Guidotti et al. *EPJ Data Science* (2018) 7:6
https://doi.org/10.1140/epjds/s13688-018-0133-0

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Discovering temporal regularities in retail customers' shopping behavior

Riccardo Guidotti[1*], Lorenzo Gabrielli[1], Anna Monreale[2], Dino Pedreschi[2] and Fosca Giannotti[1]

*Correspondence:
riccardo.guidotti@isti.cnr.it
[1]KDDLab, ISTI-CNR, Pisa, Italy
Full list of author information is available at the end of the article

Abstract

In this paper we investigate the regularities characterizing the temporal purchasing behavior of the customers of a retail market chain. Most of the literature studying purchasing behavior focuses on what customers buy while giving few importance to the temporal dimension. As a consequence, the state of the art does not allow capturing which are the temporal purchasing patterns of each customers. These patterns should describe the customer's temporal habits highlighting *when* she typically makes a purchase in correlation with information about the amount of expenditure, number of purchased items and other similar aggregates. This knowledge could be exploited for different scopes: set temporal discounts for making the purchases of customers more regular with respect the time, set personalized discounts in the day and time window preferred by the customer, provide recommendations for shopping time schedule, etc. To this aim, we introduce a framework for extracting from personal retail data a *temporal purchasing profile* able to summarize *whether* and *when* a customer makes her *distinctive* purchases. The individual profile describes a set of regular and characterizing shopping behavioral patterns, and the sequences in which these patterns take place. We show how to compare different customers by providing a *collective perspective* to their individual profiles, and how to group the customers with respect to these comparable profiles. By analyzing real datasets containing millions of shopping sessions we found that there is a limited number of patterns summarizing the temporal purchasing behavior of all the customers, and that they are sequentially followed in a finite number of ways. Moreover, we recognized *regular* customers characterized by a small number of temporal purchasing behaviors, and *changing* customers characterized by various types of temporal purchasing behaviors. Finally, we discuss on how the profiles can be exploited both by customers to enable personalized services, and by the retail market chain for providing tailored discounts based on temporal purchasing regularity.

Keywords: Temporal shopping profile; Individual collective analysis; Customers segmentation; Retail customers shopping behavior analysis; Collective perspective

1 Introduction

The availability of huge amount of retail data stimulates challenging questions that can be answered only by a deep and accurate analysis of different aspects related to customers' shopping sessions. Retail data is a complex type of data containing various dimensions: *what* customers buy, *when* and *where* they make the purchases and which is the *relevance* of the purchase in terms of money spent or number of items purchased. The choice of

analyzing a set of dimensions rather than another one depends on the kind of phenomena to be investigated. Considering all the dimensions can lead to very complex models or to weak generalizations. The most important dimension for understanding how customers schedule their shopping time is obviously the temporal one. Most of the works in the literature focus on *what* customers buy [1, 2]. Just a few of them have exploited also the temporal dimension as a feature for enriching their models based primarily on the items purchased [3–6]. However, it has been given very few attention to the temporal dimension of shopping sessions considered on its own in order to extract a customer model which helps in understanding the purchase temporal habits. As a consequence, by using methods proposed in the literature it is not possible to capture the temporal purchasing patterns of each customers, which put in correlation their temporal habits with other information such as the amount of expenditure and number of purchased items. This knowledge about the customers is important because enables different marketing strategies tailored to the temporal and systematic behavior of each customer, and new innovative services for the customer based on recommendations for shopping time schedule and for increasing her awareness. To the best of our knowledge, there is not previous work focusing only on the temporal dimension and using it as the main building block to construct an individual temporal purchasing profile. We do not claim that ignoring the items purchased and/or the shopping location may lead to an advantage, but we show that observing only the temporal dimension is crucial to completely understand the different times and expenses adopted by the customers when they go to shopping.

In this paper, we are interested in understanding *whether* and *when* a customer *typically* makes retail purchases. Which of these temporal aspects of the shopping behavior are more systematic? Which are the *regular* sequences of the temporal patterns? To this aim, we propose a *temporal purchasing profile* able to describe the regular and characteristic temporal behaviors of an individual customer. Indeed, the *individual person* is the key element that lies in between a single purchase and a whole customers population. Each individual has her own regularities and habits outlining her behavior and making her a unique part of the mass. The analysis of individuals provides the basis for understanding the common regular patterns in the purchasing behavior also at *collective* level. Thus, our aim is to define individual and collective temporal profiles which can be employed for the analysis of the temporal dimension of the customers' shopping sessions. In particular, our models enable a customer segmentation which considers the temporal components of purchases and permit to perform explorative analyses of individuals under a new point of view. We conduct different case studies by using the defined methodological framework with the aim of discovering customers temporal purchasing patterns and grouping the customers' profiles to identify sets of customers with similar temporal behavior.

The main contributions of this work are the following: *(i)* we define the *temporal purchasing profile* as the set of temporal *footprints* and *sequence of footprints* summarizing *whether* and *when* a customer typically purchases and we provide the method for extracting these profiles; *(ii)* we define the *collective perspective* for making comparable the individual and not-comparable profiles, so that the shopping routines shared by different customers can be analyzed; *(iii)* we show the application of the whole analytic framework on a set of case studies a real datasets, one of them containing 7 years of retail data for $91k$ customers; *(iv)* we observe how the individual profiles and the collective perspective allow to separate the customers into well defined groups.

Our methodological framework empowers the discovery of various customers segmentation with respect to different temporal aspects. In particular, our study reveals three main typical collective behaviors characterizing the whole collection of customers on the basis of when they go to the shopping center: *daily* spending behavior capturing purchases made every day; *one-shop* spending behavior, characterizing a regularity with a week containing a predominant shopping session; and an *occasional* spending behavior, describing a not habitual shopping sessions related to a very small expenditure amount. Among *one-shop* spending behaviors the analysis captures a further classification in with respect to the expenditure amount: *normal* spending behavior less than €50, *high* spending behavior with a typical expenditure between €50 and €100, and *higher* spending behavior with an expenditure higher than €100. By analyzing the number of different purchasing behaviors at individual and collective level we identify two categories of customers that we name *regular* and *changing*: a customer represented with a high number of temporal purchasing behavioral patterns is classified as *changing*, while a customer with a small number of temporal purchasing behaviors is classified as *regular*. Finally, we found various and diversified regular sequences explaining how the customers typically combine and follow their shopping behavior with respect to the temporal point of view.

The rest of the paper is organized as follows. In Sect. 2 we discuss the related literature. The individual profile is defined in Sect. 3, while in Sect. 4 we describe the approach to provide the collective perspective. In Sect. 5 we illustrate the case study showing the effectiveness for data analytics of the proposed models and methods. Section 6 discusses about possible applications of the proposed approach. Finally, Sect. 7 reports conclusion and future research directions.

2 Related work

While most of the works in the literature analyze the customer behavior by considering only aspects related to the items bought and without a specific focus on the temporal dimension, the analysis of the customer purchasing behavior presented in this paper uses the temporal dimension as a fundamental building block. In the literature there are various works analyzing and trying to predict the customer behavior. To this aim, they generally take into consideration aspects related to the items bought and not the time: which items the customers frequently buy, which items they pay attention within a store, the changes of customer behavior on the bases of what they buy, etc. To the best of our knowledge no paper analyzes only the temporal dimension of customers shopping sessions to understand *whether* and *when* customers typically make *characterizing* and *regular* purchases.

Data mining is the classic approach used to analyze purchasing behaviors [1, 2]. However, it is generally hard to create a comprehensive model of overall customer behavior since each individual acts according to a personal utility function depending on various factors, that can be described by different types of user data. Therefore, multiplex approaches are used to understand the customer behavior integrating multiple data sources and multiple types of data [7] to reach the combined prediction results. In [8] the authors propose to represent the customer purchasing behavior using a directed graph retaining temporal information in a purchase sequence. In particular, given a target product, they build the graph for that product attaching as nodes the purchased products in subsequent shopping sessions, putting on the edges the days passed between two consecutive purchases and concatenating products belonging to the same category (e.g. two different types of milk). Then, they apply a graph mining technique on such networks to extract

and analyze the occurring frequent patterns. In [9] it is shown how signals of RFID (Radio-Frequency IDentification) can be exploited to detect and record how customers browse stores, which items of clothes they pay attention to, and which items of clothes they usually match with. In [10] the authors propose a mixture of non-homogeneous Poisson processes to discover the latent customer groups and conduct the soft-membership customer segmentation based on the dynamically observed purchase behavior. Yet, they exploit the temporal dimension to explain the purchasing patterns of the products at a global scale and their outcomes do not explain the customers behavior with respect to time. The temporal dimension of purchasing habits is exploited in [11] to understand how predictable are consumers in their merchant visitation patterns by using a Markov model for predicting the customer's next shop location. Note that in these works the customer behavior is generalized at global level, while our model is personal and describes the customer habits and preferences in a concise way.

Other recent studies analyze purchasing data to understand changing in the customer behavior [12, 13] and if a customer will switch from one brand to another one [14]. In particular, in [12] the authors integrate variables such as recency, frequency, and monetary/demographic variables to establish a method for mining changes in shopping behavior. In [14] it is developed a method for extracting useful knowledge from individual purchase history of customers by combining information fusion techniques with data mining to predict whether a customer switches from one brand to another, or becomes loyal to a brand, and when a customer is likely to defect to a competitor. The models and methodology we propose not only help in understanding the changing in customers' temporal shopping behaviors and their cyclic succession, but also unveil the regularities of these changing if detected.

The temporal component of customer purchases is analyzed in [3–5]. In [4] it is examined the role of personal characteristics in time spent shopping. In particular, is analyzed the roles of time perceptions, brand and store loyalty, social, physical well-being, and demographic variables in predicting reported shopping time, including the hours spent at search and purchase. In [5] are studied changes in cluster characteristics of supermarket customers over a 24 week period by performing a temporal analysis that tries to detect the migrations of the customers from one group to another group. The temporal analysis presented is based on conventional and modified self organizing maps. In [3, 6] the authors use a sophisticated version of entropy to study the customers' behaviors in retail data both from the basket and the spatio-temporal point of view. In particular, they define a procedure to group similar baskets by exploiting a frequent pattern mining approach and creating in this way classes of probability for certain sets of items frequently purchased together. Their discovery is that *predictable* customers are also the more *profitable* ones. Although these papers consider the temporal dimension, we highlight that in our work the time is the focus of the models and it is not simply used as a feature for a temporal analysis. Also in [15], an analysis of the sequences of purchases exploiting Zipf-like distributions leads to the detection of five consumer groups. Customers in each group resulted to be also similar with respect to their age, gender, total expenditure, etc. Moreover, there are works aimed at understanding the behavior of customers in online shopping [16–19].

Finally, another set of works adopting the temporal dimension in shopping session is related to the task of recommending the items for the next basket. To solve this problem in the literature various methods have been adopted: collaborative filtering [20], Markov

chains [21], supervised classification algorithms [22], deep neural networks [23], and temporal frequent pattern mining algorithms [24]. However, all these methods only exploit the temporal dimension but do not provide a way to understand how the time affects the customers decisions and which are the typical temporal shopping patterns.

Our definition of temporal purchasing footprint of a customer is similar to the definition of user profile introduced in [25]. In [25] the authors extract from the call detail records of each user a profile summarizing the calls performed by the user. Their aim is to estimate the proportion of city users that can be classified as residents, commuters or visitors. We underline that the "profile" defined in [25] is just an aggregation of (a count of) the number of calls performed by the user along the various months separated between weekend and weekdays in a specific geographical area. It does not take into account any notion of behavior or regularity, it is a sort of "status" of the user. On the other hand, the profile defined in our work models a set of temporal purchasing patterns highlighting different ways of acting for the customers. It is able to explain the different behaviors adopted by the individual customers and does not report just their status.

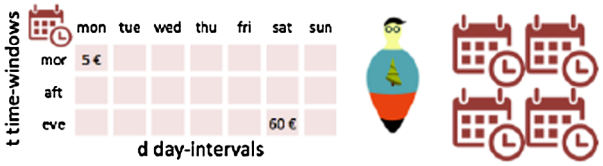
3 Individual temporal purchasing profile

The analytic process we propose for the extraction of the *temporal purchasing profile* assumes as input retail data, i.e., a collection of *shopping sessions* describing the shopping activities of a set of customers. In Fig. 1 we summarize the definitions and the notations introduced in this section. Since in this work we are not considering the shopping location or the content of the basket [3], we model a shopping session s as a tuple $s = \langle customer, timestamp, amount \rangle$ (see Fig. 1-*(a)*) For each customer, we summarize the temporal information of a set of shopping sessions by introducing the notion of *temporal purchasing unit* (unit in short) defined as follows:

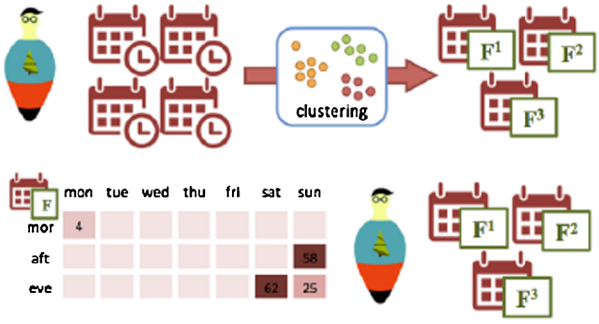
Definition 1 (Temporal purchasing unit) Given a period τ of $\bar{d}$ days, a *temporal purchasing unit* U of a customer c is a matrix $U \in \mathbb{R}^{t \times d}$, where d is the number of day-intervals in τ with $d \leq \bar{d}$, t is the number of time windows considered for each day-interval, and U_{ij} estimates the relevance of the purchases in the i-th time window of the j-th day-interval.

A temporal purchasing unit it is not a collection of shopping sessions. It is an object modeling an aggregation related to set of shopping sessions in the form of a matrix which contains the amounts spent by a customer in a certain discretized time interval with zeroes where noting was purchased. With *day-interval* we indicate any aggregation of days, e.g., single day, weekdays-weekend, etc., while with *time window* we refer to any aggregation of hours, e.g., single hour, morning-afternoon, etc. Given a period τ, each unit U captures the relevance of the customer's shopping sessions during the period τ discretized into specific temporal slots. The *relevance* can be expressed by any value related to shopping sessions: the total amount spent, the number of items bought, etc. Figure 2 illustrates two examples: on the *(left)* side each day of the week is considered separately and as relevance is used the money spent (the same happens in Fig. 1-*(b)*), while on the *(right)* side the day-intervals are weekend and weekdays and the relevance is the number of items purchased. The time granularity and the type of relevance depends on the aim of the analysis and are set when the framework is instantiated to analyze a real dataset. In our case study we adopt the amount spent as relevance for the units.

(a) A *shopping session* is a tuple $s = \langle customer, timestamp, amount \rangle$. Each customer has a sequence of shopping session $S = \langle s_1, \ldots, s_n \rangle$ containing the shopping sessions performed.



(b) A *temporal purchasing unit* $U \in \mathbb{R}^{t \times d}$ summarizes a set of shopping sessions happening in the time period modeled by the unit. Given S, for each customer we obtain the sequence of time units $\widehat{S} = \langle U^{(1)}, \ldots, U^{(m)} \rangle$.



(c) For each customer, we cluster the similar temporal purchase units obtaining a set of temporal purchasing footprints $\mathcal{F} = \{F^{(1)}, \ldots, F^{(k)}\}$. A *temporal purchasing footprint* $F \in \mathbb{R}^{t \times d}$ is a matrix, where F_{ij} contains the typical relevance of the pattern in the i-th time window of the j-th day-interval.

(d) For each customer we derive the *temporal purchasing footprint sequence* $\widehat{F}$ by replacing in $\widehat{S}$ the units with the corresponding footprints in $\mathcal{F}$.

(e) For each customer c, we define the *temporal purchasing profile* as $\mathcal{P}_c = \{\mathcal{F}, \widehat{F}\}$, where $\mathcal{F}$ is the set of footprints and $\widehat{F}$ is the corresponding footprint sequence.

Figure 1 From shopping sessions to temporal purchasing profile

Given customer c, her sequence of temporally ordered shopping sessions $S = \langle s_1, \ldots, s_n \rangle$, the time window granularity t, the day-interval granularity d, and the width of the time period τ, then S can be *segmented* into an ordered sequence of units $\widehat{S} = \langle U^{(1)}, \ldots, U^{(m)} \rangle$ with

Figure 2 Temporal purchasing units. Examples of temporal purchasing units with different day-intervals and relevance: (left) single day and money spent, (right) weekdays-weekend and number of items

$m \leq n$, where each temporal purchasing unit $U^{(i)}$ aggregates a set of shopping sessions with respect to t and d.

Our goal is to *summarize* for each customer the knowledge contained in $\widehat{S}$ in a *temporal purchasing profile* describing the customer's typical temporal behaviors. In order to define the profile we need to extract the "distinctive" temporal purchasing behaviors of customer c, i.e. her purchasing footprints.

Given the sequence of units $\widehat{S}$ of customer c, we can *detect groups* of units which are similar with respect to a distance function δ based on the concept of temporal alignment and with respect to the relevant values considered. Thus, given a group G of similar units we define a *temporal purchasing footprint* F (footprints in short) as the *representative* of the group G. Each footprint F captures a temporal behavior characterizing the customer. We define the temporal purchasing footprint as follow:

Definition 2 (Temporal purchasing footprint) Given a sequence $\widehat{S} = \langle U^{(1)}, \ldots, U^{(m)} \rangle$ and a distance function $\delta : \mathbb{R}^{t \times d} \times \mathbb{R}^{t \times d} \rightarrow \mathbb{R}$ we define the *temporal purchasing footprint groups* $\mathcal{G} = \{G_1, \ldots, G_k\}$ as a partitioning of $\widehat{S}$ into disjoint sets of similar footprints and, given a group $G = \{U^{(1)}, \ldots, U^{(q)}\}$, we name *temporal purchasing footprint* the centroid F of G. F is the matrix $F \in \mathbb{R}^{t \times d}$ such that

$$F_{ij} = \frac{1}{|G|} \sum_{\forall U^{(h)} \in G} U^{(h)}_{ij} \quad \forall i = 1, \ldots, t, j = 1, \ldots, d.$$

We refer to the *set of footprints* of a customer with $\mathcal{F} = \{F^{(1)}, \ldots, F^{(k)}\}$. Note that, for the extraction of the footprints we are not considering the order of the units in $\widehat{S}$. Figure 1-*(c)* shows an example of footprint using the days of the week as granularity for day-interval and morning-afternoon-evening as time-windows: the darker the cell the higher the relevance which corresponds to the amount in this case.

Given the groups $\mathcal{G}$ and the footprints $\mathcal{F}$ of a customer, we can replace each temporal purchasing unit in $\widehat{S}$ with the footprint representing the group to which it belongs to. We name the new sequence *temporal purchasing footprint sequence* (footprint sequence in short, see Fig. 1-*(d)*). For example, given $\mathcal{F} = \{F^{(1)}, F^{(2)}\}$, $\mathcal{G} = \{G_1, G_2\}$ where $G_1 = \{U^{(1)}, U^{(4)}\}$, $G_2 = \{U^{(2)}, U^{(3)}\}$, if $\widehat{S} = \langle U^{(1)}, U^{(2)}, U^{(3)}, U^{(4)} \rangle$, than the corresponding footprint sequence is $\widehat{F} = \langle F^{(1)}, F^{(1)}, F^{(1)}, F^{(2)} \rangle$. We define the temporal purchasing footprint sequence as:

Definition 3 (Temporal purchasing footprint sequence) Given a customer c, her sequence of units $\widehat{S}$, her groups $\mathcal{G}$ and her footprints $\mathcal{F}$, we define the *temporal purchasing footprint sequence* as the sequence $\widehat{F}$ obtained replacing in $\widehat{S}$ the units with the corresponding footprints in $\mathcal{F}$ according to $\mathcal{G}$.

Algorithm 1: *extractIndividualProfile*

Input : S—sequence of temporally ordered shopping sessions, τ—width of the time period covered by each footprint, d—days-interval to be set in a footprint, t—time windows to be set in a footprint, *rel*—relevance function, δ—distance function

Output: $\mathcal{P}$—temporal purchasing profile

1 $\widehat{S} \leftarrow segmentIntoUnits(S, \tau, d, t, rel)$
2 $\mathcal{G} \leftarrow detectGroups(\widehat{S}, \delta)$
3 $\mathcal{F} \leftarrow \{D^{(i)} | D^{(i)} = getCentroid(G_i)\ \forall G_i \in \mathcal{G}\}$
4 $\widehat{F} \leftarrow buildSequence(\widehat{S}, \mathcal{G}, \mathcal{F})$
5 **return** $\mathcal{P} = \{\mathcal{F}, \widehat{F}\}$

Finally, we define the *temporal purchasing profile* (profile, see Fig. 1-*(e)*) of a customer as:

Definition 4 (Temporal purchasing profile) Given a customer c, her sequence of units $\widehat{S}$, and a distance function δ, we define the *temporal purchasing profile* of c as $\mathcal{P}_c = \{\mathcal{F}, \widehat{F}\}$ where $\mathcal{F}$, is the set of footprints derivable from $\mathcal{G}$ detected on $\widehat{S}$ using δ, while $\widehat{F}$ is the corresponding footprint sequence,

Extracting Individual Temporal Purchasing Profile. The process for the extraction of the individual profiles is summarized by Algorithm 1. The first step is the segmentation of the sequence of temporally ordered shopping sessions S into the sequence of temporal purchase units $\widehat{S}$, given the d days-intervals, the t time-windows for each day-interval, and the relevant values returned by function *rel* (*segmentIntoUnits* function in line 1, from *(a)* to *(b)* in Fig. 1). The sequence of units $\widehat{S}$ covers a time period of width τ. Given $\widehat{S}$ and a distance function δ, the *detectGroups* function (line 2, Fig. 1-*c*) applies a clustering method to find groups of similar units on $\widehat{S}$ with respect to δ. An appropriate clustering method and distance function can be selected according to the aim of the analysis. Note that the distance function δ must allow the extraction of a centroid. Once the groups of units $\mathcal{G}$ are detected, from each group the *getCentroid* function (line 3) extracts the centroid $F^{(i)}$ representing a footprint of the customer. Then, the footprint sequence $\widehat{F}$ is built considering $\widehat{S}$, $\mathcal{G}$ and $\mathcal{F}$ using the function *buildSequence* (line 4). Finally, the temporal purchasing profile $\mathcal{P}$ returned by the algorithm (line 5) is composed by the set of distinctive footprints $\mathcal{F}$ and the distinctive sequence $\widehat{F}$ form. The computational complexity of Algorithm 1 is dominated by the complexity of the *detectGroups* function that implements a clustering algorithm.

4 Collective perspective of individual profiles

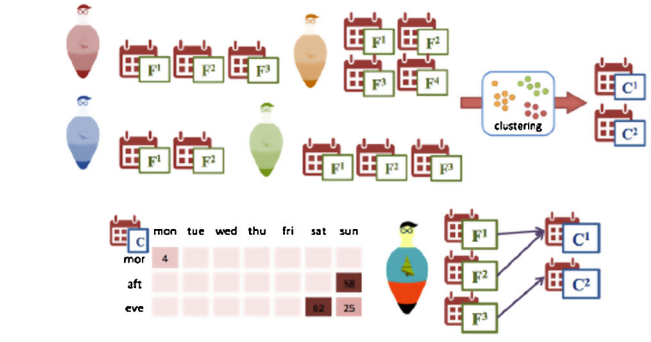
Individual profiles of different customers are not comparable because each customer can have a different number of footprints expressing different behaviors. Thus, in order to compare individual profiles of different customers we need to provide them a *collective perspective*. Given the profiles $\mathcal{P}_b$ and $\mathcal{P}_c$ of customers b and c, this means to make comparable footprints $\mathcal{F}_b$ and $\mathcal{F}_c$, and footprint sequences $\widehat{F}_b$ and $\widehat{F}_c$. We start by specifying

how to compare the footprints by defining the *collective temporal purchasing footprint* (collective footprint in short) as:

Definition 5 (Collective temporal purchasing footprint) Given a set of individual footprints $\{\mathcal{F}_c\}$ and a distance function $\delta : \mathbb{R}^{t\times d}\times\mathbb{R}^{t\times d}\rightarrow\mathbb{R}$, we define the *collective temporal purchasing footprint groups* $\mathcal{L} = \{L_1,\ldots,L_k\}$ as a partitioning of $\{\mathcal{F}_c\}$ into disjoint sets of similar footprints and, given a group $L = \{F^{(1)},\ldots,F^{(q)}\}$ of individual footprints we name *collective temporal purchasing footprint* the centroid of L, i.e., $C \in \mathbb{R}^{t\times d}$ such that

$$C_{ij} = \frac{1}{|L|}\sum_{\forall F^{(h)}\in L} F_{ij}^{(h)} \quad \forall i = 1,\ldots,t, j = 1,\ldots,d.$$

We name $\mathcal{C} = \{C^{(1)},\ldots,C^{(k)}\}$ the *set of collective footprints of all customers*. Given a customer c, her footprints $\mathcal{F}_c$, the collective footprints $\mathcal{C}$ and the collective groups $\mathcal{L}$, we denote with *customer collective footprints* $\mathcal{C}_c = \{C^{(1)},\ldots,C^{(q)}\}$ the *collective perspective* of $\mathcal{F}_c$, where $\mathcal{C}_c \subseteq \mathcal{C}$ and $\forall C^{(h)} \in \mathcal{C}_c\ \exists F^{(i)} \in \mathcal{F}_c$ s.t. $F^{(i)} \in L_h$ with $L_h \in \mathcal{L}$ and $C^{(h)}$ is the centroid of L_h. Note that two different footprints $F^{(i)}$ and $F^{(j)}$ in a collective perspective can belong to the same collective group L_h and thus, they can be represented with the same collective footprint $C^{(h)}$. In Fig. 3-*(a)* we show an example illustrating how two different individual footprints for customer c correspond to the same collective footprint.



(a) We cluster the individual footprints of all the customers obtaining the set of collective temporal purchasing footprints of all the customers $\mathcal{C} = \{C^{(1)},\ldots,C^{(k)}\}$. A *collective temporal purchasing footprint* $C \in \mathbb{R}^{t\times d}$ is a matrix, where C_{ij} contains the typical relevance of a collective pattern in the i-th time window of the j-th day-interval. We indicate with *customer collective footprints* $\mathcal{C}_c = \{C^{(1)},\ldots,C^{(q)}\}$ the *collective perspective* of $\mathcal{F}_c$ for customer c, i.e., the mapping of the individual footprint using the collective footprints.

(b) For each customer c, given the temporal purchasing profile $\mathcal{P}_c = \langle\mathcal{F},\widehat{F}\rangle$, and the collective footprints of all customers $\mathcal{C}$, the *collective perspective* of $\mathcal{P}_c$ is denoted as $\mathcal{P}_c^* = \{\mathcal{C}_c,\mathcal{R}_c\}$ where $\mathcal{C}_c \subseteq \mathcal{C}$ are the customer collective footprints, and $\mathcal{R}_c$ is the set of regular sub-sequences.

Figure 3 From temporal purchasing profile to collective perspective

We underline that we use the expression *customer collective footprints* to indicate the collective perspective of a single customer $\mathcal{C}_c$ and the expression *collective footprints of all customers* to indicate $\mathcal{C}$.

Given the collective group of collective footprints $\mathcal{L}$, and the collective footprints of all the customers $\mathcal{C}$, we can replace each individual footprint sequence contained in $\{\widehat{F}_c\}$, with the customer collective footprint representing the collective group to which it belongs to. Therefore, for each customer c her footprint sequence $\widehat{F}_c$ is mapped to an equivalent *collective temporal purchasing sequence* $\widehat{C}_c$ (collective sequence in short). Similarly to the collective footprints, the collective sequences of all the customers are comparable among each other.

Definition 6 (Collective temporal purchasing sequence) Given a customer c, her footprint sequence $\widehat{F}_c$, the collective groups $\mathcal{L}$ and the collective footprints of all customers $\mathcal{C}$, we define the *collective temporal purchasing sequence* as the sequence $\widehat{C}_c$ obtained replacing each footprint in $\widehat{F}_c$ with the corresponding collective footprint in $\mathcal{C}$ according to the groups $\mathcal{L}$.

In addition, in order to better studying the customers habits and to determine which are the most common sub-sequences we define the *regular temporal purchasing sub-sequences* (regular sub-sequences in short) as:

Definition 7 (Regular temporal purchasing sub-sequences) Given a customer c, her collective sequence $\widehat{C}_c$ and a support threshold ω, we define the *regular temporal purchasing sub-sequences* is the set $\mathcal{R}_c = \{(R_1, w_1), \ldots, (R_m, w_m)\}$, where each R_i is a sub-sequence of $\widehat{C}_c$, w_i is its support and $\forall w_i w_i \geq \omega$.

In other words, among all the possible sub-sequences of $\widehat{C}_c$, $\mathcal{R}_c$ contains only the most representative for c. For example, if the sub-sequences of $\widehat{C}_c$ are $(\langle C^{(1)}, C^{(1)}\rangle, 10)$, $(\langle C^{(1)}, C^{(1)}, C^{(2)}\rangle, 8)$, $(\langle C^{(1)}, C^{(2)}\rangle, 2)$, $(\langle C^{(2)}, C^{(1)}\rangle, 2)$, $(\langle C^{(2)}, C^{(2)}\rangle, 1)$ where the number is the support, i.e., the occurrences of that sub-sequence, than only the first two are *regular* and contained in $\mathcal{R}_c$ with $\omega = 5$. Given two customers b and c, with $\mathcal{R}_b$ and $\mathcal{R}_c$ derivable from $\widehat{C}_b$ and $\widehat{C}_c$, we can compare $\mathcal{R}_b$ and $\mathcal{R}_c$ with an appropriate distance function, e.g. *Jaccard* or *cosine* distance.

Finally, we define the *collective perspective* (see Fig. 3-*(b)*) of a profile as follows:

Definition 8 (Collective perspective) Given the profile $\mathcal{P}_c = \langle \mathcal{F}, \widehat{F}\rangle$ of customer c, and the collective footprints of all customers $\mathcal{C}$, the *collective perspective* of $\mathcal{P}_c$ is defined as $\mathcal{P}_c^* = \{\mathcal{C}_c, \mathcal{R}_c\}$ where $\mathcal{C}_c \subseteq \mathcal{C}$ are the customer collective footprints, and $\mathcal{R}_c$ is the set of regular sub-sequences.

Providing Collective Prospective to Individual Profiles. The process for providing the collective perspective to the individual profiles is summarized by Algorithm 2. Starting from the set of individual profiles $\{\mathcal{P}_c = \{\mathcal{F}_c, \widehat{F}_c\}\}$, it first employs *detectGroups* to detect from the individual profiles of all the customers the collective groups of footprints $\mathcal{L}$ (line 1). Then from each group in $\mathcal{L}$ the *getCentroid* function (line 2) extracts the centroid $C^{(i)}$. The union of the centroids forms the collective footprints of all the customers $\mathcal{C}$. Then, using $\mathcal{C}$ and the groups in $\mathcal{L}$, for each customer c in line 4 it maps the individual footprints $F^{(i)} \in \mathcal{F}_c$

Algorithm 2: *provideCollectivePerspective*

Input : $\{\mathcal{P}_c = \{\mathcal{F}_c, \widehat{F}_c\}\}$—purchasing profiles of all customers, δ—distance function,
Output: $\{\mathcal{P}_c^* = \{\mathcal{R}, \mathcal{C}_c\}\}$—collective perspective of all customers

1 $\mathcal{L} \leftarrow detectGroups(\{\mathcal{F}_c\}, \delta)$
2 $\mathcal{C} \leftarrow \{C^{(i)} | C^{(i)} = getCentroid(L_i)\ \forall L_i \in \mathcal{L}\}$
3 **for** *each customer c* **do**
4 $\quad \mathcal{C}_c \leftarrow mapIntoCollective(\mathcal{F}_c, \mathcal{L}, \mathcal{C})$
5 $\quad \widehat{C}_c \leftarrow buildSequence(\widehat{F}_c, \mathcal{L}, \mathcal{C}_c)$
6 $\quad \mathcal{R} \leftarrow extractRegularSubsequences(\widehat{C}_c)$
7 **end**
8 **return** $\mathcal{P}_c^* = \langle \mathcal{C}_c, \mathcal{R}_c \rangle$

to a collective footprints $C^{(k)} \in \mathcal{C}$ creating in this way the collective perspective $\mathcal{C}_c$ of the footprints $\mathcal{F}_c$. By using the function *buildSequence* (line 4), the collective perspective is provided to the footprint sequence $\widehat{F}_c$ generating the collective sequence $\widehat{C}_c$ by means of $\mathcal{C}_c$. The function *extractRegularSubSequences* (line 5) extracts from $\widehat{C}_c$, the regular sub-sequences of the customer $\mathcal{R}_c$. Finally, the collective perspective $\mathcal{P}_c^*$ for all the customers is returned.

We implement the exaction of the regular sub-sequences using a suffix tree [26]. Given a customer c, her collective sequence $\widehat{C}_c$ is transformed into a string where each character corresponds to the label of a customer collective footprint. Hence, we generate a suffix tree for each customer. Following a branch of the tree from the root to a leaf we can read a sub-sequence R_i and, on the leaf, we have the support w_i of the sub-sequence generating that branch. We set the support threshold ω in a data-driven way by observing the support distribution among the sub-sequences. In particular, we apply a technique known as "knee method" [27] setting ω to the value of the knee. Given a set of pairs composed of items and their support this method sorts the pairs according to the frequencies and returns the most representative, i.e., the pairs with a support greater or equal than the support ω corresponding to the *knee* in the curve of the ordered frequencies. In this way ω is different for each customer and driven by personal data. For each customer, we cut the suffix tree considering only the *regular* sub-sequences, i.e., the sub-sequences R_i with support $w_i \geq \omega$. The complexity of Algorithm 2 is dominated by the maximum between the complexity of the clustering algorithm (*detectGroup*) and the complexity of the construction of the suffix tree (*extractRegularSubsequences*) [26].

5 Case studies

In this section, we apply the proposed framework for temporal purchases analysis on a real world datasets. We show the individual temporal purchasing footprints, the effect of the collective perspective, and we analyze the most common regular sub sequences for the customers segmented in similar groups. We underline that the proposed framework, as well as the other analytical approaches described in Sect. 2, are designed to extract knowledge from the data. All of them are not assessing a task which can be quantified (e.g. a prediction or a classification). As consequence, the proposed framework is not eligible

for comparison against these other methodologies. However, it is possible to instantiate the same framework on various datasets characterized by different properties.

In line with [24], we present a main case study on a private real big dataset of shopping sessions. Then, we show that the same framework can be easily exploited for the analysis of other public available datasets. In the main case study we show the overall potentialities of the proposed framework, while on the other datasets we highlight the modeling of the framework together with the principal findings and similarities with the main case study.

We underline that the proposed methodological framework for the analysis of the temporal dimension of shopping has the goal of providing profiles and behaviors for analyzing this kind of information. As consequence, similarly to other approaches described in Sect. 2, we do not compare our methodology against other methods, but we show that the framework can be instantiated to analyze other similar datasets.

The rest of this section in organized as follows. In Sect. 5.1 we illustrate all the datasets analyzed. Section 5.2 details the model setting for the principal case study, and in Sects. 5.3, 5.4 and 5.5 we report the results for the individual footprints, collective footprints and sequences, respectively. By using additional information, in Sect. 5.6 we show that the various clusters, obtained only using the temporal dimension, are capturing diversified qualitative aspects of the customers (e.g., age and profession). Section 5.7 proves that the methodological framework can be instantiated for other case studies and that similar results to the principal case study are found. Finally, in Sect. 5.8 we demonstrate the clustering stability with respect to different choices of the number of collective footprints.

5.1 Real datasets

In our principal case study we analyze a retail dataset provided by *UniCoop Tirreno*[a] an Italian supermarket chain. The customers of UniCoop Tirreno can obtain a fidelity card enabling them to discounts. Through this fidelity card, the company can tie each shopping session to the customer whom made the purchase. We analyze 49,590,010 shopping sessions of about 91k customers that occurred from the 1st January 2007 to the 1st June 2015 in the area of Leghorn province. All the customers respect the following minimum constraints: *(i)* they performed on average at least a shopping session per month along the observation period (i.e., 12 shopping sessions per year), *(ii)* they performed at least 10 shopping sessions in different months for each year of the dataset. During the years observed the province of Leghorn has an average population of about 343,000 inhabitants. Assuming an average size of 2–3 people per household, we estimate a coverage of 60% of the population. Figure 4*(upper left)*: reports the number of shopping sessions per customer. The mode is $\sim$350, meaning that customers usually visit the shops around once a week. In the *(upper right)* plot is shown the total amount spent per time of the day. Finally, in Fig. 4*(lower left)* is reported the total amount spent per weekday. It appears that customers have a preference for shopping in days close to the weekend.

In the literature there are very few similar public available transactional datasets providing detailed temporal information about shopping sessions. Examples are *Ta-Feng*[b] and *T-Mall*[c] datasets. Ta-Feng dataset covers products like food, office supplies and furniture. It contains 817,741 transactions registered in four months (from 2000-11-01 to 2001-02-27) and belonging to 32,266 customers. T-Mall dataset records four months (from 2014-04-15 to 2014-08-14) of online transactions of an online e-commerce website. It is conceptually different from the previous datasets because a transaction does not model a purchase but

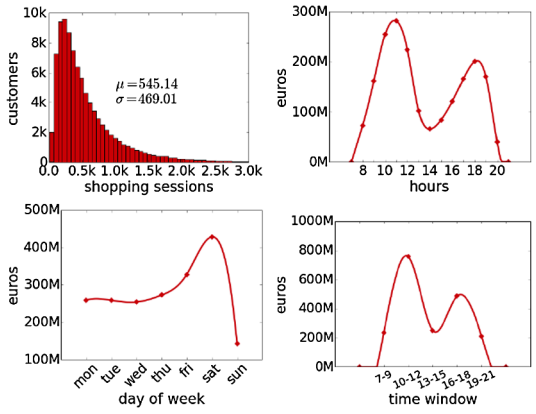


Figure 4 Dataset statistics. Distributions of number of shopping sessions per customer (upper left), amount spent per hours (upper right), amount spent per day of week (lower left), and amount spent per time window (lower right)

the fact that a set of items have been observed in the same day. It contains 4298 transactions belonging to 884 users and 9531 brands considered as items. Even though these datasets refer to a time period remarkably shorter than the period observed in *UniCoop* dataset, in order to show how our methodological framework can be easily instantiated in other case studies, we report in Sect. 5.7 some crucial analytical results on Ta-Feng and T-Mall datasets.

5.2 Model setting

As humans we operate under the cadence of a *seven-day week* [28]. This cycle of activity is deeply rooted in human experience and in our psychological habits. Indeed, the weekdays alternation drives our routinary life. Together with the previous observations, these are the reasons why we decided to adopt the week as time unit and to set the period $\tau = 7$ and the number of days-intervals $d = 7$. With respect to the time windows, by observing Fig. 4*(upper right)* we notice an *M-shaped* pattern: most of the shopping sessions happen in the morning or after working hours. As consequence, in order to capture all the phases of this curve, we summarize this trend using the data-driven time windows, containing all the phases of growth and decrease of the curve and so, by setting $t = 5$ and dividing the time as follows: 7–9, 10–12, 13–15, 16–18, 19–21 how highlighted in Fig. 4*(lower right)*. Note that, without this time discretization and adopting a finer granularity (e.g. a time slot every hour) the discovered profiles would not be sufficiently easy to read, this because we might have duplication of profiles for customers shopping in the same time slots. On the other hand, by adopting a coarser granularity (two times slot for morning and afternoon shopping sessions) we might miss some crucial differences in the temporal shopping behavior which are highlighted by the findings in the following sections. As relevance function *rel* we used the *sum* of the *amount* spent.

We implemented *detectGroups* in Algorithm 1 and Algorithm 2, using the *k-means* clustering algorithm [27]. K-means algorithm requires to specify the number of clusters k.

For the extraction of the individual profiles Algorithm 1 does not take as input the number of cluster k, but the algorithm automatically estimates the number of clusters, i.e., the number of individual footprints for each customer, by running k-means for $k \in [2, 50]$ and

selecting as number of cluster the k which can be considered the "knee" of the Sum of Squared Error (SSE) curve. The idea is to find the best k for each customer and not using the same value for everyone. In particular, we select as *knee* the point on the SSE curve having the maximum distance from the straight line passing through the minimum and the maximum point of the SSE curve. On the other hand, for *detectGroups* in Algorithm 2 we used k-means with $k \in [2, 150]$ and, yet using the knee method, analyzing the SSE curve we select $k = 45$ as number of collective groups. In both cases, as distance function δ we used the *cosine distance* because unlike the Euclidean or Manhattan, it does not suffer the problem of sparseness. Indeed, typically a customer purchases one or two times per week generating very sparse footprints F.

Therefore, by applying in sequence Algorithm 1 and Algorithm 2 we obtain for each customer her profile P_c and the collective perspective P_c^*.

5.3 Individual footprints analysis

In this section we analyze the individual temporal purchasing footprint $\mathcal{F}_c$ contained in the temporal purchasing profiles $\mathcal{P}_c$ extracted employing Algorithm 1 on all the customers. Empty footprints are clustered together by default and are represented with an empty footprint. The extraction time is about 0.5–1.0 seconds per customer, depending on the number of non empty footprints.

In Fig. 5*(upper left)* is reported the distribution of the number of customers without empty footprints, i.e., the weeks for which at least a purchase is performed. It results to be quite uniform: the customers are various with respect tot the number of purchases performed. Figure 5*(upper right)* shows the distribution of the number of individual footprints. This distribution has a Gaussian shape and its mode is ∼8: most of the customers have eight distinct individual footprints capturing distinct temporal shopping behaviors. We also notice that about 80% of the customers must be represented using more than five footprints. This happens because even though a customer makes purchases on a certain day and time window, she can spend sometimes €50, sometimes €70 and sometimes €90. These three behaviors appear in the same time slot but they are "distinct" due to the different nature of the amount spent, and they have for the customer a different meaning.

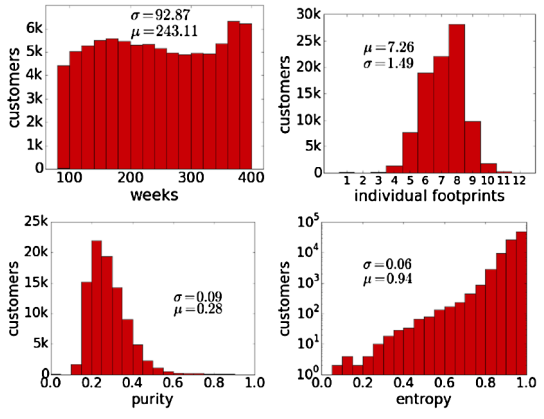


Figure 5 Individual footprints analysis. Distributions of customers per: number of weeks (upper left), number of individual footprints (upper right), purity (lower left), entropy (lower right)

The last two plots in Fig. 5 show the distributions of two indicators of regularity: purity *(lower left)* and entropy *(lower right)*. The *purity* indicates how much the customer is pure in terms of footprints [27]: how many units are represented by the most frequent footprint. On the other hand, the *entropy* indicates how much a customer is heterogenous in terms of footprints [29]: how much the units are balanced with respect to the footprints from which they are represented. We calculate the purity and the entropy according to the following formulas:

$$purity = \max_{G_i \in \mathcal{G}} \big(\sup(G_i)\big), \qquad entropy = -\sum_{G_i \in \mathcal{G}} \sup(G) \log\big(\sup(G)^{-1}\big) / \log\big(|\mathcal{G}|\big),$$

where $\sup(G_i) = |G_i|/|\hat{S}|$ is the relative support of a footprint, i.e., the number of footprints belonging to the group represented by G_i. We notice in Fig. 5 how distribution of the purity is a Gaussian with mode ∼0.2, while the entropy has a long tailed distribution with mean 0.94 and low standard deviation. Thus, for many customers the most frequent footprint captures about 20% of the units, while in general the units per footprint are well balanced causing in general an high entropy.

In light of this, the upcoming analysis aims at understanding how the collective perspective impacts the individual footprints, and the observed indicators.

5.4 Collective footprints analysis

In this section we analyze the customers' collective footprints representing the collective perspective obtained using Algorithm 2 on the individual profiles $\mathcal{P}_c$.

First of all we analyze the same indicators of the previous section for the collective footprints. In Fig. 6*(upper left)* we report the distribution of the number of customers per collective footprints $|\mathcal{C}_c|$. The fact that two individual footprints $F^{(i)}$ and $F^{(j)}$ under the collective perspective can be represented by the same collective footprint $C^{(k)}$ affects in a not negligible way a relevant number of customers. Indeed, instead of the Gaussian distribution of Fig. 5*(upper right)* we discover the bi-modal distribution of Fig. 6*(upper left)*. This recalls the *explorers* and *returners* phenomenon observed in mobility data [30]: when

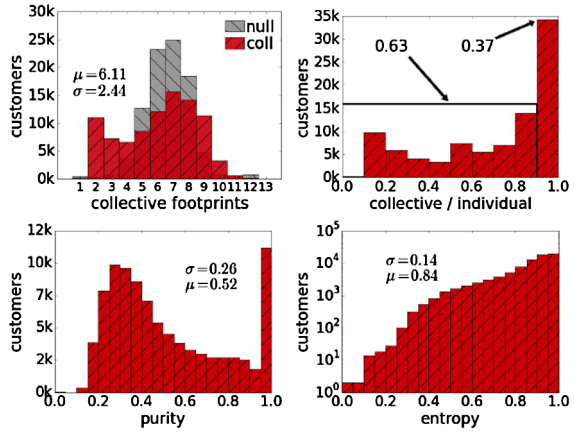


Figure 6 Collective footprints analysis. Distributions of customers per: number of collective footprints and null model (upper right), ratio between number of collective footprints and individual footprints (upper left), purity (lower left), entropy (lower right)

there is a collective threshold for all the customers analyzed, either fixed and imposed like in [30] (with $K = 2, 4, 6, \ldots$), or data-driven [31] like in our case, a dichotomy of two very distinctive behaviors appears from the models capturing the habits of the customers.

In order to prove that this result is not casual, yet in Fig. 6*(upper left)* we report the distribution generated by a *null model*: for each customer, each individual footprint is randomly assigned to a collective footprint. In other words, we preserve the number of individual footprints for each customer and the number of collective footprints, while we destroy the assignment returned by the clustering, and we label each individual footprints using a randomly selected collective footprints instead of using the one of the cluster where it should belong to. This *null-distribution* (in grey) has a Gaussian shape with mode $\sim$7 and, due to the difference with the bi-modal, it allows us to state that the bi-modal distribution is not a casual result. Hence, the bi-modal distribution statistically delineates two subsets of customers: the *regular* customers represented by a limited set of collective footprints, and the *changing* customers requiring a higher number of collective footprints.

Using 4 as threshold we obtain a 27%–83% partitioning. By comparing the individual with the collective footprints we can discriminate between two well defined groups of customers. *Regular* customers are more predictable than *changing* customers since they can adopt a smaller range of temporal footprints. Figure 6*(upper right)* illustrates the distribution of the ratio between the number of collective footprints and individual footprints $|\mathcal{C}_c|/|\mathcal{F}_c|$. For 37% of the customers each individual footprint belongs to a different collective group, while for the rest the collective perspective changes the personal definition of their behavior putting two different individual footprint in the same collective group.

Also the distributions of purity and entropy, the plots in bottom line of Fig. 6, are remarkably different from those in Fig. 5. For purity we can observe a novel group of $\sim 10k$ pure customers, while for entropy, even though the distribution remains long-tailed, now the units are more unbalanced towards a few set of footprints representing the whole customer purchasing behavior. Moreover, we observe the growth of the standard deviation σs for both measures. In addition, the average purity for a *regular* customer is 0.94, while it is just 0.19 for a *changing* customer. We notice a similar effect for entropy: the average entropy for a *regular* customer is 0.65 while it is 0.91 for a *changing* customer. This confirms the higher unpredictability of changing customers. The regular-changing partitioning is the first segmentation that emerges by employing our methodological framework. In the following we move over *changing-regular* looking for other interesting temporal segmentations of the customers.

We now visually show the expressive power of the footprints. Figure 7 reports a subset of the 45 collective temporal purchasing footprints of all the customers $\mathcal{C}$ obtained.[d] The number in the bottom left square indicates how many customers have an individual footprint $F^{(i)}$ represented by that particular collective footprint.

In Fig. 7, with the exception of the collective footprints *(29)* and *(38)*, all the collective footprints describe a *one-shop* behavior, i.e., most of the customers perform only one purchase per week. However, the day and time window of these *one-shop* purchases is spread among the various possible choices. For example, customers with a footprint represented by *(1)* spend about €37 on Sat10–12, those having a footprint represented by *(14)* spend about €49 on Fri10–12, and those represented by *(4)* spend about €55 on Fri16–18. As

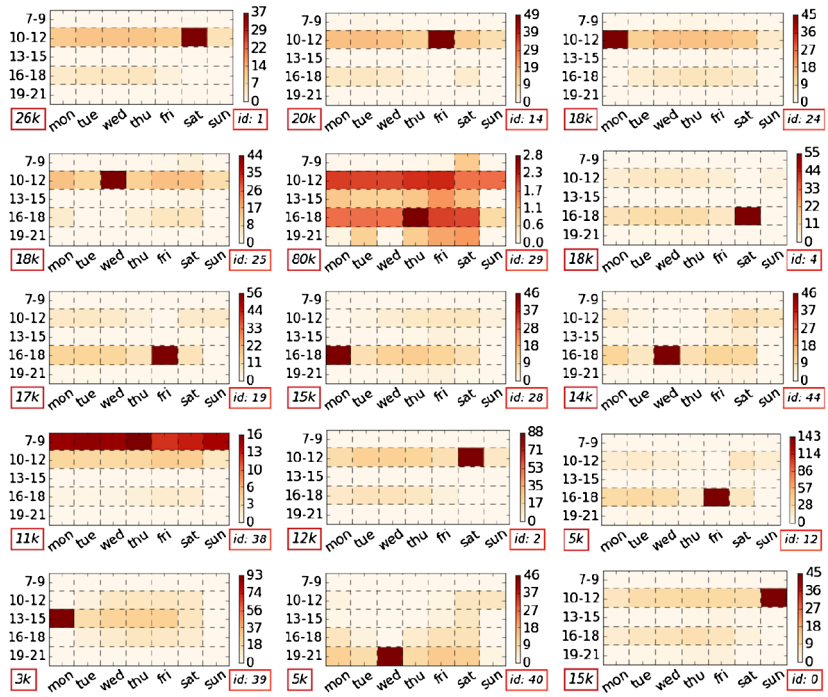


Figure 7 Collective temporal purchasing footprints. Selection of the 45 collective temporal purchasing footprints depicted as heatmaps: the darker the color, the higher the total amount spent (in €). On the *x* axis are reported the days of the week, while on the *y* axis are reported the five time slots

anticipated by the M-shape in Sect. 5.1, the two time windows mostly used are 10–12 and 16–18, but there are also some collective footprints in "unusual" time windows, e.g. *(39)* and *(40)*.

We notice that a shopping behavior for the same day and time window can be captured by different collective footprints (e.g. *(1)–(2)*, *(19)–(12)*) with a different maximum amount. Thus, we classify these *one-shop* spending behaviors in three classes according to maximum amount spent. We name *normal* spending footprint the collective footprints lower than €50, *high* spending footprint those between €50 and €100, and *higher* spending footprint those with a peak higher than €100.

Moreover, collective footprint *(29)* captures occasional shopping sessions where a maximum of €3 is spent nor in a specific day nor in a specific time window. However, 87% of the customers have the behavior described by this collective footprint. This indicates that even though each customer has a quite regular *one-shop* footprint, she sometimes makes purchases following an *occasional* spending footprint.

Finally, collective footprint *(38)* captures the behavior of customers that every morning (7–9) make a purchase spending at most €16. We name this behavior *daily* spending footprint. The customers having this footprint can be retirees who go to the shopping center every morning to satisfy only their daily needs, or workers going to the supermarket before work for buying their lunch. This is the second segmentation unveiled from the analysis of the temporal purchasing profiles.

The last analysis consists in discovering which are the most common orders in which these collective footprints are adopted by the customers.

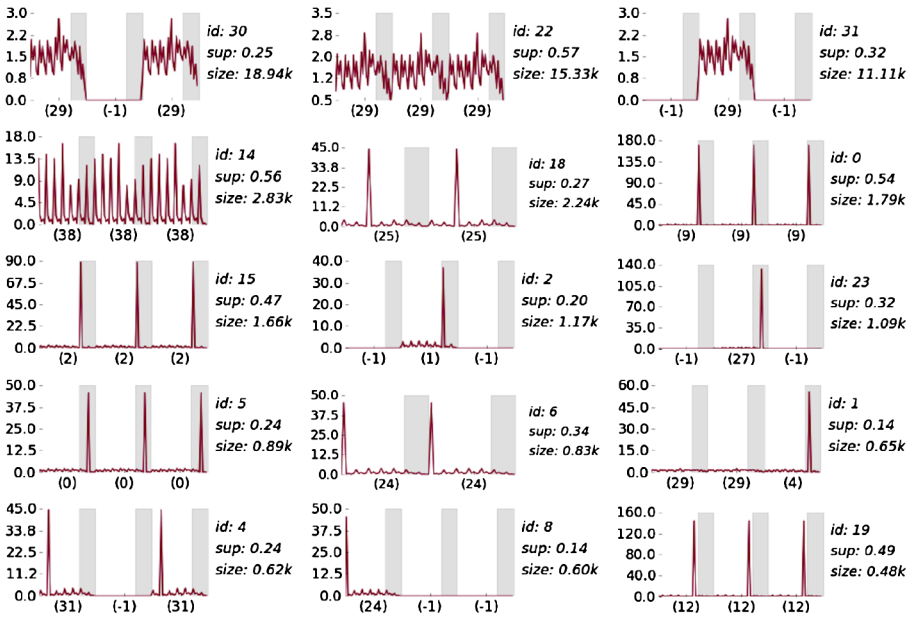


Figure 8 Regular sub-sequences. Selection of representative regular sub-sequences of the medoids of the collective clustering on the regular sub-sequences. Gray rectangles highlights the weekends, *(–1)* indicates the footprint modeling the no-shopping behavior

5.5 Collective sequences analysis

In this section we analyze the regular sub-sequences $\mathcal{R}_c$. They can help in unveiling an additional customer segmentation which describes for each group the common orders in which the collective footprints are repeated. Therefore, we partition the customers by using the regular sub-sequences $\mathcal{R}_c$ with *k-medoids* clustering algorithm [32] varying $k \in [2, 80]$, and with the knee method we select $k = 33$ clusters. We show in Fig. 8 some medoids with interesting sub-sequences. We highlight that these sub-sequences are not expressing the fact that the customers belonging to that cluster always behave in that way, but they are describing one of the most common behavior for these customers, i.e., their temporal *routines* in purchasing.

Most of the clusters are characterized by a repetition of the same collective footprint, e.g. clusters 14, 18, 0, 15, 5, 6 and 19 in Fig. 7. Customers belonging to these clusters have a preferred moment to shop and/or they need to shop in that particular moment. This behavior is probably driven by their weekly time table. The fact that there are not no-shopping behavior separating these *one-shop* behaviors is a signal that they consume all the products purchased and they need to re-purchase every week spending approximately the same amount. Cluster 14 reveals that also the *daily* spenders repeat regularly their footprint through the weeks. Clusters 30, 22 and 31 capture different permutations of collective footprints *(29)* and *(–1)*, *(–1)* indicates no-shopping. These customers generally purchase in subsequent weeks without a regular pattern. Indeed, they are mostly *changing* customers. Moreover, cluster 30 follows a Yes-No-Yes^e (Y-N-Y) pattern (complementary to 31), while customer in cluster 22 buys every week (Y-Y-Y). Clusters 2 and 23 capture two different repetitions of *one-shop* footprint following a N-Y-N pattern, i.e., these customers depletes her storage in the first week, go to shopping in the second week, and consumes the novel supplies in the third week. Cluster 4 is complementary to 2. Finally, clusters 1 and 8

are specular each other with a N-N-Y and Y-N-N pattern. In the first the customers do not purchase for two weeks and then, spent about €60 on Saturday morning of the third week, while the customer of cluster 8 spend €45 on Monday morning and then do not need to purchases for two weeks. This last analysis shows another possible segmentation driven by the temporal sequences.

5.6 Learning more about the customers in different groups

In this section we further characterize the different groups of customers discovered in the UniCoop dataset. A fist partitioning, which is related not to a specific behavior but rather to the number of collective temporal purchasing footprints adopted, is the one into *regular* customers and *changing* customers. Besides the differences highlighted in the previous sections, we observe the amount spent by the two groups which does not depend on the partitioning into *regular* or *changing*. In particular, the standard deviation of the amount spent by a *regular* customer is 7.61, while it is 32.38 for a *changing* customer. This means that temporal regularity reflects into the spending regularity. The regular customers having few different collective footprints are also regular with respect to the amount spent (low standard deviation of the amount spent). On the other hand, changing customers that can follow many different collective footprints are also more eclectic with respect to the amount spent (high standard deviation of the amount spent). This fact may be reasonable; indeed, if a customer has only small variations in the times she goes to shopping, then it is likely that her regularity depends on the fact that she has a purchasing periodic plan that allows her to consume all the bought items. This leads to an high probability that every time the customer goes to the shopping center will need approximately a *comparable* set of items to those in the previous purchase. This is not the case if the time between two shopping session varies a lot. In our claim is that it is likely that a regular customer with respect to the temporal dimension is also a regular customer with regards the items bought.

Another categorization we purse is the one related to the various collective footprints. By leveraging additional partial information on some customers we can discover who are the customers belonging to the various clusters. In particular, we observed the age of the customers at their last purchase and the declared profession for some selected clusters: *(1)*, *(24)*, *(29)*, *(38)*, *(40)*, *(0)*. The results are depicted in Figs. 9 and 10 for age and profession respectively. In Fig. 9 we show the age distribution for the customers having a collective

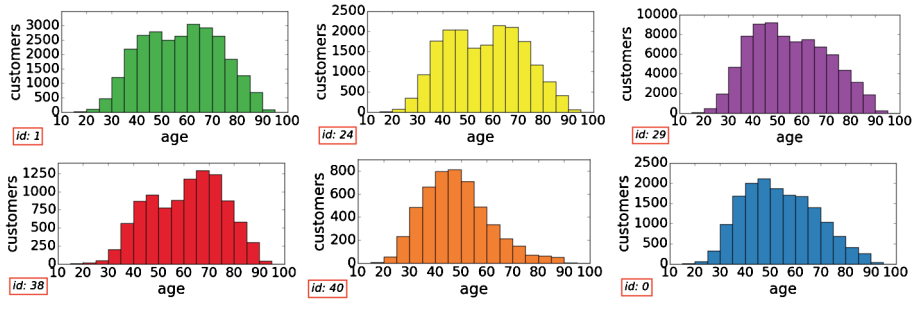


Figure 9 Customers age exploration. Distribution of age for the customers having a collective footprint among *(1)*, *(24)*, *(29)*, *(38)*, *(40)*, *(0)*

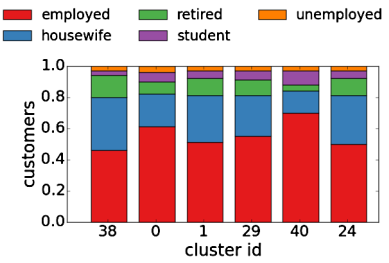


Figure 10 Customers profession exploration. Percentages of different professions for the customers having a collective footprint among *(1)*, *(24)*, *(29)*, *(38)*, *(40)*, *(0)*

footprint in such clusters. We can notice how the shopping temporal habits and regularities are driven by the age which assume markedly different distributions. For example the collective footprint *(38)* (daily and early morning shopping behavior see Sect. 5.4) contains a large group of old people and, by looking at Fig. 10, the large percentage of *retired* customers (∼15%). On the other hand, collective footprint *(40)* has on average younger customers and the largest portion of *student* (∼10%) and employed (∼70%). Finally, the largest groups of *housewife* have collective footprints *(24)*, *(29)* and *(28)* where in each of them purchases are performed consistently in morning hours.

5.7 Methodological framework portability

In this section we show that our framework can be instantiated for other case studies based on data with different characteristics with respect to the Unicoop data. In particular, the application of our methodology must take into consideration the fact that in the data there is no information about both the purchasing time and the amount spent. However, similar results with respect to the principal case study are found.

As for the principal case study, also for Ta-Feng and T-Mall datasets we adopt the week as time unit and set the period $\tau = 7$ and the number of days-intervals $d = 7$. However, since the time of the shopping is not available, we cannot model it and we consequently use a unique time window ($t = 5$). Moreover, also the amount spent is not available, thus as relevance function *rel* we used the *sum* of the *number of items* purchased in each shopping session. As before, we implemented *detectGroups* in Algorithm 1 and Algorithm 2, using the *k-means* clustering algorithm [27]. For the extraction of the individual profiles we estimated the number of clusters by running k-means for $k \in [2, 40]$, and as before, selecting as number of individual footprints the k corresponding to the "knee" of the SSE curve. Similarly, for *detectGroups* in Algorithm 2 we used k-means with $k \in [2, 145]$ and, yet using the knee method, analyzing the SSE curve we select $k = 12$ for Ta-Feng and $k = 15$ for T-Mall as number of collective footprints. For each customer we get both her profile P_c and the corresponding collective perspective P_c^*.

In this section we analyze the individual temporal purchasing footprint $\mathcal{F}_c$ contained in the temporal purchasing profiles $\mathcal{P}_c$. In Fig. 11(left column) we report the distribution of the number of individual footprints for Ta-Feng (top) and T-Mall (bottom) datasets. Similarly to the case of UniCoop, both distributions are Gaussian with a mode of 4 and 5, respectively. In Fig. 11(right column) we show the distribution of the number of customers per collective footprints $|\mathcal{C}_c|$. Again, instead of Gaussian distributions we found bimodal distributions delineating the previously observed *regular* and *changing* customers. By adopting 3 as threshold of the number of collective footprints to separate regular and changing behaviors, we obtain a ∼34%–66% partitioning. Looking at the standard devi-

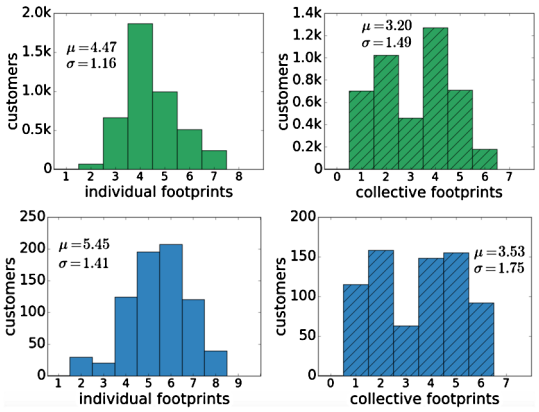


Figure 11 Ta-Feng and T-Mall collective footprints analysis. Distributions of customers per number of individual footprints (left) and number of collective footprints (right), for Ta-Feng (top) and T-Mall (bottom) datasets

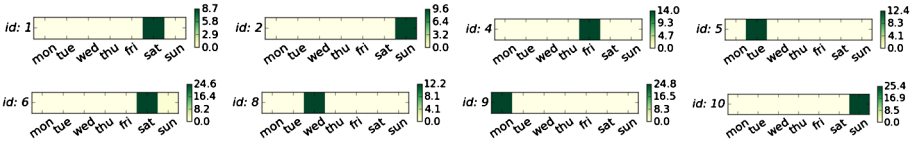


Figure 12 Ta-Feng collective temporal purchasing footprints. Selection of the 12 collective temporal purchasing footprints for Ta-Feng dataset depicted as heat-maps: the darker the color, the higher the number of items purchased. On the *x* axis are reported the days of the week

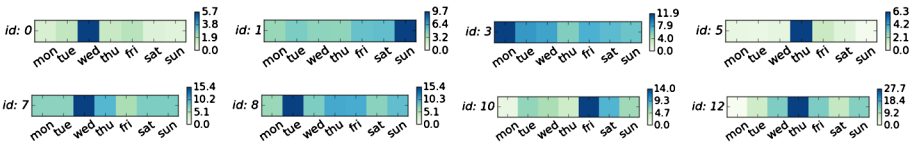


Figure 13 T-Mall collective temporal purchasing footprints. Selection of the 15 collective temporal purchasing footprints for T-Mall dataset depicted as heat-maps: the darker the color, the higher the number of items purchased. On the *x* axis are reported the days of the week

ations of the number of items purchased, we have a difference less marked than for Uni-Coop case but it is still present: in Ta-Feng it is 2.37 for regular and 3.50 for changing, while for T-Mall it is 0.88 for regular and 1.75 for changing. This is another confirmation that temporal regularity reflects into the spending regularity.

In Figs. 12 and 13 we report a subset of the collective temporal purchasing footprints of all the customers obtained for Ta-Feng and T-Mall, respectively. All the collective footprints for Ta-Feng (Fig. 12) capture a one-shop temporal purchase. For some days we have different typical number of items: both collective footprints *(2)* and *(10)* refer to a peak on Sunday. On the other hand, even though for T-Mall we also have one-shop collective footprints, these are "softer", showing that a lower level of shopping session appears also in days different from the one corresponding to the peak. In both cases we can find more than one collective footprint for each day of the week, but the days in which the peak is higher are those in the weekend: Friday, Saturday and Sunday.

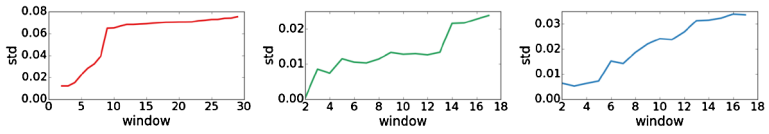


Figure 14 Number of collective footprints validation. Standard deviation of the silhouette coefficients for number of clusters k in a growing window with respect to a selected number of clusters: $k = 45$ for UniCoop, $k = 12$ for Ta-Feng and $k = 15$ for T-Mall. On the x axis is reported the size of the window, i.e., the number of silhouette coefficients considered with respect to the selected. On the y axis is reported the standard deviation of the silhouette coefficients in the window centered in the selected k

5.8 Collective footprints validation

In this section we proof that the technique adopted for selecting the number of clusters, both for the individual and collective footprints is robust: small variations in the number of clusters is not changing the overall conclusions of the paper. To assess this task we used an alternative external validation measure to the sum of squared error adopted for selecting the number of clusters. The silhouette coefficient is another useful criterion for assessing the natural number of clusters in a set of data [27]. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

In the following we analyze the choice of the number of collective footprints for the various datasets. For each dataset we calculated the silhouette when varying the number of clusters k. We select the number of collective footprints as specified in the previous sections and around that selected k^* we define a growing window of number of clusters. For each window we calculate the standard deviation of the silhouette coefficients in the window around k^*. In Fig. 14 we report the standard deviation of the silhouette coefficient (std) when varying the dimension of the window. A similar behavior is observed for all the datasets. A consistent variation in the standard deviation appears when the window becomes larger than 5–10 depending on the dataset. This means that if we had selected a value of $k^* \pm 5$ we would have obtained the same overall conclusions with comparable collective footprints and results.

As additional validation of the clustering results, we perform a cross-validation of the clustering process as follows. We divide the dataset into p partitions of equal size. For every iteration we name *training* T the dataset composed of $k - 1$ partitions, and *test* T' the dataset composed by the remaining partition. We run K-Means on T with the same number of clusters k detected in the analysis reported above. Then, we assign the remaining points in T' to the clusters of T, using a nearest neighbor approach with respect to the centroids (i.e., the collective footprints), obtaining in this way a labeling L_1. After that, we run K-Means on T', still with the same number of clusters k, and we obtain a labeling L_2 for the same test points. In this way, we can use this labeling as ground truth to evaluate the others and repeat the evaluation for every partition used as test set. As evaluation measures we report in Table 1 the results obtained using a set of measures that indicate the agreement of two independent label assignments strategies on the same dataset [33].

- *(NMI)* Normalized Mutual Information score: is an normalization of the Mutual Information score that measure the mutual information between two clusterings.

Table 1 Collective clustering cross-validation with respect to Normalized Mutual Information score, Adjusted Rand Index score and V-Measure score. For each measure is reported the mean and the standard deviation among five runs

Dataset	NMI	ARI	VM
Coop	0.73 ± 0.015	0.57 ± 0.044	0.72 ± 0.016
Ta-Feng	0.78 ± 0.054	0.78 ± 0.012	0.78 ± 0.055
T-Mall	0.65 ± 0.033	0.66 ± 0.077	0.65 ± 0.033

- *(ARI)* Adjusted Rand Index score: computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.
- *(VM)* V-Measure score: is the harmonic mean between homogeneity and completeness [33].

All these measure range between 0 and 1. They return a value of 1 when the two assignments L_1 and L_2 are identical, i.e. perfectly matched, while random assignments (independent labelings) have an expected value of 0. The results reported in Table 1 are the means and the standard deviations for a cross-validation with five partitions.

6 Exploitation

The temporal purchasing profiles can be employed for a wide range of applications. In the previous section we showed how different groups of behaviors and customers can be easily identified by exploiting the profiles and their collective perspectives. Both the customers themselves and the retail market chains adopting this methodology can gain useful insights from the analysis of the knowledge extracted with the temporal purchasing footprint. In the following we discuss possible applications of the methodologies proposed in this paper and benefits that can be drawn from the customers at individual level and from the retail market chain at collective level.

A first example is to provide to the customers a visualization of their typical temporal purchasing patterns to make them conscious of their behavior. The retail market chain could furnish to each customer a *personal dashboard*, like those theorized into [34, 35]. The visualization of the customer's temporal purchasing profile $\mathcal{P}_c$ might contribute to the improvement of the customer self-awareness. The self-knowledge might lead the customer to change her temporal habits with the possibility of saving money in the case a more regular behavior brings to spend a lower amount of money. Clearly, the change is possible if this is compatible with the daily time schedules and the customer desires to change. For example, a customer could discover to be a *regular* or *changing*. The customer could react by assuming a more *regular* purchasing behavior trying to adopt more *one-shop* patterns. Moreover, the customer could also *monitor* her shopping sessions understanding if her purchasing behaviors are remaining stables or over time. Besides this, since as it is shown in the paper the individual perspective alone is not sufficient to really understand *who we are*, the integration with the collective perspective provides a way to compare the individual behavior with those of the other customer, and to better understand who we are with respect to the others. For example, the collective perspective $\mathcal{P}_c^*$ could reveal to the customer c that she is very similar to the mass in terms of day and time window but with a typically *higher* amount spent. Hence, c could try to change her weekly habits in order to experience less crowded shopping sessions. Another individual service which can be derived from the temporal purchasing profile is a sort of *shopping reminder*.

Knowing the temporal purchasing habits the system can interact with your personal calendar and remind/alert you that in x days a certain amount is going to be spent. Finally, considering the typical purchasing behaviors of all the customers, an individual customer can be helped by providing *tailored recommendations for shopping time schedule* suggesting to her to anticipate or to postpone the purchasing time in order to find less queue at the supermarket checkout.

On the other hand, the retail market chain can exploit the collective footprints of all the customers $\mathcal{C}$ and a *customers segmentation* like the one showed in Sect. 5.5 to offer *personalized discounts*. For example, the retail manager could employ the collective footprints by promoting for each customer the shopping in her favorite day and time window by applying a tailored discount. Thus making the regular customers even more regular and also more profitable [3]. Furthermore, the analysis of the regular sub-sequences enables the retail manager to push customers which generally alternate one-shop weeks with no-shopping weeks in performing consecutive one-shop weeks in order to obtain special *temporal discounts*. Finally, the potential *predictive power* of the model could be capitalized by shop managers. For example, the knowledge of the collective footprints could be used to improve the overall service like re-organizing the shifts of the employees, or re-scheduling the disposal and replacement of the products on the shelves during opening hours. In addition, going back to possible personalized services offered by the shop manager, for each customer her individual temporal purchasing footprints can be used as features to improve existing recommender systems or for predicting the next time that an individual will perform a shopping session.

7 Conclusion

In this paper we have proposed an approach to extract the regularities characterizing the temporal purchasing profile of customers. We have proposed the *temporal purchasing profile* formed by the temporal purchasing footprints, and by the sequences in which these footprints take place. Then, we have described the approach to make the profiles comparable among different customers by providing the collective perspective to them. The collective perspective have enabled the analysis of many possible segmentations of the customers. The general methodological framework is applied to a case study regarding retail customers where we considered a week as temporal unit. Our extensive analysis of the case studies revealed that for most of the customers the vision of the individual profile is different from its collective perspective. Thus, using this information customers can be classified into *regular* and *changing* according to the number of behaviors needed to describe them. Moreover, we have outlined the typical patterns summarizing human behavior in scheduling the shopping time and their repetition through time.

The analytical results show that our framework enables the segmentation of customers with respect to different point of views. For example, we discovered segmentations based on: (a) the number of collective behaviors; (b) the shopping time and the amount of the expenditure; and (c) the frequency of the sequential order of specific behaviors.

Then we would like to extend the methodological framework in order to test the predictive power of the temporal profile by predicting when the next shopping will take place and how much will be the amount spent. Finally, in collaboration with UniCoop Tirreno, we would like to implement a web dashboard where a customer can provide her fidelity card number and visualize the patterns forming her temporal purchasing profile.

Acknowledgements
We thank UniCoop Tirreno and Walter Fabbri for allowing us to analyze the data and to publish the results.

Funding
This work is partially supported by the European Community's H2020 Program under the funding scheme INFRAIA-1-2014-2015: 654024 SoBigData, http://www.sobigdata.eu/.

Abbreviations
Unit U, Temporal Purchasing Unit; Footprint F, Temporal Purchasing Footprint; Sequence $\widehat{F}$, Temporal Purchasing Footprint Sequence; Profile P, Temporal Purchasing Profile; Collective Footprint C, Collective Temporal Purchasing Footprint; Collective Sequence $\widehat{C}$, Collective Temporal Purchasing Sequence; Regular Sub-Sequences R, Regular Temporal Purchasing Sub-Sequences.

Availability of data and materials
A sample of the source code implementing the proposed methodological framework, a sample of the dataset used in the case study, and additional results will be available at publication time at the following link https://github.com/riccotti/CustomerTemporalRegularities.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors conceived the framework and the analysis and the various case studies. RG and AM elaborated the theoretical models. RG and LG implemented the framework and performed the analysis of the case studies. All authors analyzed and discussed the results and contributed to the manuscript. All authors read and approved the final manuscript.

Author details
[1]KDDLab, ISTI-CNR, Pisa, Italy. [2]University of Pisa, Pisa, Italy.

Endnotes
[a] https://www.unicooptirreno.it/.
[b] http://www.bigdatalab.ac.cn/benchmark/bm/dd?data=Ta-Feng.
[c] http://102.alibaba.com/competition/addDiscovery/index.htm.
[d] All the collective footprints can be found at https://github.com/riccotti/CustomerTemporalRegularities.
[e] *Yes* indicates at least one purchase, *No* indicates no purchases.

Publisher's Note

Received: 26 July 2017 Accepted: 26 February 2018 Published online: 06 March 2018

References
1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: SIGMOD, Washington, D.C., pp 207–216
2. Kim E, Kim W, Lee Y (2003) Combination of multiple classifiers for the customer's purchase behavior prediction. Decis Support Syst 34(2):167–175
3. Guidotti R, Coscia M, Pedreschi D, Pennacchioli D (2015) Behavioral entropy and profitability in retail. In: IEEE international conference on data science and advanced analytics (DSAA). IEEE, pp 1–10
4. McDonald WJ (1994) Time use in shopping: the role of personal characteristics. J Retail 70(4):345–365
5. Lingras P, Hogo M, Snorek M, West C (2005) Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. Inf Sci 172(1–2):215–240
6. Guidotti R (2017) Personal data analytics: capturing human behavior to improve self-awareness and personal services through individual and collective knowledge
7. Chen Z-Y, Fan Z-P (2012) Distributed customer behavior prediction using multiplex data: a collaborative MK-SVM approach. Knowl-Based Syst 35:111–119
8. Yada K, Motoda H, Washio T, Miyawaki A (2004) Consumer behavior analysis by graph mining technique. In: Knowledge-based intelligent information and engineering systems, 8th international conference, KES 2004. Proceedings. Part II, pp 800–806
9. Shangguan L, Zhou Z, Zheng X, Yang L, Liu Y, Han J (2015) ShopMiner: mining customer shopping behavior in physical clothing stores with COTS RFID devices. In: Proceedings of the 13th ACM conference on embedded networked sensor systems. ACM, New York, pp 113–125
10. Luo L, Li B, Koprinska I, Berkovsky S, Chen F (2016) Discovering temporal purchase patterns with different responses to promotions. In: Proceedings of the 25th ACM international on conference on information and knowledge management. ACM, New York, pp 2197–2202
11. Krumme C, Llorente A, Cebrian M, Pentland A, Moro E (2013) The predictability of consumer visitation patterns. Sci Rep 3:1645
12. Chen M-C, Chiu A-L, Chang H-H (2005) Mining changes in customer behavior in retail marketing. Expert Syst Appl 28(4):773–781
13. Song HS, kyeong Kim J, Kim SH (2001) Mining the change of customer behavior in an Internet shopping mall. Expert Syst Appl 21(3):157–168

14. Hamuro Y, Katoh N, Edward IH, Cheung SL, Yada K (2003) Combining information fusion with string pattern analysis: a new method for predicting future purchase behavior. In: Information fusion in data mining. Springer, Berlin, pp 161–187
15. Di Clemente R, Luengo-Oroz M, Travizano M, Vaitla B, Gonzalez MC (2017) Sequence of purchases in credit card data reveal life styles in urban populations. arXiv:1703.00409
16. Padmanabhan B, Zheng Z, Kimbrough SO (2001) Personalization from incomplete data: what you don't know can hurt. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 154–163
17. Hansen T, Jensen JM, Solgaard HS (2004) Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior. Int J Inf Manag 24(6):539–550
18. Lin H-F (2007) Predicting consumer intentions to shop online: an empirical test of competing theories. Electron Commer Res Appl 6(4):433–442
19. Van den Poel D, Buckinx W (2005) Predicting online-purchasing behaviour. Eur J Oper Res 166(2):557–575
20. Koren Y (2010) Collaborative filtering with temporal dynamics. Commun ACM 53(4):89–97
21. Rendle S, Freudenthaler C, Schmidt-Thieme L (2010) Factorizing personalized Markov chains for next-basket recommendation. In: Proceedings of the 19th international conference on world wide web. ACM, New York, pp 811–820
22. Cumby C, Fano A, Ghani R, Krema M (2004) Predicting customer shopping lists from point-of-sale purchase data. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 402–409
23. Yu F, Liu Q, Wu S, Wang L, Tan T (2016) A dynamic recurrent model for next basket recommendation. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 729–732
24. Riccardo G, Giulio R, Luca P, Giannotti F, Pedreschi D (2017) Market basket prediction using user-centric temporal annotated recurring sequences. In: Data mining (ICDM), 2017 IEEE 17th international conference on. IEEE
25. Furletti B, Gabrielli L, Renso C, Rinzivillo S (2012) Identifying users profiles from mobile calls habits. In: Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, New York, pp 17–24
26. Giegerich R, Kurtz S (1997) From Ukkonen to McCreight and Weiner: a unifying view of linear-time suffix tree construction. Algorithmica 19(3):331–353
27. Tan P-N, Steinbach M, Kumar V et al (2006) Introduction to data mining. Pearson Education, Upper Saddle River
28. Zerubavel E (1989) The seven day circle: the history and meaning of the week. University of Chicago Press, Chicago
29. Shannon CE (2001) A mathematical theory of communication. Mob Comput Commun Rev 5(1):3–55
30. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6:8166
31. Guidotti R, Trasarti R, Nanni M, Giannotti F, Pedreschi D (2017) There's a path for everyone: a data-driven personal model reproducing mobility agendas. In: IEEE international conference on data science and advanced analytics (DSAA). IEEE, pp 1–10
32. Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. North-Holland, Amsterdam
33. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res 11:2837–2854
34. de Montjoye Y-A, Shmueli E, Wang SS, Pentland AS (2014) openPDS: protecting the privacy of metadata through safeanswers. PLoS ONE 9(7):e98790
35. Vescovi M, Moiso C, Pasolli M, Cordin L, Antonelli F (2015) Building an eco-system of trusted services via user control and transparency on personal data. In: Trust management IX. Springer, Berlin, pp 240–250