Recognizing Residents and Tourists with Retail Data Using Shopping Profiles

Riccardo Guidotti(✉) and Lorenzo Gabrielli

KDDLab, ISTI-CNR, Pisa, Italy
{riccardo.guidotti,lorenzo.gabrielli}@isti.cnr.it

Abstract. The huge quantity of personal data stored by service providers registering customers daily life enables the analysis of individual fingerprints characterizing the customers' behavioral profiles. We propose a methodological framework for recognizing residents, tourists and occasional shoppers among the customers of a retail market chain. We employ our recognition framework on a real massive dataset containing the shopping transactions of more than one million of customers, and we identify representative temporal shopping profiles for residents, tourists and occasional customers. Our experiments show that even though residents are about 33% of the customers they are responsible for more than 90% of the expenditure. We statistically validate the number of residents and tourists with national official statistics enabling in this way the adoption of our recognition framework for the development of novel services and analysis.

Keywords: Residents tourists classification
Customer shopping profile · Retail data · Spatio-temporal analytics
Data mining

1 Introduction

The availability of huge amount of personal data stimulates challenging questions that can be answered with data mining methodologies. Given a service providing spatio-temporal information, a question addressed in the last years for enabling the development of analysis and applications for social good pursues the goal of *recognizing* if a user of can be categorized as a *resident* or as a *tourist*. Since ground truth categories are generally missing, the resident-tourist classification is a very hard problem because supervised learning algorithms can not be applied. Instead, unsupervised approaches recognizing typical aspects of residents and visitors are used. Indeed, a number of previous works propose unsupervised methods by analyzing phone call data [5,7,12] or social media [4,10].

In this paper we propose an unsupervised methodological *framework* able to *recognize* residents and tourist by exploiting retail shopping data. Retail data is a very complex type of data containing various dimensions: *what* customers buy,

B. Guidi et al. (Eds.): GOODTECHS 2017, LNICST 233, pp. 353–363, 2018.
https://doi.org/10.1007/978-3-319-76111-4_35

when and *where* they make the purchases and which is the *amount* of the purchase. Most of the works in the literature focus on *what* customers buy [1], while just a few of them exploit the *spatio-temporal dimension* [9,15]. To the development of our *recognition framework* we define a *temporal purchasing profile* capturing the shopping habit of a customer in terms of *when* and *where* she purchases and *how much* she spends. Then, we develop *data-driven* heuristic rules for categorizing the temporal shopping profile of each customer as `resident`, `tourist`, `occasional` and `rare`. In our vision `resident` customers live in a place close to their favorite store and show a continuous, uniform and considerable presence over all the monitored periods; `tourist` customers are people staying in a place close to a store only for a limited time but repeatedly along different periods; `occasional` customers purchase on a certain store in a not very frequent way but uniformly over all the monitored periods; `rare` customers performed only a single purchase on the monitored periods.

We instantiate our recognition framework for a case study on real retail transactions. The dataset analyzed was provided by *UniCoop Tirreno* which serves more than a million of customers on the west coast of central Italy. The dataset contains retail market data from January 2007 to December 2015. After customers categorization as `resident`, `tourist`, `occasional` and `rare`, we perform a range of exploratory analysis. Our main findings reveal that `residents` are responsible for more than the 90% of the expenses even though they are only the 33% the customers population, while the presences and expenses of `tourists` become significant especially in summer months and on the stores situated in vacation places, while in the other periods and places the largest part of the revenues is generated yet by `residents`. `Occasional` customers desultorily purchases at UniCoop stores but constitute 30% of the active customers. Moreover, we statistically validate our estimations of residents and tourists by performing correlation analysis and comparing our trends against official national statistics.

Summarizing, this work advances the achievements of existing works (Sect. 2) by *(i)* formalizing and generalizing the concepts of *temporal purchasing units* and *temporal shopping profile*, *(ii)* defining data-driven heuristic rules to *categorize customers* as residents, tourists or others (Sect. 3), *(iii)* showing an instantiation of the framework on a real case study for retail data which reports interesting findings, and proves reliable quantification with respect to official statistics (Sect. 4). Finally, we illustrate which are the applications that could exploit our recognition framework, and we outline future research directions (Sect. 5).

2 Related Work

Residents and tourists recognition has been accomplished for different purposes in different data domains. In the literature there exists a large set of works classifying users as resident and tourists by using social media and phone call data.

In [4], the authors separate users in residents and visitors in order to *study migration patterns* and to analyze the spread of the influenza like illness infection

by monitoring Twitter posts. The approach developed in [10] for *investigating global mobility patterns*, assigns as the country of residence of a user the one in which she tweeted the most, while she is a visitor in all the other countries.

Other approaches, like the one presented in this paper, make the separation with the purpose of *analyzing the different characteristics* between the locals and the visitors. In order to provide space-time visual analytics of where the Seattle locals tweet and what they talk about, in [2] the authors profiles a user as local or visitor, by counting the days in which a user tweeted inside or outside Seattle. Similarly, to mine the mobility patterns of tourists in Florence, in [8] all the users who posted geolocated Flickr photos for less than 30 days in the province are considered tourists. In [14] the users are classified as resident and tourists by analyzing the number of user's active days in a specific area.

However, the methods of discrimination between residents and tourists described above are nothing more than a set of more or less sophisticated rules that are primarily dictated by common sense rather than by a data-driven finding. In addition, a simple rule-based approach to partition locals and visitors may hinder to derive some more useful information obtainable by conducting a deeper analysis on the data to derive the behavior of the users.

To overcome the aforementioned limitations, the authors of [5,6] defined how to build individual profiles based on mobile phone calls such that the profiles are able to characterize the calling behavior of a user. By analyzing these profiles three categories of users are identified: *residents*, *commuters* and *visitors*. In [7] this characterization is strengthen by aggregating users having a similar calling behavior with the *k-means* clustering algorithm [19]. The centroid of each cluster is compared with pre-defined prototypes representing the categories of interest, then, each cluster is classified by means of the associated prototype. The proposed framework aims at defining residents and tourists profiles by outlining the data-driven methodology identifying the users habits described in [5–7,12]. In [13] is proposed an alternative unsupervised procedure for estimating the tourists presence in an area over a specific period of time.

Finally, it is worth to underline that residents-tourists partitioning is useful for a myriad of studies ranging from economy to sociology to mobility. In [11] it is observed the economic impact of special events in locals and visitors. In [3] it is analyzed the relationship between place attachment and landscape value studying subgroups of resident and visitors having different levels of reliability with respect to the measures considered in the analysis. Moreover, interesting understanding of tourism lifecycle models comes from [16,18] where the impact that locals and tourists have on museums and shopping spaces are investigated.

3 Residents and Tourists Recognition

In this section we define the analytical framework able to classify the customers of a retail market chain into data-driven behavioral categories.

The type of data analyzed by the framework consists in customer's shopping sessions. A *shopping session* contains information about *(i)* which customer

made the purchase, *(ii)* all single items composing the basket, *(iii)* in which shop the transaction happened, *(iv)* the time and the *date* of the shopping session. We underline that in every retail market chain the purchases can be individually assigned to a customer only if a fidelity card is used for the purchases.

In order to recognize residents and tourists among the customers, the framework considers only the information relative to the *date* of the shopping sessions. The other information are exploited to characterize the categories recognized.

Customer Shopping Model. For each customer, we summarize a set of shopping sessions by introducing the *temporal purchasing unit* (unit in short):

Definition 1 (Temporal Purchasing Unit). *Given a period τ, we define the* temporal purchasing unit *u of customer c as a vector $u \in \mathbb{R}^M$, where M is the number of time-parts considered, and u_i contains the number of purchases performed by customer c in the i-th time-part.*

With *time-part* we indicate any aggregation of days, e.g. day, weeks, months, etc. Given a period τ, for every customer we can observe a set U_c=$\{u^{(1)}, \ldots, u^{(N)}\}$ of temporal purchase units. For example, in Fig. 1 are shown three temporal purchasing units for a customer where as period τ is used the year and as time-part the month, i.e., $M = 12$. Every line represents the number of purchases: the higher the line, the higher the number of purchases in a month.

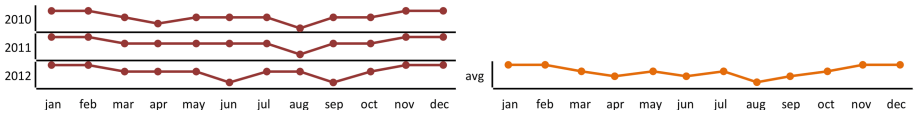

Fig. 1. *Left:* temporal purchasing units. *Right:* temporal purchasing profile. The higher the line the more purchases are done in that month.

We define the *temporal purchasing profile* (profile in short) of a customer as:

Definition 2 (Temporal Purchasing Profile). *Given the set of temporal purchase units U_c=$\{u^{(1)}, \ldots, u^{(N)}\}$ of customer c, we define the* temporal purchasing profile *$p^{(c)}$ of c as an aggregation of U_c into a vector $p^{(c)} \in \mathbb{R}^M$, where M is the number of time-parts considered, and $p_i^{(c)} = \frac{1}{N} \sum_{u^{(j)} \in U_c} u_i^{(j)}$ $\forall\, i = 1, \ldots, M$ contains the average number of purchases performed by c in the i-th time-part.*

The approach for the extraction of the profiles can be summarized as follows. The *first* step is the *construction* of the temporal purchase units from the shopping sessions considering M time-parts for each period τ. Then, the *second* step consists in aggregating the temporal purchase units $U^{(c)}$ of each customer to obtain the temporal purchasing profiles $p^{(c)}$ according to the previous definition.

Unsupervised Customers Classification. Our aim is to find prototypes describing the behavior for `residents`, `tourists`, `occasional`, and `rare`.

As preprocessing step, we label as `rare` all the customers that have performed only one purchase in the purchasing profile $p^{(c)}$, i.e., such that $\sum_i^M p_i^{(c)} = 1$.

Given a set of non rare customers $C = \{c_1, \ldots, c_N\}$ each one with her temporal purchasing profile $P = \{p^{(c_1)}, \ldots, p^{(c_N)}\}$, our objective is to find a partitioning $\mathcal{G} = \{G_1, \ldots, G_K\}$ of P into K disjoint sets. In other words, we want to cluster the customers in C according to their profile such that customers having a similar temporal shopping behavior belong to the same cluster.

To perform such partitioning we employ the well known *k-means* clustering algorithm [19]. Since in real applications the temporal purchasing profile is in fact a not sparse vector with a treatable size, we run k-means by employing the *euclidean distance* [19] as distance function between two profiles.

From the partitioning we obtain K clusters $\mathcal{G}=\{G_1, \ldots, G_K\}$ of similar customers, and K centroids $g^{(1)}, \ldots, g^{(K)}$ calculated by computing the mean for each cluster, i.e., $g_i^{(k)}=\frac{1}{|G_k|}\sum_{p^{(c)}\in G_i} p_i^{(c)}\ \forall\ i = 1, \ldots, M,\ \forall\ k = 1, \ldots, K$.

The last step consists on assigning for each *not rare* customer $c_i \in C$ a behavioral label by observing the centroid $g^{(k)}$ corresponding to the cluster G_k the customer c_i belongs to. We indicate the centroid $g^{(k)}$ of customer c with the notation $g^{(c)}$, i.e., $p^{(c)} \in G_k$. After having empirically observed various different shapes and values for the trends described by the centroids $\{g^{(1)}, \ldots, g^{(K)}\}$ we identified three categories: `resident`, `tourist` and `occasional` (see examples in Sect. 4), and we developed the following heuristic to assign the label:

- if any peak is detected in $g^{(c)}$ we label c as `tourist`, otherwise;
- if $\frac{1}{M}\sum_i^M g_i^{(c)} \leq 0.5$ we label c as `occasional`, otherwise;
- $g^{(c)}$ shows a considerable and uniform trend and we label c as `resident`.

In practice, we classify as `tourist` the customers appearing only in a precise moment with respect to the time period τ and number of time-parts M, as `occasional` those with a very low average number of purchases, and as `resident` the customers with a shopping trend which is almost constant and not negligible.

4 Case Study

In this section we present a case study on a massive real dataset of customer's shopping sessions showing the effectiveness of the proposed approach in terms of quantification and qualification of the category recognized.

Dataset. We develop our case study on the *Coop* dataset provided by *UniCoop Tirreno*[1], one of the largest Italian retail distribution company. It serves more than a million of customers covering an extensive part of the Italian territory. The stores mainly cover the west coast of central Italy. The shop distribution is not homogeneous (see Fig. 2): shops are located in a few Italian regions (Tuscany,

[1] https://www.unicooptirreno.it/.

Lazio, Campania) and therefore, the coverage of these regions is much more significant, while customers from other regions usually shop only during vacation periods in these regions. The 138 stores sell about 8,000 different items.

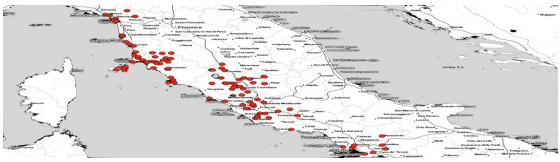

Fig. 2. Coop stores positioning on Italy west cost.

The dataset contains retail data from January 2007 to December 2015 belonging to 1,637,311 active and recognizable customers. A customer is active if she purchased during the time window, while she is recognizable if the purchase has been made using the membership card which can enable to discounts. The company is able to tie each shopping session to the card, and, for each shopping session the company knows all the information described in Sect. 3.

Customers Categorization. On top of the *Coop* dataset we instantiate the framework described in Sect. 3. First, we build the temporal purchase units from the shopping sessions with $M = 12 \times 2 = 24$ time-parts, i.e., a time-part per month considering separately weekend and weekdays, and a purchase unit for each period where $\tau = 1$ year. Given a customer c, the i-th element of the vector $u^{(j)}$ contains the number of purchases performed by c in a certain month-weekend/weekday for a certain year. Consequently, the i-th element of the temporal purchasing profiles $p^{(c)}$ contains the average number of purchases performed by c in a certain month-weekend/weekday with respect to the years c was active. Then we label as `rare` the 306,091 customers having only one purchase, and we run the *k-means* algorithm on the profiles of the remaining customers. We set the number of clusters $K = 20$ in correspondence of the position of the knee in the SSE curve [19]. Finally, we label the centroids (and the customers belonging to the clusters) according to the categorization rules defined in Sect. 3. As consequence, we highlight that small variation in the choice of K would still lead to the behaviors described in the following and delineating the four different data-driven groups (`rare`, `resident`, `tourists`, `occasional`).

We report a sample of the centroids $g^{(1)}, \ldots, g^{(K)}$ in Fig. 3. In the first row we observe a behavior typical of `resident` customers. They purchase constantly and uniformly along the whole year with a different level of repetitiveness: 3-4 times per weekday-month for the customers in *cluster 06*, 1-2 times per weekday-month for the customers in *cluster 04*. In the second row are reported centroids of customers labeled as `tourists`. They become active only in specific months: June, July, August, and they are also markedly active also on weekends. It is worth to underline that most of these tourists are in fact customers (with a

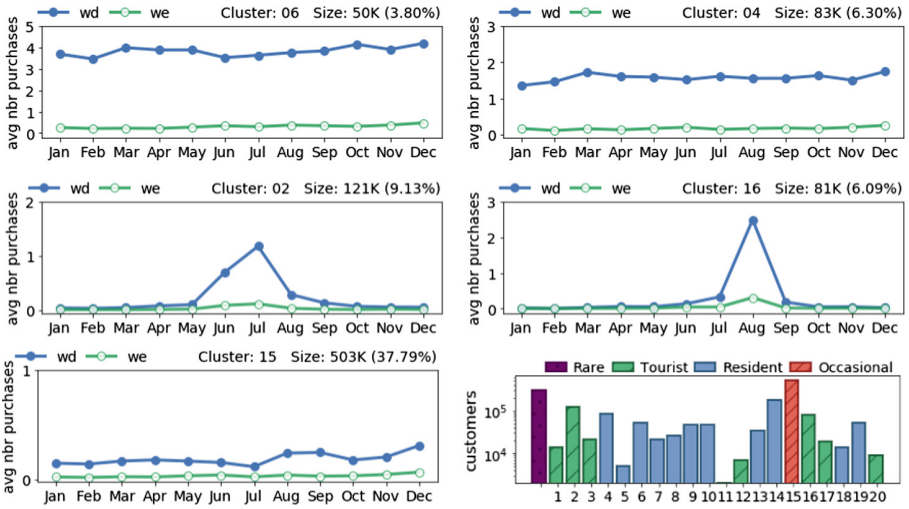

Fig. 3. Sample of clusters centroids. *First* row centroids for `residents`, *second* row centroids for `tourists`, *third* row centroid for `occasional` and clusters' sizes.

fidelity card) repeatedly appearing along different years. They are not "one-time tourists". Finally, the third row shows the centroid for `occasional` customers.

Table 1. Quantification of customers, expenditure, purchases and number of clusters.

	#customers	%customers	%expenditure	%purchases	#clusters
`residents`	554,243	33.86	91.76	93.65	11
`tourist`	273,934	16.73	02.97	02.57	8
`occasional`	503,043	30.72	05.11	03.63	1
`rare`	306,091	18.69	00.19	00.15	1

Details about the categorization are reported in Table 1. According to the Pareto principle [17], more than 90% of expenses and purchases are generated by 33% of the customers which are labeled as `residents`. Moreover, we observe that `tourists` are half of the `occasional` and generate almost the same level of expenditure and number of purchases. We remark that *all* the customers belonging to every category have a fidelity card that could have been subscribed not necessarily with *UniCoop Tirreno* but in any other *Coop* store across Italy (e.g. UniCoop Firenze, Coop Lombardia, etc.), allowing in this way every Coop (including *UniCoop Tirreno*) to trace the purchases made with the fidelity card.

Figure 4 shows how change the presences *(left)* and the expenditure *(right)* of the customers for the various categories along different months. In *summer months* we observe an increasing percentage of both `tourists` and `rare` customers, while the proportion of `residents` and `occasional` remains constant.

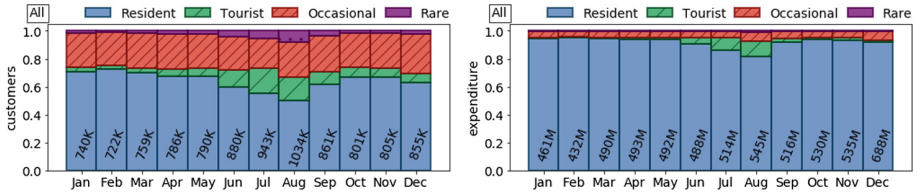

Fig. 4. Stacked bar plot with respect to relative number of customers *(left)* and relative expenditure *(right)*. The bottom numbers reports the absolute sum of the values.

This phenomenon is not due to a decrease of these last two categories, but rather to an increment on the overall number of active customers how highlighted by the absolute values on the bottom. In line with the statistics above, the strong increment in the proportion of the presences of `tourists` and `rare` customers causes only a small increment in the proportion of the expenditure. In Fig. 4 we also notice how even though `tourists` are much less than `occasional` along the year, in July and August `tourists` expends more than twice as much as `occasional` do, confirming that `tourists` are not just "occasional" sightseers but customers repeatedly following this specific behavior. Finally, it is worth to underline that August is the month with the highest number of presences but December is the month with the highest expenditure, while August is only second in terms of expenditure. Together with the categorization this indicates that in December `residents` markedly increase their expenditure while in August the expenditure increase is due to `tourists`, `occasional` and `rare` customers.

Validation of Category Quantification. We validate our estimations of `tourists` and `residents` by comparing the quantification returned by our framework using the *Coop* dataset with the official Italian data provided by *ISTAT*[2]. We remark that we are estimating if the *quantification* resulting from the unsupervised categorization is in line with the official statistics: we can not validate the *classification* because we do not have the ground truth for the customers. Moreover, not comparable methods to use as competitors are present in the state of the art. We employ two different indicators to estimate the correlations [19]. *Pearson* evaluates the linear relationship. A relationship is linear when a change in one variable is associated with a proportional change in the other variable. *Spearman* evaluates the monotonic relationship. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate.

We validate the estimation of tourists by observing the trends along different months and years. Since ISTAT does not provide the tourists for each month at regional level, we compare our estimation with the national trend (*all-istat*). How showed on Fig. 5 *(left)*, *all-istat* and the sum of the trends of Tuscany, Lazio and Campania *west-istat* are identical (Pearson and Spearman of 0.99). Figure 5 *(center)* shows a very high similarity between the `tourist`-*coop* and *all-*

[2] http://dati.istat.it/, http://demo.istat.it/.

istat tourists trends on a monthly granularity from 2010 to 2015[3]. These trends reports a Pearson of 0.91 (p-value 7.95e−29) and a Spearman of 0.90 (p-value 4.59e−28). As consequence, we can claim that *(i)* the trend of tourists quantified by the framework is a good proxy for official statistics, *(ii)* the customers in the `tourists` cluster follow the real trend of tourists and consequently they can be effectively real tourist-customers which systematically reappear across the years. Moreover, we observe in Fig. 5 (right) the `rare`-*coop* and the *all-istat* tourists trends with a Person of 0.95 (p-value 1.81e−37) and a Spearman of 0.92 (p-value 3.75e−30). This means that `rare` customers are "one-time tourists" and also in this case they are a good approximation for nowcasting official statistics.

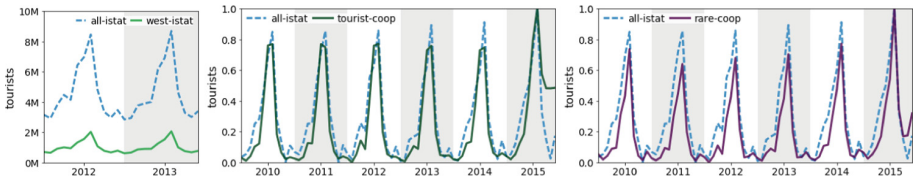

Fig. 5. Trends comparisons. ISTAT national (*all-istat*) against: (*left*) ISTAT sum of tourists in west cost (*west-istat*), (*center*) *tourist-coop*, and (*right*) *rare-coop*.

We validate the `residents` by observing the correlations with respect to different municipalities. We observe an average Pearson of 0.68 (p-value 3.97e−47) and Spearman of 0.74 (p-value 3.24e−34). For these correlations we must consider that they are tied with the adoption of *UniCoop Tirreno* stores with respect to the size of the cities and with the presence of other supermarket chains.

Single Stores Analytics. The categorization returned by the framework together with the validation obtained in the previous section, empower detailed studies of some aspects of these sets of customers from different point of views. In particular, in this section we show how the analysis of single stores reveals a different type of "audience" with respect to the four categories.

Figure 6 reports an example of two very different stores in *Livorno* and *San-Vincenzo*. In *Livorno* (upper left) we observe a negligible `rare` customers and `tourists` along over the year and a consistent percentage of `occasional` customers, while with respect to the expenditure (upper right) there is a negligible `tourists` effect with the total number of customers and expenditure remaining stable along the whole year. On the other hand, on *SanVincenzo*, the number of customers in summer months is about three times the number of customers in the other months (lower left). In addition, about 40% of the total expenditure in July and August is caused only by `tourists` (lower right). The other stores present in the dataset can be similar to *Livorno* or to *SanVincenzo*, or have a customer audience which is "in the middle" with respect to them.

[3] Missing years are not available on the ISTAT website.

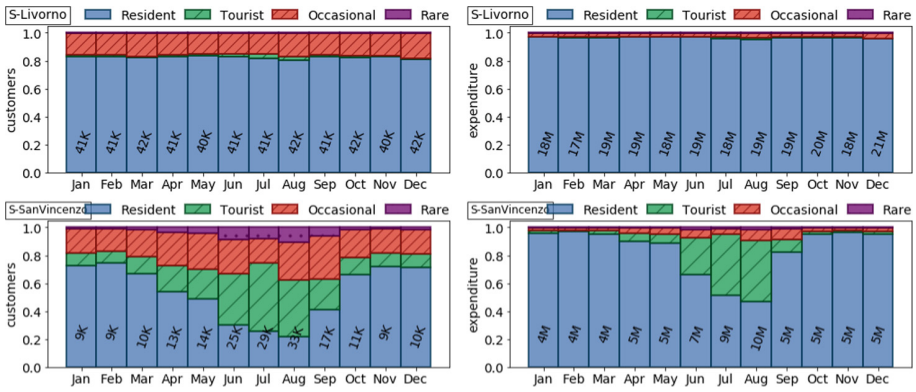

Fig. 6. Stacked bar plot with respect to relative number of customers *(left)* and relative expenditure *(right)* for two different stores: *Livorno* and *SanVincenzo*.

Therefore, the type of "audience" of a certain store can change or not along the year and, by exploiting the proposed categorization, store managers can build ad-hoc marketing strategies to maximize the profits in every month.

5 Conclusion

Nowadays various data sources, from mobile phone data to shopping transaction, are proxies for studying the user's social life. We have presented a framework for recognizing and estimating residents and tourists exploiting retail market data through the definition of temporal shopping profiles. We have applied our framework on a real massive dataset of transactions and we have observed that for Coop the residents are the customers belonging to the 20% of the Pareto principle. We have validated our estimation with national official statistics and on top of that we have showed that every store can be characterized with respect to the presences of tourists and residents and their expenditure in different months.

Our framework could be exploited in real applications for a common social good. The tourists and residents estimations can be adopted as *early approximations* of the official statistics which are provided long time after the period to which they relate. In addition, our framework enables peculiar estimations of the presences with a very fine grained spatio-temporal granularity. From the market chain perspective the customers groups can be used to better organize the product disposal and to plan targeted marketing campaigns, i.e., give to tourists special discounts on sun creams and summer products.

The framework proposed and the resulting customers categorization opens novel and interesting research directions. A very analytical step consists in better *characterizing* each group of customers to understand if they have particular preferences with respect to the time they go to shopping, to the products purchased, and to the most frequent patterns. Finally, the customers categorization

together with information about the nationality can help in better defining both the *tourists flows* but also, with respect to the residents, the *migration flows*.

Acknowledgement. This work is partially supported by the European Project *SoBigData*, 654024, http://www.sobigdata.eu. We thank UniCoop Tirreno for allowing us to analyze the data and to publish the results.

References

1. Agrawal, R., et al.: Mining association rules between sets of items in large databases. In: ACM Sigmod Record, vol. 22, pp. 207–216. ACM (1993)
2. Andrienko, G., et al.: Thematic patterns in georeferenced tweets through space-time visual analytics. Comput. Sci. Eng. **15**(3), 72–82 (2013)
3. Brown, G., et al.: The relationship between place attachment and landscape values: toward mapping place attachment. Appl. Geogr. **27**(2), 89–111 (2007)
4. Cao, G., et al.: A scalable framework for spatiotemporal analysis of location-based social media data. Comput. Environ. Urban Syst. **51**, 70–82 (2015)
5. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Identifying users profiles from mobile calls habits. In: KDD Workshop, pp. 17–24. ACM (2012)
6. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Analysis of GSM calls data for understanding user mobility behavior. In: Big Data, pp. 550–555. IEEE (2013)
7. Gabrielli, L., Furletti, B., Trasarti, R., Giannotti, F., Pedreschi, D.: City users' classification with mobile phone data. In: Big Data, pp. 1007–1012. IEEE (2015)
8. Girardin, F., et al.: Understanding of tourist dynamics from explicitly disclosed location information. In: Symposium on LBS and Telecartography, vol. 58 (2007)
9. Guidotti, R., Coscia, M., Pedreschi, D., Pennacchioli, D.: Behavioral entropy and profitability in retail. In: DSAA, pp. 1–10. IEEE (2015)
10. Hawelka, B., et al.: Geo-located twitter as proxy for global mobility patterns. Cartogr. Geogr. Inf. Sci. **41**(3), 260–271 (2014)
11. Long, P.T., Perdue, R.R.: The economic impact of rural festivals and special events. J. Travel Res. **28**(4), 10–14 (1990)
12. Lulli, A., et al.: Improving population estimation from mobile calls: a clustering approach. In: ISCC, pp. 1097–1102. IEEE (2016)
13. Mamei, M., et al.: Analysis of tourist activity from cellular data. In: POSTER (2017)
14. Manca, M., et al.: Using social media to characterize urban mobility patterns: state-of-the-art survey and case-study. Online Soc. Netw. Media **1**, 56–69 (2017)
15. McDonald, W.J.: Time use in shopping: the role of personal characteristics. J. Retail. **70**(4), 345–365 (1994)
16. Moreno Gil, S., Ritchie, J.B.: Understanding the museum image formation process: a comparison of residents and tourists. JTR **47**(4), 480–493 (2009)
17. Pareto, V.: Manual of Political Economy. Macmillan, London (1971)
18. Snepenger, D.J., Murphy, L., OConnell, R., Gregg, E.: Tourists and residents use of a shopping space. Ann. Tour. Res. **30**(3), 567–580 (2003)
19. Tan, P.-N., et al.: Introduction to Data Mining. Pearson Education India, Noida (2006)