Measuring Immigrants Adoption of Natives Shopping Consumption with Machine Learning

Riccardo Guidotti[1]✉, Mirco Nanni[2], Fosca Giannotti[2], Dino Pedreschi[1], Simone Bertoli[3], Biagio Speciale[4], and Hillel Rapoport[4]

[1] University of Pisa, Italy, {name.surname}@unipi.it
[2] ISTI-CNR, Pisa, Italy, {name.surname}@isti.cnr.it
[3] Université Clermont Auvergne, CNRS, CERDI, France, simone.bertoli@uca.fr
[4] Paris School of Economics, France, {name.surname}@univ-paris1.fr

Abstract. "Tell me what you eat and I will tell you what you are". Jean Anthelme Brillat-Savarin was among the firsts to recognize the relationship between identity and food consumption. Food adoption choices are much less exposed to external judgment and social pressure than other individual behaviours, and can be observed over a long period. That makes them an interesting basis for, among other applications, studying the integration of immigrants from a food consumption viewpoint. Indeed, in this work we analyze immigrants' food consumption from shopping retail data for understanding if and how it converges towards those of natives. As core contribution of our proposal, we define a score of adoption of natives' consumption habits by an individual as the probability of being recognized as a native from a machine learning classifier, thus adopting a completely data-driven approach. We measure the immigrant's adoption of natives' consumption behavior over a long time, and we identify different trends. A case study on real data of a large nation-wide supermarket chain reveals that we can distinguish five main different groups of immigrants depending on their trends of native consumption adoption.

Keywords: Immigrants Shopping Consumption, Human Migration Analysis, Machine-Learning-Based Measure, Adoption Trends, Integration

1 Introduction

Moving across borders exposes people to the norms that are adopted by the natives in their countries of destination. As time passes by, immigrants might progressively adopt these norms [18, 37]. Adoption is a dynamic process that requires time. A crucial choice that all immigrants must make is whether to adopt the habits of their host society [11]. Examples of this choice are decisions concerning ethnic-sounding vs. native-sounding names [1, 20], whether to change the surname [6], the language spoken [4], and whether to marry with people of the same ethnic group [34]. All these measures reflect choices that are easily observed by one's peers and, thus, potentially exposed to social sanctions and not fully reflecting one's own preferences. In addition, these measures are usually observed at one point in time, while integration is an inherently dynamic phenomenon.

Jean Anthelme Brillat-Savarin, a lawyer, politician, and famous gastronome was among the firsts to recognize the relationship between identity and food consumption. In his book [12], he wrote his well-known aphorism: "Tell me what you eat and I will tell you what you are". Following this intuition, in this paper we rely on immigrants' detailed food consumption choices from shopping retail data, over a long time after immigration. We highlight that food adoption choices are not readily observed or inferred (unless the retail stores under study are very biased, e.g. very cheap or specialized in ethnic food, which is not the case of the data used in our study), and hence less exposed to peer pressure [3]. Moreover, a measure of adoption based on food shopping consumption has a dynamic nature and can be observed several times in the host country.

We use information from retail data and from the country of birth of the customers to develop a measure evaluating whether immigrants adopt consumption habits similar to those of natives. We name this measure Native Consumption Adoption (NCA). Given a temporal granularity, we model the shopping behavior of a customer as a vector containing information about shopping habits with respect to different food categories. Then, we exploit the knowledge of the country of birth, and we build a machine learning classifier able to distinguish between natives and immigrants. In this work, instead of measuring the compliance of customers to a pre-defined model of *native behaviour*, we propose to compute the NCA as the probability of being a native that the classifier learnt from training data assigns to each customer. We adopted classification methods instead of simpler solutions based on distances between vector representations of customers' purchases because preliminary experiments showed that the latter are unable to separate natives and immigrants in reasonable ways. Finally, we estimate the NCA for the immigrants over time, and we identify different trends of adoptions of the native' shopping consumption. We name TINCA the proposed methodology aimed at discovering Trends of Immigrants' Native Consumption Adoption.

We investigate whether immigrants' consumption in terms of NCA converges towards those of natives on a case study over real data describing the purchases of the customers (immigrants and natives) of a large Italian supermarket chain between 2008 and 2015. Experimental results reveal five different groups of immigrants depending on their trends of NCA, namely increasing and stable native-refusers, strong and weak native-adopters, and native-like customers. The proposed methodology and the results that can be derived from its application can be of interest to multiple subjects. For instance, social scientists analyzing the process of integration in the host society, or collaborating with institutions and organizations responsible for the integration of immigrants. In addition, the customer management of retail sales companies can benefit from the proposed methodology and the analysis that can be done on top of the outcomes returned.

The rest of the paper is structured as follows. Section 2 summarizes related work on migration. In Section 3, we formalize the problem, while in Section 4 we define the methodology proposed to solve it. Section 5 presents experiments in the form of a case study in which we employ the proposed methodology. Finally, Section 6 concludes the paper and illustrates future research directions.

2 Related Work

Human migration has been a constant of human history, from the earliest ages until now. The study of migration spans various research fields, including anthropology, sociology, economics, statistics, and, more recently, physics and computer science. In this paper, we focus on studying the *"stay"* phase of migration [36], i.e., immigrant integration, and changes in life and habits.

The complexity of the immigrants' choice, whether to assimilate or not in terms of food consumption, has been widely analyzed in economics. In [35] it is studied how Bengali Indian households in the US converged towards the host country's norm for their breakfast eaten at home. Economists recently started to consider the analysis of immigrants' consumption as a subject ripe for empirical investigation [10] and recognized the importance of understanding immigrants' consumption behavior to assess the labor market consequences of immigration [19]. The differences in food preferences across social groups are analyzed in [7], and it is shown that internal migrants in India bring their origin-state food preferences with them during migration. In [31], the authors find that immigrants' expenditure shares for different types of food in the nineteenth-century are predicted by past relative prices in their countries of origin. Similarly, [13] finds that the current purchases of consumers who migrate across states in the US depend on both where they live currently, and where they lived in the past. Our work is also related to the economics literature on the heterogeneity in cultural traits. In [5] it is shown that in the last decades economic convergence across European countries was not accompanied by cultural convergence. The authors of [9] rely on machine learning to measure cultural distance in terms of how predictable group membership is from media consumption, consumer behavior, time use and social attitudes. They show that cultural distances in the US have remained broadly constant over time, with few exceptions.

As previously discussed, migrant integration is generally measured through indicators related to the labor market and economic status. These statistics are available with low resolution and not for all countries. A new direction is that of observing integration through big data analysis. For instance, the analysis of online social networks can allow evaluating the level of adoption of a culture. There is a vast literature regarding immigration and social networks, especially Twitter. Most of the studies start from the language in which a post is written [30]. In [29] it is proposed an approach to detect English linguistic variation and quantify its significance among geographic regions, while in [33] geolocated tweets are exploited to analyze the language diversity over different countries. Also, the topological graph-composition of social networks is analyzed for studying migration. In [27], community-centric metrics are used to study cultural assimilation as a function of the number of social ties between migrant communities and natives using the set of friendship links extracted from Facebook. In [30], a bipartite graph structure, connecting tweet languages and cities, is used for studying again cultural assimilation but from the spatial segregation

point of view. Finally, call data records (CDR) from the D4R challenge[5] are exploited in [8] to observe that integration seems to increase in time for refugees, and also that the presence of refugees influences the house market in Turkey, decreasing housing prices. The discussion above is not intended to be a complete review of data-driven analytical methods for studying immigrants integration. For a more comprehensive review, refer to [14, 36].

In addition, to complement the perspectives of works in economics, sociology and computer science, our paper is, to the best of our knowledge, the first data-driven analytical method not using twitter or CDR for estimating immigrants integration in terms of the adoption of native consumption trends. Furthermore, it is also the first work not using retail data with the final purpose of customer profiling [15, 21, 24], customer segmentation [22, 23, 39], or pattern discovery [2, 26]. We stress that, like in [17, 32], the analytical process proposed in this paper takes into account the evolution of a customer and the changes in her behavior. The work in [17] exploits behavioral and demographic variables, and a transaction database for designing measures of similarity and unexpectedness for mining change patterns to analyze the degree of resemblance at different time periods. In [32] it is proposed a customer segmentation model that allows to track the evolution of a customer, including the splitting and merging of customer groups and allowing to observe how groups evolve and how individuals shift across groups. Similar aspects are addressed in the proposed approach.

3 Problem Formulation

In this section we define the context and the problem we want to solve.

Let $P = \{p_1 \dots p_q\}$ be a set of q products, we define a basket b (or transaction) as a subset of products such that $\emptyset \subset b \subseteq P$. Given a customer i, we name $B_i^{(t)} = \langle b_1 \dots b_n \rangle$ the set of temporally ordered n baskets purchased by i in the time interval t. t can represent a week, a month, etc., and therefore the baskets in $B_i^{(t)}$ correspond to those purchased in a certain week, month, etc., respectively. With $D(b)$, we refer to the day in which the basket b is purchased. We highlight that $\forall b \in B_i^{(t)}$, $D(b)$ lies within the time interval t. Given a product p, the function $E(p)$ returns the expenditure[6] required to purchase p.

Given a finite set of countries $\mathcal{C} = \{\mathit{China}, \mathit{France}, \mathit{Italy}, \dots\}$, we define C as the function that returns the country of birth of a customer i, i.e., $C(i) \in \mathcal{C}$. Given a reference country $r \in \mathcal{C}$, and a customer i, we can distinguish whether i is a *native* customer, i.e., $C(i) = r$, or i is a *foreign* customer, i.e., $C(i) \neq r$.

Let $\mathcal{B}^{(t)} = \{B_1^{(t)}, \dots, B_N^{(t)}\}$ be the set of baskets purchased by N different customers at time interval t. Given the set of sets of baskets $\{\mathcal{B}^{(t_1)}, \dots, \mathcal{B}^{(t_k)}\}$ purchased between time intervals t_1 and t_k, and a reference country r, our goal is to measure the degree with which foreign customers adopt a shopping behavior similar to that one of native customers along time.

[5] Data for refugees of Turkey `http://d4r.turktelekom.com.tr/`.

[6] We assume that the expenditure function E also accounts for the quantity.

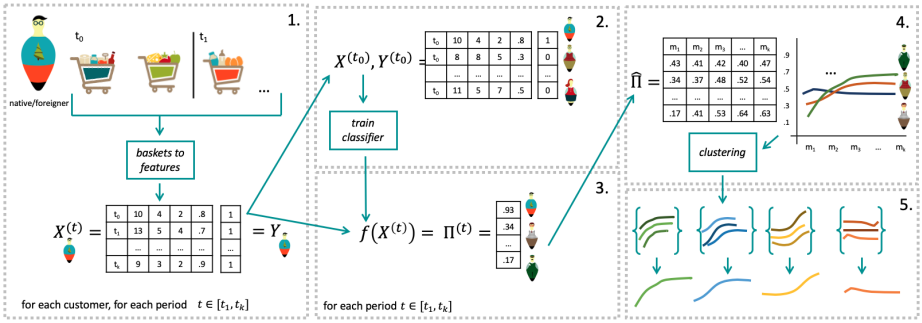

Fig. 1. TINCA analytic framework workflow. *1.* The baskets of each customer are turned into features describing the customer shopping behavior in various time periods. *2.* The feature matrix X describing the customers and their country of birth Y are used to train a machine learning classifier f. *3.* The classifier f is used to estimate the level of NCA for the various time periods $\Pi^{(t)} = f(X^{(t)}) \forall t \in [t_1, t_k]$. *4.* The NCA values of foreign customers $\hat{\Pi}$ is arranged in a matrix with respect to the time at which the customer started to purchase and a clustering algorithm is used to extract groups of customers with similar trends of NCA. *5.* From each group is extracted a representative trend of adoption/rejection of the shopping behavior of natives.

4 Trends of Immigrants Native Consumption Adoption

In this section, we describe our proposal for measuring immigrants' level of adoption of native shopping consumption over time. We name it TINCA as its aim is to discover Trends of Immigrants' Native Consumption Adoption. We indicate the Native Consumption Adoption with Π. Given the set of sets $\{\mathcal{B}^{(t_1)}, \ldots, \mathcal{B}^{(t_k)}\}$, a reference country r, a prediction function f, TINCA consists of the following 5 steps, also illustrated in Figure 1:

1. **Models** in the matrix $X^{(t)}$ the customers' shopping behavior for each time interval $t \in \{t_1, \ldots, t_k\}$, such that every row $X_i^{(t)}$ is a vector of features describing the purchases of customer i;
2. **Learns** the machine learning classifier prediction function f on the training dataset $\langle X^{(t_0)}, Y^{(t_0)} \rangle$ where $Y^{(t_0)}$ indicates if the customer i is native or not.
3. **Measures** the NCA as $\Pi^{(t)} = f(X^{(t)})$, i.e., as the probability of being classified *native* by the classifier f for each time interval t between t_1 and t_k;
4. **Groups** immigrants with respect to the fluctuations of their NCA in Π over time periods, obtaining clusters of non natives with a similar NCA evolution;
5. **Extracts** from each group a representative trend of adoption/rejection.

In the following we provide the details of each step and explain the notation adopted in the above pseudo-code.

4.1 Modeling Customer Shopping Behavior

The first step of TINCA aims to model the shopping behavior of customer i at time interval t, represented by $B_i^{(t)}$ (shortened as B, where clear from the context). We design a vector of features $X_i^{(t)}$ extracted from $B_i^{(t)}$, that capture several different aspects of individual purchase activity:

- number of products purchased: $NP = \sum_{b \in B} |b|$
- number of distinct products purchased: $DP = \left|\bigcup_{b \in B} b\right|$
- number of baskets: $NB = |B|$
- average basket length[7]: $AL = \frac{NP}{NB}$
- total expenditure: $TE = \sum_{b \in B} \sum_{p \in b} E(p)$
- average expenditure: $AE = \frac{TE}{NB}$
- average period between purchases: $AP = \frac{D(b_{NB}) - D(b_1)}{NB - 1}$
- number of purchases of product p: $NB_p = \sum_{b \in B} \mathbb{1}_{p \in b}$
- total expenditure of product p: $TE_p = \sum_{b \in B} \mathbb{1}_{p \in b} E(p)$
- average period of product p: $AP_p = \frac{\max_{b \in B|_p} D(b) - \min_{b \in B|_p} D(b)}{(NB_p - 1)}$

where the operator $\mathbb{1}_{cond}$ returns one if the condition *cond* is verified, zero otherwise; and the operator $|_p$ applied to B returns the subset of baskets containing product p, i.e., $B|_p = \{b \in B \mid p \in b\}$. The features adopted are commonly used in various work in the literature analyzing transactional data for recommendation, classification or analysis purposes [16, 17]. The first seven features capture *general* shopping behavior. Besides the total and average indicators of quantities NB, DP, NP, TE, AL and AE, we highlight how AP captures the average *frequency* of the period within which a customer makes a purchase. This is an important temporal indicator that can vary a lot even for customers having similar shopping habits in terms of items purchased. The other features, namely NB_p, TE_p, AP_p, capture the *specific* shopping behavior for each one of the various products $p \in P$. More precisely, since working with single products might lead to very sparse data, we suggest to group them into product categories collecting items of similar type, e.g., bread, pasta, tomatoes, milk, etc. This abstraction has also the effect of reducing possible effects of the market dynamics, where a product might be easily replaced by others in a short time. Details on product categories adopted and data dimensionality for our case study are provided in Section 5. An important aspect to remark is that all the features we are adopting have a clear meaning and are easily interpretable, therefore they can be exploited for further analyses.

Given $\mathcal{B}^{(t)}$, we name $X^{(t)}$ the feature matrix modeling in each row $X_i^{(t)}$ the shopping behavior of a customer among those in $\mathcal{B}^{(t)}$, where $X_i^{(t)} = \langle NP, DP, NB, AL, TE, AE, AP, NB_1, TE_1, AP_1, \ldots, NB_n, TE_n, AP_n \rangle$ for $n = |P|$. The features in $X_i^{(t)}$ describe the food consumption of customer i at time t.

[7] We consider also features derived from others, like AL and AE, since they might capture different aspects of the customer shopping behavior. Where needed, redundant features can be removed at the preprocessing stage preceding the training phase of the machine learning classifier.

4.2 Learning and Measuring Native Consumption Adoption (NCA)

The second and third steps of TINCA consist in training a machine learning classifier f for estimating the degree of adoption of natives' shopping habits by foreign customers in a specific time interval t. To this aim, our definition of the Native Consumption Adoption (NCA) score starts from the observation that customers with shopping behaviours very different from natives' will be recognized very easily as non-natives by a classifier, with a high confidence; on the opposite, the more similar to natives is the purchase behaviour, the lower will be the confidence, till the point where the customer will be recognized as native, again with confidence dependent on the closeness to native behaviours. Based on these observations, in this work we compute NCA as the probability of being recognized as a native returned by the classifier [38]. This methodology advances state-of-the-art since, to the best of our knowledge, it is the first attempt to adopt machine learning classification to implicitly build a completely data-driven model of what it means to purchase like a native, thus not depending on preconceived hypotheses or handcrafted rules.

The simplest way of implementing the principles introduced above would consist in comparing the purchase features of a customer with those of natives, for example by applying a basic k-NN classification with an Euclidean distance, working either on the raw training set or on prototypes extracted beforehand. However, preliminary experiments showed that there is not a clear distinction between the overall features distribution of natives and foreign customers (see Section 5.1) and that solutions based on simple combinations of features fail (e.g. linear regression, shown in Section 5.2), thus calling for more complex analyses of the features and requiring more sophisticated machine learning classifiers.

More in detail, let $f : X \rightarrow [0, 1]$ be the prediction function of a machine learning classifier, that takes as input the features modeling the customer shopping behavior, and returns the probability that this customer is recognized as a native. Thus, given the model of shopping behavior X_i of customer i, we indicate with $\Pi_i = f(X_i)$ her NCA, and i is recognized as a foreign when $\Pi_i \approx 0$, while i is recognized as a native when $\Pi_i \approx 1$. We obtain the prediction function f from the learning function $l : X \times Y \rightarrow f$ of a machine learning classifier. The learning function l takes as input a set of N models of shopping behavior, i.e., the feature matrix X describing the customers' shopping behaviors, and a vector Y specifying if each customer is native or not with respect to a reference country r, i.e., $Y = \langle \mathbb{1}_{C(i)=r} \rangle_{i=1,\ldots,N}$; and returns the prediction function f. Coherently with the objective of our study, the prediction function is learnt from a dataset $\langle X^{(t_0)}, Y^{(t_0)} \rangle$ containing, on one hand, vectors of features $X_i^{(t_0)}$ modeling the initial purchases of immigrants (i.e. features related to their early purchase history, which is less likely to be affected by their possible integration); on the other hand, vectors of features $X_i^{(t_0)}$ modeling the shopping behavior of natives[8].

[8] Notice that we are implicitly assuming that the food consumption habits of natives do not change over time. While not true in general, we empirically observed that it

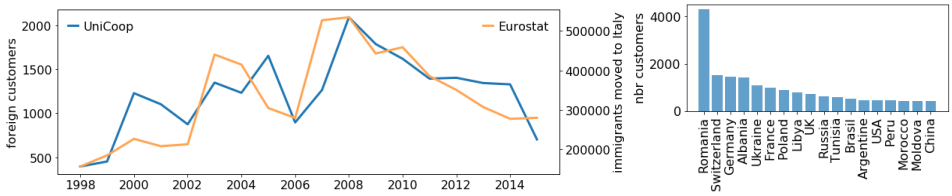

Fig. 2. *(Left)*: Eurostat trend of immigrants moved to Italy vs trend of foreign customers of UniCoop. *(Right)*: 18 most represented countries in the dataset.

	nbr baskets	tot exp	avg basket len	avg exp	avg freq
natives	3.95 ± 0.13	180.63 ± 13.67	8.21 ± 0.30	40.55 ± 2.42	3.29 ± 0.09
immigrants	4.39 ± 0.16	175.73 ± 12.59	7.26 ± 0.22	33.95 ± 1.93	3.28 ± 0.13

Table 1. Number of baskets, total expenditure, average basket length, average expenditure, and average frequency means and standard deviations for natives and immigrants.

4.3 Grouping and Monitoring NCA Trends of Foreign Customers

The objective of steps four and five is to monitor how the NCA values of foreign customers (who are the focus of our study) evolve in time. We implement these steps by exploiting a centroid-based clustering [38] that simultaneously *groups* immigrants with respect to their NCA trends and computes *representative trends* as centroids. Each customer i might start her purchase history at a time $S(i)$ that differs from other customers. In particular, while some of them start with our observation period (i.e., $S(i) = t_1$), others might start later (i.e., $t_1 < S(i) \leq t_k$). In order to take that into consideration, and to perform unbiased comparisons of the trends, before clustering the NCA trends we align them with respect to $S(i)$. More formally, let $\Pi^{(t_1)}, \ldots, \Pi^{(t_k)}$ be the NCA for all the customers for each time interval t between t_1 and t_k. We create a matrix $\hat{\Pi}$ where each row $\hat{\Pi}_i$ corresponds to a foreign customer i, i.e., $C(i) \neq r$, and each column j corresponds to the j-th time period of the customer starting from her $S(i)$, i.e., $t = S(i)+j-1$. The trend analysis focuses on the first m time periods of $\hat{\Pi}$, where m is a parameter. The missing values in the NCA sequences (either because the customer has less than m time periods, or because of gaps) are replaced with the most recent available NCA value. Thus, through the matrix $\hat{\Pi}$ we can compare the NCA trends of foreign customers. Indeed, $\hat{\Pi}$ is used as input for the clustering algorithm and for extracting the representative trends for the various clusters. In particular, the representative trends are cleaned and studied exploiting methods from time series analysis [28].

holds for the vast majority of customers in our data. Studying natives' evolution in time is part of our future works.

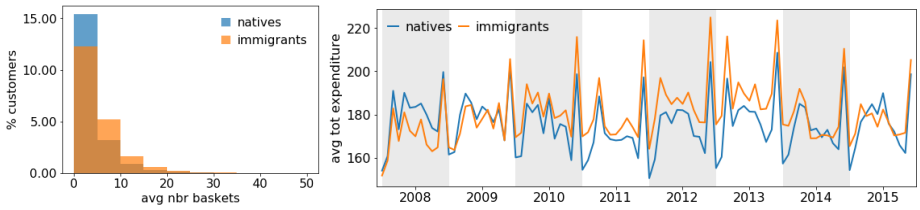

Fig. 3. *(Left)*: distributions of the average number of baskets for natives and immigrants. *(Right)*: trends of average total expenditure for natives and immigrants.

5 Experiments

In order to use TINCA[9] to study if and how immigrants' consumption converges towards those of natives, we exploited the purchases of food products made by a high number of documented immigrants (and natives) in *UniCoop Tirreno*[10], a large Italian supermarket chain. Thus, we set *Italy* as reference country r. Customers are provided with a loyalty card which allows linking different shopping sessions. We consider only customers labeled as *resident* according to [22], i.e., the customers who stably perform a minimum number of purchases over the years. In particular, we analyzed about 30 millions of transactions made by about 160k customers in 128 different shops, over the years 2008-2015. In our experiments we cover a month with each time interval t, while we use a set of products P with $|P| = 100$ distinct products where each product refer to a certain group of semantically similar items, i.e., "milk" identifies all the different types of milk[11]. Thus, we have a dimensionality of 307 for the feature matrix X describing the customers shopping behavior. This choice with time intervals of one month, and with the selected machine learning classifier avoids any issue related with data sparsity for the classification task and the other steps.

5.1 Data Analysis

Before discussing the results obtained with TINCA, we briefly present the *UniCoop* dataset and the reasons for using it to study immigrants' adoption of native consumption. Figure 2 *(left)* shows the trend of immigrants moved to Italy according to Eurostat[12], and the trend of memberships with UniCoop Tirreno of foreign-born customers. This high correlation[13] confirms the suitability of the

[9] The source code of TINCA is available here: `https://github.com/riccotti/TINCA`.
[10] `https://www.unicooptirreno.it/`, data: `https://sobigdata.d4science.org/`.
[11] The 100 product groups are available in the shared repository. The grouping was performed manually to respect the implicit semantic meaning. Each product models on average 1.9 ± 2.0 categories of items of the UniCoop dataset. The largest product groups are those modeling "bread", "fish", and "vegetables".
[12] Eurostat data: `https://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics`.
[13] Pearson of 0.75 and Spearman of 0.78 in both cases with p-value < 0.0005.

	F1-score	Recall	Precision
RF	.37 ± .02	.58 ± .10	.29 ± .05
DT	.35 ± .01	.55 ± .09	.26 ± .01
LR	.10 ± .02	.05 ± .01	.70 ± .03

Fig. 4. Average performance and its variability on the twelve training datasets for the class not native.

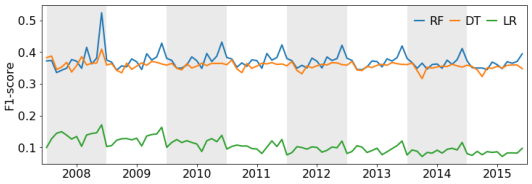

Fig. 5. F1-score along months when the classifiers are applied to estimate the values of NCA.

UniCoop dataset for this kind of study. We report in Figure 2 *(right)* the 18 most represented countries out of the 158 present in the dataset.

Table 1 and Figure 3 highlight how it is not trivial to distinguish between natives, i.e., Italians in this case, and immigrants in the *UniCoop* dataset. Table 1 reports means and std. dev. for some measures which are contained in the vector of features describing a customer: these values are quite similar and do not separate natives and immigrants due to the variability in consumption behavior. Figure 3 shows the distributions of the average number of baskets *(left)*, and the trends of average total expenditure *(right)*. In both cases we observe that natives and immigrants have rather similar distributions, without a strong separation.

5.2 Machine Learning and Classification Performance Analysis

In this section, we provide details for the training of the machine learning classifiers, and we analyze their performance. We remark that our final objective is to obtain a model with sufficient discrimination power to identify and monitor in time native-looking vs. immigrant-looking behaviours. While, obviously, the more accurate are the models, the better, our emphasis is on capturing the non-native class, which is fundamental for the analysis. As it will be shown, this is a difficult classification problem.

We account for seasonality [23] by training a machine learning classifier f for each month of the first year, i.e., $f^{(1)}, f^{(2)}, \ldots, f^{(12)}$ for January, February, etc., respectively. As the overall dataset is highly imbalanced, with only 3% of immigrants, we adopt an undersampling strategy to reach a better equilibrium and mitigate the well-known issues of classification with small minority classes. In particular, as learning datasets $X^{(1)}, X^{(2)}, \ldots, X^{(12)}$ we consider vectors of features selected as follows: *(i)* for all immigrants, consider the vectors modeling the first year of purchases; and *(ii)* take a random selection of vectors for natives. The result of the random undersampling is a set of training datasets (one per month) that contain 20% of immigrants and 80% of natives. Experiments are performed using a 10-fold cross validation approach for evaluation purposes and for parameters estimation over each month. Then, for each month the best classification models $f^{(1)}, f^{(2)}, \ldots, f^{(12)}$ are trained on the the entire learning datasets $X^{(1)}, X^{(2)}, \ldots, X^{(12)}$ of the first year. Finally, they are tested and employed to estimate the NCA on all the subsequent years.

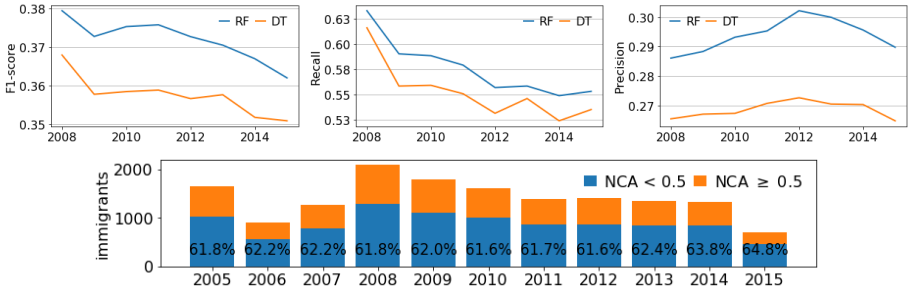

Fig. 6. *Top*: Average trends of performance for DTs and RFs along years. *Bottom*: Number of immigrants members from 2005 to 2015 divided by NCA. The percentages specify the portion of immigrants classified as *not natives*, i.e., with NCA < 0.5.

Among existing machine learning classifiers [38], we test Decision Trees (DT) and Random Forests (RF), because of their (partial) interpretability, since understanding the reasons behind the classification is one of our side goals [25]. We optimize the parameters selection search for maximizing the F1-score [38] with respect to the not native class $C(r) \neq r$. Also, we adopt Linear Regressors (LR) as a baseline. In Figure 4, we report the average performance over the twelve months on the training data with respect to the class not native, i.e., immigrant. We observe that the baseline (LR) yields a relatively high precision, but a very poor recall, basically showing its inability in most cases to capture our class of interest, and also leading to a very low F1-score. DT and RF both largely outperform the baseline in terms of F1-score, with RF showing slightly better values than DT on all the measures. Hence, RF results to provide the best trade-off in performances, with a high recall (it captures 60% of non-native customers) and a lower yet acceptable precision. Further improving the performances is not the focus of this study, and will be the subject of future work.

Figure 5 confirms these intuitions showing the trends of F1-score across the various months when the classifiers $f^{(1)}, f^{(2)}, \ldots, f^{(12)}$ are applied to estimate the values of NCA. From Figure 5 it is also clear that RFs, like in [9], overcome other classifiers and better fit our purposes. Moreover, we observe how, in certain months, it seems easier to distinguish natives from immigrants. Finally, Figure 6 *(top)* shows the average F1-score, precision, and recall of DTs and RFs over the years. Besides remarking that RFs perform better than DTs, this supports the correctness of our intuition of monitoring changes in purchase habits. Indeed, we notice a drop in the F1-score due to a drop in the recall, while precision remains stable, with just a slight improvement. The decrease of the recall from 0.64 to 0.55 indicates that, as time passes, RFs decrease their power in recognizing immigrants based on the patterns learnt at the beginning of the observation period. The most natural explanation of this effect is that immigrants changed their shopping behavior, and started adopting consumptions closer to those of natives. We can discard the possibility that immigrants arriving in more recent years have preferences closer to natives than earlier ones, by observing Figure 6

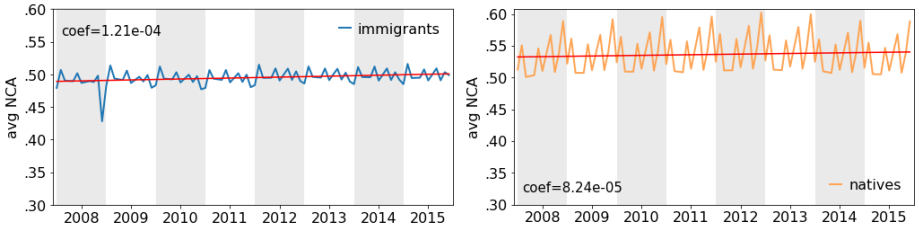

Fig. 7. Average NCA for immigrants *(left)* and natives *(right)*. In red are reported the linear regression trends and their coefficients.

(bottom): more than 60% of immigrants becoming customers in recent years have in their first year of membership an NCA lower than 0.5.

5.3 Trends of Native Consumption Adoption Analysis

In this section, we analyze the trends of NCA observing their evolution across time. Before presenting the results obtained with TINCA relative to groups of immigrants having similar trends of NCA, we show that indeed NCA is different for immigrants and natives. In Figure 7 we report the average NCA for immigrants (left) and natives (right). As expected, the NCA for immigrants is lower than 0.5, and for natives is higher than 0.5. We highlight that the limited difference between the NCA for immigrants and natives is due to the not easy task resolved by the machine learning classifier. Figure 7 also highlights in red the linear regressions (and their coefficients) showing that the average NCA of immigrants tends to 0.5, making immigrants more and more indistinguishable from natives.

We implement the centroid-based clustering of TINCA by exploiting the K-Means algorithm [38] to simultaneously group immigrants having similar NCA and to extract a representative trend of adoption/rejection. As already mentioned, we align trends of NCA of different customers with respect to the time $S(i)$ at which customer i started to perform purchases, that is, the number of *months since membership subscription*. We fill gaps in trends of NCA due to months without purchases by linear interpolation. As distance function, we rely on the Euclidean distance[14]. We apply K-Means only for immigrants, and we run K-means ranging the number of clusters from 2 to 150. We select 18 as the best number by observing the knee in the Sum of Squared Error (SSE) curve reported in Figure 8 *(bottom right)*. Finally, in order to remove seasonality and noise, the representative NCA trend of each group is computed through time series decomposition [28] applied to the centroid of every cluster.

In the following, we analyze the 18 trends of NCA (Figure 8) aggregated into five groups according to the trend's coefficient and its distance from the classifi-

[14] We leave to future works the study of the effect of other specific functions for time series clustering like dynamic time warping.

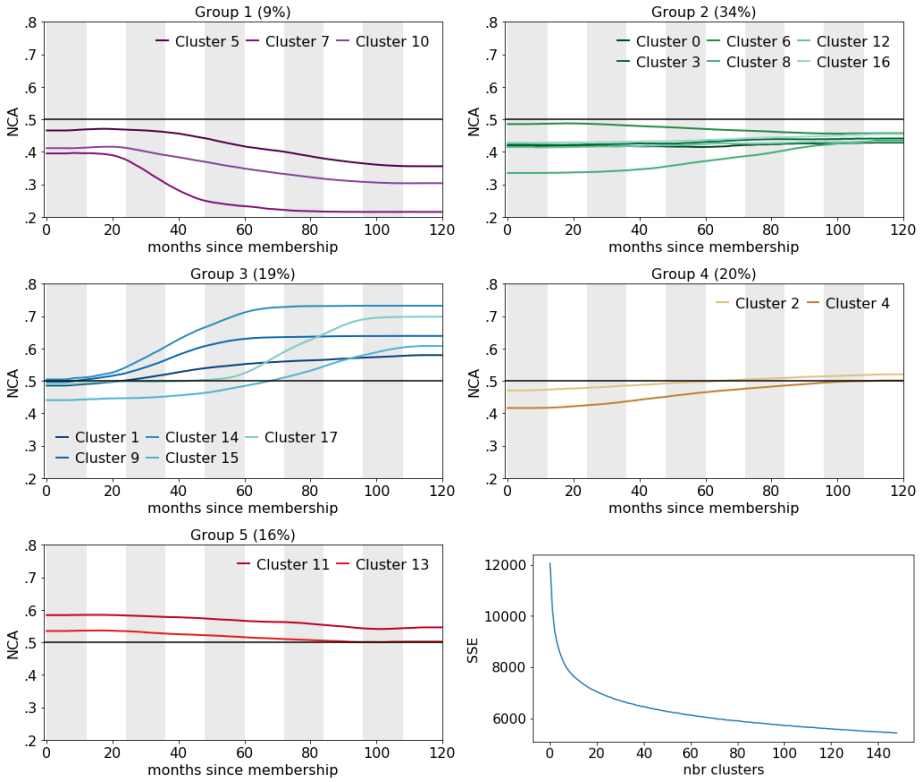

Fig. 8. Trends of NCA grouped according to the coefficient of the trend and its distance from the NCA classification threshold, i.e., 0.5. The size of each group is reported as percentage in the titles. The bottom right plot reports the Sum of Squared Error curve.

cation threshold, i.e., 0.5. We report each country within the group that captures the largest share of its customers, focusing on most representative ones[15].

- **Group 1 - Increasingly Native-Refuser**. In the first group, we find immigrants, classified as immigrants that maintain their "status" of immigrants and also increase their discrepancies against natives after 20/40 months. It is the smallest group detected with only 9% of all the immigrants. This group captures large portions of customers from *Iraq* and *Uzbekistan*.
- **Group 2 - Stable Native-Refuser**. In the second group, the largest one with 32% of immigrants, we observe customers classified as immigrants that maintain this classification but assessing their degree of NCA at about 0.45,

[15] For each group we emphasize the countries having the largest relative number of customers in that group normalized on the total number of customers from that specific country. Focusing on countries with larger absolute presence would be less interesting, as a few countries with overall very large presence (e.g. Romania, Switzerland and Germany) would simply overwhelm the others in all groups.

i.e., not far from natives. In this group, we mainly find customers from *Albania*, *China*, *Germany*, *Poland*, *Romania*, *Russia*, and *Ukraine*.
- **Group 3 - Early Native-Adopters**. In this group, the customers are initially classified as immigrants but after 10 months, or after 60 months depending on the cluster, they start to be clearly classified as natives with an NCA ranging from 0.6 to 0.75. In this group, we mainly find customers from *Brazil*, *Croatia*, *Denmark*, *Eritrea*, *Norway*, and *Slovakia*.
- **Group 4 - Late Native-Adopters**. Also in this group the customers are initially classified as immigrants but after about 80 months they are classified as natives with an NCA slightly above 0.5. We mainly find customers from *Argentine*, *France*, *Ethiopia*, *Libya*, *Switzerland*, *UK*, and *USA*.
- **Group 5 - Native-Like Customers**. Finally, in the last group we have immigrants which are never classified as such. Indeed, the trends of NCA remain stably above 0.5. In this group we mainly find customers from *Bangladesh*, *Georgia*, *Czech Republic*, and *Sweden*.

Thanks to TINCA, we are able to retrieve these distinctions in the adoption of native's shopping consumption. It is interesting to observe how the countries reported for each group do not match any specific geographical area.

6 Conclusion

In this work we investigated if the analysis of retail data through machine learning classifiers can help in understanding how much immigrants adopt natives' shopping consumption. We accomplished this task by designing TINCA, a methodology aimed at discovering Trends of Immigrants Native Consumption Adoption measuring the native adoption level by means of machine learning classifiers. Experiments on a real dataset revealed that foreign-born customers stably resident in Italy, i.e., immigrants, can be distinguished into five different groups depending on their trends of native consumption adoptions.

By design, the methodology and the results obtained are clearly of interest for the research community on social studies, and for the public bodies managing the integration of immigrants in the territory. However, we also expect that the information extracted can be useful for the retail sale companies themselves to improve their customer management, since they provide insights on the needs and behaviour of a significant portion of customers that are typically difficult to grasp, as their purchase habits derive from a time-evolving mix of culture, traditions, taste and financial resources.

This work is preliminary and it is intended to be extended in various directions besides testing it on other available datasets with similar characteristics and with different reference countries. First, we would like to validate the results with null models and extensive clustering evaluation. Second, we could add useful features taking into account spatio-temporal aspects like the typical hour and day of purchase, the number of visits at small-size and large-size shops, and a distinction between high-prices vs. low-prices products. Third, since the convergence between immigrants and natives might be explained by immigrants

switching over time from low-quality to higher quality products, we could incorporate in the customer models aspects characterizing the variety of different products purchased, like the relative price within the product category. In a way this could also mean to consider aspects related to the economic status of the customers. Fourth, we could improve the performance of the machine learning classifier with a finer parameter tuning, by adopting other classifiers like support vector machines or deep neural networks, or by modeling customers with products in the word2vec fashion [40]. Fifth, we would like to explain the reasons for the classification by interpreting the decisions made by the classifiers [25] and therefore describe the retrieved groups also with respect to food consumption and products purchases. Finally, we would like to deepen the study with respect to specific countries by either developing classifiers aimed at recognizing specific nationalities, or by clustering the trends for each country separately.

Acknowledgment

This work is partially supported by the European Community H2020 programme under the funding schemes: H2020-INFRAIA-2019-1: Res. Infr. G.A. 871042 *SoBigData++*, G.A. 825619 *AI4EU*, G.A. 761758 *Humane AI*, and G.A. 780754 *Track &Know*. We thank UniCoop Tirreno for providing the data, and Roberto Zicaro for preliminary studies on the proposed methodology and analysis.

References

1. R. Abramitzky, L. P. Boustan, et al. Cultural assimilation during the age of mass migration. Technical report, National Bureau of Economic Research, 2016.
2. R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
3. G. A. Akerlof et al. Identity economics. *The Economists' Voice*, 7(2), 2010.
4. R. Alba et al. Only english by the third generation? *Demography*, 39(3):467, 2002.
5. A. Alesina, G. Tabellini, and F. Trebbi. Is europe an optimal political area? Technical report, National Bureau of Economic Research, 2017.
6. M. Arai et al. Renouncing personal names: An empirical examination of surname change and earnings. *Journal of Labor Economics*, 27(1):127–147, 2009.
7. D. Atkin. The caloric costs of culture: Evidence from indian migrants. *American Economic Review*, 106(4):1144–81, 2016.
8. S. Bertoli, et al. Integration of syrian refugees: insights from D4R. In *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 179–199. Springer, 2019.
9. M. Bertrand and E. Kamenica. Coming apart? cultural distances in the united states over time. Technical report, National Bureau of Economic Research, 2018.
10. G. J. Borjas. The analytics of the wage effect of immigration. *JoM*, 2(1):22, 2013.
11. G. J. Borjas. *Unraveling the immigration narrative*. N&C, 2016.
12. J. A. Brillat-Savarin. *Physiologie du goût*. Charpentier, 1841.
13. B. J. Bronnenberg et al. The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, 102(6):2472–2508, 2012.
14. J. R. Bucheli, M. Fontenla, and B. J. Waddell. Return migration and violence. *World Development*, 116:113–124, 2019.

15. D. Chaffey, F. Ellis-Chadwick, R. Mayer, and K. Johnston. *Internet marketing: strategy, implementation and practice*. Pearson Education, 2009.
16. B. P. Chamberlain, A. Cardoso, et al. Customer lifetime value prediction using embeddings. In *ACM SIGKDD*, pages 1753–1762, 2017.
17. M.-C. Chen, A.-L. Chiu, and H.-H. Chang. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4):773–781, 2005.
18. F. Docquier, et al. *Emigration and democracy*. The World Bank, 2011.
19. C. Dustmann et al. Labor supply shocks, native wages, and the adjustment of local employment. *The Quarterly Journal of Economics*, 132(1):435–483, 2017.
20. R. G. Fryer Jr and S. D. Levitt. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, 119(3):767–805, 2004.
21. R. Guidotti, M. Coscia, D. Pedreschi, and D. Pennacchioli. Behavioral entropy and profitability in retail. In *2015 IEEE DSAA*, pages 1–10. IEEE, 2015.
22. R. Guidotti and L. Gabrielli. Recognizing residents and tourists with retail data using shopping profiles. In *GOODTECHS*, pages 353–363. Springer, 2017.
23. R. Guidotti, L. Gabrielli, A. Monreale, et al. Discovering temporal regularities in retail customers' shopping behavior. *EPJ Data Science*, 7(1):1–26, 2018.
24. R. Guidotti, A. Monreale, M. Nanni, et al. Clustering individual transactional data for masses of users. In *KDD*, pages 195–204. ACM, 2017.
25. R. Guidotti, A. Monreale, S. Ruggieri, et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
26. R. Guidotti, G. Rossetti, et al. Personalized market basket prediction with temporal annotated recurring sequences. *IEEE TKDE*, 31(11):2151–2163, 2018.
27. A. Herdağdelen, B. State, L. Adamic, and W. Mason. The social ties of immigrant communities in the united states. In *ACM WEBSCI*, pages 78–84, 2016.
28. R. J. Hyndman, et al. *Forecasting: principles and practice*. OTexts, 2018.
29. V. Kulkarni, et al. Freshman or fresher? quantifying the geographic variation of language in online social media. In *AAAI ICWSM*, pages 615–618, 2016.
30. F. Lamanna, M. Lenormand, et al. Immigrant community integration in world cities. *PloS one*, 13(3):e0191612, 2018.
31. T. D. Logan and P. W. Rhode. Moveable feasts: A new approach to endogenizing tastes. *manuscript (The Ohio State University)*, 2010.
32. L. Luo, et al. Tracking the evolution of customer purchase behavior segmentation via a fragmentation-coagulation process. In *IJCAI*, pages 2414–2420, 2017.
33. A. Magdy, T. M. Ghanem, M. Musleh, and M. F. Mokbel. Exploiting geo-tagged tweets to understand localized language diversity. In *GeoRich*, pages 1–6, 2014.
34. Z. Qian et al. Social boundaries and marital assimilation: Interpreting trends in racial and ethnic intermarriage. *American Sociological Review*, 72(1):68–94, 2007.
35. K. Ray. *The Migrants Table: Meals And Memories In*. TUP, 2004.
36. A. Sîrbu, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, pages 1–20, 2020.
37. A. Spilimbergo. Democracy and foreign education. *AER*, 99(1):528–43, 2009.
38. P.-N. Tan, et al. *Introduction to data mining*. Pearson Education India, 2016.
39. M. Wedel and W. A. Kamakura. *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, 2012.
40. B. Yoshua, D. Réjean, V. Pascal and J. Christian, *A neural probabilistic language model*, Journal of machine learning research, 3:1137–1155, 2003.