Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 76 (2024) 397–408

Transportation Research Procedia

www.elsevier.com/locate/procedia

12th International Conference on Transport Survey Methods

Combining a regional household survey and passive data streams for longitudinal monitoring purposes

Elodie Deschaintres[a]*, Catherine Morency[a,c,d], Martin Trépanier[b,c,d]

[a]*Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, Montreal, QC H3T 1J4, Canada*
[b]*Department of Mathematical and Industrial Engineering, Polytechnique Montréal, Montreal, QC H3T 1J4, Canada*
[c]*Mobilité Research Chair, Montreal, QC H3T 1J4, Canada*
[d]*Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montreal, QC H3T 1J4, Canada*

Abstract

More and more passive big data are available, but travel surveys are still widely used in transportation. This coexistence brings questions about the role of each data source and new data fusion challenges. Within this context, this paper aims to combine the Montreal household survey and passive data streams to provide longitudinal monitoring of aggregated travel behaviors. The proposed methodology, based on time series decomposition, consists in applying the trend and seasonal variations detected in passive data to the typical fall weekday observed in the survey. This allows traditional mobility indicators, such as modal shares, to be projected and annualized.

Keywords: cross-sectional survey; passive big data; data fusion; time series decomposition; typical day; longitudinal monitoring

1. Introduction

Large-scale household travel surveys have long been conducted to assist in transport and mobility planning. In addition, over the past few years, the rapid development of information and communication technologies have made available a wide variety of passive datasets, from GPS coordinates, sensor and count data to cell phone, social media, smart card or other travel transactional data (e.g., shared mobility services). However, despite this multiple-source data landscape, transport research is still suffering from a lack of both longitudinal and multimodal data.

* Corresponding author.
E-mail address: elodie.deschaintres@polymtl.ca

2352-1465
10.1016/j.trpro.2023.12.064

Indeed, there is no single database describing the continuous use of all modes of transport at the same time. On the one hand, trips made with all modes are gathered in traditional surveys, but multiple low-frequency and irregular modes are not sufficiently observed. Furthermore, the information is cross-sectional (i.e., collected at specific points in time, with gaps of several years between them), and the trip diaries are pooled to represent an average "typical" weekday. Therefore, such data does not account for the variability of mobility behaviors, and the compiled information quickly becomes dated and less representative of current conditions. On the other hand, passive data are longitudinal, hence available continuously over long periods, but they are collected independently and heterogeneously for each mode; thus, they make the analysis of interactions between modes difficult.

Data fusion is one solution put forward in the literature to take advantage of all data sources and compensate their respective limitations. In this line, this paper aims to enrich a regional household travel survey (the Montreal Origin-Destination survey, conducted every five years) with passive data streams to allow longitudinal monitoring of the use of several modes of transport. It is the continuation of a previous work that demonstrated there are large similarities between the Montreal survey and passive datasets regarding the patterns of use of three modes, expanded and aggregated over several months; the same fluctuations, especially the weekly rhythms, were revealed in the two data types (Deschaintres et al., 2022). This first analysis validated there is potential to combine the two data sources, and the next step is to apply variations detected in passive data to mobility indicators measured with traditional surveys to cover a longer period. To this end, this paper proposes an automated procedure based on time series decomposition. This method is applied to the 2013 Montreal survey and validated with the 2018 survey. The results allow annualizing the daily modal shares of three modes and following their evolution over five years, between the two travel surveys, based on the seasonality and the long-term trend highlighted in passive data.

The remainder of this paper is organized as follows. First, some background elements regarding data fusion are provided. Then, the methodology is described, including all datasets used in this work and the processing steps. Finally, the results are exposed, and the paper is concluded by emphasizing the contributions and suggesting further investigations.

2. Background

Data fusion (or integration) is not new in transportation. Surveys have long been combined, for instance to merge revealed and stated preference data (Hensher et al., 1998), or census and travel survey data (Venigalla, 2004), the latter case being also involved in the weighting of sampled data. However, the use of this concept has intensified in recent years. Three causes were advanced by Bayart et al. (2009). First, travel surveys have become less representative because of declining response rates and less complete/accurate answers. This trend can be explained by different factors, in particular the exposure to a growing number of survey requests, causing fatigue effects, but also an increasing anxiety about revealing personal information. In addition, difficulties are encountered in reaching certain population subgroups. Therefore, recent examples of survey combination include merging several survey modes, or integrating a special-purpose satellite survey (e.g., targeting students) with a central household travel survey (Verreault and Morency, 2018; Wang et al., 2021). The second cause for the growing interest in data fusion according to Bayart et al. (2009) is the increasing awareness of the complexity of urban phenomena. This leads researchers to cross transportation databases with data from different fields, such as urban planning and health, or from economic, environmental, and social dimensions. Finally, the third cause (the one behind this paper) is the increasing availability of observation systems to collect passive data.

This emerging data is commonly seen as an opportunity to complement traditional data, especially surveys (Bonnel and Munizaga, 2018; Miller et al., 2018). The comparison of the two data types (passive data versus surveys) highlights the potential benefits of merging them. Surveys are cross-sectional and often based on single-day trip diaries, whereas emerging data is longitudinal and provides very precise spatial-temporal information. There are alternatives to cross-sectional surveys, e.g., multiday, panel and continuous surveys - see (Ampt, 2013; Ortúzar et al., 2011) for definitions and a review -, but they are much more expensive and laborious than emerging data, collected passively and therefore without respondent burden. However, emerging data is unidimensional, focused on mobility without contextual information (due to confidentiality and privacy concerns), while traditional surveys contain socio-demographic attributes of the respondents, which are necessary to explain mobility behaviors (Bayart et al., 2009; Cherchi and Bhat, 2018). Consequently, emerging data and traditional surveys answer the questions

"how?" and "why?" respectively (Callegaro and Yang, 2018). Other arguments for why emerging data cannot replace traditional surveys include concerns about data quality, completeness, and representativeness, as well as data access (Bonnel and Munizaga, 2018; Miller et al., 2018). In particular, few works benefit from enough data to study simultaneously different transport services (Morency et al., 2018). Furthermore, emerging data is generally massive (big data) and therefore difficult to handle, and data collection methods are still-evolving (Miller et al., 2018).

Therefore, both data types have pros and cons which can be exploited in a data fusion framework to maximize their value and overcome their limitations. To goal of the data fusion problem is to convey additional information from a donor database to a receptor database thanks to variables that are common across both databases. Data fusion procedures can be classified in two categories: exact matching, when two records (one from each database) are matched using a unique identifier, or statistical matching, based on explicit or implicit models (Bayart et al., 2009; Zhu et al., 2018). However, there are few applications of these methods in a transportation context. This lack may be due to several technical challenges, such as the heterogeneity of data formats and usage units (or travel indicators). In particular, as passive data was initially produced for functions other than mobility analysis and modeling, it does not contain trips as defined in travel surveys but transactions, counts, GPS points or other specific information (Bonnel and Munizaga, 2018). Beyond this semantic dissimilarity, there are also temporal and spatial incompatibilities between the data sources, which may have been collected with different resolution levels. In particular, the catchment areas and the corresponding reference population are different between travel surveys and passive data (residents of the whole region versus users, including residents and visitors, of a given system) (Miller et al., 2018). All these disparities lead to a lack of common variables and to the need of data aggregation. Therefore, many authors agree that new tools are required to allow the integration of various data sources, as well as to better define the role of both traditional and emerging data (Cherchi and Bhat, 2018).

The most common applications of data fusion with passive data can be found in road traffic engineering and intelligent transportation systems (El Faouzi and Klein, 2016). Several passive datasets were also used together to cross-validate or complete some information, such as origins and/or destinations (Giraud et al., 2016; Lovelace et al., 2016). However, for the reasons previously mentioned, examples of data integration between passive data and traditional surveys are much sparser. Nevertheless, some authors mined household surveys and smart card data for comparison or validation of spatial and temporal distributions (Munizaga et al., 2014; Spurr et al., 2015), and enrichment procedures (model-based or by implicit matching) were developed to add trip attributes, such as trip purpose (Kusakabe and Asakura, 2014), or sociodemographic information (Grapperon et al., 2016), from surveys to passive data. Therefore, these works capitalized on a travel survey (donor) to complete information in big data (receptor) or both data sources (both as donor) were valued for validation/comparison purposes, but none of them conveyed information from passive data (donor) to a traditional survey (receptor). Yet, the core-satellite data collection framework proposed in several papers (Miller et al., 2018; Wang et al., 2021) consider the household travel survey as the core and passive datasets as ancillary. This is also the approach adopted in this work, which aims to apply the variations captured in longitudinal data to a cross-sectional survey.

3. Data and Methodology

3.1. Data description and processing

This paper benefits from two types of data. First, the Origin-Destination (OD) regional travel survey of Montreal, Canada, is a telephone (and web since 2018) survey conducted every five years in the metropolitan area of Montreal to record the trips of about 4-5% of all resident households over one weekday in the fall. The last two surveys, namely the 2013 and 2018 ones (78,731 and 75,266 households respectively), are used in this work. Both were weighted by applying a margin calibration method (category totals being drawn from the Canadian census) to portray representative travel behaviors on a typical weekday. Therefore, they provide cross-sectional information; they allow calculating average daily mobility indicators, such as modal shares, for the 2013 and 2018 falls.

Secondly, several passive data streams are harnessed to describe the longitudinal variations in the use of three modes of transport from January 1, 2013 to December 31, 2018: validation data for the subway, and count data for bicycles and private cars. The subway network of Montreal is composed of 4 lines and 68 stations. STM *(Société de Transport de Montréal*), the transit authority, provided us with all tap-in validations (regardless of the support: smart

card or ticket) collected by the automatic fare collection system, called the OPUS system, at the entrance of all subway stations over the study period. Cycling data comes from the Eco-counter system of the municipal organization of the city of Montreal (*Ville de Montréal*), while highway and bridge counters from the CIR system of the Quebec Ministry of Transportation (MTQ: *Ministère des Transports du Québec*) are handled for traffic data. The cycling counts include personally owned bicycles and (station-based) bikesharing, operated by BIXI in Montreal, while only private car counts (without trucks) are extracted from the classified traffic data. However, in both datasets, there are only a few counters with almost continuous information between 2013 and 2018. Two counters by mode are selected, based on the number of missing observations and on their location (counters on bridges are preferred for the car because they are typically less affected by congestion). The cycling counters are located near the city center, while the car counters are positioned on bridges connecting the island of Montreal to the suburbs. Therefore, by their location, these counters mainly capture commuting trips, which are typically well represented in the Montreal survey (Spurr et al., 2015).

Some data preprocessing steps are required to aggregate the use of each mode (in number of validations or counts) into daily times series over six years. Validations and counts are summed by 24 hours, from 4:00 a.m. to 28:00 p.m. (4:00 a.m. the next day): this is the definition of a day in the OD survey, where trips collected from 24:00 to 27:59 are trips that began the day after the one being surveyed and that typically allow for the conclusion of a travel chain. When validations are missing over several hours and for several stations, the daily sum is also considered missing in the subway data. The selected cycling and car counts are first summed by counter, i.e., without distinguishing the direction, then by mode. When a daily observation is missing for one counter, then the sum with the other counter is also considered missing. At the end, $365 \times 5 + 366 = 2191$-day time series per mode are obtained, with 3, 8 and 122 missing values for the subway, cycling and private car respectively. Therefore, all datasets used in this work are incomplete, but also heterogeneous in nature (validations and counts in the passive data, versus trips in the OD survey). Moreover, the cycling and car daily volumes are not representative of the whole region. However, the developed method allows for missing values and outliers, and only the variations in the passive data (neither the values nor the metric) are of interest in the proposed framework.

3.2. Method overview

The fusion procedure developed to combine the Montreal OD survey and passive data streams is presented in Fig. 1.

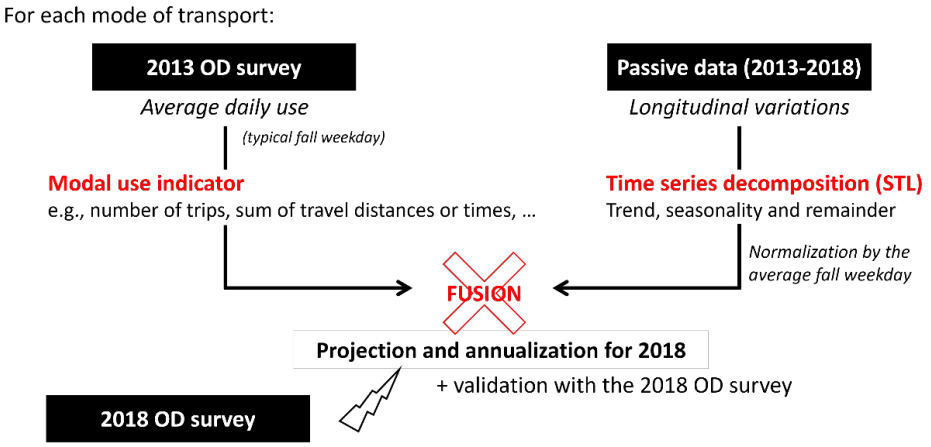

Fig. 1. Data fusion procedure between the Montreal Origin-Destination (OD) survey and passive data

This consists in applying the longitudinal variations detected in the passive dataset (converted into a daily time series) to the typical daily use observed in the OD survey, e.g., the 2013 survey, for each mode of transport. The variations in mode use measured in the passive data, available from 2013 to 2018, are highlighted using a time series decomposition method. This method breaks down the data into three types of components (a long-term trend, seasonality, and remainder) linked by an additive or multiplicative equation, depending on whether the components

are added or multiplied together to return to the original data. More specifically, the STL (Seasonal and Trend decomposition using Loess) algorithm is employed. This approach is detailed in the following section. The components are then normalized by the average fall weekday, leading to standardized coefficients which are applied to a modal use indicator, such as the number of trips made with a given mode, calculated from the 2013 OD survey. This results in projecting and annualizing the typical use indicator observed in the 2013 fall over six complete years, including day-to-day fluctuations. The latest OD survey (2018) is finally used to validate the results by comparing the predicted average number of trips in the 2018 fall to the value estimated from the survey. In this way, the longitudinal patterns extracted from the passive data allow filling the 5-year gap between the two surveys, and accounting for the variability of mobility behaviors over time rather than a typical fall weekday.

3.3. Seasonal and Trend decomposition using Loess

There are several time series decomposition methods, but the STL method, developed by Cleveland et al. (1990), is chosen in this work because it has many advantages over the others. In particular, STL is more flexible and allows for missing values, which are present in our data. The goal of the method is to decompose a time series Y_t into a trend component T_t (i.e., low frequency variations in the data and long-term changes in level), one or several seasonal components S_t (i.e., variations in the data given a seasonal frequency) and a remainder component (i.e., residuals) R_t. The additive form is written as follows, for all t observations:

$$Y_t = T_t + S_t + R_t$$

STL does not provide facilities for multiplicative decomposition, but this can be reached by taking the logarithm of the data, such as:

$$\log(Y_t) = \log(T_t) + \log(S_t) + \log(R_t) \quad \text{or} \quad Y_t = T_t * S_t * R_t$$

To get this decomposition, STL applies a sequence of smoothing operations based on locally weighted regression, or Loess (i.e., local estimation of a polynomial of degree d for each value from its q nearest neighbors, with weights decreasing with the distance from the value). The full algorithm is detailed below. It encompasses an inner loop, made up of several steps to estimate the trend and the seasonal components, nested inside an outer loop, which calculates the remainder component and can reduce the influence of aberrant points (outliers); thus, it allows for robust estimation. If needed, robustness is achieved by iterated weighted least-squares fitting in the loess smoothing operations: the neighborhood weights are multiplied by robustness weights (the largest values in absolute of the remainder are given the lowest robustness weights).

STL - GENERAL ALGORITHM

Outer loop (n_o iterations)
- **Inner loop** (n_i iterations)
  - Step 1: Detrending ➔ $Y_t - T_t$
  - Step 2: Cycle-subseries (for seasonality) smoothing with $q = n_s$ and $d = 1$ ➔ C_t
  - Step 3: Low-pass filtering of the smoothed cycles-subseries with $q = n_l$ and $d = 1$ ➔ L_t
  - Step 4: Detrending of the smoothed cycles-subseries ➔ $S_t = C_t - L_t$
  - Step 5: Deseasonalizing ➔ $Y_t - S_t$
  - Step 6: Trend smoothing with $q = n_t$ and $d = 1$ ➔ T_t

Calculation of $R_t = Y_t - T_t - S_t$ and robustness weights for each observation (outlier detection)

STL - DECOMPOSITION PARAMETERS

n_p: number of observations in each period or cycle of the seasonal component | based on the data properties
n_o: number of iterations in the outer loop | $n_o = 5$ *or* 10 if robustness is needed, $n_o = 0$ otherwise
n_i: number of iterations in the inner loop | $n_i = 1$ if robustness is needed, $n_i = 2$ otherwise
n_s: smoothing parameter of the seasonal component | odd integer greater than or equal to 7; based on diagnostic plots and/or prior knowledge
n_l: smoothing parameter of the low-pass filter | $n_l = [n_p]_{odd}$
n_t: smoothing parameter of the trend component | $n_t = [1.5n_p/(1 - 1.5/n_s)]_{odd}$

Notation: $[x]_{odd}$ is the smallest odd integer greater than or equal to x

The inner loop is composed of six interconnected steps which aim to prevent the trend and seasonal components to compete for the same variations in the data. First, the data is detrended, the trend being zero in the initial pass of the loop (step 1). The detrended series is then used in steps 2, 3 and 4 to assess the seasonal component. n_p, the number of observations in each period (or cycle) of the seasonal component, depends on the data properties. For instance, if the series is monthly with an annual periodicity, then $n_p = 12$. Cycle-subseries are the subseries of values at each position of the seasonal cycle, for instance the January values, the February values, etc. These cycle-subseries are first smoothed by loess and all smoothed values are gathered in the temporary series C_t (step 2). Next, a filter consisting of the application of several moving averages followed by a loess smoothing is applied (step 3), which leads to L_t. Finally, the smoothed cycle-subseries are detrended by subtracting L_t from C_t, resulting in the seasonal component S_t (step 4). The initial data are then deseasonalized (step 5), and the trend component is obtained by loess smoothing of the deseasonalized series (step 6). Therefore, three loess smoothing operations are included in the procedure (steps 2, 3 and 6). If the time series has two or more seasonal components, the procedure is repeated: the components are successively estimated (from the shortest-period component to the longest-period component), subtracted out, and the next component is determined from the residuals. The final trend component is then computed at the end of all iterations.

Cleveland et al. (1990) prescribed values for almost all parameters (number of iterations and smoothing parameters) of the procedure, based on the mathematical properties of the decomposition. Note that all polynomial degrees involved in the smoothing operations are equal to 1; this means that the components are locally linear. Only the choice of n_s, the smoothing parameter of the seasonal component (i.e., the number of consecutive observations to be used when estimating C_t at step 2), which is intrinsically linked to n_t, the smoothing parameter of the trend component, is less straightforward. This parameter controls the amount of variation in the data that makes up the seasonal component. In other words, it determines how rapidly the component can change. The cycle-subseries can be plotted to decide how smoothly the values should evolve from one cycle to another (the higher n_s is, the smoother or lower the evolution). Therefore, the choice of n_s is generally based on diagnostic plots and a priori knowledge of the data.

See (Cleveland et al., 1990) for more methodological details and (Hyndman and Athanasopoulos, 2018) for practical applications using the R statistical software package *fpp2*, that is used in this work. Furthermore, Wang et al. (2006) proposed an indicator whose value is between 0 and 1 to measure the strength of each component:

$$F_T = \max_t\left(0; 1 - \frac{Var(R_t)}{Var(T_t+R_t)}\right) \text{ and } F_S = \max_t\left(0; 1 - \frac{Var(R_t)}{Var(S_t+R_t)}\right) \text{ where } Var \text{ is the variance}$$

The strength of the component is higher when F is closer to 1, i.e., when the deseasonalized data ($T_t + R_t$) or the detrended data ($S_t + R_t$) have much more variation than the remainder component (R_t).

3.4. Projection and annualization

In this paper, the components extracted from the passive data with the STL method are used to calculate projected and annualized curves (in daily number of trips) by mode. The projected curve, defined at all t observations (days) from 2013 to 2018, is equivalent to the trend component standardized by its average value over the 2013 OD survey period ($\bar{x}$ being the notation for the mean), and then applied to the typical number of trips calculated from the 2013 OD survey:

$$Projection^{mode}_{t\,(2013\rightarrow2018)} = \frac{T_{t\,(2013\rightarrow2018)}}{\overline{T_{t\,(2013\ OD\ survey\ period)}}} * \#trips_{2013\ OD\ survey}$$

The predicted number of trips in the 2018 OD survey is computed by averaging the resulting values over the 2018 OD survey period. Therefore, this method allows estimating a projection that is not affected by seasonality (which is also the case in the OD survey), nor by data outliers.

Similarly, the annualization for a specific year (e.g., 2018) without the trend influence is made by applying the seasonal component (which can be a sum or a multiplication of several components), standardized by its average value in the 2018 OD survey period, to the typical number of trips calculated from the 2018 OD survey. The equation differs between the additive and multiplicative forms, the former being more complicated and requiring additional standardization to be able to add values with the same metric. For both forms, the equation is formulated

so that the mean of the annualized curve over the 2018 OD survey period is equal to the typical number of trips observed in the 2018 OD survey.

$$Annualization\ [add]^{mode}_{t\ (2018)} = \#trips_{2018\ OD\ survey} + \left(S_{t\ (2018)} - \overline{S_{t\ (2018\ OD\ survey\ period)}}\right) * \frac{\#trips_{2018\ OD\ survey}}{\overline{Y_{t\ (2018\ OD\ survey\ period)}}}$$

$$Annualization\ [mult]^{mode}_{t\ (2018)} = \frac{S_{t\ (2018)}}{\overline{S_{t\ (2018\ OD\ survey\ period)}}} * \#trips_{2018\ OD\ survey}$$

4. Results

4.1. Passive data decompositions

The STL method is first applied to each passive dataset. Since the data was aggregated at the daily level, two seasonal components are estimated: one for the weekly seasonality and one for the annual seasonality, with $n_p = 7$ and $n_p = (365 * 5 + 366)/6 = 365.17$ respectively. The calculation of the annual periodicity is based on the study period, composed of five years with 365 days and one year (2016) with 366 days. Note that it is not possible to delete February 29, 2016, without impacting the weekly seasonality. The crucial choice of n_s is supported by diagnostic plots (available upon request to the authors). The changes from one year to another or one week to another are assumed to be slow (or smooth); thus, n_s should be high. In addition, an iterative loop (successively testing different odd values of n_s from 7 to 99) is executed to find the optimal value in the validation of the procedure in Fig. 1, i.e., giving the best projection for the 2018 fall average indicator (or the lowest error). The selected parameters are provided in Table 1 for each mode, decomposition form and seasonal component. All other parameters are determined using the prescribed values. Robust estimation ($n_o = 15$) is chosen because there may be outliers in the passive data, especially in the cycling and car counts.

Table 1. Selected values of the smoothing parameters of the seasonal components for each mode and decomposition form

Decomposition form	ADDITIVE		MULTIPLICATIVE	
Mode / Season	Week	Year	Week	Year
SUBWAY	33	33	45	79
BICYCLE	7	7	47	79
CAR	13	7	19	53

The results of the application of the method to the subway tap-in validations are displayed in Fig. 2, for both (additive and multiplicative) decomposition forms. The raw data (time series of the daily number of validations) is illustrated in the first row, while the following graphs represent the different components of the decomposition. Since the two decompositions are very similar, common comments are made below.

The trend component, plotted in the second position, indicates a slow growth in the total daily subway demand, except in 2015. This long-term curve can be compared to the annual ridership published by the STM in open official reports and similar profiles will be found. However, here the trend is described with one value per day, and not only one value per year. Note that the trend (and the seasonal components) is available at all days, even when data is missing; this is possible thanks to the loess smoothing.

Unlike other decomposition methods, STL allows the seasonal components to change over time. Therefore, in this application, this makes it possible to capture variability from day to day, week to week and year to year. This flexibility allows us to see, for the subway, that the difference between weekdays and weekends is lower during the summer; the decrease in total ridership is rather considered in the annual seasonality. However, the corresponding seasonal component is quite regular from one year to another between 2013 and 2018. The most important dissimilarities are due to public holidays, which are not on the same day for all years.

Regarding the last component (remainder), we observe an increase in the residuals during the summer and holiday seasons. This could be explained by temporary factors, such as special events/activities or weather. As the estimation is robust, the impact of these specific cases on the calculation of the other components is reduced.

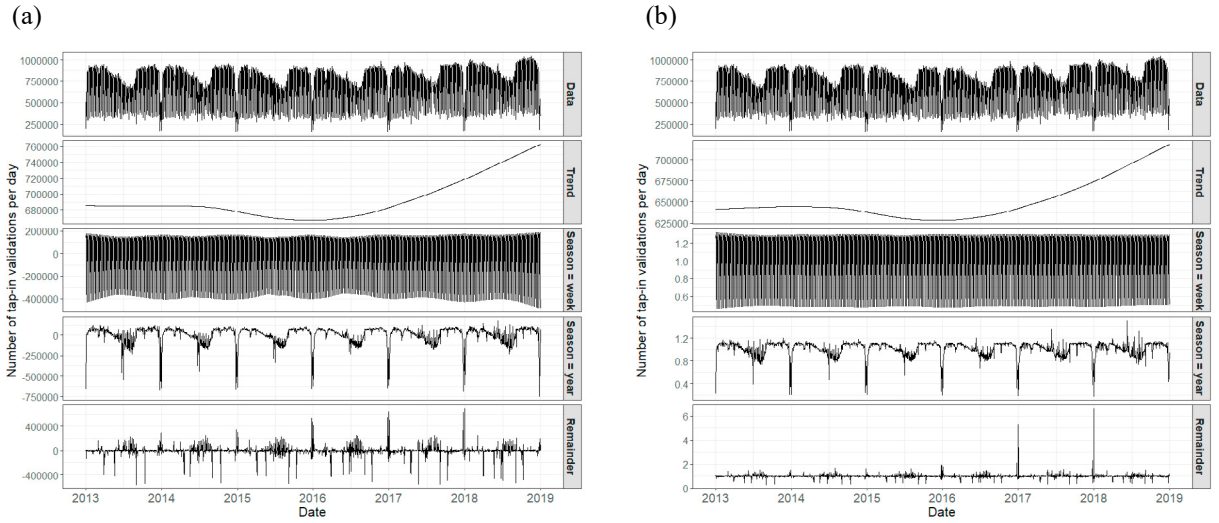

Fig. 2. (a) Additive and (b) multiplicative STL decompositions of the subway tap-in validations

The decompositions obtained from cycling and car counts are exposed in Fig. 3. Only the multiplicative form is presented because this form gives the best projection for 2018 (see the following section). For cars, the long-term trend is declining, with an acceleration in 2018. The seasonal components have points in common with those of the subway, but they are less regular. Indeed, the smaller n_s values in Table 1 indicate that changes are quicker. The usefulness of robust estimation is demonstrated since the unusual counts in 2015 have led to higher residuals. For bicycles, the trend is much less clear. The weekly seasonality varies over time, but the annual seasonality is very pronounced.

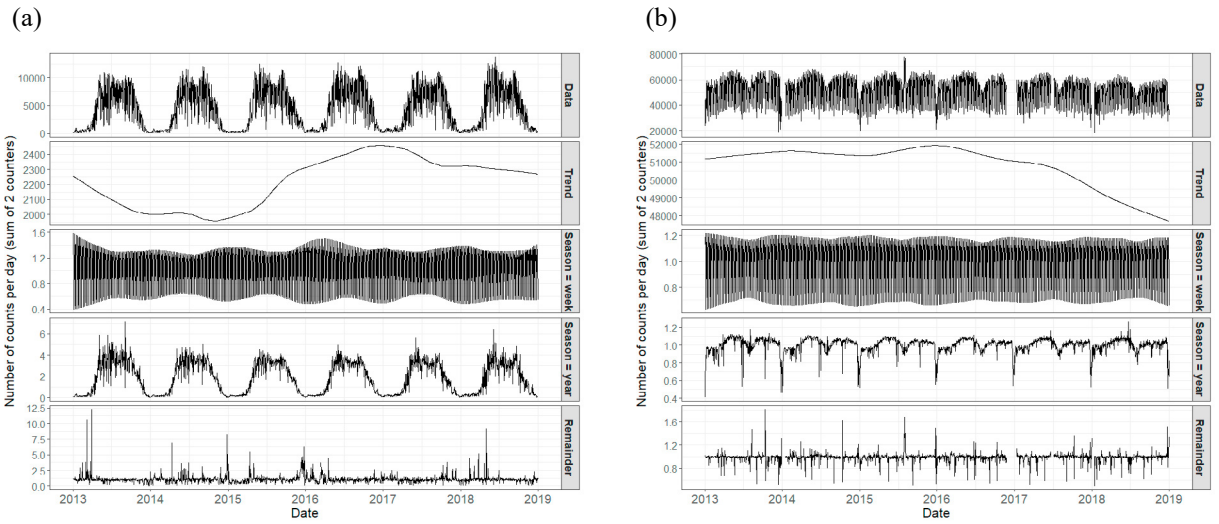

Fig. 3. Multiplicative STL decomposition of the (a) cycling and (b) car counts

This is confirmed by the strength indicators in Table 2 (for the multiplicative decomposition only, but the results are very similar for the additive form), which reveal that the strongest annual seasonality is measured for cycling. The strength of the seasonal components is analogous between the subway and car, and important weekly rhythms are highlighted. However, the trend is much lower for all modes; this can be due to the length of the study period (6 years), which may not be sufficient to observe strong long-term changes.

Table 2. Strength of the trend and seasonal components (F indicator) in the multiplicative data decomposition of each mode

Component	SUBWAY	BICYCLE	CAR
Trend	0.05	0.05	0.07
Weekly seasonality	0.84	0.36	0.81
Annual seasonality	0.47	0.89	0.50

4.2. Projected and annualized modal shares

Using the trend of the previous decomposition for each mode, the projected average number of trips on the 2018 fall weekdays is evaluated and compared with the observed number in the 2018 OD survey. The resulting error rates between projection (with optimal choice of the parameter n_s) and observation are provided in Table 3 by mode and decomposition form. In addition, these percentages are compared with those obtained when applying the traditional (disaggregated) demand projection method of MTQ (Quebec Ministry of Transportation). This method projects the weighting factor of each individual trip according to the anticipated evolution of socio-demographic characteristics. A Furness adjustment is also made to modulate the spatial distribution of the work-related trips. The error reported in Table 3 is the comparison between the number of trips calculated with the 2016 projected factors (estimated from the 2013 OD survey) and the observed number in the 2018 OD survey (weighted with the 2016 census).

Table 3. Error rates between the projected and observed number of trips on the 2018 fall weekdays

Method	SUBWAY	BICYCLE	CAR
MTQ method (2016 projected factors)	1.8 %	4.5 %	8.6 %
Fusion procedure - Additive form	3.4e-04 %	11.1 %	6.5 %
Fusion procedure - Multiplicative form	1.8e-03 %	9.4e-04 %	6.4 %

In all cases (except one), the errors assessed with the proposed procedure are smaller than with the MTQ's traditional method. Therefore, our method works pretty well. This was not initially obvious because of all the semantic, temporal, and spatial disparities between passive data and traditional surveys which were described in the background and data section. For the subway, the difference is almost zero, regardless of the decomposition form, whereas the errors are larger for bicycles and cars, but the count datasets are of lower quality and these modes may be more difficult to forecast. Nevertheless, the multiplicative decomposition seems to perform better, that is why it is chosen for the following analyses. This better performance may be because the logarithm helps reduce the influence of extreme values, resulting in a more Gaussian distribution for local regression.

The projected and annualized curves derived from the multiplicative decomposition of the subway validations (by applying the trend or seasonal components to the typical number of trips observed in the OD survey) are shown in Fig. 4. The black lines account for the numbers observed in the 2013 or 2018 OD survey.

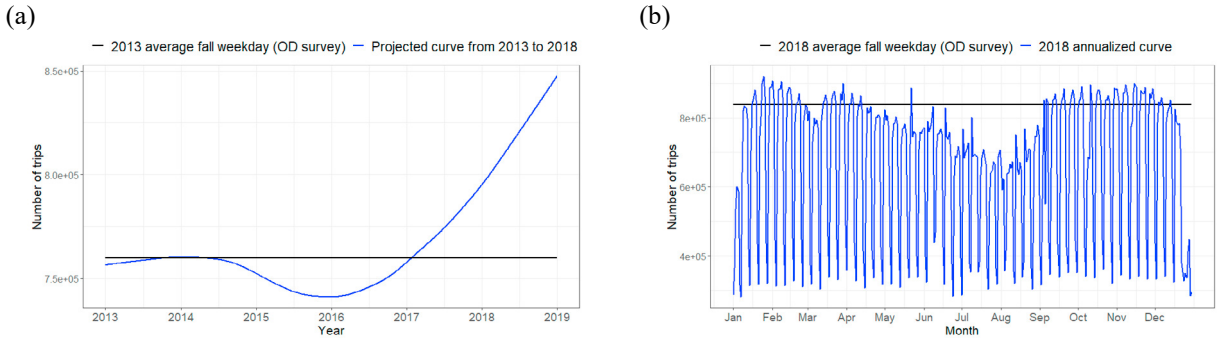

Fig. 4. (a) Projected and (b) annualized curves of the total number of trips per day in the subway

The same curves are computed for each mode, then the corresponding modal shares (in proportions of total trips) are calculated. The projected modal shares from 2013 to 2018 are represented in Fig. 5. Therefore, rather than having only one point in 2013 and one point in 2018 (i.e., the typical daily modal shares observed in the two surveys), the method allows for a longitudinal tracking of what happened during the 5-year gap. The variations are quite small, but this was excepted because of the slow changes in aggregated mobility behaviors.

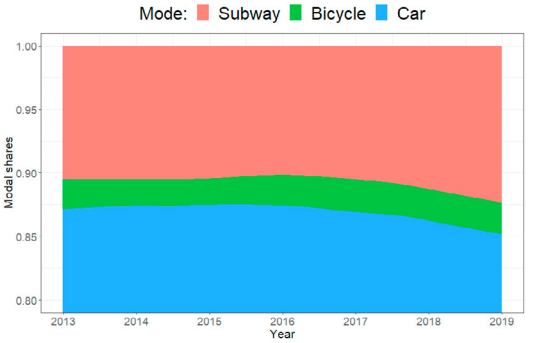

Fig. 5. Daily modal shares of three modes projected from 2013 to 2018

The annualized (or seasonalized) modal shares for 2018, as well as the average week across the 2018 fall, are exhibited in Fig. 6. Therefore, instead of an average number for the whole year, the method allows for a profile which varies by day of the week and day of the year. Looking at the variations by day type over an average week of the 2018 OD survey period plus weekends, we see that modal shares are not constant, even from Monday to Friday. This questions the representativeness of the "typical" weekday of the OD survey, already investigated by Verreault and Morency (2011).

(a) (b)

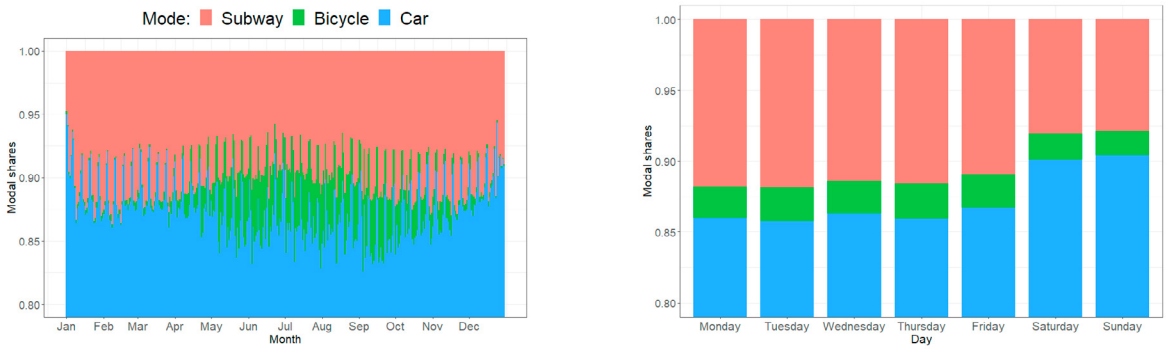

Fig. 6. (a) Annualized daily modal shares of three modes in 2018 and (b) average modal shares on a 2018 fall week

In addition, the variations of the daily modal shares selected over the 2018 fall on weekdays only (hence without considering the variability between weekdays and weekends) are described in Table 4. The results prove there is a lot of variations around the mean, especially for cycling.

Table 4. Variations across the 2018 fall of the daily modal shares (on weekdays only)

MODE	Minimum	Mean	Maximum
SUBWAY	7.8 % (- 33 %)	11.5 %	13.0 % (+ 13 %)
BICYCLE	0.2 % (- 91 %)	2.3 %	6.3 % (+ 167 %)
CAR	82.6 % (- 4 %)	86.2 %	90.1 % (+ 5 %)

5. Conclusion

This work proposes a framework to convey temporal information from passive data streams (donor) to a cross-sectional single-day household travel survey (receptor). Combining these two types of data, especially in that direction (i.e., to enrich a survey), has been little explored in the literature because of a lack of common variables between the data. However, in this paper, aggregation at the daily level allowed us to create links between them. The developed method, based on STL decomposition of time series, consists in applying longitudinal fluctuations (locally-linear trend and seasonality) extracted from passive data to the typical daily use observed in the survey for a given mode of transport. This method has suitable properties to work with passive big data: it allows for missing values, and can be robust to outliers, thus compensating data incompleteness and quality weaknesses. In addition, the computation is easy and fast, which makes it possible to deal with massive datasets. In this paper, the procedure was applied to the Montreal OD survey, conducted every five years on a typical fall weekday, for three modes: subway, cycling and private car. While it was only possible to calculate average numbers of trips by mode for the fall periods of 2013 and 2018 from the surveys, the method allowed monitoring their evolution over the 5 year-gap and estimating a profile which accounts for annual and weekly variations. In this way, modal shares, which are typically available on a one-off basis, can become continuous, and therefore both multimodal and longitudinal information can be obtained thanks to the complementarity between traditional surveys and emerging data. From this perspective, the role of traditional surveys is to collect average patterns of use of the modes and contextual information, whereas passive data supplements them by capturing variability over the year and between two surveys. This may obviate the need for longitudinal surveys (at least at the aggregate level), which suffer from additional economic and technical challenges. However, as new mobility services are still underrepresented in travel surveys, more information must be collected on these modes to include both traditional and emerging modes.

Further research involves testing the proposed framework with other modes, such as bus or taxi, and carrying out more disaggregated analyses: per subway station (spatial component), per fare product or for specific times (e.g., peak hours). The variations assessed around the mean thanks to the passive data could also be leveraged to develop variability factors to partially correct the current weighting factors in the OD survey and consider day-to-day or weekly variations over the fall. This could allow generating a distribution of values rather than an average value from the same data. In addition, modeling the residuals of the time series decomposition or integrating the estimation of a regression component in the STL procedure (by iterative weighted fitting) would be interesting to include the effects of weather, holidays, and events. Instead of a regression component, machine learning methods (e.g., random forests or neural networks) could also be investigated and combined with the decomposition procedure (Qin et al., 2019). This would allow for non-linearity and nonparametric specification, thus overcoming two potential limitations of the current method. Going even further, a longitudinal model could be developed to model the deseasonalized times series (rather than simply observe it in the data) and then forecast the future trend. In this paper, the trend was adjusted (interpolated) with two observations, one from the 2013 OD survey and one from the 2018 OD survey, but the ultimate objective would be to project this curve beyond 2018 using the passive data to forecast the number of trips that will be collected in the 2023 OD survey. Predicting the trend effect may depend on slow time-varying predictors, such as population evolution, but also on specific events such as supply improvements and changes in legislation which are implemented at a given date. The latter could be modeled using structural change point detection (Yeh and Lee, 2019). This forecasting exercise is even more relevant than ever in the context of the COVID-19 pandemic. Indeed, the changes which are currently impacting travel behaviors make it more difficult to predict the modal shares that will be observed in five years. However, a trend-based method may no longer be valid when such unexpected events occur. With more recent passive data, the effect of the health crisis on the trend and seasonality could also be analyzed. Teleworking, widely adopted in Montreal since March 2020, may have a strong impact on the weekly rhythms. Different scenarios for projecting mobility could then be proposed.

Acknowledgements

The authors wish to thank all the providers of the passive datasets used in this paper: STM - *Société de Transport de Montréal* (Montreal transit authority) for smart cart data, *Ville de Montréal* (municipal organization of the city of Montreal) for cycling count data and MTQ - *Ministère des Transports du Québec* (Quebec Ministry of

Transportation) for traffic count data, as well as the *ARTM - Autorité Régionale de Transport Métropolitain* (Metropolitan regional transportation authority) for giving access to the household surveys. They also acknowledge the financial support of the Canada Research Chairs (CRC), Quebec Research Funds (FRQNT) and Institute for Data Valorization (IVADO) programs.

References

Ampt, E., 2013. Workshop Synthesis: Longitudinal Methods: Overcoming Challenges and Exploiting Benefits, in: Zmud, J., Lee-Gosselin, M., Munizaga, M., Carrasco, J.A. (Eds.), Transport Survey Methods. Emerald Group Publishing Limited, pp. 393-406.

Bayart, C., Bonnel, P., Morency, C., 2009. Survey mode integration and data fusion: methods and challenges, in: Bonnel, P., Lee-Gosselin, M., Zmud, J., Madre, J.-L. (Eds.), Transport Survey Methods. Emerald Group Publishing Limited, pp. 587-611.

Bonnel, P., Munizaga, M.A., 2018. Transport survey methods - in the era of big data facing new and old challenges. Transportation Research Procedia. 32, 1-15. https://doi.org/10.1016/j.trpro.2018.10.001.

Callegaro, M., Yang, Y., 2018. The role of surveys in the era of "big data", The Palgrave handbook of survey research. Springer, pp. 175-192.

Cherchi, E., Bhat, C., 2018. Workshop Synthesis: Data analytics and fusion in a world of multiple sensing and information capture mechanisms. Transportation Research Procedia. 32, 416-420. https://doi.org/10.1016/j.trpro.2018.10.059.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: a seasonal-trend decomposition. Journal of official statistics. 6, 3-73

Deschaintres, E., Morency, C., Trépanier, M., 2022. Cross-analysis of the variability of travel behaviors using one-day trip diaries and longitudinal data. Transportation Research Part A: Policy and Practice. 163, 228-246. https://doi.org/10.1016/j.tra.2022.07.013.

El Faouzi, N.-E., Klein, L.A., 2016. Data Fusion for ITS: Techniques and Research Needs. Transportation Research Procedia. 15, 495-512. https://doi.org/10.1016/j.trpro.2016.06.042.

Giraud, A., Trépanier, M., Morency, C., Légaré, F., 2016. Data Fusion of APC, Smart Card and GTFS to Visualize Public Transit Use. CIRRELT. Retrieved from https://www.cirrelt.ca/documentstravail/cirrelt-2016-54.pdf.

Grapperon, A., Farooq, B., Trépanier, M., 2016. Information fusion of smart card data with travel survey. CIRRELT. Retrieved from https://www.cirrelt.ca/documentstravail/cirrelt-2016-59.pdf.

Hensher, D., Louviere, J., Swait, J., 1998. Combining sources of preference data. Journal of Econometrics. 89, 197-221. https://doi.org/10.1016/S0304-4076(98)00061-X.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: principles and practice, 2nd ed. OTexts, Melbourne, Australia.

Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. Transportation Research Part C: Emerging Technologies. 46, 179-191. https://doi.org/10.1016/j.trc.2014.05.012.

Lovelace, R., Birkin, M., Cross, P., Clarke, M., 2016. From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. Geographical Analysis. 48, 59-81. https://doi.org/10.1111/gean.12081.

Miller, E.J., Srikukenthiran, S., Chung, B., 2018. Workshop Synthesis: Household travel surveys in an era of evolving data needs for passenger travel demand. Transportation Research Procedia. 32, 374-382. https://doi.org/10.1016/j.trpro.2018.10.067.

Morency, C., Trépanier, M., Saunier, N., Verreault, H., Bourdeau, J.-S., 2018. Using 5 parallel passive data streams to report on a wide range of mobility options. Transportation Research Procedia. 32, 82-92. https://doi.org/10.1016/j.trpro.2018.10.014.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies. 44, 70-79. https://doi.org/10.1016/j.trc.2014.03.008.

Ortúzar, J. de D., Armoogum, J., Madre, J.L., Potier, F., 2011. Continuous Mobility Surveys: The State of Practice. Transport Reviews. 31, 293-312. https://doi.org/10.1080/01441647.2010.510224.

Qin, L., Li, W., Li, S., 2019. Effective passenger flow forecasting using STL and ESN based on two improvement strategies. Neurocomputing. 356, 244-256. https://doi.org/10.1016/j.neucom.2019.04.061.

Spurr, T., Chu, A., Chapleau, R., Piché, D., 2015. A Smart Card Transaction "Travel Diary" to Assess the Accuracy of the Montréal Household Travel Survey. Transportation Research Procedia. 11, 350-364. https://doi.org/10.1016/j.trpro.2015.12.030.

Venigalla, M., 2004. Household travel survey data fusion issues. Paper presented at the National Household Travel Survey Conference: Understanding Our Nation's Travel (Vol. 1).

Verreault, H., Morency, C., 2011. Transcending the Typical Weekday with Large-Scale Single-Day Survey Samples. Transportation Research Record. 38-47. https://doi.org/10.3141/2230-05.

Verreault, H., Morency, C., 2018. Integration of a phone-based household travel survey and a web-based student travel survey. Transportation. 45, 89-103. https://doi.org/10.1007/s11116-016-9726-2.

Wang, K., Hossain, S., Habib, K.N., 2021. A hybrid data fusion methodology for household travel surveys to reduce proxy biases and under-representation of specific sub-group of population. Transportation. https://doi.org/10.1007/s11116-021-10228-x.

Wang, X., Smith, K., Hyndman, R., 2006. Characteristic-Based Clustering for Time Series Data. Data Mining and Knowledge Discovery. 13, 335-364. https://doi.org/10.1007/s10618-005-0039-x.

Yeh, C.-F., Lee, M.-T., 2019. Effects of Taichung bus policy on ridership according to structural change analysis. Transportation. 46, 1-16. https://doi.org/10.1007/s11116-017-9778-y.

Zhu, S., Amirjamshidi, G., Roorda, M.J., 2018. Data Fusion of Commercial Vehicle GPS and Roadside Intercept Survey Data. Transportation Research Record. 2672, 10-20. https://doi.org/10.1177/0361198118768516.