Figure 1 shows the distribution of the home locations of the hospital patients of the University Hospitals Leuven (UZ-Leuven). It can be seen that most of the patients come from the region around Leuven, but smaller number of patients come from across whole Flanders. In regions where other hospitals are situated, the attraction of the University Hospitals Leuven is less.

Figure 1 : Distribution of the home TAZs of the hospital patients.

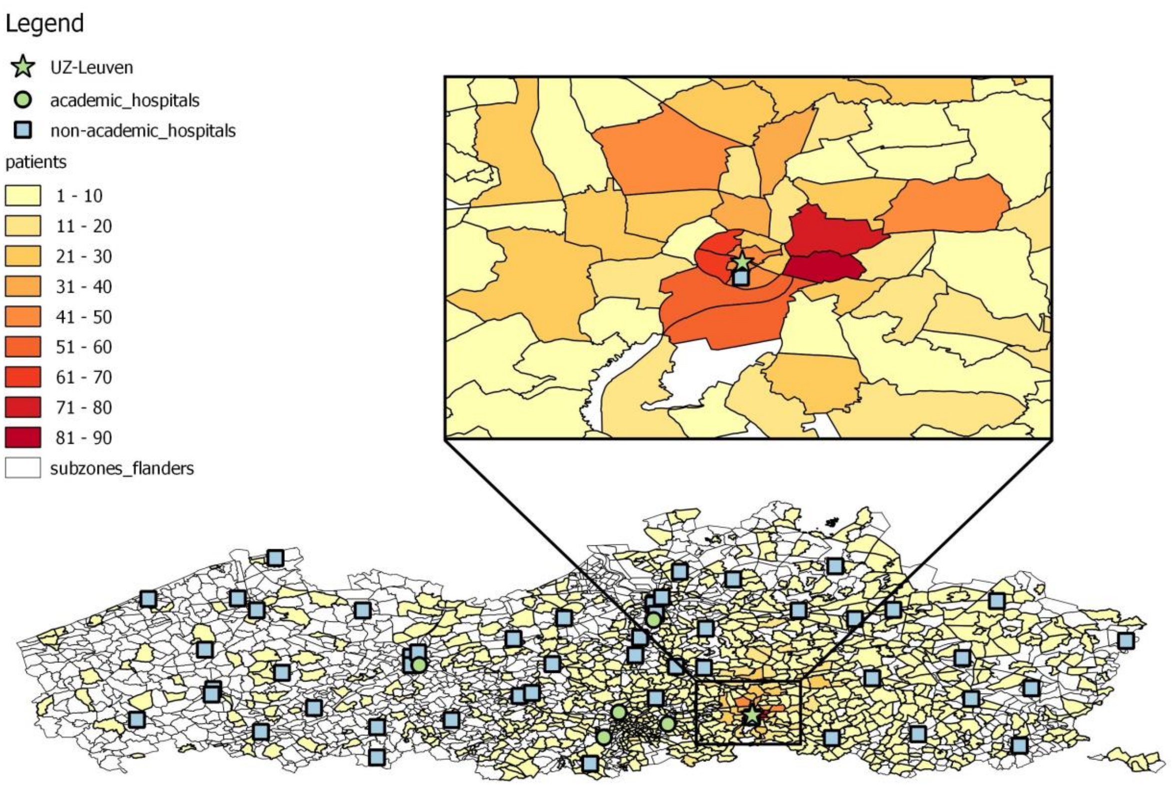
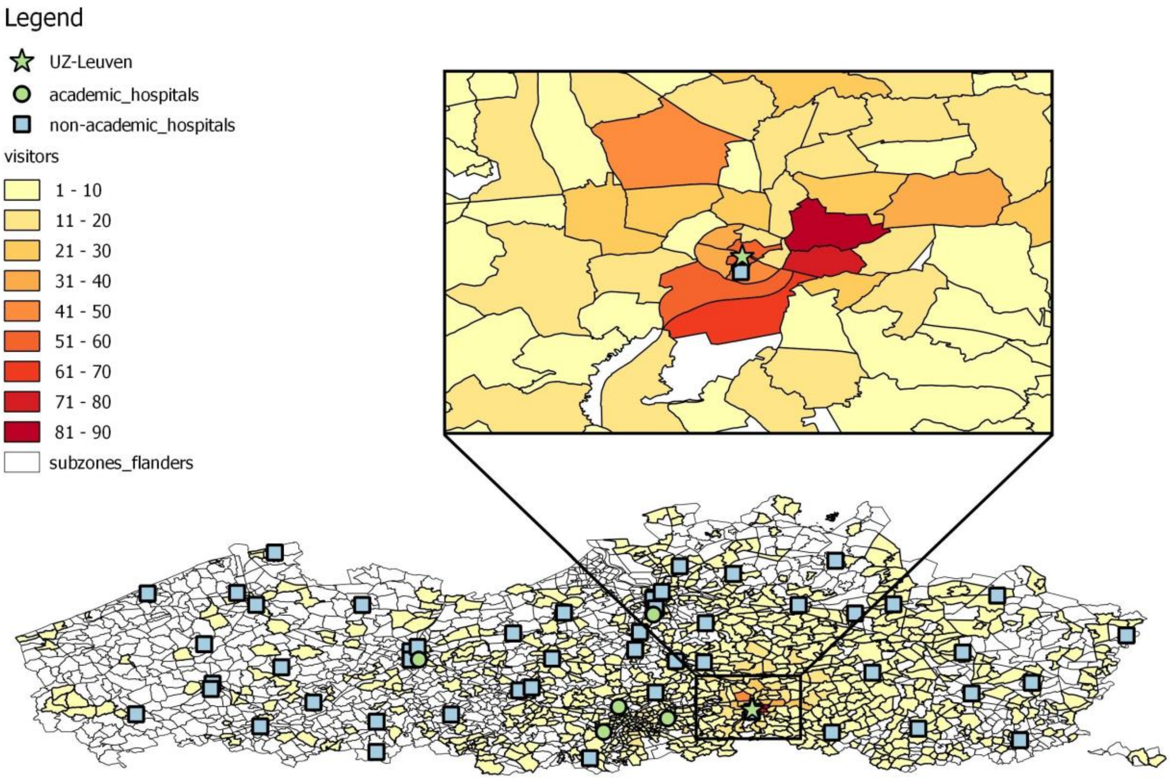


Figure 2 shows the distribution of the home locations of the hospital visitors of patients of the University Hospitals Leuven (UZ-Leuven). It can be seen that this distribution mostly corresponds with the distribution of the patients, which agrees with correlation between the squared distance and social closeness.

Figure 2 : Distribution of the home TAZs of the hospital visitors.



5. Results Validation

This section describes how to determine the optimal values for the α and β coefficients of the formulas given in Section 3.1.1. To do this we need to know the occupancy level η of each hospital. This is the number of patients (excluding consultations) going to a hospital divided by the number of beds of the hospital. For the University Hospitals Leuven this value is equal to 0.90 and we will assume that this value is identical for all other hospitals. Now we can express the total number of patients to be sampled for all hospitals by

$$N_{pat} = \sum_{h \in H} S_h \cdot \eta$$

Let P denote the set of individuals, i.e., the population. The expected value, after sampling, for the number of patients for a hospital h can then be expressed by

$$\begin{aligned} E_{pat}(h) &= \sum_{i \in P} Prob\big(h \big| p(i)\big) \cdot Prob\big(p(i)\big) \\ &= \sum_{i \in P} \frac{a_{h,i}}{\sum_{x \in H} a_{x,i}} \cdot Prob\big(p(i)\big) \\ &= \sum_{i \in P} \frac{a_{h,i}}{\sum_{x \in H} a_{x,i}} \cdot \frac{\sum_{x \in H} S_x \cdot \eta}{|P|} \end{aligned}$$

The expected number of patients for each hospital should approximate the effective number of patients. In other words, the squared error between these two values for each hospital should be minimal.

$$
\begin{aligned}
(\alpha,\beta) &= arg\ min_{\alpha,\beta} \sum_{h\in H} \left[E_{pat}(h) - S_h \cdot \eta\right]^2 \\
&= arg\ min_{\alpha,\beta} \sum_{h\in H} \left[\sum_{i\in P} \frac{a_{h,i}}{\sum_{x\in H} a_{x,i}} \cdot \frac{\sum_{x\in H} S_x \cdot \eta}{|P|} - S_h \cdot \eta\right]^2 \\
&= arg\ min_{\alpha,\beta} \sum_{h\in H} \left[\sum_{i\in P} \frac{(S_h)^{\alpha} \cdot (d_{i,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{i,x})^{-\beta}} \cdot \frac{\sum_{x\in H} S_x \cdot \eta}{|P|} - S_h \cdot \eta\right]^2
\end{aligned}
$$

In the equation above, the computation of $d_{i,h}$ can be time consuming. Therefore, observe that

$$
\sum_{i\in P} \frac{(S_h)^{\alpha} \cdot (d_{i,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{i,x})^{-\beta}} = \sum_{z\in Z} \sum_{i\in P_z} \frac{(S_h)^{\alpha} \cdot (d_{i,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{i,x})^{-\beta}}
$$

Because no individual addresses are used $d_{i,h} = d_{z(i),h}$, where $z(i)$ denotes the zone where individual i lives. Let Z denote the set of zones and let N_z denote the number of inhabitants of zone z, then it follows that

$$
\sum_{i\in P} \frac{(S_h)^{\alpha} \cdot (d_{i,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{i,x})^{-\beta}} = \sum_{z\in Z} |N_z| \frac{(S_h)^{\alpha} \cdot (d_{z,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{z,x})^{-\beta}}
$$

This results in minimizing the following equation:

$$
(\alpha,\beta) = arg\ min_{\alpha,\beta} \sum_{h\in H} \left[\sum_{z\in Z} |N_z| \frac{(S_h)^{\alpha} \cdot (d_{z,h})^{-\beta}}{\sum_{x\in H} (S_x)^{\alpha} \cdot (d_{z,x})^{-\beta}} \cdot \frac{\sum_{x\in H} S_x \cdot \eta}{|P|} - S_h \cdot \eta\right]^2
$$

Since no data is available for the distribution of the hospital-home distance, the alpha and beta values cannot be determined from the data we have available. We therefore need to determine one of the values by other means.

Figure 3 shows the squared error for $\alpha \epsilon \{0, \ldots, 3\}$ and $\beta \epsilon \{1.4, \ldots, 2.6\}$. Based on the evidence in (Lambiotte et al., 2008), (Wang et al., 2011) and (Krings et al., 2009) we assume that the optimal value for $\beta = 2$. The optimal value for α, the one that minimizes the squared error, when $\beta = 2$, approximates 1. This justifies our values for α and β in Section 4.

During the experiment of Section 4 we also collected data about how many patients go to each hospital. Figure 4 shows a scatterplot of this data in relation to the number of beds. Note that the number of patients is always higher than the number of beds, this due to consultations being included in the total number of patients. Most of the points are situated near the trend line, except for the three outliers at the top left of the plot. These three outliers each represent a hospital in the surrounding of Brussels. Looking at the surrounding of these hospitals quickly shows there are multiple smaller hospitals located in the immediate environment, which we did not account for due to lack of detailed data. Not taking these outliers into account results in the correlation depicted in Figure 5, where there are no more heavy outliers. This can be seen by the correlation coefficient which has increased from 0.8565 to 0.9695. This indicates that our values for α and β approach reality.

Figure 3 : Squared error.

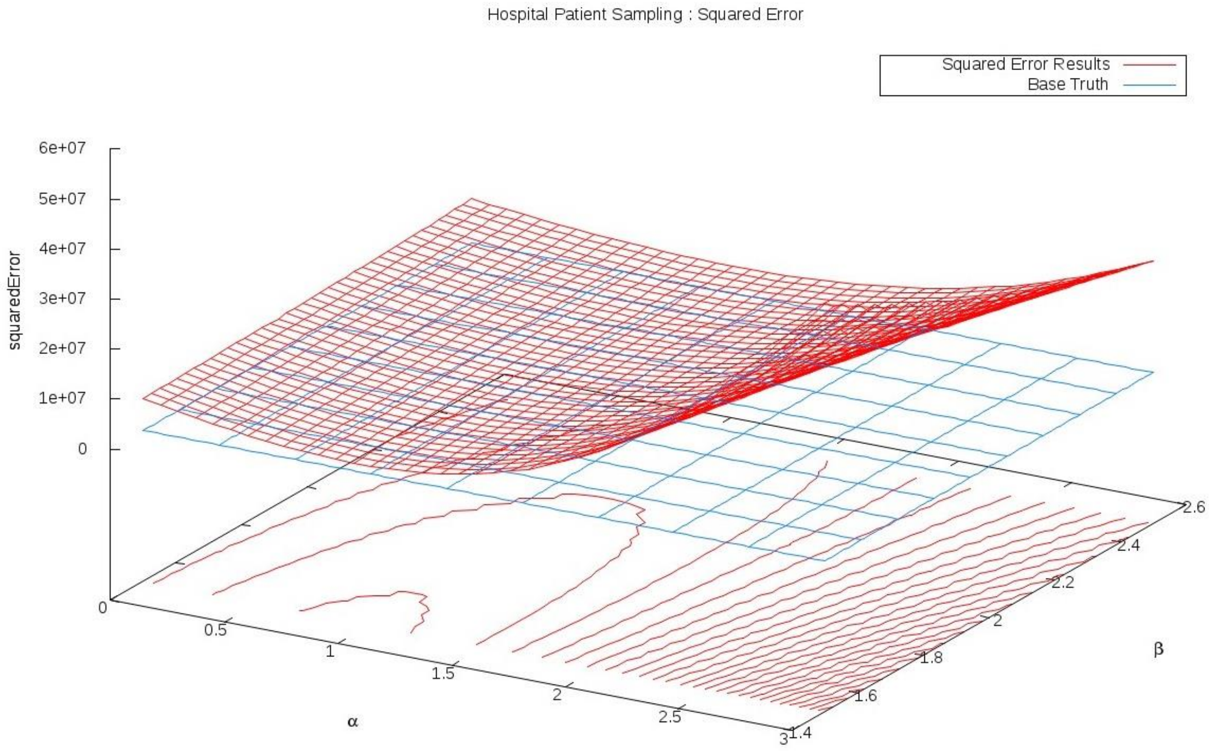


Figure 4 : Correlation between the number of beds and number of patients for all hospitals.

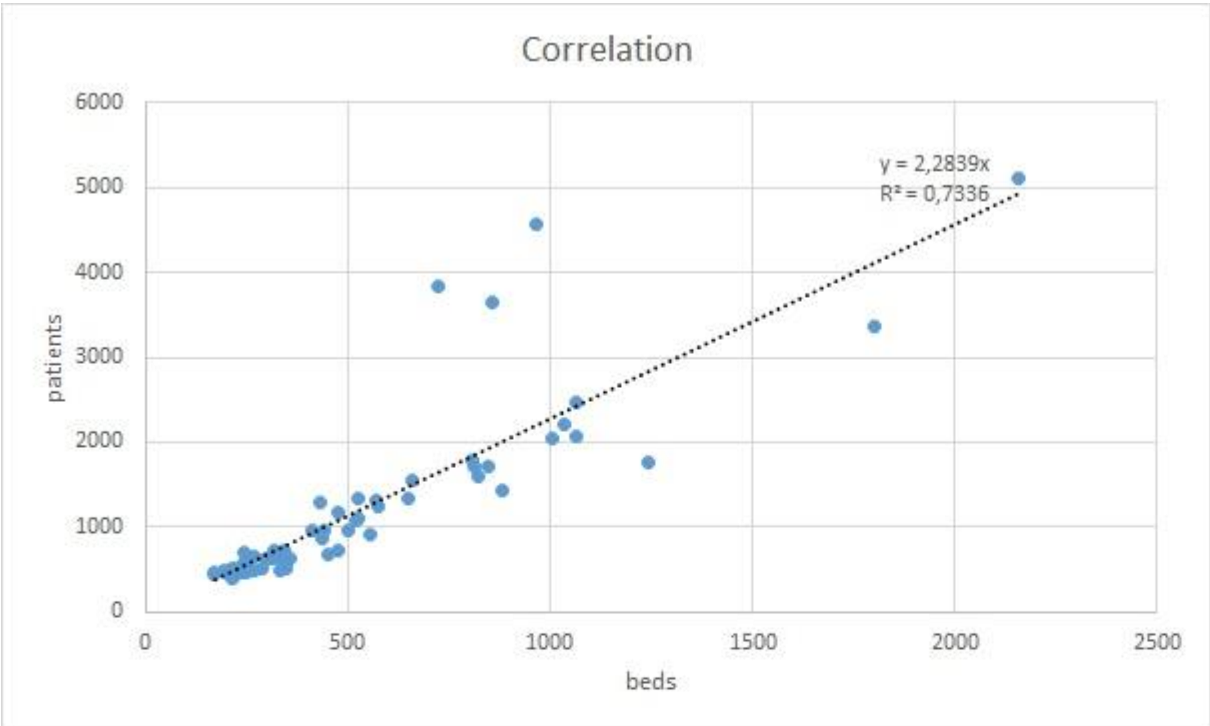
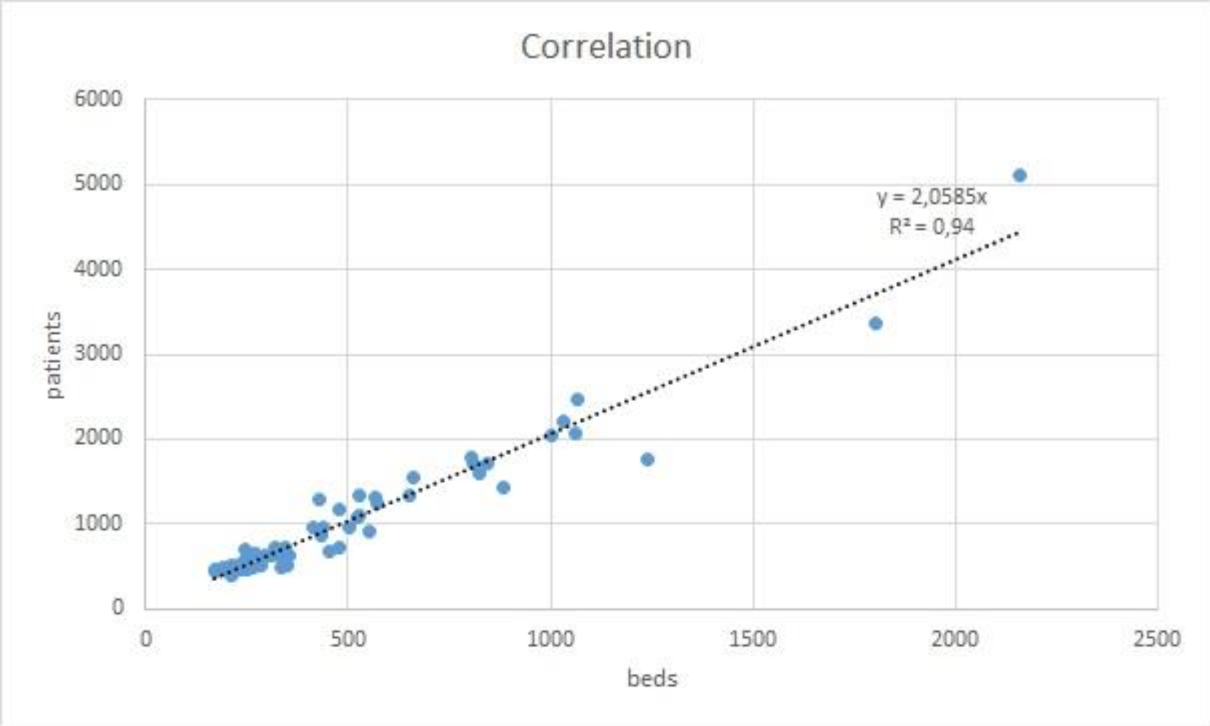


Figure 5 : Correlation between the number of beds and number of patients for all hospitals except those in Brussels.



6. Discussion

The goal of the project is to determine the demand for public transportation services at a local level (an area of 50 km by 40 km). Particular city governments and public transportation operators are interested in those estimates in order to allocate scarce resources. However, it can turn out that TAZ based modelling is not sufficient in such cases. Specific cases require micro-modelling and microsimulation. The University Hospitals Leuven are claimed to generate approximately 30k trips per day. Furthermore, it is at a distance of about 5 km from the train station at the opposite site of the historic centre of the city. The station and the hospital are connected by bus services. Evaluation of the required capacity for PT services in such context requires micro-modelling.
Therefore, we combine FEATHERS and MATSim (which is street address based) to reach a finer granularity. This means that we shall model some large sites in detail. Taking large attraction sites into account requires a large amount of specific data collection. Annual reports for the University Hospitals Leuven and studies estimating the traffic flows and parking requirements for hospitals in the US and Canada have been used as sources (Duncan and Heady, 1976). This data is not easily available, in contrast to network or public transport data, which only needs to be downloaded. One of the research goals is to find out whether this extensive data collection is worth the effort. A second research objective is to find out how to model the boundary problem, i.e., the effect of what happens in the area surrounding the study area. Finally, the experience with the specific schedule adaptation used in this project contributes to our understanding of how to build a multi-day agenda generator consisting of separate planning and scheduling components.

7. Conclusion and future work

The first stages of a project to determine the demand for public transportation in a city using micro-modelling and micro-simulation are reported. Schedules (daily plans) for synthetic individuals are generated using the FEATHERS activity-based model. Academic hospital

patients and their visitors are sampled from that population making use of data collected from annual reports for the hospital at hand and from international literature.
The next step for the hospital patients/visitors consists of schedule adaptation and PT demand generation. Modelling the hospital personnel requires generation of schedules for people working in time-shifts.
This work serves as a basis for ongoing research focusing on multi-modality, schedule adaptation and alternative collective transportation facilities that require micro-modelling due to the required level of granularity and the focus on thin flows.

8. Acknowledgement

The research reported was partially funded by the IWT 135026 Smart-PT : Smart Adaptive Public Transport (ERA-NET Transport III Flagship Call 2013 “Future Travelling").
Glenn Cich (Hasselt University) contributed to this work by technical support and results discussions.

Bibliography

Duncan M. R. and Heady E. O. *A programming model for analysis of nonmetropolitan hospital services systems and application of the model*, 1976

Knapen L., Bellemans T., Usman M., Janssens D. and Wets G., *Within day rescheduling microsimulation combined with macrosimulated traffic*, **Transportation Research Part C: Emerging Technologies**, 45, pp. 99-118, 2014

Krings G., Calabrese F., Ratti C. and Blondel V. D., *Urban Gravity: a Model for Intercity Telecommunication Flows*, **Journal of Statistical Mechanics: Theory and Experiment**, 2009

Lambiotte R., Blondel V. D., de Kerchove C., Huens E., Prieur C., Smoreda Z. and Van Dooren P., *Geographical dispersal of mobile communication networks*, **Physica A: Statistical Mechanics and its Applications**, 387, pp. 5317-5325, 2008

Rieser M. and Nagel K., *Combined agent-based simulation of private car traffic and transit*, **12th Conference of the International Association for Travel Behaviour Research (IATBR)**, 2009

Wang D., Pedreschi D., Song C., Giannotti F. and Barabási A.-L., *Human Mobility, Social Ties, and Link Prediction*, **Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 1100-1108, 2011