Article

Not peer-reviewed version

Applying Data Analytics to Analyze Activity Sequences to Assess Fragmentation in Daily Travel Patterns—A Case Study in the Metropolitan Region of Barcelona

Lídia Montero * , Lucia Mejia-Dorantes , Jaume Barceló

Posted Date: 14 August 2023

doi: 10.20944/preprints202308.0949.v1

Keywords: Travel behavior; sequence analysis; gender; equity; classification analysis; spatial analysis; synthetic populations

Article

Applying Data Analytics to Analyze Activity Sequences to Assess Fragmentation in Daily Travel Patterns—A Case Study in the Metropolitan Region of Barcelona

Lídia Montero [1,*], Lucía Mejía-Dorantes [2] and Jaume Barceló [1]

[1] Universitat Politècnica de Catalunya (UPC); lidia.montero@upc.edu, jaume.barcelo@upc.edu
[2] Independent consultant; mejia-dorantes@web.de
* Correspondence: author: lidia.montero@upc.edu

Abstract: Sequence analysis is a robust methodological framework that has gained popularity in various fields, including transportation research. It provides a comprehensive approach to understanding the dynamics and patterns of individual behaviors over time. In the context of the Metropolitan Region of Barcelona, applying sequence analysis to the mobility surveys offers valuable insights into the sequencing and order of travel activities and modes, shedding light on the complex interrelationship between individuals, their travel choices, and the built environment. The Barcelona travel surveys collect detailed data on individuals' travel behavior, such as trip purpose, duration, mode of transportation, and origin-destination pairs. Sequence analysis allows for examining travel behaviors as dynamic processes, unveiling travel patterns' underlying structure and evolution in a day. A data analytics methodological approach is described; it enables the identification of common travel patterns and the exploration of variations across different demographic groups or geographical regions. Sequence analysis reveals insights into the factors influencing mode choice and potential opportunities for sustainable transport interventions. The paper proposes a methodological approach to discover homogeneous travel behavioral segments from diaries included in travel surveys in order to refine transport policies to selected segments by transportation planners and authorities.

Keywords: travel behavior; sequence analysis; gender; equity; classification analysis; spatial analysis; synthetic populations

1. Introduction

The need for a detailed understanding of transport demand and its behavioral patterns has prompted an interest in a better understanding of mobility, transportation, their relationships, interdependencies, and the factors that shape them. This point of view has led to paying attention to the activities necessary for the dynamics of our societies and to fulfill personal needs, focusing on the locations and the time the activities occur in a day.

The transportation system ensures the connectivity of the territory. It provides the means to access the activities subject to the time and distance constraints between the activities' locations, determined by the topography of the region's geography and its transportation network. According to [1], the transportation system is the tool that makes accessibility feasible and satisfying. In that way, the mobility demand, leaving aside marginal trips for other purposes, is a derived demand for the relevant journeys. As highlighted by [2], "Accessibility to the realization of activities is thus at the core of the process, and mobility must then ensure the completion of accessibility, where citizens and goods must reach destinations to satisfy needs and have access. Apart from the spatial dimension, individual characteristics shape citizens' decisions regarding place access". Different lifestyles determine these decisions [3], as well as socioeconomic, ethnicity, and gender issues [4].

Sequence analysis (SA) is a relevant approach for travel behavior analysis, as it describes the fragmentation and daily patterns in terms of strings of activities and the transitioning from one activity to another as well as the amount of time spent in each activity, as many researchers, such as [5,6], observe.

Two decades ago, [7] revealed that gender difference in travel patterns is linked to employment status, household structure, child care, and maintenance tasks. They found that travel patterns of men and women are much similar when considering families without children; when comparing multi-person households' males and females show more significant differences and are the highest for those living in homes with children. Over the past two decades, numerous studies have been conducted on travel behavior, showing gender as an influential factor in travel decision-making [8].

Their methodological approach is based on analyzing travel surveys, including a travel diary. Most travel surveys collect information about an individual (socioeconomic, demographic, etc.), their household (size, structure, relationships), their transportation habits, and a diary of their journeys (their start and end location, start and end time, travel modes, purposes of travel, etc.) on a given day, usually in a labor day.

Major travel surveys are conducted in metropolitan areas typically once a decade. Some metropolitan regions conduct a panel survey, which interviews the same people year after year to see how their particular behavior evolves. Traditional travel models rely almost exclusively on cross-sectional data, so individual/household travel surveys designed to capture people's behaviors and attitudes simultaneously have always been the most appropriate data collection tool. Either Spring or Fall are the selected seasons to collect household travel/activity surveys. These seasons coincide with the most common traffic data collection periods. In addition, they represent periods when schools are in session and when potential respondents are least likely to be away from their homes, typically on vacations.

Travel surveys are usually complex surveys. The organization and expansion of the data for analysis requires special care. The survey data can typically be analyzed using several units, including households, individuals, trips, or activities. A vast literature about the topic exists, and a well-cited reference is the FHWA Travel Survey Manual [9].

A travel survey file contains at least two basic tables: individual-related data and all trip-related data for trip-makers included in the sample in a given period (commonly using a single day).

GPS-based household travel surveys are becoming more prevalent in Europe and North America. Such surveys require independent household travelers to carry GPS devices such as loggers or phone-based applications. These surveys enable the collection of more accurate and precise personal travel behavior data, mainly in combination with prompted recall interviews.

Data collection is expensive, time-consuming, and not always straightforward, so care is needed in the planning, designing, and conducting surveys. Without this attention, resources – time, people, and money – can easily be wasted for little gain. High-quality and relevant data are essential for analyzing and supporting policy formulation and decision-making. Poor quality or inappropriate data are to the detriment of informed decision-making.

Traditional demand models rely on a trip-based approach, and travel diaries collect trips done on daily-based. Stopher [10] modified the travel diary concept to an activity diary, in which instead of asking first the question of what trip was made, and the purpose for the journey, the activity diary asked the trip-maker what was the next activity done and the transportation mode. Travel diaries are trip-based, while activity diaries focus on what the respondent did rather than on the places where they were doing such activities.

In 1995, the North Central Texas Council of Governments pioneered a time-use diary [11]. The primary difference between the activity diary and the time-use diary is that travel becomes another activity rather than a means to reach an activity. This diary has yet to become popular among Transportation Agencies [12].

Travel diaries are still one of the primary outputs from Household Travel surveys, and the sequence analysis addressed in this paper examines places visited by a person during a day jointly with the duration of activities at each site, travel mode episodes, and time spent to reach these places.

Furthermore, in the results reported in a forthcoming work [13], gender, education, and age are remarkable factors in switching between activities. Moreover, the authors also observe that, in addition to gender and other relevant factors, a deeper understanding of the spatial effects is necessary by exploring the relationships between the built environment and travel behavior and the influence of the spatial component on the sequences.

This study aims to continue with this line of analysis by exploring the components influencing the fragmentation of daily travel patterns in the Metropolitan Region of Barcelona. This paper is structured as follows: Section 2 sets up the context and datasets used in the computational experiments. Section 3 describes the methodological approach. Section 4 presents the results, and Section 5 draws the main findings and conclusions.

2. Materials and Methods

2.1. Context

Sequence analysis (SA) has become an invaluable tool in the realm of social sciences, offering insights into the structured occurrence of social events. At the forefront of this field stands Andrew Abbott, widely acclaimed as a trailblazer for his pioneering work in developing fundamental concepts and methodologies. His contributions extend beyond the mere ordering of events by historians and encompass how quantitative research addresses sequences in social processes. Throughout various publications, ranging from [14–18], Abbott elaborates on the evolution of these concepts and methodologies. According to Abbott [17] (pp 428), he asserts that "Social reality happens in sequences of actions with constraining or enabling structures [...]. It is a matter of particular social actors, in particular social places, at particular social times." This statement encapsulates the essence of sequence analysis, highlighting its focus on the interplay between activities, their context, and the temporal dimension in which social events unfold. By recognizing the significance of sequences and their inherent constraints and possibilities, researchers gain a deeper understanding of complex social phenomena. Abbott's groundbreaking work continues to shape and enrich the field, fostering new perspectives and avenues for exploration in social sciences.

This study is based on sequence analysis (SA). It analyses a series of time points at which a subject can move from a discrete "state" to another. States are usually based on people’s activities in places they visit and stay during the day, as graphically described in **Error! Reference source not found.**.

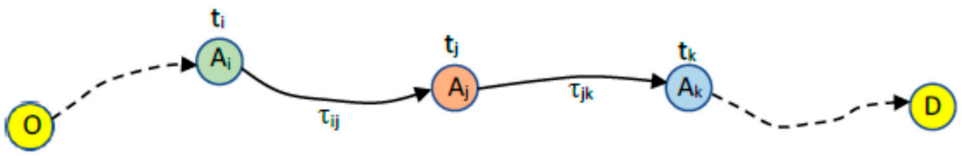


Figure 1. Sequences of activities.

Where it is the time spent in activity i, and τ_{ij} is the travel time between the place where activities i and j occur. Travel activities between these places are also considered a "state.” An example of a sequence could be HOME-GoToSchool-SCHOOL-GoToWork-WORK-GoToSchool-SCHOOL-GoToOther-OTHER-GoHome-HOME.

SA may offer valuable insights by highlighting differences or similarities among groups. Studer and Ritschard [19] highlight some features that are helpful in comparing activity sequences:

- Experienced states: These refer to each alternative activity in the sequence, such as being at home, work, school, or traveling by car, public transport, or other. State sequences can provide essential information that highlights group differences or similarities.
- Distribution refers to the total time allocated to each state within a sequence.
- Timing: This is the specific moment when each state appears within the sequence.
- Duration: This pertains to the time spent in each successive experienced state.
- Sequencing refers to the specific order in which distinct successive states occur. A sequence represents an ordered string of activities spanning a particular period.

Table 1 shows daily sequences for three hypothetical individuals in a travel survey, including the considered activities and travel modes and duration of the episodes. The activity proposed list is A (Escorting), C (occasional activity), H (staying at home), S (going to school/university), O (recurrent daily activities such as shopping and visiting family), and W (work) and travel modes grouped into TW (Walking), TB (Cycling), TP (Public transport), TC (private vehicle) and TM (e-Scooter, Segway). For example, in unit 1, the traveler spends 560 minutes from the start of the day at home, then reaches an occasional activity that lasts 110 min by public transport (80 min), to continue with a 5 min recurrent activity (O) reached walking (25 min) and a new occasional activity that lasts 210 min walking after a 30 min trip (TW) and finally the same public transport episode (TP) appears at the end of the day (60 min), arriving home and remaining there until the end of the day (23:59). Variety of activities is minimal for unit 2 (3 activities, repeated in the daily sequence), but 5 activities are included in unit 1.

Table 1. Daily-travel pattern example for three units in a travel survey.

Unit	Daily activity sequence	Used time per episode (min)
1	H-TP-C-TW-O-TW-C-TP-H	560-80-110-25-5-30-210-60-360
2	H-TW-O-TW-O-TW-H	660-60-480-10-10-10-210
3	H-TW-O-TW-H-TW-O-TW-H-TW-O-TW-H-TW-O-TW-H	600-2-28-3-27-2-58-2-28-2-13-2-373-2-28-2-268

Nevertheless, from travel diary files to daily activity sequence definition, a systematic data processing and analysis has to be performed, including some critical decisions affecting sample unit characteristics and the daily trip attributes as:

- Education. A qualitative variable is usually coded with many levels that group into a factor defining primary, secondary, and higher education groups.
- Professional activity. Either retired, unemployed, housemaker, student, etc.
- Age groups. They are grouped according to match local authorities commonly defined groups.
- Trip purpose. This feature is recorded in detail, but some categories can be grouped to simplify summarized results depending on the aim of the addressed study. Again, the grouping has to be consistent with the undergoing analysis presented by local authorities.
- Travel mode. A qualitative variable usually consists of many categories. Travel demand modeling needs ad-hoc grouping, depending on the aim of the analysis. Since travel mode analysis is critical, we will detail some strategies to define the principal mode of a trip and the day principal mode.

In travel behavior, the entire daily sequence of activities and travel may be quantitatively described by some indicators developed by numerous researchers [5,20–22]. These indicators allow for the analysis and measurement of activity durations and transition rates from one activity to the other, thus providing insights into the diversity and complexity of sequences. The indicators developed to address the limitations of not considering the ordering and number of state changes are the following:

The Entropy provides a measure of the variety in daily schedules in terms of an amount of the "prediction of the uncertainty". Whereas it accounts for the proportion of time allocated to each state during the day, it does not consider the number of state transitions.

On the other hand, the turbulence index depicts the intricacy of the daily schedule as a measure of variability in terms of different activities, the order of these activities, and the variance of the duration of these activities in a day. It is directly related to the fragmentation of time concept, indicating a lack of self-time and stress.

The Complexity index is based on the entropy and transitions within a sequence. It considers the order of successive states, measured by transitions, and the distribution of different states. This index is a normalized score [0,1].

doi:10.20944/preprints202308.0949.v1

Finally, the travel time ratio (TTR) represents the trade-offs people make between travel time and activity time, as it accounts for the total travel time in a day divided by the sum of the total time outside the home plus the total travel time in a day.

A detailed description can be found in [13].

2.2. Case study

The Metropolitan Region of Barcelona (RMB), depicted in **Error! Reference source not found.**, comprises about 200 municipalities. The 18 municipalities near Barcelona city define the EMT subarea (Primary Crown including Barcelona city divided into ten districts). The Metropolitan Area of Barcelona (AMB, Secondary Crown) subarea comprises 36 municipalities, including EMT, and ten districts for Barcelona's city, which accounts for 3.2 million inhabitants. It has a well-scattered public transportation network with over 200 bus lines, 4,000 stops, ten metro lines, 15 railways, and two tramway lines. More than 9 million trips are carried out every day. The rest of the RMB area consists of 164 municipalities and 1,848,514 inhabitants. More details may be found in [2].

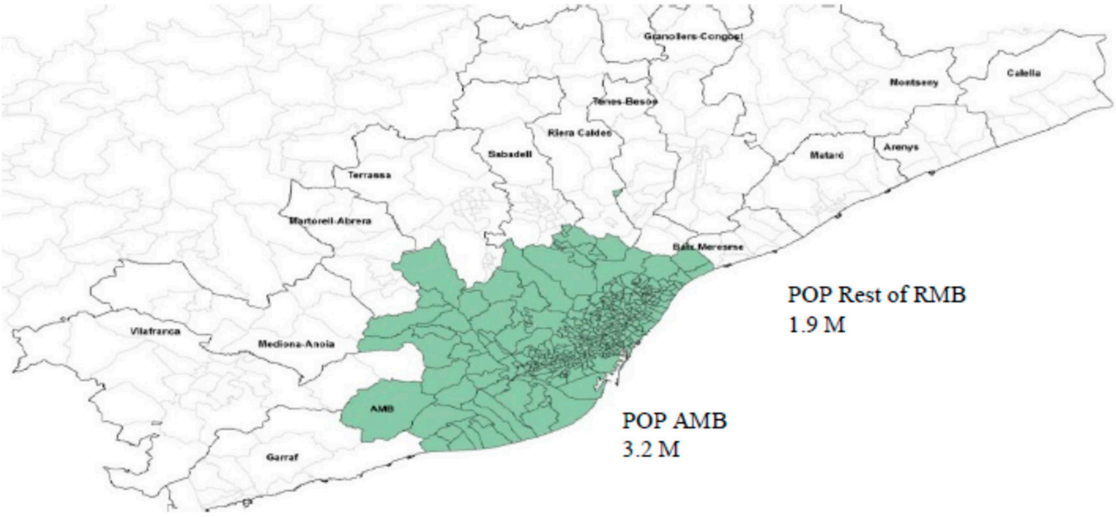


Figure 2. Barcelona Metropolitan Region (RMB) Study Area: Transportation Analysis Zones.

The area is divided into the TAZ-EMO Zoning System (522 transportation zones) for transportation planning purposes. The travel survey's zoning system is TAZ-EMEF (see **Error! Reference source not found.**), and each TAZ-EMEF zone aggregates several TAZ-EMO zones.

Table 2. Number of zones in the TAZ-EMEF Zoning System.

Crown	TAZ-EMEF
Barcelona City	10
Rest of Primary Crown (ETM)	17
Rest of AMB	18
Rest of RMB	128
Rest of Barcelona Province	134

2.3. Data Sets

In this study, we make use of four consecutive surveys for the Barcelona Metropolitan Region (RMB) "Working Day Mobility Survey (EMEF)"[23] from 2018 to 2021. Individual characteristics and the list of trips made the day before are included.

Additionally, to the trip purpose and travel mode necessary to build up the daily activity sequence data, we include the following relevant individual information available in the surveys:

- Education: A qualitative variable that groups education into basic, secondary, and higher education levels.
- Professional activity: Retired, unemployed, homemaker, student, and active.
- Gender: Either male or female.
- Age groups: the survey includes 16–29, 30–44, 45–64, and 65 and more.
- Other factors include car availability, residential area, modal use frequency, etc.

2.4. Data Processing

The following points summarize the undertaken processes and decisions:

- Data orchestration is needed to account for the four EMEF sources because they were delivered independently, and the recorded fields differ. The orchestration of EMEF datasets selects common subsets of fields and reorders them appropriately. Although EMEF data allow access to specific periods of the day, data orchestrations address the total number of daily trips.
- Characteristics of trip makers for EMEF datasets are gender (2 categories), age group (16–29, 30–44, 45–64 and 65, and more). EMEF 2019, 2020, and 2021 datasets do not contain residential zone for individual units (trip makers). Still, it has been imputed using the origin zone for the first trip in the day in home-based trips. This means that some units lack TAZ-EMEF residential area (only residential county is known); this subsample is less than 5% of the sample size.
- EMEF datasets contain characteristics such as education level (None, Primary, Secondary, and Higher) and professional activity group (student, active, unemployed, retired, and non-active). Unfortunately, family size and structure are missing for 3 out of 4 EMEF travel surveys. These data are being included in the EMEF surveys after 2021, and therefore they could not be consistently analyzed until the near future
- The maximum number of modes collected for any trip is 3. Travel time for each trip segment is unavailable, just the overall trip travel time (minute units).
- Individual sample sizes by year are 9.930, 9.934, 10.024, and 10.028, respectively, for 2018 to 2021 in RMB. The total number of trips in the sample is 39.318, 40.276, 34.714, and 35.209, respectively, from 2018 to 2021. After filtering professional drivers and inconsistent data, the total sample size for individuals is 37.877 units. Travel surveys are cross-sectional; no panels are available.
- The number of considered activities and travel modes consists of 11 elements: A (Escorting), C (occasional activity), H (staying at home), S (going to school/university), O (recurrent daily activities such as shopping, visiting family) and W (work) and travel modes grouped into TW (Walking), TB (Cycling), TP (Public transport), TC (private vehicle) and TM (e-Scooter, Segway).

3. Methodological approach

The proposed methodological approach is based on deriving activity sequences derived from travel diaries and analyzing travel behavior patterns. Most travel surveys collect information about an individual (socioeconomic, demographic, etc.), their household (size, structure, relationships), their transportation habits, and a diary of their journeys (their start and end location, start and end time, travel modes, purposes of travel, etc.) on a given day, usually in a working day.

The approach to be applied to travel surveys and daily travel behavior relies on the following steps:

- Data Preprocessing: The data needs to be preprocessed before applying sequence analysis. This task involves cleaning the data, handling missing values and multivariate outliers, and organizing the data into sequences based on the time order of activities. Each individual's sequence of activities becomes a series of ordered events. Quantitative time-fragmentation indicators are elaborated.
- Sequence Mining: Data analytics algorithms are applied to identify common patterns and sequences within the dataset produced by the data processing step. These algorithms can reveal frequent sequences, such as common travel patterns or recurring combinations of activities. Activity sequences are qualitative time series; there are some proposals to quantify the degree of similarity between sequences in literature. Nevertheless, we have selected a data analytics approach and considered likeness after projecting activity sequences in a real space resulting

from multiple correspondence analysis (MCA). Euclidean distances are applied to assess the similarity between projected sequences.

- Travel Behavior Comparison: Sequence analysis allows for comparing sequences between individuals or groups. By comparing sequences, researchers can identify typical or representative travel behavior patterns that can help understand variations in travel behavior based on demographic characteristics, such as age, gender, or socioeconomic status.
- Clustering and Typology Development: clustering on projected activity sequences obtained by MCA identifies distinct groups or clusters of individuals based on their travel behavior patterns. After clustering individuals with similar projected sequences, we identify typologies or travel behavior profiles representing different population segments.

Statistical analysis of sequences and fragmentation indicators allows analysis in greater depth, as indicated in the workflow shown in **Error! Reference source not found.** (the methodological workflow is inspired by [13]). Some potential analysis line relies on developing general linear models using a fragmentation indicator as a target variable and quantitative and qualitative explanatory variables such as gender, education, day principal mode, etc. The marginal effects of explanatory variables help clarify the multivariate association with the characteristics of the individual units.

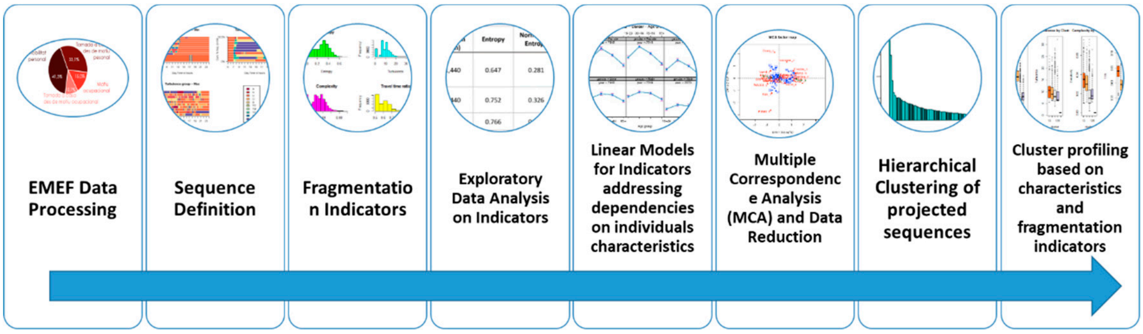

EMEF Data Processing
Sequence Definition
Fragmentation Indicators
Exploratory Data Analysis on Indicators
Linear Models for Indicators addressing dependencies on individuals characteristics
Multiple Correspondence Analysis (MCA) and Data Reduction
Hierarchical Clustering of projected sequences
Cluster profiling based on characteristics and fragmentation indicators

Figure 3. Methodological workflow.

It is worth noting that the sequences for all sampled units in the EMEF 2018 to 2021 travel surveys weighted by its expansion factor are produced using the TraMineR package in RStudio [20,24]. Afterward, fragmentation indicators (entropy, turbulence, complexity, and TTR) are elaborated from daily travel sequences using functions in the TraMineR package for RStudio.

Moreover, the principal transport mode for a trip and daily primary mode labeling is determined based on PCA and unsupervised classification [25].

3.1. Descriptive Analysis

Firstly, we address a basic descriptive analysis of fragmentation indicators for activity sequences of sample units in 2018 to 2021 EMEF Travel Surveys.

The exploratory data analysis of the fragmentation variables includes the following comments:

- Distribution of fragmentation variables for trip-makers and non-trip-makers.
- Univariate and multivariate outlier detection based on robust Mahalanobis distance [26].
- Spearman correlation coefficient between fragmentation variables with/without multivariate outliers to assess the association between selected fragmentation variables.

3.2. Data dimension reduction

We apply a data reduction technique to the dataset daily sequences by minute in such a way that the number of columns (the number of minutes by the number of activities (11)) retained are the first N=500 factorial axes in the Multiple Correspondence Analysis (MCA in FactoMineR package in R [27,28]) that account for more than 90% of data variability from 6 to 24h. Our input matrix to MCA is a 37,877 x 11,880 matrix (*complete disjunctive encoding)* containing activities from the selected alphabetic list (11 options), where each column represents one category of the minute factor (11 possible levels accounting for activities plus transportation options) and the output is a 37,877 x 500 matrix.

This procedure is beneficial for handling such a large dataset, which in our case comprises daily sequences. Our database thus remains highly detailed, as we maintain relevant information at 1 min based, rather than aggregating the timeframe, which would lead to losing information. Activity recording each 1 min of the day is the most suitable methodological approach that maintains the level of detail.

The multiple correspondence analysis projection procedure ensures the utilization of all available information (minute-to-minute activities) and avoids distortions during dimensionality reduction. The reduction is based on the extended Kaiser criteria [29], with the retained factorial axes accounting for more than 90% of data variability, thus ensuring representativeness. These factorial axes have been used to project the sequences into the new $\mathbb{R}^{500}$ space, achieving a 95%-dimension reduction and the possibility of addressing non-supervised clustering based on real numbers (instead of qualitative variables with 11 categories).

Afterward, we project data sequences in the N-dimension factorial space. Sequence projections are vectors of N real numbers. Then, we apply a hierarchical clustering technique for data discovery of clusters showing similar daily sequences. Clusters can be profiled based on characteristics and numerical variables from 2018 to 2021.

For clustering analysis, the similarity/dissimilarity matrix has millions of cells (37,877 x 37,877) that contain the dissimilarity scores for the sequences of each person in the working sample. The method seqdist() from the TraMineR package in R allows us to calculate the dissimilarity matrix based on several metrics [19] to the original minute sequences. However, it is not feasible in the original space due to large memory requirements. For this reason, we have applied multiple correspondence analysis (MCA) for data reduction instead of principal component analysis (PCA is suitable for numeric variables) to detect underlying structures in the dataset before clustering.

After the data reduction of all activity sequences (including multivariate outliers), we use a clustering technique to group sequences of activities with similar dissimilarity scores obtained from the sequence comparison after projection. The sample units' fragmentation indicators and characterization variables help interpret the clusters. We determine the final number of clusters by using an optimized criterion for balancing within-group similarity and between-group dissimilarity. Specifically, we apply Hierarchical Clustering (HC) [25] to the reduced projected data of daily travel patterns in the minute activity matrix. Each cluster comprises points that are more similar than those in other groups. The hierarchical clustering method in the FactoMineR package can reduce the computational burden by starting the agglomerative process on a heuristic partition that represents 10% of the original length. We cut the hierarchical agglomerative tree at a degree of similarity of almost 40%, using a balanced combination of commonly used techniques, such as the between the sum of squares to the total sum of squares, gap statistic, and silhouette methods.

3.3. Principal travel mode definition

In metropolitan areas, trips might be composed of several modes: leaving home in a car as a non-driver to reach a bus stop and, at some point, transferring to the train and arriving at the destination in a 5 min walk. The concept of principal mode is tricky and usually involves some decision-making. The maximum number of user modes is a design parameter in the survey named K; sometimes, the travel time spent in each stage is unknown. Let us assume that the sequence of used modes in a trip is mode1, …, mode K.

Rules of Assignment of principal trip mode (gmode) for K=3:

- If mode1 is defined and mode2 is None, then the principal mode gmode is "code1".
- If mode1 and mode2 are defined, and mode3 is None, then gmode is "code1:code2". For example, mode 1 is the car as a driver, and mode 2 is the bus, then gmode becomes "C:B".
- If mode1, mode2, and mode3 are defined, then gmode is "code1:code2:code3". For example, mode1 is the car as a driver, mode2 is the train, and mode3 is the bus, then gmode becomes "C:T:B".

- Repeat the process until the maximum number of stages has been considered.
- If data preparation shows some drawbacks as: mode1 and mode3 are None and mode2 is defined then gmode is defined as "code2".
- If mode2 and mode3 are defined, and mode1 is None, then gmode is "code2:code3".

If the trip segment duration is known, then "principal mode assignment" can be based on the mode taking the most extended period. Otherwise, it can be applied after a preprocessing step:

1. Identify *gmode* frequencies once performing a reduction of the number of possibilities based on unordered sets. For example, using car and bus would be C:B and assimilated to B:C (alphabetical order of the set code modes). Any mode composition involving W (Walking) is also set to the non-walking mode. For example: "T:W" is designated as "T" (train).
2. The number of occurrences of each code for each trip survey is addressed, and a Principal Component Analysis applies to the data matrix composed of n rows (as many as total trips in the sample) and as many columns as different mode codes. An unsupervised clustering analysis after principal components defines the final number of clusters meaning groups of transportation modes seen at individual trips. Thus, cluster representative modal combinations set the principal travel mode.

Each individual's principal set of modes along a working day can be helpful in labeling modal preferences in the expanded population. The analyst can apply the same steps to define "day principal mode" (*dpmode*), taking all the daily trips.

3.4. Fragmentation variable profiling

Each fragmentation variable determines a significant global association with quantitative and qualitative variables characterizing sample units, in this case, individuals. The quantitative variables are the number of daily trips and the total travel time in a day, and the qualitative variables are gender, education, professional activity, principal daily mode, declared modal preferences, etc. Each fragmentation variable quantifies those quantitative variables related to sample units, those qualitative variables where the fragmentation indicator mean is not homogenous for all categories, and those categories involving an indicator mean significantly different from the overall mean at 99% confidence.

Afterward, we use the Tukey multiple comparisons of means at 95% confidence. It is a particular case of Multiple Comparison Test (MCT). Tukey's Honestly Significant differences (HSD) can be used under equal variance assumption. The Tukey test is considered a reliable method to detect the difference during pairwise comparison (less conservative than others when applied to small samples) and increases the probability of rejecting the null hypothesis when small group sizes are present (this is not a problem in our data set since comparisons by years are needed and year sizes are large enough to justify theoretical assumptions). Tukey's HSD implemented in R [30] is the Tukey-Kramer test (original Tukey's HSD modified to cope with unbalanced data).

4. Results

The following sections present the results of the different analyses undertaken in this study.

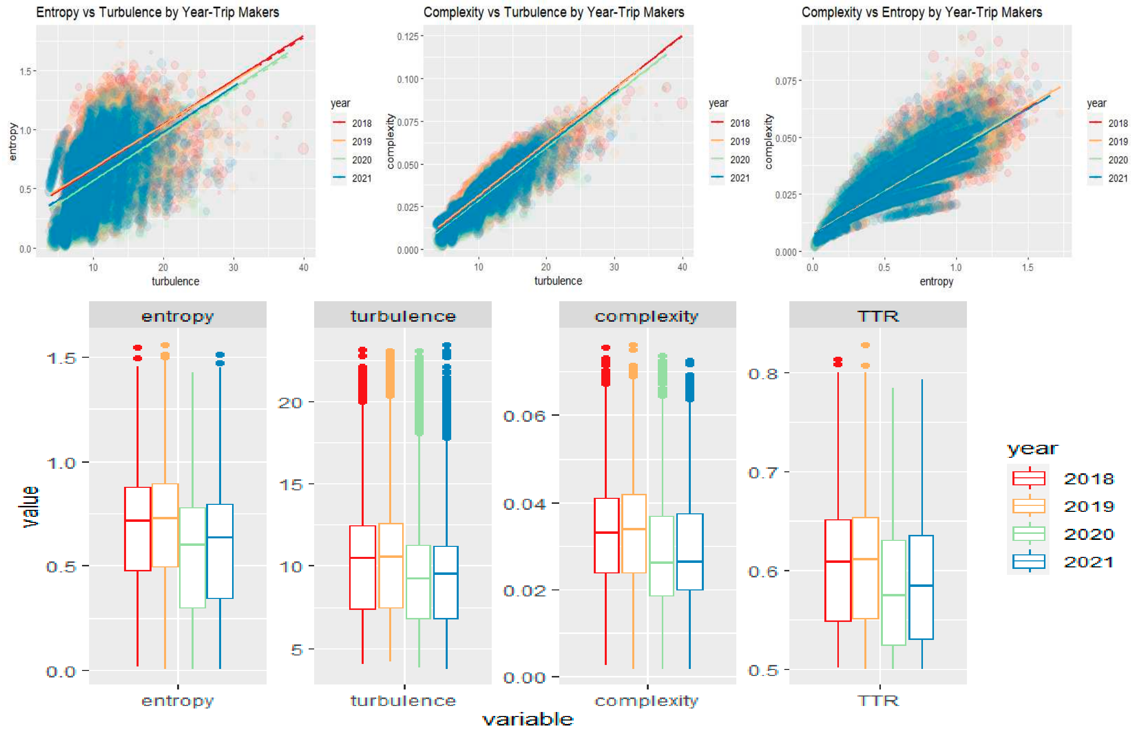
4.1. Descriptive analysis

This section shows the descriptive analysis of fragmentation indicators between 2018 and 2021. We derive fragmentation indicators from daily travel sequences using functions in the TraMineR package for RStudio [31]. **Error! Reference source not found.** shows fragmentation indicators calculated for daily sequences shown in **Error! Reference source not found.**.

Table 3. Daily-travel pattern example for three units of the working sample based on an alphabet of 11 activities.

Id	Daily activity sequence	Used time per episode (min)	Total duration (min)	Entropy	Turbulence	Complexity	TTR (Travel time ratio)
1	H-TP-C-TW-O-TW-C-TP-H	560-80-110-25-5-30-210-60-360	1,440	0.776	8.947	0.03020	0.610
2	H-TW-O-TW-O-TW-H	660-60-480-10-10-10-210	1,440	0.689	8.519	0.02846	0.623
3	H-TW-O-TW-H-TW-O-TW-H-TW-O-TW-H-TW-O-TW-H	600-2-28-3-27-2-58-2-28-2-13-2-373-2-28-2-268	1,440	0.272	16.751	0.02919	0.526

Error! Reference source not found. shows the fragmentation indicators. At the top, bivariate plots take the whole sample of trip makers. At the bottom, histograms for the same variables exclude multivariate outliers at a 99% confidence level based on robust Mahalanobis distance. Non-parametric Spearman correlation coefficients among fragmentation indicators excluding multivariate outliers show a direct association between them, being the more intense the complexity-turbulence pair: 0.6 (entropy x turbulence), 0.91 (complexity x turbulence) and 0.85 (complexity x entropy).



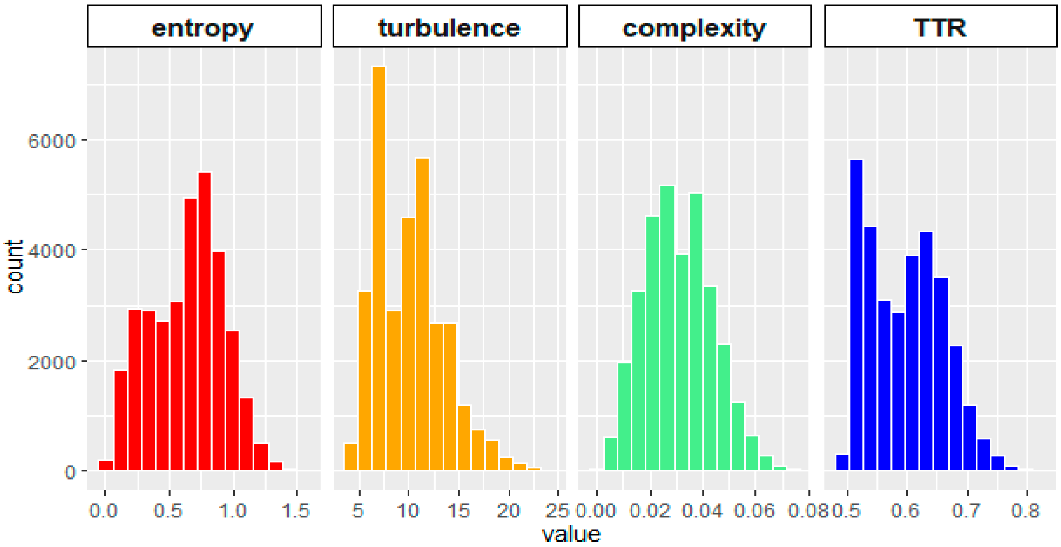


Figure 4. Fragmentation indicators. At the top, bivariate plots for all trip makers, and at the bottom, histograms excluding multivariate outliers (99% confidence).

The entropy indicator has a maximum when all possible activities appear in a sequence, and the total duration for each activity in the alphabet is the same. In a sequence where staying home takes all the daily minutes, entropy is 0 (minimum), and the maximum for an 11-activity alphabet is 2.40. Fragmentation of daily time into many episodes weighted by duration is accounted by turbulence; i.e., many episodes with a small period mean much stress to cover all duties. In contrast, complexity considers the number of transitions between episodes and the number of different activities represented, and their total duration. Stress in daily life is captured mainly by turbulence since the transition between activities uses at least one transport mode. In comparison, complexity combines entropy and turbulence characteristics.

To understand the meaning of these fragmentation indicators, we select a subset of trip makers belonging to the 1%, 50%, and 99% percentile for turbulence indicators. Results are presented in **Error! Reference source not found.**.

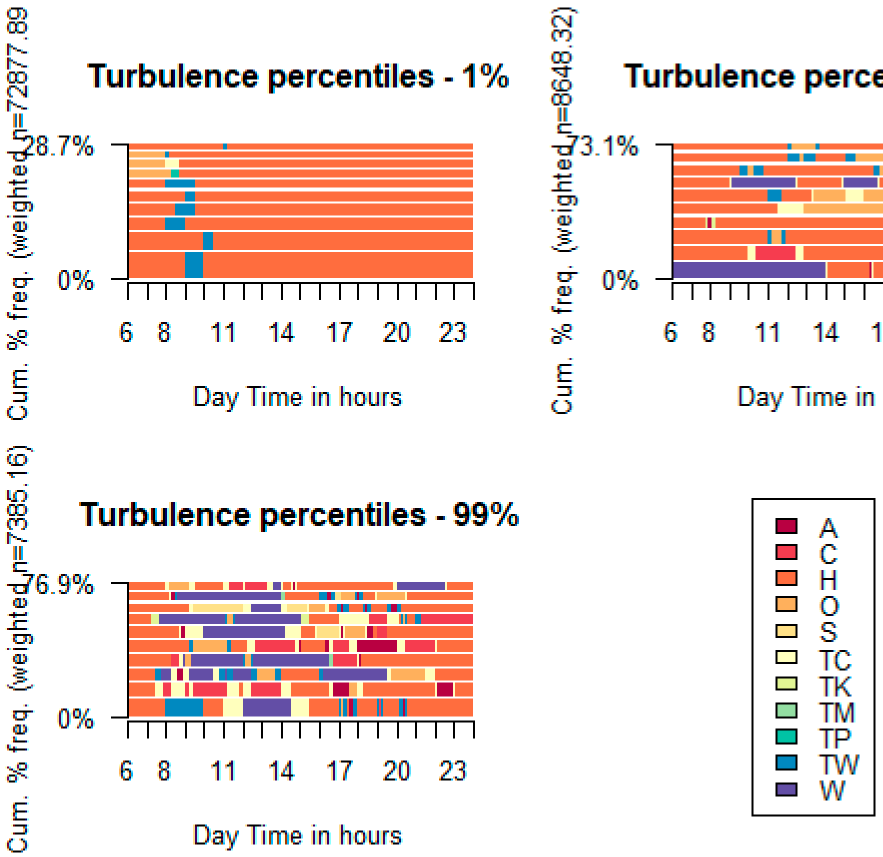


Figure 5. Some activity sequences are in the 1%, 50%, and 99% turbulence percentile groups. In the Figure, the Trip-makers subset and multivariate outliers are excluded.

The radar plot in **Error! Reference source not found.** shows the four fragmentation indicators, total trips, and total travel time in a day relative to the means over the years. The Covid-19 effects can be seen in 2020 fragmentation variables: they got the smallest values, while 2018-2019 show the greatest ones; while 2021 recovery was not as expected, revealing that some behavioral changes that could be latent before Covid-19, and potentiated by the pandemics, seem to remain after it. This will be clarified once data for 2022 and future years can be analyzed.

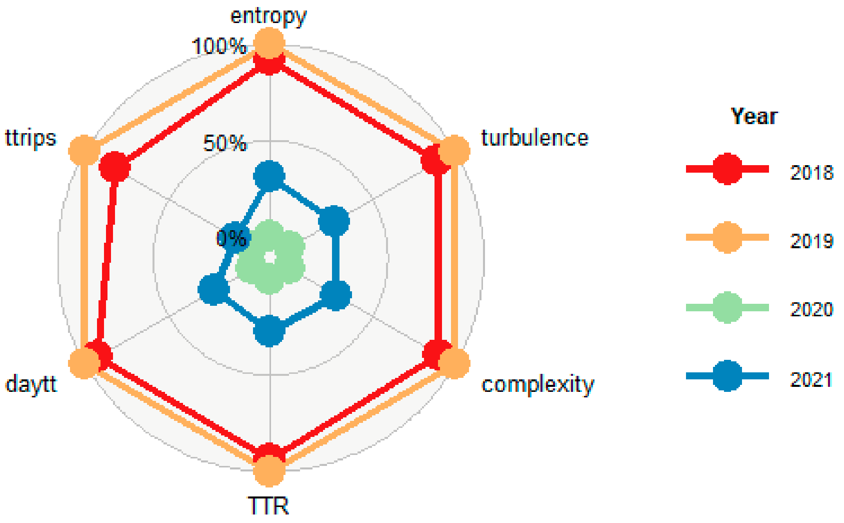


Figure 6. Fragmentation indicators according to EMEF travel survey year. Ttrips and daytt mean the total number of trips and total time traveling in a day, respectively.

4.2. Modal frequency and residential area

Travel surveys offer multiple possibilities for multivariate exploratory analysis, and this section illustrates how declared modal use frequency (generic modal preferences) is connected to some individual characteristics such as residential place. These findings open a new line of research where spatial representation is critical. The first factorial plane of multiple correspondence analysis (see **Error! Reference source not found.**) applied to individual modal preferences and residential areas shows an increasing frequency of use of public transport and rarely frequency of car use for Barcelona city residents (BCN). In contrast, frequent car use preferences and non-active modes are located in the outer metropolitan region (RMB). The vertical axis has a spatial meaning (negative to positive values as external to an inner metropolitan area), and the horizontal axis has a temporal meaning (positive values belong to 2020-21 years, clearly separated from 2018-19 negative values).

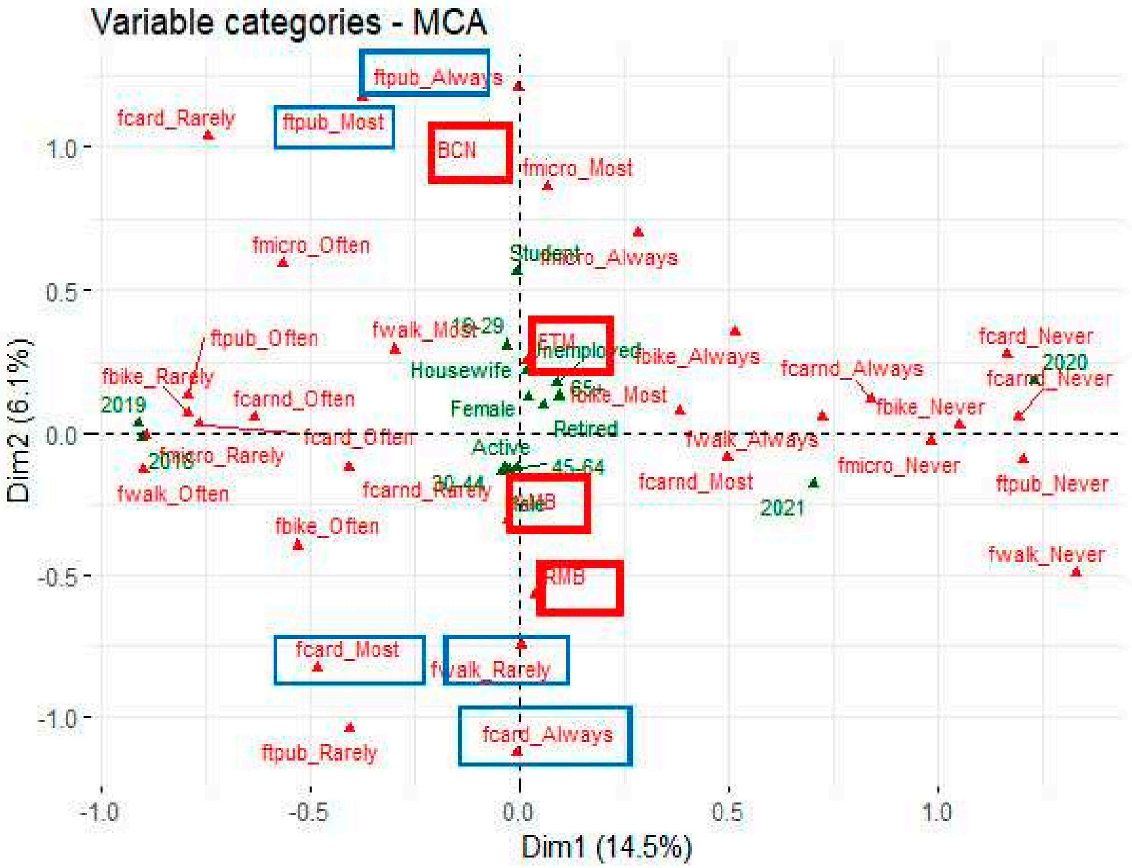


Figure 7. First Factorial Plane: modal preferences and residential areas across years. Residential areas are highlighted in the red box and modal preferences in the blue box. Modal preferences legend: fcard stands for car as a driver, fcarnd car as a non-driver, fwalk-walking, ftpub-public transport, fbike-bike, fmicro –eScooter/Segway. Residential areas: BCN – Barcelona city, EMT-Primary Crown, AMB-Metropolitan Area of Barcelona, and RMB – Metropolitan Region of Barcelona. Year, gender, age group and activity factors are shown in green.

4.3. Linear models for fragmentation indicators

One analysis line that addresses variables affecting a fragmentation indicator is linear modeling. For example, in the case of the complexity indicator, linear model results show a significant dependency (marginal effect once all the rest of significant variables are included) on year, gender by age-group interaction, activity, residential area and education.

Error! Reference source not found. shows the marginal effects of gender by age-groups across years on complexity indicator. Women show significantly greater complexity for the 30-44 age-group than men (all years except 2020), while the opposite is clearly seen for women's '65+' age-group (any year).

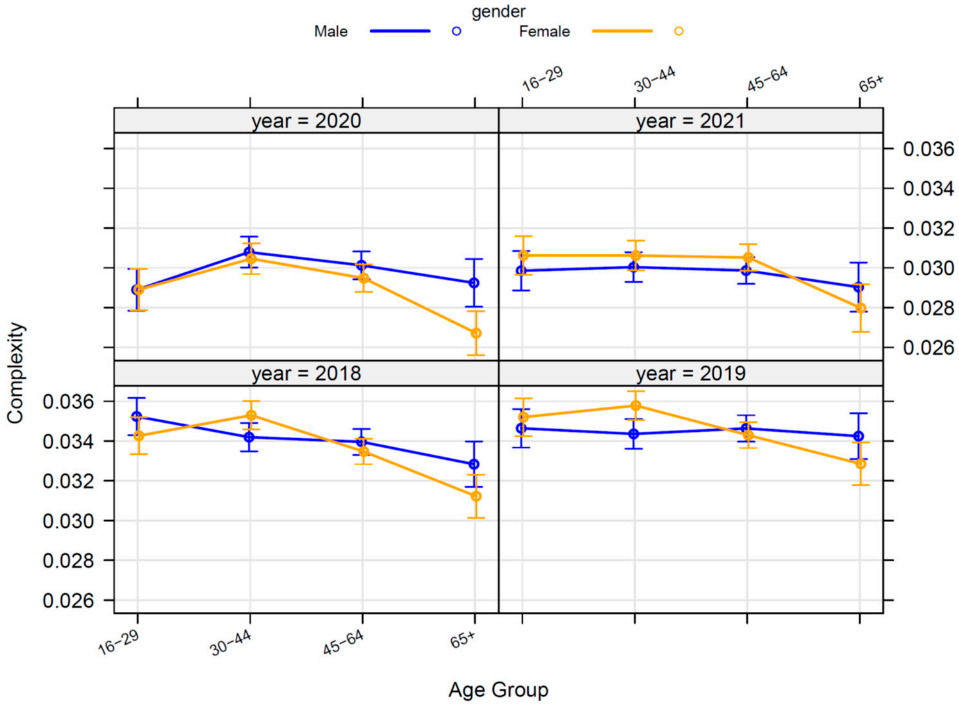


Figure 8. Marginal effects of age-group by gender across years on complexity indicator.

A second analysis involving the travel time ratio (TTR) indicator as the target variable in a linear model shows a pattern that compares 2018-19 data against 2020-21, where TTR in 2021 does not seem to recover to values obtained before Covid-19 spread. Education is an important factor, higher education profiles spent a greater part of their daily time out of home before Covid-19 spread than the rest of education levels (none, primary and secondary), in contrast to 2021 data (see **Error! Reference source not found.**). Higher educated people increase their time at home during 2020 and 2021, teleworking has a remarkable impact in this group after Covid-19, differences across residential areas have been minimized according to 2021 data. The higher educated group increased home-stay time from 2019 to 2020 in a more remarkable way than the rest of education groups where teleworking was not an option.

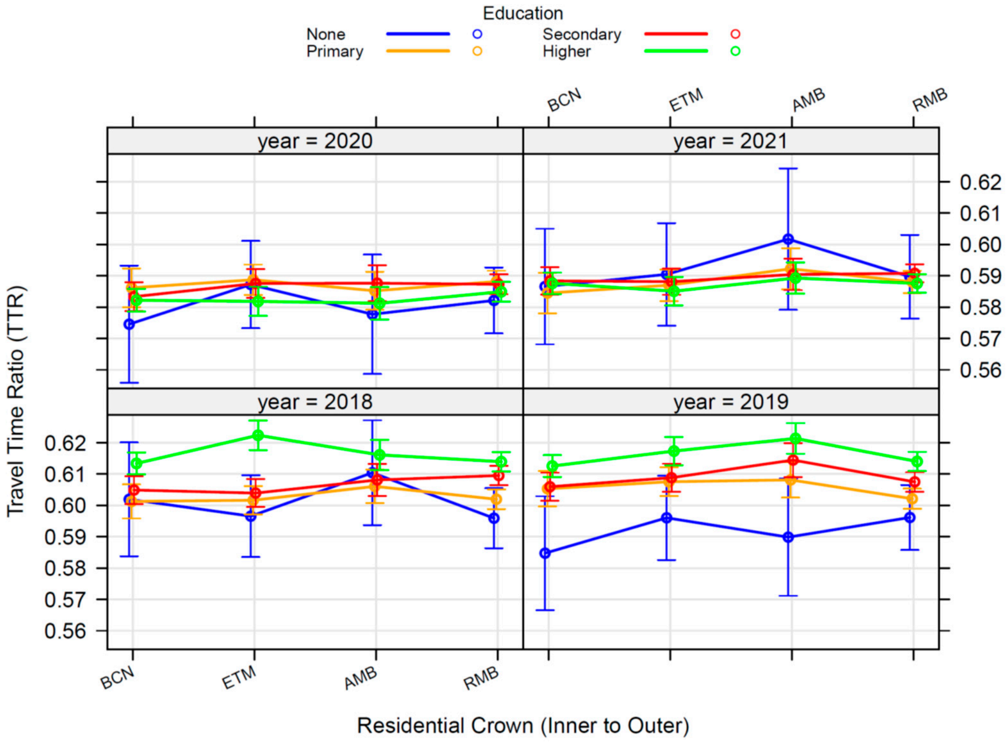


Figure 9. Marginal effects of residential area by education across years on TTR.

4.4. Principal mode

As suggested in the Methodological Section 3, one step in data processing relies on defining the day principal mode (*dpmode*). We elaborate this variable according to the frequencies of transportation modes recorded in the transitions between activities in the sample (2018 to 2021) and unsupervised classification. The first factorial axis separates private transport (negative values) and public transport (positive values) day principal mode. In contrast, the second factorial axis splits the sample into more pedestrians (negative values) and fewer pedestrians (positive values) (see **Figure 10**).

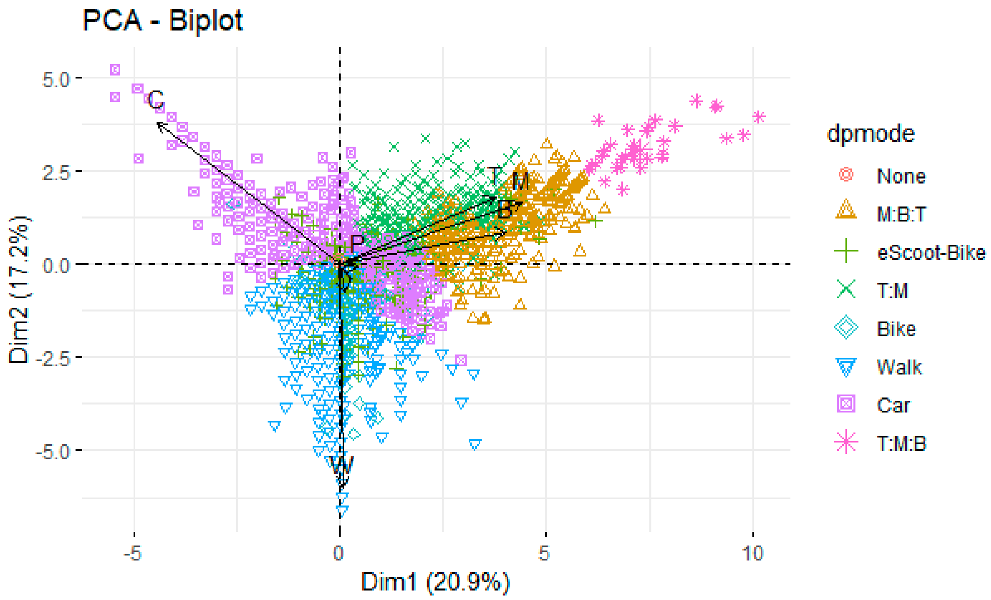


Figure 10. Day principal mode clusters based on the number of activity occurrences in daily trips.

4.5. Clustering

The five most significant clusters exhibit distinct distributions of entropy, turbulence, complexity, and travel time ratio, highlighting the divergent behaviors of individuals within each cluster regarding these indicators (see **Error! Reference source not found.**). These clusters account for 37% of the sample variability, with a median size of 145 units, 90% comprising less than 800 units, and 90% of the clusters containing more than ten units. The greatest cluster consists of non-trip makers (4,190 units), with 0 values for entropy, turbulence and complexity, and TTR is 0.5 (this cluster does not appear in the sequence state analysis shown in **Error! Reference source not found.**). The hierarchical clustering method was employed after dimensionality reduction through multiple correspondence analysis factorial projections of sequences according to standard multivariate data reduction techniques, as explained in Section 4.2 since it mitigates biases resulting from information loss.

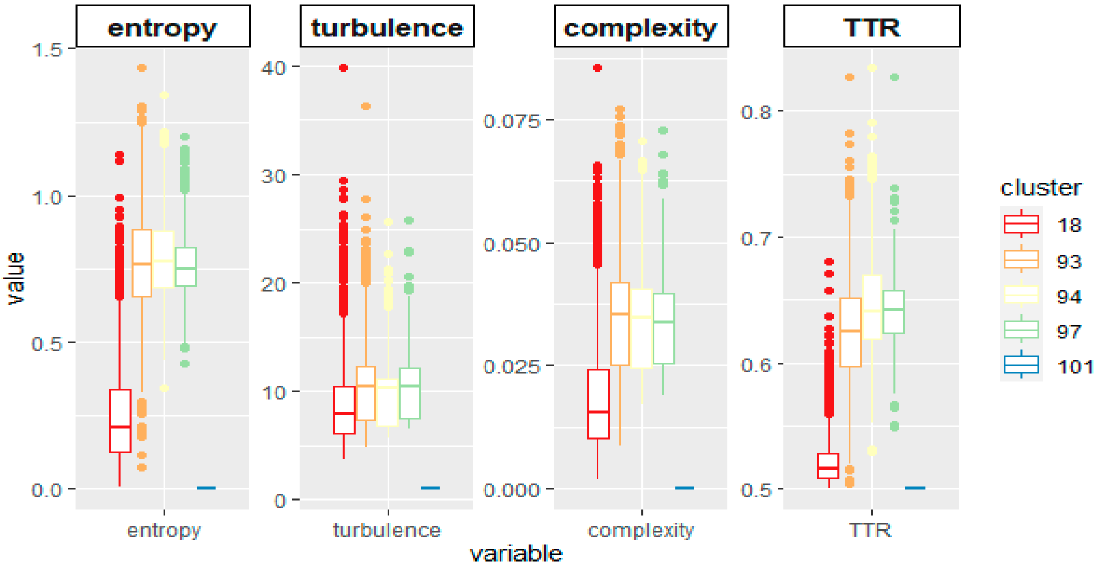


Figure 11. Fragmentation indicators in the five greatest clusters after activity sequence classification.

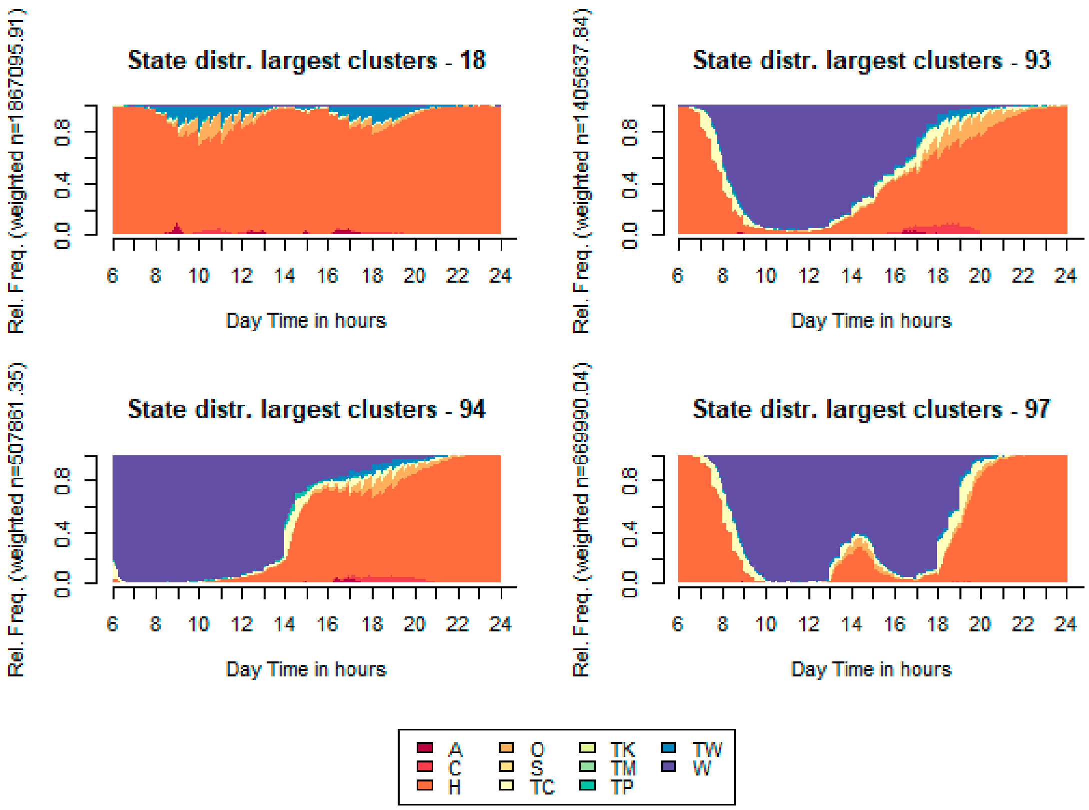


Figure 12. State distribution for the four greatest clusters after activity sequence classification.

Clusters 93, 94, and 97 show people mainly involved in working activity; nevertheless, patterns are not the same. Cluster 94 includes people whose shifts start very early in the morning, and after-work activities, while people in clusters 93 and 97 start their working activity later in the morning, and 97 includes units that break the shift to lunch at home, not in the working place as cluster 93 units do. The profiling of these clusters reveals additional characteristics of these units.

Although we do not include profiling details in this paper, a summary list including these findings is:

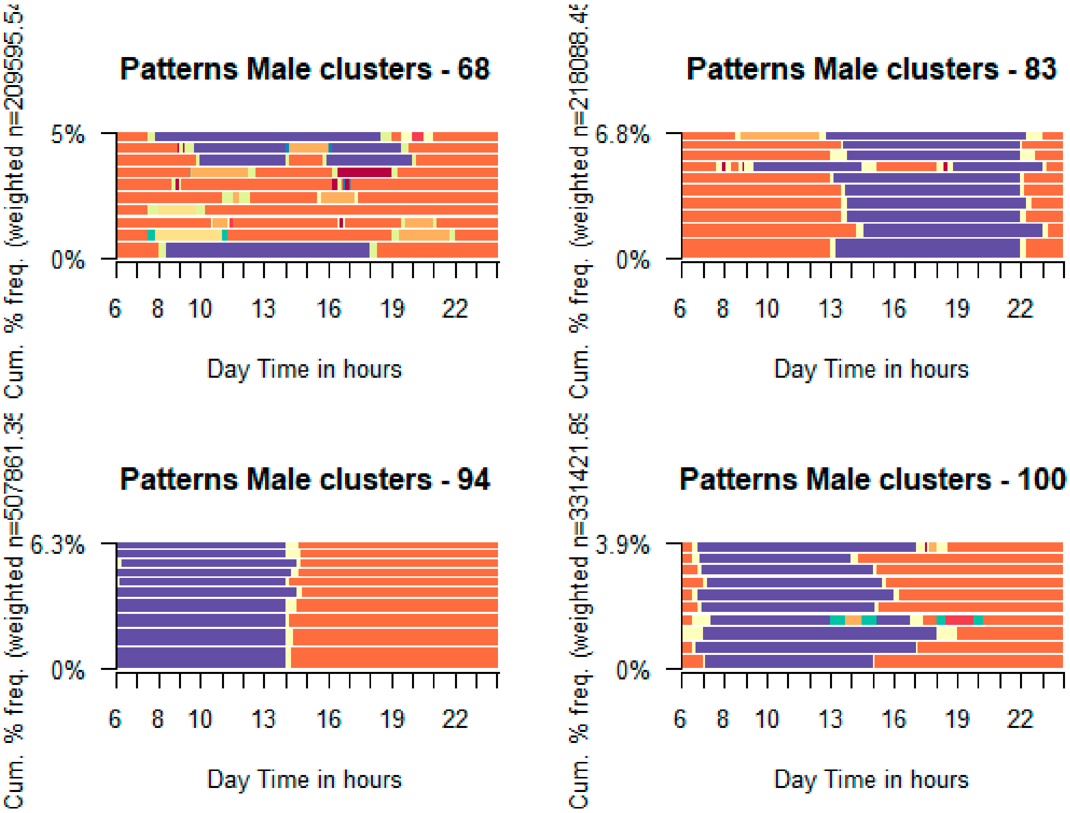
- Cluster 18. Retired, over 65 years, primary education or handicapped, origin rest of Spain.
- Cluster 93. High education, professionally active, origin Catalonia in 30-44 age group, private car use 13 points over the overall mean.
- Cluster 94. Primary or Secondary education, professionally active, origin foreigner in the 30-44 or 45-64 age groups, private transport use 15 points over the overall mean.
- Cluster 97. High education, professionally active, origin Catalonia in 30-44 or 45-64 age groups, private car use 26 points over the overall mean.

An analysis in greater depth of clusters overrepresented by males and those by females shows activity sequences with remarkable differences (see **Error! Reference source not found.**). In the clusters containing more than 65% of males, early morning and afternoon shifts (clusters 94 and 83) and extended shifts (cluster 100) with occasional escorting activities are seen. Nevertheless, in the clusters grouping female units, escorting is persistent, especially in clusters 48 and 69 in the after-school period (usually after her working activity); cluster 90 shows a public transport commuter mode for arriving at the working place. Profiling details are not included in this paper. However, the main findings can be summarized as follows:

- Cluster 68. E-scooter users, unemployed or student, and Barcelona city residents.
- Cluster 83. 16-29 age group, secondary education, active, car-depending users being RMB residents.
- Cluster 94. Primary education, professionally active, origin Catalonia engaged in non-flexible job schedules and public transport use, mostly Primary Crown or AMB residents.

doi:10.20944/preprints202308.0949.v1

- Cluster 100. High education, professionally active, flexible work schedule, and private car use 26 points over the overall mean being RMB residents.
- Cluster 34. Retired over 65 age group, or unemployed young people or students living in Barcelona city.
- Cluster 48. Primary education, unemployed, mostly escorting activity using car in RMB area.
- Cluster 69. 30-44 age group, homemakers, mostly escorting activity using car and resident in RMB or AMB areas.
- Cluster 90. Higher education, non-flexible work schedule, public transport users, and residents of Barcelona city. Foreign origin is overrepresented.



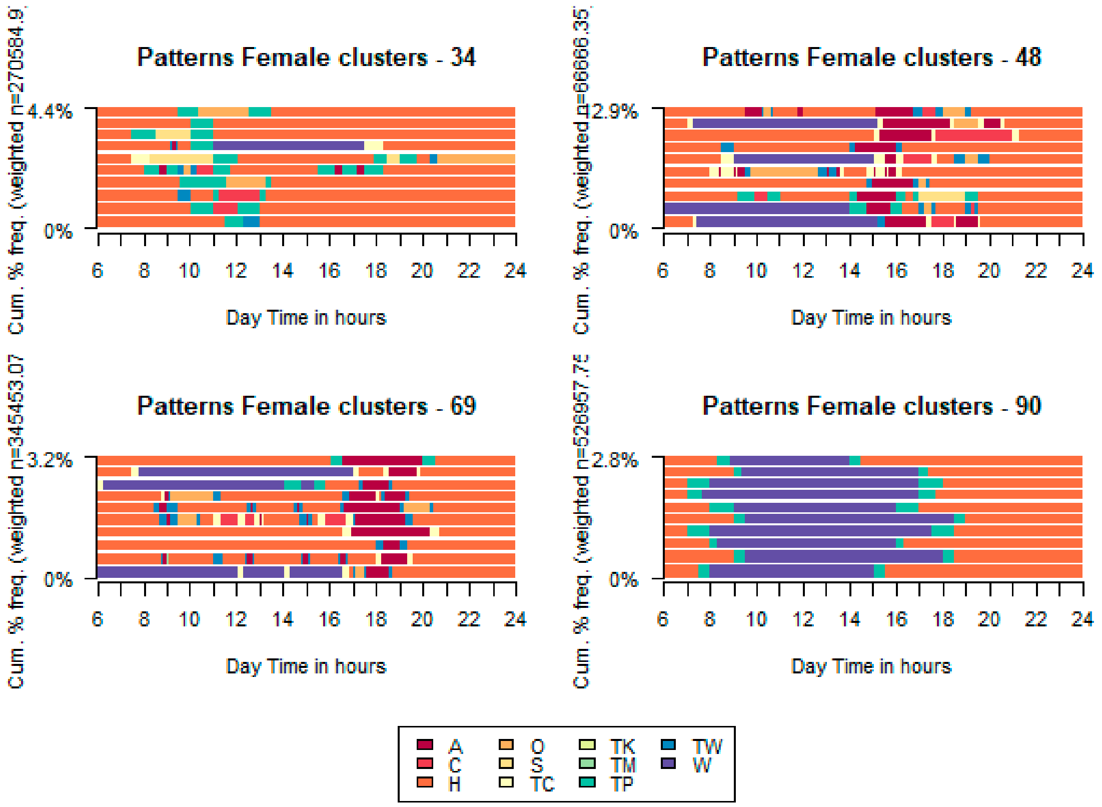


Figure 13. Daily sequence activity according to clusters overrepresented by Males and Females.

Fragmentation indicators are helpful to complement the former interpretation (**Error! Reference source not found.**). We see mean fragmentation indicators on radar plots for some selected clusters, either male or female over-represented. Cluster 68 contains male patterns involving considerable turbulence and complexity indicators, no remarkable entropy, and TTR values, in contrast to male clusters 83, 94 and 100. In the case of female clusters, cluster 34 shows shallow fragmentation indicators compared to clusters 48, 69 and 90.

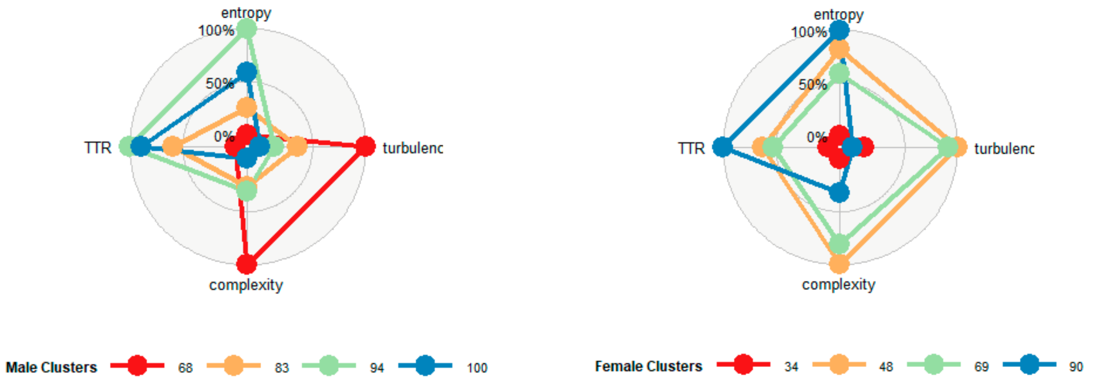


Figure 14. Fragmentation indicators for the 8 clusters over-represented by male and female population.

5. Discussion and Conclusions

Sequence analysis is a statistical method used to analyze and interpret patterns in sequential data. When applied to travel surveys and daily travel behavior, sequence analysis helps researchers understand the sequential order and dynamics of activities undertaken during travel and their interconnections. It allows for a detailed examination of the sequences of activities individuals engage in, such as commuting, work, leisure, and other daily routines.

We applied sequence analysis to EMEF travel surveys and daily travel behavior according to the following steps:

- Data Preprocessing: Each individual's sequence of activities becomes a series of ordered events. Entropy, turbulence, complexity and TTR (travel time ratio) indicators have been elaborated using TraMineR method in RStudio. Regarding fragmentation variables, 1190 units out of 37877 are multivariate outliers (3%), they have not been discharted, but used as supplementary observations when applying data analytics methods.
- Sequence Mining: Data analytics algorithms are applied to identify the profiles of fragmentation indicators within the EMEF dataset. Data reduction based on MCA allows to project activity sequences defined at minute level into a multivariate real space reducing the computational burden. Euclidean distances are applied to assess the similarity between projected sequences.
- Sequence Comparison: Based on fragmentation indicators as target variables, linear models highlight variations in travel behavior based on demographic characteristics, such as age, gender, or socioeconomic status.
- Clustering and Typology Development: clustering on projected activity sequences identifies distinct segments or clusters of individuals based on their travel behavior patterns. We have obtained 10% of the clusters being greater than 800 sample units.
- After clustering individuals with similar projected sequences, we have developed typologies or travel behavior profiles focussing on clusters over and under represented by males and females. All activity sequences have been considered in the clustering process leading to many small clusters grouping multivariate outliers. We have also paid attention to the four largest clusters.
- Large cluster typologies can inform transportation planning and stakeholders about policy-making and allowing to focus into targeted interventions by segments. In the case study, 10 clusters group more than 50% of activity sequences. Immobility affects 11% of the population.
- Modal frequency use and residential area have a remarkable association that will be addressed in future research. Built environment also seems to play a critical role.

Characterization of activity sequences will be refined as soon as household composition is available and new yearly travel surveys are processed (from 2021). Our agreement with ATM will give us access to 2022 data when they become available, hopefully before 2024. Then the forthcoming work will check whether the conjecture about the behavioral changes is correct.

Author Contributions: Conceptualization, Lídia Montero, Lucía Mejía-Dorantes and Jaume Barceló; Formal analysis, Lídia Montero, Lucía Mejía-Dorantes and Jaume Barceló; Funding acquisition, Lídia Montero and Jaume Barceló; Methodology, Lídia Montero, Lucía Mejía-Dorantes and Jaume Barceló; Software, Lídia Montero; Supervision, Jaume Barceló; Writing – original draft, Lídia Montero; Writing – review & editing, Lucía Mejía-Dorantes and Jaume Barceló.

Funding: This research was funded by PID2020-112967GB-C31 Spanish R+D Programs and by Secretaria d'Universitats-i-Recerca-Generalitat de Catalunya- 2021 SGR 01252 Information Modeling and Processing.

Data Availability Statement: Restrictions apply to the availability of EMEF data. Data was obtained from Autoritat del Transport Metropolità (ATM) and can not be distributed without the permission of ATM.

Acknowledgments: EMEF datasets were kindly shared by the Autoritat del Transport Metropolità (ATM). Their contribution to our research is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rodrigue, J.P.; Comtois, C.; Slack, B. *The geography of transport systems*; Routledge, Ed.; Taylor and Francis: London, UK, 2016; ISBN 9781317210108.
2. Mejía-Dorantes, L.; Montero, L.; Barceló, J. Mobility Trends before and after the Pandemic Outbreak: Analyzing the Metropolitan Area of Barcelona through the Lens of Equality and Sustainability. *Sustain.* **2021**, *13*, 7908, doi:10.3390/SU13147908.

3. Lyons, G.; Mokhtarian, P.; Dijst, M.; Böcker, L. The dynamics of urban metabolism in the face of digitalization and changing lifestyles: Understanding and influencing our cities. *Resour. Conserv. Recycl.* **2018**, *132*, 246–257, doi:10.1016/j.resconrec.2017.07.032.
4. Mejía-Dorantes, L.; Soto Villagrán, P. A review on the influence of barriers on gender equality to access the city: A synthesis approach of Mexico City and its Metropolitan Area. *Cities* **2020**, *96*, doi:10.1016/j.cities.2019.102439.
5. McBride, E.; Davis, A.; Goulias, K. Fragmentation in Daily Schedule of Activities using Activity Sequences: *Transp. Researc Rec.* **2019**, *2673*, 844–854, doi:10.1177/0361198119837501.
6. McBride, E.C.; Davis, A.W.; Goulias, K.G. Exploration of Statewide Fragmentation of Activity and Travel and a Taxonomy of Daily Time Use Patterns using Sequence Analysis in California: *Transp. Res. Rec.* **2020**, *2674*, 38–51, doi:10.1177/0361198120946011.
7. Nobis, C.; Lenz, B. Gender Differences in Travel Patterns: Role of Employment Status and Household Structure. In Proceedings of the Research on Women's Issues in Transportation; TRB Publications Office 1073-1652, 2004; pp. 114–123.
8. Baratian-Ghorghi, F.; Zhou, H. Investigating Women's and Men's Propensity to Use Traffic Information in a Developing Country. *Transp. Dev. Econ.* **2015**, *1*, 11–19, doi:10.1007/S40890-015-0002-5.
9. Systematics, C. Travel Survey Manual. **1996**, doi:10.21949/1404543.
10. Stopher, P.R. Use of an activity-based diary to collect household travel data. *Transportation (Amst).* **1992**, *19*, 159–176.
11. Goldenberg, L.; Stecher, C.; Červenka, K. CHOOSING A HOUSEHOLD-BASED SURVEY METHOD: RESULTS OF THE DALLAS-FORT WORTH PRETEST PRESENTATION ABSTRACT. **1995**.
12. Stopher, P.R.; Greaves, S.P. Household travel surveys: Where are we going? *Transp. Res. A* **2007**, *41*, 367–381.
13. Montero, L.; Mejía-Dorantes, L.; Barceló, J. The role of life course and gender in mobility patterns: A spatiotemporal sequence analysis in Barcelona (In review). *Eur. Transp. Res. Rev.* **2023**, *In Review*.
14. Abbott, A. Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Hist. Methods A J. Quant. Interdiscip. Hist.* **1983**, *16*, 129–147, doi:10.1080/01615440.1983.10594107.
15. Abbott, A.; Forrest, J. Optimal Matching Methods for Historical Sequences. *J. Interdiscip. Hist.* **1986**, *16*, 471, doi:10.2307/204500.
16. Abbott, A.; Tsay, A. Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociol. Methods Res.* **2000**, *29*, 3–33, doi:10.1177/0049124100029001001.
17. Abbott, A.; DeViney, S. The Welfare State as Transnational Event: Evidence from Sequences of Policy Adoption. *Soc. Sci. Hist.* **1992**, *16*, 245, doi:10.2307/1171289.
18. Leszczyc, P.T.L.P.; Timmermans, H. Unconditional and conditional competing risk models of activity duration and activity sequencing decisions: An empirical comparison. *J. Geogr. Syst.* **2002**, *4*, 157–170, doi:10.1007/s101090200083.
19. Studer, M.; Ritschard, G. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures on JSTOR. *J. R. Stat. Soc. A* **2016**, *179*, 481–511.
20. Gabadinho, A.; Ritschard, G.; Müller, N.S.; Studer, M. Analyzing and Visualizing State Sequences in R with TraMineR. *J. Stat. Softw.* **2011**, *40*, 1–37, doi:10.18637/JSS.V040.I04.
21. Elzinga, C.H.; Liefbroer, A.C. De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *Eur. J. Popul.* **2007**, *23*, 225–250, doi:10.1007/S10680-007-9133-7/FIGURES/2.
22. Ritschard, G. Measuring the Nature of Individual Sequences. *Sociol. Methods Res.* **2021**, 1–34, doi:10.1177/00491241211036156.
23. OMC Working day mobility survey (EMEF) of the ATM of the Barcelona area Available online: https://omc.cat/en/w/surveys-emef.
24. RStudio-Team RStudio: Integrated Development for R. 2020.
25. Husson, F.; Lê, S.; Pages, J. *Exploratory multivariate analysis by example using R Analysis*; Chapman &.; 2010; ISBN 9781138196346.
26. Mahalanobis, P.C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49–55.
27. Husson, F.; Josse, J.; Lê, S.; Mazet, J. FactoMineR: Exploratory Multivariate Data Analysis with R 2008.
28. R Development Core Team R: The R Project for Statistical Computing 2021.

29. Kaiser, H.F. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* **1960**, *20*, 141–151.
30. Tukey, J.W. (John W.; Tukey, J.W. (John W.; Brillinger, D.R.; Cox, D.R. (David R.; Braun, H.I. *The collected works of John W. Tukey*; Wadsworth Advanced Books & Software: Belmont Calif., 1984; ISBN 9780534033033.
31. Gabadinho, A.; Ritschard, G.; Studer, M.; Müller, N.. Mining sequence data in R with the TraMineR package: A user s guide 2011, 129.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.