Transportation Research Record

A MARKOV CHAIN MONTE CARLO APPROACH FOR ESTIMATING DAILY ACTIVITY PATTERNS

--Manuscript Draft--

Full Title:	A MARKOV CHAIN MONTE CARLO APPROACH FOR ESTIMATING DAILY ACTIVITY PATTERNS
Abstract:	Determining the purpose of trips is a fundamental information to evaluate travel demand during the day and to predict longer-term impacts on the population's travel behavior. The concept of tours is the most suited to consider the value of a daily scheduling of individuals and travel interdependencies. However, the meticulous care required for both collecting data of high quality and interpret results of advanced demand models are frequently considered as major drawbacks. The objective of this study is to incorporate into a standard trip-based model some inherent concepts of activity-based models in order to enhance the representation of travel behavior. The main focus of this work is to infer, employing utility theory, the trip purpose of a population, at a zonal level. Making use of Markov Chain Monte Carlo, a set of parameters is estimated in order to retrieve tour-based components of the demand. The main advantages of this methodology are the low requirements in terms of data, as no individual information is used, and the interpretability of the model. Estimated parameters of the priors characterize a utility-based probability function for departure time, which allows to have a dynamic overview of the demand. In order to account for the tour consistency of travel decisions, an activity duration constraint is added to the model. The proposed model is applied to the region of Luxembourg city and the results show the potential of the methodologies for dividing an observed demand, based on the activity at destination.
Manuscript Classifications:	Planning and Forecasting; Traveler Behavior and Values ADB10
Manuscript Number:	
Article Type:	Publication & Presentation
Order of Authors:	Ariane Scheffer
	Claudia Bandiera
	Guido Cantelmo
	Ernesto Cipriani
	Francesco Viti

A MARKOV CHAIN MONTE CARLO APPROACH FOR ESTIMATING DAILY ACTIVITY PATTERNS

Ariane Scheffer
University of Luxembourg
2, Avenue de l'Université
L-4365 Esch-sur-Alzette - Luxembourg
Tel: (+352) 46 66 44 5485; Email: ariane.scheffer@uni.lu

Claudia Bandiera
Università "Roma Tre"
Via V. Volterra 62 – 00146 - Roma - Italy
Email: cla.bandiera@stud.uniroma3.it

Guido Cantelmo
University of Luxembourg
2, Avenue de l'Université
L-4365 Esch-sur-Alzette - Luxembourg
Tel: (+352) 46 66 44 5593 Fax: (+352) 46 66 44 35593; Email: guido.cantelmo@uni.lu

Ernesto Cipriani
Università "Roma Tre"
Via V. Volterra 62 – 00146 - Roma - Italy
Tel: (+39) 06 57 33 34 06; Fax: (+39) 329/0559286; Email: ernesto.cipriani@uniroma3.it

Francesco Viti
University of Luxembourg
2, Avenue de l'Université
L-4365 Esch-sur-Alzette - Luxembourg
Email: francesco.viti@uni.lu

Word count: 6.954 words text + 2 tables x 250 words (each) = 7.454 words

Submission Date
01/08/2018

ABSTRACT

Determining the purpose of trips is a fundamental information to evaluate travel demand during the day and to predict longer-term impacts on the population's travel behavior. The concept of tours is the most suited to consider the value of a daily scheduling of individuals and travel interdependencies. However, the meticulous care required for both collecting data of high quality and interpret results of advanced demand models are frequently considered as major drawbacks. The objective of this study is to incorporate into a standard trip-based model some inherent concepts of activity-based models in order to enhance the representation of travel behavior. The main focus of this work is to infer, employing utility theory, the trip purpose of a population, at a zonal level. Making use of Markov Chain Monte Carlo, a set of parameters is estimated in order to retrieve tour-based components of the demand. The main advantages of this methodology are the low requirements in terms of data, as no individual information is used, and the interpretability of the model. Estimated parameters of the priors characterize a utility-based probability function for departure time, which allows to have a dynamic overview of the demand. In order to account for the tour consistency of travel decisions, an activity duration constraint is added to the model. The proposed model is applied to the region of Luxembourg city and the results show the potential of the methodologies for dividing an observed demand, based on the activity at destination.

Keywords: Markov Chain Monte-Carlo, Travel Demand Estimation, Utility Theory, Trip Purpose, Tours, Activity-Based Models

INTRODUCTION

The inherent complexity of people's mobility needs has direct consequences on understanding and modelling their travel behavior. Driven by this reason, sophisticated demand models emerged during the last decades (*1*) to tackle this issue. While traditional trip-based models (TBM) currently remain widely adopted to forecast travel demand (*2*), they provide a coarse representation of the demand, which makes them inadequate for planning purposes (*3*). The main problem is that, while researchers agree that travel needs raise from the demand for activities and services (*4*), conventional TBM do not account for trip-purpose (*5*). This weakness is however offset by the ease of application and the reasonable approximation of traffic flows. Furthermore, trip-based origin-destination demand flows are the dominant input for advanced dynamic traffic assignment models (DTA), which are the most established tool for planning, optimizing and managing transportation networks (*6*).

To compensate for these limitations, the last decades have witnessed intensive research efforts in developing Activity-Based Models (ABM) and tools capable of representing individual mobility on large scale systems (*7*). Theoretically attractive, they propose an in-depth representation of the demand but tend to be harder to apply (*8*). In fact, in order to handle the linkage among various activity-travel decisions, this family of models usually rely on synthetic agents, reproducing a population usually based on a sample. When the synthetic population is well-representative of the real one and consistent, the model will provide more reliable results (*9*). That quality depends on highly precise and detailed information which is usually hard to gather both because of availability and privacy issues (*10*). Even though new methods which are not sample based appeared (*10*), this step of creating a realistic population is crucial (*11*).

The goal of this paper is to introduce some of the distinctive characteristics of ABM (*12*) within a classic TBM representation of the demand, using advanced sampling techniques. This large variety of methods has been applied since years and for diverse use in transport modelling (*13*): from synthetic population (*14*) and qualification of agents in disaggregated models (*15*) to traffic modelling (*16*) for instance . We use it here in order to refine the typical representation of the population without the burden of collecting extra data. Specifically, we show that it is possible to heed inter-dependencies between trips considering tours and inserting a utility-based departure time choice model (*17*). To do so, the global daily demand is separated into a number of functions, each of them being one component of a home-based tour.

By including daily activity patterns within a flow-based demand model, the proposed methodology enhances the representativeness of the demand and the consistency of traffic flows in time and space. Including utility-theory in the model presented in (*18*) allows to have a better representativeness of the estimated parameters and thus to refine better the information. The objective of this study is to see in which condition, adding such a meaning helps achieving better results and enhances the behavioural interpretation. A case study on a synthetic database for the city of Luxembourg is used to validate this model. We show that attracted and generated demand can be represented through tour-specific flows and that purpose-dependent macroscopic demand can be identified.

BACKGROUND

Travel demand models can be classified in two main groups, named ABM and TBM in the following of this paper. We discuss strength and limitations of both families and describe briefly how they consider trip purposes and activity chains. In addition, a glimpse to trip purpose inference in the era of big data is proposed.

By nature, the main goal of **Activity Based Models (ABM)** is to model activity-travel patterns. Timmermans at al. (*1*) distinguishes four main types of ABM, which are: (i) constraints-based, (ii) utility-maximizing, and (iii) computational process and (iv) microsimulation models. A first attempt to use utility-maximizing theory to derive tours and stops during a day for a household is proposed in (*19*). Various formulation and classes of utility have been developed since then: they may be function of time of day or function of the duration of the activity, the utility often considers the benefits gained by doing an activity but also the disutility of travelling towards it. Individuals aim at maximizing this relation (*20*).

The aforementioned utility-theory has already been put into practice in the context of **Trip-Based Models (TBM)**. Specifically, some authors showed that it is possible to include purpose specifications within a departure time choice model to obtain a stronger behavioral representativeness (*21*). After calibrating the utility-based departure time choices through sample data, some authors proposed to use this approach to model activity scheduling and trip-purposes within conventional flow-based representation of the demand (*22*). The care on activities and scheduling often settles inside a combined estimation of various travel choices for adding consistency inside flow-based models. As for the concept of tours, it allows to account for activity durations (*23*) and simultaneously model both morning and evening commute departure times, with an activity-based vision of flows (*21*). The inclusion of activities inside dynamic origin-destination (OD) matrices can also result from processing spatiotemporal information of individual. Alexander et al. (*24*) use for example mobile phone data to reconstruct purpose-dependent matrices after identifying activity type and location, based on call detail records (CDR) and using them instead to traditional travel surveys.

This application of **big data collection**, belongs to the family of OD matrix derivation. However, because most big data don't usually give information about the activity performed at the end of the trip (*25*), a lot of searches have been done to estimate activity types at destination. Many sources of information are used to this end. GPS data (*26*, *27*) which are either be collected through data loggers inside private vehicles (*28*) or taxi trajectories (*25*), automated fare collection, notably smart card (*29*) or mobile phone data (*30*) are examples of those. All of them containing rich spatial information, many methodologies are based on the trajectory analysis conceptualized by (*31*). Yet, various other information is included in order to complete the insight of the trips. Most of them identify points of interests (POIs) and link the trajectory to spatiotemporal information. Both time and duration of the stop help to distinguish an activity performed at the POI (*27*). These methodologies, even though they apply to passive collection methods, rely on many additional information which can either be included in the collection method, like fare card type (*32*) and observation frequency (*24*), or external, like household surveys (*30*, *33*), OD data and weather information (*34*).

Even if the methodologies apply to various modes of transport (taxi, public transport, private cars) they always keep a microscopic approach, focusing on agents and relying on individual's information. A related issue is that those users are not representative of the whole population and that few of their characteristics are observable (*35*).

MODEL FORMULATION

The proposed methodology leverages a Markov Chain Monte Carlo (MCMC) to calibrate a utility-based departure time choice model and derive purpose-dependent OD flows. Concretely, the flow towards and from a specific Traffic Analysis Zone (TAZ) is divided according to the activity at origin and destination, over a day, without distinguishing individual users.

Utility-based departure time choice model

We assume that the departure time choice is made according to a chain of scheduled activities for which a time and a place is preferable. Following the general framework proposed in (*20*), we define the overall utility as the sum of two components:

$$U = (U^T + U^A) \tag{1}$$

Where U is the overall utility during the reference time period (e.g. a day), U^T represents the disutility of travelling and U^A the utility of performing one or more activities "n". In this paper, we only use the positive element of this formulation, which can be calculated as

$$U^A = \sum_n U^{A,n} \tag{2}$$

Where $U^{A,n}$ is the utility of performing a certain activity n and it is usually formulated as a time-dependent function, so that utility associated to a certain time interval t can be mathematically calculated. This means that users will choose a departure time that maximizes the utility derived from the activities defined in their schedule (*17*) as in the following equation.

$$U^{A,n}(t) = \frac{\gamma_n \beta_n (U_n^{max})}{exp[\beta_n(t-\alpha_n)] + (1 + exp[-\beta_n(t-\alpha_n)])^{\gamma_n+1}} \tag{3}$$

Where $U^{A,n}(t)$ is a function of the following parameters:

- U_n^{max} : maximal utility accumulated for a determined activity;
- α_n : position on the temporal axis;
- β_n : variance around the saturation point;
- γ_n affects the position of saturation.

Figure (1) shows the influence of the four parameters of the utility function: they are the central element of the model.

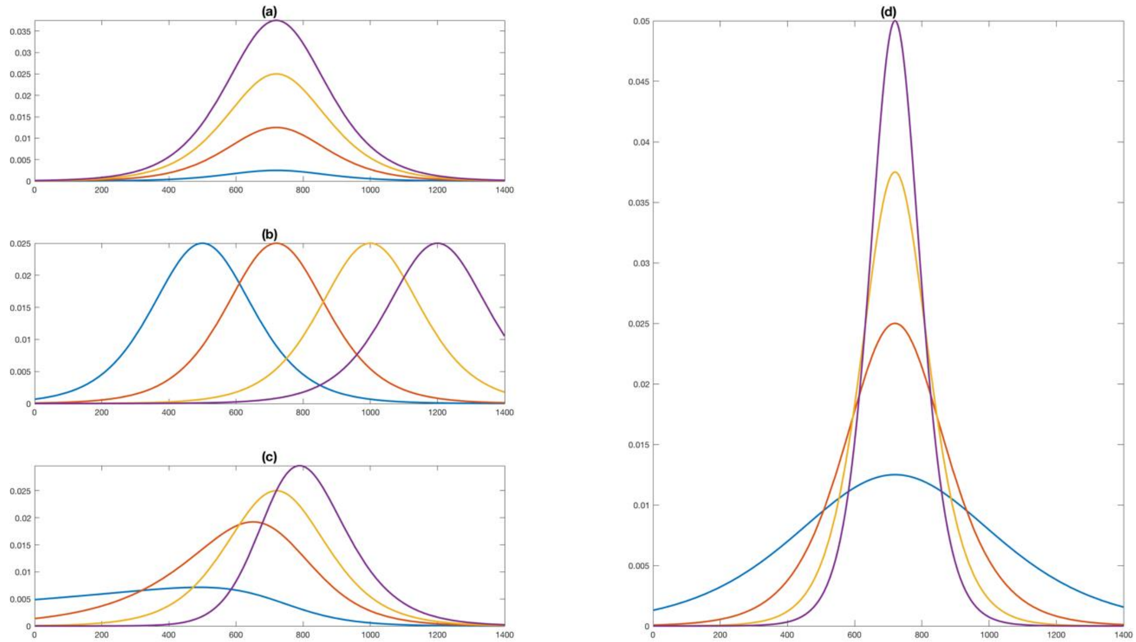


FIGURE 1 Effect of the parameter (a) U_{max} (b) alpha (c) gamma (d) beta

In the context of a tour-based estimation, the total utility is calculated according to the equation (4), where we can see that the total utility is derived from the integrals of all three curves, in function of the limits set by a pair of departure and arrival time (expressed in minutes).

$$U(t_1, t_2) = \int_0^{t_1} U_1^A(t)dt + \int_{t_1+t_t}^{t_2} U_2^A(t)dt + \int_{t_2+t_t}^{1440} U_3^A(t)dt \tag{4}$$

In order to translate the utility and the departure time choice into a probability, a multinomial Logit is used as in the following equation, where U_k is a generic marginal utility calculated for a pair of departure times.

$$P_k = \frac{exp(U_k)}{\sum_j exp(U_j)} \tag{5}$$

The MCMC - Markov Chain Monte Carlo

Given the departure time choice model, the main idea and contribution of this work is to use a Markov Chain Monte Carlo (MCMC) approach to calibrate the parameters of its utility functions without using a sample dataset. In practice, we consider that a group of curves can fully describe a tour of activities, i.e. a set of trips where the origin and the destination are located at the same place. Then, our approach exploits the MCMC model to estimate optimal parameters of these curves that best fit the observed OD flows. Without entering into details, it is important to stress inputs and assumptions behind this algorithm before introducing the activity identification step.
The first assumption of the model regards the number of activities and tours to be considered i.e. the number of probability curves considered as primitives to the complete demand. For each of these probability curves, a probability function P_k and – in case of a utility-based departure time choice model – a function $U^{A,n}(t)$ are also required. This is the first strong assumption of the

model as the result will be tied to the chosen format. The selected form can be of different type for each component. Once the shape is selected, the number of parameters to estimate can be calculated. A given distribution is controlled by a given amount of factors. Among those, some can be known and fixed, other will be the concern of the estimation. In any case a starting value is selected.

Then the link is chosen to combine these curves together. The weight of each activity can be given by an a priori proportion or by the number of users, estimated in the procedure.

The last assumption, which is of very high importance, is the prior of each parameter of interest, defined in the previous step. As the name suggests, the prior is the a priori information which describes the degree of knowledge we have about the values and our belief about the distribution. Again, this probability curve is different for each of the parameters to be estimated and its form will influence the possible variations. If the prior is very informative, e.g. when it has a narrow distribution around a specific value, the result will be very dependent to the initial knowledge. Otherwise, the end value will be influenced more by the observed data.

Indeed, the prior $\mathbb{P}(\Theta)$ is used in the Bayes formula (6) together with the likelihood $\mathbb{P}(x/\Theta)$ in order to calulate the value of each parameter for every iteration the posterior $\mathbb{P}(\Theta/x)$, based on both observed data and parameters' values.

$$\mathbb{P}(\Theta/x) = \frac{\mathbb{P}(x/\Theta).\mathbb{P}(\Theta)}{\mathbb{P}(x)} \tag{6}$$

The MCMC in practice

Once all these parameters are fixed, the goal of the MCMC is to reconstruct the probability distribution, based on event observations. At each iteration of the sampling, a new distribution is proposed. A set of variables is selected and the function obtained is used for calculating the likelihood. Then a confrontation between the current and proposed values results in the updated parameters. In this application, the evidences consist of the observed traffic flow by time of the day and the likelihood is calculated based on the aggregate output of the MCMC.

$$Likelihood = \sum \frac{-1}{2}(P_{estimated} - Demand)^2 \tag{7}$$

The complete score is in this case:

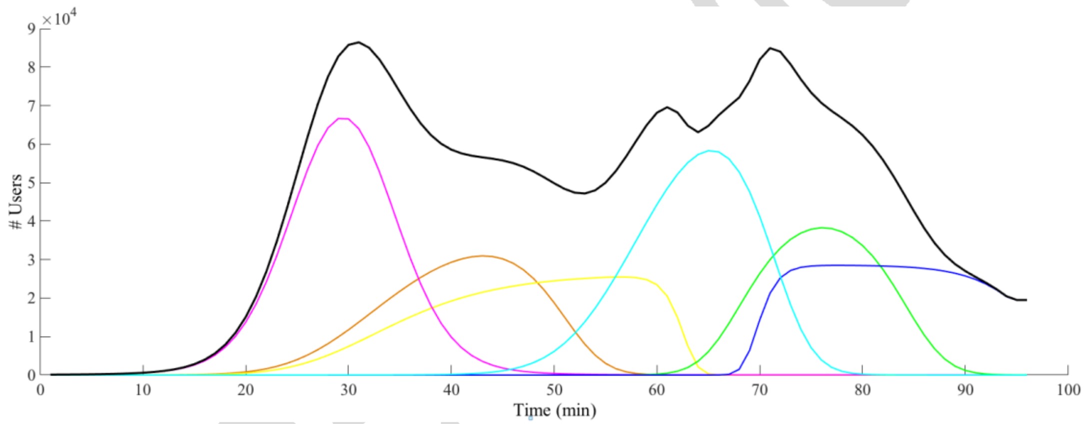
$$\begin{aligned} Score = \frac{Likelihood}{Weight} + \sum \log\big(N(\alpha)\big) + \sum \log\big(U(\beta)\big) + \sum \log\big(N(\gamma)\big) \\ + \sum \log(N(U_{max})) + \sum \log(N(Demand)) \end{aligned} \tag{8}$$

The result consists of the likelihood together with the plausibility of the selected parameters with respect to the form of their prior. It can be noted that in this formulation the likelihood is weighted. The reason we added this factor is to balance the effect of the observed data with respect to the assumptions on the different parameters. If the factor is smaller than one, it will enhance the impact of evidences, otherwise the prior will have a stronger influence on the estimation. Once this comparison is done, the proposed values are either kept and used as a starting point for the next iteration, or a new set of parameters is proposed based on the previous one. This way, at the end of the process, the algorithm outputs a distribution for each parameter, rather than converge towards a value.

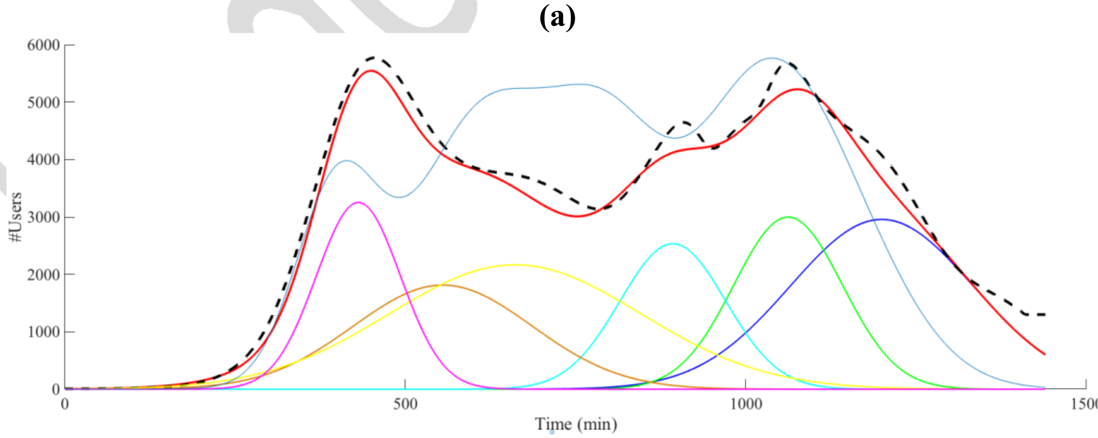
The duration of the process varies with the number of parameters to estimate. They influence the number of iterations needed before having a good approximation of the values of interest as it increases with the complexity of the target functions. Also, the initial value of each parameter and the starting function are of paramount importance to make the procedure faster.

Duration constraint

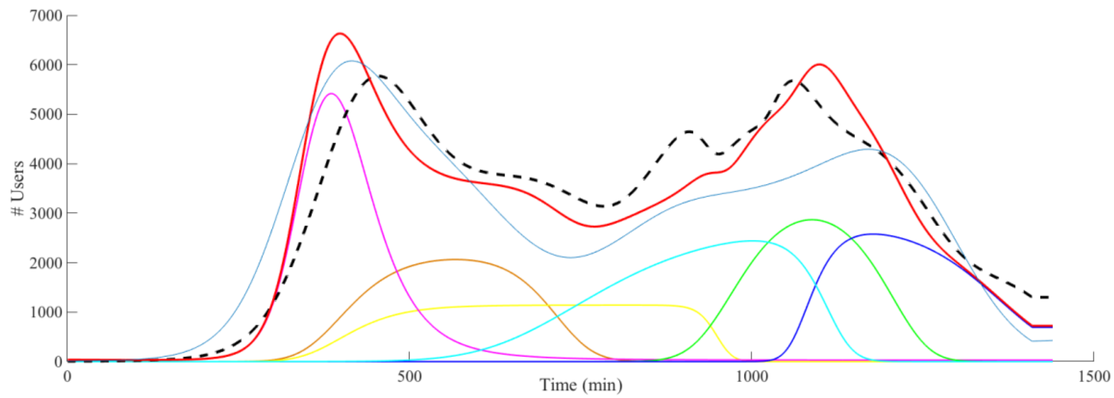
Following this procedure, we assess the possibility of using utility-based functions, their advantages and limitations. Specifically, the first problem is that utility functions have usually many parameters, meaning that the MCMC is likely to over-fit the data and provide a poor estimation of the mobility demand. In this section, we introduce a constraint that considers activity duration to reduce this problem. Another possibility could be to use simplified probability functions – such as the Gaussian distribution - that have a lower number of parameters. However, this simple distribution cannot capture complex human behavior.



(a)



(b)



(c)

FIGURE 2 (a) Reference demand (b) Gaussian decomposition (c) Utility decomposition

In order to support this point, we first compare the utility-based model and the Gaussian distribution on a synthetic aggregated demand. In order to form this reference, we considered three home-based tours: work related, maintenance i.e. not recreational personal trips, and leisure. The synthetic demand is made using the formulation described above. The assumption that the demand is formed this way creates a realistic curve and permits the evaluation of MCMC limitations in the framework of tours. Indeed, we see that on such an experiment, a Gaussian distribution still gives an overall better fitting. Figure (2) shows the reference (black), starting estimation (blue) and final estimation (red), all other curves are activity specific. We can see there that after 10 000 repetitions, the procedure seems to give good results and the output is very similar to the reference one. However, in spite of the good general representation of the demand, the Gaussian functions represent only in a simplistic way the complexity of departure time choice and participation to activities as can be seen by the six individual primitives. The two parameters of this distribution, the variance 'σ' and the mean 'μ', are only partially representative of the departure time and its dispersion. All curves have the same shape, which is extremely regular and cannot reproduce the heterogeneity of the demand. As for the second model, even though the computation time is slightly higher because of the double number of parameters, it is preferable not only because it uses inherently the utility function but also because the curves have a more realistic shape with respect to the temporal distribution of trips. Nevertheless, one can observe the deficiency of the method and an upgrade of the model is required to adequately reproduce mobility choices. We believe that removing degrees of freedom to the system is necessary to avoid considering unrealistic solutions. As a matter of fact, the two departure times of a tour are not correlated and there is no link between the estimation of the probabilities coming out from the same trip chain. It is clear that considering the curves separately is a strong weakness as it omits that duration of activities at destination influence the departure time of following trips.

In order to correlate two curves, the most straightforward solution is to insert a *minimal duration* for each activity type. It is important to stress that the proposed constraint works only as lower bound. For instance, if a minimum duration of 6 hours for activity work is considered, users can still spend a longer time without any penalty.

To implement this constraint, the departure times intervals are considered as pairs: one for going to do the activity and the other one to leave the place where the activity was performed. The joint probability of departure is still estimated through the Logit model. In this case, an even more particular care has to be given to the parameter α_n because it influences the tie between the two curves of a tour. In case of an inappropriate prior, the results can become implausible.

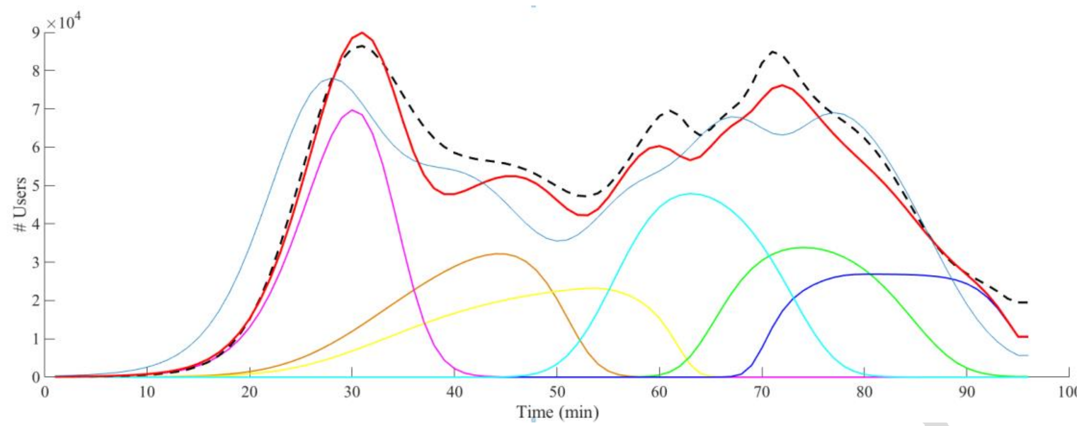


FIGURE 3 Decomposition with the duration constraint

TABLE 1 Result parameters of the synthetic experiment

	Parameters	U_n^{max}	α_n	β_n	γ_n	Demand
Tour 1	Reference					900 000
	Estimated					820 713
Home	Reference	10	250	0.01	1	
	Estimated	9.59	618	0.008	1.11	
Work	Reference	10	650	0.01	1	
	Estimated	9.83	650	0.02	1.27	
Home	Reference	10	1200	0.02	1	
	Estimated	11.01	1257	0.02	0.74	
Tour 2	Reference					700 000
	Estimated					612 943
Home	Reference	10	250	0.01	1	
	Estimated	9.59	199	0.008	1.11	
Maintenance	Reference	10	900	0.02	1	
	Estimated	10.27	864	0.02	1.13	
Home	Reference	10	1400	0.02	1	
	Estimated	9.95	1417	0.02	1.06	
Tour 3	Reference					600 000
	Estimated					599 583
Home	Reference	10	250	0.01	1	
	Estimated	9.59	199	0.008	1.11	
Leisure	Reference	10	1000	0.06	1	
	Estimated	9.17	995	0.05	0.91	
Home	Reference	10	1600	0.02	1	
	Estimated	8.19	1501	0.03	1.07	

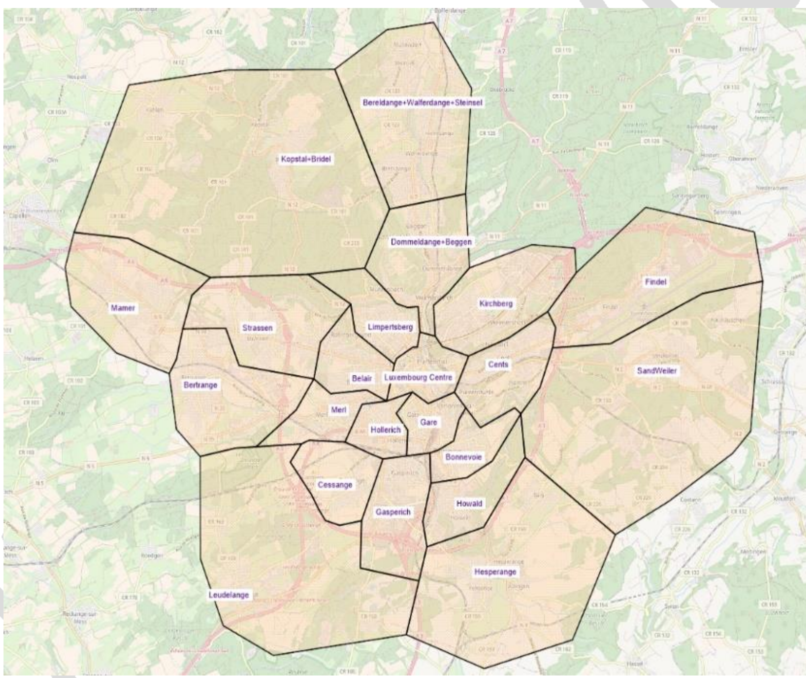
We can see here that the model gives very good results and that pairs of curves are very close to the synthetic reference. Furthermore, table (1) shows that the parameters are very well approximated. Introducing the duration constraint brings us to a new level of detail where secondary activities are well represented. The improvement in comparison to the first step is noteworthy. As it is, the model can be applied to different kind of input, and was for example tested with CDR data. However, for comparison purpose, a validation with synthetic data containing information about activity is described in the following section.

CASE STUDY (VALIDATION)

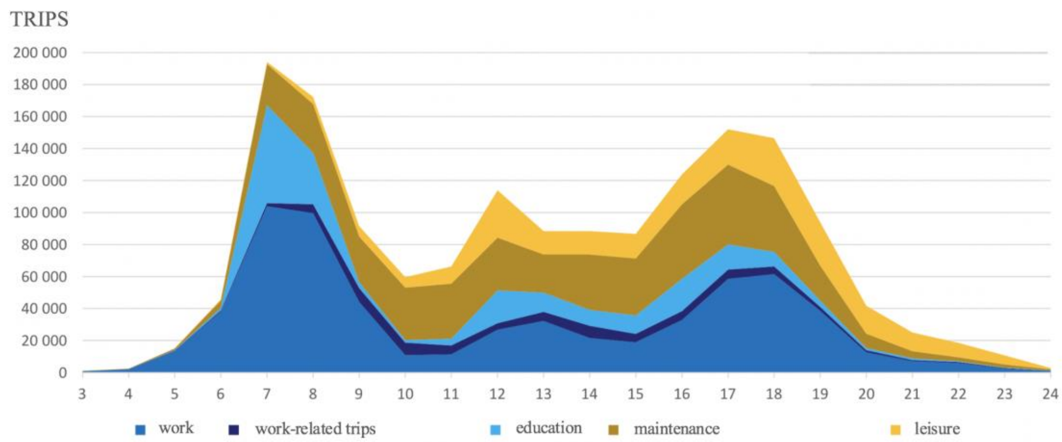
In order to evaluate the methodology with real-world data, we use a synthetic dataset produced for a Luxembourgish case study (*36*) based on a travel survey collected in Belgium in 2008 (*37*) and the ellipses methodology (*38*). The model we use as reference gives us the output of a gravity model accounting for the activities. Obviously, a hurdle in comparing these data with MCMC's output is that the gravity model provides attracted and generated trips according to purpose, while the proposed approach retrieves tours based on the overall OD flows. However, the main components are similar i.e. activities, daily demand, traffic zones.

Environment and dataset

The study area is the city of Luxembourg and its surrounding which together are divided in 22 zones (Figure 4a).



(a)



(b)

FIGURE 4 (a) Luxembourg city and zones (b) Demand profile according to "Enquête Luxmobil 2017"

As explained in the previous section, the first parameter to choose is the number of handled activities. The available database recognizes 7 activities (Home, work, school, shopping, drop off/ pick up, leisure, eat and other) which we cluster in order to obtain three groups. Assemble trip purposes together is a way to increase the number of observations by family of activities while reducing the number of necessary functions to evaluate. The drawback is that a cluster of activities typically have a less significant profile. Nevertheless, the total estimated demand being lower than the full signal, we assume the underestimation to represent these secondary activities which are harder to brand or pull away from the characteristic distribution.

In order then to have a usable signal as input of the algorithm we consider the complete dynamic OD matrix, without information about the activity type. Evidences used for evaluating the likelihood of the MCMC are composed of both the attracted and generated trips to (conversely from) a zone. We assume indeed that the destination of the first trip of a tour will be the origin of the next trip and vice versa. To take this fact into consideration, the total demand added here is halved afterwards as two trips amount to one tour and so to one individual car.
The time space considered is a full day, from 0h to 24h with a one-hour interval.

Hypotheses of the model

To take advantage of the MCMC and boost the performances of the algorithm, we need a number of hypotheses for starting the model. As mentioned before, the considered tours are only home-based and three activities are considered. A first assumption is that the complete demand is a summation of the six estimated curves. With respect to the presented methodology, some additional hypotheses need to be specified.

Prior information

A prior following a normal distribution is selected for U_n^{max}, α_n and γ_n. It means that we know the value of the parameter should be close to already known points. This precision is dulled by the variance which is also selected for each prior. In the case of β_n, the parameter can oscillate between an upper and a lower bond over the iterations. An initial set of values is carefully selected to fit all the zones. The starting point is only scaled by the total number of trips in the zone as available in the input.

The demand is the only parameter which varies from one zone to the other. The overall volume is rounded, in order to fit better both attracted and generated trips and split with respect to a priori proportions between work-related, maintenance or leisure trips. For simplicity, the same proportions are taken for every zone and values are an estimation based on a national travel survey conducted by the Luxembourgish Government in 2017 (Figure 4b). This prior also has a standard deviation, which means that values proposed during the MCMC can exceed the starting number. To avoid an overestimation of the total demand, 10% of the initial data is subtracted at the beginning of the process. The obtained value is applied for the two correlated curves of a same tour.

The duration constraint necessary for linking the two function together have been derived from the features "*popular times*" and "*visit duration*" from Google. An average of minimal typical stay in a selection of major POIs' inside the study area gives the results presented in table (2). The minimal duration for work instead derives from (*33*), as such activity is harder to qualify this way.

Algorithm parameters

In this application, the likelihood parameter reduces the influence of the observed values to avoid overfitting the data. The MCMC is able to reproduce the signal in any way, even if it means going away from the provided a priori information. In contrast, and because the proposed model offers a strong behavioral interpretation, we aim at accentuating the prior's effect.

Because the data are not as smooth as in the synthetic experiment, a higher number of iterations is required to achieve good results and we stop the MCMC after 50 000 run of the algorithm. This number on the one hand offers good approximation and before everything stable results, on the other hand it remains in an acceptable computation time. For this case study of a 24 hours signal, we estimated the four parameters of six curves in an average of 40 minutes per zone. It is important to remind that the different zones can be calculated in parallel.

The last parameter to be chosen is extremely important because it impacts the whole MCMC process. The threshold value represents the degree of acceptance of the proposed set of parameters.

TABLE 2 Parameters of the MCMC

		Work	**Maintenance**	**Leisure**
U_n^{max}	initial value	10		
	μ	10		
	σ	0.5		
α_n	initial value	$[250; 850; 1275]$	$[400; 900; 1300]$	$[300; 725; 1225]$
	μ	$[250; 850; 1275]$	$[400; 900; 1300]$	$[250; 725; 1225]$
	σ	10		
β_n	initial value	$[0.2; 0.02; 0.02]$	$[0.2; 0.06; 0.04]$	$[0.02; 0.02; 0.02]$
	Upper	$[0.2; 0.02; 0.02]$	$[0.2; 0.06; 0.04]$	$[0.02; 0.02; 0.02]$
	Lower	0.05		
γ_n	initial value	1		
	μ	1		

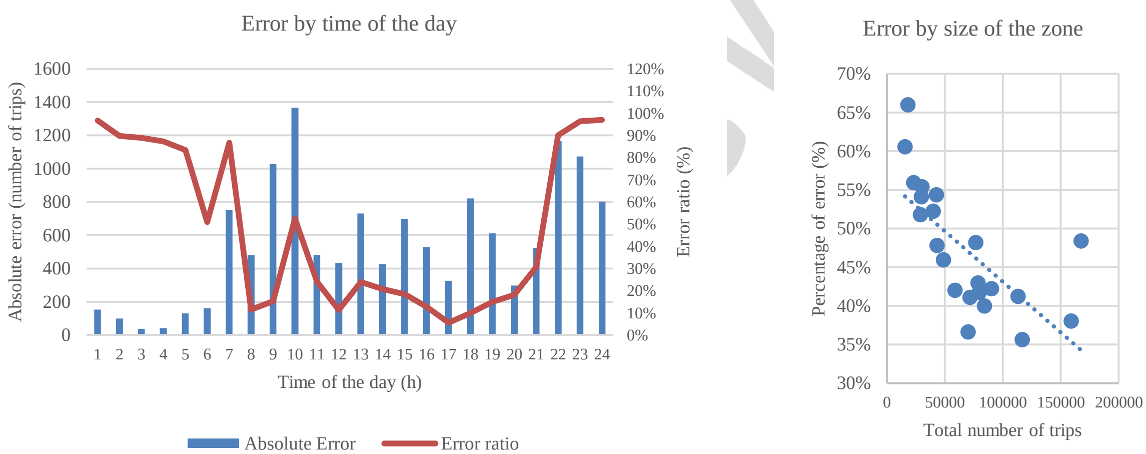
	σ	0.25		
Demand	Proportion	50%	30%	20%
	σ	10		
Minimal duration (min)		360	25	90
Likelihood factor		100 000		
Number of iterations		50 000		
Threshold		0.002		

Results

The results of the 22 zones fluctuate with the type of distribution of the demand by activity type. For the sake of simplicity, all zones were subject to the same procedure, all with the same parameters from table (2).

Dynamic estimation

The first indicator to evaluate the capacity of the methodology in reproducing the demand is to see how, from the starting curves and observed signal, the algorithm was able to reproduce the global daily demand. The easiest indicator is the difference, hour by hour, between the obtained data and the real distribution for a zone, without looking at the activity types. The following figure (Figure 5a) shows the average on all the zones of the estimation error along the day.



(a) (b)

FIGURE 5:(a) Error for all zones by time of the day (b) Error by size of the zone

We can see here that the model performs well in the afternoon but is not able to reproduce the edges of the demand. This is due to the chosen form, which does not insert a tail in the function. Indeed from 11 PM until 4 AM, almost 100% of the demand is missing. The performance for the morning peak is interesting because the MCMC can estimate extremely well the 8AM-9AM peak but not the periods just before and after. If we look closer at the results, zone by zone, it appears that these two periods are underestimated, it means that the model considers a peak which is usually too sharp with respect to the actual demand.

In the majority of other time intervals, the error is due to an overestimation for most of zones. Figure (5b), shows that the error decreases significantly with an increasing number of observations and so that MCMC is less adequate for small zones. These two considerations confirm that the model is well adapted for estimation of time periods and zones where we can observe a large number of trips.

Zonal improvement from starting point

Other parameters influence the quality of the results. For example, if the first estimation is already wrong for a zone, the resulting curve will also have the highest errors. Nevertheless, in the estimation all cases have a considerable improvement of the daily error with respect to the starting point as the following figure (Figure 6) shows for zone 1 to 22.

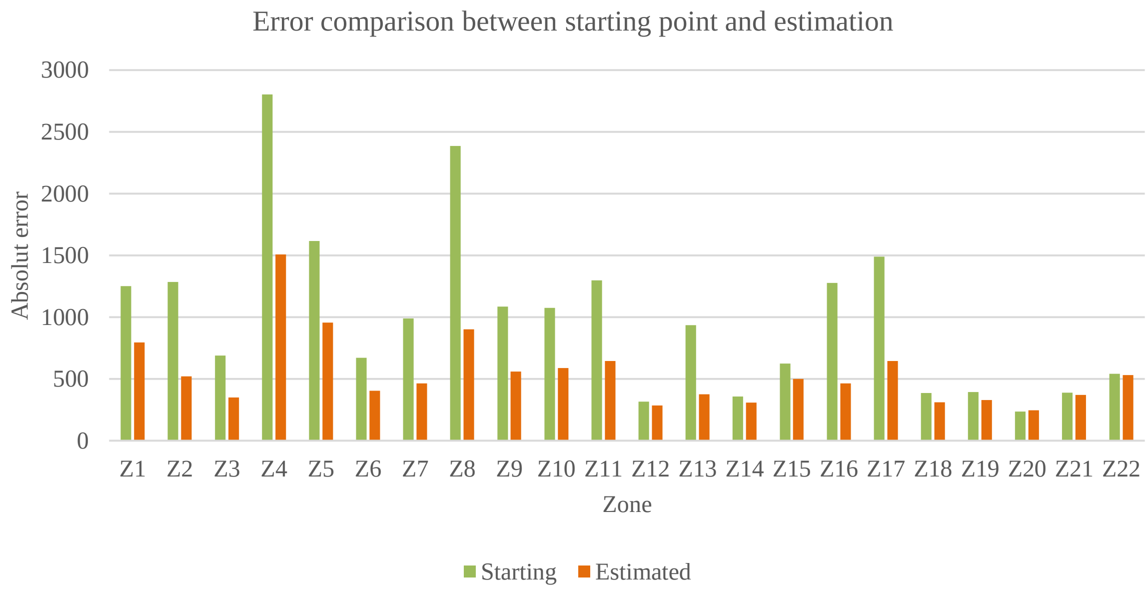


FIGURE 6 Improvement of the error between starting point and estimation

Once we validated the global ability of the model to reproduce the daily signal, we used the real data from Luxembourg to estimate, on an activity point of view if the inferred trip-purpose are rightly correlated to the data. To do so, four zones are selected as example because they offer a good overview of the different types of zones and quality of results obtained.

- Belair (Zone 2) asymmetric demand (evening peak more evident) and high number of trips;
- Cents (Zone 7) typical demand, with two peaks a hump at midday with an average number of trips;
- Findel (Zone 14) very sharp demand with three peaks and a very low number of trips;
- Bertrange (Zone 16) more atypical demand with an average number of trips;

On the following figure, we can see the real demand: the blue curve, the starting point: the black curve and the final estimation: the red curve.

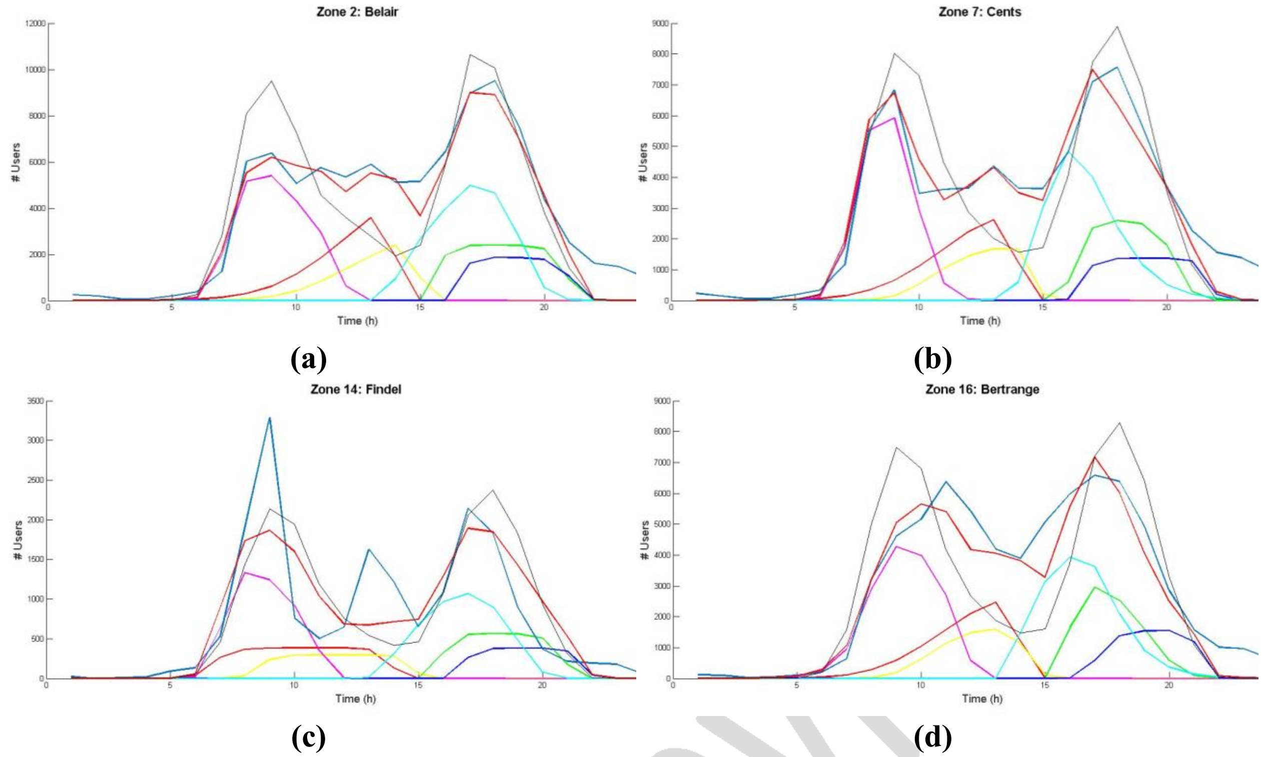


FIGURE 7 Estimated decomposition of (a) zone 2 (b) zone 7 (c) zone 14 (d) zone 16

Zones 2 and 7 have a shape which represent a good part of the 22 zones. We can see on figures (7a b) that the model was able to reproduce these, based on the probability functions as defined in the previous section. The results of zone 16 let us think that the model is flexible enough to answer to a more unusual signal. In opposition, zone 14 corroborates the weakness when too few observations are available.

As we can see, no matter the shape of the global demand, the proposed model gives parameters defining a complete curve very close to the original one. However, some of the zones do not obtain a strong improvement with respect to the initial demand. In order to go forward with the evaluation of the results, we separated the zones in two groups, based on a threshold estimated with respect to the calculated mean square error. Concretely, five zones with an improvement higher than 20% in the first curve of work-related trips are put apart. This activity is indeed typically well-represented and has a steady distinctive shape, for all zones.

Activity identification

For comparing the reference data to the output of the model, the estimated demand has been separated depending on whether the first curve is attracted and the second generated or vice versa. A likelihood was calculated for all alternatives and the highest value was selected. All eight combinations of attraction and generation are calculated and the mean square error allows to decide on the most suitable solution. Only the first curve of the tour is considered for this step and compared with the real attracted and generated data. Once the type is selected the opposite is allocated to the second curve, in order to reproduce the consistency of the tour. The following figure (Figure 8) shows the comparison of the six estimated curves with respect to the reference data for both groups of zones. We can see that the good zones have a better improvement for all of them and that most of the activities' identification ends up with the same range of error.

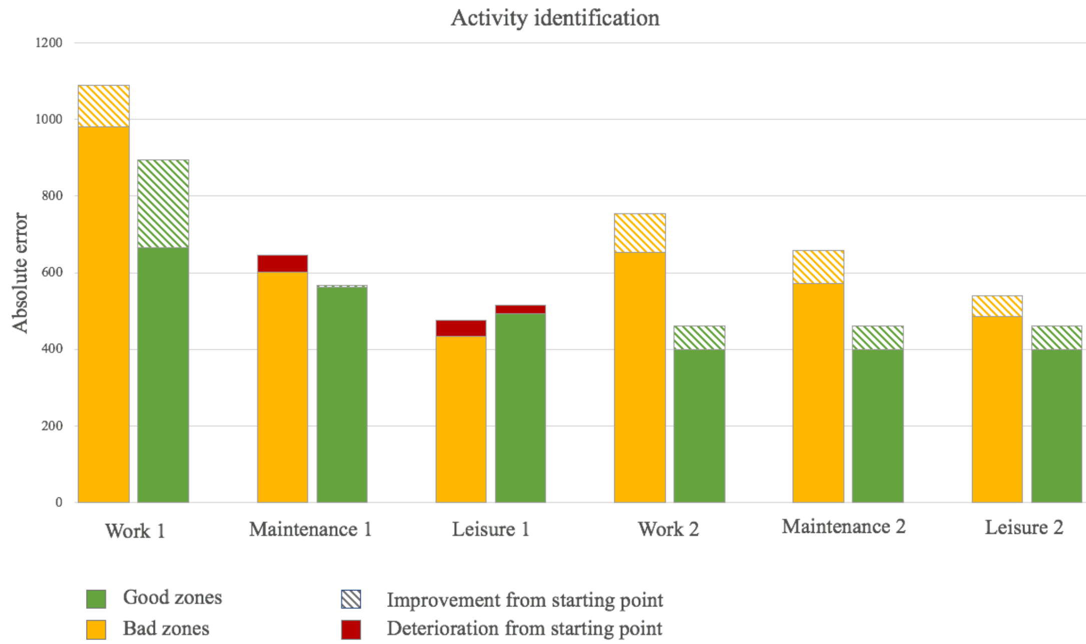


FIGURE 8 Absolute error improvement for the six curves for two groups of zones

In addition, figures (9) shows that if the MCMC is able to reproduce well the function of work then the model is also able to reduce the deterioration for activities which are less recognized, which is a strong improvement. This is usually the case when the error is already low at the starting point, i.e. when the prior is well-defined. That observation accentuates the importance that has to be given in the selection of the prior.

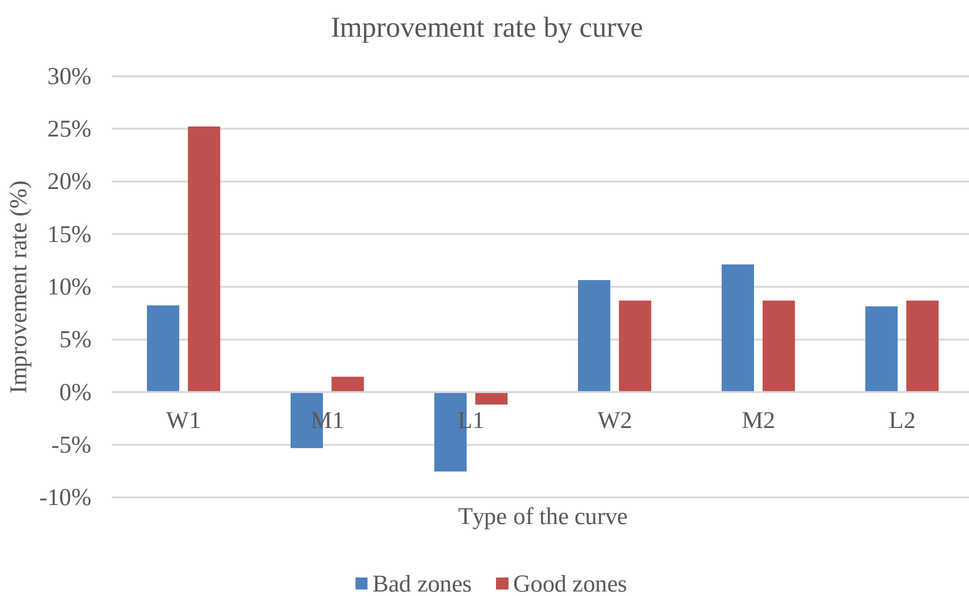


FIGURE 9 Improvement rate by curve for two groups of zones

Luxembourg City Center

One of the five aforementioned zones is the city center of Luxembourg, for which the improvement was very high for most of the activities. The example of the real demand in the area gives an insight on the estimation of the side activities.

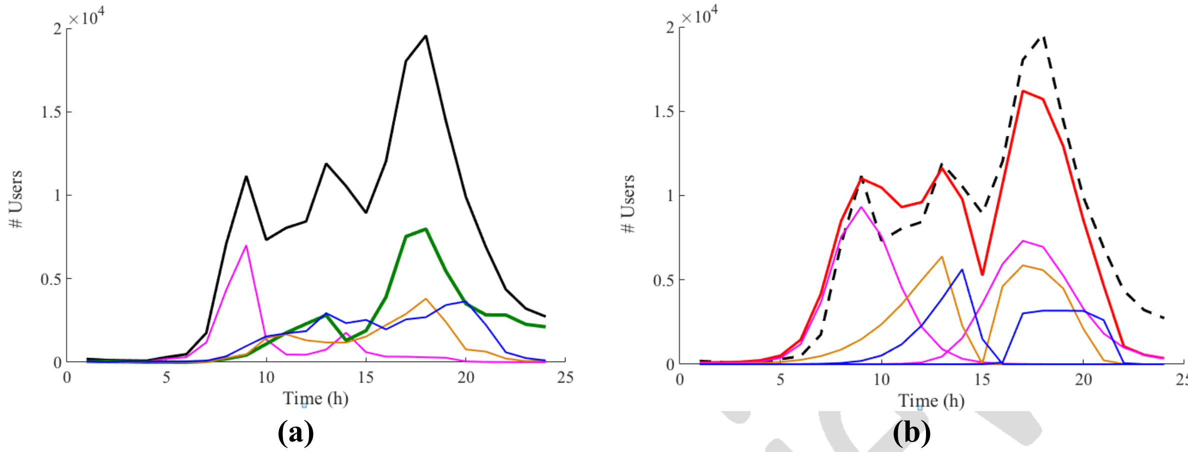


FIGURE 10 (a) Reference demand by activity type (b) Decomposition resulting from MCMC

A look to the reference data (Figure 10a) shows the obvious complexity of demand modeling for other activities than work. Indeed, a group of many "flat" functions are inconvenient for such a model. Nevertheless, if no uniform function can reproduce the base in that study, the significative peaks are identified during the MCMC and the improvement with respect to the starting point is solid. On the last figure, the pink curve represents activity work, orange is maintenance and bue leisure. The green line is in this case the generation for activity home. We can see that the results are close to the real data but mostly overestimated. This is due to the fact that we consider 100% of one tour type being either attracted or generated, and the comparison is done with the adequate portion. This unique example highlights the complexity of defining specific attributes when limiting dramatically input data.

CONCLUSION

In this paper we propose a model based on advanced sampling methods, specifically MCMC, in order to determine activity types based on traffic signal. Its specifications are based on a departure time choice model derived from the utility associated to the participation to given activities. The concept of tour is handled by the combined estimation of two curves, each of them associated to one trip, and the integration of a duration constraint. A synthetic experiment offers extremely positive results in activity identification. In order to validate the methodology with real data, the proposed model was tested on a Luxembourgish case study, with dynamic OD flows. The MCMC as defined shows interesting results for dividing an aggregated demand in activity-specific flows. Despite the complex shape of the signal, the utility-based probabilities prove to be adequate for reproducing a whole day signal. This property is more valid when the number observation is high enough. Inserting strong constraints on the probability form allows to have a better interpretation of the results. These constraints also make the model unable to reproduce distributions being away from their inherent form, like uniform demands. Indeed, results confirms the strong impact of the prior. This means that for better results, distinct sets of prior's parameters could be chosen for different zones. Nonetheless, as the probability curves are calculated with the current model, the results when combined with the actual dynamic OD matrices, can give a useful interpretation to the flows.

However, when the distributions are not typical enough or when a zone does not have a strong residential or conversely business district type for example, it is extremely hard to distinguish generated from attracted trips. It is indeed clear that a zone will both be a destination to work for a certain amount of people and the origin for another part of the population. An improvement of the model would be to evaluate a convolution of the two type of sequences for each tour and each zone. This aspect requires to have more specific data about the area or a richer input signal.

Acknowledgements
This research has been funded by the European Union (European Regional Development Fund) through FEDER (European Regional Development Fund) project MERLIN (R-AGR-3313-10-C).

Author Contribution Statement

The authors confirm the contribution to the paper as follows: study conception and design: A. Scheffer, C. Bandiera; Methodology: A. Scheffer, C. Bandiera G. Cantelmo F. Viti. Analysis and interpretation of results: A. Scheffer, C. Bandiera G. Cantelmo; draft manuscript preparation: A. Scheffer, C. Bandiera, G. Cantelmo, F. Viti, E. Cipriani. Authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Timmermans, H., T. Arentze, and C.-H. Joh. Analysing Space-Time Behaviour: New Approaches to Old Problems. *Progress in Human Geography*, Vol. 26, No. 2, 2002, pp. 175–190. https://doi.org/10.1191/0309132502ph363ra.
2. McNally, M. G. The Four-Step Model. In *Handbook of Transport Modelling*, Emerald Group Publishing Limited, pp. 35–53.
3. Lindveld, C. D. R. Dynamic O-D Matrix Estimation: A Behavioural Approach. 2003.
4. Axhausen, K. W., and T. Gärling. Activity□based Approaches to Travel Analysis: Conceptual Frameworks, Models, and Research Problems. *Transport Reviews*, Vol. 12, No. 4, 1992, pp. 323–341. https://doi.org/10.1080/01441649208716826.
5. Cantelmo, G., F. Viti, E. Cipriani, and M. Nigro. A Utility-Based Dynamic Demand Estimation Model That Explicitly Accounts for Activity Scheduling and Duration. *Transportation Research Part A: Policy and Practice*, Vol. 114, 2018, pp. 303–320. https://doi.org/10.1016/j.tra.2018.01.039.
6. Peeta, S., and A. K. Ziliaskopoulos. Foundations of Dynamic Traf®c Assignment: The Past, the Present and the Future. p. 33.
7. Balmer, M., K. Meister, M. Rieser, K. Nagel, and K. W. Axhausen. *Agent-Based Simulation of Travel Demand: Structure And . . .* 2008.
8. Moeckel, R., L. Huntsinger, and R. Donnelly. From Macro to Microscopic Trip Generation: Representing Heterogeneous Travel Behavior. *The Open Transportation Journal*, Vol. 11, No. 1, 2017, pp. 31–43. https://doi.org/10.2174/1874447801711010031.
9. Synthetic Populations: Review of the Different Approaches. *ResearchGate*.
10. Barthelemy, J., and P. L. Toint. Synthetic Population Generation Without a Sample. *Transportation Science*, Vol. 47, No. 2, 2013, pp. 266–279. https://doi.org/10.1287/trsc.1120.0408.
11. Guo, J. Y., and C. R. Bhat. Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2014, No. 1, 2007, pp. 92–101. https://doi.org/10.3141/2014-12.
12. Lin, D.-Y., N. Eluru, S. T. Waller, and C. R. Bhat. Integration of Activity-Based Modeling and Dynamic Traffic Assignment. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2076, No. 1, 2008, pp. 52–61. https://doi.org/10.3141/2076-06.
13. Calvert, S. C., H. Taale, M. Snelder, and S. P. Hoogendoorn. Application of Advanced Sampling for Efficient Probabilistic Traffic Modelling. *Transportation Research Part C: Emerging Technologies*, Vol. 49, 2014, pp. 87–102. https://doi.org/10.1016/j.trc.2014.10.013.
14. Beckman, R. J., K. A. Baggerly, and M. D. McKay. Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice*, Vol. 30, No. 6, 1996, pp. 415–429. https://doi.org/10.1016/0965-8564(96)00004-3.
15. Ben-Akiva, M., H. N. Koutsopoulos, C. Antoniou, and R. Balakrishna. Traffic Simulation with DynaMIT. In *Fundamentals of Traffic Simulation* (J. Barceló, ed.), Springer New York, New York, NY, pp. 363–398.
16. Szeto, W. Y., Y. Jiang, and A. Sumalee. A Cell-Based Model for Multi-Class Doubly Stochastic Dynamic Traffic Assignment. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 26, No. 8, 2011, pp. 595–611. https://doi.org/10.1111/j.1467-8667.2011.00717.x.
17. Ettema, D., and H. Timmermans. Modeling Departure Time Choice in the Context of Activity

Scheduling Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1831, 2003, pp. 39–46. https://doi.org/10.3141/1831-05.

18. Scheffer, A., G. Cantelmo, and F. Viti. Generating Macroscopic, Purpose-Dependent Trips through Monte Carlo Sampling Techniques. *Transportation Research Procedia*, Vol. 27, 2017, pp. 585–592. https://doi.org/10.1016/j.trpro.2017.12.111.
19. Adler, T., and M. Ben-Akiva. A Theoretical and Empirical Model of Trip Chaining Behavior. *Transportation Research Part B: Methodological*, Vol. 13, No. 3, 1979, pp. 243–257. https://doi.org/10.1016/0191-2615(79)90016-X.
20. Yamamoto, T., S. Fujii, R. Kitamura, and H. Yoshida. Analysis of Time Allocation, Departure Time, and Route Choice Behavior Under Congestion Pricing. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1725, 2000, pp. 95–101. https://doi.org/10.3141/1725-13.
21. Li, Z.-C., W. H. K. Lam, and S. C. Wong. Bottleneck Model Revisited: An Activity-Based Perspective. *Transportation Research Part B: Methodological*, Vol. 68, 2014, pp. 262–287. https://doi.org/10.1016/j.trb.2014.06.013.
22. Adnan, M. Linking Macro-Level Dynamic Network Loading Models with Scheduling of Individual's Daily Activity–Travel Pattern. In *Chapters*, Edward Elgar Publishing.
23. Zhang, X., H. Yang, H.-J. Huang, and H. M. Zhang. Integrated Scheduling of Daily Work Activities and Morning–Evening Commutes with Bottleneck Congestion. *Transportation Research Part A: Policy and Practice*, Vol. 39, No. 1, 2005, pp. 41–60. https://doi.org/10.1016/j.tra.2004.04.005.
24. Alexander, L., S. Jiang, M. Murga, and M. C. González. Origin–Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 240–250. https://doi.org/10.1016/j.trc.2015.02.018.
25. Gong, L., X. Liu, L. Wu, and Y. Liu. Inferring Trip Purposes and Uncovering Travel Patterns from Taxi Trajectory Data. *Cartography and Geographic Information Science*, Vol. 43, No. 2, 2016, pp. 103–114. https://doi.org/10.1080/15230406.2015.1014424.
26. Huang, L., Q. Li, and Y. Yue. Activity Identification from GPS Trajectories Using Spatial Temporal POIs' Attractiveness. 2010.
27. Spinsanti, L., F. Celli, and C. Renso. *Understanding People's Activities by Places Visited*.
28. Wolf, J., S. SchöUnfelder, U. Samaga, M. Oliveira, and K. Axhausen. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1870, 2004, pp. 46–54. https://doi.org/10.3141/1870-06.
29. Jun, C., and Y. Dongyuan. Estimating Smart Card Commuters Origin-Destination Distribution Based on APTS Data. *Journal of Transportation Systems Engineering and Information Technology*, Vol. 13, No. 4, 2013, pp. 47–53. https://doi.org/10.1016/S1570-6672(13)60116-6.
30. Zhou, G. The Analysis of Intercity Passenger Mode Choice Behavior Based on Probability Choice Modal. In *ICLEM 2010: Logistics for Sustained Economic Development - Infrastructure, Information, Integration - Proceedings of the 2010 International Conference of Logistics Engineering and Management*, No. 387, 2010, pp. 3161–3167.
31. Spaccapietra, S., C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A Conceptual View on Trajectories. *Data & Knowledge Engineering*, Vol. 65, No. 1, 2008, pp. 126–146. https://doi.org/10.1016/j.datak.2007.10.008.
32. Lee, S. G., and M. Hickman. Trip Purpose Inference Using Automated Fare Collection Data.

Public Transport, Vol. 6, No. 1–2, 2014, pp. 1–20. https://doi.org/10.1007/s12469-013-0077-5.

33. Erath, A., and A. Chakirov. Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport. 2012. https://doi.org/10.3929/ethz-a-007328823.
34. Alsger, A., A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman. Public Transport Trip Purpose Inference Using Smart Card Fare Data. *Transportation Research Part C: Emerging Technologies*, Vol. 87, 2018, pp. 123–137. https://doi.org/10.1016/j.trc.2017.12.016.
35. Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 285–299. https://doi.org/10.1016/j.trc.2016.04.005.
36. Carrese, F. *A Within Day Dynamic Gravity Model That Integrates Data Driven and Time Geography Approaches*. Master Thesis.
37. Castaigne, M. Behaviour and Mobility during the Week "BMW."
38. Sprumont, F., and F. Viti. The Effect of Workplace Relocation on Individuals' Activity Travel Behaviour. 2017.