HAL
open science

Discovering Characterizing Regions For Consumer Products

Shashwat Mishra, Vincent Leroy, Sihem Amer-Yahia

To cite this version:

Shashwat Mishra, Vincent Leroy, Sihem Amer-Yahia. Discovering Characterizing Regions For Consumer Products. 2015. hal-01190980

HAL Id: hal-01190980

https://hal.science/hal-01190980v1

Preprint submitted on 2 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Discovering Characterizing Regions For Consumer Products

Shashwat Mishra
Univ. Grenoble Alpes
CNRS, LIG, Grenoble, France
Email: shashwat.mishra@imag.fr

Vincent Leroy
Univ. Grenoble Alpes
CNRS, LIG, Grenoble, France
Email: vincent.leroy@imag.fr

Sihem Amer-Yahia
Univ. Grenoble Alpes
CNRS, LIG, Grenoble, France
Email: sihem.amer-yahia@imag.fr

***Abstract*—Consumer behaviour holds special importance in the retail industry. Consumer location impacts consumer behaviour by dictating purchase trends. This paper investigates the problem of examining product sales across a chain of stores to extract the geographic regions that *characterize* a product. Characterizing region for a product is a coherent geographic region where the consumers actively consume the said product. We introduce DICE, a diffusion-based technique to uncover all such regions for a given product, when they exist. In contrast to current state of the art, DICE involves minimal usage of parameters and shows remarkable tolerance to noise. We present experiments conducted on real datasets from a general commercial supermarket in France. Empirical evaluation and user-studies establish that the presented method significantly outperforms its natural baseline and previous state of the art approaches.**

I. Introduction

Customer receipt analysis is a common task in the retail industry and has proven to be very effective for recommending products, implementing loyalty programmes 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 across multiple stores that contrasts sales of a given product in different stores to identify regional trends. In order to do that, we introduce the notion of *characterizing region* for a product, and address the challenge of discovering them. To the best of our knowledge, the presented body of work is the first to address such a problem.

Understanding consumer behaviour is of vital importance for marketers and consumer-goods companies. Much work has thus been devoted to analyze 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.

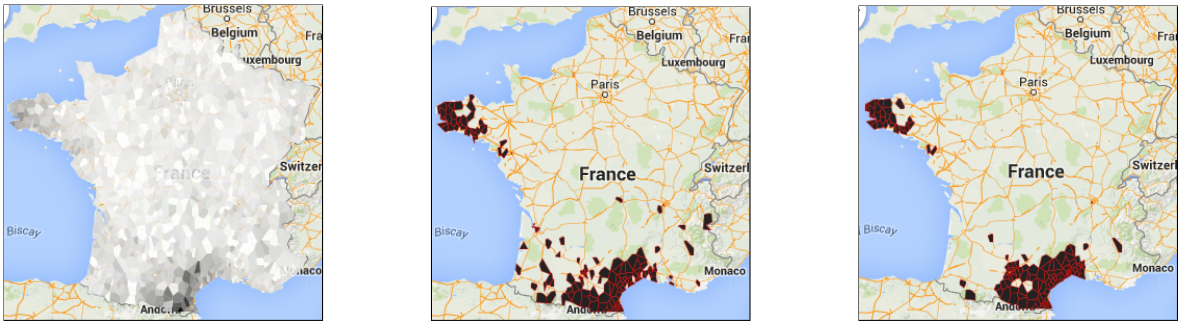
In this paper, we complement these approaches by introducing a *product-centric* analysis that combines in-store sales with geographic proximity to identify regional trends for a product. Such analysis is critical to product marketing. Popular shopping trends do indeed affect a customer's decision of 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. The ability to combine trends with geo proximity can be very useful for cross-store management and provisioning.

Consider Figure 1a which shows the heatmap computed using the sales of a certain product[1] in various stores across France. At each store, the total units sold for the product is normalized by the total units sold for all products. 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. One application with great potential is the decision to expand or not a product's availability in some regions. 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.

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 Figure 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 (Figure 1a) they offer a fragmented view of the product's regional trends. Product sales across

[1] *The product in example is a type of canned tuna sauce that is hidden for confidentiality purposes.*



(a) Sales heatmap (b) Top-K (c) DICE

Fig. 1. Inadequacy of TopK for uncovering characterizing regions. Figure 1a shows the heatmap for a product where darker color implies higher sales. Figures 1b and 1c show the characterizing regions extracted using the natural baseline (TopK) and our method (DICE). The proposed method returns smoother regions while preserving "holes".

stores show high variance (noise) leading to several stray maximas 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] [2] [3] [4]. The current state of the art is established in [5] where the authors study tagged photographs to extract regions for geo-entities. The key idea is to represent the tag's frequency distribution across the map as an image and identify contours that demarcate sharp changes in the frequency. These contours are further subjected to a series of complex image processing operations to obtain the characterizing region. Region discovery for retail products is not amenable to such a solution as it assumes the existence of a sharp boundary (in terms of sales) between the characterizing region and its surroundings. As we see later in Section IV, this assumption seldom holds true for retail products.

This analysis suggests that both, the natural baseline and the current state of the art for related problems, are insufficient for uncovering 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.

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 demonstrated in Figures 1b and 1c, the obtained results are significantly better than the natural baseline. In Section IV, we present our experiments that conclusively establish the significant gains achieved over the natural baseline and the current state of the art.

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. Further, the proposed technique offers high quality results when compared to the natural baseline and the previous state of the art. We validate our results via both, empirical comparisons that report $> 25\%$ gain over previous methods and user studies that establish that majority of the users prefer DICE over competing methods and show wide agreement ($> 70\%$) with characterizing regions returned by DICE.

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 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.

Our paper is organized as follows. In Section II, we formalize our data model and state the problem of identifying the set of characterizing regions of a product. Section III describes the design process for DICE. Section IV is dedicated to our experiments and findings. The related work is summarized in Section V. We conclude in Section VI.

II. Data Model and Problem

A. Model

We have a set of *products*, $\mathcal{P} = \{p_1, p_2...\}$, which represent distinct items available in a typical departmental 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, $\mathcal{S} = \{s_1, s_2...\}$, 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 Figure 2.

We say that two stores are *connected* if corresponding 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 (non-zero) units of certain products at a store. We assume the existence of a set of transactions, $\mathcal{T} = \{t_1, t_2...\}$, such that each transaction $t \in \mathcal{T}$ 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.

Using $\mathcal{T}_s$ to denote the set of transactions occurring at store s, the total units of product p sold at s is computed as:

$$sales(p, s) = \sum_{t \in \mathcal{T}_s} f(p, t) \tag{2}$$

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:

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

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

$$D_p = \{d_p(s_1), d_p(s_2)...d_p(s_{|\mathcal{S}|})\} \tag{4}$$

B. Problem

Our aim is to identify, given a product p, a set of stores, $\mathcal{S}_p \subseteq \mathcal{S}$, such that the stores in $\mathcal{S}_p$ have high product density and result in *coherent* regions. Note that a product may have more than one coherent region, as shown in Figure 1c. We refer to each such coherent region as a *characterizing region* and $\mathcal{S}_p$ 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, $\mathcal{S}_p$ 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 $\mathcal{S}_p$ is thus undefined. That is, for example, the case for generic products that are consumed uniformly across all stores, such as *chewing gum*, *toilet paper* etc.

The technique we describe in this paper assumes the existence of characterizing regions for the product. Subsequently, 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 Section IV.

III. Methodology

Our objective is to analyze the density distribution in order to identify characterizing regions for the product. We focus on geographically coherent regions, defined in Section II as a large 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 some "holes" within a region represent valuable marketing information.

As stated previously, the natural approach of selecting top k stores works well for products with sharply defined characterizing regions but it is not a viable solution for 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 behaviour.

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.

Diffusion over graphs has been well studied in different version 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, given k, 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), which characterize different types of social interactions [7] [8] [9] [10] [11] [12]. While the IC model assumes independent interaction for each pair of nodes, the LT model uses threshold-based behaviour.

To design DICE, we select the linear threshold model of influence propagation as it allows us to incorporate influence from multiple influencing agents 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 (Figure 3b). These stores are marked as active and the remaining stores as inactive. The

(a) Store locations (b) Voronoi map

Fig. 2. Distribution of stores across France. Fig 2a shows the exact location of all stores while Figure 2b shows the 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.

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 (Figure 3c). DICE then extracts coherent regions as significant connected components covering the activated stores (Figure 3d). We now go over the salient features of IP models and discuss them in the context of our problem in the following subsections.

A. Seeds selection

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

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

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 will not have any negative 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 Section IV reveal that stable results are returned by DICE for large range of values for m.

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

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

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.

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

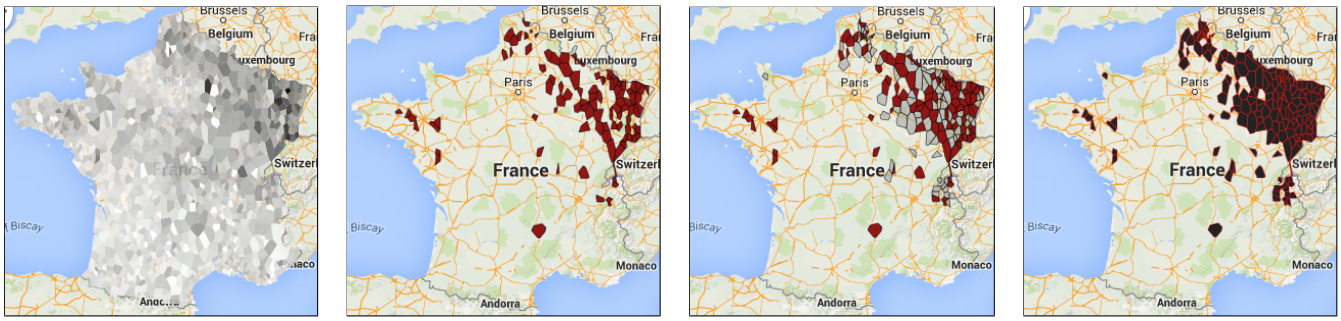
We ran experiments to evaluate different choices for weighting the stores (including $w_s = 1$). Logarithmic formulation, as in Eq. 7, delivered best results. Intuitively, Eq. 7 assigns higher weights to stores with greater sales thus allowing them to spread more influence. The store with greatest number 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 $\mathcal{S}_p$, and is defined as follows:

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

At any point of the recursive influence spread, the total influence received by a store s is evaluated as $\sum_{x \in \mathcal{S}_p} b_{s,x}$, 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. As Figure 2a shows, stores are far from being evenly geographically distributed: populated regions, around major



(a) Heatmap (b) Seeds (c) Diffusion (d) Characterizing regions

Fig. 3. Illustration of different stages in DICE. Normalized density distribution (represented as heatmap in Figure 3a) forms the starting point. Selection of the desired size for the seed-set (m) results in selection of top-m stores as seeds (Figure 3b). The activated stores influence other stores potentially activating them (Figure 3c). The region constituted by the set of activated stores is reported as the characterizing region (Figure 3d).

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 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 ($w_{s_i}.d_p(s_i)$), 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 to $w_{s_i}.d_p(s_i)$.

We remark that DICE does not completely eradicate the fragmentation issue incurred by the natural baseline (TopK). $\mathcal{S}_p$ 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 we see in Section IV, DICE significantly mitigates the fragmentation incurred by the natural baseline (TopK).

C. Detailed algorithm

The details of DICE are presented in Algorithm 1. $\mathcal{S}_p$ is first initialized to be the set of seeds following the approach described in Section III-A, 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 last iteration A_{prev} (line 11) following the update rules defined in Section III-B. 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 $\mathcal{S}_p$ are then updated at the end of each influence propagation iteration. After convergence, DICE returns the set of active stores $\mathcal{S}_p$ (line 21).

Algorithm 1 Diffusion based Iterative Characterizing region Exploration (DICE)

Input: p: product
Input: m: seed-set parameter
Output: $\mathcal{S}_p$: Set of stores that form the characterizing regions

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

IV. Results

In this section we report on experiments conducted to study the behaviour of DICE and evaluate its performance in comparison to 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 dataset, (Section IV-A) and the test-bench (Section IV-B). This is followed by a succinct description of the competing methods, natural baseline (TopK) and the previous state of the art (SSR) in Section IV-C. Subsequently we discuss the various experiments in Section IV-D.

TABLE I. STATISTICS OF RETAIL DATASET.

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

A. 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 1st, 2012 to July 31st, 2014. The data consists of a large number of *csv* files in different domains such as *stores*, *transactions*, *products*, *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 data model described in Section II fits perfectly to the dataset.

We subjected the raw data to some cleaning operations which we describe next.

1) 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 ($|\mathcal{T}_s| \geq 20k$). This results in more than 1400 stores. In Figure 2, we show the location of obtained stores on the map of France and the associated Voronoi map.

2) 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 Figure 4, we 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 I reports the final cardinalities of the sets. We examine over 1 billion transactions across more than 1400 stores for more than $10k$ products.

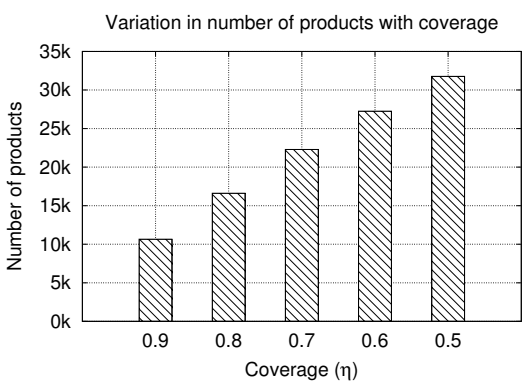


Fig. 4. Distribution of products w.r.t. fraction of stores covered (η). The X-axis shows the different coverage values. The Y-axis shows the number of products with coverage $\geq$ specified value.

B. Platform

All experiments were run on a 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.

C. Competitors

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

1) TopK: TopK is the trivial algorithm for discovering characterizing regions given product's density distribution. It returns the stores with the k highest densities as the result $\mathcal{S}_p$. The choice of k is non-trivial. By design, the baseline lays no 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 suitable k. However, for most products the assumption does not hold good and resulting characterizing regions are fragmented and erroneous.

Note that TopK shares certain aspects with DICE as the latter uses the former for seed-selection which bootstraps the diffusion process.

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", "poland") from a large volume of tagged photographs.

The method starts by gridding the map and computing frequency of the given tag in each cell of the grid. This gives a representation of the tag's frequency-distribution over the map in the form of 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 grid cells 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 Section II. A photograph corresponds to a transaction and the tags within the photograph correspond to products. The two models differ slightly in that photographs can manifest at any arbitrary point on the map, our 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 4 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.

D. Experiments

We conduct three different types of experiments. The first experiment studies the impact of parameter m, the *seed-size* for DICE. The remaining two experiments compare the performances of DICE, TopK, and SSR using an empirical approach and a user-study respectively.

1) Impact of seed-size parameter: DICE admits a single parameter, the initial seed-set size, m. $|\mathcal{S}_p|$ is guaranteed to increase with increase in m because a higher value of m implies a bigger seed-set (which by design is part of the result). Furthermore, a bigger seed-set implies more influence on other stores. We study the impact of seed-set size on the characterizing regions for different products.

We observe that for the same choice of m, certain products diffuse more than others. We illustrate this with the help of an example. Figure 5 shows the characterizing regions, obtained via DICE, for two different products for two different 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 II we show the size of the final characterizing regions which establishes that the former product grows significantly more than the latter.

This behaviour 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.

2) Empirical Comparison: Evaluation of the quality of returned characterizing regions is difficult in the absence of ground truth. Note that relevant previous studies [4] [5] have exclusively dealt with extraction of colloquial boundaries for geographic concepts (such as "germany"). It can be easily assumed that the region extracted for tag "germany" 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 will be restricted to certain locales.

<u>Ground Truth</u>*:* We attempt to fill the void created by the absence of ground-truth with the aid of domain experts. 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 $p \in \mathcal{P}$, 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 absence of data, the administrative region of *Corsica* was omitted. For each of the remaining 21 administrative regions of France, we obtained 2 products that were expected to be characterizing in the said region. This resulted in 42 ground truth pairs of the form $\langle p, r\rangle$.

TABLE II. IMPACT OF SEED-SIZE ON SIZE OF CHARACTERIZING REGIONS ($|\mathcal{S}_p|$).

Product	$\lvert\mathcal{S}_p\rvert$ ($m = 50$)	$\lvert\mathcal{S}_p\rvert$ ($m = 100$)
peculiar meat	71	175
chewing gum	56	121

<u>Evaluation Scheme</u>*:* 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. This measure, *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.

<u>Parameterization</u>*:* 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 m for which the resulting regions are nearly the same. This is expected because if the supplied value of m is less than, but close to, the optimal value, the iterative diffusion-based design of DICE would ensure the discovery of all relevant stores 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.

TopK suffers from a serious limitation when compared to SSR and DICE. The characterizing regions for all products contain the same number of stores, k, as fixing k fixes the number of stores in the returned regions. Single value of k may not be suitable for several products simultaneously. Thus, choosing k becomes a non-trivial task. 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 diffusion phase are different and product specific.

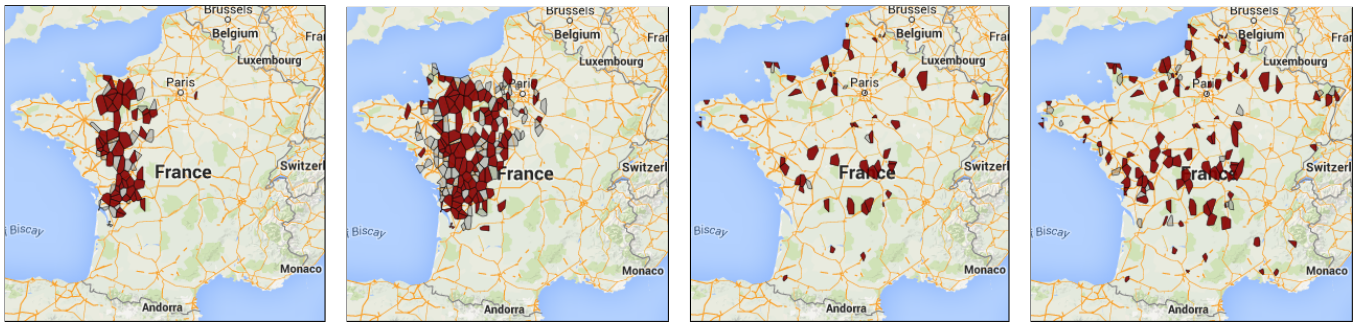
We give TopK the flexibility to return characterizing regions of different sizes in the following manner. For each product, TopK uses the number of stores in the characterizing regions returned by DICE as the value of k i.e. $k = |S_p^{\text{DICE}}|$. Thus, TopK is executed with different values of k 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 III, 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]. We explore the range $[0, 1000]$ for σ and $[0, 200]$ for ρ. In Figure 6, we show the variation in Jaccard overlap achieved by SSR (averaged over all ground-truth pairs) w.r.t its parameters, in 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 III. We abstain from listing the default value for TopK since instantiating DICE instantiates TopK.

TABLE III. 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}



(a) peculiar meat, $m = 50$ (b) peculiar meat, $m = 100$ (c) chewing gum, $m = 50$ (d) chewing gum, $m = 100$

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 grey. The seed-set stores are more localized for product *peculiar meat* than for *chewing gum*. Consequently the former encounters more growth (larger number of grey stores) at both values of m during the diffusion step.

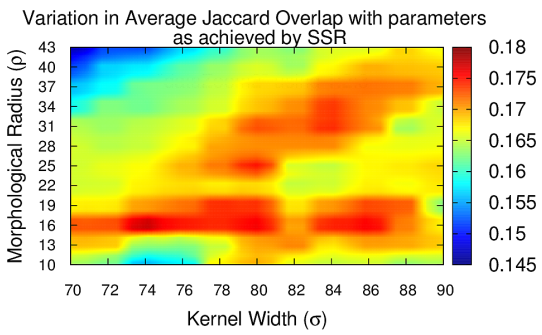


Fig. 6. Parameter space exploration for SSR [5]. Optimal Average Jaccard Overlap is obtained at *Kernel Width*=74 and *Morphological Radius*=16.

Result: Finally, we present the results of the empirical comparison in Figure 7. DICE achieves significantly higher values for Average Jaccard overlap (*Average*) than TopK and SSR. 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 anonymized products, *detergent* and *canned tomatoes*, in Figures 8 and 9.

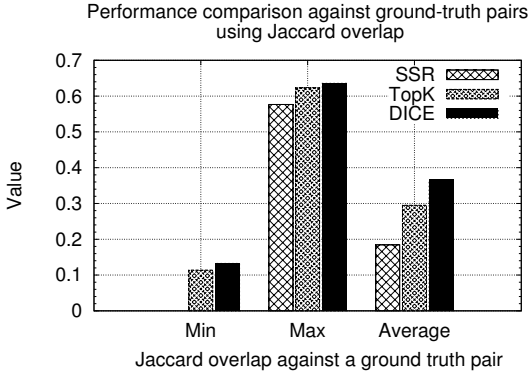


Fig. 7. Performance evaluation of TopK, SSR and DICE against the 42 ground truth pairs. Seed-set size for DICE (m) was fixed at 100. For TopK, $k = |\mathcal{S}_p^{\text{DICE}}|$. Parameters for SSR were fixed at $\sigma = 74$ and $\rho = 16$.

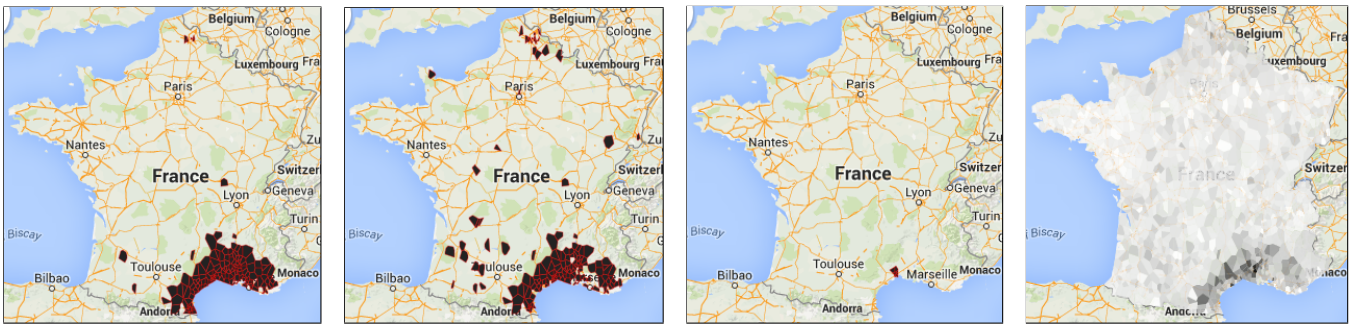
In Figure 8, results from TopK, while consistent with the corresponding heatmap, Figure 8d, are expectedly more fragmented. Whereas SSR (Figure 8c) picks up high frequency variation within the characterizing region and results in a characterizing region that is too small.

As compared to the product in Figure 8, the product in Figure 9 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 (Figure 8d and Figure 9d). This increased noise has adverse effects on both TopK and SSR. For TopK, Figure 9b, the severity of incurred fragmentation increases substantially. SSR, Figure 9c, results in too large a region, covering almost all of France. This occurs as the high frequency variations give rise to several contours (demarcating sharp fall in product density) in the binary map of SSR which unify into a single connected component post morphological closing.

Note that DICE is also adversely affected by increased noise as established by Figures 9a and 8a. 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*.

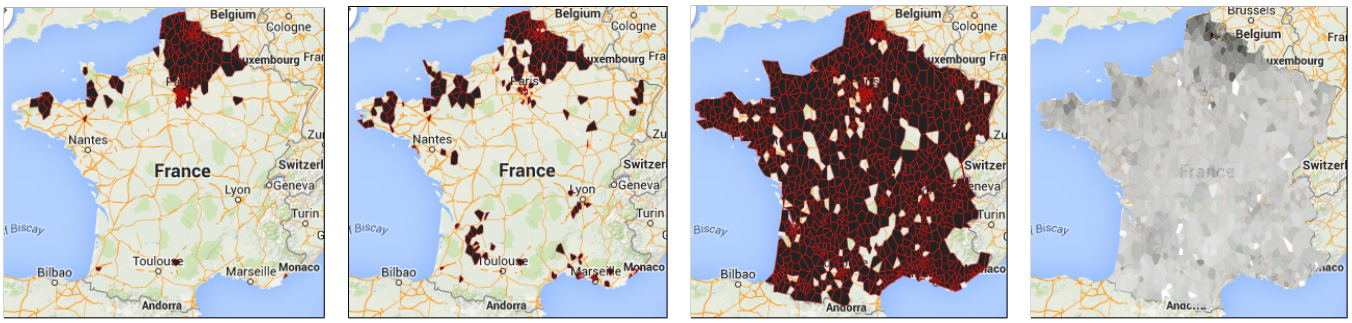
We also remark that SSR was observed to consistently result in regions that are either too small or too large. Figures 8c and 9c present the two different extremes achieved by SSR.

To substantiate our argument about fragmentation, we quantified the incurred fragmentation (as the average pairwise distance between stores in $\mathcal{S}_p$) and measured it for different choices of seed-set size (m). We show the obtained result in Figure 10. 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.



(a) DICE (b) Top-K (c) SSR (d) Heatmap

Fig. 8. Illustration of characterizing regions obtained for the ground-truth pair *(detergent, Languedoc-Roussillon)*. While TopK (Figure 8b) significantly overlaps with *Languedoc-Roussillon*, it includes several distant stores. SSR (Figure 8c) picks up the sharp density variation within the characterizing region and results in a region that is too small. DICE delivers the best result (Figure 8a). Jaccard overlap with the boundaries of Languedoc-Roussillon region of France are 0.41 (DICE), 0.31 (TopK) and 0.02 (SSR).



(a) DICE (b) Top-K (c) SSR (d) Heatmap

Fig. 9. Illustration of characterizing regions obtained for the ground truth pair *(canned tomatoes, Picardie)*. TopK (Figure 9b) results in high fragmentation and SSR (Figure 9c) results in a very large region. Jaccard overlap achieved by the 3 approaches with the French administrative region of *Picardie* are 0.36 (DICE), 0.17 (TopK) and 0.04 (SSR).

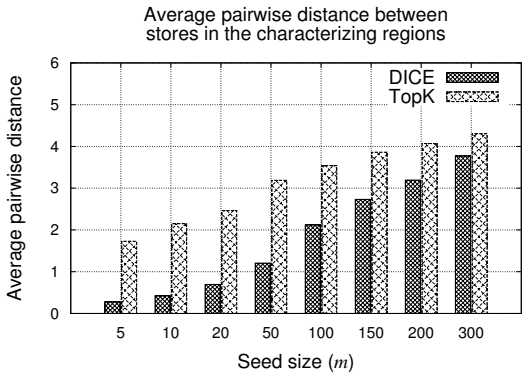


Fig. 10. Mitigation of fragmentation by DICE. Y-axis shows the average pairwise distance between stores in $\mathcal{S}_p$. X-axis shows the seed-size parameter (m) for DICE. $\mathcal{S}_p^{TopK}$ is more fragmented than $\mathcal{S}_p^{\text{DICE}}$.

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 8 experts from the retail domain who 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 IV 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 *aesthetic bias* that influences them to vote for the method that returns aesthetically 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 no suffer from this bias.

Independent Evaluation: The aim of the second user study is to perform an independent evaluation of DICE. For each

TABLE IV. USER-STUDY: COMPARATIVE EVALUATION

	DICE	TopK	SSR
Users	46.55	21.13	32.32
Experts	47.46	34.19	18.35

TABLE V. 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

product in a randomly generated list of 10, 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 expressed on a scale of 1 to 5. Results were averaged over all users and all products.

We present the final figures in Table V. 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 *aesthetics bias* as *Experts* are more likely to show agreement with non-trivial shaped regions.

V. RELATED WORK

Uncovering characterizing regions for geotagged data is a relatively new problem. Most works have dealt with examining different types of user generated content, such as *photographs, blogs, tweets etc.* to unearth colloquial boundaries [1] [2] [3] [4] [5].

The influence maximization problem [6] [7] [8] [14], 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] [15] etc. However, no previous work has investigated the applicability of such propagation models towards discovering prominent regions for retail products.

To the best of our knowledge, the presented body of work is the first to address the problem of uncovering characterizing regions for consumer products. The application of influence propagation models to uncover such regions has not been investigated in past works.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present DICE, an algorithm that extracts characterizing regions for a given consumer product using retail transaction logs. We investigate the application of decay-based, weighted, influence propagation models to achieve enhanced smoothing. 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. As a consequence of this study we have developed a tool that can be used to extract and analyze colloquial characterizing regions for consumer products.

Our empirical evaluation and user studies on customer receipts of a large food supply chain in France show that DICE discovers meaningful, coherent, product-centric characterizing regions and outperforms its competitors. This opens several avenues for future work. In particular, we seek to explore the following two directions. We intend to combine pattern mining techniques with characterizing regions to extend our focus from single product to *set of products*. We also seek to correlate found regions with health indicators for different regions to gather information on customers' wellbeing. Both directions open challenging computational questions that necessitate the development of scalable algorithms to enable the production of such analytics efficiently.

REFERENCES

[1] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *SIGIR 2007*, pp. 103–110.

[2] T. Rattenbury and M. Naaman, "Methods for extracting place semantics from flickr tags," *Transactions on the Web*, vol. 3, no. 1, pp. 1:1–1:30, 2009.

[3] D. J. Crandall, L. Backstrom, D. P. Huttenlocher, and J. M. Kleinberg, "Mapping the world's photos," in *WWW 2009*, pp. 761–770.

[4] S. Intagorn and K. Lerman, "Learning boundaries of vague places from noisy annotations," in *SIGSPATIAL 2011*, pp. 425–428.

[5] B. Thomee and A. Rae, "Uncovering locally characterizing regions within geotagged data," in *WWW 2013*, pp. 1285–1296.

[6] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD 2003*, pp. 137–146.

[7] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *ICDM 2010*, pp. 88–97.

[8] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: an efficient algorithm for influence maximization under the linear threshold model," in *ICDM 2011*, pp. 211–220.

[9] Ç. Aslay, W. Lu, F. Bonchi, A. Goyal, and L. V. S. Lakshmanan, "Viral marketing meets social advertising: Ad allocation with minimum regret," *PVLDB*, vol. 8, no. 7, pp. 822–833, 2015.

[10] N. Barbieri and F. Bonchi, "Influence maximization with viral product design," in *SIAM 2014*, pp. 55–63.

[11] M. Ye, X. Liu, and W. Lee, "Exploring social influence for recommendation: a generative model approach," in *SIGIR 2012*, pp. 671–680.

[12] C. Zhou, P. Zhang, J. Guo, X. Zhu, and L. Guo, "UBLF: an upper bound based approach to discover influential nodes in social networks," in *ICDM 2013*, pp. 907–916.

[13] T. Ye and S. Kalyanaraman, "A recursive random search algorithm for large-scale network parameter configuration," in *SIGMETRICS 2003*, pp. 196–205.

[14] D. Kempe, J. M. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *ICALP 2005*, pp. 1127–1138.

[15] J. Guo, P. Zhang, C. Zhou, Y. Cao, and L. Guo, "Personalized influence maximization on social networks," in *CIKM 2013*, pp. 199–208.