POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title: A robust datawarehouse as a requirement to the increasing quantity and complexity of travel survey data

Auteurs: Authors: Pierre-Léo Bourbonnais, & Catherine Morency

Date: 2018

Type: Article de revue / Article

Référence: Citation: Bourbonnais, P.-L., & Morency, C. (2018). A robust datawarehouse as a requirement to the increasing quantity and complexity of travel survey data. Transportation Research Procedia, 32, 436-447. https://doi.org/10.1016/j.trpro.2018.10.054

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL: https://publications.polymtl.ca/10610/

Version: Version officielle de l'éditeur / Published version
Révisé par les pairs / Refereed

Conditions d'utilisation: Terms of Use: CC BY-NC-ND

Document publié chez l'éditeur officiel
Document issued by the official publisher

Titre de la revue: Journal Title: Transportation Research Procedia (vol. 32)

Maison d'édition: Publisher: Elsevier

URL officiel: Official URL: https://doi.org/10.1016/j.trpro.2018.10.054

Mention légale: Legal notice: © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 32 (2018) 436–447

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Steering Committee for Transport Survey Conferences

A robust datawarehouse as a requirement to the increasing quantity and complexity of travel survey data

Pierre-Léo Bourbonnais*, Catherine Morency

Polytechnique Montreal, C.P. 6079, Station Centre-Ville, Montreal, QC, Canada H3C 3A7

Abstract

This research proposes a travel datawarehouse using dimensional modelling for promoting a more understandable structure, generating comparable results, providing faster access to data and accelerating publication of highlights. The adaptation of dimensional modeling to travel data encourages a better structure while integrating, enriching and enhancing data. It provides automated data processing and validation. The proposal of a dimensional model for travel data follows the expected development of transportation planning tools.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the International Steering Committee for Transport Survey Conferences (ISCTSC).

Keywords: travel, survey; datawarehouse; fact; dimension; Kimball; Origin-Destination; software; household-based; individual-based; person-based; highlights; animated maps; trip

1. Introduction

Whether it is to increase efficiency and sustainability or to plan for change, it is essential for the various actors involved in travel planning to draw an accurate picture of travel behaviors of the population that uses the transportation systems. The increasing amount of data, the increasing number and complexity of the software platforms used, and the increasing accountability of stakeholders for their decisions raise questions about the efficient use of human resources in this area.

* Corresponding author.
E-mail address: leo.bourbonnais@polymtl.ca

2352-1465 © 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the International Steering Committee for Transport Survey Conferences (ISCTSC).
10.1016/j.trpro.2018.10.054

The processing, validation and storage of travel data are the essential steps enabling analysts and planners to study travel behavior and to propose models and tools that support the management and the optimization of transportation networks. In this domain, individual-based analysis and the object-oriented approach have partially responded to the technological challenges of recent decades (Trépanier, 1999). However, due to the increase in the amount and sources of travel data (Rousseau, 2016), the burden and challenges associated with travel data fusion have only increased (Bayart et al. 2009; Mohammadian, Zhang, 2007; Venigalla, 2004). Moreover, the need to integrate new forms of transportation (carsharing, bikesharing, ridesourcing, multimodality, etc.) and new indicators of sustainability (Sioui, 2014; Sioui et al. 2013) has made data processing and storage even more complex. Thus, travel analysis has become a more burdensome task, often resulting in long delays in the production of highlights and difficulty in analyzing trends in travel behaviors. However, in the area of business intelligence, the development of datawarehouses designed to facilitate analysis has accelerated and simplified the integration of data from multiple sources (Jukic, 2006; Kimball, 1997). At the same time, it helped to simplify access to databases. In addition, these warehouses have enabled data enrichment and the generation of innovative visualizations through automated procedures and pre-calculations, while encouraging comparative and trend analyzes with comparable and durable attributes (scales and ranges of values that do not change from one dataset to another and from year to year).

This paper starts with a literature review that includes a general presentation of dimensional modelling. Then, a dimensional model for travel data is proposed and defined. The process of data validation and enrichment is then explained with examples of automatic generation of highlights and animated maps of trips. Finally, the conclusion includes the main contributions, the limitations, and the future enhancements that could make the travel datawarehouse even more efficient for travel behavior analysis.

2. Paradigms and definitions

This section first presents definitions surrounding the collection of travel data, and then exposes the main paradigms surrounding the management of travel data. It follows with an introduction to the dimensional modeling borrowed from the field of business intelligence for the establishment of high performance datawarehouses. Finally, the relation between datawarehousing and the creation of visualization objects like animated trips maps is explained.

2.1. Origin-Destination travel surveys

Origin-Destination surveys are an important source of data for the study of travel behaviors. During interviews, respondents are asked to provide their socio-demographic profile and to report the places they visited during a specified period while taking care to mention the activities as well as the modes of transportation they used to get there. This information provides an objective diagnosis of the travel behavior of the population.

2.1.1. Household-based and person-based survey

In a household survey, it is customary to collect the demographic characteristics of each member of the household and the trips they made during the specified period. Typically, these surveys involve only one respondent in the household, and this person is required to disclose information for the other household members as well. When a person-based survey is conducted, the interview involves only one individual and the goal is usually to collect travel data for this person only. Nevertheless, it is standard practice to collect minimal demographic information on the other household members for comparability with other surveys, for modelling (household structure being an important determinant of behaviors) or for sample weighting purposes.

2.1.2. Harmonization of travel surveys

In order to cope with the proliferation of questionnaires and survey methodologies in Europe in particular, the COST (European Cooperation in Science and Technology) action has proposed a reference document to harmonize methods of travel surveys (Armoogum, 2014). This standardization effort is promising. Although one of the objectives of this document is to propose harmonized data, the methodologies for processing, validating and storing data are not defined.

2.1.3. Travel objects modeling

Following the totally disaggregated approach to validate, analyse and model of data from travel surveys, Trépanier (1999) proposed an entity-association diagram representing the objects collected in a typical Origin-Destination household survey (Figure 1).

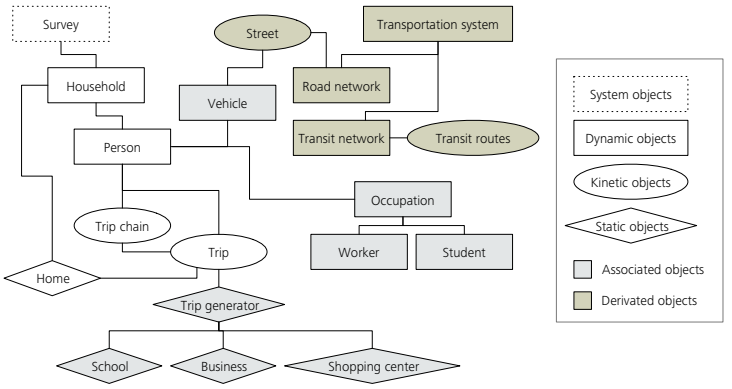

Figure 1 Entity-association diagram of objects related to travel behavior analysis (Trépanier, 1999) (reproduced with permission of the author)

Some years ago, the Ministry responsible for regional surveys in the Quebec province proposed a relational model called SAQE for data produced by travel surveys. It uses the same elements but also includes tables associated more specifically with the interview process. In addition, the data model lists the attributes collected for each object (table). To help understand the sequence of trips and the influence trips have on one another during the day, the concept of trip chains has been studied and defined by Valiquette (2010). Trip chains are not part of the SAQE model. However, they are generated after the data collection process.

While the actual structure of travel databases allows transportation planners to fulfill their analyses in a disaggregated manner, there is a need for solutions that are more robust and more comparable between surveys, and data structures that are more understandable and that provide faster access to enriched and validated data.

3. Datawarehouses and dimensional modeling

A datawarehouse is a data storage tool designed to facilitate and accelerate the analysis and enhancement of validated data generated from multiple sources. There are two main approaches to data management within a datawarehouse: the normalized model and the dimensional model proposed by Kimball & Ross (2013), both based on the relational data model. The normalized model respects the normal forms to avoid information redundancy. However, it increases the number of tables, leads to multiple joining processes during queries, and does not facilitate the interpretation of the results by looking at the schema itself. The dimensional model was designed to facilitate data analysis and interpretation, at the cost of some data redundancy, greater storage space and higher initial processing time. Queries from a datawarehouse based on the dimensional model are much faster because the tables have already been partially denormalized and in some cases aggregated for some analyzes. The database therefore does not have to join all the tables included in a normalized relational schema. In fact, the dimensional model recommends allowing only one degree of separation between the tables (single join), with limited exceptions. Figure 2 illustrates this fundamental difference between the two models. For example, the normalized model on the left has three degrees of separation between the sales transaction table and the representative table, while the dimensional model has only one degree of separation for the same relationship.

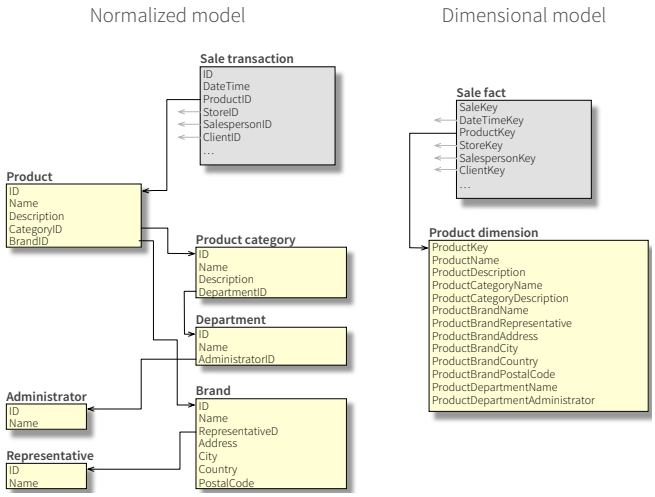

Figure 2 Comparison between the normalized model and the dimensional model

3.1. Dimensional modeling

According to Kimball & Ross (2013), dimensional modeling is an optimal response to two main goals: providing data that is understandable and quickly perform the requested queries. Dimensional models are also called star schemas. At the center of the star is a fact table representing the elementary objects of the field of study. For example, for a retail business, a record in the fact table is registered for each sale of a specific product at a given time (a sale transaction). At the points of the star, one finds the dimensions (the set of descriptions and elements of contexts related to the fact). In the case of a retail sale, the model will contain one table for the dates, a second for the products and their description, a third for the salespersons, a fourth for the customers, another for the stores, and so on. Figure 3 shows the star schema based on the dimensional model for a retail business. By instance, for the product dimension, the attributes of the joined dimensions are integrated into the table, whereas on the normalized side these attributes are separated into several distinct tables.

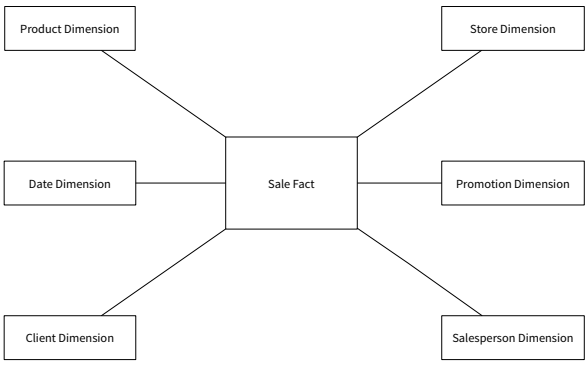

Figure 3 Example of a dimensional star schema for a retail sale business

3.2. Facts and dimensions

The fact table contains only measurable values (usually numerical values) and foreign keys to link the fact to its associated dimensions. Since it is the table that contains the largest number of records, the rule of not including text or unnecessary information reduces its size significantly. Most of the time, only fact tables should be joined to other tables (dimensions). The dimension tables contain textual descriptions, numerical attributes, and usually pre-calculated fields that enrich the data for a better understanding and easier analysis. Since they are joined by using keys

in the fact table, and the same dimension is normally associated with more than one fact, the number of records in the dimension tables is relatively small compared to the number of records in the fact table. It is important to note that the dimension tables are denormalized. For example, in the case of a product-specific dimension, product category information is embedded directly into the dimension. As a result, the information is repeated for each product in the same category. This paradigm increases the size of the dimension tables, but allows a better understanding of the dimensional attributes while promoting better performance during queries.

3.3. Datawarehouse and visualization

The datawarehouse allows information to be structured to facilitate knowledge acquisition and to simplify data processing from one domain to another (Adamson, 2010). The more the amount of data increases, the greater the need for a common structure and comparability between different data sources becomes crucial to perform rigorous analyzes under tight timeframes. In fact, the structure of a datawarehouse based on dimensional modeling allows scientists and the general public to select custom datasets in order to conduct their own explorations (Manoochehri, 2013). Buliung & Morency (2010) explain that visualization is an exploratory exercise that can encourage the proposition of new research hypotheses and, when well framed, prevent misinterpretations by non-specialists. Clarity, precision and efficiency are the most important criteria for valuing a visualization object and facilitating the understanding of complex human and physical phenomena. More specifically, geovisualization (MacEachren & Kraak, 2001) integrates the spatial dimension with visualizations and is an essential tool for the exploration and analysis of spatial data. Entire research groups and associations of visualization specialists (ARUP; Citylab; van der Wijk) study and propose visualization objects to clarify issues related to transportation. According to Pack (2010), simply making data more accessible by means of efficient visualization objects can convince the agencies in charge of transportation management to finance special programs for the creation of visualization tools. However, creating such tools requires a lot of time and resources. In addition, the new features offered are accompanied by increasing expectations from users, which in turn increases the resource and time requirements even more. However, to limit this escalation, a certain degree of automation in the creation of the visualization objects is required. Once the data is stored in a datawarehouse, many visualization objects can be constructed. However, the creation of these visualizations - animated maps of trips for example - requires considerable effort and tedious manual work (Harrower, 2004). Automating the creation of animated maps would reduce this burden and promote their dissemination. Knowing that the primary purpose of visualizations is to reveal trends and relationships that cannot be deduced without them (Meirelles, 2013), simplifying the process of creating such representations accelerates the acquisition of knowledge and allows a better understanding of the phenomena that verbal or written communication and conventional static illustrations do not reveal.

4. Travel datawarehouse

To contextualize the creation of a travel datawarehouse, the proposed modeling process is first outlined. The attributes that ensure the comparability of data from one survey to another are defined and the data enrichment process is described. The processing and validation methods used to convert the collection data to the datawarehouse are also specified. Thereafter, the characteristics of the different tables in the dimensional model are finally detailed.

4.1. Dimensional modeling of travel data

The first step in designing a dimensional model is to determine the elementary objects, that is, the granularity levels of the system. These are the elements that will be found in the fact tables. In the study of travel behavior, we find the following objects, from the grossest to the finest: the household, the person, the trip chain, the trip, the visited place (the trip's origin and destination), and the trip segment (segment of a trip using a unique mode). In travel behavior analysis, the elementary object is not always the trip segment. In fact, some analyzes do not require such precise information, and in that sense, may be satisfied with higher granularity (the household, the person, the trip chain, the trip or the visited place). As mentioned in Kimball & Ross (2013), the classification of the different objects of a system into facts or dimensions depends on the context. It is possible to find objects that are both fact and dimension. The

classification depends on the needs of the user. For example, an analyst who wishes to study the demographics of people regardless of the specific trips made by each person will be able to benefit from a model integrating persons as a fact table. Another, interested in schedules, durations and travel distances, but not by the unique modes used, will favor a model in which trips are the facts. Consequently, for the sake of flexibility, a distinct dimensional model is proposed for each of the main objects of the system. In this way, the analyst will simply choose the level of granularity desired and will perform its queries in the most efficient and effective way in consideration of his needs. The disadvantage associated with this choice is the greater amount of disk space required to store the data. Each fact table includes all keys referring to its associated dimensions, as well as boolean fields and a natural key (NaturalKey suffix) to retrieve the original object in the original survey database. Dimensional modeling usually provides only one degree of separation between facts and dimensions tables. However, there is an exception to this rule since the date and time dimensions can be referenced by keys in the dimension tables, resulting in a second degree of separation. This relaxation of the rules for this particular case simplifies the model and is specified in the Kimball documentation. Finally, a Universally Unique Identifier (UUID) that complies with RFC 4122 and ISO / IEC 9834-8: 2005 standards is included in each of the tables in order to integrate data from multiple surveys in a single database and ensure the uniqueness of each record.

4.2. Comparable attributes

One of the main objectives of creating a travel data warehouse is to facilitate comparison of data from one survey to another. To do this, a set of comparable attributes is proposed. The first characteristic of these attributes is their durability (they are also called durable attributes), that is, they are designed not to be altered or, at the very least, very rarely. If a change to one of the comparable attributes is planned, an update of all surveys stored in the datawarehouse must be performed to maintain the comparability of the data over time. The second feature is the inclusion of all possibilities to allow for rigorous comparisons. For example, a comparable mode attribute should include all possible modes, not only locally but also internationally. For the same reason, particular care must be taken in the definition and the proposed choices for each comparable attribute to avoid ambiguities. Some comparable attributes and dimensions have been proposed in this research. However, the final choice of comparable attributes and dimensions to be integrated into the datawarehouse must be the result of a possible consensus among stakeholders in the field of travel surveys.

4.3. Data enrichment

The creation of a datawarehouse is accompanied by a process of data enrichment. In this dimensional model proposal for travel data, several fields are pre-computed during import and several new attributes are added to facilitate and enrich the analyzes. For example, in the household dimension, several descriptive statistics are added: the average age of the household, the number of women and men, the number of students, workers and retirees, the number of members of each age group, the average and total number of trips, the average and total distances traveled by all members of the household and the minimum number of vehicles required in the household to complete the car trips declared.

4.4. Data validation

When filling a datawarehouse, the process of cleaning and importing data is called ETL (Extract, Transform, Load) (Adamson, 2010; Kimball, Ross, 2013). The ETL module must be updated when new data sources appear or when the imported files' format changes. As part of this research project, two separate data sources were used to construct the first version of the ETL module: data from various web surveys as well as data from various telephone surveys held in the Quebec region. The data is validated and processed to be enriched and imported into the datawarehouse. A configuration file indicates the required fields and the conditions of validity of the different information declared by the respondents. When one of the rules is not followed, a record is added to the collection database's audit table. All validations are flexible and may vary from survey to survey depending on the needs of the administrators and analysts.

4.5. Data consistency

Denormalizing data from the collection database into the datawarehouse dimensions is accompanied by a significant problem in terms of data consistency. For example, if a user changes the name of an activity, it must be changed in the activity dimension and in the tables that include activity names (visited places and trip chains, among others). However, in a dimensional model, it is the responsibility of the ETL module to verify the consistency of the data by making sure to modify the values in all affected tables. In that sense, the datawarehouse must be used read-only by all users. In fact, the ETL module must be the only instance with write permissions on all tables.

4.6. The fact tables and their dimensions

In this section, each fact and dimension is defined. Then, the fact tables (household, person, visited place, trip chain, trip and segment) are presented in a schema with their dimensions and described by their main attributes.

4.6.1. The survey and sample dimensions

Two dimensions are associated with all the fact tables. First, there is the survey dimension, that includes characteristics of the survey for which the different travel objects were collected (name and type of the survey and the organizations in charge). Second, there is a dimension describing the sample associated with each household and respondent who participate in the survey. This dimension includes attributes on recruitment and collection methods, sample sizes, and recruitment periods.

4.6.2. The date dimensions

In the date dimensions, several attributes are added allowing different sorting and filtering methods associated with dates. These include the name of the day, the number of the day in the week (1 to 7), the month (possible values from 1 to 31) and the year (possible values from 1 to 366) and the week number in the month (1-4 or 5) and the year (1-52 or 53). Another field indicates whether the date is part of the week (Monday to Friday) or the weekend (Saturday or Sunday). Finally, the year, month and day are also provided separately in their respective fields. Using this dimension, it is for instance possible to perform a query that groups trips by type of day (week or weekend) without having to group first by day of the week. This makes it possible to simplify queries and groupings, especially when several other sorts or filters are added to the query. The date dimension also serves to standardize all dates used in the datawarehouse, ensuring data integrity and comparability. In the date table, there is only one record per single date and thus the dimensions and facts that reference the same date always have the same date key.

4.6.3. Time and extended time dimensions

The time dimensions make it easier to group or sort by the hours or time periods of the day. A dimension table is created for the 24 hours of the day (0 to 23), while a second includes the entire 1440 minutes (0 to 1439) of a day. In these tables, we find attributes that represent the time, with or without the minutes, in several formats. For example, distinct fields are provided for the time in 24-hour international format and 12-hour AM / PM format. Another column includes the number of minutes since midnight, facilitating sequential analyzes that no longer require conversion or specific functions to measure time intervals. The extended time dimensions are similar, with the difference that they include hours from 0 to 47 hours (0 to 2879 minutes). The addition of these extended time dimensions ensures better consistency in the analysis of night trips since the format recommended by the General Transit Feed Specification (GTFS) for transit schedules can exceed 24 hours (one trip at night can start at 24:20 and end at 26:50, which is equivalent to 00:20 and 02:50 in the morning the next day).

4.6.4. Geographic object dimensions

Using the import interface, survey administrators can upload geographical data (shapefiles or geojson files commonly used for storing geographic data). Once the file is imported, the objects that belong to it and their category will be inserted into the geographic object dimension table of the datawarehouse. When processing data, each record with a geographic type attribute (home, usual place of work or study, visited place, etc.) will be linked to all geographic objects of the polygonal type associated (by means of a spatial intersection). A separate column is created for each

imported geographic object category. The attributes of the geographical objects will thus be available for all queries made on the associated fact tables. For example, if the census tracts of the survey territory are imported including the population and weight of each age cohort, for example, each home and visited place reported by the respondents will be intersected with the corresponding census tract zone and corresponding attributes and geography will be associated.

4.6.5. Network routing dimension

The trip network routing dimensions are proposed for collecting information related to routing calculations for a particular trip. Routing for each trip is automatically calculated and stored for each of the main modes of transportation (walking, cycling, car, public transit, P&R using car and public transit, and B&R using bike and public transit). This facilitates comparisons between different modes for the same trip, without being limited to the mode declared by the respondent, and travel times for each mode can be fed to modal choice models. For public transit routing, the boarding and alighting stops, as well as each route and mode used (bus, tram, subway, train, etc.) are also recorded. In addition, adjusted travel times for walking and cycling according to demographic characteristics of the respondent are also included when typical speeds are provided for each gender and age groups.

4.6.6. Household fact

For each household, a dimension representing the home location (geographic location, declared address, corrected address, etc.) is associated with the fact table. The household dimension includes its characteristics (household size, number of available cars, type of dwelling, etc.).

4.6.7. Person fact

The person fact table contains all respondents and valid household members. Each person is associated with a dimension representing the home, a household dimension, two dimensions for the usual places (one for the usual place of work and the other for the usual place of study, both being optional), a person dimension that includes the person's characteristics (age, age group, gender, main occupation, possession of a driver's license, possession of transit pass, language of interview, etc.). This person dimension also has comparable attributes that are the same for all surveys stored in the datawarehouse (comparable age group, comparable occupation, etc.) as well as pre-calculated descriptive statistics on the trips made. Network routing dimensions are also included for the trip from home to the usual place of work /study, and the trip from the usual place of work/study to home. Finally, the person dimension of the survey respondent (also called a proxy when this person provides answers for the other members) is also included.

4.6.8. Visited place fact

Following the same logic as the person-made table, the visited place table includes a home, household, person dimensions as well as the dimensions of the usual place of work/study of the person who visited the place. In addition, there is a dimension describing the activity carried out at the visited place as well as the date and time dimensions of arrival and departure.

4.6.9. Trip chain fact

In addition to the dimensions described above (home, household, person, usual places), the trip chain table is associated with the start and end date and time of the trip chain, and the main activity carried out at the anchor point of the trip chain (accompanied by the comparable main activity). The trip chain dimension includes attributes related to the chain category, as defined by Valiquette (2010) (constraint / unconstrained, simple / complex / open), the number of loops, the number of trips, distinct modes, segments, the duration of the main activity and the total distance traveled.

4.6.10. Trip fact

In addition to the dimensions of home, household, person, usual places, the trip fact table (Figure 4) includes the associated trip chain dimension, as well as dimensions for the origin and destination (visited places) and the dates and times for the start and end of the trip. In the case where the trip has at least one modal transfer between car or bicycle and a public transport mode, one dimension represents the transfer location. The trip dimension includes different attributes such as the number of segments, the number of unique modes used, the trip category (single mode or

multimode), the type of parking place for segments made as a driver, the driver type for trips made as a passenger, the distance traveled and the declared trip duration. The comparable mode category dimension describes the mode class imputed per the mode(s) declared (single mode or bimodal).

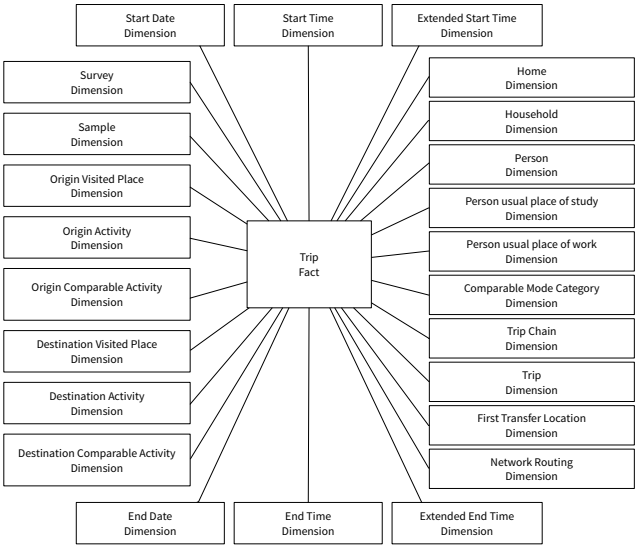

Figure 4 Trip Fact

4.6.11. Segment fact

The trip segment fact follows the same pattern as the one proposed for trip, but also includes dimensions for the public transit routes used and the boarding and alighting stops along the way. The vehicle dimension is added when the characteristics of the vehicle used in the segment are part of the questionnaire (vehicle type, capacity, year, make, model, odometer, etc.). Similarly, if information is available on the driver, a dimension representing the driver is included. The mode dimension includes attributes for the declared mode (name and description, comparable mode category, etc.).

4.7. Custom attributes

When a survey includes questions that are not present in a conventional Origin-Destination questionnaire, the datawarehouse relational schema is modified to include the attributes associated with the customized questions configured for the survey. To do this, the ETL module compares the list of questions in the original survey database with the list of attributes for each of the dimensions.

4.8. Weighting

Weighting is the process of expanding data to allow inference to the reference population. Weighing factors are assigned to each sample unit according to the sampling rate obtained for each geographical zone and/or socio-demographic group for which a census is available. In addition, if the analyst wishes to assign multiple weighing factors to the same object, the platform may add additional columns to include these factors.

4.9. Processing and validation

Processing and validating data in preparation for generating highlights and exporting data files is often time-consuming and difficult to implement. Indeed, for the 2013 Origin-Destination Montréal regional survey, for example, more than a year was required between the end of the data collection process and the production of highlights for 231,014 validated trips (Agence métropolitaine de transport, 2015), or approximately 630 trips per day on average. In

comparison, from the completion of the Toronto University Student Survey interviews in 2015, fewer than 10 days were required to process, validate and enrich the data and then generate highlights for 36,710 trips (approximately 3,700 validated trips per day on average, for a relatively comparable questionnaire). Although several validations must be done manually, the interview observation and validation module allows an administrator to enter directly into the respondent's interview to correct or verify the declared information. It is also possible to view the trips of the respondents for each interview on a single map per household. Whenever a change is made to the database during validation, the history is retained.

5. Visualization from datawarehouse

In the context of recent surveys used to implement the travel datawarehouse, animated maps of trips were produced for dissemination and analysis purposes. These maps allow the public and the analysts to visualize in time and space the trips made during an average weekday, in accelerated. In fact, it shows the traces of people and vehicles, with the aim of illustrating the use of the transportation networks of the area studied. Figure 5 shows a screenshot of one of these maps.

Figure 5 Animated trips map showing the trips reconstructed from the complete sample of the 2011 Québec regional survey
Full video: https://vimeo.com/162916534

The presentation of highlights is the first exercise to disseminate the results to the public and stakeholders after conducting a travel survey. The highlights should serve as a starting point for the more sophisticated and comprehensive analysis process that will follow. At the same time, the highlights must be adapted to their main audience, i.e. to all respondents, but also to the general population and the decision-makers and transportation planners who must take informed decisions in the light of the results obtained. Since the public typically does not possess a sharp knowledge of the field of study and, especially, rarely has the time to understand the methodology involved, it is essential that highlights are based on rigorous and objective analysis of the data, and are presented in a clear and friendly format. The developed platform includes a module for generating and presenting highlights. This module includes a set of graphs and comparative tables that facilitate the visualization of the data collected during travel surveys. Using the interface, it is possible to compare the results between different surveys and between different geographical zones.

6. Conclusion

The increasing complexity in processing and managing data collected through travel surveys has stimulated the need to simplify the validation and storage of these data. Although totally disaggregated modeling has accurately represented the objects associated with travel behavior, the increasing amount of data collected from multiple sources has only exacerbated the management challenge. As a solution to this problem, the main objective of the platform presented in this paper was to make the processing of travel data simpler, more flexible and faster, while enabling data enrichment and automated production of visualization objects. Thanks to the adaptation of the dimensional modeling commonly used in business intelligence, the implementation of a travel datawarehouse has been proposed with the aim of facilitating the integration, validation and visualization of travel data.

6.1. Limitations

The main limitation of using a dimensional model for travel data is the relative rigidity of the model. Although the dimensional model proposed for the travel data structure is a step forward, the model remains rigid on some aspects. It is difficult to integrate new dimension tables, new units or network routing calculation methods that were not included in the original design without modifying the ETL module and the database schema. One of the greatest difficulties in assessing the contributions of the platform was the comparison of the proposed methods of processing, validation and storage with those used elsewhere in the field. These methods are rarely documented, and when they are, they remain accessible only to the survey partners and administrators. Knowing this, the publication of methodologies for processing, validating and storing data upstream of the dissemination of results is strongly encouraged.

6.2. Future studies and proposed development

The question of the comparability of surveys requires further study. How do we deal with discrepancies in response choices, how to avoid ambiguities, what would be the loss of precision in responses if most attributes were converted into comparable attributes? Although the developed platform makes it easier to manage comparable attributes, the issues associated with the comparison of surveys must be assessed and measured. The future enhancements to the platform and to the travel datawarehouse are numerous. A discussion with travel survey managers around the world would help standardize the comparable attributes and choose the best enrichments of data that should be included by default in all surveys imported into the platform. Also, new visualization objects should also be proposed and implemented into the analysis dashboard. Regarding security, researches are needed to ensure confidentiality and enforce data protection and integrity in the travel datawarehouses. A large amount of work would be required to provide precise and complete documentation, with clear definitions of all the variables stored in the various tables, not to mention the need to include the methodology, the sampling process and the fusion strategies involved in each survey recorded in the datawarehouse. The sharing of travel data enriched and validated using the proposed structure has the general objective of facilitating the analysis of travel behaviors, feeding the models and simulations, and especially enhancing and extending the analysis power of travel data that was collected in the past and that will be collected in the future.

References

Adamson. 2010. Star Schema: The Complete Ref. Tata McGraw-Hill Education.
Agence métropolitaine de transport. 2015. Enquête Origine-Desination de Montréal 2013. Montréal: Agence métropolitaine de transport.
Andrienko, G., Andrienko, N., Bak, P., Keim, D.,Wrobel, S. 2013. Visual Analytics of Movement. Springer Science & Business Media.
Andrienko, N., Andrienko, G. 2006. Exploratory Analysis of Spatial and Temporal Data. Springer Science & Business Media.
Armoogum, J., INRETS. 2014. Survey Harmonisation with New Technologies Improvement, SHANTI.
ARUP. Bureau of Transport Statistics Data Visualisation. Retrieved March 10, 2016, from http://www.arup.com/Projects/Bureau_Transport_Statistics_Data_Visualisation.aspx
Bayart, C., Bonnel, P., Morency, C. 2009. Survey Mode Integration and Data Fusion: Methods and Challenges. In P. Bonnel, M. Lee-Gosselin, J. P. Zmud, J.-L. Madre (Eds.), (pp. 587–611). Presented at the Transport Survey Methods: Keeping Up with a Changing World, Bingley.
Buliung, R. N., Morency, C. 2010. "Seeing Is Believing": Exploring Opportunities for the Visualization of Activity–Travel and Land Use Processes in Space–Time
Citylab. Retrieved March 10, 2016, from http://www.citylab.com/
Harrower, M. 2004. A Look at the History and Future of Animated Maps. Cartographica, 39(3), 33–42.
Jukic, N. 2006. Modeling Strategies And Alternatives For Data Warehousing Projects. Communications of the ACM, 49, 1–8.
Kimball, R. 1997. A dimensional modeling manifesto. Dbms, 10, 58–70.
Kimball, R., Ross, M. 2013. The Data Warehouse Toolkit. John Wiley & Sons.
MacEachren, A. M., Kraak, M.-J. 2001. Research Challenges in Geovisualization. Cartography and Geographic Information Science, 28(1), 3–12.
Manoochehri, M. 2013. Data Just Right. Addison-Wesley.
Meirelles, I. 2013. Design for Information. Rockport Pub.
Mohammadian, A., Zhang, Y. 2007. Investigating transferability of national household travel survey data. Transportation Research Record: Journal of the Transportation Research Board, 1993, 67–79.
Pack, M. L. 2010. Visualization in transportation: challenges and opportunities for everyone., 30(4), 90–96.
Páez, A., Gallo, J., Gallo, Buliung, R. N., Dall'erba, S. (Eds.), 2010, Progress in Spatial Analysis (pp. 119–147). Berlin, Heidelberg: Springer Berlin Heidelberg.
Rousseau, G. 2016. New Directions in Travel Surveys: Big Data, Smartphones, and Stated Preference. Presented at the Transportation Research Board th Annual Meeting, Washington D.C.
Sioui, L. 2014. Pour une approche pragmatique et opérationnelle de la mobilité durable: Concept, méthodes et outils.
Sioui, L., Morency, C., Trépanier, M. 2013. How Carsharing Affects the Travel Behavior of Households: A Case Study of Montréal, Canada. International Journal of Sustainable Transportation, 7(1), 52–69.
Trépanier, M. 1999. Modélisation totalement désagrégée et orientée-objet appliquée aux transports urbains. Publications de Polytechnique Montréal.
Valiquette, F. 2010. Typologie des chaînes de déplacements et modélisation descriptive des systèmes d'activités des personnes. Publications de Polytechnique Montréal.
van der Wijk, J. Flowing City. Retrieved March 10, 2016, from http://flowingcity.com/
Venigalla, M. 2004. Household Travel Survey Data Fusion Issues. Presented at the Transportation Research Board.