Explaining the Product Range Effect in Purchase Data

Diego Pennacchioli[1,2], Michele Coscia[2,3], Salvatore Rinzivillo[2], Dino Pedreschi[4], Fosca Giannotti[2]
[1] IMT Institute for Advanced Studies, Piazza San Ponziano 6, Lucca, Italy
[2] KDDLab ISTI-CNR, Via G. Moruzzi 1, Pisa, Italy, Email: {name.surname}@isti.cnr.it
[3] CID - Harvard Kennedy School, 79 JFK Street, Cambridge, MA, US, Email: michele_coscia@hks.harvard.edu
[4] KDDLab University of Pisa, Largo B. Pontecorvo 3, Pisa, Italy, Email: pedre@di.unipi.it

***Abstract*—In our market society, buyers are considered rational entities, driven by two utility functions: i) the amount of money spent, a universal quantity to be minimized; and ii) the individual needs to satisfy, a personal quantity, varying from person to person, to be maximized. In this paper, we propose an analytic framework based on big data to measure the personal utility function and we prove that this function has a stronger effect on customer behavior than the price. By focusing on the purchases in an Italian supermarket chain, we discover and describe a *range effect* of products: the more sophisticated the needs they satisfy, the more cost the customers are willing to pay to buy them, in terms of distance to travel more than in terms of the price of the item itself. We exhibit a striking empirical evidence of this theory by tracking the geographical information about points of sale and customers, in a large dataset containing tens of thousands of customers and thousands of products. We create a data mining framework able to scale to possibly hundreds of thousands, or millions, of customers and to let emerge from the data the knowledge about the actual range of each product. As an application of this finding, we show how it is possible to accurately predict how long a customer will travel (or which shop she will choose) to buy a product, as a function of the product's sophistication.**

I. Introduction

In the economic literature, market society is considered driven by rationality and the expression of this rationality is the price system. According to this view customers are rational beings: they try to minimize the amount of money they are spending, while at the same time maximizing the amount of goods they are purchasing [1]. Therefore, price is a generic utility function that each customer tries to minimize, and it is the same for everybody. However, customers are also driven by their own personal needs and desires [2]. Many of these needs are shared with other customers, such as the basic needs for survival, but many others are intimately bound to each individual and possibly different from the ones of everybody else. A customer is driven both by a generic utility function (cost minimization) and by a personal utility function (fulfillment of unique desires).

If we are able to quantify the personal utility function for each customer, then we can address a question with repercussions on a seller's market strategy: which function will win the arms race in influencing the purchase behavior of a customer, the generic one or the personal one? If the generic one is stronger, then a seller is forced to compete mostly on the price; while if a customer's needs are more important, then it is the quality of the choice that matters the most.

Here, we develop an analytic framework based on mining big customer transaction data, aimed to quantify the strength of both utility functions. We test the customer behavior in terms of distance traveled, under the assumption that customers want to minimize their travel length. We observe that customers do not always go to the closest supermarket: there is a *range effect* for each product, due to the intrinsic characteristics of the product. To explain and predict the range effect we propose a method to compare the strength of the generic and the personal utility function in the customer's mind. This comparison boils down to the question: given that customers travel on average x meters to buy product p, are they doing that because p is expensive or because p satisfies very particular needs?

While the price is an explicit information of the product, the needs the product itself is satisfying are not. We quantify them by evaluating the *sophistication* of each product and customer, following [3]. We find that the sophistication of a product is better than the price in explaining a customer's behavior.

We provide empirical evidence of these claims with real world data about customer behavior. We analyze digital traces of customer purchases in the database of a large supermarket chain in Italy. We focus on a single supermarket chain, Coop, the largest Italian supermarket company, and on a single Italian city. Coop is a cooperative and most of its customers are members that receive discounts and promotions through fidelity (membership) cards. With these fidelity cards, the company is able to recognize the different purchases of the same customer. Moreover, when registering for the fidelity card, the customer is giving to the company some personal information, including the home address.

We show how our proposed product sophistication index is a better explanatory variable of a product range than the price. The more sophisticated is the need a product satisfies, the longer a customer will travel to purchase it on average, almost regardless of its price. Intuitively, this means that to buy bread people will just settle with the closest shop where it is available, while to buy blank DVDs, with roughly the same price and available in all the supermarkets of the chain, a customer will travel a significantly longer distance. While the product range concept may be quite intuitive, in this paper we provide a system able to quantify it better than just assuming that it is proportional to the product price.

There are many consequences for sellers from the ability of predicting a product's range. For instance, to know the range of all the products of a supermarket implies that the supermarket's marketing strategies can be tailored according to the distance of a customer from the nearby points of sales. Customers far away from a point of sale need to be stimulated on more sophisticated needs, while nearby customers may be more susceptible to more basic needs.

A second application is in point of sale placement, as we can use our methodology in conjunction with the central place theory [4]. Besides the construction costs, each point in the city space is altering the minimum distance between a customer and a product. Therefore, given the range effect, each point in the city space has one optimum in its product assortment. In this paper, we provide the proof that this problem can be formally addressed to find a good approximate solution.

The final contribution of this paper is to show how to accurately predict how long a customer will travel (or which shop she will choose) to buy a given product, as a function of the product's sophistication. In other words, product sophistication reveals as a powerful predictor feature for a challenging predictive task, because most people shop preferably at the closest store for most products, so it is difficult to accurately characterize for which products a customer will travel more.

These applications have to cope with the enormous amount of data flowing every day in the real world. For this reason, we create a scalable framework, using a data mining approach similar to the one at the basis of the PageRank [5], that is able to analyze networks with hundreds of million of nodes. To increase the interpretability of our results, we narrow our questions to the customer base of a city. The total amount of customers we are tracking is $60,366$, buying $4,567$ products. However, we are able to scale to large numbers, having applied our framework on an entire Italian province, including more than $300,000$ customers [6]. With our framework, it is possible to let emerge from the data the knowledge about the actual range of each product.

The rest of the paper is organized as follows. We deal with the related literature on economics and data mining in Section II. In Section III, we present in detail our dataset. Section IV presents proofs of the range effect in customer purchases, along with some possible explanations. We provide a final explanation of the range effect in Section V, first by introducing product and customer sophistication and then putting these concept in relation with the distances traveled by customers. In Section VI we create a classifier able to predict customer movements. Section VII concludes the paper, with insights about future developments.

II. Related Works

In this paper we address the problem of explaining customer behavior in terms of how customers decide to move given what they want to buy. We do so by using data mining tools. Thus, the paper relates mainly to two parts of the literature: studies of customer behavior and data mining for marketing and for the analysis of spatial data.

The first field has been tackled mainly in the economy literature. As noted in the introduction, behavioral economy has focused its attention on the rational choices of the entities in the market [7], [2], [1]. Customer behavior study is also a classical marketing problem [8]. However, the recent studies about the customer movements and purchases are more focused on visually inspecting the movements of people inside a shop [9]. This line of research is present also on the computer science side [10]. In computer science, there are also examples of GIS approaches to business intelligence [11], of recommender systems for customer retention [12] and spatial analysis of customer-to-business communications [13]. In computer science [14] is a comprehensive book explaining the relations between economy and a network-based analysis.

Marketing applications are historically one of the most natural testing ground for Data Mining [15], [16]. Our main aim, understanding the links between customers and products, has been tackled in data mining: by analyzing them in a multidimensional space [17], by mining frameworks to understand customer behavior [18], [19], [20] and by defining a data-driven customer segmentation [21]. However, these works have in common the aim of the specific description of single customers, rather than finding a broader and general pattern in the data, that is the main focus of this paper. Data mining has been widely used also in other generic problems related to geographical systems. For example, a network mining approach has been used to detect the borders of human mobility [22], of tweet's topics [23] and trajectory pattern analysis [24], [25].

III. Data

The dataset we used is the retail market data of Coop, one of the largest Italian retail distribution company. The conceptual data model of the data warehouse storing the retail data is depicted in Figure 1.

The whole dataset contains retail market data in a time window that goes from January 1st, 2007 to December, 31st 2011. The active and recognizable customers in that interval are $1,066,020$. A customer is active if she has purchased something during the data time window, while she is recognizable if the purchase has been made using a membership card. The 138 stores of the company cover an extensive part of Italy, selling $345,208$ different items.

Each data entry contains information about a product item bought by a customer in a specific store. We focus on the following three dimensions of each data item: marketing category, store category and customer attributes.

Marketing category, is used to classify products: it is organized as a tree and it represents a hierarchy built on the product typologies, designed by marketing experts of the company (see Figure 1 for a list of hierarchy levels). The top level of this hierarchy is called "Area" and it is split in three fundamental product areas: *Food*, *No Food* and *Other*. The bottom level of the marketing hierarchy, the one directly on top of the leaves of the tree, is called *Segment* and it contains $7,003$ different values. Each item has a classification in this hierarchy and, thus, we can exploit such tree to choose the most suitable level of aggregation of products. For example, at the Segment level we identify with "Sugar-free Orange Juice" both the

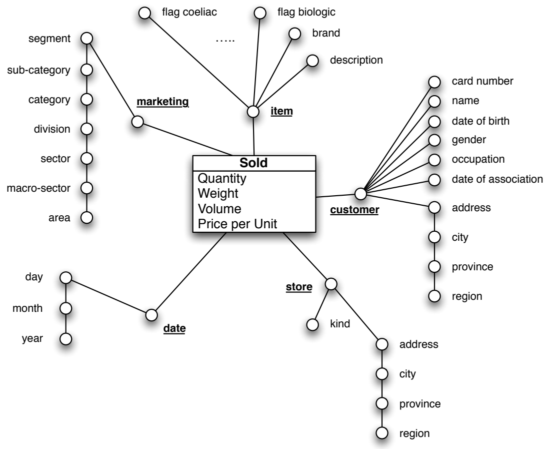


Fig. 1: The Conceptual Data Model (star schema) of the Data Warehouse

liter and half-liter bottle items. Since we are not interested in distinguishing the different packaging of the same product we aggregate all our products at the *Segment* level. In this way, we aggregate products that are logically equivalent, thus reducing a possible source of noise. To reduce potential outliers, we exclude from the analysis all the products (segments) that are either too frequent (e.g. the shopping bag) or meaningless for the purchasing analysis (e.g. discount vouchers, errors, segments never sold, etc.).

Store category allows to group the retail stores according to their size (expressed in terms of product assortment, physical shop size and number of employees). There are three distinct categories: *iper*, *super* and *gestin*, in decreasing order of size. For each store we also have several attributes and, in particular, its geographic position (Figure 1).

Customer attributes contains demographic information about a customer and are collected at the moment of obtaining the membership card. In particular, in the following we will use mainly the geographic position of the customer, derived via geocoding from her address.

Since our analysis is based on distances between customers and stores, we focus our presentation only on one metropolitan area, to be able to interpret more easily the results and avoid the problem of people in the border of more cities. However, it is important to notice that our framework can easily scale for larger data collections [6]. We chose a city motivated by the strong penetration of customers for our retail distribution company. The customer base is not only large, but very committed to the brand, being "members" of COOP, and not just getting discounts: we can fairly assume that the people we are studying make most of their purchases in this chain. For each of the five stores we selected all the customers within a radius of $5km$ from each store.

The resulting dataset contains $60,366$ customers and $4,567$ segments, with $107,371,973$ total purchases[1].

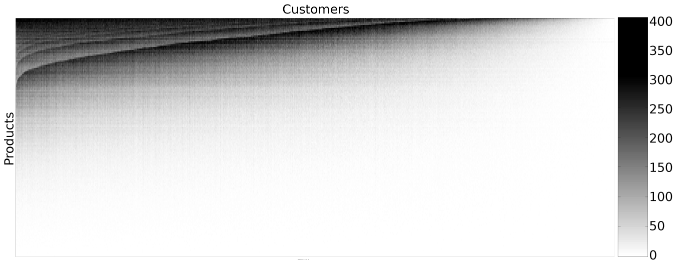


Fig. 2: The purchase matrix M_{cp}.

From the purchase data we derive a relation M_{cp}, represented as a binary matrix, where the entry (c_j, p_i) evaluates to 1 if customer c_j has bought a significant amount of product p_i. In this case we cannot simply state the existence of a single purchase of c_j for p_i, since this may generate excessive noise. We need a mechanism to evaluate how meaningful is a purchase quantity for each product p_i for each customer c_j. This evaluation is done using the concept of lift [15], that is related to association rule mining. Given a couple of itemsets (X, Y), the lift of the couple is defined as follows:

$$\text{lift}(X, Y) = \frac{\text{supp}(X, Y)}{\text{supp}(Y) \times \text{supp}(X)},$$

where $\text{supp}(I)$ is the relative support of the itemset I. The relative support of itemset I is the number of times all $i \in I$ are purchased together over all the transactions.

In our case, we force a particular condition: the itemset X always contains one item (the customer c_j); the itemset Y always contains one element (the product p). Therefore, $\text{supp}(c_j, p_i)$ is the relative amount of product p_i bought by customer c_j, $\text{supp}(p_i)$ is the relative amount sold of product p_i to all customers and $\text{supp}(c_j)$ is the relative amount of all products bought by customer c_j. So, if c_j bought 100 total items, p_i has been sold in 200 items, c_j bought 10 p_i items and the entire transaction dataset contain 1,000 purchases, $\text{supp}(c_j, p_i) = \frac{10}{1000}$, $\text{supp}(c_j) = \frac{100}{1000}$, $\text{supp}(p_i) = \frac{200}{1000}$ and $\text{lift}(c_j, p_i) = \frac{0.01}{0.2 \times 0.1} = 0.5$.

Lift takes values from 0 (when $\text{supp}(c_j, p_i) = 0$, i.e. customer c_j never bought a single instance of product p_i) to $+\infty$. When $\text{lift}(c_j, p_i) = 1$, it means that customer c_j did buy product p_i in the quantity we would have expected if her purchases were random. If $\text{lift}(c_j, p_i) < 1$ then customer c_j purchased the product p_i less than expected, and viceversa. Therefore, higher than 1 lift values imply that the purchases are meaningful. The values in M_{cp} are then:

$$M_{cp}(c_i, p_j) = \begin{cases} 1 & \text{if lift}(c_j, p_i) > 1; \\ 0 & \text{otherwise.} \end{cases}$$

The purchase matrix M_{cp} is depicted in Figure 2, where rows and columns are sorted according to the volume of sales: from left to right customers are sorted according to how many products they bought and from top to bottom products are sorted according to how much they are sold.

In Figure 2, the columns of the matrix are the $60,366$ customers and the rows are the $4,567$ products. We depicted a

[1] This dataset has been made available along with all the framework coding at `http://www.michelecoscia.com/?page_id=379`. Customer and product IDs are obfuscated for privacy and business protection reasons.

Product	AVG Distance (in meters)
Pizza	809
Packed Salads	1, 576
Frozen Side Dishes	2, 437
School Notebooks	3, 511
Travel Books	5, 523

TABLE I: A selection of the more basic products according to their PS values.

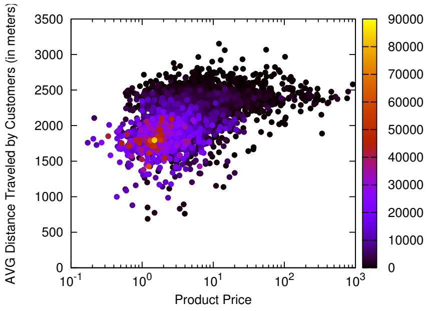


Fig. 3: Average distance traveled to get a product with a given price.

compressed view of the matrix, where each data dot represent a 30×30 square of the original matrix and the gray gradient represents how many $1s$ are present in that section of the matrix, for space constraints.

IV. The Range Effect

The assumption of this paper is that customers modify their shopping behavior according to their relative position w.r.t the shop they are going to. A customer may decide to buy or not buy a given product because it is close enough or too far away from the shop. We call this phenomenon the *range effect* of a product. Table I reports the average distance traveled for purchasing a product. We can find products for which customers traveled more than 5 kilometers on average, other products for which the average distance is less than 1 kilometer and many other products generated a variety of average distances. There are two trivial explanations of this fact: it is driven by price and/or by the frequency with which a product needs to be purchased.

We expect that customers will travel more to purchase products that are more expensive, for many possible reasons (they require higher quality, they may be not available around them, and so on). We check this hypothesis by plotting for each purchase the price of an item against the average distance that a customer traveled to get the product. This plot is depicted in Figure 3: the price is on the x axis (in logarithmic scale), while the distance traveled is on the y axis. The price is recorded in Euros. Each dot is a purchase and we color it accordingly to how many purchases are represented by the same price and by the same distance.

Intuitively, it would make sense to plot just one point per product, as we want to know the average distance traveled by customers given a product price. However, this would disproportionately weigh the purchases of products sold less frequently, or the purchases made by customers who buy only a handful of products. Filtering out these purchases also would not make sense, as our purchase matrix is triangular: there are many products sold in small quantities and many customers who purchase only few products. Therefore the behavior of these shoppers is important. By plotting each single purchase, we know that we are weighing each customer behavior by its fair proportion of purchases.

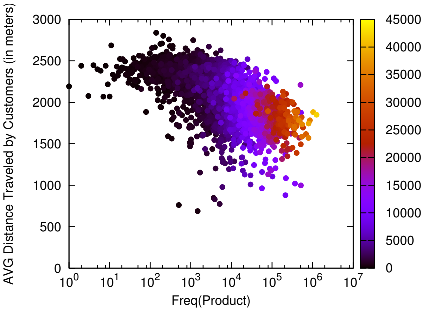


Fig. 4: Average distance traveled to get a product with a given popularity.

The connection of a customer to a product is created with the procedure described in Section III, therefore we are only considering connections generated when the quantity of product p bought by customer c_i is significant. A customer c_i may have bought product p_j in different shops, say s_1, s_2, s_3, s_4. In this case, we weigh each distance traveled with the amount of purchases made using the following formula:

$$d(c_i, p_j) = \sum_{\forall s \in S} \frac{p_j(c_i, s) \times d(c_i, s)}{p_j(c_i, *)},$$

where S is the set of all shops, $d(c_i, s)$ is the distance between customer c_i and shop s, $p_j(c_i, s)$ and $p_j(c_i, *)$ is the amount of purchases of product p_j made by customer c_i in shop s and in general, respectively. This procedure has been done for the plots depicted in Figures 3, 4, 5(a) and 5(b).

Products with the same price are bought by customers placed at different distances w.r.t the shop. Given a price, we average the distance traveled by the customers buying the products with that exact price. By averaging, we lose the ability of describing each single customer and we just describe the behavior of the system in its entirety. We do so because the single customer is bounded by the place where she lives, thus each single customer carries a noisy information, and we can make sense of it only by looking at the global level.

From Figure 3 we can conclude that price plays a role in driving customer decisions of traveling a given distance for a product. The correlation here looks weak, but positive: customers travel more if they need to buy a more expensive product. We calculate a log-normal regression[2] using the function $f(x) = a \log x + b$. In this regression, $R^2 = 17.25\%$, meaning that we can explain 17.25% of the variance in the distance traveled using the price.

To check if the frequency of purchase can explain the distance traveled by customers, we repeated the same analysis,

[2] This and all other regressions have been calculated with the *leastsq* function of the *SciPy* module for Python.

using the number of purchases of a product instead of the price. We depicted the plot in Figure 4. The correlation here is negative: the more frequently a product needs to be bought, the smaller the distance a customer will travel for it. We calculate a regression with the function $f(x) = a \log x + b$ and we obtained $R^2 = 32.38\%$.

As a conclusion of this section, we can state that the price plays a small role in predicting the distance a customer will travel for purchasing a product, by increasing it. If a product is needed more frequently then it drives (down) the distance a customer will travel to buy it, regardless of the price. However, there is a large amount of variance that remains unexplained. In the next section, we provide one possible explanation.

V. Explaining the Range Effect

In this Section we tackle the problem of explaining the *range effect* for products. Our theory states that customers travel more to buy a product if the product can satisfy a more sophisticated need and/or they have sophisticated needs in general. To do so, first we need to formally define what exactly product and customer sophistication are, and we do that in Section V-A. Then, we provide evidences that the product and the customer sophistication are variables able to better explain the distance traveled by customers, in Section V-B. Finally, in Section V-C we provide explanations of why our product sophistication index is better predictor of customer behavior.

A. Product and Customer Sophistication

The basic intuition to quantify the sophistication level of products and customers is that more sophisticated products are by definition less needed, as they are expression of a more complex need. To be considered "sophisticated" a product needs to satisfy two constraints: 1) it has to be sold to few customers; and 2) the customers buying it have to buy all products that are less sophisticated than it. The logic is that each product satisfies a need and a customer buys a product if and only if she already satisfied all more basic needs. Figure 2 shows that the data align with this theory: the columns in the right part of the matrix are customers buying only few products and these products are more or less bought by everyone, thus they are basic. For this reason, we need to evaluate at the same time the level of sophistication of a product and of the needs of a customer using the data in the purchase matrix, and recursively correct the one with the other. We adapt the procedure of [3], adjusting it for our big data.

We calculate the sums of the purchase matrix for each customer ($k_{c,0} = \sum_p M_{cp}(c,p)$) and product ($k_{0,p} = \sum_c M_{cp}(c,p)$). To generate a more accurate measure of the sophistication of a product we need to correct these sums recursively: this requires us to calculate the average level of sophistication of the customers' needs by looking at the average sophistication of the products that they buy, and then use it to update the average sophistication of these products, and so forth. This can be expressed as follows: $k_{N,p} = \frac{1}{k_{0,p}} \sum_c M_{cp} k_{c,N-1}$. We then insert $k_{c,N-1}$ into $k_{N,p}$ obtaining:

$$k_{N,p} = \frac{1}{k_{0,p}} \sum_c M_{cp} \frac{1}{k_{c,0}} \sum_{p'} M_{cp'} k_{N-2,p'}$$

$$k_{N,p} = \sum_{p'} k_{N-2,p'} \sum_c \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}$$

and rewrite this as:

$$k_{N,p} = \sum_{p'} \widetilde{M}_{pp'} k_{N-2,p'},$$

where:

$$\widetilde{M}_{pp'} = \sum_c \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}.$$

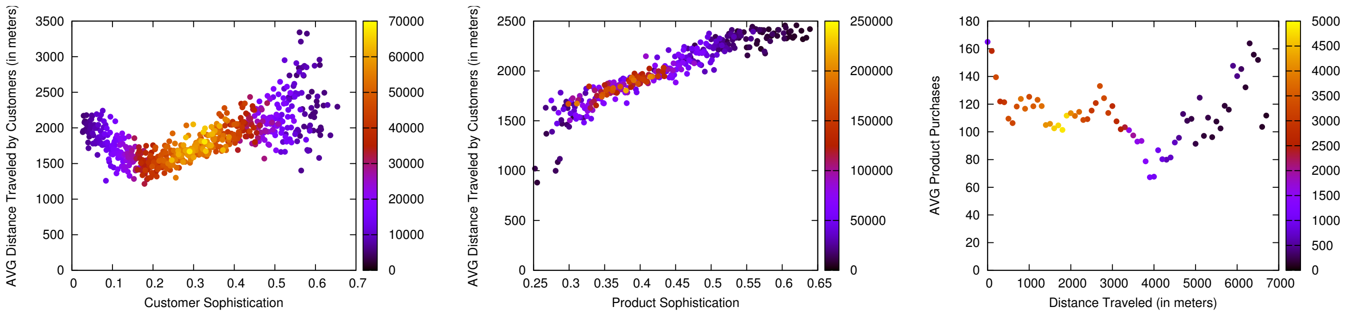
We note in the last formulation $k_{N,p}$ is satisfied when $k_{N,p} = k_{N-2,p}$ and this is equal to a certain constant a. This is the eigenvector associated with the largest eigenvalue (that is equal to one). Since this eigenvector is a vector composed by the same constant, it is not informative. We look, instead, for the eigenvector associated with the second largest eigenvalue. This is the eigenvector associated with the variance in the system and thus it is the correct estimate of product sophistication.

However, this formulation is very sensitive to noise, i.e. products that are bought only by a very narrow set of customers. To calculate the eigenvector on the entire set of products generates a small amount of products whose sophistication level is seven orders of magnitude larger than the rest of the products. This variance provokes the other sophistication estimates to be flattened down to the same values and therefore not meaningful. However, we do not want to simply cut the least sold products, as we aim to create a full product hierarchy, including (especially) also the least sold products. To normalize this, we employ a three step strategy. First, we calculate the eigenvector on a restricted number of more popular products (purchased by at least a given threshold δ of customers). Then we use the estimate of the sophistication of these products to estimate the sophistication of the entire set of customers (that is, as defined before, the average sophistication of the restricted set of products they buy). Finally, we use the estimated sophistication of the customers to have the final sophistication of the entire set of products, again by averaging the sophistication of the customers buying them. Hence, we define the product sophistication index (PS) of product p as:

$$PS(p) = \frac{\vec{K}(p) - \min(\vec{K})}{\max(\vec{K}) - \min(\vec{K})},$$

where $\vec{K}$ is the eigenvector of $\widetilde{M}_{pp'}$ associated to the second largest eigenvalue, normalized as described above. With this strategy, PS takes values between 0 and 1. The Customer Sophistication CS is calculated using the very same procedure, by estimating $k_{c,N}$ instead of $k_{N,p}$.

In Table II we report a selection of the least sophisticated products, i.e. the ones with the lowest PS values, in the purchase matrix. The less sophisticated products should be intuitively the ones covering the most basic human needs,



(a) Average distance traveled to get a product with a given sophistication index

(b) Average distance traveled by a customer with a given sophistication index

(c) Average number of purchases by customers at a given distance from the shop

Fig. 5: Sophistication and customer behavior against the distance from the shop.

p_i	PS
Regular Bread	0.252
Red Meat	0.266
Artichokes	0.275
Pasta	0.275
Rabbit Meat	0.278

TABLE II: A selection of the more basic products according to their PS values.

p_i	PS
Winter Suit 3-12yo	0.796
TV 29”	0.769
DVD Readers	0.754
Hair Spray	0.742
8mm Cameras	0.739

TABLE III: A selection of the more sophisticated products according to their PS values.

and this intuition is confirmed by the reported products: bread, vegetables, meat and pasta (remember that this is an Italian chain). On the other hand, Table III reports the most sophisticated products, i.e. the ones with the largest PS values, that intuitively should be products satisfying high-level non-necessary, probably luxury, needs. In fact, what we find in Table III are hi-tech products (televisions, DVD readers and cameras), fashion products and very specific clothing.

As a last note, we do not provide a time and space complexity evaluation of this methodology for two reasons. First, the calculation of product and customer sophistication is mostly similar to the PageRank calculation [5], that has been applied to large matrices describing the entire Web. Second, we have published a technical report [6] in which we describe this procedure and we apply it to three different datasets with more than $300,000$ customers and hundreds of millions of total purchases. For these two reasons, we claim that our system is able to scale and to analyze very large datasets.

B. Sophistication and Range

To understand if the product and/or the customer sophistication is influencing the distance a customer will travel to purchase the product she needs, we generate the same plots shown in Section IV. The plots are depicted in Figure 5(a-b). We recall that in these plots each data point is a purchase.

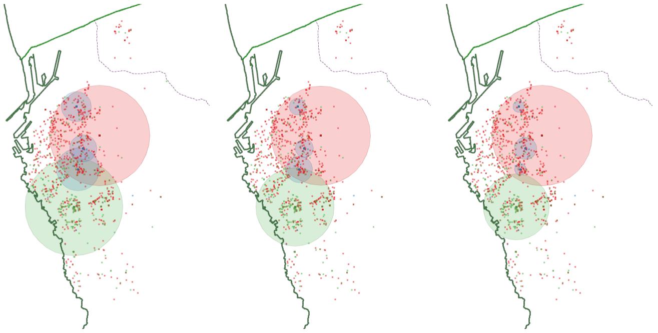
In Figure 5(a) we test the relationship between the distance traveled and the customer sophistication: we calculate the average distance traveled by customers (y axis) to get to the shop against their sophistication value (x axis). In this case, the x axis has not a logarithmic scale, as the relationship is linear.

We can see that the relationship between distance traveled and customer sophistication looks non-linear. From a value of sophistication of 0 to around 0.2 the relationship is negative, while it is clearly positive afterwards. The sole conclusion we have is that there is some kind of relation, but we do not have an explanation for it.

For this reason, we move on in depicting the product sophistication (x axis) against the average distance traveled by the customers to purchase the given product (y axis) in Figure 5(b). In this case, the relationship is clear: the more a product is sophisticated, the more customers will travel to buy them. The product sophistication has a normal distribution, but less sophisticated products are more sold, given the triangular shape of the matrix. This fact explains why most of the data points are in the left part of the plot: most purchases are generated for low sophistication products. We calculated a linear regression, for which $R^2 = 85.72\%$. This R^2 is more than twice higher than the R^2 obtained with the purchase frequency, explaining much better the variance in the distances traveled by customer.

A possible objection is that the distance is influencing the number of products purchased by a particular customer, and this would invalidate the explanatory power of the product sophistication index. We already saw in Section IV that the distance and the frequency of purchase are somewhat related, but this relationship cannot fully explain what we see in Figure 5(b). However, this objection is focused on the customer, not on the product: it states that the customer-shop distance may have a strong positive or negative correlation with the number of items purchased on average by he customer.

We depict this relationship in Figure 5(c): the x axis is the distance of a customer from the shop and on the y axis we have its average number of products purchased. Customers at the same distance may have purchased different quantities of products, so we average them and we color the data point accordingly to how many customers it represents. As we can



(a) Pasta (b) Breaded Frozen Fish (c) Health Testers

Fig. 6: The customer-shop distribution for customers buying different products.

see, there is no relationship at all between distance and number of products purchased. For this reason, we can reasonably state that customers tend to travel more to purchase more sophisticated products.

C. Customer Behavior

We now provide a possible explanation of customer behavior. Customers tend to buy products with a low sophistication level in the closest possible shop. However, when they are in need to buy a more sophisticated product, they do not choose the closest shop even if the shop has the product they look for (and, given the fact that we are considering shops of the same chain, the quality level of the products is identical).

We provide a visual argument for this explanation. In Figures 6(a-c) we generated three maps representing the purchases of three different products. We chose products with different sophistication level: in Figure 6(a) we focus on a very low sophisticated product (pasta), in Figure 6(b) we focus on a medium-low sophisticated product (breaded frozen fish) and Figure 6(c) we focus on a medium-high sophisticated product (health testers, like pregnancy or insulin indicators). Each dot in the map is a location in which we found one or more customers that has bought the given product. The color of the dot represent the shop type in which the customer went for her purchase. The colored circles are centered on the position of the given shop and their radius is the median distance traveled by customers to purchase the product in that shop.

As we saw in Section III, shops have an attribute "type", that encodes the category of the shop, a proxy of its size. In Figures 6(a-c), the customers in red went to the "iper" shop (the largest in our data), the customers in green went to the "super" shop (smaller than the "iper"), while customers in blue went to one of the three "gestin" shops (smaller than a "super"). As we can see, the smaller shops have quite some range in attracting customers who need the lowest sophisticated product. However, as the sophistication of the product increases, the number of customers going at those shops becomes lower and lower. The red circle keeps its radius, while the green and blue circles tend to shrink.

Instead of relying on three examples out of the $4,567$ products, we report this trend in Table IV. For each shop, we record the average product sophistication of the products sold in that shop in significant quantities (column "AVG PS"). We also record the average distance traveled by customers to buy products in significant quantities in that shop (column "AVG Distance"). The "Shop Type" column refers to the shop classification explained in Section III.

Shop Type	AVG PS	AVG Distance
Iper	0.49	$2,392$
Super	0.46	$1,721$
Gestin	0.43	869

TABLE IV: Average PS values and average customer distance for the shops in our dataset.

We can see that indeed there is a difference between the complexity of the "iper" (red) shop with the "super" (green) shop, and another significant difference between the "super" shop and the rest of the "gestin" shops. This significant difference is also reflected in the average distance traveled by customers: almost 2.4 kilometers to get to the "iper" shop, more than 1.7 kilometers for the "super" shop and less than 900 meters for the "gestin" shops. Table IV proves that large shops are objectively more sophisticated than smaller ones and suggests that are also subjectively considered so by customers.

The conclusion we draw is that the average sophistication of the products in a shop is influencing customers' decisions: when they need a more sophisticated product they are prone to decide to go to a larger shop with higher sophistication even if that product is also present in the smaller shops.

In the next section, we put our finding into practice.

VI. Customer Behavior Prediction

In the previous sections we showed how the product sophistication can be used to describe the average customer behavior better than the price of a product or how frequently the product is needed. However, as noted, the average customer and all the separate individual customers are different entities. If the knowledge about the average customer cannot be used for predictions, then the product sophistication cannot be used in practice. Aim of this section is to show that predictions based on the product sophistication can achieve a significant improvement over the ones based on price and frequency of purchase.

To do so, we provide the following problem definition:

Definition 1: Let D be a set of triplets (c, p, s). Each triplet represent the purchases of product p made by customer c, and $s \in \{s_1, s_2, s_3, s_4, s_5\}$ is the target shop. We want to build a classifier that, given some features of c and p, returns the value of s.

In other words, for each customer c and product p we know the shop s in which c usually goes to buy p, and we want to predict s using information about c and p. For each customer, the feature we calculated is the weighted average distance that c usually travels to buy all the products she needs. We used the formula:

$$\bar{d}(c_i) = \frac{1}{W(c_i)} \sum_{p \in P(c_i)} w_p \times d(c_i, s),$$

Shop ID	Shop Type	Row Share
s_1	Iper	53.67%
s_2	Super	32.08%
s_3	Gestin	7.62%
s_4	Gestin	2.91%
s_5	Gestin	3.72%

TABLE V: The distribution per shop of the filtered dataset for the classifier.

where w_p is the weight of product p, $P(c_i)$ is the set of products bought by customer c_i and $W(c_i) = \sum_{p \in P(c_i)} w_p$. We used three different classes of weights, based on the product price, quantity and sophistication, thus generating three attributes for each customer. We repeated the same procedure for the products, by using the same weighted average distance, using $C(p_j)$ (the set of customers buying product p_j) instead of $P(c_i)$. In the end, we obtained three features also for the product, based again on price, frequency of purchase and product sophistication.

Our dataset is heavily unbalanced on the larger shop, that attracts most of the purchases and contains products that are not present in any other shop. We then filter the data, to focus only on those cases where the prediction task is harder. For this reason, we defined three constraints that each entry in our test data has to satisfy.

1) If the product is sold only in one shop, the prediction task is trivial. Thus, we want to consider only the products that are sold at least once in each of the five shops.
2) We consider only customers with a diversified shopping behavior. If the customers always went to the same shop, the prediction of its movements is trivial. For this reason, we select only the customers who purchase significant quantities of products in at least two different shops.
3) If a customer purchased the same product in two different shops we only kept the entry corresponding to the shop where he purchased the largest quantity of the product, as the classifier will output only one shop and therefore could not achieve a perfect accuracy.

The entries in our dataset, the triplets (customer, product, shop), satisfying all three constraints are $10,412,391$. In Table V we report the share of the rows of our filtered dataset whose target variable takes one of the possible five values, corresponding to the five shops. From Table V we know that we can build a naive classifier that always returns "s_1" as a result, and we would get an accuracy of 53.67%.

Given the size of the dataset, we extracted samples containing 5% of the entries (around $500,000$) and we performed our prediction tasks on these samples. The results we show are consistent in our samples. We created our classifier using the c4.5 algorithm [26]. To validate our results, we used the k-fold cross-validation method, by setting $k = 10$. We divided our data sample by putting two thirds of the data in the training set and the remaining data in the test set.

We depicted the lift charts of our classifiers in Figures 7(a-c). In the lift chart, on the x axis we have to total population fraction and on the y axis we have the population fraction that has been classified correctly. Since the correctly classified population has as upper bound the population itself, the perfect predictor that achieves a 100% accuracy is the bisector that goes from $(0, 0)$ to $(1, 1)$, and we depict it in all figures with a blue line. The naive classifier, that always returns s_1 as result is depicted with a black line. The area between the blue line and the black line is where a model that improves over the baseline should lie.

In Figure 7(a) we consider as first model a classifier based on the product price. The red line shows that this classifier makes only an incremental improvement over the baseline, with an overall accuracy of 59.03%. We added the product sophistication information to this classifier (green line) showing a further accuracy improvement, ending up with an overall value of 65.87%.

We repeated the same analysis, this time using a classifier based on the frequency of purchase of a product. We can see that Figure 7(b) looks very similar to Figure 7(a): again the classifier based on on the frequency (red line) improves to an overall accuracy of 60.09%, while adding the product sophistication information (green line) bring to an overall accuracy of 67.91%.

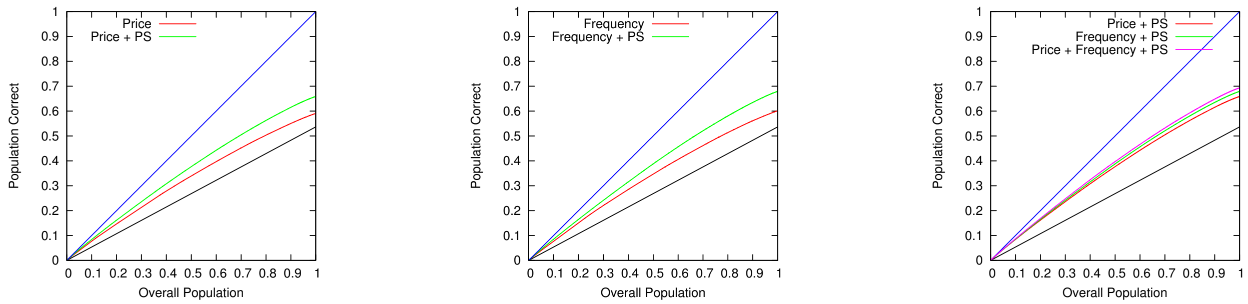
We also point out in Figure 7(c) that a classifier including all the available information does not significantly improve the accuracy. Especially comparing to the frequency of purchase and product sophistication classifier (green line in Figure 7(c)), the increased accuracy of the classifier including also the price information (purple line) is very low. The overall accuracy of this model is 69.33%.

As a conclusion, we saw that the product sophistication adds significantly to the accuracy of the predictions based on price ($+6.84\%$) and on the frequency of purchase ($+7.82\%$). Adding the price information to this last classifier provides too a marginal improvement, but lower than the one provided by the product sophistication itself ($+1.42\%$). Therefore, the product sophistication is a very strong factor that not only explains on average customer movements, but can be effectively used to increase customer behavior predictions at the level of the single customers.

VII. Conclusion and Future Works

In this paper we addressed the problem of explaining and predicting customer behavior when shopping to large retailers. We showed that products have what we call a *range effect*: for some products, customers travel long distances, while for other products they settle down with the closest shop. We ruled out as possible explanations of this phenomenon the price of a product and the frequency of purchase. We introduced a new measure, namely the product sophistication, that is able to better explain customer movements: it is because products satisfy more complex needs, not because they are more expensive or they are needed less often, that customers travels more. We also showed as this additional information provides a significant boost of the accuracy in predicting in which shop a given customer will go buying a given product.

This paper paves the way to many future developments. First, our prediction accuracy is good, but it may be improved,



(a) Price

(b) Frequency

(c) Global

Fig. 7: Lift charts showing the increase in predicting performance obtained using product sophistication.

by using more sophisticated measures such as the radius of gyration [27], [28] of customers and products. Second, we analyzed a static snapshot of retail, but it would be interesting to analyze the evolution of customer behavior. Finally, following [3], to create a network of products based on the customers buying them may lead to further insights.

Acknowledgments. We thank the supermarket company Coop and Walter Fabbri for sharing the data with us and allowing us to analyze and to publish the results. We also gratefully thank Muhammed Yildirim for his support. This work has been partially supported by the European Commission under the FET-Open Project n. FP7-ICT-270833, DATA SIM.

References

[1] M. Hollis and E. J. Nell, *Rational Economic Man*. Cambridge University Press, 2007. [Online]. Available: http://books.google.com/books?id=9G23xUw7NEIC

[2] E. Fernandez-Huerga, "The economic behavior of human beings: The institutional/post-keynesian model," in *Journal of Economic Issues*, vol. 42, 2008, p. 709.

[3] R. Hausmann, C. Hidalgo, S. Bustos, M. Coscia, S. Chung, J. Jimenez, A. Simoes, and M. Yildirim, "The atlas of economic complexity," *Boston. USA*, 2011.

[4] W. Christaller, *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. University Microfilms, 1933. [Online]. Available: http://books.google.com/books?id=elAiAAAAMAAJ

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[6] D. Pennacchioli, M. Coscia, F. Giannotti, and D. Pedreschi, "Calculating product and customer sophistication on a large transactional dataset," Technical Report, 2013.

[7] P. Diamond and H. Vartiainen, *Behavioral Economics and Its Applications*. Princeton University Press, 2008. [Online]. Available: http://books.google.com/books?id=1-SVhlC9mVoC

[8] R. Stock and W. Hoyer, "An attitude-behavior model of salespeoples customer orientation," *Journal of the Academy of Marketing Science*, vol. 33, pp. 536–552, 2005. [Online]. Available: http://dx.doi.org/10.1177/0092070305276368

[9] A. J. Newman, D. K. Yu, and D. P. Oulton, "New insights into retail space and format planning from customer-tracking data," *Journal of Retailing and Consumer Services*, vol. 9, no. 5, pp. 253 – 258, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969698902000103

[10] H. G. N. Rai, K. Jonna, and P. R. Krishna, "Video analytics solution for tracking customer locations in retail shopping malls," in *KDD*, 2011, pp. 773–776.

[11] J. Pick, *Geographic Information Systems in Business*. Idea Group Pub., 2005. [Online]. Available: http://books.google.com/books?id=4hZrCVw7WUIC

[12] M. Gorgoglione, U. Panniello, and A. Tuzhilin, "The effect of context-aware recommendations on customer purchasing behavior and trust," in *RecSys*, 2011, pp. 85–92.

[13] R. Hariharan, J. M. Loh, J. Shanahan, and K. Yamada, "Spatial probabilistic modeling of calls to businesses," in *GIS*, 2010, pp. 466–469.

[14] D. A. Easley and J. M. Kleinberg, *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[15] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993. [Online]. Available: http://doi.acm.org/10.1145/170036.170072

[16] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed. *Wiley Computer Publishing, Apr. 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471470643

[17] X. Li, J. Han, X. Yin, and D. Xin, "Mining evolving customer-product relationships in multi-dimensional space," in *ICDE*, 2005, pp. 580–581.

[18] A. C. M. Fong, B. Zhou, S. C. Hui, J. Tang, and G. Hong, "Generation of personalized ontology based on consumer emotion and behavior analysis," *T. Affective Computing*, vol. 3, no. 2, pp. 152–164, 2012.

[19] B.-E. Shie, H.-F. Hsiao, P. S. Yu, and V. S. Tseng, "Discovering valuable user behavior patterns in mobile commerce environments," in *PAKDD Workshops*, 2011, pp. 77–88.

[20] R. Kumar, Y. Lifshits, and A. Tomkins, "Evolution of two-sided markets," in *WSDM*, 2010, pp. 311–320.

[21] M. Ester, R. Ge, W. Jin, and Z. Hu, "A microeconomic data mining problem: customer-oriented catalog segmentation," in *KDD*, 2004, pp. 557–562.

[22] M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi, "Optimal spatial resolution for the analysis of human mobility," *ASONAM*, 2012.

[23] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *WWW*, 2012, pp. 769–778.

[24] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *KDD*, 2011, pp. 1082–1090.

[25] Y. Liu, Y. Zhao, L. Chen, J. Pei, and J. Han, "Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 11, pp. 2138–2149, 2012.

[26] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[27] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[28] L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti, "Understanding the patterns of car travel," *Eur. Phys. J. Special Topics*, no. 215, pp. 61–73, 2013.