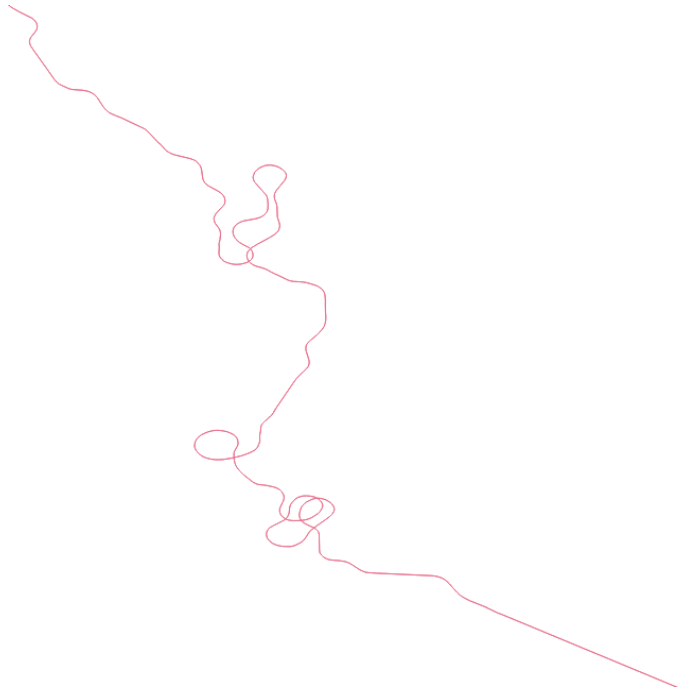


Geodata-cleaning of the Swiss Mobility and Transport Microcensus 2021

Matthias Balmer

Antonin Danalet

Nicole A. Mathys

STRC conference paper 2022 **Mai 11, 2022**

STRC | **22nd Swiss Transport Research Conference**
Monte Verità / Ascona, May 18-20, 2022

Geodata-cleaning of the Swiss Mobility and Transport Microcensus 2021

Matthias Balmer
Fundamental Policy Questions Section
Federal Office for Spatial Development ARE
CH-3003 Bern
matthias.balmer@are.admin.ch

Antonin Danalet
Fundamental Policy Questions Section
Federal Office for Spatial Development ARE
CH-3003 Bern
antonin.danalet@are.admin.ch

Nicole A. Mathys
Fundamental Policy Questions Section
Federal Office for Spatial Development ARE
CH-3003 Bern
and
University of Neuchâtel
CH-2000 Neuchâtel
nicole.mathys@are.admin.ch

Mai 11, 2022

Abstract

The Swiss government regularly conducts the Mobility and Transport Microcensus (MTMC). It captures the travel behavior of the Swiss population in the largest scale and most comprehensive mobility survey for Switzerland. The MTMC is therefore a unique and crucial input for policymaking and research. More than 193'000 trips based on 57'000 persons were recorded and routed in 2015. In this paper, we give an insight into the systematic quality control of the 2021 MTMC geodata. We show the adjustments and control mechanisms applied. We focus on the decomposition of trip legs into their domestic and foreign part, the verification of the routing distance and the handling of distances estimated by the respondents. With this paper, we increase transparency of the data-cleaning process, show the necessary data adjustments and encourage constructive contributions from researchers. A challenge is to clean the data (defining part of the data as not plausible, imputing information and removing impossible events) without removing rare events and adding biases (e.g., by filling a missing part of a path using the shortest path). We hope this paper eases the work with the data and provides clarity: MTMC data available to researchers are not raw data.

Keywords

Swiss Mobility and Transport Microcensus, MTMC, transport behaviour, data cleaning, geodata, data cleansing, microrecensement mobilité et transports, MRMT, Mikrozenzus Mobilität und Verkehr, MZMV

Suggested Citation

Balmer, M., Danalet, A. and N. A. Mathys (2022) Geodata-cleaning of the Swiss Mobility and Transport Microcensus 2021, paper presented at the 22nd Swiss Transport Research Conference, Monte Verità, Ascona, Switzerland

Picture on the title page: Albula Railway

Acknowledgements

The authors would like to thank Hanja Maksim and Jean-Luc Muralti, from the Federal Statistical Office (FSO), for the discussions and helpful comments. We also would like to thank our colleagues at the Federal Office for Spatial Development (ARE)) who gave us feed-backs.

Contents

List of Tables

List of Figures

1 Introduction

In this paper, we give an insight into the quality control and plausibility checks of the geodata in the Swiss Mobility and Transport Microcensus (MTMC). We show the adjustments and control mechanisms. We focus on the trip legs[1] and discuss the decomposition of trip legs into their domestic and foreign parts, the verification of the routing distance and the handling of estimated distances.

The aim of the plausibility checks is to find wrong or inconsistent data in order to obtain an unbiased and proper final data set. In this sense, the data available to researchers, cantons, practitioners and other federal offices (after signing a data protection contract) are not raw data. The data have been cleaned and made plausible. With this paper, we increase transparency and show the data adjustments performed. The paper makes clear which spatial variables are changed and how. This should facilitate the work with the data and the interpretation of results by researchers.

1.1 Structure of the paper

We divide the paper into four main parts. We give a brief insight into the MTMC in Section 2. Questionnaire, interview modalities, first plausibility checks directly in the interview and the data structure are presented.
Section 3 then gives an overview of the plausibility checks. We distinguish between systematic and manual corrections.
In Section 4, we go into more depth on checking the trip legs.
We end the paper with selected challenges (Section 5) and the conclusions (Section 6).

1.2 Motivation for cleaning the geodata

In the end, we want to have accurate and precise data, reflecting real world mobility behaviour. The distance and the number of trips are particularly important to us. Optimally, we record the effective distance and the actual number trip legs (see Footnote 1).

[1]Here we use the term "trip legs" to translate the German term "Etappen" ("étapes" in French). Since there is no exact English translation, we use this term. The concept of trips legs is explained in chapter 3.2.3 in BFS and ARE (2017).

To come close to this, we check the data at various points. First, directly in the interview (see Section 2.3). Then, we check the data continuously during the survey year to detect possible systematic errors. These may include problems in the routing tool or questions that are misinterpreted during the interviews. Depending on the problem, it could be corrected directly or integrated in the next MTMC edition. Finally, we check the data after the interviews. This paper focuses on the post-interview geodata-cleaning. The systematic data-cleaning thus allows coherence and minimizes errors.

Given the size of the survey, we can not manually check all the details and we have to automatize parts of the process. For that, we rely on threshold values and multiplicative factors. We define limits within which we believe that data are consistent and plausible. If values are beyond the chosen limits, an initial correction is applied automatically. The distinction between wrong data and rare cases is particularly difficult. After correction, we validate the data again and observations remaining implausible are checked manually.

1.3 Ensuring consistency of the time series

To ensure consistency in the time series of the MTMC, we are not free in the definition of plausibility checks and the determination of the limits. The performed checks and thresholds have therefore largely been taken over from previous editions of the MTMC. The thresholds and methods used in 2015 are applied as far as possible, so that differences between 2015 and 2021 are due to differences in the data and not in the methods. The publication of the data-cleaning process is done for the first time with the present paper.

2 The Mobility and Transport Microcensus (MTMC)

The MTMC is the largest scale and most comprehensive statistical survey of the travel behaviour of the Swiss resident population. It is conducted every five years by the Federal Statistical Office (FSO) and the Federal Office for Spatial Development (ARE). The market and social research institute LINK is responsible for the data collection in the 2021 edition. With a sample of more than 55'000 persons and a long list of questions, this is the largest national-level survey about travel behaviour in Switzerland. This survey

has been performed since 1974[2]. Data were collected most recently in 2015 (see BFS and ARE (2017) for further information), 2020 (briefly, see below) and 2021.

The 2021 MTMC survey had been actually planned for 2020. At the beginning of 2020, the personal interviews could take place as planned. However, when public life largely came to a standstill in March 2020 in the wake of the first wave of the Covid pandemic and public transport services were severely restricted, the survey had to be interrupted and postponed for a year. Due to the termination in March 2020 and the restart of the survey in January 2021, MTMC data from two directly consecutive years (before and after the start of the pandemic) are available for a period of several weeks. A comparison of the two data sets can be found in BFS and ARE (2021).

2.1 Content and questionnaire

The MTMC contains questions about:

- the socioeconomic characteristics of households and individuals
- mobility tools (vehicles and public transport season tickets)
- daily mobility (trips on a given reference day)
- occasional journeys (day trips and trips with overnight stays)
- preferences for transport policies in Switzerland (see Danalet *et al.* (2022)).

A short version of the questionnaire can be found in BFS (2017). For each trip leg carried out on the reference day and among lots of questions, respondents are asked about the departure and arrival times, the transport mode chosen, the destination (geocoded using a GIS tool), the type of activity at destination, etc.

For the 2021 survey, about 55'000 persons, representative of the Swiss resident population and randomly selected within geographic quotas, were surveyed by computer-assisted telephone interviewing (CATI). The trip legs on the reference day of the interviewees are routed directly during the interview (see Section 2.2). This also allows the verification of the distances during the interview (see Section 2.3) and after (see Section 4.4).

[2]For a detailed history of the MTMC until 2000, please see Simma (2003).

2.2 Routing during the phone interviews

Since 2010, the routes of the trip legs on the reference day are collected as geodata in the MTMC (BFS and ARE, 2017; BFS, 2018). Based on the information given on the phone (transport mode, start and arrival location of the trip leg, start and arrival time of the trip leg), a route is computed, verified on the phone with the person and, if needed, corrected. Ohnmacht *et al.* (2012) discuss the methodological advances of the route-recording in detail. In 2015, for more than 90% of the trip legs a route was found[3]. For cars the percentage of found routes is even higher (approx. 98%), for public transport significantly lower (approx. 80%).

For trip legs on foot, there is generally no verification. If it is a round trip (same departure and arrival location, mostly home, e.g., going for a walk), the interviewer asks for the most distant point and defines it by clicking on the map or by selecting a predefined geolocation in a list (e.g., an address).

For trip legs by bike, the most attractive and the fastest routes are computed based on travel time and slope. The interviewer picks the one corresponding the most to the actually performed route. Then, the interviewer asks for two verification points on the route. If a verification point described by the interviewee is not already on the route, the route will be computed again to include the verification point.

Quite similarly, for trips by cars and motorbikes, the fastest and shortest routes are computed, the interviewee chooses between the two routes and gets asked for two verification points along the route. These two verification points are only asked for routes longer than three kilometers.

For trips by public transport, all trip legs are routed using the official time table.

2.3 Online data-cleaning

The live routing allows for a real-time check of the distances recorded directly in the interview. In case of implausible data, the interviewers have the possibility to inquire directly. In this way, incorrect information can be significantly reduced.

[3] "A found route" means that a route ID was present, regardless of whether the route itself was plausible or not.

All trip legs by car, motorcycle, bus, train, streetcar and bicycle are routed during the interview. In the following cases, an estimate on the distance traveled is additionally requested from the respondents:

- for all trip legs by bicycle and on foot
- for trip legs by car, motorcycle, bus, train, streetcar, an estimated distance is requested if:
  - the routing did not work,
  - the routed distance is less than 3km (for motorized private transport),
  - the departure or arrival address is not precise,
  - the trip leg crosses the border,
  - the trip leg is a round trip,
  - the routed distance is not plausible.

In the last case, plausibility is defined by computing the detour that the respondent did in comparison with the distance as the crow flies: if the routed distance is too large in comparison to the distance as the crow flies, it is defined as not plausible enough to avoid directly asking the respondent about the distance. More precisely, if the routed distance d_{routed} is in the range $d_{asthecrowflies} < d_{routed} < d_{asthecrowflies} * df_{mode}$, where $d_{asthecrowflies}$ is the distance as the crow flies (great-circle distance) and df_{mode} is the detour factor, the routed distance is considered plausible and the estimated distance is not asked. In 2015, an estimated distance was asked for just under 50% of the trip legs by car.

The detour factors *df* for the MTMC 2021 are shown in Table 1. The calculation of the detour factors is based on Chalasani *et al.* (2005). They are calculated according to categories of routed distances and recalculated based on MTMC 2015 data. The detailed results and the code are available on github (Danalet, 2020).

Detours taken can be justified by many factors: natural obstacles, the desire to take a less congested or faster route or the desire to choose a more pleasant route (e.g., sightseeing, tourism). By asking the respondent to estimate the distance, we do not prevent the possibility to make such detours. We only check if the transport mode and the departure and arrival time of the trip leg are coherent with such a detour. In some cases, the coordinates of the departure or arrival point or the verification points along the route are wrong. Problems might also be related to missing links in the network used by the routing tool. To check if the distance is coherent with the mode and departure and arrival times, we compute the speed.

Table 1: Detour factors (correction factors for distance as the crow flies)

Routing distance (in km)	$df_{\text{individual modes}}$	$df_{\text{public transport}}$
between 0.5 and 5 km	1.43	1.31
between 5 and 10 km	1.5	1.37
between 10 and 25 km	1.63	1.35
between 25 and 50 km	1.84	1.37
between 50 and 75 km	1.82	1.31
between 75 and 100 km	1.74	1.39
more than 100 km	1.93	1.65

Source: Danalet (2020) based on Chalasani *et al.* (2005)

We verify the speed corresponding to both, the routing and the estimated distances. Speed is calculated using the distance and the duration of the trip leg (using the departure and arrival times, see Section 2.1). In general, the speed limits have been deliberately set at high levels to take into account possible inaccuracies in the network used (TomTom for roads, networks used in the Swiss national passenger transport model (NPTM) for rail and bike). The applied speed limits can be found in Table 2. The limits are based on the values of the past MTMCs and were agreed upon in the involved offices of the Federal Department of the Environment, Transport, Energy and Communications (DETEC).

Generally, the routed distances are used for the calculation of the final distances used in the MTMC. The routed distances are preferred over estimated distances. The error in routed distances is considered consistent with the respondents' estimated distances, whose bias is idiosyncratic. To our knowledge, estimated distances are not used by those who order the raw data (when routed data are available). For the offline plausibility check of the data (see Section 4.4 for more details) we however use the estimated distances.

2.4 Data structure

Figure 1 shows the simplified data structure. The structure also serves as the basis for the subdivision of the plausibility check. Each data set is briefly described below. Since the variable names are originally in German we also provide the German titles.

Households (Haushalte): The top level of the data structure is formed by the *households*.

Table 2: Speed limits for verification of routed and estimated distance

		Error	Warning				
		Limits	<2 km	2-5 km	5-10 km	10-25 km	>25 km
Walking	Min	0.2	0.8	1	1	1	1
	Max	30	10	10	10	10	10
Bicycle, slow Ebike	Min	0.2	4	5	5	5	5
	Max	40	30	30	30	30	30
Fast Ebike	Min	0.2	4	5	5	5	5
	Max	60	40	40	40	40	40
Light Motorcycle	Min	0.2	5	10	10	10	10
	Max	60	50	50	50	50	50
Motorcycle	Min	0.2	10	15	20	20	20
	Max	200	60	80	80	100	100
Car, Truck, Coach	Min	0.2	10	12	15	20	25
	Max	200	60	80	80	100	100
Bus, Tram	Min	0.2	4	4	5	10	10
	Max	150	40	50	60	60	60
Train	Min	0.2	5	5	10	15	20
	Max	400	60	80	120	120	120
Boat, Cable car	Min	0.2	1	1	1	1	1
	Max	150	80	80	80	80	80
Plane	Min	0.2	50	50	50	50	50
	Max	2000	1000	1000	1000	1000	1000
Others	Min	0.2	1	1	1	1	1
	Max	250	100	100	100	100	100

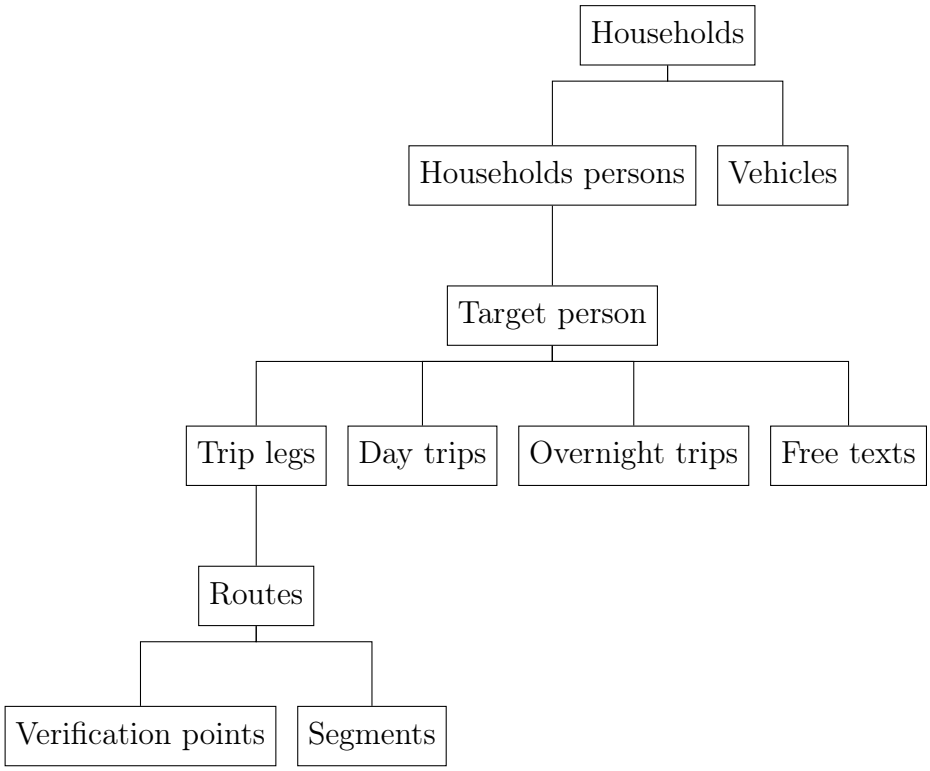
The *households* file contains information concerning an entire household, e.g., number of persons in the household, number of vehicles or geographical location. Each household is assigned a unique identification number.

Household persons (Haushaltspersonen): The information on the persons of the household (e.g., age, sex, possession of driving license) is recorded in this file.

Target persons (Zielpersonen): Each household is assigned one target person who is questioned about her daily mobility, occasional journeys and attitudes towards transport policy in Switzerland. The *target person* file contains information related to the target person, such as marital status, labour market status, ownership of public transport subscriptions, etc.

Vehicles (Fahrzeuge): The *vehicles* file contains information on all cars and motorbikes

Figure 1: Data structure of the MTMC



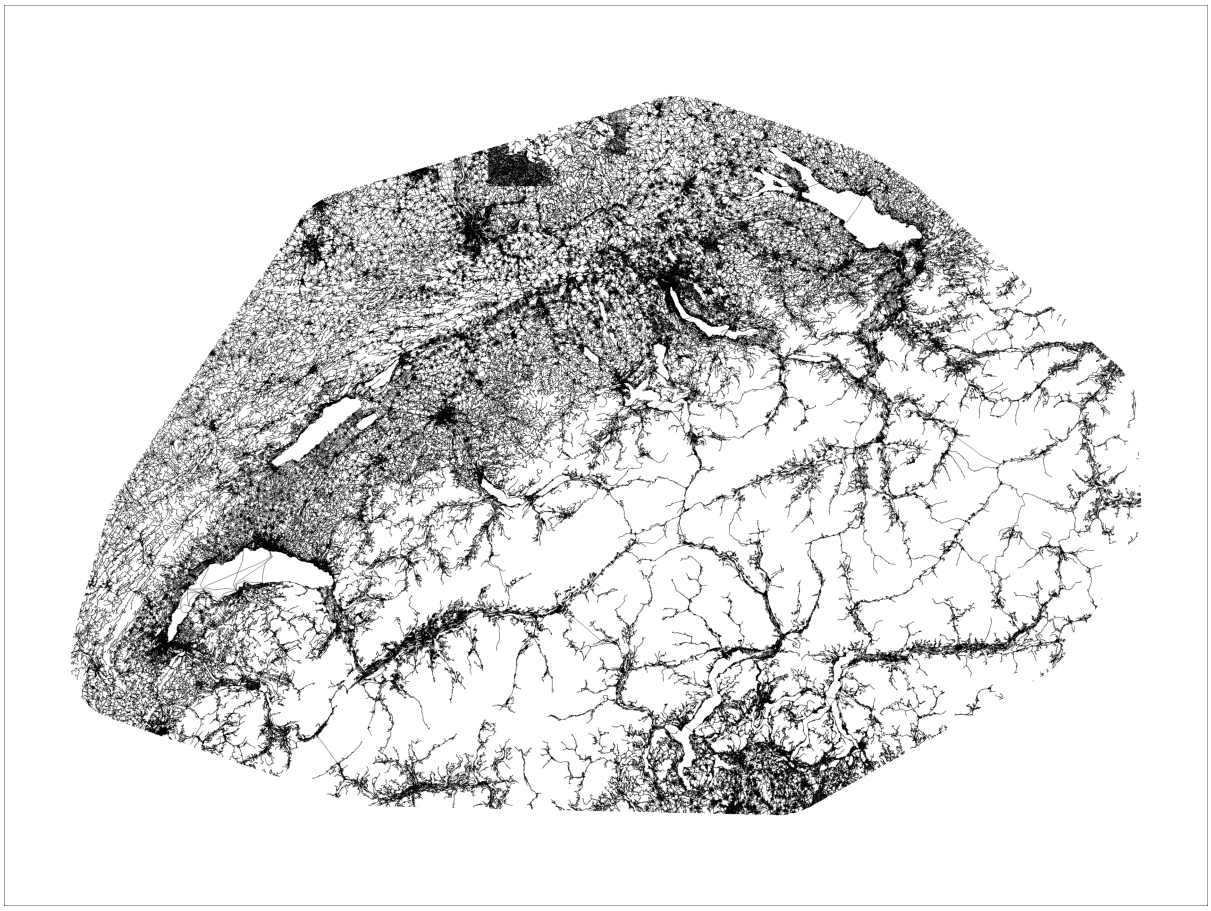
of a household, e.g., year of introduction, engine capacity, speedometer reading, etc. If a household has several cars (or motorcycles), these are numbered consecutively.

Trip legs (Etappen): The trip legs represent the smallest unit of the daily mobility. They have a minimum length of 25 meters and are covered by a single mean of transport. If the mean of transport is changed, a new trip leg begins. Changes of location within buildings do not constitute trip legs. As the route choice is recorded geographically in the MTMC 2021, the *trip leg* data record contains diverse route information for each trip leg.

Routes (Routen): Each trip leg covers a route between the starting point and destination, if a trip leg could be successfully routed in the MTMC 2021. The *route* file contains the associated geodata. Routing is done for all trips on the territories of Switzerland and Liechtenstein. In order to be able to route trips starting and ending in Switzerland, but going through the border in between, the road network has been extended to a convex hull of Switzerland, to which a band of 20 km was added (see Fig. 2).

Verification points (Verifikationspunkte): As part of the route verification of the MTMC

Figure 2: Road network of the convex hull of Switzerland, to which a band of 20 km was added.



2021, at least two verification points were requested on the phone in the case of motorized individual transport with a distance greater than 3 kilometres. In addition, for trip legs crossing the Swiss border, the file contains the geographical details of the border point.

Segments (Segmente): The *segments* file contains all segments of the routing for motorised private transport (for successfully routed trip legs on the street network).

Day trips (Tagesreisen): 30% of the randomly selected target persons are asked about their daily trips in the additional module *day trips* (number of day trips, purpose of day trips, choice of means of transport, etc.). The number of day trips per person is recorded for a reference period of 14 days, whereby detailed information is recorded for a maximum of three randomly selected day trips.

Trips with overnight stays (Reisen mit Übernachtung): The additional module *trips*

with overnight stays, which 30% of randomly selected target persons answer, contains questions on travel behaviour (purpose of the trips, choice of means of transport, etc.). The number of trips with overnight stays is recorded for a period of 4 months, whereby a maximum of three randomly selected trips per target person are recorded in detail.

Free texts (Freitexte): Interviewees and interviewers have the opportunity to provide comments as free text at the end of the survey.

3 Overview of the data-cleaning

We divide the plausibility checks into systematic and manual checks. The systematic plausibility checks are applied to all data and rely on threshold values and factors (see Section 4.4). We define limits within which we believe that data are consistent and plausible. If values are beyond the chosen limits an initial correction is applied automatically. After these corrections, we validate the data again. Observations remaining implausible are checked manually.

The systematic plausibility checks are performed in the software R. The code is publicly available on github: `github.com/AREschweiz/microcensus-geodata-cleaning`.

3.1 Systematic data-cleaning

We structure the systematic plausibility checks according to the input data from the FSO and LINK institute (see also Fig. 1). In this paper we focus on the plausibility checks of the trip legs (see Section 4). Nevertheless, we give hereafter a brief insight into the verification of the occasional journeys, i.e., the day trips and trips with overnight stays. We do however not address the plausibility checks of the other data sets (e.g., *households*, *target persons*, etc.).

3.1.1 Day trips and trips with overnight stays

The plausibility checks for the data sets *day trips* and *trips with overnight stays* are similar. The respondents have to estimate the distance of the whole trip (meaning the distance of the outward journey, the return journey and the sum of the distance of the journeys at the destination). The estimated distance is compared with a comparison distance. This comparison distance d_{comp} is calculated as: $d_{comp} = d_{asthecrowflies} * 2 * 1.1$. Where we use the distance as the crow flies between the starting point and the end point if there are no intermediate stops. If there are intermediate stops, the distance as the crow flies is equal to the sum of the distances as the crow flies between the subsequent points. The comparison distance is to be understood as a lower bound and a conservative measure of the distance made.

The values used for the lower limits are:

- for *day trips*: $0.7 * d_{comp}$
- for *trips with overnight stays*: $0.8 * d_{comp}$

For the upper limits we use:

- for *day trips*: $6 * d_{comp}$
- for *trips with overnight stays*: $4 * d_{comp}$

If the estimated total distance is outside the limits, it is replaced with the comparison distance (d_{comp}). A comparison of the raw data with the final data shows that respondents tend to estimate the distance of their journeys as too short. For the MTMC 2015 about 30% of the *day trips* and 40% of the *trips with overnight stays* have been estimated as too short (according to the limits above). Replacing the too short distances with d_{comp} leads to an increase of the journey distance by a factor of 1.9 for the *day trips* and 1.6 for the *trips with overnight stays*.

3.2 Manual data-cleaning

Manual plausibility checks and corrections are inevitable and necessary. We distinguish two main types of manual data processing:

- necessary modifications indicated in the *free texts* and
- filtered observations which are still implausible after the systematic quality control.

At the end of the survey, the target persons can leave remarks and also the interviewers can address open points or issues in the form of free texts. The *free texts* are analysed manually and based on that it is decided whether manual changes are necessary.

The R code for the automated data-cleaning stops running if observations are still not plausible after the systematic checks. In particular, we check observations with implausible distance, speed and/or duration. Also, missing coordinates are added manually. It would go too far to list all manual adjustments here[4].

4 Cleaning of trip legs

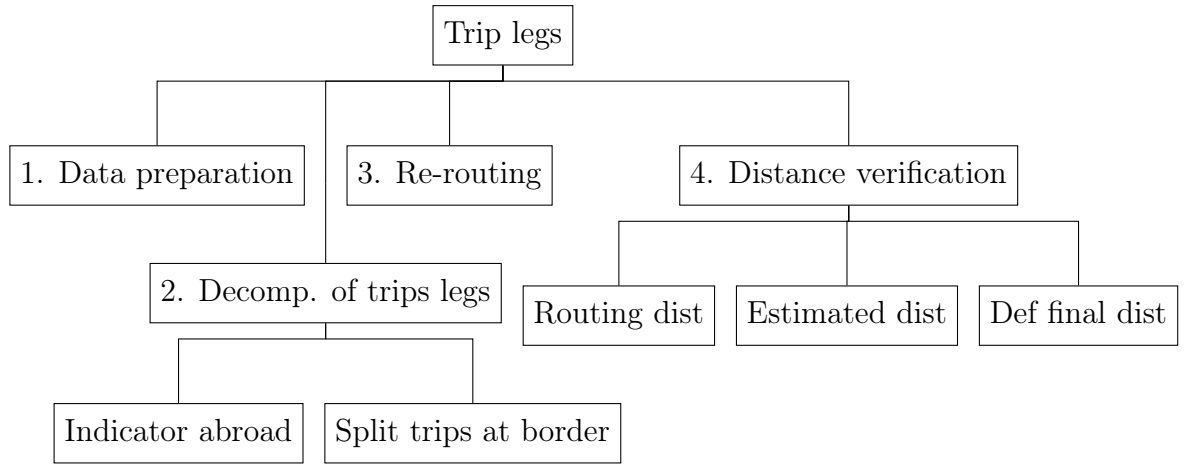
The *trip leg* data set contains all trip legs carried out by all respondents on their reference day. This is the main part of the plausibility check performed by the ARE. The plausible distance of each trip leg is the single most important output of this check. It leads in the end to the key-figure from the MTMC - the average daily distance traveled per capita - and is therefore crucial for further analyses. Hence, most of the checks focus on validating the distance of the trip legs.

As mentioned in the introduction, the plausibility checks base on what was done in previous MTMCs. Changes in the data due to changes in methodology are therefore limited. Figure 3 gives an overview of the performed checks on the trip legs.

The main challenge is to draw the line between wrong data and rare cases. Imagine the following example: A person travels from A to B via C. Travelling via C is a rather big detour. The challenge for us is to judge whether the person did indeed travel via C or if the via point was recorded incorrectly by the interviewer. We have two strategies: i) Comparing the routing and estimated distance with the distance from the Swiss National passenger transport model (NPTM) (allows us to judge whether the distance is plausible). ii) Comparing the distance with speed (allows us to judge whether it is likely that the trip has been done in the recorded time).

[4]For more details see `github.com/AREschweiz/microcensus-geodata-cleaning`

Figure 3: Plausibility checks trip legs



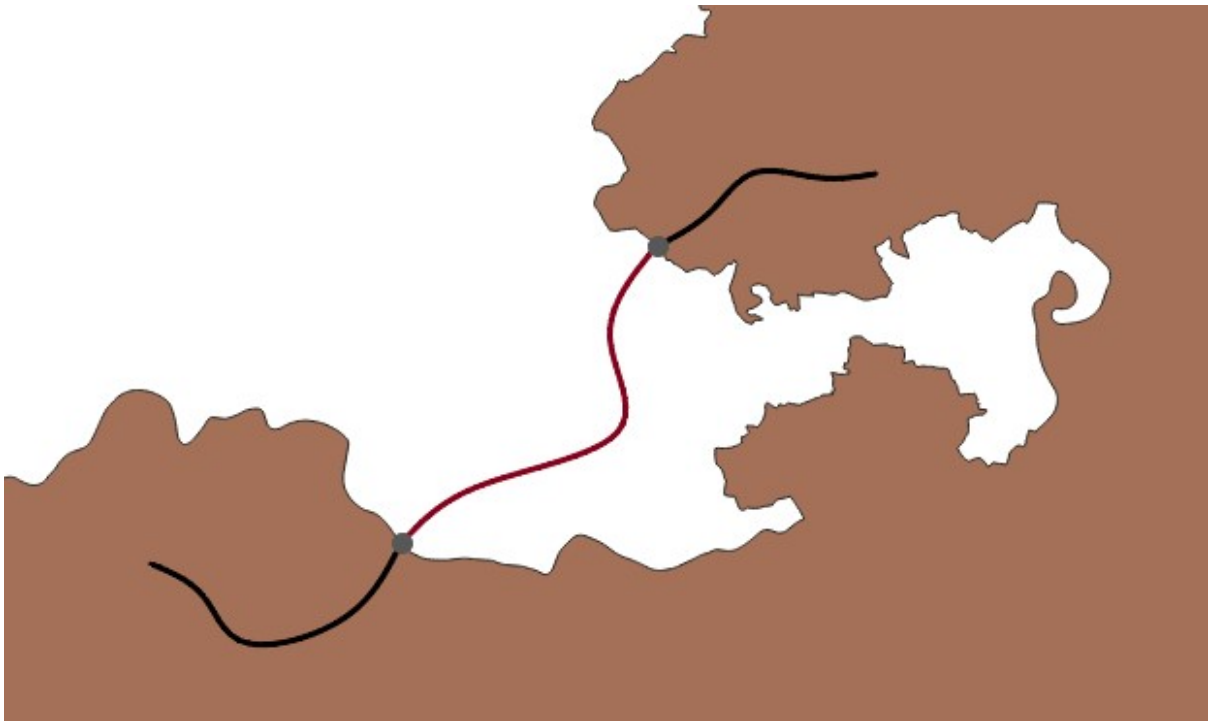
Following, we show each single step of the plausibility checks for trip legs. We show which variables are adjusted under which conditions and how.

4.1 Data preparation and computation of spatial variables

After the preparation of the data, the manual corrections of the *trip leg* data coming from the *free texts* are included. This ensures that the plausibility checks are also performed on the manual corrections. After that, we check if the coordinates of the starting and arrival points are existing and valid. Missing coordinates are added manually (based on the available information, but normally address). Missing municipality numbers (defined by the FSO) are completed based on the coordinates of origin and destination.

To ensure a harmonized and up-to-date data basis, the spatial variables on the nomenclature of territorial units for statistics (NUTS 3 regions) and the codes of the degree of urbanization (DEGURBA) are recalculated. For this purpose, the coordinates of origin and destination are superimposed on the NUTS 3 respectively DEGURBA map and assigned accordingly.

Figure 4: Decomposition of trip legs into domestic and abroad



4.2 Decomposition of trip legs into domestic and abroad

For the final analysis we have to be able to distinguish the distances made on the territories of Switzerland and Liechtenstein and abroad.[5] This is important for the final results and the key-figures e.g., distance travelled in Switzerland. Trip legs crossing the border have therefore to be split at the border crossing. If the trip takes place within the convex hull (see Fig. 2) the data contain the re-partition of kilometres, abroad and in Switzerland.

Thus, this section does not only describe a pure plausibility check, but also how we split the trips crossing the border. If a trip leg crosses the border once we receive two trip legs, if twice three trip legs and so forth. For cases, where the trip ends or starts directly at the border, one trip leg gets subtracted. Trips crossing borders other than the Swiss one are not decomposed. These border crossings are not relevant for the purposes of the MTMC. Figure 4 gives an example with two border points and three resulting trip legs. The parts of the trip leg in Switzerland are shown in black, the part abroad in red.

For trip legs crossing the border once the aim is to receive two distinct trip legs: one

[5]When we refer to domestic and foreign in the following, domestic means Switzerland and Liechtenstein. Also, with Switzerland actually Switzerland and Liechtenstein is meant is this section.

in Switzerland and the other entirely abroad - split at the border point. We distinguish between trips legs with and without routing. Trip legs with routing are within the convex hull and thus, the distance in Switzerland is contained in the data. The kilometers abroad can be easily calculated as the difference between the total distance and the distance in Switzerland.

For trips without routing, the distances for the two "sub-legs" are calculated using the ratio of the distance made in Switzerland and the whole distance. The distances of the parts made in Switzerland are calculated with help of the interzonal distance matrix of the NPTM[6]. If a trip leg is not within the NPTM distance matrix we use the distance as the crow flies to calculate the ratio.

For trips legs crossing the border more than once we cannot simply calculate the distance abroad as difference between total distance and distance in Switzerland. We instead split the routes each time they cross the border (see also Fig. 4). Said that, this only applies to trip legs with a valid routing. Trip legs without routing are treated as having one border crossing. For each original trip leg with routing we calculate the length and the number of the "sub-legs".

After that we create a data frame with the right number of copies per trip leg crossing the border several times. That means that for a trip leg crossing the border twice three sub-legs are created. Next, we start to change the relevant variables of the sub-legs (e.g., purpose of the trip, coordinates of the border points, etc.). We start with the first sub-leg and replace the information of the destination by the information about the border point. For the last sub-leg we replace the information of the origin. For the middle legs we replace the information about both, origin and destination.

With this procedure we ensure that we record the distances on Swiss territory and distinguish them from the ones abroad.

4.3 Re-Routing

Free texts can contain information about wrong or unreasonable routing. These routes are computed again. The aim of this process is to verify the distance. Thus, we replace

[6]For more information see Swiss national passenger transport model (NPTM) and Justen *et al.* (2020).

the "wrong" routing distance by the recalculated distance. For the re-routing we use the TomTom data for all street modes and the network from the NPTM for trains.

4.4 Verification of routing and estimated distances

The aim of this section is to check whether the routed and estimated distances of the trip legs are plausible. A first verification has already been performed during the interviews (see Section 2.3). For the offline plausibility check, we first calculate the distance as the crow flies for all trip legs. The distance as the crow flies provides a consistent measure and is recalculated after the integration of the manual corrections made earlier. We start to check the distances of the trip legs and identify the implausible ones. This process is explained in more detail below.

4.4.1 Verification of routing distances

Next, we check whether the routing distance is plausible and we identify the round trips[7]. Duration and distance of the trip legs are used to calculate the average speed. To verify the plausibility of the routing distance we compare the calculated speed with the lower and upper speed limits by mode given in Table 3. The routing distance is assumed to be plausible if the calculated speed lies within the limits.

If the distance as the crow flies is greater than 5 metres and the routing distance is greater than thirty times the distance as the crow flies, the code stops and lists the problematic cases. We also mark all trips legs which have an estimated distance and a routing distance which is greater than fifty times the estimated distance as implausible.

In later steps we adjust the duration of identified trip legs. In order to do that, we first calculate the available time windows before and after the trip leg, as shown in Fig. 5.

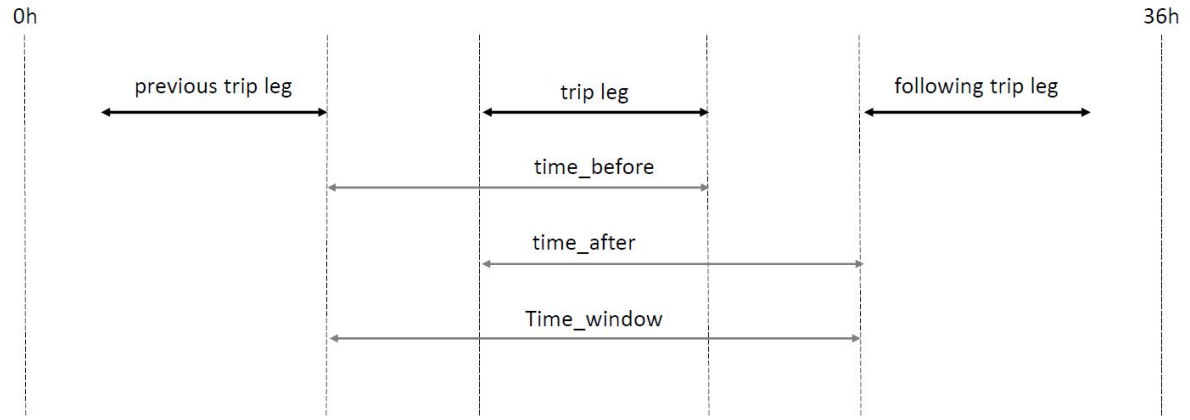
We calculate the following time windows:

[7]Round trips are trip legs with identical origin and destination.

Table 3: Speed limits per transport mode

Transport mode	Limit min (km/h)	Limit max (km/h)
Walking	0.2	30
Bicycle, slow EBike	0.2	40
Light Motorcycle, fast EBike	0.2	60
Motorcycle	0.2	200
Car, Truck, Coach	0.2	200
Bus, Tram	0.2	150
Train	0.2	400
Boat, Cable car	0.2	150
Plane	0.2	2000
Vehicle-like devices	0.2	150
Others	0.2	250

Figure 5: Time window concept



- Time_before: Interval between the time of arrival of the previous trip leg and the time of arrival of the trip leg.
- Time_after: Interval between the start of the trip leg and the start of the next leg.
- Time_window: Time window in which the trip leg necessarily takes place. This is the interval between the end of the previous trip leg and the beginning of the next trip leg.

4.4.2 Verification of estimated distances

In the next step, we compare the estimated distance of trips in Switzerland with the following two measures in order to verify whether it is plausible:

- the distance from the Swiss national passenger transport model (NPTM) and
- the speed calculated on the basis of the estimated distance.

Check estimated distances using distance as proxy
We compare the distance from the NPTM with the estimated distance for trips in Switzerland with geocoded start and end points. We first compute the interzonal distance with the NPTM for trip legs which are not round trips, are not made by plane, have an estimated distance and a known mode of transport.

We save the distance from the NPTM in the variable *comparison_distance*. If the distance is shorter than 5 km, we replace it by the distance as the crow flies multiplied by a factor depending on the transport mode. For walking, cycling, vehicle-like devices and others we do not use the NPTM distance and take the distance as the crow flies multiplied with a factor for all distances. Table 4 shows the used factors.

Table 4: Correction factors for comparison distance

Transport mode	Factor
Walking	1.2
Bicycles, Ebikes, Vehicle-like devices, Others	1.3
Light Motorcycle, Motorcycle	1.2
Car, Truck, Coach	1.3
Bus, Tram/Metro, Taxi, Uber-like	1.2
Train	1.3
Boat, Funiculaire	1.1

Source: ARE based on Chalasani *et al.* (2005)

For trip legs by train, the estimated distance is plausible when it is between 0.8 and 1.7 times the comparison distance. If the estimated distance is not plausible, the comparison distance is used instead. If a trip leg was done by another transport mode, the estimated distance is plausible when it is between 0.8 and 10 times the comparison distance. Again, if the estimated distance is not plausible, the comparison distance is used instead. Like that, we have a reasonable comparison distance for all the here described trip legs.

Check estimated distances using speed as proxy

We also check the distances by comparing the estimated distance with the speed calculated on the basis of the estimated distance. We check whether the calculated speed is plausible based on the thresholds in Table 3.

We compute the speed again and check whether now it is plausible. If not, we alter the duration and then the departure and arrival times. Thus, we replace the duration, assuming an average speed per mode whereas the duration cannot be longer than the time window between the previous and the next trip leg (see Fig. 5). The duration is then computed based on average speed and rounded to the minute.

After that we update start and end times, taking into account the new trip duration. We reorder trip legs based on departure time and we "re-recompute" speed. If still implausible, we check the identified trip legs manually.

4.4.3 Define final distance

Now we can define the final distance used for analysis. We call it r_{dist} and distinguish the following cases:

- if the routing distance is plausible, r_{dist} takes the value of the routing distance
- if the routing distance is not plausible, but the estimated distance is plausible, r_{dist} takes the value of the estimated distance multiplied by a factor (see Table 5)
- if the routing distance and the estimated distance are not plausible, r_{dist} takes the value of the distance from the NPTM
- and if the distance is not available in the NPTM, the distance as the crow flies multiplied by a factor is used (see Table 1).

We finally check, if r_{dist} has been defined for all trip legs. The remaining cases we check manually.

Table 5: Correction factors estimated distance

Routig distance (in km)	Factor individual modes	Factor public transport
<1km	1.01	0.97
between 1 and 3	0.91	0.94
between 3 and 10	0.87	0.88
between 10 and 20	0.89	0.8
more than 20	0.98	0.94

Source: ARE based on Ohnmacht *et al.* (2012)

5 Challenges

This paper presents the geodata-cleaning of the MTMC 2021. Consistency in the time series is one of the objectives of this process. In order to be able to compare the results, we cannot drastically change the data-cleaning method from one MTMC to the next. On the other hand, the MTMC survey method has been continuously updated and each improvement needs a new approach to data-cleaning. Below we present some examples of challenges that arise.

Some challenges were already present in 2015. For example, when recording a trip on a highway, the interviewer might click on the wrong side of the highway. The routing tool then computes the shortest path going through this point, generating a long route including leaving the highway, taking it again in the other direction in order to go through the verification point and then leaving it again and changing direction again (see Fig. 6). Interviewers were made aware of this problem at the beginning of the 2020/2021 MTMC in order to limit this undesirable phenomenon.

In 2020/2021 the MTMC features for the first time a routing tool allowing routing abroad, in a zone close to the Swiss border (convex hull of Switzerland plus 20 kilometer buffer). This creates new challenges. Until 2015, interviewees crossed the border in one specific point, since the question was "where did you cross the border". In 2020/2021 trip legs with several border crossings are possible. This asks for new geodata-cleaning approaches. Additionally, due to data imprecisions of roads and the border, some routes following the border generate many border points and many very small segments of routes (see e.g., Fig. 7).

Figure 6: Wrong verification point on the wrong side of the highway: If you look carefully, you see two lines

Figure 7: A route following the border. There is officially only one border point...

Cleaning the data from a smartphone app will be a future challenge. The potential of such an app for the recording of the mobility on the reference day will be tested in 2022. The aim is to test whether an app based mobility recording on the reference day can replace the CATI method for a subsample of the MTMC 2025. Similar approaches have already been applied in Switzerland (e.g., Gao *et al.* (2021)). New challenges include in particular the high number of GPS points and the possible outliers.

6 Conclusions

Geodata-cleaning: Too much or too little?
It is difficult to define the right amount of data-cleaning and it will remain a challenge to automatically identify the wrong observations and to distinguish them from rare and special cases. Although we try to keep the same methodology in the plausibility checks to allow for comparable time series, we try to incorporate new approaches, findings and developments continuously.

Geodata-cleaning is also quality control
Verification of the chosen routes, distances and times directly in the interview are important and allow for follow-up questions. In this way, errors can be efficiently avoided and, at the same time, special and rare cases can be validated by the respondent herself.
The examination of the data during the survey year allows the early detection of possible systematic errors. In this way, problems in the routing tool or in the questionnaire can be identified. Thus, the focus of this paper, the post-interview plausibility check, forms a final check to find and correct the observations that are still implausible. It will not only improve the final data, but also the way we design the next MTMC in 2025. Finally, comparing aggregated statistics of 2021 with previous MTMCs and with other data sources will be yet another step in the quality control process.

Geodata-cleaning faces challenges
New approaches such as data collection using smartphone apps are gaining importance and allow for additional insights into mobility behavior. They will also require new geodata-cleaning methods.

Geodata-cleaning must be transparent to guarantee quality
This paper provides insights into the plausibility process, the thresholds used and creates

transparency. We hope that this facilitates working with the data. To keep the quality of the survey high, feedback for further improvements is very welcome.

7 References

BFS (2017) *Mikrozensus Mobilität und Verkehr 2015: Kurzversion Fragebogen*, no. 5606052, Jun 2017.

BFS (2018) *Rapport méthodologique: plan d'échantillonnage, taux de réponse et pondération*, no. 4262242, Neuchâtel, Jan 2018.

BFS and ARE (2017) *Verkehrsverhalten der Bevölkerung*, no. 1840477, Bundesamt für Statistik (BFS), Neuchâtel, May 2017, ISBN 978-3-303-11262-5.

BFS and ARE (2021) *Auswirkungen der Covid-19-Pandemie auf das Mobilitätsverhalten*, no. 19244128, Neuchâtel, Oct 2021.

Chalasani, V. S., O. Engebretsen and K. W. Axhausen (2005) Precision of geocoded locations and network distance estimates, *Journal of Transportation and Statistics*, **8** (2) 1–15, ISSN 1094-8848.

Danalet, A. (2020) Detour factors from the mobility and transport microcensus 2010 and 2015, `https://github.com/antonindanalet/detourfactor`.

Danalet, A., A. Erath, T. Ohnmacht and N. A. Mathys (2022) Attitudes towards transportation policy in Switzerland: a new choice experiment, paper presented at the *22nd Swiss Transportation Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.

Gao, Q., J. Molloy and K. W. Axhausen (2021) Trip purpose imputation using gps trajectories with machine learning, *ISPRS International Journal of Geo-Information*, **10** (11) 775, ISSN 2220-9964.

Justen, A., A. Danalet and N. A. Mathys (2020) Das neue Schweizer Personenverkehrsmodell, in C. Laesser, T. Bieger and K. W. Axhausen (eds.) *Schweizer Jahrbuch für Verkehr 2020*, SVWG Schweizerische Verkehrs-wissenschaftliche Gesellschaft, IMP-HSG Institut für Systemisches Management und Public Governance der Universität St.Gallen, St.Gallen, Schweiz, ISBN 3-906532-32-1.

Ohnmacht, T., K. Rebmann and A. Brügger (2012) Swiss Microcensus on Mobility and Transport 2010, paper presented at the *12th Swiss Transportation Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.

Simma, A. (2003) History of the Swiss travel surveys, paper presented at the *3rd Swiss Transportation Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.