Development of a Swiss LCV model
Technical report (development version)

Raphaël Ancel Andreas Justen Nicole Mathys

Federal Office for Spatial Development

September, 28th 2022

Internal notes about the current version:

- Data preparation process: state each time how many observations (number and %) were dropped.

1 Introduction

Light commercial vehicles (LCVs) are defined by their targeted usage (the carriage of goods) and by their gross weight, which must not exceed 3.5 tonnes. Their usage goes however beyond the carriage of goods. In a survey conducted in 2013 in Switzerland, 53% of the surveyed LCVs indicated "service" as one of their main usages [Federal Statistical Office, 2015a]. LCVs have gained the attention of public authorities because of their central role in city logistics and home deliveries. They are also becoming increasingly common: the annual number of LCV registrations in Switzerland has surged by 82% between 1999 and 2019 [Federal Statistical Office, 2022] and the most recent Swiss transport Outlook [Federal Office for Spatial Development, 2021a] projects a 58% increase of LCV vehicle kilometers between 2017 and 2050.

The transport modeling team within the Swiss Federal Department of the Environment, Transport, Energy and Communications aims at developing a national model for LCVs (hereafter LCV-model). The main objectives are (i) to provide a reasonable description of the current average daily traffic at the link-level for the road network and (ii) to permit the translation of long-term demographic and economic projections into traffic forecasts suitable for infrastructure planning.

The LCV-model should integrate with the existing models dedicated to the mobility of persons [Federal Office for Spatial Development, 2020b] and heavy freight [Federal Office for Spatial Development, 2015], in particular for congestion modeling. The main challenges come from the large heterogeneity of vehicle usages, the limited available data and the fact that many LCV trips are organized into tours, which should be accounted for. The envisioned approach is inspired from Hunt and Stefan [2007], which allows for a rather fine description of the touring behavior.

This paper is articulated as follows. Section 2 clarifies the scope of this paper and provides a more extensive definition of LCVs. Section 3 provides a brief review of the literature dedicated to freight and service traffic modeling. Section 4 describes the data available. Section 5 presents the global model structure and the criteria used for segmentation. Section 6 offers a more detailed description, at the module level, including the estimated numerical values and their significance. Section 7 covers additional issues related to the application, including the treatment of international traffic. Finally, Section 8 discusses the challenges related to model validation.

2 Scope

In the transport literature, transport segments and models are usually defined based on the objective of a trip (e.g. specific models for freight transport, for private mobility or for more specific purposes, for instance tourism). Here, we focus instead on a vehicle type. This is dictated by the data available. This section provides a more precise definition of the vehicle type we are interested in and of the trip purposes often associated to the use of such vehicles.

2.1 Definition of a LCV

The definition of light commercial vehicles we adopt corresponds to the homologation class N1 of the Swiss and EU classifications of vehicle types. These vehicles must be “designed to carry goods” and have a gross vehicle weight up to 3.5 tons. The Swiss ordinance on the technical requirements for road vehicles (SR 741.41) provides a more precise definition of what it means for a vehicle to be “designed to carry goods”: it “includes the vehicles equipped with additional folding seats in the load compartment for the occasional and non professional transport of persons, provided that the total number of seats, including the driver’s seat, does not exceed 9” (translated from French). Minivans like the Renault™ Kangoo™ can be classified either as passengers cars (when they have true rear seats) or as LCVs (when they have no rear seats or only folding ones).

In terms of the vehicle categories defined in the the national vehicle registre [Federal Roads Office, 2022], the adopted definition of LCVs covers the categories 30 (delivery vehicles) and 36 (light articulated vehicles) as well as the vehicles of category 38 (articulated trucks) with a gross weight up to 3.5 tons. Delivery vehicles (category 30) represent 97 % of the total.

2.2 Relevant demand segments

We expect that most of the trips carried out by LCVs correspond to business trips. Business trips can be further divided into four categories: freight, service trips with goods, service trips without goods and passenger transport services such as coaches and taxis - this typology is presented in [Federal Office for Spatial Development, 2021b] (in German). Within all business trips, the sub-segments where LCVs are the most relevant are "service with goods" and "freight". While the LCV is certainly the most important vehicle type within the sub-segment "service with goods", the "freight" sub-segment also includes many trips carried out by heavier trucks, railways, ships, and to some extent, plane. LCVs have a particular role in this segment, as they only account for a tiny share of ton-kilometers but are nonetheless very significant in terms of vehicle-kilometers traveled. LCVs may also be of relative importance in the segment of services without goods, although we expect passenger cars to be more commonly used (together with public transit and soft modes of transportation).

3 Literature review

As mentioned above, freight and service trips are rarely modeled together. Even within freight models, it is frequent to distinguish long-distance transport (where the relevant modes are generally heavy trucks, trains and cargo ships) from short-distance transport (where LCVs are more relevant). The methodological approaches have however often been directly transferred from one application to the other. We first briefly review the most common of these approaches, and then review more in detail the literature that is most relevant for the model we envision. For a broader literature review, the reader is referred to de Jong et al. [2021] for freight applications and to Ellison et al. [2017] for methods addressing service trips.

A very simple and common approach to commercial traffic modeling relies on "expanded OD-matrices". This method consists in first deriving from empirical data a base year Origin-Destination matrix and then scaling it to obtain various forecasts. Such an approach is applied for instance in the current heavy freight model of Switzerland [Federal Office for Spatial Development, 2015]. There, the base year matrix was obtained by pooling origin-destination data from vehicle-based surveys over eight years and then calibrating the resulting matrix against traffic counts for a specific year. This method provides robust forecasts that can be used for "business as usual" scenarios for infrastructure planning at a large scale (e.g. national or international), when the expected demand variations remain small. It is however not ideal to model structural changes or the effect of new policy instruments.

Another common approach consists in adjusting the traditional four-step model of travel demand. The first two steps (generation and distribution) and the last step (network assignment) are almost always handled at an aggregate scale, as when modeling the movements of people. In the context of commercial vehicles, the third step (originally mode choice) should be considered together with other logistics choices such as shipment size/frequency (for the transport of goods only), trip-chaining and departure time. These choices are more naturally handled at the level of individual establishments. Many freight models simply do not include these logistics choices - e.g. Cambridge Systematics and

COMSIS Corporation [1996]. The commercial software PTV-VISUM, which is already used in Switzerland for modelling the mobility of persons, provides with its "TB-Freight" module a partial solution. Mode choice is not modelled but tours are described in an aggregate fashion : the producer-consumer matrix is translated into a trip matrix which satisfies the demand and the vehicle-balancing constraints, while allowing to control the average number of stops per tour [PTV GROUP, 2019]. There are however many matrices that satisfy these constraints and the one produced by TB-Freight is not necessarily very realistic.[1]

Another approach in commercial traffic modeling consists in adopting a microscopic (i.e. agent-based) view to describe logistics choices - see de Jong et al. [2021] for a review of the latest developments. The most well established approach is probably the Aggregate-Disaggregate-Aggregate (ADA) freight model system [de Jong and Ben-Akiva, 2007], which focuses on freight and has been applied to several countries (Norway, Sweden, Denmark and ongoing work in Austria). It describes the choice of shipment size/frequency, of the number of legs in the transport chain and of mode and vehicle type for each leg. Yet, with a high level of detail come large data needs. The ADA approach requires a commodity flow survey to be available, describing the full trip chain. Such surveys can be expansive to conduct and are not very common. The application of ADA in Austria required such a survey to be conducted specifically for the needs of model development - see Grebe et al. [2020]. There is no such survey available in Switzerland.

To avoid the difficulties associated with the explicit consideration of shipments, Hunt and Stefan [2007] proposed a model of tours which is purely vehicle-based. This type of approach seems to be the most suitable in the Swiss case, given the data available. This model is microscopic in that it considers individual vehicles, but it is also macroscopic in that it considers only zones, and not individual establishments. Every zone is the base for a number of tours. These tours are associated to various purposes and vehicle types. The total number of tours is estimated based on land-use and economic characteristics of the zone via a regression analysis. Stops are then generated one by one, via an iterative procedure. After every stop, a module called *Next Stop Purpose* determines whether the vehicle goes back to its establishment or continues its tour. If another stop should be made, the module *Next Stop Location* determines the next destination. Yet another logit-model determines the stop duration. A schema of this modular structure is provided in Section 5.2.

We would like to highlight at this point an important difference between the approach of Hunt and Stefan [2007] and those based on the four-step model: the *Next Stop Location* module does not require any pre-estimated demand matrix (which would normally be the result of the generation and distribution steps). Instead, the destinations are chosen based on a behavioral model.

Some subsequent models [Thoen et al., 2020, Sakai et al., 2020] proposed

[1] One criticism is that TB-Freight tends to produce a large number of internal trips in isolated zones, thereby implying that the few tours serving these zones have an extremely large number of stops.

alternative formulations which explicitly describe shipments. They take as input a set of shipments (which are either real, or generated from another submodule), and generate tours compatible with these shipments. The data needs of such an approach are however far beyond what is available in Switzerland for the whole country. Indeed, these approaches require the knowledge of not only the origin and destination zones, but also of the schedule constraints and of the companies responsible for each shipment.

4 Data

4.1 LCV survey

The main data source is a vehicle-based survey conducted in 2013 in Switzerland [Federal Statistical Office, 2015a]. The survey was conducted with two types of questionnaires. Questionnaires of type 1 only contained a few questions about the mileage and the main use of the vehicle during a reference day. It is useful for calibration purposes and to estimate the total vehicle kilometers travelled by LCVs. Questionnaires of type 2 contained the same questions as those of type 1, but also questions about the individual shipments transported during that day (postal codes of origin and destination, weight, type of good, mileage of the vehicle at the origin and destination). Empty legs also had to be registered, thereby allowing for the reconstruction of all vehicle movements. This second part however was only required for vehicles that carried at least 50 kg of goods during the reference day. Overall, according to the survey report [Federal Statistical Office, 2015b]:[2]

- Questionnaires of type 1 were sent to the owners of 40'000 LCVs. The response rate was 76%.
- Questionnaires of type 2 were sent to the owners of 28'000 LCVs. The response rate was 71%, but the part with questions about individual shipments was only completed for 3'874 LCVs.

Besides, to limit the respondent burden with questionnaires of type 2, a simplified form was proposed for tours containing at least four different locations distant from each other from at most 20 km with shipments of a single type. This simplified form does not contain the information for every single shipment, but only the overall weight and the mileage at the first and last stop, as well as at the origin.

The 3'874 questionnaires with detailed data correspond to 1.2% of the entire LCV fleet. These surveys contain detailed information about 11'473 individual shipments, and summarizing information (i.e. simplified forms) for 830 tours.

[2]These were the planned numbers. Light variations in the implementation.

4.2 National vehicle register

The second most important data source is the Swiss national vehicle register [Federal Roads Office, 2022]. For this project we used the standard dataset "BEST_R". This dataset contains various pieces of information for each vehicle registered in Switzerland. The attributes that are most relevant for our application are the type of owner (private or business), the vehicle category (e.g. 30, 36 and 38 - see Section 2.1) , the curb weight, the gross weight and the postal code of where the vehicle is registered.

4.3 Additional explanatory variables

Although a large variety of land-use and socio-economic variables is available, the model mostly relies on the number of jobs (full-time equivalents) per branch and the number of inhabitants. This data is available for any geographic partitioning for the current state and it is also a common product of socio-economic forecasts, thereby allowing the model to be applied to establish forecasts. Travel time and distance matrices estimated using the national model for passenger transport [Federal Office for Spatial Development, 2020b] are also available.

5 Modeling considerations

5.1 Zones

The model estimation and application rely on two different zoning systems. For model estimation, we relied on the postal codes (state 2013), which is the most precise geographical information contained in the LCV survey. These correspond to a partition of Switzerland in 3'187 geographical areas. For the model application however, we use the zones of the national model for passenger transport [Federal Office for Spatial Development, 2020b] (about 8'000 zones). This choice simplifies the integration of the results and the network assignment.

Note that some postal codes indicated in the LCV survey do not correspond to geographical areas, but to specific large organisations - for instance the federal government. These organizations might have various locations, which are sometimes far from each other but associated to a single postal box. As we could not obtain explanatory variables of sufficiently good quality for trips involving these postal codes, the corresponding survey entries were discarded (see Appendix A.2 for more details).

5.2 Model structure

The adopted model structure is illustrated in Fig. 1, together with the one originally proposed by Hunt and Stefan [2007]. Our structure consists of five modules:

- *Vehicle Generation*: This module provides an estimate of the number of LCVs per zone, per vehicle type (curbweight), owner type (private/business)

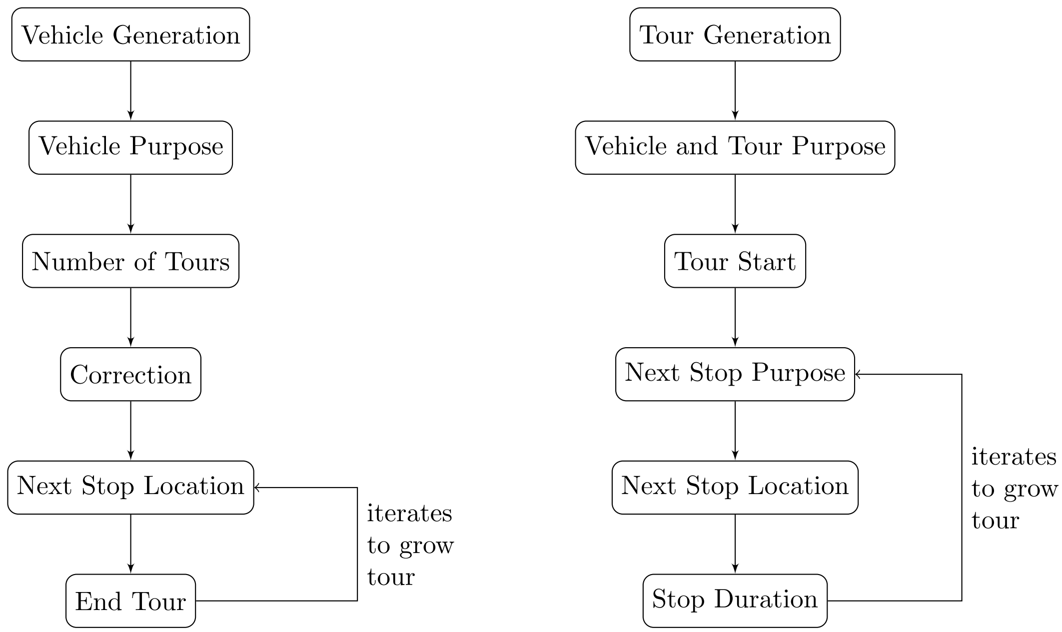

Figure 1: Two model structures: the one adopted in this paper (left), and the one proposed by Hunt and Stefan [2007] (right).

and in the case of a business owner, by economic branch (see Section 5.3 for more information about these segments).

- *Vehicle Purpose:* This module assigns a purpose (*Idle*, *Goods transport*, *Service* or *Other*) to each vehicle.
- *Number of Tours:* This module assigns a number of tours to each vehicle.
- *Correction:* This module applies a correcting factor to account for the difference in daily distance travelled between vehicles making tours with few stops (fully documented in the data) and vehicles making tours with many stops (not fully documented, and therefore not modelled).
- *Next Stop Location:* This module determines which zone must be visited next.
- *End Tour:* This module determines whether the vehicle makes an additional intermediate stop or goes back to its establishment.

The structure we propose differ from the one proposed by Hunt and Stefan [2007] in various regards. First, the module *Tour Generation* of Hunt and Stefan [2007] is replaced by a module *Vehicle Generation* and a module *Number of Tours*, to better correspond to the data available. Indeed, we have no data about the number of tours originating in each zone, but the national vehicle register tells us how many LCVs are registered in each zone and the LCV survey allows us to estimate the number of tours made by each vehicle. Second, we do not model the departure time and stop duration (no module *Tour Start* and *Stop Duration*). Third, the module *Vehicle and Tour Purpose* of Hunt and Stefan [2007] is replaced by a module *Vehicle Purpose*. This is because we only focus on LCVs (no vehicle choice) and because in our data, the purpose is a characteristic of the vehicle for the entire reference day (it is not tour-, or stop-specific). Fourth, the module *Next Stop Purpose* of Hunt and Stefan [2007] was renamed *End Tour*. In Hunt and Stefan [2007], the outcome of the module could be "Goods stop", "Service stop", "Other stop" and "Return to establishment". In our case, stops do not have purposes, so the outcome can only be "Return to establishment" or "New stop". Fifth, the added module *Correction* is necessary here to account for vehicles making many stops (such as post services). Such vehicles are not handled by Hunt and Stefan [2007] but by a separate model described in Stefan et al. [2007].

5.3 Criteria for demand segmentation

The demand segmentation is not the same for each module. It depends on the amount of data available, on the type of model estimated, and on the position of the module in the overall model structure. The vehicle purpose for instance is only determined after the module *Vehicle Purpose*, so it cannot be used in the module *Vehicle Generation*. This section presents all the criteria that come in use for the demand segmentation in the various modules.

- Owner type: private / business.
- NOGA code: indicates the economic branch for business owners. We use only the first level of the Swiss General Classification of Economic Activities (NOGA) - see Table 1. Note however that this information is only available for 58.6 % of the observations of the LCV survey (the remaining 41.4 % consist of private owners and of business owners whose branch could not be identified - for example because several companies were registered at the same address).
- Curb Weight: we only use two categories (up to 2 tons and above).
- Vehicle purpose: The respondents had many options in the questionnaire ("Goods transport", "service", "passenger transport", "rental", "measurements", "private use" and "other"). They were also allowed to select multiple answers. To avoid over-segmentation, we defined the following 3 segments:
  - *Goods Transport*: if the set of main purposes indicated by the respondent contains "Goods Transport" and does not contain "Service".
  - *Service*: if the set of main purposes indicated by the respondent contains "Service".

Table 1: First level of the NOGA classification (source: `https://www.kubb-tool.bfs.admin.ch/en`)

A	agriculture, forestry and fishing
B	mining and quarrying
C	manufacturing
D	electricity, gas, steam and air-conditioning supply
E	water supply; sewerage, waste management and remediation activities
F	construction
G	wholesale and retail trade; repair of motor vehicles and motorcycles
H	transportation and storage
I	accommodation and food service activities
J	information and communication
K	financial and insurance activities
L	real estate activities
M	professional, scientific and technical activities
N	administrative and support service activities
O	public administration and defence; compulsory social security
P	education
Q	human health and social work activities
R	arts, entertainment and recreation
S	other service activities
T	activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
U	activities of extraterritorial organisations and bodies

- *Other*: this category contains all the observations which did not fall in the first two categories.

Note that this is a "vehicle purpose", which remains the same for the entire day. It is not trip-specific.

5.4 Overview

Table 2 provides an overview of the segments, methods and outputs of the different modules.

Table 2: Overview of the different modules

Module	Method	Explanatory variables	Segmentation
Vehicle Generation	linear regression + distribution per branch	population and jobs	by curb weight and owner type
Vehicle Purpose	average rates	none	by curb weight, owner type and branch (if business)
Number of Tours	average rates	none	by curb weight, owner type, branch (if business) and purpose
Correction	average rates	none	by curb weight, owner type, branch (if business)
Next Stop Location	multinomial logit	population, jobs, travel times, land-use	by purpose
End Tour	binomial logit	number of stops, travel times, branch	by purpose (incl. branch-specific dummy variables)

6 Model estimation

6.1 Vehicle Generation

6.1.1 Current implementation

The purpose of this module is to convert a number of jobs (disaggregated by branch) and a number of inhabitant per zone into a number of registered vehicles, disaggregated by owner type and by branch for business-owned vehicles.

The segmentation criteria are:

- curbweight: up to or above 2 tons,
- owner type: business or private.

This results in four segments. The number of vehicles in each of them is indicated in Table 3.

Table 3: Number of light good vehicles registered by segment (state: June, 2022), source: National Vehicle Registre (ASTRA)

Segment	Owner	Curb weight	Vehicles	Share of LCVs
1	private	≤ 2 tons	57'886	14%
2	private	> 2 tons	51'560	13%
3	business	≤ 2 tons	136'036	33%
4	business	> 2 tons	165'899	40%

The empirical data we have come from the national vehicle register. It distinguishes the vehicles owned by organizations from those owned by private persons, but it does not include any information about the branch of business owners. To circumvent this limitation, we estimate branch-specific rates with a model. We use an ordinary least-square linear regression, with no constant term. The estimation is done at the level of postal codes.

The explanatory variables depend on the owner type. For the segments associated to private owners (1 and 2), we use only one explanatory variable, the number of residents. For the segments associated to business owners (segments 3 and 4), we use the number of jobs (Full-Time Equivalents) in various NOGA branches.

To avoid overfitting, we tried to limit the number of branches considered as potential explanatory variables. The distribution of vehicles between branches observed in the LCV Survey can help in this regard, even though the sampling scheme was not specifically designed to provide a representative proportion of each branch (see Table 4, all columns but the last one).

After excluding the 41.4 % of vehicles which are either privately owned or could not be associated to a branch, the branch that appears the most often in the LCV survey is by far the branch F (construction). In second and third position come the branches C (manufacturing) and G (wholesale and retail trade;

NOGA	Prop. of vehicles	Prop. excl. (empty)	Prop. model
A	1.1%	1.9%	16.5%
B	0.1%	0.2%	0.5%
C	8.8%	15.0%	6.0%
D	1.1%	1.8%	0.5%
E	0.6%	1.1%	1.7%
F	24.0%	41.0%	61.2%
G	8.6%	14.7%	10.3%
H	2.9%	4.9%	0.7%
I	0.7%	1.3%	0.0%
J	0.4%	0.6%	0.0%
K	0.8%	1.4%	0.0%
L	0.5%	0.8%	0.0%
M	1.7%	2.9%	0.0%
N	4.7%	8.0%	2.7%
O	1.1%	1.9%	0.0%
P	0.2%	0.3%	0.0%
Q	0.6%	1.0%	0.0%
R	0.2%	0.4%	0.0%
S	0.5%	0.8%	0.0%
T	0.0%	0.0%	0.0%
U	0.0%	0.0%	0.0%
#N/A	41.4%		

Table 4: Distribution of vehicles between branches in the LCV Survey and based on the estimated models for business-owned LCVs

repair of motor vehicles and motorcycles). Then come the branch N (administrative and support service activities) and H (transportation and storage). Together, these five branches represent 83.6% of the vehicles associated to a branch.

Based on this first analysis and on our intuition, we first tried a linear regression where the jobs in the branches I to M and O to S were all aggregated into one explanatory variable, while the other branches were taken separately. This regression led to $R^2 = 0.95$ for business-owned LCVs with a curb weight up to 2 tons and to $R^2 = 0.94$ for heavier business-owned LCVs. Two explanatory variables however had negative coefficients for each regression: the branch H (resp. the branch N) and the aggregate of I-M and O-S for light (resp. heavy) LCVs.

Because we want to be able to interpret the obtained coefficients as rates, expressed in vehicles per FTE, we can not work with negative coefficients. We thus tried new specifications, without using the jobs in these branches as explanatory variables. The results are summarized in Table 5. As all coefficients were positive, we kept these rates. The branches generating the more vehicles per employee are by far F (construction) and the branch E (water supply, sewerage, waste management and remediation activities). Although some variables had a relatively weak statistical significance, we chose to keep them in the model, because our objective is to identify a rate of vehicles/FTE for each branch. The estimates we obtain are admittedly uncertain, but they are still probably better than assuming a rate of 0 (which is the case for variables that are omitted).

We then multiplied these rate by the number of jobs (FTE) in each of these branches, to obtain a number of vehicles per branch and compare the fleet composition with the one observed in the LCV survey. The resulting distribution of business-owned vehicles between branches is described by the last column of Table 4. The proportion of vehicles registered in the branch F (construction) is even larger with the estimated model. Then come the branch A (agriculture, forestry and fishing) with 16 % of the fleet (and only 1.9 % in the LCV survey), the branch G and the branch C.

Table 5: Estimation results for *Vehicle Generation*

	Segments			
	1	2	3	4
A	-	-	0.102	0.341
B	-	-	0.022	0.282
C	-	-	0.014	0.016
D	-	-	0.023	0.034
E	-	-	0.111	0.201
F	-	-	0.271	0.278
G	-	-	0.021	0.038
H	-	-	-	0.011
N	-	-	0.037	-
population	0.0043	0.0033	-	-
Observations	3179	3179	3179	3179
R^2	0.74	0.67	0.90	0.88

The fact that some service-related branches had negative coefficients suggest that (i) few vehicles are registered in these branches and that (ii) land-use effects were captured by these variables. For instance, office-jobs of the branch C (manufacturing) are likely to generate fewer LCVs than actual manufacturing jobs, and they are also likely to be located in zones where most jobs belong to service-related branches. As we do not distinguish office jobs from other jobs, this negative effect is captured by the number of jobs in service-related activities.

6.1.2 Ideas for future work

This module would deserve additional refinements, because it captures the relation between the economic activity and the amount of LCV traffic and is therefore central when making forecasts. We list here some ideas for future developments.

Distinguishing office jobs from other jobs A more proper (but also more complex) solution to avoid negative signs would have been to explicitly account for different types of jobs within branches. An attempt at separating the number of employees of a specific branch working in offices from those working in factories was already done by coupling the Business and Enterprise Register with the Federal Register of Buildings and Dwellings (both administered by the Federal Statistical Office). The coupling did not work as expected (many industrial sites have a small office building where all employees are registered) and the negative signs remained, with the same branches.

Starting from empirical data: An alternative would be to use this module only to distribute the vehicles between branches but work with the empirical vehicle numbers from 2022 (possibly scaled), to see whether the end results (after network assignment) better match the available empirical data (especially traffic counts). If so, this would suggest a less arbitrary type of calibration, such as zone-specific calibration coefficients.

Cross-validation.

6.2 Vehicle Purpose

6.2.1 Current implementation

This module first multiplies the vehicles generated at the previous step (Vehicle Generation) by the probability that they are used on a randomly selected day. These vehicles are referred to as being "active". The active vehicles are then distributed among the three vehicles purposes (*Goods transport, Service, Other* - see Section 5.3).

These probabilities were computed using all surveys (type 1 and type 2) and are listed in Table 6. Because the standard dataset of the LCV survey only includes data about the vehicles that traveled on the reference day, we requested from the federal statistical office additional data about the vehicles that did not travel on the reference day. This allowed us to compute the proportion of vehicles that are used. This proportion is branch-specific, but not curb-weight specific (we do not have the information about the curb weight of unused vehicles). The vehicles which are used the less frequently (p(active)= 0.49, i.e. 3.43 days per week in average) are those registered in the branch B (mining and quarrying) . Those are are used the most frequently (p(active)=0.61, i.e. 4.27 days per week in average) are those of the branch D (electricity, gas, steam and air-conditioning supply).

The vehicles with a curb weight below 2 tons are more often used for services than the heavier vehicles. The vehicles registered in the branch H (transportation and storage) with a curb weight above 2 tons are those that are most often used for the purpose "Goods transport" (with a probability of 75 %). Conversely, the vehicles registered in the branch D (electricity, gas, steam and air-conditioning supply) with a curb weight below 2 tons are those that are most often used for services (86 % probability).

Note that the probabilities for the segment where the branch is "Unknown" were computed using all observations of the LCV survey where the branch is unknown. This includes vehicles owned by private persons and those owned by organizations who could not be associated to a branch. The fact that the vehicles purposes "Goods Transport" and "Service" are also predominant for this segment indicates that respondents belonging to this segment mostly used their vehicles for business purposes. Note also that when applying the model in simulation, we only apply these probabilities to vehicles owned by natural persons (i.e. we do not simulate vehicles whose branch would be unknown).

Some observations of the LWE are associated to branches for which the module *Vehicle Generation* generates no vehicle. These observations were ignored here. They represent 6.8 % of the sum of statistical weights of the vehicle file.

Table 6: Probability that a vehicle is active and probability of the various vehicle purposes

			Vehicle purposes		
Branch	p(active)	Curb weight	*Goods Transport*	*Other*	*Service*
A	0.58	≤ 2 t	0.37	0.28	0.36
		> 2 t	0.64	0.12	0.24
B	0.49	≤ 2 t	0.09	0.58	0.33
		> 2 t	0.27	0.50	0.24
C	0.51	≤ 2 t	0.30	0.05	0.65
		> 2 t	0.46	0.03	0.51
D	0.61	≤ 2 t	0.09	0.05	0.86
		> 2 t	0.17	0.13	0.70
E	0.56	≤ 2 t	0.31	0.24	0.45
		> 2 t	0.58	0.07	0.36
F	0.58	≤ 2 t	0.14	0.07	0.78
		> 2 t	0.34	0.12	0.55
G	0.52	≤ 2 t	0.38	0.08	0.55
		> 2 t	0.61	0.06	0.33
H	0.57	≤ 2 t	0.42	0.19	0.39
		> 2 t	0.75	0.07	0.18
N	0.57	≤ 2 t	0.15	0.09	0.76
		> 2 t	0.34	0.08	0.58
Unknown	0.51	≤ 2 t	0.36	0.16	0.47
		> 2 t	0.43	0.17	0.40

6.2.2 Ideas for future work

Distinction between working days and week-ends: This would be critical for the probability that a vehicle is used. We do not have the data yet about the day of the week for the vehicles which were not used, but the federal statistical office could possibly provide it to us (potentially also with the curb weight, even though it is less important).

6.3 Number of Tours

6.3.1 Current implementation

The vehicles were then converted into a number of tours, using average rates that are specific to the combination of branch, curb weight and purpose. These average rates were estimated using the object-oriented database generated based on the transport file (see Appendix A.1). The number of tours per vehicle and per day for the considered users are given in Table 7 (see Appendix A.2.2 for the list of observations that were dismissed).

Table 7: Number of tours per vehicle and per day

		Vehicle purposes		
Branch	Curb weight	*Goods Transport*	*Other*	*Service*
A	≤ 2 t	1.83	1.93	1.46
	> 2 t	1.26	1.23*	1.45
B	≤ 2 t	1.59*	1.47*	1.52*
	> 2 t	1.54*	1.23*	1.44*
C	≤ 2 t	1.72	2.34	1.52
	> 2 t	1.78	1	1.23
D	≤ 2 t	6.24	1	2.41
	> 2 t	1.98	1.23*	1.16
E	≤ 2 t	1	1.47*	2
	> 2 t	1.8	1.23*	3.35
F	≤ 2 t	1.43	1.38	1.44
	> 2 t	1.51	1.16	1.37
G	≤ 2 t	1.31	1.44	1.38
	> 2 t	1.58	1.02	1.39
H	≤ 2 t	1.37	1.47*	1
	> 2 t	1.33	1	1.15
N	≤ 2 t	1.76	1	1.62
	> 2 t	1.87	1	1.67
Unknown	≤ 2 t	1.61	1.35	1.56
	> 2 t	1.38	1.2	1.47

Note: values with an asterisk (*) were assigned using averages over all observations due to a lack of observations for these particular branches.

6.3.2 Ideas for future work

Replacing this module by a module "End Day": In practice, the number of tours made by a vehicle within a day may depend on the length of these tours. Thus, a refinement would consist in replacing this module by a module which would determine, after a tour is finished, whether a new tour should be done. This would be behaviorally more realistic, but would add stochasticity and the added-value in terms of link-level vehicle loads is unclear.

Zone dependency: The location of the vehicle's base may have an influence on the number of tours. For instance, vehicles based in remote areas may need to make longer tours, and may therefore make fewer tours within a day. Instead of implementing a "End Day" module, an alternative could be to make the number of tour and/or the number of stops per tour zone-dependent. We could for instance use the zone accessibility as an explanatory variable.

6.4 Correction

Because the modules *Number of tours*, *Next Stop Location* and *End Tour* are estimated using only observations coming from the set of vehicles who did not do any **SimplifiedTour** (see Appendix), we apply a correction to better reproduce the vehicle kilometers traveled of all vehicles. Let us denote S the set of vehicles who did at least one **SimplifiedTour** and R the set of vehicles who did not do any **SimplifiedTour**. Let n_R and n_S denote the corresponding numbers of vehicles and d_R and d_S the average daily distance traveled within each group according to the LCV survey. Let T_R denote the average number of tours made by vehicles in the set R. If we convert the $n_R + n_S$ vehicles into $(n_R + n_S)T_R$ tours, we expect to obtain a total traveled distance of approximately $(n_R + n_S)d_R$, i.e. we expect that the average tour length (in kilometers) produced by the model should be around d_R/T_R. According to the LCV survey however, the total traveled distance should be $n_R d_R + n_S d_S$. Thus, to be able to reproduce this distance, we use instead an average number of tour equal to $T_R(n_R + \frac{d_S}{d_R} n_S)/(n_R + n_S)$. Thus, the total vehicle-kilometers traveled (VKT) should be equal to

$$
\begin{aligned}
\text{VKT} &= \text{number of tours} \times \text{average tour length}, \\
&\simeq= (n_R + n_S) \times T_R \frac{n_R + \frac{d_S}{d_R} n_S}{n_R + n_S} \times \frac{d_R}{T_R}, \\
&\simeq n_R d_R + d_S n_S.
\end{aligned}
$$

These correcting factors were computed for each branch for which the module *Vehicle Generation* generates vehicles, for the two categories of curb weight. The results are listed in Table 8. This approach has the merit of avoiding a structural bias in the total distance traveled, although it does not account for the characteristics that are specific to simplified tours (e.g. different leg lengths).

6.5 Next Stop Location

The module *Next Stop Location* consists in a multinomial logit model, where the alternatives are different zones. In the estimation process, the alternatives are the zones defined by postal codes while in the application, the alternatives

Table 8: Correcting factors

	Curb weight	
Branch	≤ 2 t	> 2 t
A	0.98	1.02
B	0.98*	1.04*
C	0.99	1.01
D	1.05	1.00
E	1.00	0.93
F	1.00	0.99
G	0.98	1.02
H	0.76	1.06
N	1.01	1.01
Unknown	0.96	1.00

Note: values with an asterisk (*) were assigned using averages over all observations due to a lack of observations for these particular branches.

are the zones of the national passenger transport model. The coexistence of these two zone definitions precludes the use of alternative-specific constants.

To reduce the number of zones to be considered in the estimation process and thereby limit the computational time, we relied on stratified importance sampling. Instead of considering a choice set made of 3'187 zones, we only kept 300 (including the one chosen in the empirical data): 100 among the 200 zones which are the closest from the current one (in terms of travel time), 100 among the next 600 zones, and 100 among the furthest 2'387 zones. The differences in terms of probability to be sampled were then accounted for in the estimation process by adding $\log(p_i)$ to the utility function, where p_i must be proportional to the inverse of the probability that alternative i is sampled ($p_i = 2$ for the 200 closest zones, $p_i = 6$ for the next 600 zones and $p_i = 23.87$ for the remaining 2'837 zones) - see McFadden [1978] for a theoretical background to stratified sampling.

Let b denote the base of the tour and i the current zone. Let T_{ij} denote the travel time by LCV from zone i to any zone j (in minutes). The chosen utility specification is the following:

$$
\begin{aligned}
U_j = &\sum_{\text{land\_use}} \theta_{\text{land\_use}} \times \delta_{\text{land\_use}} \\
&+ \theta_{\text{intern}} \times \delta_{\text{intern}} \\
&+ (\theta_T \times \delta_{\#\text{stops}>1} + \theta_{T,\text{1stop}} \times \delta_{\#\text{stops}=1}) \times T_{i,j} \\
&+ \theta_{T>20} \times \delta_{T>20} \times (T_{i,j} - 20) \\
&+ \theta_{T>40} \times \delta_{T>40} \times (T_{i,j} - 40) \\
&+ \theta_{T,\text{return}} \times \delta_{\#\text{stops}>1} \times T_{j,b} \\
&+ \theta_{\text{size}} \times \log(\text{population in zone } j + \theta_{\text{jobs}} \times \text{jobs in zone } j) \\
&+ \log(p_j).
\end{aligned} \tag{1}
$$

The θ's represent the coefficients to be estimated while the δ's represent dummy variables. Instead of utilizing alternative-specific constants, we have constants relative to the land-use. The following land-use types are considered:[3]

- Low density: at most 100 inhabitants/km^2 and 100 jobs/km^2,
- Residential: more than 100 inhabitants/km^2 and a density of population at least twice as large as the density of jobs,
- Intermediary: at most 3'000 jobs/km^2 and neither a residential area nor a low density area,
- Employment Node: more than 3'000 jobs/km^2 and neither a residential area nor a low density area.

The dummy variables associated to the various land-uses ensure that only the constant $\theta_{\text{land\_use}}$ corresponding to the land-use of a zone enters in the utility function of that zone. The dummy variable δ_{intern} is equal to 1 if $i = j$, and 0 otherwise. The motivation behind it is that the travel time from a zone to itself cannot be calculated by our transport model. We therefore defined it to be zero and added the constant θ_{intern} to capture the disutility associated to travel time for internal trips.

For non-internal trips, the disutility associated to the travel time from the current zone to the candidate destination is captured via a piece-wise linear relation. For legs that are not the first of a tour, every minute of travel time between 0 and 20 min brings an additional disutility of θ_T, then every additional minute between the 21$^{\text{st}}$ and the 40$^{\text{th}}$ brings a disutility of $\theta_T + \theta_{T>20}$, and every additional minute above 40 minutes of travel time brings a disutility of $\theta_T + \theta_{T>20} + \theta_{T>40}$. The piece-wise linear relation is the same for legs that are the first of a tour, but θ_T is replaced by $\theta_{T,\text{1stop}}$.

The travel time from the candidate destination back to the base also enters the utility function when the vehicle has already done at least 2 stops.[4] It is then multiplied by the constant $\theta_{T,\text{return}}$.

[3] This is adapted from Hunt and Stefan [2007].

[4] Otherwise the return time would be the same as the travel time from the origin to the destination, which would be problematic for the estimation.

The last term is known as the "size term". The logarithm accounts for the fact that a zone is actually an aggregation of individual destinations (i.e. houses or establishments). The parameter θ_{jobs} quantifies the relative attractiveness of jobs compared to inhabitants for this segment. Segments where the customers are mostly businesses are expected to have a large θ_{jobs}.

The model specification in Eq. (1) was then estimated with the R package "Apollo" [Hess and Palma, 2019] using a dataset extracted from the LCV survey (see Appendix A.2). The zones considered for the estimation process are the geographical areas within Switzerland defined by postal codes. The matrix of travel times was obtained using the network of the national passenger transport model [Federal Office for Spatial Development, 2020a]. During the estimation process, the explanatory variables that were found to have a relatively weak explanatory power or that had large robust t-ratios were discarded. The resulting models are presented in Table 9.

Table 9: Estimation results for *Next Stop Location*

	Segments		
	Goods Transport	*Service*	*Other*
Estimated coefficients			
$\theta_{\text{land\_use}}$ Low Den	1.15 (8.1)	0.93 (4.8)	0 (-)
$\theta_{\text{land\_use}}$ Residential	0.50 (4.5)	0.38 (2.5)	0 (-)
$\theta_{\text{land\_use}}$ Interm	0.39 (4.1)	0.39 (3.5)	0 (-)
$\theta_{\text{land\_use}}$ Emp Node [ref]	0 (-)	0 (-)	0 (-)
θ_{intern}	-1.33 (8.4)	-1.5 (8.2)	-0.52 (1.5)
θ_{T}	-0.149 (21)	-0.149 (19)	-0.154 (5.0)
$\theta_{T,\text{1stop}}$	-0.132 (20)	-0.141 (19)	-0.134 (10)
$\theta_{T>20}$	0.011 (1.2)	0.017 (1.5)	0 (-)
$\theta_{T>40}$	0.071 (10)	0.064 (6.8)	0.079 (3.2)
$\theta_{T,\text{return}}$	-0.012 (3.5)	-0.044 (9.7)	-0.040 (2.3)
θ_{size}	0.724 (27)	0.760 (27)	0.539 (4.6)
θ_{emp}	13.2 (1.6)	1.24 (2.2)	0 (-)
Goodness of fit statistics			
Number of obs.	4'035	2'374	162
Ln-Likelihood (0)	-19'833	-15'575	-859
Ln-Likelihood final	-13'451	-10'535	-573
ρ^2	0.322	0.324	0.333

Note: values in parentheses denote |robust t-ratio|.

To facilitate the interpretation of the coefficients associated to the travel time to the destination, we drew the corresponding contribution to the utility function in Fig. 2. We observe that the marginal disutility of travel time decreases very slightly after 20 min for the segments "Goods Transport" and "Service", and then more significantly above 40 min for all segments, and for first and other legs. Note also that travel time is less penalized for the first legs of a

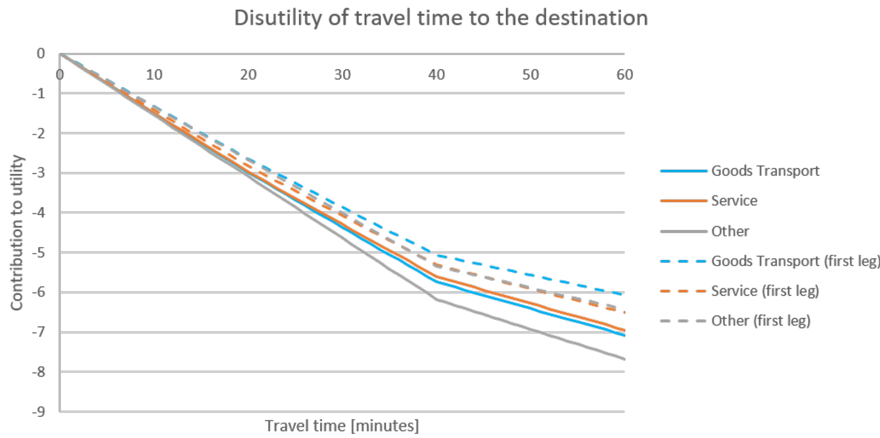

Figure 2: Contribution of travel time to the destination to the utility.

tour, in particular for goods transport. This is consistent with common delivery pattern where several destinations close to each other are grouped within a tour, producing long first and last legs and shorter intermediary legs.

The positive land-use coefficients have to be interpreted as a larger attractiveness for the corresponding land-uses **compared to employment nodes** (which was taken as reference by forcing the corresponding $\theta_{\text{land\_use}}$ to be zero), and **all things being equal** (in particular, for a given population size and number of jobs). The relatively high value of $\theta_{\text{land\_use}}$ for low density areas suggests that between 2 candidate destinations that are as far away from the current zone and from the base and that have identical number of inhabitants and jobs, the zone with the smallest population and job density (i.e. the zone with the largest area) is likely to be more attractive.

The coefficient θ_{intern} is advantageously interpreted in comparison with θ_T. As θ_{intern} has unit of utility and as θ_T has unit of utility/min, the ratio $\theta_{\text{intern}}/\theta_T$ has units of minutes. For "Goods Transport" for instance, it is equal to 8.9 minutes (10.1 min for the segment "Service" and 3.4 min for the segment "Other", but this last value has more uncertainty) . This gives an idea of the average travel time for internal trips.

The negative values of $\theta_{T,\text{return}}$ suggest that vehicles also tend to prefer zones that are closer to their home establishment, although the comparison with θ_T shows that the proximity to their current zone is more important.

The values of ρ^2 are relatively large but this is not very surprising because the ρ^2 is comparing against a model where all the coefficients to be estimated would be zero, i.e. where $U_j = \log(p_j)$. This approximates a model where, without sampling, all zones would be equally likely to be selected. It is clearly not difficult to improve upon such a model.

6.5.1 Analysis

To evaluate the consistency between the provided utility specification and the real world distribution of trip duration, we applied the estimated models on the choice situations used for the estimation process. We then averaged the estimated travel time distributions over all observations of the same segment, and compared the results with the average of the observed travel times. The results, illustrated in Fig. 3, suggest that the piece-wise linear utility specification with respect to travel time is sufficient.

6.5.2 Ideas for future work

Further quality checks: a further test would consists in checking whether the model predicts shorter trip lengths when the observed trip length is short. This could be done by grouping the observations by bins of observed trip lengths (e.g. 5-10 km) and plot the distribution of predicted trip lengths for these decisions.

Branch-specific travel time valuation: such specifications should be tested again and documented.

Dynamic panel effects: the model specification could be modified, such that the valuation of travel time for the second, third and following destination choices within a tour could depend on the travel times of the alternatives previously chosen by this vehicle. This would lead to some vehicles doing tours with only short legs, while others would do tours with only long legs, which might be more realistic. The best utility specification to model such an effect remains to be identified.

6.6 End Tour

The module *End Tour* is a binary logit model, where the outcome is the probability to end the tour after the current stop. We estimated one model for each vehicle purpose. The branch of the vehicle owner and the curb weight were integrated using specific constants (and only when their impact was significant).

The retained model specification is the following:

$$\begin{aligned}
U_{\text{return-to-establishment}} =& 0 \\
U_{\text{continue}} =& \theta_0 + \sum_{\text{branch}} \theta_{\text{branch}} \times \delta_{\text{branch}} \\
&+ \theta_{\text{CW}>2\text{t}} \times \delta_{\text{CW}>2\text{t}} \\
&+ \theta_{\text{2stops}} \times \delta_{\#\text{stops}=2} \\
&+ \theta_{\log(\#\text{stops})} \log(\#\text{stops}) \\
&+ \theta_{T_\text{return}} \times \text{travel time from current zone to base} \\
&+ \theta_{\text{accessibility}} \times \text{accessibility of the current zone,}
\end{aligned}$$

where the θ's are coefficients to be estimated, the δ's are dummy variables allowing the consideration of specific constants for specific branches (δ_{branch}),

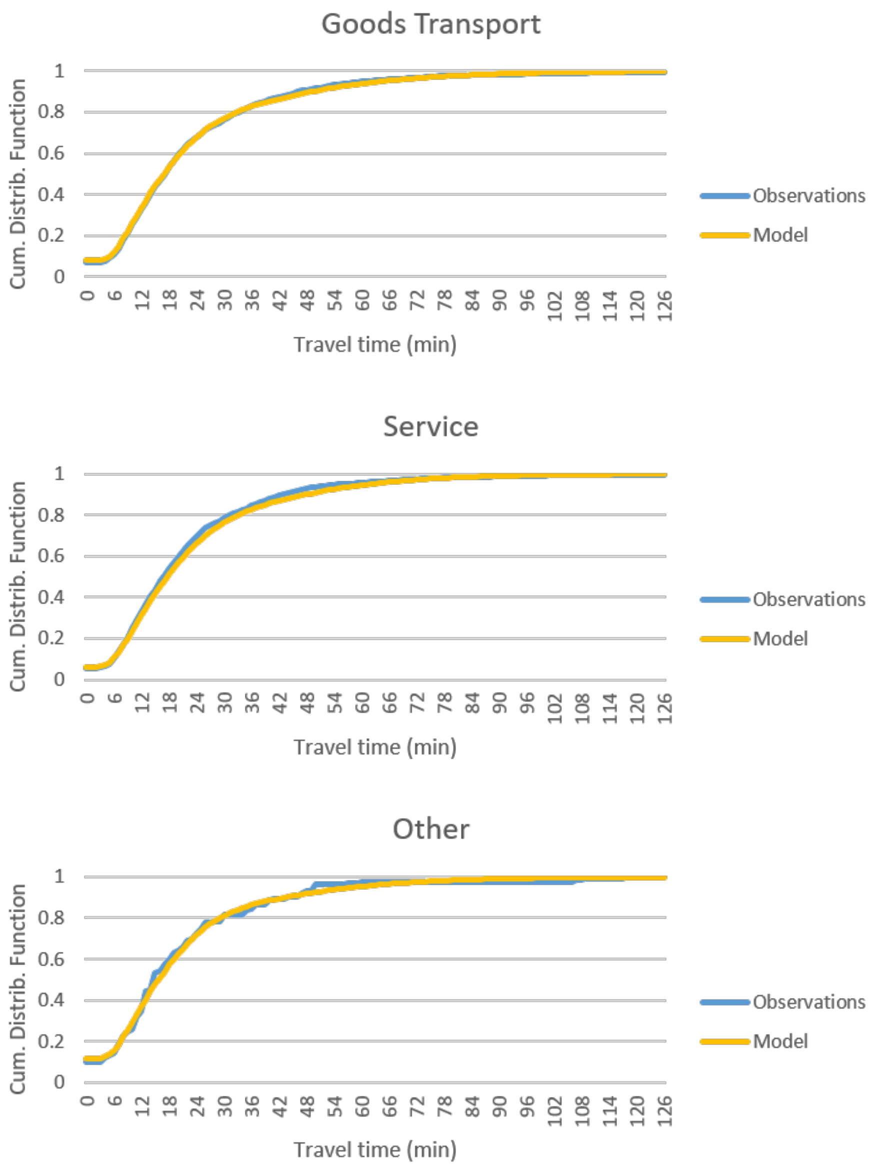

Figure 3: Next Stop Location: application of the model on the observations

for curb weights above 2 tons ($\delta_{\text{CW}>2\text{t}}$), or when the number of stops made so far (including the base) is equal to 2 ($\delta_{\#\text{stops}=2}$). The number of stops made so far also influences the probability to continue via its logarithm (log(#stops)).[5] The accessibility is computed using the following formula:

$$\text{Accessibility of zone } i = \frac{1}{40000} \sum_j A_j \exp(-0.2 \times T_{ij}),$$

where A_j is the sum of the number of jobs (full-time equivalents) and inhabitants in zone j. The factor 1/40000 is only here for numerical reasons, to ensure that the scale of this explanatory variable is consistent with the scale of other variables.

Table 10: Estimation results for *End Tour*

	Segments		
	Goods Transport	*Other*	*Service*
Estimated coefficients			
θ_0	-0.68 (2.7)	0.23 (0.74)	0.064 (0.59)
$\theta_{\text{branch = G}}$	0.68 (5.4)	0 (-)	0.64 (4.8)
$\theta_{\text{branch = H}}$	0.59 (3.6)	0 (-)	0 (-)
$\theta_{\text{branch = N}}$	0.40 (1.6)	0 (-)	0 (-)
$\theta_{\text{branch = Unknown}}$	0.25 (2.2)	0 (-)	0 (-)
$\theta_{\text{CW}>\text{ 2t}}$	0.32 (3.1)	0 (-)	0 (-)
θ_{2stops}	-1.16 (7.1)	-0.89 (2.0)	-0.88 (8.5)
$\theta_{\log(\#\text{stops})}$	0.39 (2.7)	0 (-)	0 (-)
θ_{T_return}	0.01 (4.2)	0 (-)	0 (-)
$\theta_{\text{accessibility}}$	0.09 (1.9)	0 (-)	0.11 (2.1)
Goodness of fit statistics			
Number of obs.	3'636	126	2'078
Ln-Likelihood (0)	-2'163	-86	-1'654
Ln-Likelihood final	-1'854	-81	-1'558
ρ^2	0.143	0.053	0.058

Note: values in parentheses denote |robust t-ratio|.

The ρ^2 corresponds to McFadden's pseudo-R squared. It compares the estimated model against the a model where all parameters are 0. In this case, it means returning to the establishment with a probability equal to 0.5. The positive value of ρ^2 indicates that we do better than such a simplistic model. The obtained ρ^2 remain however rather small (a rule of thumb often applied in the field considers a good ρ^2 to be above 0.2).

[5]The choice to include the logarithm rather than simply the number of stops is motivated by the intuition that the absolute difference in probability between stop n and stop $n + 1$ should decrease with n.

The branch-specific constants can be interpreted as follows: the larger this constant is, the larger the probability to do more stop. For *Goods Transport* for instance, the vehicles registered by establishments of the branches G, H, N or whose branch is unknown have larger probabilities to do more stop than vehicles of other branches (all things being equal).

Similarly, the positive value of $\theta_{\log(\#\text{stops})}$ for goods transport indicates that the longer the tours are, the less likely they are to be terminated after the current stop. This is consistent with the fact that the distribution of the number of stops in a tour tends to be fat-tailed (even if most tours only include 1 or 2 stops, some professions can do dozens of stops within a tour). The negative value of $\theta_{2\text{stops}}$ indicates that the probability to return to the establishment after only two stops (i.e. the base and one external stop) is larger than predicted by the logarithm of the number of stops for the segment *Goods transport*, and above the average of other situations for the segments *Service* and *Other*.

The positive value of θ_{T_return} for the segment *Goods Transport* indicates that the further the vehicles are from their base, the more likely they are to do one more stop before returning to their establishment. Similarly, the positive values of $\theta_{\text{accessibility}}$ indicate that vehicles are more likely to make one more stop when they are in zones with a better accessibility.

6.6.1 Analysis and validation

In order to verify that the model predictions are reasonable, we applied the estimated models on the choice situations used for the estimation process. We then averaged the probability to continue over all observations having the same number of stops made so far, and compared the results with the average of the observed choices for these same observations. The results are illustrated in Fig. 4. The match between the empirical and simulated cumulative distribution functions seems rather good. The large variations in the empirical averages for large numbers of stops are due to the small number of observations. Is is entirely expected and acceptable that the model does not reproduce these variations.

6.6.2 Ideas for future work

Cross-validation.

7 Issues regarding the model application

The different modules are now estimated. Before applying them sequentially as described in the structure illustrated in Fig. 1, some additional practical challenges have to be solved. One of them is the question of the curse of dimensionality that typically arises when one wants to model all possible tours and associate them to a probability. Section 7.1 discusses two alternative solutions to this problem. The second is the problem of the external traffic, i.e. the traffic whose origin or destination is outside the studied area, which requires a specific treatment. This treatment is described in Section 7.2.

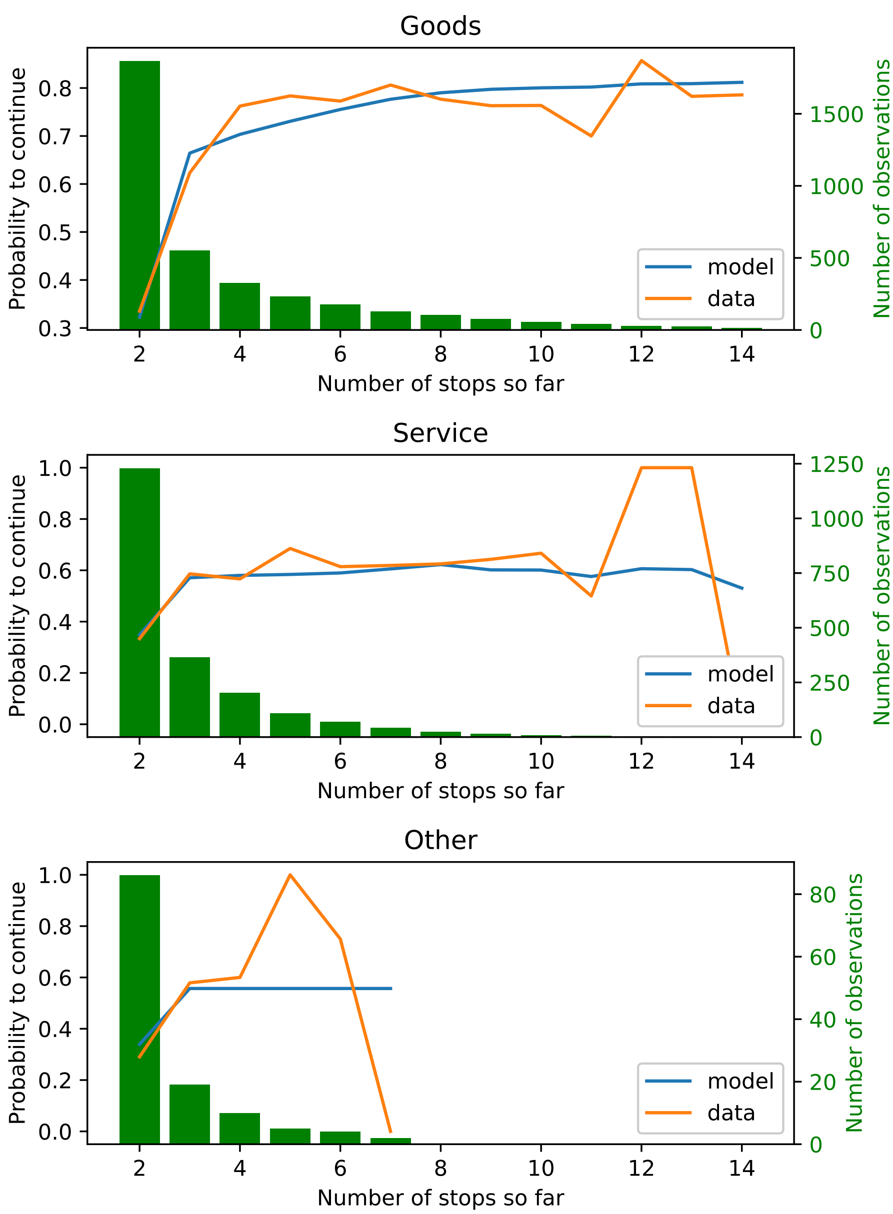

Figure 4: EndTour: application of the model on the observations

7.1 Microscopic vs macroscopic view

There are two ways to apply such a model. The approach chosen by Hunt and Stefan [2007] and most of the subsequent literature consists in interpreting the results of modules as probabilities and simulating discrete agents taking decisions iteratively until they finish their tour. The second approach consists in considering a continuum of agents and interpreting the results of the modules as proportions. The modeler then simply keeps track of the proportion of this continuum of users choosing each alternative. The advantage of this second approach is that it does not require any sampling, so that results are fully deterministic. The drawback is that all combinations of decisions must be evaluated, and that this can computationally expansive. To ensure that the number of combinations is finite, a solution consists in assuming a maximum number of legs per tour. Then a second trick to limit the computational burden consists in restricting ourselves to models which allow the aggregation of combinations. For instance the result of the proposed module Next Stop Location depends on the number of stops made so far and on the current zone, but not on the zones previously visited.

We implemented the two variants but rely for the time being on the microscopic view for the modules Next Stop Location and End Tour, because it reduces the computational time and the results were found to be reasonably stable in terms of traffic loads on important links.

7.2 International traffic

The modules described in the Sections 5 and 6 are only applied to zones inside Switzerland. Because of limited data availability, the international traffic was modelled in a much simpler way, with conventional trip generation and trip distribution, followed by a scaling step. These steps are briefly presented hereafter.

7.2.1 Trip generation

For lack of better data, trip generation rates were estimated based on the results for Swiss zones of the model described in this paper (see Sections 5 and 6). We did a linear regression of the number of trips produced (equal to the number of trip attracted) on the population and number of jobs of each zone. The zones used for this purpose are the 7'965 domestic zones of the national passenger transport model. The result is an average rate of 0.108 LCV trip per day per job (full time equivalent) and 0.068 LCV trip per day per inhabitant. The R^2 of the linear regression is 0.89. We then applied these rates on all zones (inside and outside Switzerland) of the national passenger transport model.

7.2.2 Trip distribution

Here as well, the results of the model described in this paper for domestic traffic were used instead of empirical data for estimation purposes. Let us denote X_i

the number of trips originating from/arriving into zone i. With the retained functional form, the number of trips from zone i to zone j is given by:

$$\text{Trips from zone } i \text{ to zone } j = \alpha_i \beta_j X_i X_j \exp(\gamma T_{i,j}),$$

where γ is the parameter to be estimated, while α_i and β_i must be computed using a classical Iterative Proportional Fitting algorithm. The value of γ that minimizes the root-mean-square error for the Swiss zones was found to be -0.1.

This model was then applied on the trips generated in the first step (Trip Generation), to obtain an Origin-Destination matrix for all zones (inside and outside Switzerland). The part of the matrix corresponding to trips inside Switzerland was then deleted.

7.2.3 Scaling

The Origin-Destination matrix resulting from the first two steps was then assigned to the network of the national passenger transport model. The traffic loads produced at the Swiss boundaries (where no domestic Swiss traffic is expected) were then compared to the traffic counts available. The Origin-Destination matrix was then scaled in order to match the sum of the traffic loads at all borders with traffic counts available.

8 Validation

This Section is about the validation of model as a whole (module-level validation is handled in Section 6).

8.1 Network assignment and comparison with traffic counts

The trip matrix resulting from the application of the models for domestic and international traffic was affected on the road network in VISUM. This allows the comparison of the traffic loads resulting from the model with the traffic counts on the highway network (provided by the Federal Roads Office). The comparison is however difficult, because the existing detectors based on inductive loops or laser technology cannot reliably distinguish LCVs from passenger cars. This limitation is unavoidable, because the definition of a LCV relates to the equipment inside the vehicle (the existence of true rear seats), and not to an easily measurable characteristic like its length, number of axles, mass or volume. It is therefore unclear whether data from traffic counts should be used for validation and calibration purposes.

8.1.1 Ideas for future work

Simultaneous calibration of passenger cars and LCVs As inductive loops and laser detectors are relatively good at distinguishing light vehicles (i.e. with a gross vehicle weight up to 3.5 tons) from heavier ones, a possible solution would be to validate and calibrate the segments of passenger cars and LCVs together.

8.2 Comparison with the LCV survey

- Distribution of trip duration
- Distribution of trip length
- ...

References

Cambridge Systematics and COMSIS Corporation. Quick response freight manual, 1996. prepared for the United States Federal Highway Administration, Washington, DC, USA.

G. de Jong and M. Ben-Akiva. A micro-simulation model of shipment size and transport chain choice. *Transportation Research Part B: Methodological*, 41 (9):950–965, 2007.

G. de Jong, M. de Bok, and S. Thoen. Seven fat years or seven lean years for freight transport modelling? Developments since 2013. *Journal of Transport Economics and Policy*, 55(2):124–140, 2021.

R. B. Ellison, C. Teye, and D. A. Hensher. Modelling sydney's light commercial service vehicles. *Transportation Research Part A: Policy and Practice*, 96: 79–89, 2017.

Federal Office for Spatial Development. Aggregierte Methode Güterverkehr (AMG): Methodenbeschrieb, 2015.

Federal Office for Spatial Development. Modelletablierung Nationales Personenverkehrsmodell (NPVM) 2017 - Schlussbericht, 2020a.

Federal Office for Spatial Development. Modelletablierung Nationales Personenverkehrsmodell (NPVM) 2017 - Schlussbericht, 2020b.

Federal Office for Spatial Development. Schweizerische Verkehrsperspektiven 2050: Schlussbericht, 2021a.

Federal Office for Spatial Development. Wirtschafsverkehr in urbanen Räumen, 2021b.

Federal Roads Office. National register of vehicles, 2022.

Federal Statistical Office. Erhebung leichte Nutzfahrzeuge 2013, 2015a.

Federal Statistical Office. Erhebung leichte Nutzfahrzeuge 2013: Erhebungsbericht, 2015b. FSO Number: do-d-11.05-LWE13-D21.

Federal Statistical Office. New registrations of road vehicles by vehicle group and vehicle type, 2022. FSO Number: cc-e-11.03.02.02.03.

S. Grebe, G. de Jong, C. Wampera, C. Obermayer, and R. Pompl. VMÖ–a new strategic transport model for austria. *Transportation Research Procedia*, 49: 95–106, 2020.

S. Hess and D. Palma. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 32:100170, 2019.

J. D. Hunt and K. Stefan. Tour-based microsimulation of urban commercial movements. *Transportation Research Part B: Methodological*, 41(9):981–1013, 2007.

D. McFadden. Modelling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickers, and J. W. Weibull, editors, *Spatial Interaction Theory and Planning Models*. North-Holland Publishing Company, 1978.

PTV GROUP. PTV Visum 2020 Manual, 2019.

T. Sakai, A. Romano Alho, B. Bhavathrathan, G. D. Chiara, R. Gopalakrishnan, P. Jing, T. Hyodo, L. Cheah, and M. Ben-Akiva. Simmobility freight: An agent-based urban freight simulator for evaluating logistics solutions. *Transportation Research Part E: Logistics and Transportation Review*, 141:102017, 2020.

K. J. Stefan, J. D. Hunt, J. D. P. M. McMillan, and A. Farhan. Development of a fleet allocator model for calgary, canada. *Transportation Research Record*, 1994(1):89–93, 2007.

S. Thoen, L. Tavasszy, M. de Bok, G. Correia, and R. van Duin. Descriptive modeling of freight tour formation: A shipment-based approach. *Transportation Research Part E: Logistics and Transportation Review*, 140:101989, 2020.

A Appendix - LCV Survey: data preparation

A.1 Generation of an object-oriented database and tour reconstruction

The data of the LCV survey is separated into two main files : a vehicle file (LWE_2013_vehicle.csv) and a transport file (LWE_2013_transport.csv). The vehicle file contains one row per vehicle surveyed. This row contains various information about the vehicle, its main use and its mileage on the reference day. This represents the data that is common to both survey types (1 and 2). The transport file contains one row per “transport”. A transport is a quantity of goods of a given type (e.g. food) which are transported together (i.e. picked up and dropped off simultaneously). Each row then contains information about the vehicle which carried out these transports (including a unique identifier) and information about the transport itself, such as the weight and type of the transported goods, the postal code of their origin and destination, as well as the vehicle mileage at the origin and destination. "Empty transports" must also be reported. "Empty transports" correspond to vehicle movements without any goods on board. The reporting of such movements allow the reconstruction of the entire trip chain of a vehicle over a day.

To permit the estimation of our *Next Stop Location* and *End Tour* modules, it was necessary to deduce the vehicle movements from the individual transports and to arrange them into tours. To facilitate this task, we organised the data into an object-oriented database.

The types of objects we considered are the following:

- Questionnaire
- Shipment
- Stop
- Leg
- SimplifiedTour
- ReconstructedTour

The object **Questionnaire** corresponds to a questionnaire of the LCV survey. Its own attributes include the unique ID of the questionnaire, the statistical weight considered in the LCV survey, the curve weight and load capacity of the vehicle, the main uses indicated for the reference day and the total distance traveled during that day. The object **Shipment** corresponds to a "transport", i.e. a row of the transport file. Its own attribute include the weight and type of good. The object **Stop** corresponds to a stop made a vehicle. Its own attributes include the mileage of the vehicle at this stop and the postal code. The object **Leg** corresponds to an interval between two stops. The object **SimplifiedTour** corresponds to a tour in the sense of the LCV survey, i.e. the case where vehicles did not have to fill in information about each individual transport and were allowed to write only aggregate information (such as the number of stops made in this tour and the total weight of all included transports). The object **ReconstructedTour** is made of a series of consecutive legs (the criteria for grouping them are explained further below).

Besides their own attributes, the objects can also contain links to related objects. The **Questionnaire** object contains links to the collections of **Shipments**, **Stops**, **Legs**, **SimplifiedTours** and **ReconstructedTours** associated to it. The **Shipment** and **Leg** objects contain links to the **Stops** serving as origin and destination. The **SimplifiedTour** and **ReconstructedTour** contains links to the **Legs**, **Shipments** and **Stops** which are fully contained in these tours.

In order to create this database, the transport file was first processed line by line, in order to create the collections of **Questionnaires**, **Shipments** and **Stops**. Except for rows corresponding to **SimplifiedTours** (which are characterized by a positive value in the column "NumberOfStops"), each row generates a new **Shipment** object. New **Questionnaires** are only created when processing a row with a new unique identifier. A similar procedure was followed to generate **Stops**, but the mileage (integer number of kilometers) was used as unique identifier within a **Questionnaire**.

SimplifiedTours required a special treatment as each instance is saved as three rows in the transport file. For grouped deliveries for instance, one row corresponds to the leg from the origin to the first stop and has as weight the total weight of all shipments of this **SimplifiedTour**. Another row corresponds to all intermediary legs and has as weight the half of the total weight. The third row corresponds to the leg from the last stop to the final destination (which is often the same as the origin of the first leg, but not necessarily). This last leg has a weight of zero. The intermediate legs are easily identified as parts of **SimplifiedTours** thanks to their positive value of "NumberOfStops". The identification of the first and last legs required comparing the type of good, the weight, as well as the origins and destinations with those of the intermediate legs.

The list of **Legs** and that of **ReconstructedTours** were generated only after processing all the rows of a **Questionnaire**. Legs are logically defined as the interval between two consecutive stops (after sorting the **Stops** of a **Questionnaire** by increasing mileage). The rules for identifying **ReconstructedTours** are somewhat more arbitrary. We defined the base of the vehicle as the postal code of their first Stop of the day (i.e. the one with the smallest mileage). We then defined "closed" **ReconstructedTours** as the set of legs contained between two successive stops in the vehicle's base. This is clearly not ideal, as for instance two consecutive stops in the base zone generate a tour with only one leg. When the last stop of the day does not have the same postal code as the first one, the remaining set of legs is grouped into an "open" **RecontructedTour**. Note that a **SimplifiedTour** does not necessarily start or end at the vehicle's base. It might thus be included within a **RecontructedTour**.

A.2 Preparation of datasets for model estimation

A.2.1 Data for the module NextStopLocation

The data necessary for the estimation of the module NextStopLocation is essentially a subset of the set of **Leg** objects. There are several reasons for excluding Legs:

- **legs of users who did SimplifiedTours**: As we do not have all the data to reproduce **SimplifiedTours**, we entirely exclude them from the estimation set. [6]

[6] We could technically keep some of the legs made by such users (those outside Simpli-

- **Return legs**: The last legs of closed **ReconstructedTours** do not include any destination choice (the destination is the base of the tour). They were discarded from the estimation set.

- **Postal code corresponding to large organizations**: some postal codes do not correspond to well-defined geographic areas but to large organizations. The estimates of travel times from or to such postal codes would necessarily be very approximate. We discarded the legs where either the origin or destination had such codes, as well as the legs belonging to tours where the base has such a code (because the travel time from the destination to the base is an explanatory variable).

- **Postal code corresponding to zones outside Switzerland**: Because of difficulties to obtain a comprehensive dataset for all the neighboring countries, we focused on trips within Switzerland. The legs where either the origin or destination were abroad, as well as the legs belonging to tours where the base is abroad were discarded.

The remaining legs represent choice situations for which we want to estimate the model. The file generated for the estimation process then includes the chosen alternative, various alternative-specific attributes for the chosen alternative as well for other sampled alternatives (such as the number of jobs, inhabitants, travel time from the current zone to the candidate destination and from the candidate destination to the base), as well as some general data, such as the statistical weight of this questionnaire.

A.2.2 Data for the module EndTour

The data necessary for the estimation of the module EndTour is also derived from the set of **Legs**. If the leg is a "return leg" (i.e. its destination has the same postal code as the base of this questionnaire), then the decision taken is to end the tour. Otherwise the decision taken is to make another stop. Here as well, we had to discard some observations:

- **Legs originating from the base**: to avoid stopping tours before they start, we assumed that the first leg of a tour could not be a returned leg. This does not mean that the zone of the base is excluded from the destination choice.

- **Legs of users who did SimplifiedTours**: as we do not know the intemediate stops of SimplifiedTours, we could not include the intermediate EndTour decisions in the estimation process. To avoid any bias, we exclude all decisions taken by users who did **SimplifiedTours**.

fiedTours as well as the first legs of SimplifiedTours). This would have increased the set of observations by about 10 %. We chose not to do so, so that the reproduced tours are fully consistent with the behavior of users not doing simplified tours.

- **Postal code corresponding to large organizations**: here as well, we discarded the observations where either the current zone or the base had postal codes corresponding to large organizations.
- **Postal code corresponding to zones outside Switzerland**: here as well, we discarded the observations where either the current zone or the base had postal codes outside of Switzerland.

The file generated for the estimation process contains the chosen alternative (end tour or continue) as well as various variables related to the choice situation (e.g. distance from the current zone to the base, number of stops made so far).