Colloquial region discovery for retail products: discovery and application

Shashwat Mishra¹ · Vincent Leroy¹ · Sihem Amer-Yahia¹

Received: 6 January 2016 / Accepted: 5 June 2016 / Published online: 1 June 2017
© Springer International Publishing Switzerland 2017

Abstract Consumer behavior is of quintessential importance to the retail industry. Purchase trends for retail products are often affected by consumer location. Region-specific purchase trends, such as *the French love wine*, are typically supported by anecdotal evidence. An automated technique for the discovery of such trends from retail data has so far been absent. In this paper, we address the challenge of colloquial region discovery for retail products. More specifically, we target the problem of examining product sales across a chain of stores to extract the geographic regions that *characterize* a product. We introduce DICE, a diffusion-based technique to uncover all such regions for a given product, when they exist. In contrast to the current state of the art, DICE involves minimal usage of parameters and shows remarkable tolerance to noise that is often ubiquitous in retail data. We present results of experiments conducted on real datasets from a supermarket chain in France. Empirical evaluation and user studies establish that the proposed technique significantly outperforms the natural baseline and previous state-of-the-art approaches. Further, we study the impact of time and product category on DICE and discuss use cases for application of DICE in the retail world.

Keywords Retail · Diffusion · Characterizing region · Clustering

The authors acknowledge the generous support of Project ALICIA.

✉ Shashwat Mishra
shashwat.mishra@imag.fr

Vincent Leroy
vincent.leroy@imag.fr

Sihem Amer-Yahia
sihem.amer-yahia@imag.fr

¹ CNRS, LIG, University of Grenoble-Alpes, Grenoble, France

1 Introduction

Customer receipt analysis is a common task in the retail industry and has proven to be very effective for recommending products, implementing loyalty programs for customer rewards and discounts, and optimizing stocks for retailers. Uncovering areas that mark active consumption for consumer products is key to understanding how to provision stores in different locations. In this paper, we adopt a *product-centric* analysis or how to target specific consumer groups that contrasts sales of a given product in different stores to identify regional consumption. In order to do that, we introduce the notion of *characterizing region* for a product and address the challenge of discovering such regions. To the best of our knowledge, the presented body of work is the first to address such a problem.

Understanding consumer behavior is of vital importance for marketers and consumer-goods companies. Much work has thus been devoted to analyzing retail data from *consumer-centric* and *store-centric* viewpoints. While consumer-centric analyses are used to drive product discounts and recommendations, store-centric analyses focus on shelf space management and relative product placement in the store to increase revenue.

In this paper, we complement these approaches by introducing a *product-centric* analysis that combines in-store sales with geographic proximity to identify regional trends for a product. Such analysis is critical to product marketing. Popular shopping trends do indeed affect a customer's decision of which product to buy. In addition, geographic proximity plays an important role in a customer's decision to shop at one place or another. While some customers choose a store solely based on geographic proximity, others will travel the extra mile to find some products.

Fig. 1 Inadequacy of TopK for uncovering characterizing regions. **a** Heatmap for a product where *darker color* implies higher sales. **b, c** Characterizing regions extracted using the natural baseline (TopK) and our method (DICE). The proposed method returns smoother regions (color figure online)

Consider Fig. 1a which shows the heatmap computed using the sales of a certain product¹ in various stores across France. At each store, we compute the density of the product as the total units sold for the product normalized by the total units sold for all products at the store. The *hot zones* for the product can be immediately identified as the southern and western parts of France. These regions, where the consumers actively consume the product, characterize the said product. Our aim is to identify and isolate all such regions that exist for a product. Such *characterizing regions* are vital sources of information. Besides being a holistic cross-store view of a product's sales, they offer valuable information to the retailers and manufacturers. Most importantly, a key application that is hardly enabled by other analyses is cross-store advertising: a store may advertise that some products could be found in neighboring stores in order to balance its overall supply/demand chain. Characterizing regions could also be used by government instances that go beyond a single food supply chain to study the consumption of some products or product types and correlate that information with health indicators for different regions. The ability to combine trends with geographic proximity can also be very useful for cross-store management and provisioning.

A key challenge in the identification of characterizing regions is to account for both in-store sales of a product and geo proximity of stores. The basic tasks involved in uncovering characterizing regions are spotting areas deemed as *hot zones* and identifying boundaries for the characterizing regions within the *hot zones*. While it is straightforward to spot such areas for the product in Fig. 1a, obtaining them via manual inspection may not always be trivial. Even if one could spot the *hot zones*, realizing a boundary could be difficult as there may not be a clear contour demarcating sharp

fall in product sales. Automated techniques that can examine large volume of products to uncover quality characterizing regions, when they exist, are thus highly lucrative.

A natural way to address the problem is to return the regions constituted by the k stores that exhibit highest (normalized) sales for the product. Figure 1b shows the result obtained using such (TopK) scheme. While the resulting regions are consistent with the heatmap (Fig. 1a), they offer a fragmented view of the product's regional trends. Product sales across stores show high variance (noise) leading to several stray maxima or *outliers*, i.e., stores that lie far from *hot zones* but have sales higher than some stores in the *hot zone*. TopK erroneously selects such *outliers* over stores within the characterizing regions. Thus, it lacks the necessary smoothing required to obtain coherent regions.

Uncovering colloquial regions corresponding to entities from geo-tagged data has recently drawn attention of the research community [1–4]. The current state of the art is established in [5] where the authors develop a technique (SSR) to identify regions for geo-entities. The key idea is to interpret the tag's frequency distribution over the map as an image and identify contours that demarcate sharp changes in the frequency. The extracted contours are subjected to a series of image processing operations to obtain the characterizing region. The technique implicitly assumes the existence of a sharp boundary (in terms of sales) between the characterizing region and its surroundings. Colloquial region discovery for retail products is not amenable to such a solution because the assumption seldom holds true for retail data.

This analysis suggests that both the natural baseline, TopK, and the current state of the art for related problems, SSR, fail to uncover characterizing regions for retail products. There is a clear need for a solution that is robust to the noise in the data and effectuates the necessary smoothing required to obtain coherent regions.

¹ The exact product is hidden for confidentiality purposes.

To address the insufficiencies of existing methods, we propose a new approach, coined *Diffusion-based Iterative Characterizing region Exploration (DICE)*, that combines store sales and geographic proximity to find a set of coherent regions for a given product. Intuitively, the proposed technique starts with a seed set of stores that exhibit high sales for the product and iterates over neighboring stores to find the largest contiguous regions covering all high-sales stores or stores that are contiguous to high density ones. DICE models the diffusion of *product endorsements* to neighboring stores that delivers enhanced smoothing. As established by Fig. 1c, the obtained results are significantly better than the natural baseline. We present several experiments in Sect. 4, where we compare DICE with the natural baseline (TopK) and the current state of the art (SSR) and study the behavior of DICE with respect to seed selection, time, product type.

To summarize our findings, the performance studies conducted on real datasets from retail domain suggest that a large number of consumer products have associated characterizing regions. We identified over 450 such products out of 10k. Further, the quality of regions returned by DICE is significantly better when compared to the natural baseline (TopK) and the previous state of the art (SSR). To facilitate validation, we build ground truth with the aid of domain experts. We validate our results via both empirical comparisons that report over 25% gain when compared to previous methods and user studies that establish that majority of the users prefer DICE over competing methods and are in agreement (over 70%) with the characterizing regions returned by DICE. Further, we study the impact of time and discover that certain products are characterizing during specific periods or seasons around the year, such as *fruit punch* in coastal areas during summer (Sect. 4.5.1). We also discover that certain regions exhibit a liking for a category of products as whole with no clear favorites within the category (Sect. 4.5.2). We finish by presenting two use cases that exemplify how DICE can be used by analysts in retail world (Sect. 5).

This paper makes the following contributions:

1. We define the problem of finding characterizing regions of a product and argue for its significance.
2. We design and implement a novel diffusion-based algorithm that, given a product and its sales in different stores, finds characterizing regions of that product.
3. We run performance experiments that include an empirical evaluation as well as a user study with novice and expert users that validate our findings on real datasets from a large French supermarket. We compare ourselves to the natural baseline (TopK) and the previous state of the art (SSR). Our experiments establish that DICE significantly outperforms both.

4. We study the effect of time and product category on characterizing regions and report interesting insights.
5. We discuss two use cases that are made possible by characterizing regions. In particular, we use the obtained characterizing regions to build maps based on product preferences (Sect. 5.1) and to quantify variability in preference within each product category (Sect. 5.2).

Our paper is organized as follows. In Sect. 2, we formalize our data model and state the problem of identifying the set of characterizing regions of a product. Section 3 describes the design process for DICE. Section 4 is dedicated to our experiments and findings. In Sect. 5, we present some use cases for characterizing regions. We summarize the related work in Sect. 6. We conclude in Sect. 7.

2 Data model and problem

2.1 Model

We have a set of *products*, $P = \{p_1, p_2, \dots\}$, which represent distinct items available in a typical store. Examples of products are *tuna sauce*, *red wine*, *detergent* and *cat food*. For confidentiality purposes, we do not provide specific product names and brands in this paper.

We also have a set of stores, $S = \{s_1, s_2, \dots\}$, where each store s has a pair of geo-coordinates ($s.lat$, $s.lon$) using which we define the distance between two stores s_i and s_j as below.

$$g(s_i, s_j) = \sqrt{(s_i.lat - s_j.lat)^2 + (s_i.lon - s_j.lon)^2} \quad (1)$$

We define a bijective mapping from the set of stores to a set of *cells*. The *cell* for each store is simply a polygon enclosing that store, with the property that the closest store from any point within the cell is the enclosed store. Such a set of cells is easily obtained by computing the Voronoi map over store locations. The resulting Voronoi polygons form the cells. We show an example in Fig. 2.

We say that two stores are *connected* if the respective Voronoi polygons are adjacent. For a given set of stores, the region constituted by those stores may range from being completely fragmented (when no two stores are connected) to completely unified (when all stores are contiguous). We call the latter a *coherent* region.

A *transaction* is a purchase of some (nonzero) units of certain products at a store. We assume the existence of a set of transactions, $T = \{t_1, t_2, \dots\}$, such that each transaction $t \in T$ is a 2-tuple indicating the store at which the transaction occurred and the set of products purchased in the transaction. For example, the transaction $t = \langle s_k, \{p_l, p_m\} \rangle$ indicates that products p_l and p_m were purchased at store s_k .

Fig. 2 Distribution of stores across France. **a** Exact location of all stores while **b** computed Voronoi cells. Each cell corresponds to a single store. Color of each cell is governed by the total number of transactions that occur at the store. Darker color implies more transactions (color figure online)

Using T_s to denote the set of transactions occurring at store s , the total units of product p sold at s are computed as:

$$\text{sales}(p, s) = \sum_{t \in T_s} f(p, t) \quad (2)$$

Here, f is a function that gives the units of p sold in transaction t . We assume $f(p, t) = 0$ if p was not sold in transaction t .

We can now define the *density* of a product p at a store s , $d_p(s)$, as follows:

$$d_p(s) = \frac{\text{sales}(p, s)}{\sum_{p' \in P} \text{sales}(p', s)} \quad (3)$$

We compute the density of a product p at all the stores in S . This is referred to as the density distribution, D_p , of the product.

$$D_p = \langle d_p(s_1), d_p(s_2), \dots, d_p(s_{|S|}) \rangle \quad (4)$$

2.2 Problem

Our aim is to identify, given a product p , a set of stores, $S_p \subseteq S$, such that the stores in S_p have high product density and result in *coherent* regions. Note that a product may have more than one coherent region, as shown in Fig. 1c. We refer to each such coherent region as a *characterizing region* and S_p as the set of stores that constitute the characterizing regions.

A product may not have any meaningful characterizing region. For a product p , the set of its characterizing regions, S_p , is well defined if the product exhibits *localized* overconsumption. If the consumption is relatively uniform across all stores, or the stores with high density are scattered, the product may not be coupled to a geographic region and S_p is thus undefined. That is, for example, the case for generic

products that are consumed uniformly across all stores, such as *chewing gum* and *toilet paper*.

The technique we describe in this paper assumes the existence of characterizing regions for the product. Later we show how our solution offers out-of-the-box support for segregation of products that do have characterizing regions from those that do not in Sect. 4.4.1.

3 Methodology

Our objective is to analyze the density distribution of a product in order to identify characterizing regions for the product. We focus on geographically coherent regions, defined in Sect. 2 as the largest connected component in the associated Voronoi map. Intuitively, a coherent region should initially contain several neighboring stores exhibiting a high density for the product, as well as other stores with relatively high densities. Conversely, isolated high-density stores should not cause the emergence of a region. Thus, our aim is to design a smoothing process that can expand dense regions. The main challenge in this context is to balance the smoothing that allows DICE to include some regions, while retaining characteristics of the original data distribution.

As stated previously, the natural approach of selecting top k stores based on density works well for products with sharply defined characterizing regions, but it is not a viable solution for the majority of products. Further, it enforces that all products have the same number of stores (k) in the final result, which may not reflect true behavior.

We observe that the popularity of a product at a store is an *endorsement* of the product by that store. This endorsement may affect the product's popularity in neighboring stores as stores may influence other stores. Modeling this diffusion gives us a more refined method to achieve the desired smoothing.

Fig. 3 Illustration of different stages in DICE. Normalized density distribution (represented as heatmap in **a**) forms the starting point. Selection of the desired size for the seed set (m) results in selection

of top- m stores as seeds (**b**). The activated stores influence other stores potentially activating them (**c**). The region constituted by the set of activated stores is reported as the characterizing region (**d**)

Diffusion over graphs has been well studied in different versions of the influence maximization problem first introduced by Kempe et al. [6]. The problem involves a scenario where a user initially activates certain nodes in a social graph. The activated nodes may then influence neighboring nodes and activate them. The process stops when no nodes can be activated. The objective is to find the set of k initial nodes to activate to cause maximum spread. To solve the problem, one has to assume an *influence propagation* (IP) model. The IP model specifies how the influence spreads from an active node to an inactive node. Relevant literature offers a choice between two popular IP models, independent cascade (IC) and linear threshold (LT). Put together, they characterize different types of social interactions [7–12]. While the IC model assumes independent interaction for each pair of nodes, the LT model uses threshold-based behavior.

To design DICE, we select the linear threshold (LT) model of influence propagation as it allows us to incorporate influence from multiple stores simultaneously and does not require to assume independence between stores. We now give an overview of our algorithm and discuss its two core processes.

Overview of the algorithm For a given product p , DICE starts with its (normalized) density distribution D_p . It selects an initial *seed set* of stores (Fig. 3b). These stores are marked as active and the remaining stores as inactive. The activated stores cast an influence on inactive ones which in turn may become active. The algorithm iterates and stops when no inactive stores can be activated (Fig. 3c). DICE then extracts coherent regions as significant connected components covering the activated stores (Fig. 3d). We now go over the salient features of IP models and discuss them in the context of our problem.

3.1 Seed selection

In influence propagation, the seeds represent the initial *active* entities, i.e., the entities that initiate the spread of a trend. Tra-

ditionally, influence propagation models are used to compute, given a social network, an optimal seed set to maximize overall influence spread [6]. DICE however aims at analyzing past sales records. Hence, the seed set of a product p is selected from the input data using D_p . Seed selection plays a crucial role in influence propagation. The primary goal of seeds is to bootstrap the process of diffusion. DICE selects the seeds as the stores that, given D_p , appear to be the most characterizing. $Seeds(p, m)$ is defined as the set of m stores having the highest density for product p .

$$\begin{aligned} Seeds(p, m) &\subseteq S, |Seeds(p, m)| = m \\ d_p(s) &\geq d_p(s') \quad \forall s \in Seeds(p, m), \quad s' \in S \setminus Seeds(p, m) \end{aligned} \quad (5)$$

Our aim is to design an algorithm that uses as few parameters as possible. The formalization of seed-set selection introduces m , number of desired seeds, as the only parameter admitted by DICE. In practice, the method is robust and only needs to select a sufficient amount of seeds to ensure that each characterizing region of p contains a few of the selected seeds. As long as this condition is met, influence spreads in those regions resulting in their discovery by the method. Minor increase in number of seeds does not have an impact. However, selecting too many seeds tends to add noise to DICE's results as the resulting large number of outliers may unify into erroneous regions. Our experiments in Sect. 4 reveal that stable results are returned by DICE for a large range of values for m .

3.2 Influence spread

Influence propagation recursively marks entities as active as they get influenced by a trend. Active entities spread influence to other entities, which in turn may become active. Propagation ends when no new entity gets activated. In our model, given a product p , the set of active stores S_p is initialized

to $Seeds(p, m)$. A non-seed store s becomes active and is added to S_p if the influence it receives exceeds its threshold θ_s^p . We define the value of this threshold as the difference between the (normalized) product density at s and the lowest (normalized) density of a seed.

$$\theta_s^p = \left(\min_{x \in Seeds(p, m)} d_p(x) \right) - d_p(s) \quad (6)$$

DICE assigns a weight factor w_s to each store s . Since larger stores attract more customers, and therefore constitute a stronger source of evidence for finding characterizing regions, w_s scales with the total number of sales at s , with a logarithmic damping factor.

$$w_s = \log \left(1 + \frac{(e-1) \sum_{p \in P} sales(p, s)}{\max_{x \in S} \sum_{p \in P} sales(p, x)} \right) \quad (7)$$

We ran experiments to evaluate different choices for weighting the stores (including $w_s = 1$). Logarithmic formulation, as in Eq. 7, delivered best results. Intuitively, Eq. 7 assigns higher weights to stores with greater sales, thus allowing them to spread more influence. The store with the greatest number of sales is assigned a weight of 1.

Here, e is the base of the natural logarithm. Upon its activation, a store s_i influences another store s_j by b_{s_j, s_i} , which contributes to the possibility of recursively adding more stores to S_p , and is defined as follows:

$$b_{s_j, s_i} = \frac{w_{s_i} \cdot d_p(s_i)}{g(s_i, s_j)^2 \sum_{\substack{x \in S, \\ x \neq s_i}} \frac{1}{g(s_i, x)^2}} \quad (8)$$

At any point of the recursive influence spread, the total influence received by an inactive store s is evaluated as $\sum_{x \in S_p} b_{s, x}$ and is compared against the threshold θ_s^p .

DICE focuses on regional trends, so we postulate that a store with high density (for a product p) influences other stores and attempts to *induce* the popularity of the product in its neighboring stores. We set the influence of a store on another to decay with the squared distance between these stores. Figure 2a shows that stores are far from being evenly geographically distributed: populated regions, around major cities, have a lot of stores, while rural areas are much sparser. The distance between neighboring stores varies significantly, which is amplified by the squared factor in the decay model. This requires DICE to be able to take into account the surroundings of a store s to parameterize $b_{*, s}$, such that isolated stores may still spread a significant amount of influence, while limiting it on populated regions to avoid snowball effects. The chosen formulation (Eq. 8) essentially amounts to selecting the total influence that a store can spread ($w_{s_i} \cdot d_p(s_i)$) and distributing it among stores according to

their distance. This ensures that the maximum influence cast by any store s_i over any store s_j is bounded by $w_{s_i} \cdot d_p(s_i)$.

We remark that DICE does not completely eradicate the fragmentation issue incurred by the natural baseline (TopK). S_p may still contain stores that are clear *outliers*, i.e., lie far from *hot zones*. However, the fraction of such stores is very small as compared to the stores that form connected components. As shown in Sect. 4, DICE significantly mitigates the fragmentation incurred by the natural baseline (TopK).

3.3 Detailed algorithm

The details of DICE are presented in Algorithm 1. S_p is first initialized to be the set of seeds following the approach described in Sect. 3.1, line 1. I is the set of remaining stores that can potentially become active through propagation (line 2). The accumulated influence for each store in I is initialized to 0 (line 4). DICE then loops until I converges by monitoring A_{prev} , the set of stores activated in the last iteration (line 7). DICE updates the influence accumulated by inactive stores of I with the influence they receive from stores activated in the last iteration A_{prev} (line 11) following the update rules defined in Sect. 3.2. If the influence received by an inactive store exceeds its threshold (line 13), it becomes active and is added to A_{cur} , the set of stores activated in this current iteration. I , A_{prev} and S_p are then updated at the end of each influence propagation iteration. After convergence, DICE returns the set of active stores S_p (line 21).

Algorithm 1 Diffusion-based Iterative Characterizing region Exploration (DICE)

Input: p : product, m : seed-set parameter

Output: S_p : Set of stores that form the characterizing regions

```

1:  $S_p \leftarrow Seeds(p, m)$ 
2:  $I \leftarrow S \setminus S_p$ 
3: for all  $s \in I$  do
4:    $s.inf \leftarrow 0$ 
5: end for
6:  $A_{prev} \leftarrow S_p$ 
7: while  $A_{prev} \neq \emptyset$  do
8:    $A_{cur} \leftarrow \emptyset$ 
9:   for all  $s \in I$  do
10:    for all  $x \in A_{prev}$  do
11:       $s.inf \leftarrow s.inf + b_{s, x}$ 
12:    end for
13:    if  $s.inf \geq \theta_s^p$  then
14:       $A_{cur} \leftarrow A_{cur} \cup \{s\}$ 
15:    end if
16:  end for
17:   $I \leftarrow I \setminus A_{cur}$ 
18:   $A_{prev} \leftarrow A_{cur}$ 
19:   $S_p \leftarrow S_p \cup A_{prev}$ 
20: end while
21: return  $S_p$ 

```

4 Results

In this section, we report on experiments conducted to study the behavior of DICE and evaluate its performance in comparison with its competitors. All experiments are conducted on a real dataset that consists of retail transaction logs. Our performance experiments conclusively establish that DICE significantly outperforms both TopK and SSR. According to our user studies, the percentage of users that prefer DICE, TopK and SSR are 47, 21 and 32%, respectively. The percentage gains achieved by DICE over TopK and SSR, using empirical measure, are 25 and 99%, respectively.

In the following headings, we first describe the setup, (Sect. 4.1). This is followed by a succinct description of the competing methods, natural baseline (TopK) and the previous state of the art (SSR) in Sect. 4.3. Subsequently, we discuss the various experiments in Sect. 4.4.

4.1 Setup

4.1.1 Data

We conduct our experiments over logs generated at a popular general commercial supermarket that owns over 1800 outlets in mainland France. The logs were gathered over a period of 27 months from May 1, 2012, to July 31, 2014. The data consist of a large number of *csv* files in different domains such as *stores*, *transactions*, *products* and *customers*.

Each line in *stores* corresponds to a store and consists of a unique store id and associated meta data (number of employees, geo-coordinates, opening date, etc.). In similar fashion, each line in *transactions* and *products* contains an entry for unit transaction and unit product, respectively. The employed data model is in agreement with the discussion under Sect. 2. Raw data were subjected to some cleaning operations as described next.

4.1.2 Stores

Stores with fewer transactions may result in superfluous densities. We filter the set of stores to retain those that have a sufficient number of transactions ($|T_s| \geq 20k$). This results in more than 1400 stores. In Fig. 2, we show the location of obtained stores on the map of France and the associated Voronoi map.

4.1.3 Products

Availability of certain products was limited to a select few stores. Our focus in this study is on products that are widely available yet over-consumed in certain locales. We filter the original set of products to retain only those that are available in at least η fraction of the stores (coverage). In Fig. 4, we

Fig. 4 Distribution of products w.r.t. fraction of stores covered (η). The X-axis shows the different coverage values. The Y-axis shows the number of products with coverage $\geq$ specified value

Table 1 Statistics of dataset (post filtering)

Set	Cardinality
Number of products ($ P $)	10,650
Number of stores ($ S $)	1,426
Number of transactions ($ T $)	1,108,748,098

show the number of products for different values of η . We fix η to 0.9 obtaining roughly 10k products. We remark at this point that not all products are expected to have characterizing regions.

Table 1 reports the final cardinalities of the sets. We examine over 1 billion transactions across more than 1400 stores for more than 10k products.

4.2 Platform

All experiments were run on an Intel E5-2650L machine with 32 GB RAM, running CentOS 6.5. Each algorithm was implemented in Java and executed on JRE 1.6.0_31.

4.3 Competitors

In the following, we introduce the natural baseline (TopK) and the current state-of-the-art approach (SSR) to uncover characterizing regions for products.

4.3.1 TopK

TopK is the trivial algorithm for colloquial region discovery. It returns the stores with the k highest densities as the result S_p . This forms the natural baseline for discovering characterizing regions given a product's density distribution. The choice of k is non-trivial. By design, the baseline lays no

Fig. 5 Impact of seed size (m) on two different products, *peculiar meat* and *chewing gum*. Initial seed-set stores are indicated in red while those added during diffusion in gray. The seed-set stores are more localized for product *peculiar meat* than for *chewing gum*. Consequently, the

former encounters more growth (larger number of gray stores) at both values of m during the diffusion step. **a** peculiar meat, $m = 50$, **b** peculiar meat, $m = 100$, **c** chewing gum, $m = 50$, **d** chewing gum, $m = 100$ (color figure online)

emphasis on obtaining stores that result in *coherent* regions. If the characterizing regions are sharply defined, i.e., the density of stores that comprise the regions is greater than all stores outside the regions, TopK delivers considerably good results given a suitable k . However, for most products the assumption does not hold and resulting characterizing regions are fragmented. Note that TopK shares certain aspects with DICE as the latter uses the former for seed selection which bootstraps the diffusion process.

4.3.2 Scale space representation (SSR)

SSR [5] establishes the current state of the art for characterizing region discovery. The authors attempt to uncover colloquial region boundaries for location related Flickr tags (such as “germany” and “poland”) from a large volume of tagged photographs.

The method starts by dividing the map into uniform grids and computing frequency of the given tag in each grid. This gives a two-dimensional representation of the tag’s frequency distribution over the map, which is interpreted as an image. In the next step, the method identifies all contours in the image across which the frequency changes abruptly. This is done by subtracting a high blurred version of the image from a low blurred version of the image. This step results in a binary map that highlights the aforementioned contours.

In the final step, the binary map is subjected to *contour filling* and *morphological closing* to arrive at closed connected *blobs*. Of potentially several resultant blobs, the authors retain the largest. The grids that constitute this blob are returned as the characterizing region. The approach thus returns a single characterizing region per product which may not be true for all products.

Note that the data model above closely resembles the one described in Sect. 2. A photograph corresponds to a transaction and the tags within the photograph correspond to products. The two models differ slightly in that pho-

tographs can manifest at any arbitrary point on the map, while the transactions are constrained to occur at fixed locations (stores).

Note further that the instantiation of smoothing and image processing operations in SSR requires specification of multiple parameters. The approach is thus highly parametric (requiring specification of up to four parameters). The quality of obtained regions is sensitive to the choice of parameters; thus, it is required that the system be trained on some supplied ground truth to obtain optimal parameter combination.

4.4 Experiments

We conduct three different sets of experiments. The first set of experiments study the impact of seed size (m) on DICE (Sect. 4.4.1). The second and third sets of experiments compare the performances of DICE, TopK and SSR using an empirical approach (Sect. 4.4.2) and a user study, respectively (Sect. 4.4.3).

4.4.1 Impact of seed-size parameter

DICE admits a single parameter, the initial seed-set size, m . Intuitively, $|S_p|$ is expected to increase with increase in m , because a higher value of m implies a bigger seed set which (a) by design is part of the result and (b) implies greater number of initially active stores. However, different products are expected to behave differently for the same seed size m depending on whether they exhibit meaningful characterizing regions. We present a set of experiments to examine these aspects of the seed size parameter.

First, we study how different products behave differently for the same value of m . The growth incurred during diffusion depends on the relative location of m seeds which may be different for different products. We illustrate this with the help of an example. Figure 5 shows the characterizing regions, obtained via DICE, for two different products for two differ-

Table 2 Impact of seed size on size of characterizing regions ($|S_p|$)

Product	$ S_p (m = 50)$	$ S_p (m = 100)$
<i>Peculiar meat</i>	71	175
<i>Chewing gum</i>	56	121

ent values of m . While the first product is a *peculiar meat*, the other is a common *chewing gum*. We mark the stores that form the initial seed set in red. Note that the stores that comprise the seed set are more localized for *peculiar meat* than for *chewing gum*. In Table 2, we show the size of the final characterizing regions which establishes that the former product grows significantly more than the latter. This behavior allows us to segregate products that have meaningful characterizing regions from those that do not by using a suitable threshold. For the following discourse, where we evaluate the quality of extracted regions, we restrict ourselves to products that resulted in meaningful characterizing regions.

Next, we study how the characterizing region varies with growth in m . Ideally, we would expect the number of stores added during the diffusion to remain the same for a range of m as it removes the burden of selecting a *good* seed size. To demonstrate that DICE achieves this behavior, we conduct a study where we study products in pairs. Each pair consists of two products: one with a meaningful characterizing regions and one without. For each product in the pair, we compute the difference in the size of characterizing regions and the seed size (thus, the number of stores added during diffusion) across different seed sizes. Figure 6 shows the results obtained for the pairs $\langle \text{Porc}, \text{Baguette} \rangle$ and $\langle \text{Butter}, \text{Diapers} \rangle$ where the former product is the one with meaningful characterizing region. Analyzing the first pair, we observe that as the seed size is increased, *Porc* first reaches a plateau ($m \in [20, 140]$) and then grows gradually. In contrast, *Baguette* grows gradually with m throughout. We conclude that for seed size values demarcating the plateau, the size of the characterizing region does not change significantly. We obtain similar results for the second pair where the plateau

is observed for *Butter* after a period of initial growth. DICE thus affords stability in the result for products with meaningful characterizing regions, thereby making it easier to use.

4.4.2 Empirical comparison

Evaluation of the quality of returned characterizing regions is difficult in the absence of a ground truth. Previous studies on similar theme [4, 5] have exclusively dealt with extraction of colloquial boundaries for geographic concepts (such as “germany”) for which the ground truth is easily available. The region extracted for tag “germany” should overlap with the international boundaries of the eponymous country. However, for consumer products, it is difficult to affirm with certainty that the over-consumption of a certain product should be restricted to certain locales.

Ground truth We attempt to fill the void created by the absence of ground truth with the aid from our industrial partner. We construct a sequence of 2-tuples consisting of products and French administrative regions, i.e., every tuple is of the form $\langle p, r \rangle$ such that $p \in P$, r is an administrative region of France and p is expected to be characterizing in r . Experts were instructed to report only such pairs which could be argued for with high confidence. On account of the absence of data, the administrative region of *Corsica* was omitted. For each of the remaining 21 administrative regions of France, we obtained two products that were expected to be characterizing in the said region. This resulted in 42 ground-truth pairs of the form $\langle p, r \rangle$.

Evaluation scheme Each of DICE, TopK, and SSR is evaluated against each pair $\langle p, r \rangle$ of ground truth by measuring the Jaccard overlap (in terms of area) between the characterizing regions for p (as returned by the method) and the French administrative region r . Subsequently, the Jaccard overlap is averaged over all the pairs. *Average Jaccard overlap* serves as the primary empirical measure for comparing the three methods where higher values are desirable. We also compute and report the maximum and minimum over all ground truth pairs.

Fig. 6 Variation in number of stores added during diffusion with seed size (m). Products with meaningful characterizing regions plateau for a range of seed size values (*Porc* for $m \in [20, 140]$ and *Butter* for $m \in [60, 140]$). Generic products not tied to a region show linear dependence on seed size m

Table 3 Default parameters for experiments

Algorithm	Parameter	Value
DICE	Seed-set size (m)	100
SSR	Kernel width (σ)	74
SSR	Morphological radius (ρ)	16
SSR	Gaussian kernel size	45
SSR	ε	5×10^{-4}

Parameterization A key benefit of DICE is its dependence on a single parameter, the initial seed-set size (m). We observed that for most products, there exists a range of seed size (m) for which the resulting regions are nearly the same (Sect. 4.4.1). Thus, we keep the seed-set size fixed at 100 for all subsequent experiments as it returned quality results for a wide choice of products.

A serious limitation in TopK when compared to SSR and DICE is that the characterizing regions for all products contain the same number of stores, k . Fixing k fixes the number of stores in the returned regions. A single value of k may not be suitable for several products simultaneously. Choosing k is thus non-trivial. DICE does not suffer from this limitation because though the seed-set size m is same for all products, the number of stores added during the diffusion phase is product specific.

We give TopK the flexibility to return characterizing regions of different sizes in the following manner. For each product, TopK uses the number of stores in the characterizing regions returned by DICE as the value of k , i.e., $k = |S_p^{\text{DICE}}|$. Thus, TopK is executed with different values of k for different products. One could argue that this gives undue advantage to TopK by giving it a calculated guess for k . As we show later, despite the added advantage, TopK is outperformed by DICE.

In contrast, SSR requires proper specification of up to four parameters. Of the four parameters listed in Table 3, two, Gaussian kernel size and ε , are fixed to values specified in [5]. Optimal values for the remaining two parameters are obtained using recursive random search [13] over the ground-truth pairs. We explore the range $[0, 1000]$ for σ and $[0, 200]$ for ρ . In Fig. 7, we show the variation in Jaccard overlap achieved by SSR (averaged over all ground-truth pairs) w.r.t its parameters, in the vicinity of the optimal values. The optimal aggregate Jaccard overlap achieved by SSR is 0.18.

We list the complete set of parameters for all three algorithms with their default values in Table 3. We abstain from listing the default value for TopK since instantiating DICE instantiates TopK.

Result Finally, we present the results of the empirical comparison in Fig. 8. DICE achieves significantly higher values for average Jaccard overlap (*Average*) than TopK and SSR.

Fig. 7 Parameter space exploration for SSR [5]. Optimal average Jaccard overlap is obtained at *Kernel Width* = 74 and *Morphological Radius* = 16 **Fig. 8** Performance evaluation of TopK, SSR and DICE against the 42 ground-truth pairs. Seed-set size for DICE (m) was fixed at 100. For TopK, $k = |S_p^{\text{DICE}}|$. Parameters for SSR were fixed at $\sigma = 74$ and $\rho = 16$

The relative gains are 19.6 and 99.02%. DICE also achieves the highest value for maximum Jaccard overlap (*Max*) and minimum Jaccard overlap (*Min*) against a ground-truth pair.

For deeper examination, we explicitly present and discuss the characterizing regions returned for two products, *detergent* and *canned tomatoes*, in Figs. 9 and 10. From Fig. 9, we establish that TopK (Fig. 9b) is consistent with the heatmap (Fig. 9d), but more fragmented, whereas SSR (Fig. 9c) picks up high-frequency variation within the characterizing region and results in a characterizing region that is too small. DICE (Fig. 9a) most accurately captures the characterizing regions for the said product.

As compared to the product in Fig. 9, the product in Fig. 10 has less sharply defined characterizing region (i.e., the product density in the characterizing region is not significantly higher than elsewhere). This can be ascertained from their respective heatmaps (Figs. 9d, 10d). This increased noise has adverse effects on both TopK and SSR. For TopK, Fig. 10b,

Fig. 9 Illustration of characterizing regions obtained for the ground-truth pair (*detergent*, *Languedoc-Roussillon*). While TopK (b) significantly overlaps with *Languedoc-Roussillon*, it includes several distant stores. SSR (c) picks up the sharp density variation within the charac-

terizing region and results in a region that is too small. DICE delivers the best result (a). Jaccard overlap with the boundaries of *Languedoc-Roussillon* region of France are 0.41 (DICE), 0.31 (TopK) and 0.02 (SSR)

Fig. 10 Illustration of characterizing regions obtained for the ground-truth pair (*canned tomatoes*, *Picardie*). TopK (b) results in high fragmentation and SSR (c) results in a very large region. Jaccard overlap

achieved by the three approaches with the French administrative region of *Picardie* are 0.36 (DICE), 0.17 (TopK) and 0.04 (SSR)

the severity of incurred fragmentation increases substantially. SSR, Fig. 10c, results in too large a region, covering almost all of France. This occurs as the high-frequency variations (ubiquitously present in the heatmap) give rise to several contours (demarcating sharp fall in product density) in the binary map of SSR. Upon morphological closing, the contours unify into a single connected component.

Note that DICE is also adversely affected by increased noise as established by Figs. 9a and 10a. However, the mitigating effect of diffusion-based design counters the introduced noise enabling DICE to deliver best results for both products. The Jaccard overlap as achieved by DICE, TopK and SSR are 0.41, 0.31, 0.02 for *detergent* and 0.36, 0.17, 0.04 for *canned tomatoes*.

We also remark that SSR was observed to consistently result in regions that are either too small or too large. Figures 9c and 10c present the two different extremes achieved by SSR.

To substantiate our argument about fragmentation, we quantified the incurred fragmentation (as the average pairwise distance between stores in S_p) and measured it for

different choices of seed-set size (m). We average the results over all products that were observed to give meaningful characterizing regions by DICE (455 products). Figure 11 shows the results. TopK incurs more fragmentation than DICE. In either case, the fragmentation increases with an increase in seed-set size, since larger seed sets result in larger characterizing regions that are spread over a large area. At higher values of m , both methods return similar regions and thus incur the same amount of fragmentation.

4.4.3 User study

We conduct two different types of user studies. The first study (comparative evaluation) is done to perform a human assessment of the quality of the results obtained by the three methods. The second study (independent evaluation) captures the utility of results returned by DICE.

Each study was independently conducted on two different groups of respondents, *Users* and *Experts*. The group *Users* consisted of 30 novice users from France, while the group *Experts* consisted of eight experts from the retail domain who

Fig. 11 Mitigation of fragmentation by DICE. Y-axis shows the average pairwise distance between stores in S_p . X-axis shows the seed-size parameter (m) for DICE. S_p^{TopK} is more fragmented than S_p^{DICE}

Table 4 User study: comparative evaluation

	DICE	TopK	SSR
Users	46.55	21.13	32.32
Experts	47.46	34.19	18.35

had in-depth knowledge of products and their consumption trends. We now present the two user studies.

Comparative evaluation The first study provides a comparison of the *goodness* of the three methods as perceived by average users and marketing experts. We generated, randomly, a list of 20 products and the corresponding characterizing regions returned by all three methods. For each product, the respondents were tasked with indicating which of the three methods best reported the characterizing regions for the product. The maps were anonymized and randomly permuted to eliminate any bias. In Table 4, we present the votes received by each method averaged over all user and products.

DICE establishes itself as a clear favorite among both *Users* and *Experts*. While SSR outperforms TopK for *Users*, the opposite holds for *Experts*. We attribute this observation to the fact that *Users* exhibit an *esthetic bias* that influences them to vote for the method that returns esthetically pleasant characterizing regions. SSR, due to its region selection step, returns a single contiguous region for each product which is appealing to *Users*. *Experts*, being more informed about the products' consumption trends, do not suffer from this bias.

Independent evaluation The aim of the second user study is to perform an independent evaluation of DICE. For each product in a randomly generated list of ten, users were asked to express their level of agreement that the shown map (obtained via DICE) correctly highlights all parts of France expected to be characterizing for the product. Agreement levels were

Table 5 User study: independent evaluation

	Not agree	Weak agree	Moderate agree	Agree	Strong agree
Users	24.05	23.53	28.47	18.33	5.62
Experts	10.24	34.58	27.62	19.52	8.04

expressed on a scale of 1–5. Results were averaged over all users and all products.

We present the final figures in Table 5. More than 70% of *Users* and 80% of *Experts* show some level of agreement with the results. *Experts* show greater agreement than *Users* which may again be attributed to the *esthetic bias* as *Experts* are more likely to show agreement with non-trivially shaped regions.

4.5 Qualitative evaluation

Our discussion so far made some implicit assumptions. In particular, we assumed that (a) product preferences do not change with time and (b) customers can only exert preferences for individual products and not product categories. In practice, neither of these assumptions need be true.

Region-specific product preferences could potentially vary with time. For example, the southern regions of France, particularly those close to the coast, consume more *table wine* than other regions in summer, whereas in winter the consumption is roughly uniform throughout France. Further, it is possible that a certain geographic region may prefer all products under a certain category. This is known to be the case for *Bretagne* region of France that prefers most pork-based products and has no clear favorites within that category. Uncovering such insights requires us to factor in the product hierarchy before we execute DICE.

In this section, we break these assumptions and explore the dimensions of time and product category.

4.5.1 Effect of seasons

While some product preferences remain stable with time, certain others often vary with the time of the year. This is especially true for seasonal vegetables, festival delicacies or beach products such as *sun screen*. It is therefore expected that the presence of characterizing regions, for a certain product, could be conditional on the time period.

We study the behavior w.r.t. time by measuring the changes in the characterizing regions for different seasons. To do this, we define two seasons, *summer* (consisting of June, July and August) and *winter* (consisting of November, December and January). We use the timestamp available with each transaction to generate sales for each season.

Table 6 Product region pairs with seasonal dependence

Product	Region	Season
Pistachios	West	Winter
Edible Cream	South Central	Winter
Sauce(x4)	East	Winter
Lemon Ice cream	South East	Summer
Mint Ice cream	South East	Summer
Fruit Punch	South Central	Summer

Our aim is to identify products that have meaningful characterizing regions in exactly one of the two seasons. We achieve this by simply comparing the two characterizing regions (one for each season) for all products. In Table 6, we show examples of such products along with the region and the season for which they were observed to be characterizing. We observe that coastal areas in the south of France consume more *lemon or mint flavored ice cream* in summer. West, in particular Bretagne, prefers pistachios but only in winter.

To take a deeper look, we present in Fig. 12, the characterizing regions obtained for the two seasons for a type of *sauce* and *fruit punch*. While *sauce* has no meaningful characterizing regions in summer (Fig. 12a), in winter it becomes characterizing in the east (Fig. 12b). The opposite holds true for *fruit punch* which exhibits characterizing regions in coastal areas during summer but not in winter (Fig. 12c, d). This latter observation is intuitive as *fruit punch* is a typical coastal drink and southern France experiences more beach activity than other regions during the summer season.

Note that identifying products whose characterizing regions differ across seasons is just one of the many objectives that are made possible with the introduction of time. One could, for example, pick a finer temporal granularity and monitor shifts in the observed characterizing region with

Table 7 Products hierarchy

Depth	Description	#Nodes
0	Root	1
1	Domain	11
2	Cluster	44
3	Family	650
4	SubFamily	3774

time. Such insights, enabled by DICE, could make valuable information to retail personnel tasked with designing promotion campaigns, marketing schemes, etc.

4.5.2 Effects of product hierarchy

So far, the target of our interest has been individual products. An important component of the universe of products (of which individual products form the fundamental building block) is product categories. It is certainly possible for regions to have preference for a product category in general and have no clear favorites within the category. In this section, we conduct experiments to explore this aspect in greater detail. To do this, we first introduce some terminology associated with the product hierarchy.

Our industrial partners provide us with a hierarchy of product categories H . The product hierarchy is a tree (of depth 4) where each node represents a concept. The coarseness of concepts decreases with increase in depth. For example, depth 1 contains nodes such as *boutique*, while depth 4 contains nodes such as *strawberry lip balm*. At depth 0, we have a single node, *root*. Table 7 shows the distribution of number of nodes at various depths. We use H_i to denote the set of nodes at depth i .

Each product $p \in P$ is associated with exactly one node at depth 4. This gives us immediately, for each p , a set of ancestors (one at each depth) that denote a fine-to-coarse categorization of product p . We refer to the ancestor of product p at depth i as H_i^p .

Fig. 12 Variance in characterizing region with seasons. We show the characterizing region obtain for two products during winter and summer. **a, b** Results for *sauce*. **c, d** Results for *fruit punch*. While former product exhibits characterizing region in winter, the latter shows preference for summer

Fig. 13 Illustration of normalization by hierarchy on different products. **a, b** Global and hierarchy heatmap for peanut oil. **c, d** Same for sunflower oil. Normalization by hierarchy has little effect on *hot regions* for *peanut*, while it dissolves the said regions for sunflower oil

Fig. 14 Illustration of normalization by hierarchy on different products. **a, c** Global heatmaps for two different brands of *edible creams*. Note that the two brands agree on the hot region. **b, d** Respective hierar-

chy heatmaps. While hierarchy heatmap of brand *B* shows no hot regions, that of brand *A* translates to the south

Recall the *density* of a product at a store, Eq. 3. Note that the sales are normalized by the total sales of *all* the products at that store. An alternative could be to normalize *by hierarchy*. Given a product, p , and a suitable choice of depth, i , we could recompute the hierarchy normalized density of p at store s as below.

$$d_p(s) = \frac{\text{sales}(p, s)}{\sum_{H_i^q = H_i^p} \text{sales}(q, s)} \quad (9)$$

If consumers in a certain region r exhibit preference toward a category (H_i^p) instead of the product (p) hierarchy normalized density distribution should contain no characterizing regions. If preference is specific to the product, the characterizing regions should persist. Discovering such insights necessitates comparing results obtained from the two normalization alternatives. We refer to the heatmap of a product obtained by normalizing over all products as the *global heatmap* whereas that obtained by normalization w.r.t. hierarchy as the *hierarchy heatmap*.

We present an example in Fig. 13 where we show the heatmap for two different types of oil, *peanut* and *sunflower*. In France, it is widely believed that the south prefers oil-based cuisine. Consistent with domain knowledge, global

heatmaps for both products report hot regions in southern parts of France (Fig. 13a, c). While the hot zones for peanut oil persist in the hierarchy normalized map (Fig. 13b), they disappear for sunflower oil (Fig. 13d).

Note that it is also possible for a product that has no apparent hot zones in the global heatmap, to have one in the hierarchy heatmap, as we show in the example in Fig. 14. We study the global and hierarchy heatmaps for two different brands of edible cream, *A* and *B*. Both brands show similar hot zones in the global heatmap, Fig. 14a, c. When normalized by the product category *creams*, hot regions dissolve for brand *B* but, more interestingly, for brand *A* they shift southward. This new region that appears in the hierarchy heatmap but not the global heatmap for product *A* is essentially a region constituted by consumers who do not show preference toward the product *A* in general, but when faced with buying a cream, prefer *A* over other creams.

5 Characterizing regions: application

In this section, we present some use cases for direct application of characterizing regions in retail. Our aim is to showcase how the analyst can benefit from DICE. We ask ourselves

Fig. 15 Reorganizing regions based on product preferences. Realized by hierarchical clustering of stores, using characterizing products (Eq. 10) to compute similarity between stores. All merges with similarity $\geq \tau$ were realized. Each color represents a maximally merged node during hierarchical clustering (community) **a** $\tau = 0.1$, **b** $\tau = 0.2$ (color figure online)

the following two questions. *Q1: Can administrative boundaries be reorganized based on product preferences?* and *Q2. Can characterizing regions help quantify market competition in different domains.* In the following headings, we take a deeper look at both questions and attempt to answer them using DICE as a black box.

5.1 Q1: Map repartitioning

Geographic regions are typically exhaustively partitioned into smaller geographic regions to ease governance and administration. The administrative boundaries are realized using a complex mix of meteorological-socio-cultural-geo factors such as climate, language, landscape and rivers. However, product preferences may transcend administrative boundaries. The reason could be either of two; administrative boundaries are not always strictly aligned along contours that demarcate sharp change in the aforementioned factors, or there may be hidden factors that drive the consumption of the said product(s). It has been shown previously that popular understanding of geographic regions is fuzzy and does not adhere to administrative separation [14, 15].

Segmenting geographic regions on the basis of their product preferences (as opposed to administrative boundaries) helps marketeers address the provisioning and logistics challenges in a refined manner. Retail companies are interested in defining “regions” on the basis of homogeneous product preferences. It would therefore be interesting to reorganize the administrative regions of a country based on product preferences. We now describe how this can be achieved using DICE by clustering places not on the basis of spatial proximity but product preferences.

Recall that the set of stores constituting the characterizing regions for a product p is specified as S_p . For each store $s \in S$, we compute set P_s as defined below.

$$P_s = \{p \mid p \in P, \quad s \in S_p\} \quad (10)$$

Equation 10 states that for a store s , the set P_s contains each such product p where s was part of the characterizing regions of the product, S_p . We say that each product $p \in P_s$ is a characterizing product of s and the P_s characterizes the store s .

We quantify the *product preference similarity* between two stores s_1 and s_2 by computing the similarity between the sets P_{s_1} and P_{s_2} . This is then used to perform a hierarchical clustering of the stores. We traverse the generated dendrogram to identify *communities*. A *community* is a *maximally merged node* in the dendrogram, i.e., all such nodes where the similarity $\geq \tau$ for all descendents and $< \tau$ for all ancestors. We use N to denote the number of distinct *communities* obtained for a chosen threshold τ .

Figure 15 shows the results of clustering obtained for two different choices of τ where each *community* is shown in a different color. Expectedly, increasing τ increases N as the strictness of merge increases with τ . We observe that the resultant communities are almost always contiguous, i.e., they are not fragmented and form a connected whole. This is interesting, especially considering that neither the similarity comparison (Eq. 10) nor the hierarchical clustering imposes any contiguity constraints. Of the 16 communities shown in Fig. 15a, only 1 is fragmented.

We also notice that the discovered boundaries are different from the administrative boundaries. We observe several interesting intricacies. The southwest region of France that consists of two distinct regions (*Languedoc-Roussillon* and *PACA*) emerges a single region (Fig. 15a). We conclude that the two administrative regions must show wide agreement in their product preferences. The region of *Rhone-Alpes* is almost perfectly realized at $\tau = 0.1$.

This is similar to the approach in [16, 17] where the respective authors cluster points of interest (POIs) based on whether

Fig. 16 Illustration of variation in number of communities for different product categories. We report the number of products under each category within parentheses. Note that *Cheese* which offers > 200 alternatives to choose from, consistently results in greater number of communities than *Hygienic Paper*, which offers fewer options

or not they are visited by the same set of people. As opposed to clustering on people, we cluster on products. The surfacing of contiguous communities despite any explicit contiguity constraints ascertains that consumers in vicinity of each other share product preferences and that the resultant regions transcend the established administrative boundaries.

5.2 Q2: Market competition

We could limit the set of characterizing products of a store to a certain category of products. For a store s , we can compute the set of characterizing products P_s that belong to the category *wine*, using the following equation.

$$P_s = \{p \mid s \in S_p, p \in P \text{ s.t. } \exists j \text{ s.t. } H_j^p = \text{wine}\}$$

Categories that contain large number of products offer more *option* to consumers who intend to make a purchase in that category. This may imply greater diversity and hence greater number of communities, N . To verify this, we conduct some experiments and present our results in Figs. 16 and 17.

In Fig. 16, we show the variation in the number of communities N with change in clustering threshold τ for two different product categories, *Cheese* and *Paper*. While the former contains over 300 individual products, the latter contains less than 50. We observe that *Cheese* consistently results in larger number of communities across all values of τ than *Paper*. We therefore conclude that there is greater diversity in preferences toward *Cheese* than *Paper*.

However, product categories with comparable number of products need not agree on N . If a category is dominated by a small number of *market leaders*, preferences would vary little across the stores. We verify the presence of this artifact and show an example in Fig. 17 where we compare the categories

Fig. 17 Illustration of variation in number of communities for different product categories. We report the number of products under each category within parentheses. While both categories have rough agreement on the number of products, *animal products* result in larger number of merged regions at all values of τ as compared to *mens' hygiene*

Mens' Hygiene and *Animal Food*. The former category is known to have market leaders. While both categories have comparable number of products (224, 297, respectively), the latter category results in greater number of merged regions at all values of τ . For categories with comparable number of products, the number of communities that emerge from hierarchical clustering, N , could be a measure to compare the market competition in the two categories.

6 Related work

Recent years have witnessed a sharp increase in the volume of user-generated, geo-tagged data available to researchers, primarily due to the emergence of location-based services such as Foursquare or Twitter and the ubiquitous use of smartphones. Thus, there is a large volume of work centered around proliferation of such data to various ends. Most previous studies can be broadly classified into following two categories.

The first category targets *region characterization* or discovering the features that are most characteristic of a given region. The region, supplied by the user, is typically a well-defined administrative entity with known boundaries. In [18], the authors study similarity in Foursquare check-ins of various users in a city to characterize neighborhoods. The underlying idea is that a neighborhood is characterized by the type of places contained within the neighborhood (restaurants, pubs, museums) as well as by the people that frequent these venues. The authors show that the discovered *livehoods* evolve with time. In [19], the authors study Flickr photographs to generate the tags (features) that best describe geographic entities (region) such as *USA* or *New York City* using a probabilistic model. Subsequently, they

use the obtained tags to identify region pairs that are similar but geographically distant.

The second category targets *region discovery* wherein the aim is to generate a colloquial region given a keyword. Previous studies have targeted several different data sources such as photographs, blogs and tweets. [1–5]. The state of the art is established in [5] where the authors first generate a frequency histogram and use an image segmentation- based technique to isolate the characterizing region. In [16, 17], the authors use check-ins to cluster points of interest (POIs) within a city. The core idea is that locations frequented by similar users should be similar. The bears resemblance with the use case described under Sect. 5.2.

The influence maximization problem [6–8, 20], due to its generic formulation, has been widely studied under various use cases. Consequently, the IC and LT influence propagation models have been successfully applied to address viral marketing [9, 10], recommendation systems [11], identification of influential peers in social graph [12, 21], etc.

Retail transactions, by nature, are user-generated and geo-tagged (since the location of stores is known). To the best of our knowledge, no previous studies have exploited this artifact to effectuate region discovery for retail products. In this paper, we make a case for the application of propagation models as a tool to uncover colloquial regions for products.

7 Conclusion

In this paper, we unearth colloquial regions for retail products. We introduce the notion of characterizing region of a product, which is a region marking active over-consumption of the said product. We study the existing methods and report that they fail to handle the large amount of noise present in retail data. We propose a new solution, DICE, that attempts to leverage influence propagation models as smoothing operators. Intuitively, DICE starts from a seed set of stores with high sales for a product and iterates over neighboring stores to find the largest contiguous regions covering all high-sales stores or stores that are contiguous to high density ones.

We conduct experiments on real data from a large super-market chain in France. Empirical evaluation and user studies establish that DICE discovers meaningful characterizing regions and significantly outperforms the natural baseline and the current state of the art. We study the impact of time on characterizing regions and ascertain that emergence of characterizing regions for a product can be time dependent owing to seasonal variation in product preferences. We also discover that some regions exhibit preferences toward product categories as opposed to individual products. We show how DICE can be used to reorganize administrative boundaries based on product preferences and compare diversity in preferences across product categories.

To the best of our knowledge, this is the first study that aims at discovering colloquial regions for retail products. We envisage two future research directions (a) improving DICE by experimenting with the seed selection process, diffusion model or threshold instantiation and (b) targeting the application of DICE in additional use cases under retail. We hope that the promising results would open avenues for future work.

Acknowledgements The authors acknowledge the generous support of Project Datalyse.

Funding was provided by Agence Nationale de la Recherche (FR) (ANR-13-CORD-0020).

References

1. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from Flickr tags. In: Proceedings of the ACM SIGIR conference on research and development in information retrieval, pp. 103–110 (2007)
2. Rattenbury, T., Naaman, M.: Methods for extracting place semantics from Flickr tags. *Trans. Web* **3**(1), 1:1–1:30 (2009)
3. Crandall, D.J., Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M.: Mapping the world's photos. In: Proceedings of the International conference on World Wide Web, pp. 761–770 (2009)
4. Intagorn, S., Lerman, K.: Learning boundaries of vague places from noisy annotations. In: Proceedings of the ACM SIGSPATIAL International conference on advances in geographic information systems, pp. 425–428 (2011)
5. Thomee, B., Rae, A.: Uncovering locally characterizing regions within geotagged data. In: Proceedings of the international conference on World Wide Web, pp. 1285–1296 (2013)
6. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ACM SIGKDD International conference on knowledge discovery and data mining, pp. 137–146 (2003)
7. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: Proceedings of the IEEE International conference on data mining, pp. 88–97 (2010)
8. Goyal, A., Lu, W., Lakshmanan, L.V.S.: SIMPATH: an efficient algorithm for influence maximization under the linear threshold model. In: Proceedings of the IEEE International conference on data mining, pp. 211–220 (2011)
9. Aslay, Çigdem, Lu, W., Bonchi, F., Goyal, A., Lakshmanan, L.V.S.: Viral marketing meets social advertising: ad allocation with minimum regret. *PVLDB* **8**(7), 822–833 (2015)
10. Barbieri, N., Bonchi, F.: Influence maximization with viral product design. In: Proceedings of the SIAM International conference on data mining, pp. 55–63 (2014)
11. Ye, M., Liu, X., Lee, W.-C.: Exploring social influence for recommendation: a generative model approach. In: Proceedings of the ACM SIGIR conference on research and development in information retrieval, pp. 671–680 (2012)
12. Zhou, C., Zhang, P., Guo, J., Zhu, X., Guo, L.: UBLF: an upper bound based approach to discover influential nodes in social networks. In: Proceedings of the IEEE International conference on data mining, pp. 907–916 (2013)
13. Ye, T., Kalyanaraman, S.: A recursive random search algorithm for large-scale network parameter configuration. In: Proceedings of the ACM SIGMETRICS/international conference on measurement and modeling of computer systems, pp. 196–205 (2013)

14. Montello, D.R., Goodchild, M.F., Gottsegen, J., Fohl, P.: Where's downtown? Behavioral methods for determining referents of vague spatial queries. *Spat. Cogn. Comput.* **3**(2–3), 185–204 (2003)
15. Lynch, K.: *The Image of the City*. MIT Press, Cambridge (1960)
16. Barbieri, N., Bonchi, F., Manco, G.: Influence-based network-oblivious community detection. In: *Proceedings of the IEEE International conference on data mining*, pp. 955–960 (2013)
17. Srivastava, S., Pande, S., Ranu, S.: Geo-social clustering of places from check-in data. In: *Proceedings of the IEEE International conference on data mining*, pp. 985–990 (2015)
18. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.M.: The livelihoods project: utilizing social media to understand the dynamics of a city. In: *Proceedings of the International conference on web and social media*, pp. 58–65 (2012)
19. Kafsi, M., Cramer, H., Thomee, B., Shamma, D.A.: Describing and understanding neighborhood characteristics through online social media. In: *Proceedings of the International conference on World Wide Web*, pp. 549–559 (2015)
20. Kempe, D., Kleinberg, J.M., Tardos, É.: Influential nodes in a diffusion model for social networks. In: *Proceedings of the International colloquium on automata, languages, and programming*, pp. 1127–1138 (2005)
21. Guo, J., Zhang, P., Zhou, C., Cao, Y., Guo, L.: Personalized influence maximization on social networks. In: *Proceedings of the ACM conference on information and knowledge management*, pp. 199–208 (2013)