DATA ERRORS AND THEIR IMPACTS, THE CASE OF DATELINE

Kees van Goeverden, Bart van Arem, Rob van Nes
Delft University of Technology, Transport & Planning Department

1. INTRODUCTION

In the scientific publication process much attention is given to correct performance of statistical analyses. Generally, less attention is paid to the quality of the data that are used for the analyses. However, just like methods that are statistically not fully correct, errors in data will make the results less accurate. This paper deals with the impact of data errors on the outcomes of analyses. The discussion is based on the data that are collected in just one project, the DATELINE-project. This is a survey on long distance travelling by European residents, carried out in a large number of countries by a large number of organisations. Such a complex project that is directed at data collection, is liable to several kinds of data errors and a good sample for analysing their impact. Still, limitation to one project gives the findings a limited value. Similar studies on other (kinds of) datasets can enlarge the knowledge on the impact of data errors.

Data problems in surveys can result from inaccurate reporting by respondents and errors in coding/entering and handling data. In this paper we focus on the different kinds of errors in the data and we will not discuss the well-known problem of underreporting in long distance travel surveys (Kuhnimhof *et al*, 2009) that has been observed for DATELINE as well (Hautzinger *et al*, 2005). We go into the latter problem in van Goeverden *et al* (2014).

Next sections discuss the DATELINE-project, errors in the data, the way we corrected a number of the errors, and the impact of the errors on the results of analyses. Two types of analyses are discussed: a descriptive analysis describing volume indicators of long distance travelling, and a statistical analysis on associations between variables. A priori we hypothesise that just correcting data errors will have little impact on travel volume but may significantly influence the results of statistical analyses.

2. THE DATELINE PROJECT

The DATELINE-project is a survey on long distance travelling by European residents and was carried out in 2001/2002 in 16 countries: the 15 EU-countries at that time and Switzerland. It was one of the projects in the 5th

framework programme of the EU. The project is of exceptional importance, because it is the only EU-wide long distance survey ever conducted. Knowledge on long distance travelling is highly relevant because this travel segment covers a large part of mileage of persons and contributes significantly to the environmental problems related to travelling.

The survey is conducted on household level in about half of the countries and on person level in the other countries. The household survey includes data on personal characteristics and journeys of all members of the addressed households, the person survey includes only data on the addressed persons and their journeys, and some characteristics of their households. Data are collected in two phases. In the first phase data on households, persons, and regular journeys are registered. The second phase includes data on trips, excursions, and commuting journeys.

The different registered kinds of movements demand for an explanation.

- Regular journeys are round-trips from home or another address in the home region to a destination and back to home or another address in the home region. Regular journeys exclude commuting.
- Trips are parts of a regular journey travelled within one day. A trip in DATELINE can exist of several trips as defined in most regular travel surveys (movement between two activity places). Most journeys in DATELINE exist of two trips, an onward trip and a return trip, but the number of trips in a journey can be significantly larger.
- Excursions are day round-trips made from one of the trip destinations (these include the journey destination).
- Commuting journeys are the onward and return trips to the workplace or school/university.

A journey made by several persons travelling together is in DATELINE sometimes defined as one journey (regular journeys in the household survey) and sometimes as a number of journeys that equals the number of travellers (person survey and commuting journeys). In this paper, we adopt these definitions, but when we analyse the impact on travel volume (Section 4.1), we calculate the volumes always on person level.

In DATELINE only long distance movements are relevant. Long distance is defined as 100 km or more as the crow flies.

The survey is retrospective. The reporting period for the journeys is 1 year for holidays, 3 months for business and other private journeys, and 4 weeks for commuting. The reporting period for trips and excursions equals the period for the corresponding journeys.

The number of households in the whole survey, including the person survey, is 56,000, the number of persons is 87,000, the number of regular journeys in the survey period is 71,000. The database includes in addition 27,000 journeys outside the survey period, but these are generally not used for analyses and almost always lack relevant information like the used mode(s). The number of commuters that report about their commuting trips is 500, and the number of reported excursions is 4,000.

The data are stored in 8 related thematic databases. One database includes information on the survey process, the second includes household characteristics, the third characteristics of the reporting person or of all persons in the reporting household if the survey is on household level, the fourth characteristics of the regular journeys, the fifth information on which persons of a household participated in a journey (only for the survey on household level), the sixth characteristics of the trips, the seventh characteristics of the excursions, and the eighth characteristics of commuting journeys.

An additional (ninth) database that is used for DATELINE is a geodatabase that includes information on topographical locations and larger administrative areas all over the world. The information includes the name(s), coordinates, and some other characteristics. In the case of an area, e.g. a country, the coordinates are close to the geographical centre. Each item, settlement or area, has a unique geocode. These geocodes are used for defining the locations in the DATELINE-databases, like the home city in the household database and the origin, destination and return locations in the journey database.

We found that the data are well structured and convenient to explore.

3. DEFICIENCIES OF THE DATA

Employing the data, we encountered a growing number of problems that result from errors in the data. In this section we will give an overview of these problems for three stages in data processing: collecting/reporting, coding/entering the collected data, and data management/manipulation. We will describe the characteristics of the problems and the magnitude. We will also indicate how we corrected a number of the errors. Because we found that in all three stages errors affect the selection of locations, we add an overview of the aggregate magnitude of these errors.

3.1 Data collecting

Encountered problems regarding data collection include two cases of inaccurate reporting by respondents. Both refer to the locations in the journey database. The first is the observation, that the distance from the home city to the most distant trip destination is sometimes significantly longer than the distance to the reported journey destination (1-2% of the journeys). In most of these cases, the destination of the first trip was reported as the journey destination. There are even examples of respondents who reported the airport where they boarded a flight to a far-away country as the journey destination.

The second source for inaccurate reporting regards the origin and return locations of the journeys. Respondents that made a journey by airplane or a long distance ferry sometimes reported the name of the airport/airport city or seaport at the home side of the journey as the origin and/or return location. They did so for 4% of the origin locations and 2% of the return locations. Presumably, the actual origin and return locations were the home cities.

3.2 Coding and entering data

When entering data of large samples like those of DATELINE, certainly a number of mistakes will be made. Mistakes like typing errors are difficult to detect and we can only state that employment of the data gave us not the feeling that this kind of errors was frequently made. It is our impression that the data were entered quite accurately.

However, there are two other kinds of errors that are worthwhile to discuss. These concern the identification of topographical locations and the calculation of distances.

Identification of locations

The identification of locations is usually a problem in travel surveys and was so in DATELINE as well. In DATELINE many locations were identified: the home city of the respondent, the origin, destination, and return locations of the regular journeys, and the destination locations of the trips, excursions and commuting journeys. The identification was implemented by adding a geocode to each location, corresponding to a location in the geodatabase.

To start, we have to explain how we know (or better: assume) that frequently the wrong locations were selected. We have no definite information about where the respondents live or where they travel to. The only information is the name as it is spelled in the databases and characteristics of journeys, trips, etc. These characteristics *exclude* reported distances, except for the commuting database. Reported distances would have been helpful in identifying locations. Our assumption regarding wrongly selected locations is based on a number of improbable cases. The journey database has

numerous examples of destinations that are unimportant settlements but have the same name as a large city or touristic hotspot and are highly attractive for DATELINE respondents. An example is the small village of Berlin in the German state Schleswig-Holstein; this is in DATELINE by far the most attractive settlement in its nuts3-region. Other improbable cases are journeys by land mode (car, train, bus) to overseas continents. Though this can be due to an error in the coding of the mode(s), usually it is a matter of incorrect identification of the destination location. In many of these cases, a European city with the same name is a much more probable destination than the selected location in, for instance, the USA. Impossible cases are locations of home cities that are not located in the nuts1-area that is surveyed. The DATELINE survey is organized by nuts1-region and the household database reports this region. It is peculiar that in DATELINE location identification errors include home cities of the respondents. Apparently, those who addressed the respondents did not (always) share their information about the home locations with those who entered the geocodes in the databases.

In discussing the magnitude of the identification errors in DATELINE, we have to compare incorrect locations with correct locations. However, as stated before, we have no definite information about which locations are correct. Figures on the magnitude are based on comparison of the identified locations in DATELINE with locations that we *assume* are the correct ones. For that reason, we identified a large number of the reported locations alternatively. We assumed that, in the case several locations have the same name but just one is highly probable, the latter is always correct. So we identified the destination “Berlin” always as the German capital and “New York” always as the largest city of the USA. If there are several locations with the same name that are rather probable to be the correct one (e.g. “Neustadt” in Germany), we tried to identify the correct destination from other journey characteristics like purpose of the journey, duration, and location of the origin. If several probable locations were left, we kept the identification in DATELINE and assumed that this is correct.

Our identification was limited to locations that have a unique name, locations with the name of a European city with at least 100,000 inhabitants, locations that are at least five times reported as journey destinations, home cities that are located in the wrong nuts1-region, and destination locations of journeys where tests indicate that they are likely to be wrong (like journeys by car to America). These include 97% of both the home locations of the households and the origin/return locations of the journeys, and 99% of the journey destinations. For the trip and excursion destinations the percentages are lower (95% and 82%), in the small commuting database we identified the locations of all destinations (100%).

We assume, and have the strong feeling, that our identification of locations is much more accurate than the original identification. Assessing the magnitude of errors by comparing the original ones with our guesses might give a good approximation of the real magnitude. When discussing the magnitude, we limit the comparison to the locations that we identified and focus on the home cities and the destination locations in the journey database.

There are four reasons why the identification in DATELINE is sometimes not correct: 1) there are different locations with the reported name, 2) the reported name is not included in the geodatabase for the correct location, 3) an error is made when typing the geocode, and 4) there is an error in the coordinates of the geodatabase. We will discuss these four causes for errors.

Several locations have the same name

Locations that have no unique name include 31% of both the home locations and the journey destinations. Assuming that our identification always is correct, 7% of the home locations and 13% of the destination locations with a not unique name were wrongly identified in DATELINE. In most cases of wrong identification, the country is correct and sometimes we had the impression that the person who entered data in DATELINE had information about the smaller administrative region. However, within the country or region the selection was more or less randomly. A ‘random’ selection frequently implies that the first record in the geodatabase including the reported name in the reported country or region was selected. The locations in the database are mainly ordered from north to south and no relation between probability of being the correct location and place in the database may be assumed. Then, given the number of different locations with the same name in the same country, in 40% of the cases the name that is first mentioned in the geodatabase is likely to be the correct one. However, in DATELINE the proportions are 54% for the home cities and 59% for the journey destinations.

There is a positive relation between the quality of the identification and importance of or familiarity with the city. Table 1 displays the proportion of assumed incorrect location choices for European cities with different sizes and destinations outside Europe. The figures are limited to locations that have not a unique name.

Table 1 Assumed incorrect identification of locations

	Home city	Destination city
Locations in Europe:		
>500,000 inhabitants	1.8%	4.6%
100,000-500,000 inhabitants	6.8%	8.0%
<100,000 inhabitants	11.8%	18.3%
Locations outside Europe		33.3%

The reported name is not included in the geodatabase for the correct location

A reported name is sometimes not included in the geodatabase if the spelling in DATELINE differs from that in the geodatabase, or simply if a location is missing in the latter. Usually, the names of non-administrative areas like lakes or isles are not included but still reported as journey destinations. In DATELINE, this problem was solved in different ways. The selected locations are sometimes locations with another name close to the reported city, sometimes larger cities in the same region but often on a longer distance (e.g. 30 km), sometimes locations with similar names but usually located in a quite different region, and sometimes the larger area where the reported location is located (in most cases the country).

It sometimes happens that the reported name is included in the geodatabase, but that the location that is likely to be the correct one is missing. In that case frequently one of the locations in the geodatabase with the reported name was selected. An example are journeys with the reported destination “Wales”. We assume, that always the country Wales in the United Kingdom is the correct location. However, this country is missing in the geodatabase. Still, the database includes one location with the name “Wales” in the UK, a small settlement in England. The journey database includes 20 journeys to “Wales”. For 16 journeys the English settlement was selected, for 2 journeys the geocode of the UK, and for the remaining 2 journeys a city in the country of Wales.

If the geodatabase does not include the reported name in the reported country, but does include the name in one or more other countries, there is a good opportunity that one of the locations with this name in another country was selected. This is one of the causes for selecting locations in the wrong continent.

Typing errors in the geocode

This type of error is rare, but has important consequences. If the typed code is included in the geodatabase, the corresponding location is selected. This may be situated in a quite different area of the world. If the typed code is not included in the geodatabase, the ‘centre’ of the earth is selected; that is the crossing of the equator and the prime meridian, somewhere west of Africa in the Atlantic Ocean. This is a second source for selecting locations in the wrong continent.

Errors in the coordinates of the geodatabase

In a few cases, the coordinates in the geodatabase include an error. If in DATELINE the correct city was selected, the geographical location is still wrong. An example is the English city of Penrith. The latitude in the database

is 53.65, but it should have been 54.65. The distance between both locations is 110 km. An error that is valid for several cities in France is the absence of the minus sign for locations west of the prime meridian. The latter has no consequences for the DATELINE locations, because we found that in such cases cities with another name close to the correct geographical location were selected. We have the impression that in France first another geodatabase was used for the selection of a location and after that the nearest location in the DATELINE geodatabase was looked up.

Calculation of distances

The DATELINE respondents were not asked for trip and journey distances, except for the distance to their workplace or school/university. Crow fly distances were added to the journey, trip and excursion databases, calculated from the geocodes of the origin and destination locations. However, we observed that the calculated distances are not accurate; presumably there was an error in the formula that was used for the calculation. The deviations from the correct calculated distances are not large and never exceed 15%. In a large majority of the cases, 96% of the journeys in the journey database, the distances were underestimated. The underestimation increases when the direction of the journey turns from east-west to north-south. The average underestimation is 6-7%.

3.3 Data management

After entering the data, many data were manipulated. Mistakes in the manipulation process are the source of some other, sometimes strange, errors. In this section we discuss some errors that clearly result from incorrect data manipulation, and errors which origin we do not know but that *could* have to do with mistakes in data manipulation.

Transferring data to new databases

Some errors arose when data were transferred to other databases. One eye-catching problem when looking at the data regards the naming of locations in the Swiss journey database. The names include only the first two letters of the full name. We received a few older versions of the databases and found the explanation. In one older database, the location names were filled in the fields for the country codes (origin, destination or return country). These fields have just two positions and can include no more than two letters. Presumably, this happened when transferring data to this database; the data were copied to the wrong fields. Later this was repaired by again transferring the data to a new database and putting the data in the right fields, but the letters behind the first two letters of the names were definitely lost.

A second, serious problem regards the data for the French and the Walloon. Most countries of destinations outside Western-Europe got the wrong country

code. Concurrently, the fields of the destination names became empty. We found that in one of the older databases the country codes were correct. Apparently, the error was introduced when transferring data to another database. However, also in the older database, the destination names are missing. The wrong country codes in the DATELINE databases are alphabetically close to the correct codes. For instance, journeys to Argentina (code 'AR') got the code for Austria ('AU'). This error creates the strange phenomenon that many French would travel to unattractive areas like the uninhabited Clipperton Island, while no French journeys would be made to the French overseas departments like Reunion. The error relates to 2.5% of the journeys of the French and the Walloon. This is not so much, but the impact on the assumed journey distances is large. Most of the assumed destinations are located in a wrong part of the world. This error is a third cause for selecting locations in the wrong continent.

We observed some errors that possibly are caused by mistakes in data transfers in the Spanish and Portuguese surveys. For both countries, the geocodes of the home locations are wrong for 25-30% of the households. In the Spanish database the correct city is always replaced by the same wrong city (e.g. Madrid is always replaced by Villamalea, Barcelona always by Cripan). It seems a matter of movement from many to few: several other Spanish cities are (always) replaced by Villamalea or Cripan. In the Portuguese data such regularities are not found. These errors exist only in the household databases. The origin and return cities in the journey database, usually the home cities, do not contain this error.

In the Spanish trip and excursion databases sometimes foreign cities are replaced by Spanish cities. For instance, Paris is sometimes replaced by Paterna de Rivera, Amsterdam by San Rafael del Rio. The country of the city is related to the Spanish area where the replacement city is located. French cities are always replaced by cities that are located in Andalucia (and mainly in Cadiz), cities in more northern countries by cities located in Comunidad Valenciana (mainly Castellon).

Deletion of short distance journeys

Based on the calculated distances, journeys shorter than 100 km were removed from the journey database. However, due to both the incorrect calculation of distances and mistakes in the selection of locations, two kinds of errors had been made. First, some journeys with an actual distance <100 km were retained, second, some journeys with an actual distance >= 100 km were removed. Due to the general underestimation of distances, the proportion of journeys that was erroneously retained because of the incorrect distance calculation is marginal: 0.04%. The number of retained short

distance journeys due to wrong selection of locations is larger but still small: 0.35%.

The number of erroneously removed journeys –journeys which assumed distance was shorter than 100 km but which actual distance was longer– is likely to be larger. Assessing this number is difficult, because most data about these journeys are lost. Two sources are available. First, the household database includes information on the total number of deleted journeys for each of the three main purposes (holiday, business, and other private). The total number of deleted journeys is 4956, 7% of all journeys. Second, some information about 2380 of the deleted journeys (48%) has been maintained. The information is limited to household id, names of origin and destination, and (wrongly) calculated distance. We added geocodes to the origin and destination names and recalculated the distances. We found that 27% of these journeys were wrongly deleted due to the inaccurate distance calculation, and another 8% due to incorrect identification of one of the locations. Assuming that these percentages are valid for the remaining 52% of the deleted journeys as well, 1.85% of the original journeys in the journey database was erroneously deleted because of incorrect distance calculation, and another 0.58% was erroneously deleted because of wrong selection of locations. However, we do not know whether the 2380 journeys are representative for all deleted journeys.

Because of the lack of sufficient information about the deleted journeys, we decided to correct for erroneously deleted journeys in a way that does not use the available information on deleted journeys. We developed expansion factors that we added to the journeys in the journey database that would have been deleted if the presented (wrongly calculated) distance would have been the actual distance and the same error would have been made in calculating the distance. Take as an example a journey with a calculated distance of 105 km that is underestimated by 8%. Now we define three distances: D1 is the actual, correctly calculated distance (114.1 km in the example), D2 the presented wrongly calculated distance (105 km), and D3 the distance that would have been calculated if D2 was the actual distance (96.6 km). The expansion factor should expand for the journeys which actual distance is D2 (>= 100 km) but that were removed because the calculated distance (D3) was shorter than 100 km. We developed next expansion factor:

EF = GF**(D1-D2) + 1

Where:

EF: expansion factor

GF: growth factor of the journey frequency when the distance becomes shorter
D1: the actual journey distance (in the example 114.1)
D2: the wrongly calculated distance (in the example 105).

The expansion is only applied to journeys where D2 >= 100 km and D3 < 100 km. In the analyses that use distance information, we use for the expanded part of the journey distance D2 and for the not expanded part distance D1.

Analysing the journey frequency in the distance range 115-215 km, we found that the frequency is highly stable for holiday journeys, while the frequency of journeys for both business and other private purposes increases by about 1% when the distance decreases by 1 km. Therefore, we made GF equal to 1 for holiday journeys, for other journeys we assumed the value 1.01. Figure 2 shows the resulting journey numbers. One curve includes the journeys in the journey database, the second the journey frequency if the deleted journeys >=100 km with distance information are added to them ("short journeys (1)"), the third the journey frequency if additionally deleted journeys >= 100 km without distance information are added ("short journeys (2)"), assuming that their frequency distribution is equal to that of the short journeys with distance information, and the fourth the frequency of expanded journeys as described before.

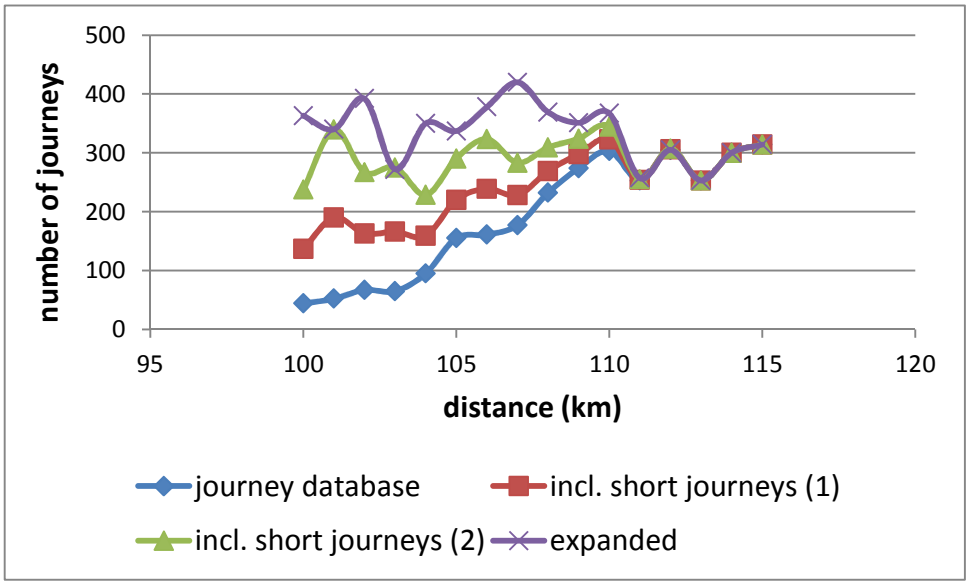

Figure 2: Journey frequencies between 100 and 115 km

The number of journeys in the journey database declines strongly when the distance decreases from 115 to 100 km. This is a clear indication that something is wrong; an increase at decreasing distance would be expected. If the deleted long distance journeys with distance information are added, there is still a (smaller) decline. Adding the other deleted long distance journeys, the

figure becomes rather stable, but here is certainly no increase. The frequency of expended journeys shows a small increase, and that is what might be expected. The expanded numbers are somewhat higher than the sum of journeys in the journey database and all erroneously deleted journeys. This means that the expansion does not only corrects for the erroneously deleted journeys, but 'finds' even more journeys that were not included in the original journey database. We hypothesize that this is a correction for the fact that respondents sometimes do not report journeys if the distance is slightly longer than the minimum distance because in their perception the distance is too short. Frei *et al* (2010) report a similar effect in the KITE-survey.

Absence of intercontinental journeys of the Finnish

Intercontinental journeys of the Finnish are missing in the journey database. We do not know the cause or reason for this. Certainly, it will affect the registered mileage of the Finnish significantly. We tried to repair this deficiency by adding expansion factors to the intercontinental journeys of the Swedish. We choose these factors so that the ratio between the number of visits of the Swedish and that of the Finnish to a continent equals the ratio of visit numbers that are published in the continental reports of the World Tourism Organization (2005 and 2006).

Commuting journeys

The commuting data include several serious errors. We discuss them together in this subsection.

Who compares the outcomes regarding long distance commuting in the different countries as published in Deliverable 7 of the DATELINE-project, will observe that two groups of countries exist: countries with relatively many long distance commuter journeys per person per year (0.2-1.0) and countries with relatively few commuter journeys pppy (0.02-0.06). The explanation is a different manipulation of the data. Respondents were asked to report their long distance commuting journeys in the past 4 weeks. Therefore, in the commuting database the journeys of respondents from a number of countries were expanded with a factor 12 to create numbers for a whole year. However, for other countries such an expansion was not performed, and the 4 weeks-numbers were presented as yearly numbers in the deliverable.

Additionally, there are two other major problems with the data that are valid for nearly all countries. The first is, that it was not noticed that a lot of the journeys in the commuting database had a crow fly distance shorter than 100 km (35%). The commuting database is the only database including origin and destination locations where no crow fly distances were calculated.

The second problem is, that a large part of the long distance commuters is missing in the commuting database. This conclusion can be drawn from the person database that includes reported distances to workplace or school/university of the respondents. Assuming that reported distances longer than 128 km corresponds to a crow fly distance longer than 100 km, 48% of the long distance commuters are not included in the commuting database.

The two latter problems partly balance out, but the aggregate result is an underestimation for Europe as a whole. The results per country vary widely. If we compare our estimations of long distance commuter journeys –these exclude journeys <100 km in the commuter database, include estimated journey numbers for long distance commuters that are missing in the commuter database, and are expanded to annual totals using an expansion factor 365/28– with the published figures, the latter vary from being 1.7 times higher (France) to 34 times lower (UK).

3.4 Aggregate effect of incorrect selection of locations

In all stages of data processing, some errors affected the selection of locations. This section discusses the aggregate effect. In assessing the proportions of incorrect selections, we define the correct selection of inaccurately reported locations by respondents as correct (Section 3.1). There is one exception. If the most distant trip destination of a journey is significantly more distant from the home city than the reported journey destination, we indicate them as incorrect. We define ‘significantly more distant’ as the case that the distance to the most distant trip destination is either more than two times as long or longer than 400 km than the distance to the reported journey destination.

We estimate that next percentages of locations are not correct:

- 3% of the home locations in the household database if the Portuguese and Spanish sections where the locations were wrongly replaced are left out. Including these sections, 11% is not correct.
- 3% of the origin and return locations, and 8% of the destination locations in the journey database; for 11% of the journeys one of the locations is not correct.
- 7% of the destinations in the trip database. These include the return locations of the journeys (usually the home location).
- 14% of the origin locations and 20% of the destination locations in the excursion database; for 30% of the excursions one of the locations is not correct.
- 3% of the destinations of journeys with completed destination information in the commuting database. These exclude journeys where the geocode for the destination location is missing (30% of the journeys; in most cases the destination name is provided) and journeys

where the geocode and name of the home city erroneously are filled in the corresponding fields for the destination city (3% of the journeys, including all journeys of the Danish; any information on the actual destination is missing).

Generally, the consequences of incorrect selection of locations will be more serious the larger is the topographical mismatch. We define the latter as the distance between the selected location in DATELINE and the location that we assume to be the correct one. Figure 1 shows the distribution of these distances for both the origin and destination locations in the journey database. The figure is limited to the journeys where the two assumed destinations differ and the distance between both is more than 10 km (3% of all origins and 8% of all destinations).

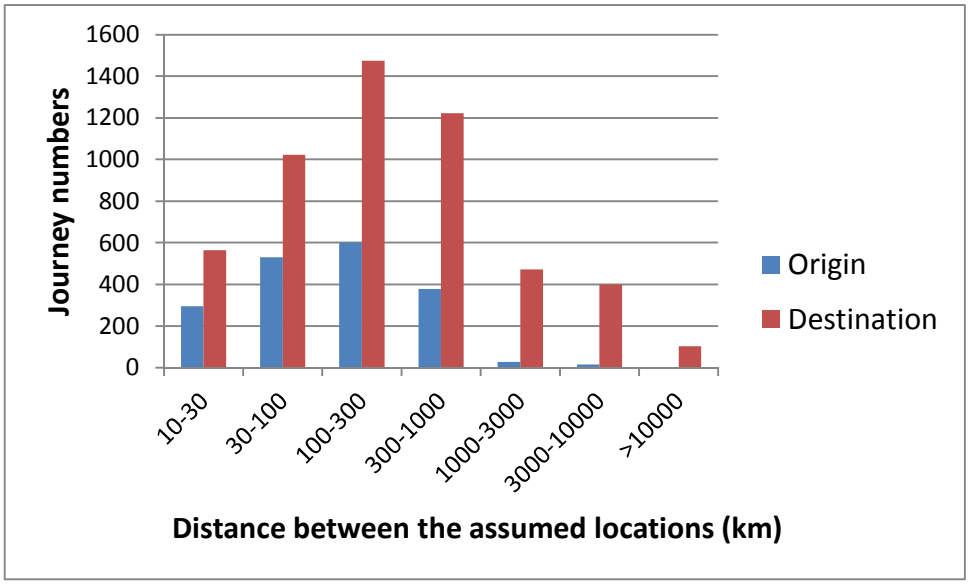

Figure 1: Mismatch of journey origin and destination locations

4. CONSEQUENCES OF THE ERRORS

A pressing question is: do the errors in the data affect the outcomes of analyses significantly? We examined the impacts on a) registered travel volume and b) results of a modal choice analysis. We estimated the impacts by comparing the results between using the original DATELINE data and the data that include our corrections. We did this on an aggregate level, using observations of all 16 DATELINE countries.

4.1 Impacts on travel volume

Table 2 shows the impacts of the corrections on both journey numbers and mileage in two databases: the journey database and the (smaller) excursion database. Results for trips are rather comparable to those for regular journeys because trips and journeys are closely related. Presenting European figures

for the large errors regarding commuting is not so useful, because these errors vary widely for the different countries.

Table 2 Impacts of some corrections, using index numbers

	Journeys >=100 km		Excursions >= 100 km	
	Numbers	Mileage	Numbers	Mileage
No correction	100	100	100	100
Correcting locations	100	99	78	34
Correcting distance calculation	104	106	89	39
Adding intercontinental journeys of the Finnish	104	107	-	-

The balanced impact of **correcting locations** on journey numbers is negligible. The numbers of journeys that were wrongly assumed to be either longer or shorter than 100 km are of the same magnitude. The mileage decreases a little, indicating that incorrect identification of locations more frequently produces a distance that is too long than one that is too short.

The impacts of correcting locations on long distance excursions is significant. The number of excursions decreases by more than 20%, the mileage reduces even to one third. For excursions two problems play a role. First, a relatively large part of the destination locations was not correct selected. Excursion destinations are frequently no well-known places or natural sites like lakes and mountains that are not included in the geodatabase. Second, incorrect identification of the origin or destination location of an excursion will generally lead to a distance that is (far) too long. The distances of long distance excursions are usually relatively short, in the range of 100-300 km. A location error that makes the distance unpredictable has a high probability to produce a significant longer distance than the actual distance.

The **correction of distance calculation** has a clear but not large positive balanced impact on journey numbers. The removal of journeys that are erroneously assumed to have a distance shorter than 100 km explains a small part of the observed underregistration by DATELINE (Hautzinger *et al*, 2005). The impact of mileage is somewhat larger, and the impact on the number of excursions is significant. Many excursions have distances close to 100 km and can convert from just below to just above 100 km.

Finally, **adding intercontinental journeys for the Finnish** has a minor effect on the aggregate European level. For the Finnish the impact on mileage is large: nearly 40%.

4.2 Impacts on results of a modal choice analysis

We analysed the choice for the train mode in long distance journeys using the data of both the original and corrected journey databases. We used a binary logit model and selected 13 exogenous variables on household and journey level. Selecting variables on person level is difficult, because in long distance journeys frequently persons with different characteristics travel together. Some variables have identical values in the two databases (like car ownership), others can have different values which can lead to different results. The latter are the spatial variables home country, destination country, size of the home city, size of the destination city, domestic/international journey, and distance. We split up the distance in classes, assuming that the influence of distance is not linear. The analysis is limited to journeys within Europe that are no longer than 1500 km crow fly.

When trying to perform this analysis, we directly encountered an additional deficiency in the DATELINE data. One of the selected exogenous variables is the number of participants in the journey. However, this variable is missing in the person survey. The modal choice analysis proved that this is a serious shortcoming. Analysing the data of the household survey demonstrates that this variable is the most important explanatory variable for train choice. We performed for this reason two analyses: one for all countries excluding the most influential variable, and one for the countries in the household survey including all selected variables.

We performed the two analyses each two times, one using the data of the original database, and one using the data of the corrected database. The outcomes are not encouraging for data correction. The results that are based on the corrected data differ just marginally from those that are based on the original data. The R^2 improved hardly (Table 3), in all four estimations all variables were significant, and the ranking of the variables is nearly equal and differs only for some less important variables.

Table 3 R^2 values

	All countries, 12 variables		Household survey, 13 variables	
	Cox and Snell R^2	Nagelkerke R^2	Cox and Snell R^2	Nagelkerke R^2
Original journey data	0.107	0.206	0.115	0.212
Improved journey data	0.108	0.207	0.117	0.213

Explanations for the absence of significant differences are:

- Not considering the recalculated distances that affect this variable somewhat in nearly all observations, the proportion of observations

where one or more variables differ is small, only 8% in the selected data (journeys < 1500 km within Europe). And in most cases just one variable differs: distance.

- The spatial variables that sometimes have different values in the two databases have only a moderate influence on train choice. The two most important spatial variables, size of the destination city and domestic/international journey, rank 3 and 4 (behind the number of participants in the journey and car ownership). The influence of distance, the most frequently adapted variable, is relatively small; it ranks 7.
- The observations that presumably have the most negative impact on the performance of the estimations are left out of the analysis. These include journeys to destinations that are erroneously assumed to be located in another continent.
- The analysis excludes the impact of deletion of journeys that are somewhat larger than 100 km. This impact cannot be assessed because essential information about these journeys is missing.
- Variables indicating the level of service and price of the alternative modes were not included in the analysis. We assume that train choice is strongly related to the competitiveness of the train variables that indicate this would differ for journeys where the identification of one or more locations is different. Inclusion of such variables to the two databases is might increase the differences in the fit of the analyses.

Looking in more detail at differences between categories of a variable –the categories are compared with a selected reference category–, for two spatial variables notable different influences are observed. In the analysis that includes all countries, the significance of the coefficients of the home countries that differ significantly from the reference country (6 and 7 countries in the original and corrected databases; the reference country is Austria) is nearly always higher when using the corrected data. More important is the different influence of the distance variable. The corrected data produce a stronger influence in the range 700-1100 km and a higher upper limit of distance that differs significantly from the reference class (900 versus 1100 km). Table 4 displays the estimated coefficients and p-values for the distance classes.

Table 4 Influence of the distance

Distance class (km), 100-200 km is the reference	All countries, 12 variables				Household survey, 13 variables			
	Original data		Corrected data		Original data		Corrected data	
	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.
200-300	0.357	0	0.358	0	0.264	0	0.271	0
300-400	0.511	0	0.475	0	0.35	0	0.335	0
400-500	0.625	0	0.668	0	0.248	0.001	0.322	0
500-600	0.801	0	0.633	0	0.564	0	0.235	0.012
600-700	0.934	0	0.979	0	0.611	0	0.637	0
700-800	0.948	0	1.027	0	0.433	0.003	0.619	0
800-900	0.841	0	1.065	0	0.693	0	0.872	0
900-1000	0.26	0.077	0.65	0	0.161	0.412	0.348	0.042
1000-1100	0.275	0.114	0.574	0	0.022	0.928	0.46	0.017
1100-1200	0.155	0.448	0.123	0.54	0.277	0.287	-0.138	0.612
1200-1300	-0.461	0.084	-0.026	0.906	-0.256	0.472	0.212	0.427
1300-1400	-0.828	0.01	-0.577	0.06	-0.895	0.062	-0.655	0.168
1400-1500	-1.092	0.01	-0.355	0.215	-1.033	0.093	-1.156	0.03

We have two explanations. First, the distance class limits differ in the two databases because in the original database the calculated distances are generally too short. Therefore, in the original database each distance class includes a number of journeys that are classified in the next higher class in the corrected database. This is also valid for the reference class. Second, due to wrong identification of locations, the selection of journeys <1500 km within Europe includes in the original database journeys that actually have a much longer distance where the probability of plane use is nearly 100%. Inclusion of these journeys lowers the train share and may consequently lower the upper limit of the distance range where the train is considered to be competitive. A good assessment of this upper limit is relevant for policies that promote train use and focus on distances where the train is a feasible mode.

5. CONCLUSION

The hypothesis in the introduction, that correcting data errors will have little impact on travel volume but will significantly influence the results of statistical analyses, is not validated by the research in this paper. There are examples of large impacts on travel volumes (in particular commuting trips and excursions), while the impacts on the results of the modal choice analysis are mainly marginal. Still, one can argue that the impacts on the number and mileage of regular journeys –these comprise most of the long distance travel volume– is rather small, while a statistical analysis on not corrected data *can* produce results that are clearly wrong.

A number of practical lessons can be drawn from the deficiencies in the DATELINE data:

- In a large data-project like DATELINE it is advisable to charge one partner with the task to control the quality of all produced data.
- Be careful when data are transferred to another database. Always check whether the data are correct included in the new database.
- If observations are removed from a database because they are assumed not relevant, keep the data. It might appear later, that the indication for relevance was not correct. Apart from that, such observations could be useful for non-general analyses.
- People that enter location id's in databases including home locations of the respondents should be informed about the addresses of the respondents.
- One can consider to ask the respondents about distances of journeys and trips. It increases the burden for the respondent with the danger of less accurate reporting or larger non-response, but reported distances are very helpful in the identification of destination locations.
- Register in long distance surveys always the number of participants in the journeys. This is a key variable in modal choice analyses.

The findings in this paper are based on the analysis of data errors of just one project, though this project includes several data collections that each appear to have their own errors. A similar research on other data projects may give different results and a more general insight in the impacts of data errors on the outcomes of analyses.

REFERENCES

DATELINE (2003) *Deliverable 7, Data Analysis and Macro Results*

Frei, A., Kuhnimhof, T., and Axhausen, K.W. (2010) Long distance travel in Europe today: Experiences with a new survey, *89th Annual Meeting of the Transportation Research Board*, Washington DC

Goeverden, C.D. van, Arem, B. van, Nes, R. van (2014) Volume and characteristics of long-distance travelling, paper to submit at the *International Conference Climate Change and Transport*, Karlsruhe, forthcoming

Hautzinger, H., Stock, W., and Schmidt, J. (2005) *Erstellung von Microdatenfiles zu Ein- und Mehrtagesreisen auf Basis der Erhebungen MiD und DATELINE, Schlussbericht*, Institut für angewandte Verkehrs- und Tourismusforschung e.V., Heilbronn/Mannheim

Kuhnimhof, T., Collet, R., Armoogum, J., and Madre, J-L. (2009) Generating Internationally Comparable Figures on Long-Distance Travel for Europe, *Transportation Research Record*, **2105**, pp. 18-27

World Tourism Organization (2006) *Tourism Market Trends Africa, 2005 Edition*, Madrid

World Tourism Organization (2006) *Tourism Market Trends Americas, 2005 Edition*, Madrid

World Tourism Organization (2005) *Tourism Market Trends Asia and the Pacific, 2004 Edition*, Madrid

World Tourism Organization (2005) *Tourism Market Trends Middle East, 2004 Edition*, Madrid