



Research Internship Report

presented by

Gabrielle MEASSON

Internship from 01/07/2026 until 31/08/2026 in:

Transport and Mobility Laboratory, EPFL

Subject:

Data analysis for the creation of synthetic containers

Director of the laboratory: **Professor Michel Bierlaire**

Supervisors: **Candice Baud** and **Pavel Ilinov**

In the scope of the **Laidlaw Research and Leadership Programme**

Abstract

Freight demand models are moving from aggregate flows between zones towards individual shipments, and that shift depends on what statistical offices publish. This report asks which of the variables such a model needs can actually be obtained for Switzerland. I built a list of variables from the literature, inventoried the road and rail freight catalogue of the Federal Statistical Office, ranked the series with a scoring rule I adopted, and opened twenty of them to check what they really contain. Four of the twenty titles describe their content inaccurately, the files my ranking put at the top are mostly not the ones that proved useful, and no published table describes a single shipment.

Contents

Abstract	1
1 Introduction	3
1.1 Context	3
1.2 The project	3
1.3 Objective of this work	4
1.4 Definitions	4
2 Literature review	6
2.1 The link between population and freight	6
2.2 Other models of synthetic freight demand	7
2.3 What these references need	9
3 Methodology	10
3.1 Phase 1: from reading to a list of variables	10
3.2 Phase 2: finding the catalogues	11
3.3 Phase 3: building and ranking the inventory	11
3.4 Phase 4: opening the files	12
4 Results	13
5 Discussion and conclusion	17
6 Acknowledgements	20
A Appendix	22
A.1 Inventory files	22
A.2 The twenty files, full grid	22
A.3 Calculations	22

Chapter 1

Introduction

1.1 Context

Freight transport is an essential part of the transport system. Goods have to move from where they are produced to where they are consumed, and they use the same roads and the same railway lines as passenger traffic. This movement has a cost for society: congestion, noise, and greenhouse gas emissions. Many countries therefore try to influence how freight moves. To know whether a given policy can work, and how much traffic it would actually change, we need models that are able to predict freight demand.

Predicting freight demand is more difficult than predicting passenger demand, and the main reason is data. For passengers, national travel surveys describe who travels, how often, and by which mode. Nothing equivalent exists for freight. Firms are not surveyed in the same way, and information about shipments is commercially sensitive, so it is rarely published. What is publicly available is mostly aggregate: total tonnes transported between regions, by commodity group and by mode.

This has consequences for the models. Current freight demand models rely heavily on aggregate data, which limits their capacity to capture the detailed logistics decisions made by individual firms. A firm decides how much to send at a time, how often to send it, and by which mode. These are exactly the decisions a transport policy tries to influence, and they are no longer visible once flows have been aggregated into tonnes between zones.

1.2 The project

Passenger transport modelling has faced a similar problem and has addressed it with synthetic populations. Instead of modelling flows between zones, a list of individual agents is generated, each with realistic attributes, in such a way that the population as a whole reproduces the statistics that are known. The same idea can be applied to freight: rather than generating synthetic travellers, one generates synthetic freight units. This analogy is what motivates the approach followed in this project, and the models that have already been built on it are reviewed in Chapter 2.

The broader objective of the project is to move toward a more disaggregate representation of freight demand by generating synthetic freight units that remain consistent with available data. In practice, this means combining several sources: statistics describing firms, such as their location, their sector of activity and their size, and aggregate freight flows between zones. A first version of such a model already exists for Switzerland. For it to progress, it needs Swiss data, both to be calibrated and to allow its output to be compared with reality.

1.3 Objective of this work

My work focuses on the data foundations of this approach. The question I try to answer is a practical one: which of the variables that a freight demand model needs can actually be obtained for Switzerland, and which ones cannot? I first reviewed the literature in order to identify the key variables used in modern freight demand models. I then searched for datasets that could provide these variables, and evaluated their structure and their quality. Finally, I compared the datasets I retained with the variables the model requires, in order to identify which information is available and where the gaps are. Part of this material was also presented as a poster at the end of the internship, and two of the files discussed in Chapter 4 were first opened for it.

The rest of this report follows these steps. Chapter 2 reviews freight demand models and the data they use, and extracts from them the variables of interest. Chapter 3 describes how I searched for datasets and the criteria I used to rank them. Chapter 4 presents what the datasets I opened actually contain. Chapter 5 discusses what these data allow, what they do not, and what remains to be done.

1.4 Definitions

A few terms are used throughout the report and are worth fixing here.

A *shipment* is a quantity of goods sent by one firm to one receiver at one time. It is the unit the model tries to generate. A truck can carry several shipments on the same journey, so the load of a vehicle is not the weight of a shipment.

Aggregate data describes totals, for example the tonnes moved between two regions in a year. *Disaggregate* data describes individual units. No source used in this report is disaggregate in that sense. When I call one table more detailed than another, I mean that it is less aggregated, not that it reaches the level of a shipment.

The *catalogue* is the list of statistical products published online by the Federal Statistical Office. One *object* in that catalogue is one published item, which may be a table, a graph, a press release or a glossary entry. One *series* is one table republished over several years, so that one series groups several files.

Two shapes of table are distinguished throughout. An *origin-destination matrix* is what lets a demand be placed on a map. It gives flows from each place to each other place, so many tonnes from one region to another. A *breakdown* gives a national total split along one or two

dimensions, with no origin and no destination: the tonnage of the whole country divided by type of goods, or by distance travelled, or by type of vehicle. It says how much and of what kind, never from where to where. The two words are the ones I used in the inventory itself, and almost every published table is one or the other.

Goods are classified by international nomenclatures. The current one, used since 2008, has twenty groups. The one used before it has ten chapters. The two do not map onto one another group by group, which matters in Chapter 4 because the only table that crosses place and commodity for Swiss vehicles uses the older one.

Tonnages in this report are given in thousand tonnes unless stated otherwise, as in the source files. A *tonne-kilometre* is one tonne carried over one kilometre, so dividing tonne-kilometres by tonnes gives an average distance.

Chapter 2

Literature review

This section reviews the work the project builds on. It first presents the approach followed by the project itself, which transfers a method from passenger transport modelling to freight. It then presents other models that generate disaggregate or synthetic freight demand, with a focus on the data each of them used.

2.1 The link between population and freight

Passenger transport modelling and freight transport modelling have followed the same path, but not at the same speed. Both started with aggregate flows between zones. Passenger modelling then moved to synthetic populations, where a list of individual agents is generated in such a way that the population as a whole reproduces the statistics that are known, and where behaviour emerges from the choices of these agents. Freight modelling has been slower to follow, mainly because the observations needed to build and calibrate such models are rarely available.

The project follows this path for freight. It builds on an existing agent-based model of Swiss inland freight transport, in which the decision-maker is the container itself. This assumption makes the model operational, but a container does not decide anything. The project replaces this decision-maker by the carrier, who is the actor that actually chooses how goods are transported (Bierlaire and Ilinov, 2025).

The method used to represent this choice comes directly from passenger modelling. It is the household activity scheduling problem, in which a household decides how each of its members will travel during the day, choosing between the vehicles it owns and public transport. The analogy is explicit: the carrier plays the role of the household, and each shipment plays the role of a member of that household. Just as a household decides who takes the car and who takes the train, the carrier decides which shipments travel in which vehicle, and which ones are grouped together with shipments from other carriers at a consolidation hub (Bierlaire and Ilinov, 2025). Grouping shipments in one's own vehicle corresponds to using the family car, and grouping them at a hub corresponds to using public transport.

The model is organised in three steps, which correspond to the generation, distribution and mode assignment steps of the classical four-step framework (Ilinov, 2026). The first step assigns

a role, sender or receiver, and a commodity type to each Swiss establishment. The second step estimates how many tonnes are exchanged annually between each pair of establishments. The third step turns these annual flows into individual shipments.

The output of these three steps is a synthetic population of shipments. In the first version of the model, a shipment is a vector of seventeen attributes. Five of them belong to the carrier and are the same for every shipment it sends: its location, its size category, the vehicles available to it, the types of freight it can ship, and its number of shipments. The twelve others belong to the shipment itself: the location and the size of the receiver, the type of freight, the weight, the required vehicle type, the value in Swiss francs, the requested start time, the latest acceptable delivery time, whether that delivery time is a hard or a soft constraint and, if soft, the acceptable time window, the available transport modes, and the availability of internal and of external consolidation (Bierlaire and Ilinov, 2025).

The version currently implemented uses six of them: sender, receiver, commodity type, weight, departure time and road vehicle type (Ilinov, 2026). This report focuses on which of these variables can actually be obtained for Switzerland. Chapter 3 keeps the ones a survey of vehicles could plausibly carry and groups them into eight variables, which are what the rest of the report follows.

The model uses three datasets as input (Ilinov, 2026). An aggregate origin-destination dataset, which gives the quantity shipped from one geographical zone to another. The STATENT dataset, which gives the approximate location, the number of employees and the industry type, with NOGA codes, of firms in Switzerland. And regression factors estimated from the Federal Office for Spatial Development.

2.2 Other models of synthetic freight demand

The Freight Analysis Framework. Hwang et al. (2021) is not a research article but a technical report documenting a national statistical product: the freight flow database of the United States. Its goal is to give a complete picture of freight movements between American regions, by commodity and by mode, in tonnes and in dollars, for 132 zones and with forecasts up to thirty years ahead. It is built on the Commodity Flow Survey, a national survey of shippers conducted every five years, which covers about 71% of the total value. Eleven sectors are out of the scope of that survey, among them farm-based agriculture, municipal waste, construction debris, retail and services. They are estimated separately, and together they account for 39.7% of the total tonnage. In addition, more than 40% of the cells of the full origin-destination-commodity-mode matrix are not published, for confidentiality or reliability reasons, although the one-dimensional totals are complete. These missing cells are estimated with an iterative proportional fitting procedure constrained by the known margins. The difference with the project is one of level: this database describes flows between zones, and stops there. It is the type of source the project uses as an input, not the type of output the project aims to produce.

SimMobility Freight. Sakai et al. (2020) present an agent-based simulator of urban freight,

applied to Singapore. Its goal is to evaluate logistics solutions such as delivery consolidation or time window regulations. It is the most detailed model reviewed here: it represents establishments with their location, number of employees, floor area, function type and industry type, and it follows freight from the annual quantity produced or consumed by an establishment, to commodity contracts between a supplier and a receiver, and finally to individual shipments. It then simulates the choice of a carrier, the construction of vehicle tours, the route taken and the parking space used for the delivery. The models were estimated mainly on the 2013 Tokyo Metropolitan Freight Survey, which covers about 43,000 establishments together with their shipment records, and completed with GPS traces and driver surveys covering 16,657 delivery stops in Singapore for the delivery time windows. The structure is close to the one used by the project, which borrows from it the chain from annual quantity to contract to shipment. The difference lies in the data: this level of detail was obtained from a large establishment survey, and the authors state that the parameters they could not estimate, in particular for the choice between full and partial truckloads, were assumed.

Truck tour synthesis from public data. Davis III et al. (2024) propose a method to generate a synthetic population of truck tours for New York City. Their starting point is that simulators such as SimMobility require detailed establishment data that public agencies do not have, so their goal is to obtain a comparable output using public data only. They generate 470,000 candidate tours over 47 industry groups and assign flows to these tours with an entropy maximisation model solved by an iterative balancing algorithm. Their inputs are the number of employees by industry and by zone, freight trip generation formulas taken from a national guidebook, the Freight Analysis Framework for the flows entering and leaving the city, payload factors to convert tonnes into trucks, traffic counts, road network data and speed data by time of day. Each generated tour carries an origin, a sequence of stops, a departure time, a truck class, a delivered weight and estimated emissions. Compared with the project, the modelling unit is the tour rather than the shipment, and the level of information on establishments is lower: employees are known by zone, not by establishment. The authors report an average error of 10.2% against cross-borough traffic counts, state that no data on delivery time windows was available, and describe their model as valid at the mesoscopic level.

From a synthetic population of persons to parcel demand. Hörl and Puchinger (2023) present a pipeline that generates parcel deliveries for the region of Lyon, in order to evaluate the performance of last-mile delivery services. It is the reference where the link between population and freight is the most direct: the synthetic population of persons is not an analogy but the input of the freight model. They start from an existing synthetic population, which describes the location and size of each household, the socio-professional category and age of each person, and their activity chain during the day. They combine it with a local survey on remote purchases, of which only the published margins were available to them. An iterative proportional fitting procedure distributes these averages across households, and a Poisson distribution turns them into a discrete number of parcels per household per day. The activity chains also indicate when household members are at home, which gives the time windows during which a parcel

can be received. The difference with the project is the segment: these are business-to-consumer parcel deliveries, generated by households, whereas the project models business-to-business flows generated by establishments. The authors state that no information was available on the size or the weight of parcels, so all parcels are assumed to be identical.

An establishment-based freight survey. Alho and de Abreu e Silva (2015) report on a survey of 604 retail establishments conducted in the centre of Lisbon in 2013. Unlike the other references, this is not a model but a data collection exercise, carried out in order to estimate the parking demand generated by deliveries. The survey collected the characteristics of each establishment, including its sales area, storage area and number of employees, the number of weekly deliveries and their distribution over two-hour periods of the day, the type of vehicle delivering and the usual duration of a delivery, the parking location used, and a description of the ordering process, including the urgency of deliveries, the interval between orders and who decides the delivery time. It was first attempted online, which produced ten answers in two months, and had to be carried out through face-to-face interviews instead. Quantities are recorded as a number of items, where a bag, a box or a crate counts as one item regardless of its size. Relative to the project, this reference is of a different nature. It does not produce a model. It shows what an establishment survey provides, and the attributes the project cannot implement are the ones such a survey records. The authors also report that establishment size is only weakly correlated with the number of deliveries received, and that 76% of the establishments describe their delivery routine as undefined.

2.3 What these references need

What I take from all of these, the project documents and the five published references together, is a list. About the firms, they need the location of an establishment, its number of employees, its floor area, the function of the site and its sector of activity. About the flow, they need an origin, a destination, a type of goods, an annual tonnage, a mode and a distance. About the individual shipment, they need a weight, a frequency, the vehicle it requires, the way the goods are packed, the hour it is meant to leave and the hour by which it must arrive. And about the decision behind the shipment, they need things only a survey of firms ever records: how urgent a delivery is, how long before it the order was placed, and who chooses the hour.

Two of the published references are explicit about what they could not get. Davis III et al. (2024) state that no data on delivery time windows was available to them. Hörl and Puchinger (2023) state that nothing was known about the size or the weight of a parcel, so all parcels are assumed identical. These admissions are as useful as the lists themselves, because they show that the gap I was asked to describe for Switzerland is not a Swiss peculiarity.

This list is the raw material of the next chapter, where it is grouped and cut down to the variables a survey of vehicles could plausibly carry.

Chapter 3

Methodology

This chapter describes how I looked for data. The work fell into four phases, which I number here and refer to by number in the rest of the report. I did not plan them in advance. Looking back, this is the order in which they happened.

3.1 Phase 1: from reading to a list of variables

I started from the literature rather than from the data. Reading the references of Chapter 2, I asked each of them the same question: what data did it need in order to run? The answer was easier to find than I expected, because every one of these papers says at some point which information it had and which it had to assume. The list that came out of that reading is the one given at the end of Chapter 2.

I then grouped these variables by the stage of the model they serve. The model described in Chapter 2 works in three steps, so the three groups follow it: who ships and how much, who ships to whom, and how the shipment leaves.

The Swiss freight statistics are a survey of vehicles and of the journeys they make. Anything that describes a decision, a contract, or the internal workings of a firm cannot appear in such a survey, whatever its quality. So the whole ordering process went out, and with it the hard or soft nature of a deadline, the delivery time window, both forms of consolidation, the floor area and the function of the site.

Eight variables were left, and they are the grid used from here on. Serving the first step of the model, who ships and how much, there are three. The *shipper and receiver*, meaning who sends the goods, who takes delivery, and where each of them is. The *commodity type*, meaning what kind of goods travel. And the *quantity*, meaning how much leaves a place over a year. Serving the second step, who ships to whom, there is one: the *origin and destination* of the flow. Serving the third step, how the shipment leaves, there are four. The *shipment size*, meaning the weight of one shipment rather than the load of a truck. The *mode*, meaning whether the goods go by road, by rail, or by a combination of the two, which is the choice the model has to represent. The *timing*, meaning when the shipment leaves. And the *vehicle and container*, meaning the kind of vehicle the shipment requires and the way the goods are packed.

The three versions of the attribute list, with seventeen attributes wanted, then eight modelled, then six implemented, come from the project documents (Bierlaire and Ilinov, 2025; Ilinov, 2025, 2026). I did not establish them, I read them.

3.2 Phase 2: finding the catalogues

Swiss freight data is published by the Federal Statistical Office in an online catalogue organised by theme. Road freight carried by heavy vehicles forms one theme, and rail freight forms another, which already gathers rail, combined transport and transalpine traffic together. Seeing the two themes side by side, and seeing how much each of them held, I decided to build two separate inventories rather than one.

I also decided to inventory everything rather than pick out a few tables I had heard of. It was not a methodological choice. I simply could not find my way otherwise. It turned out to matter, because describing the tables that exist does not show which ones are missing, and the question I had been given was about what is missing.

3.3 Phase 3: building and ranking the inventory

I knew what I wanted the inventory to look like before I knew how to build it. One line per statistical product rather than one line per file, and on each line the things I kept having to open a file to find: the period, the measured variable, the geography, the form, the vehicle universe and the publication date. Reading the catalogue page by page would not have produced that in the time I had, so I built it with the help of a large language model. I described what I wanted, it extracted the catalogue through the public interface, which returns every published object with its title, its period and its download link, and it filled the columns from those titles. A few more columns were proposed to me and I kept them.

The road theme gives 699 objects, of which 592 are tables. The three rail, combined transport and transalpine themes give 103 entries covering 99 distinct objects, four definitions being filed under two themes at once, and 28 of them are tables. That is 798 distinct objects described, and 620 tables. Files republished every year were then grouped into series, which gives 89 series for road and 35 for rail, 124 in total.

What I did not have was a way of ranking them. I asked for the tables that are recent and that carry the most information at once. The formula that came back gives most weight to two things, whether a table says where the goods go and whether it says which goods they are, and much less weight to two others, being recent and having a long history. On the road side it adds three points for an origin-destination matrix, three for an explicit commodity nomenclature, two for a coarser breakdown by kind of goods, one for an edition published in 2023 or later, and one for a series running over at least ten editions. The road scale therefore stops at eight, and nothing in the catalogue reaches it. The best score is six, and the two points that are always missing are the same two. No table that crosses place with commodity is recent, and none of them has been published for long.

The first two items are conditions rather than qualities. Without a geography, a flow cannot be placed anywhere. Without a commodity breakdown, a model that treats each commodity separately has nothing to feed any of its modules. Being recent is comfortable, but a recent series without either dimension is of no use.

The rail inventory needed a different rule, because no rail series carries both a matrix and a commodity breakdown, so the road formula would have flattened the whole catalogue. It gives three points for a matrix, two for a commodity group, one for a breakdown by country, one for a recent edition, one for a table rather than a graph, and one for a series of at least fifteen years. Nothing on the rail side reaches more than four. Scores from the two inventories are therefore not comparable with one another.

Everything in these two files is therefore read from titles, and none of it is read from the files themselves. I knew that while I was building them, and I wrote it into the inventory as a warning. Checking it is the next phase. The two inventory files are in the appendix.

3.4 Phase 4: opening the files

A classification read from titles is a claim about what a file contains, not an observation of it. Swiss statistical titles are explicit and usually enumerate the dimensions they cross, but a title can be incomplete, and it can also be misleading. I therefore opened files and compared their contents with their description.

I opened twenty of the 124 series. The number is a consequence of the time available rather than of a rule. I did not choose the twenty best. I chose a spread: the series that carry both a geography and a commodity breakdown, the two files I had already used for the poster presented at the end of the internship, several that had never been opened, and several ranked near the bottom of my own list, in order to find out whether the ranking was wrong.

Each file was read against the same grid. What does it actually contain, is the announced crossing real or are two separate totals placed side by side, which nomenclature and at which level of detail, how fine are the zones, which unit, which vehicles are covered, what is published about reliability, does the file hold a dimension its title does not announce, and what has to be computed before anything can be used. And for each of the eight variables, whether the file carries it at all, and if so at what level.

Chapter 4

Results

The inventory itself is the first result, and its shape matters more than its size. Of the 592 road tables, a hundred are matrices that say where the goods come from and where they go. On the rail side there are six, all of them in combined transport, and only two measure the rail leg itself, the other four measuring the road legs that feed it. Conventional wagonload freight has no origin-destination table at all. Everything else in both catalogues is a national total broken down along one or two dimensions.

The first file I opened in detail is the one I later used for the poster. It reports goods carried by type of goods and by distance, for Swiss heavy vehicles in 2024. It crosses the twenty commodity groups of the current nomenclature with six distance classes, from under 25 kilometres to more than 200. The point that matters is easy to miss. The file contains no distance at all. It contains tonnages placed in distance bands.

To obtain one number per commodity I had to build it. I took the midpoint of each band, weighted it by the tonnage sitting in that band, and divided by the total. The last band is open, above 200 kilometres, so it has no midpoint and I had to choose one. I took 250. The result for all goods is 41.1 kilometres. Taking 225 instead gives 40.6 and taking 300 gives 42.0, and that band carries under two per cent of the tonnage, so the choice is not what drives the answer. The assumption that does more work is the one I could not avoid, that tonnes are spread evenly inside each band.

Later, in a series I had not put among the twenty, I opened a file that reports tonnes and tonne-kilometres side by side, by commodity group, for the same year and the same vehicles. There the average distance is a division rather than a reconstruction, and it comes to 40.6 kilometres over all traffic, international journeys included. The two figures do not cover exactly the same thing, since mine is domestic only, but they sit close to one another and the second one took a minute to obtain. I had spent an afternoon rebuilding something of the same order that was already published two tables away. A classification read from titles tells you which dimensions a table crosses. It does not tell you that a quantity you were about to estimate is printed in full somewhere else, and my own shortlist had not even kept that table.

The remaining files were read against the same grid. Table 4.1 lists all twenty, with the score my inventory had given them from their title alone and what each of them turned out to hold.

Table 4.1: The twenty series opened. The score is the one my inventory assigned from their title, before any of them was opened, and a higher score means a table looked more promising. The road and the rail inventories use different scoring rules, described in Section 3.3, so a rail score is not comparable with a road score. Matrices give flows between places, breakdowns give national totals split along one or two dimensions. The rail files are numbers 10, 11 and 16.

No.	File	Year	Score	What it holds
<i>Origin-destination matrices</i>				
1	B18	2003	6	The only Swiss-vehicle matrix crossing region with commodity
2	C18	2003	6	The foreign half of the same year, so the two add up
3	C18	2014	6	The same crossing, superseded by its own 2019 edition
4	C18	2019	6	The most recent region by commodity matrix, foreign vehicles only
5	A11	2019	6	The crossing at country level, both fleets together
6	C20	2019	6	Subtracted from the previous one, it rebuilds the Swiss fleet
7	B18	2024	5	A recent regional matrix, with no commodity at all
8	B20	2024	5	Imports and exports by country, no flow between two foreign countries
9	P18	2016-20	3	The finest zones of the twenty, canton by canton
10	K124	2016-22	4	The only rail matrix in tonnes, combined transport only
11	K104	2016-22	4	The same in intermodal units, which gives tonnes per unit
<i>Breakdowns</i>				
12	B26	2024	2	Tonnage by economic sector, without saying whose sector
13	B26	2013	2	The note that answers it, the sector is the vehicle holder's
14	B26	2019	2	Redundant with 2013, apart from a control total
15	B02	2024	1	Payload, empty running, and the average load of a heavy vehicle
16	TP-ZR	1990-2025	3	The only rail commodity breakdown, hidden in a public transport file
17	B22	2024	4	Distance by commodity, domestic traffic only
18	B12	2024	4	Domestic, import and export by commodity
19	B13	2024	2	How the goods are packed, containers and pallets
20	E25	1993-2024	1	The annual control total, both fleets, transit included

Opening the files changed how I read my own ranking. The score was computed from the titles the Federal Statistical Office publishes with each table, which came down with the catalogue itself, and from nothing else. It measured what a table announces, not what it holds. That was a deliberate choice, since reading 124 series any other way was not possible in the time I had, and the inventory carries it as a warning. The two turned out to drift much further apart than I thought they would.

The files I placed at the top do contain what they announce. The score never asked which

trucks a table covers, how recent it is, or whether it holds something its title does not mention. Four of the six best-scored files describe foreign trucks, which carry about one tonne in seven, and the only one of the six that covers Swiss trucks on their own is from 2003.

The drift also runs the other way. The one file that let me compare the shipment weight the model assumes with a Swiss figure scored one, the lowest value in the set. A score built from titles tells you what a title announces, and nothing about what a table is worth.

The titles were not always dependable either. Four out of the twenty describe their content inaccurately. A file listed under public transport statistics turned out to hold the only table breaking rail freight down by commodity group in the whole rail catalogue, over a series running from 2008 to 2025. A matrix between cantons is described in the catalogue as cumulated over five survey years, while the file itself gives a yearly average of those years, which is not the same quantity. A matrix announced as being between countries holds no flow between two foreign countries. And a table of performance by economic activity does not say, in its 2024 edition, whose activity is being recorded.

My own inventory was wrong twice as well, and it took opening the files to see it. Three matrices I had filed as crossing Swiss regions with foreign countries also hold the flows between Swiss regions, the more useful half of what is in them. And I had recorded that the rail catalogue carries no detailed commodity breakdown, which the public transport file contradicts. Both mistakes come from the same place as the inaccurate titles, which is reading a title instead of a file.

The one table that crosses Swiss regions with commodity for Swiss vehicles is reproduced in Table 4.2. Each column shows what a region loads, as a share of everything it sends. Building materials come first in every region without exception, from 48.4 per cent in the North-West to 73.5 per cent in Ticino, and what a region adds on top of that base is what distinguishes it. The North-West is the only region above five per cent in chemicals, which fits the industry around Basel. Two cautions come with the table. The file publishes a confidence interval for each cell, and the Ticino column is the least certain of the set, so its smaller figures should be read as clearly below average rather than as precise values. And the data are from 2003, in a nomenclature of ten chapters that has since been replaced by one of twenty groups, so the rows cannot be matched to any recent file line by line.

I expected a lot from one file, and it turned out not to be what I thought. Swiss firms are coded by sector of activity, in the same nomenclature the business register uses, and the catalogue holds a series that reports freight by sector. On paper that is the only place where the freight statistics and the register speak the same language. The first stage of the model has to decide how much freight a firm of a given sector sends, and a table of that kind would say it.

The 2024 edition does not say whose sector it records. The 2013 and 2019 editions do, in a note: the sector is the one of the holder of the vehicle. So when the truck belongs to a transport company the tonnage counts under transport and storage, whatever was inside, and that section holds about half of the total. The tables also stop at the twenty-one broad sections, where the register goes finer, and none of them separates firms carrying their own goods from firms carrying other people's. What I was looking for and what these tables record are not the same

Table 4.2: What each region loads, by commodity, in 2003. Share of each commodity in what a region loads, in per cent, Swiss heavy goods vehicles, computed from the published tonnages. The last column covers all traffic carried by Swiss heavy vehicles, including the small share loaded abroad. One chapter with negligible tonnage is omitted.

	Lake Geneva	Mittel-land	North-West	Zurich	East	Central	Ticino	Swiss average
Agricultural products	4.6	6.0	4.2	2.7	4.5	4.9	1.4	4.6
Food and fodder	14.4	14.1	10.7	9.0	14.2	12.4	7.5	12.2
Petroleum products	6.8	3.7	7.3	4.1	5.2	2.6	4.8	4.8
Ores and metal waste	0.4	0.5	1.4	0.4	0.8	0.4	0.5	0.7
Metal products	1.4	1.3	3.0	1.1	2.3	0.7	0.8	1.7
Building materials	54.0	55.6	48.4	64.3	51.2	60.0	73.5	55.7
Fertilisers	0.2	0.1	0.4	0.2	0.6	0.1	0.1	0.2
Chemicals	3.4	1.8	5.4	1.0	1.8	1.6	1.1	2.8
Machinery and manufactured goods	14.8	16.9	19.1	17.3	19.3	17.5	10.4	17.3
Total	100	100	100	100	100	100	100	100

thing.

A second file shows the gap between what is counted and what the model needs, and it does so with two numbers that look comparable and are not. The file reports performance by vehicle type and payload class, and from it the average load works out at 8.64 tonnes, obtained by dividing tonne-kilometres by the vehicle-kilometres run with a load. An earlier version of the model set an average shipment weight of 13 tonnes, taken from Danish data for want of a Swiss figure (Ilinov, 2025). These two quantities describe different objects. One is what a heavy vehicle carries on a journey, the other is what one shipment weighs, and a vehicle can carry several shipments. None of the twenty files allows one to be converted into the other.

The twenty files answer the eight variables in three different ways. Five of them can be read somewhere, who sends and who receives, the type of goods, the tonnage, the origin and destination of the flow, and the vehicle. None is read at the level the model needs, because all of it is published as totals. Senders and receivers are known by region or by canton, never as a firm. The commodity mix is national. The origin and destination are crossed with the type of goods only in the 2003 matrix. The vehicle is described per journey, never per shipment.

The sixth variable, the mode, is a case of its own. The mode is the means of transport, road or rail or a combination of the two. Every file covers one of them and says so in its title, so the mode of a published flow is never in doubt. The difficulty is that the model has to choose between road and rail, and choosing needs the same flow measured both ways. No file among the twenty does that. The two rail matrices carry no commodity breakdown, the one rail table that carries a commodity breakdown has no geography, and nothing links either of them to a road table.

The last two variables are not in the files. The weight of an individual shipment appears in none of the twenty, and neither does the hour at which a shipment leaves, although the quarter of the year does. Three more things the references of Chapter 2 asked for are absent as well: the frequency of shipments, the size of the receiving firm and the value of the goods.

Chapter 5

Discussion and conclusion

The question I set out with was whether Swiss public statistics can supply the variables a disaggregate freight model needs. The honest answer is not yet, and the summer leaves a much clearer picture of what the catalogue holds and where it runs out.

It is worth being fair to the catalogue before being critical of it. Several of the eight variables can be read from it without difficulty. The spatial structure of road freight is published every year at the level of the seven major regions and, in a pooled form, at the level of the cantons. The commodity mix is published every year in twenty groups. The distance profile is published by commodity, and the vehicle side is well covered too. For a model that reasons about flows between zones, this is a reasonable foundation.

All of that describes the road. Rail freight is not published in the same way, and there is far less of it. The road catalogue offers 592 tables and a hundred origin-destination matrices. The rail, combined transport and transalpine catalogues offer twenty-eight tables and six matrices between them, none of which says which goods are moving, and the single rail table that does say it has no geography and files more than half of the tonnage it reports for 2025 under goods that could not be identified. Within this catalogue, then, a model that has to choose between road and rail would be calibrated on one side and assumed on the other. I did not look beyond the catalogue, and my own rail inventory names three places where more might be found, the interactive database of the statistical office, the Federal Office of Transport, and the railway companies themselves. The imbalance is therefore firm for published tables and open for the rest.

The shipment is where it stops. None of the twenty files describes one. Every road file among them counts vehicles and the journeys they make. That is not an oversight. The statistic is a survey of vehicles, so a shipment is not an object it was built to observe. The gap this leaves is a real one. The 2024 files give an average load of 8.64 tonnes for a heavy vehicle, while the model assumes a shipment weighs 13 tonnes, and the two cannot be compared, because one vehicle can carry several shipments.

The way forward is the one the literature already takes. What cannot be read has to be generated, and the generated values have to remain consistent with the totals that are published. The Freight Analysis Framework, described in Chapter 2, fills more than forty per cent of the

cells of its national matrix in exactly this way, constrained on margins that are complete even where the crossings are not. The parcel model built for Lyon does the same at household level, and the truck tour synthesis built for New York at the level of a tour. A disaggregate model for Switzerland therefore needs to generate plausible shipments while remaining consistent with observed aggregate totals.

Two numbers in the model come from outside the current statistics, one from an older study and one from abroad, and Swiss data can be held against one of them.

The first is the distance a shipment travels. The project documents take it from a Swiss study of 2008, which reports an average of 42 kilometres and almost three quarters of internal shipments under 50 kilometres (Ilinov, 2026). The 2024 files give 40.6 kilometres, and 76 per cent of domestic tonnage at 50 kilometres or less. The two are close enough for the borrowed figure to hold, with one caution, since the study counts shipments and the file counts tonnes. And the same file gives that distance commodity by commodity, something the model does not have today.

The second is the weight of a shipment, set at 13 tonnes from Danish data (Ilinov, 2025). No Swiss file measures that object, as the average load of a heavy vehicle has just shown, so this figure cannot be checked against anything. It can only be left in place and flagged.

My own work has four limits that a reader should keep in mind. The classification of the catalogue was read from titles rather than from file contents. Four titles out of the twenty I checked turned out to be inaccurate, and my own coding of three others was wrong in the same way. The part I never opened probably holds more of both. I opened twenty series out of 124, so 104 series and the greater part of 620 tables have not been looked at, and the twenty were chosen by me rather than drawn at random. The grid of eight variables is my own construction from six sources, and another reader would have cut it differently. And the extraction of the catalogue and the first pass of classification were done with a language model, with my criteria and my checks, but not by hand.

Part of what is left to do is simply to continue. The inventory holds 104 series I did not open, and the four inaccurate titles I did find work in both directions, since a file can also hold a dimension its title does not announce. Some of the variables I recorded as absent may therefore be answered somewhere in the tables nobody has looked at yet, and the inventory was built so that the next person can go straight to them.

The gap that remains is at shipment level, and closing it would take a different kind of source. Lisbon's establishment-based freight survey, described in Chapter 2, records the weekly number of deliveries, the hour they arrive and the process behind the order. It also shows what such a survey costs. Its online form produced ten answers in two months and had to be replaced by face-to-face interviews. And even then it would not solve everything, since it records quantities as a number of items rather than in tonnes, and it measures only a weak relation between the size of an establishment and the deliveries it receives, which is the assumption on which the generation step of every model reviewed here rests.

Looking back, what I take from the summer has less to do with freight than with how I work. I had to learn what doing research actually involves: organising a project over two months,

keeping track of what I had already done, and setting my own criteria rather than applying someone else's. I also had to understand a subject I knew nothing about, which meant reading papers for what they needed rather than for what they concluded. And I spent time stuck, first on the catalogue and then on the files, long enough to learn that being lost is not the same as being blocked.

Chapter 6

Acknowledgements

I would first like to thank my supervisors, Pavel Ilinov and Candice Baud, who guided me throughout the summer. I am grateful to Professor Michel Bierlaire for welcoming me into the Transport and Mobility Laboratory, and to my colleagues there for the kindness they showed me.

A special thanks goes to the Laidlaw Foundation and to the Laidlaw team at EPFL, especially Dan-Thi Nguyen. This internship, and the two months of research that came with it, would not have taken place without them.

Finally, thank you to the people I met in the cohort, for their company and friendship.

Bibliography

- Alho, A. R. and de Abreu e Silva, J. (2015). Lisbon’s Establishment-based Freight Survey: revealing retail establishments’ characteristics, goods ordering and delivery processes. *European Transport Research Review*, 7(2):16.
- Bierlaire, M. and Ilinov, P. (2025). Incorporating decision-making into the Freight Demand Model for Switzerland. Internal project document, Transport and Mobility Laboratory, EPFL, September 2025.
- Davis III, H., Landes, H., Namdarpour, F., Yang, H., Chow, J. Y. J., and Ozbay, K. (2024). Aggregate urban truck tour synthesis from public data. *Transportation Research Part A: Policy and Practice*, 185:104107.
- Hörl, S. and Puchinger, J. (2023). From synthetic population to parcel demand: A modeling pipeline and case study for last-mile deliveries in Lyon. *Transportation Research Procedia*, 72:1707–1714.
- Hwang, H.-L., Lim, H., Chin, S.-M., Uddin, M., Biehl, A., Xie, F., Hargrove, S., Liu, Y., and Wang, C. (2021). Freight Analysis Framework Version 5 (FAF5) Base Year 2017 data development technical report. Technical Report ORNL/TM-2021/2154, Oak Ridge National Laboratory.
- Ilinov, P. (2025). Synthetic Population of Shipments. Attributes. Internal project document, Transport and Mobility Laboratory, EPFL, December 2025.
- Ilinov, P. (2026). Synthetic Population of Shipments. Vol 2. Internal project document, Transport and Mobility Laboratory, EPFL, May 2026.
- Sakai, T., Alho, A. R., Bhavathrathan, B., Dalla Chiara, G., Gopalakrishnan, R., Jing, P., Hyodo, T., Cheah, L., and Ben-Akiva, M. (2020). SimMobility Freight: an agent-based urban freight simulator for evaluating logistics solutions. *Transportation Research Part E: Logistics and Transportation Review*, 141:102017.

Appendix A

Appendix

A.1 Inventory files

Two spreadsheets accompany this report.

- `Inventaire_OFS_transport_marchandises_route.xlsx`. The road freight catalogue. Sheet *Series* holds the 89 series with their score and tier and is the working sheet. Sheet *Detail* holds the 592 tables, one line per year, with a download link for each. Sheet *Couverture* crosses geography with commodity breakdown.
- `Inventaire_OFS_fret_ferroviaire.xlsx`. The rail, combined transport and transalpine catalogue, with the same structure, 35 series and 103 entries.

A.2 The twenty files, full grid

The summer produced far more material than a report can carry. Besides the two inventories, I have the twenty files themselves, downloaded and kept as they were published, and the grid in which I read them, one line per file, with its structure, its nomenclature, its spatial resolution, its unit, its vehicle universe, what it publishes about reliability, and the answer given for each of the eight variables. Putting all of that here would have made the report unreadable, so it stays in the working folder, and whoever needs it can have it.

A.3 Calculations

All figures quoted in Chapter 4 are computed from the published files. The three that involve a choice are recorded here.

Average haul distance from distance bands. Tonnages in each of the six domestic distance bands are multiplied by the midpoint of the band and the sum is divided by the total domestic tonnage. Midpoints used: 12.5, 37.5, 75, 125, 175 and 250 kilometres. The last band is open

above 200 kilometres and 250 was chosen. The result for all goods is 41.1 kilometres. Using 225 gives 40.6 and using 300 gives 42.0.

Average haul distance read directly. Tonne-kilometres divided by tonnes, by commodity group, for the same year and the same vehicles. This covers all traffic, international journeys included. The result for all goods is 40.6 kilometres.

Average load of a heavy vehicle. Tonne-kilometres divided by the vehicle-kilometres run with a load, which are themselves the total vehicle-kilometres less those run empty. All three quantities are published in the same table, for Swiss heavy vehicles in 2024. The result is 8.64 tonnes. The choice here is the denominator. Dividing by all vehicle-kilometres, empty running included, gives 6.3 tonnes, which is not the load of a vehicle that is carrying something.