GESTÃO & PRODUÇÃO

ISSN 0104-530X (Print)
ISSN 1806-9649 (Online)

ORIGINAL ARTICLE

Mapping regional business opportunities using geomarketing and machine learning

Mapeamento de oportunidades de negócio regionais utilizando geomarketing e aprendizado de máquina

Marcelo Fernando Felix de Oliveira[1], Pedro Henrique Melo Albuquerque[1], Peng Yao Hao[1], Pedro Alexandre Henrique[1]

[1]Universidade de Brasília – UnB, Laboratório de Aprendizado de Máquina em Finanças e Organizações, Campus Darcy, Brasília, DF, Brasil. E-mail: felixmarcelo.mf@gmail.com; pedroa@unb.br; peng.yaohao@gmail.com; pedroalexandre.df@gmail.com

How to cite: Oliveira, M. F. F., Albuquerque, P. H. M., Hao, P. Y., & Henrique, P. A. (2020). Mapping regional business opportunities using geomarketing and machine learning. *Gestão & Produção*, 27(3), e4158. https://doi.org/10.1590/0104-530X4158-20

Abstract: The objective of this study is to develop a quantitative tool, based on Machine Learning and Geomarketing to identify business opportunities and contribute to the strategic process of local choice of franchises' network selecting regions that have a high demand forecast and a lack of product supply. In addition, we conducted a qualitative analysis of the selected business places based on defined criteria. This prediction is given by constructing a consumption pattern, defined by a classifier, based on the characteristics of the reserved rights. Initially, for a better understanding on this subject, a theoretical background was made covering the main concepts about Geomarketing and Machine Learning and its applications. After that for a demonstration of the results, we opted for the application of the method for the market of fine chocolates (Cacau-Show) in the Distrito Federal. The main databases used in this paper were *Pesquisa de Orçamentos Familiares* and from *Instituto Brasileiro de Estatística e Geografia* (IBGE). As a result, the Standardized Spend was obtained, which indicates the requirement for each Censitar Sector, as georeferenced information of the competition, containing 44 stores that have as their main product of fine chocolate, and as digital meshes of the Federal District. The crossing is available for the elaboration of a map that facilitates the identification of the business opportunities for the market of fine chocolates in the Distrito Federal, Brazil.

Keywords: Geomarketing; Support vector machine; Regional economics; Estrategic franchises; Supervised learning.

Resumo: O objetivo deste estudo é a elaboração de uma ferramenta quantitativa, baseada em técnicas de Geomarketing e Aprendizado de Máquina, capaz de identificar oportunidades de negócio e contribuir para o processo estratégico de escolha locacional de uma rede de franquias, selecionando regiões que possuam uma alta previsão de demanda e uma carência na oferta do produto. Além disso, realizou-se uma análise qualitativa dos pontos comerciais selecionados com base em critérios definidos no decorrer do trabalho. Essa previsão se dá pela construção de um padrão de consumo, definido por um classificador, baseado nas características dos indivíduos que costumam comprar o produto. Inicialmente, para um melhor entendimento a respeito do assunto, foi feito um embasamento teórico abarcando os principais conceitos sobre Geomarketing e Aprendizado de Máquina e suas aplicações. Em seguida, para a demonstração dos

Received Sept. 18, 2017 - Accepted June 29, 2018
Financial support: None.

resultados, optou-se pela aplicação do método para o mercado de chocolates finos (Cacau-Show) no Distrito Federal. As principais bases de dados utilizadas neste trabalho foram provenientes da Pesquisa de Orçamentos Familiares e do Censo Demográfico, ambos desenvolvidos pelo Instituto Brasileiro de Estatística e Geografia (IBGE). Como resultado, obteve-se o Gasto Padronizado, que indica o nível de demanda para cada Setor Censitário, as informações georreferenciadas da concorrência, contendo 44 lojas que possuem como principal produto o chocolate fino, e as malhas digitais do Distrito Federal. O cruzamento dessas informações permitiu a elaboração de um mapa que facilita identificação das oportunidades de negócio para o mercado de chocolates finos no Distrito Federal.

Palavras-chave: Geomarketing; Máquinas de suporte vetorial; Economia regional; Franquias estratégicas; Aprendizado supervisionado.

1 Introduction

Location is a crucial item for the success or failure of an enterprise. It is known that a representative portion of the entrepreneurs who closed their companies did not know the number of customers they would have and their consumption habits; the number of competitors present in the region; and the best location for the installation of their business. These data express the importance of choosing the point for the future of an organization.

According to Cliquet (2006), territorial coverage at both regional and international levels is at least as important as sales volume to determine the strength of a store network. In this context, and considering the need for a network of franchises to have a judicious and strategic decision process in determining a commercial point, this paper has the following research problem:

“Demonstrate the potential of Geomarketing, combined with Machine Learning techniques, to improve locational choice strategies in a network of franchises”.

Thus, the objective of this article is the elaboration of a quantitative method that assists the locational decision process of a network of franchises from the use of Geomarketing and Machine Learning techniques. During the course of the study, a basket of products was prepared according to what is offered by the franchise network chosen and that are also present in the POF 2008-2009. In addition, an explanatory model was drawn up based on the variables contained in the POF 2008-2009 that were relevant to explain the consumption of the product offered by the franchise. This was done to construct a machine learning model capable of predicting consumer behavior according to the information collected by the POF 2008-2009. After that, the decision function estimated by the classifier was applied in the database of the Demographic Census of the Brazilian Institute of Geography and Statistics (IBGE, 2010) of 2010, to obtain forecasts about the consumption of the population of the Federal District. Finally, a mapping construction was carried out, with the help of Qgis software and Geomarketing techniques, of the competition and the distribution of the demand for the chosen product along the territory of the Federal District; as well as to indicate, through eliminatory and classificatory criteria, the places that represent business.

According to Dobbs and Hamilton (2007 apud Machado, 2016) and Wright & Stigliani (2012 apud Machado, 2016), the external environment and the conditions of supply and demand may or may not favor the growth of a company. A competitive market is characterized by uncertainty and its complexity requires the use of a large number of information for decision making (Janssen, 2009 apud Machado, 2016).

It is essential for any company to know its market in detail. The combination of a database that provides behavioral information to potential clients - such as their consumption habits, income, schooling, among others - and their spatial information results enable the manager to know how they behave, what their characteristics are and where these individuals are located. This union of information allows a more comprehensive analysis and is made possible through the techniques of Geomarketing.

This knowledge about consumers, plus the Geomarketing tool, information about where and how these customers are graphically placed in the marketplace, is of great value for implementing a marketing approach. According to Cliquet (2006), there is a growing need for a more precise understanding of the market, which manifests itself through increasing and specific segmentations. This segmentation is due to the gradual fragmentation of the population and the need to define a differentiated strategy for each segment.

It is therefore a very useful tool for managers, assisting in Decision Making and facilitating the identification of opportunities and threats to your. In addition, it is possible to affirm that Geomarketing techniques contribute to a better allocation of the available resources (Baviera-Puig et al., 2009; Mangini et al., 2017); to reduce the risks involved in the process of opening a new commercial spot, preventing issues such as the lack of demand, poor public acceptance or excessive competition; and to optimize organizational results, since it is able to identify the ideal environment for business development and, in the case of franchises, target regions with a coverage gap, which ends up representing a cost for a network of stores.

The application of machine learning in the geomarketing literature is scarce. Although it is not yet a widespread subject in Brazil, it is an internationally consolidated methodology that presents a wide variety of disciplines that make use of this tool as an object of study, such as Agriculture, Health and Engineering. This technology is present in banks, social networks, applications, and even in the way computers and smartphones work to offer a personalized service to the owner. All this is possible because the machine can accumulate experiences and learn from them, and is thus able to make predictions. The potential about the junction between these two technologies is great, considering that the model can, through a georeferenced database, learn how consumers behave and return, also in this format, a forecast of how the individuals of a given region will behave in front of the opening of a new business in their residential area.

In this sense, the main theoretical motivation for the resolution of this work is the lack of exploration of this field of knowledge and the contribution that this study can provide for the academic development in this area.

2 Theoretical reference

The advancement of techniques and tools in the area of cartography has been increasingly useful and effective in the most diverse areas of knowledge. The amount of data that can be inserted in the maps is increasing, which has coused the loss of space by the analogue system (manual,) since the tendency is the insertion of multiple information in the same map. The increase in the demand for an increasing representation of data within a map, parallel to the advance of computers, led to the emergence of a powerful tool, Geographic Information Systems (GIS).

The optimal location of an establishment, according to Daskin (1995) apud Briozo & Musetti (2015) is the one that maximizes the service provided. In addition, the strategic analysis of the possible points to invest in a business is essential for the survival of the business, since, in most cases, it directly impacts its productive capacity (Kloose & Drexl, 2005 apud Briozo & Musetti, 2015).

2.1 GIS applications

According to Burrough and McDonnell (1998) apud Carnasciali & Delazari (2011, p. 107), Geographic Information Systems (GIS) are "[…] a set of tools to collect, store, retrieve, transform and visually represent spatial data". They have the ability to integrate data from diverse sources with a georeferenced database, in order to perform diverse and complex analyzes from the result of such integration. The ability to integrate and combine information makes the Geographic Information System capable of generating new representations that aid in the decision making process. Having initially been developed with a computational tool capable of storing, manipulating, and allowing analyzes and presentations of geographic data, GIS, with technological advances and the reduction of its costs, have become a strongly used tool for solving problems that require precise spatial analysis, as is the case in the context of market analysis (Martin et al., 2005 apud Fan & Collischonn, 2014)

With the development of Geographic Information Systems and the rise of machines to the stage of modern technological advancement, Geoprocessing - "[…] a set of knowledge destined to the treatment of information concerning objects, occurrences or phenomena that are associated with relative positions of the Earth's surface" (Furlan, 2011, p. 98) - has become a dynamic element in the process of knowledge and representation of the earth's surface.

Considering the amount of information that can be generated - such as identifying market segments, in-depth knowledge of each consumer group, their demands and specificities, as well as their location in space - the importance of geomarketing tends to consolidate more and more, since it is vital for companies, especially in times of crisis, as is the current Brazilian scenario, the use of techniques and tools that make it possible to attract customers and improve sales strategies and advertising aimed at maximizing results.

2.2 Geographic Information Sistem (GIS) and marketing.

By joining the concepts of geography and marketing, the first one being the territorial distribution of phenomena and the second as "[…] the act of knowing the market of action of an organization, to subsequently offer, in an innovative and creative way, products and services that this market wants" (Zela, 2004 apud Furlan, 2011, p. 100), defines geomarketing as "[…] the discipline that studies the existing relations between Marketing strategies and policies and the territory or space, where the institution, its customers, suppliers and points of distribution are located "(Davies, 1976 apud Furlan, 2011, p. 100).

The basis for successful relationship marketing is the identification of specific customer groups that have homogeneous characteristics in detail. In order to identify these segments, a very large number of information is necessary to know the particularities of each group and to be able to satisfy their needs. The segmentation

process is long and complex, since it requires the confirmation that the segments exist, the determination of their characteristics and location so that, from this information, it is possible to elaborate ways of allocating each customer in the correct segment (Shepard, 1993).

The GIS, and consequently the Geomarketing, integrate three types of files: database; geographic files; and point files. The database is information purely external to the company, containing, for example, economic, demographic and social market data. The geographic archives contain the geographical entities defined by their coordinates used in the maps production. The third type of file is the union of the first two, where the data collected is associated with its geographic location. The combination of these three files makes possible the creation of maps and the application of colors, patterns and symbols, simultaneously representing several types of data (Aranha, 1996). The final result is the analysis of market potential, segmentation, customer location or identification of the best commercial point for an organization

2.3 Expert systems as a subsidy for local decision making

The decision about the geographical location of a company is decisive for its success or failure. Therefore, deciding the location of a venture requires strategic decisions based on careful studies, not just on matters of common sense or expert advice. To this end, there are computer tools and techniques available to support this decision process, such as Geographic Information Systems (GIS) and Expert Systems (ES).

From the various definitions of GIS and geomarketing presented in this paper it is possible to affirm that this is a very useful tool to aid decision making in spatial analysis. The possibility of combining data on the location of businesses and service stations, both from the network itself and from competitors, with socioeconomic data, such as income, schooling and demographic density, and making forecasts based on this information represents an advance in the analysis of data, which previously, in isolation, did not generate great benefits for organizations. This differential in the analysis is possible through Machine Learning methodologies.

2.3.1 Machine learning as a mechanism of expert systems

Feigenbaum, a leading expert on Systems Specialists (ES), according to Waterman (1983) and Harmon & King (1988), defines an Expert System as an "[…] intelligent computer program that uses inferential knowledge and procedures to solve problems that require human expertise for your solution". According to Waterman (1986), SE are computer programs that manipulate knowledge to solve problems efficiently in a specific area.

An expert system is then composed of an extensive base of knowledge and rules on a given subject and an inference processor, which uses the basis to draw conclusions and produce judgments about that subject. The machine interprets and decides how the rules should be used and in what order, thus inferring new knowledge (Genaro, 1986).

In this study, Machine Learning will be used as a Specialist System mechanism. According to Monard & Baranauskas (2003), a Machine Learning System is a computer program that can make decisions based on accumulated experiences.

This program uses inductive inference to derive new knowledge and predict future events. Induction is a form of logical inference that allows the generalization of a validated model for a specific sample. For this reason, it is necessary to be cautious in choosing the quantity and quality of the examples that will be presented, since this may cause the hypotheses generated to be of little value and not preserve the truth (Monard & Baranauskas, 2003).

Inductive learning is divided between supervised and unsupervised. In the first case, training data is provided to the inductor (learning algorithm) which individually contains a number of characteristics as well as the class associated with them. That is, in the supervised learning is provided for the algorithm a set of examples, where each one is associated to a group of characteristics that define a certain class, belonging to a discrete (nominal) set of classes {C1, C2, .. . In this way, the induction algorithm will be able to correctly determine the class of a new example that presents only its group of characteristics as information (Monard & Baranauskas, 2003).

In non-supervised learning, the inductor analyzes the examples provided and tries to determine if there is any way to group them, forming the so-called clusters. After this step, an analysis is usually necessary to identify what each grouping means in the context of the problem being studied.

According to Michalski (1983 apud Monard & Baranauskas, 2003) and Kubat, Bratko, Michalski (1988 apud Monard & Baranauskas, 2003), learning systems can be classified into two broad categories:

I. Black-box systems: systems that do not present clear internal results about the concept created. That is, their internal representation and recognition process can not be easily interpreted by humans;

II. Knowledge-oriented systems: They aim to create outputs in the form of symbolic structures that are understandable by humans.

In general, the specification of the problem occurs and the selection of the set of examples that will serve as input to the inductor. After induction, a classifier capable of making future decisions is generated based on the information provided to it in the first step. Afterwards, a classifier validation is made, where, considering its precision, changes are made to the problem specification and data selection, in order to improve the system as a whole.

Another interesting way to illustrate the Machine Learning process is to imagine a graph containing all the examples offered to the inductor. In this scenario, the objective of the classifier is to identify which class belongs to each example and to separate them in a linear way, as shown in Figure 1:

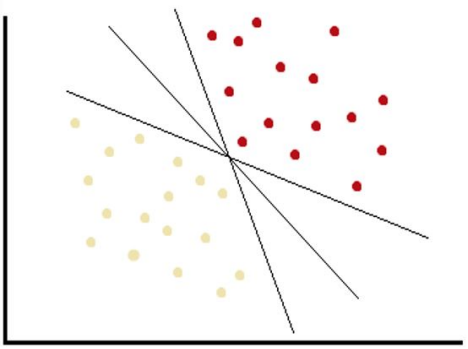

Figure 1. Separation of examples. Source: Elaborated by the authors.

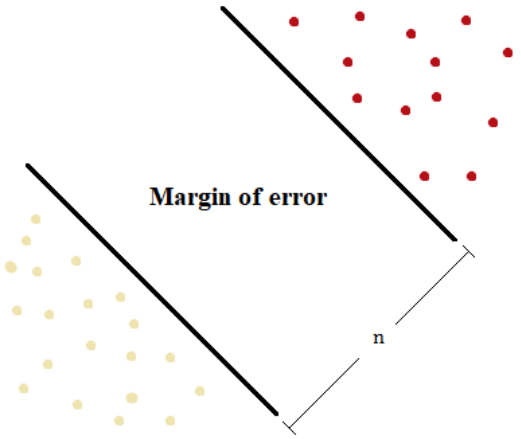

Figure 2. Maximization between the lines. Source: Elaborated by the authors.

As you can see, there are endless ways to separate the classifiers. Therefore, the classifier will find the one that provides the largest margin between classes, aiming to increase the distance (n) between the limits of each class. The higher the value of n the lower the probability of classification error, as illustrated in Figure 2.

3 Search methods and techniques

3.1 Search overview

The databases used in this study were the Pesquisa de Orçamentos Familiares (POF) 2008-2009 and the Demographic Census of 2010, made available by the Brazilian Institute of Geography and Statistics (IBGE). According to IBGE (2010), the POF 2008-2009 aimed at the composition of domestic budgets by collecting data on consumption habits, spending allocation and income distribution, also considering the characteristics of households and people interviewed. Also according to the IBGE, the Demographic Census is a research carried out every ten years, where researchers from the Brazilian Institute of Geography and Statistics visit all households in the country applying questionnaires that aim to measure population density and know the profile of the Brazilian population .

Considering that this research has as objective the elaboration of a quantitative method that assists the locational decision process for a network of franchises, it is important to emphasize that the choice of the franchise to be studied has a secondary importance within the construction of the results, since the the purpose is to construct a method that can be reproduced for any institution that is in a process similar to what is being presented.

The activity chosen for the application of the method was the commercialization of chocolates, which has one of its main franchises the Cacau Show network.

The population was then defined as all the stores, located in the Federal District, which have as their main product the fine chocolate, as well as all individuals residing in the Federal District and consuming chocolate. As for the sample, 44 chocolate shops

were selected in the Federal District and all the individuals identified as chocolate consumers in the 2008-2009 Family Budget Survey.

It is also necessary to have a precise segmentation of the market. As previously seen, Machine Learning, according to Monard & Baranauskas (2003), uses inductive inference to generalize models and predict future events based on a set of examples supplied to the inductor. In this way, the data obtained based on the 2008-2009 IBGE POF will serve as input for the Machine to recognize patterns among chocolate consumers and then be able to predict the demand for the product within the region studied.

Defined the demand for the desired product and the sample of franchises and companies within the Federal District that sell similar products or that may be considered competitors, it will be possible to identify business opportunities for the installation of a new franchise.

3.2 Basket of products

For the elaboration of the basket of products, all the products related to the consumption of fine chocolates found in POF 2008-2009 were selected. After the selection, the final basket of products related to the consumption of fine chocolates obtained a total of 62 items.

3.3 Selection of variables

After the elaboration of the Basket of Products, the next step was to select the variables that can influence the consumption of chocolate. The company Ipsos carried out in 2015 a research whose objective was to identify the main motivations of consumers to buy chocolate. In his research, conducted with the focus on chocolate, it was found that age is a very relevant factor when trying to understand the consumption of chocolate in Brazil. Their data show that 89% of respondents, aged between 13 and 19, claim to consume chocolate, while only 42% of respondents over 60 years old had a positive response regarding the consumption of this product, showing that age discrimination can reveal quite different patterns, and the inclusion of this variable in the model tends to increase its explanatory power.

In addition, it has been found that the consumption of chocolate also varies according to the gender of the individual. The survey shows that 71% of women answered yes to chocolate, while only 64% of men responded in the same way. Women also consume the product more often, as 35% of respondents said they eat chocolate at least once a week, and for men that number drops to 30%. In a similar way to that observed with consumers of different age groups, the heterogeneity observed between genders is also relevant for the problem analyzed in this study.

Another variable that can influence the consumption of chocolate is income. Chocolate is a commodity of normal consumption - that is, its consumption tends to grow as the income of the buyer increases; on the other hand, it is a food commodity, so it is expected that at very high levels of income consumption of this product will tend to show marginally smaller increases. Therefore, income not only decisively influences the consumption of chocolate, but also constitutes a potentially non-linear relationship, which justifies its inclusion in the present study. Thus, the variables that will be

considered as influencers of the consumption of chocolate are thus defined: age, gender and income.

3.4 Data treatment and application of the method

After the selection of the variables, the Machine learning process was started, with the objective of obtaining, based on the POF 2008-2009 data, the consumption of individuals found in the Demographic Census of 2010, carried out by the Brazilian Institute of Geography and Statistics (IBGE, 2010).

With this in mind the machine training process, the first step was to read the 2008-2009 POF data. The work was done using the respondents expenses information, as well as the socioeconomic information of these, both made available on the IBGE website.

The goal was to make the machine use the data from POF 2008-2009 to establish a consumption pattern for the chocolate market. This process was divided in two stages: the first was the training and the second the classifier validation.

This division aims to avoid underfitting, which occurs when few representative examples are offered to the learning system, making the hypothesis fit very little to the training set, and overfitting, which is characterized by an over-adjustment of the hypothesis in relation to the training set. Therefore, from the sample manipulation used in the training stage, it is possible to induce hypotheses that best fit the training set, which compromises its performance in new examples. Following the same reasoning, a hypothesis in the underfitting situation performs poorly on a test set and presents a very small performance improvement in the training set. In order to avoid these two extremes, the database was divided into two parts, with 70% being for the training of the machine and 30% for the validation of the classifier. (Monard & Baranauskas, 2003, p. 46).

Briefly explaining the method, we have an example $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i) = (\vec{x_i}, y_i)$ that has m attributes x_{ij} and each attribute corresponds to a coordinate in the description space $1 \leq i \leq n$ and n represents a number of exemples. In addition, each coordinate defined by an attribute x_{ij} is inserted in a region of the description space, which was associated, by the classifier, with a class y_i. (Monard & Baranauskas, 2003, p. 45).

Another important point in a machine training is the definition of the error rate of a classifier h, as well as its accuracy. The error rate is nothing more than a comparison between the true class of the example and the label imputed by the classifier, generally represented by the formula $Err(h) = \frac{1}{n}\sum_{i=1}^{n} y_i \neq h(\vec{x_i})$, where the operator E retorns the velue 1 when the conditiono E is true and 0 otherwise. This ratio reports the error rate of the classifier. Therefore, in order to define the level of precision, $Acc(h) = 1 - Err(h)$. (Monard & Baranauskas, 2003)

In this work, because it is a regression problem, the comparison between the real class of the example and the assignment given by the machine was done through the Mean Squared Error, given by (1):

$$mse\_err(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(\vec{x_i})\right)^2 \quad (1)$$

It is important to note that we want the construction of classifiers with a low error rate (1) in relation to the test set. Therefore, the calculation of the precision (or error) of the hypothesis focused on the 30% of the sample destined to the validation step.

Having done the division of the database, the next step was the creation of the parameter list, C and σ, which will be used in the learning of the machine. Both parameters are set by the method user. The regularization parameter C represents, to a greater or lesser degree, the importance level of the classification errors generated by the classifier, while the parameter σ is related to the accuracy of the classifier.

For the definition of each pair of parameters, it was used the one that had the lowest Medium Square Error ($mse\_err(h)$). This causes a reduction in the difference between the actual value and that assigned by the classifier, making the machine more accurate and closer to reality.

3.4.1 Method application

This causes a reduction of the difference between the actual value and that assigned by the classifier, causing the machine to become more complete. The machine learning has been completed, and the consumption behavior of chocolate in the Federal District has begun. As seen in the data processing session, it can be seen that the POF 2008-2009 and the 2010 Demographic Census of the IBGE (Distrito Federal) are organized in such a way that, in both bases, the variables of age, gender and income behave in the same way, and only POF 2008-2009 has information related to the expense (expense) of chocolate. Thus, the machine, in a simple way, compared the two bases, applying the pattern of consumption established from the POF 2008-2009 in the 2010 Census. In other words, individuals with similar characteristics of age, gender and income, hypothetically, will have patterns of consumption are also similar. And close to reality.

After the learning process, the behavior of consumption of chocolate in the Federal District was started. As seen in the data processing session, it can be seen that the POF 2008-2009 and the 2010 Demographic Census of the IBGE (Distrito Federal) are organized in such a way that, in both bases, the variables of age, gender and income behave in the same way, and only POF 2008-2009 has information related to the expense (expense) of chocolate. Thus, the machine, in a simple way, compared the two bases, applying the pattern of consumption established from the POF 2008-2009 in the 2010 Census. In other words, individuals with similar characteristics of age, gender and income, hypothetically, will have similar consumption patterns.

The POF 2008-2009 is a nationwide survey and was used in its entirety to establish a consumption pattern. As for the 2010 census, which is also a national survey, only data referring to the Federal District were used.

After the application phase, determine the Standardized Spending (GP) for each census sector of the Federal District. The GP follows the same logic of the standardized score and was defined here as the difference between the Expenditure forecast (G), assigned to a particular census sector, and the Average Expenditure ($\bar{G}$) registered in the Federal District, divided by the standard deviation (σ), as shown in Equation 2 below:

$$GP = \frac{(G - \bar{G})}{\sigma} \quad (2)$$

This equation allows the comparison of the relative position of each census sector, in the case of chocolate expenditure, in relation to the others. Thus, the closer to zero the GP of the census sector, the closer it will be to the Federal District Average Expenditure ($\bar{G}$).

Thus, through the index (GP), the census tracts were classified according to their level of demand for chocolate, as shown in the Table 1 below:

Table 1. Classification of census sectors by demand.

N	COLOR	INTERVAL OF GP	CLASSIFICATION
1		0.00000 - 0.08522	Extremely Low Demand
2		0.08523 - 0.08617	Very Low Demand
3		0.08618 - 0.08685	Low Demand
4		0.08686 - 0.08738	Low Medium Demand
5		0.08739 - 0.08787	Medium Demand
6		0.08788 - 0.08838	High Medium Demand
7		0.08839 - 0.08904	High Demand
8		0.08905 - 0.08997	Very High Demand
9		0.08998 - 0.09285	Extremely High Demand

Fonte: Elaborated by the authors.

Table 1 shows how the classification of the census tracts was constructed, considering nine levels of demand. At the lowest level, the census tracts that presented a Standardized Spend between 0.00000 and 0.08522 have a demand for very low chocolate compared to the rest of the Federal District. For the highest level, those who had their GP between 0.08998 and 0.09285 were considered as regions of very high demand for the product.

4 Results and discussion

4.1 Business and opportunities

After the training step and calculating the standardized consumption for each census tract of the Federal District, the process of recognition of the points that could be classified as business opportunities was started. This stage consists in identifying the areas that are predisposed to chocolate consumption, taking into account the information generated by the Machine Learning method and the application of consumption patterns from POF 2008-2009 in the demographic census database of IBGE (2010) for the Federal District.

In addition to the information obtained in the previous session on the profile of who consumes chocolate and what are the demographic characteristics of these individuals, some qualitative criteria have been defined so that any point can be classified as a business opportunity. It was therefore established that a business opportunity should respect the following criteria:

ELIMINATORY CRITERIA

Being within a region classified by IBGE (2010) as urban;

Being located in a commercial area.

Be part of a census sector that, according to its GP, presents a high, very high or extremely high demand classification.

CLASSIFICATORY CRITERIA

The higher the GP forecast the better the business opportunity compared to the others.

Referring to the abovementioned item “b” of the eliminatory criteria, which deals with the restriction of competition in the same commercial establishment, there is a need for a detailed definition of which stores, present in the Federal District, will be considered as competition for a new cacao-show’s franchise.

Thus, we continued with the search, through Google Maps, for stores that have a proposal similar to the one presented by Cacau Show. The result of this research was a list of 44 stores, where 15 are from the Cacau Show itself, since the allocation of these stores close to each other can generate a process of “cannibalism” within the network. In addition, 9 stores belong to the Brazil Cacau franchise, 10 stores to the Kopenhagen chain, 3 stores to Kaebisch Chocolate and the brands Aguimar Ferreira Bombons, Baby Chocolates, Brigadeirando, Chocolataria Gramado Brasília, Dulce Patagonia, Chocolate Factory and Chocolate Stans have a store each.

Next, the Qgis program was used to create the map. This software allows the practical connection between the Standardized Spending (GP), competition and the digital meshes of the Federal District. As a result of this data crossing, a colored map was created according to the GP (Table 1) and marked with the location of each of the 44 stores that sell fine chocolates from the Federal District.

With this, a general analysis of the DF was made, with the purpose of making possible the identification of points of opportunity for the market of fine chocolates of the Federal District. As shown on the map below, there are regions that can be easily identified as having low competition and an interesting demand for opening a new store. In a fast and strategic way, considering the information on the forecast demand and the competition made available by the map, it is possible to define these regions as possible points of opportunity.

Figure 3 shows that even with the high number of competitive stores, there are easily identifiable points through the darker shades of red, which should be considered as possibilities for opening a new business. As a result, a complex analysis is generated, considering demographic aspects that influence the forecast of demand for the product, and at the same time efficient can reduce in a relevant way the regions that must be analyzed in depth quickly, saving time and availability of organizational resources in the locational selection process.

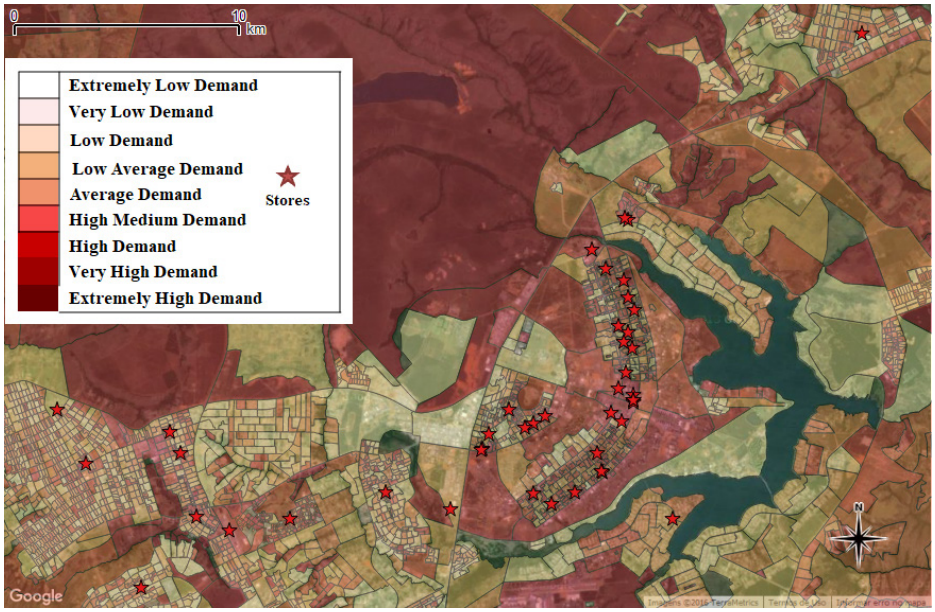

Figure 3. General demand forecast map for chocolate in the Federal District.
Source: Elaborated by the authors.

5 Conclusion and recommendations

5.1 Conclusion

The proposed method consisted in constructing a quantitative tool capable of identifying business opportunities in the Federal District. For this, the fine chocolates market was chosen and the data treatment of the Family Budget Survey 2008-2009 was started, creating the desired variables and eliminating unnecessary information for this analysis. Within the 2008-2009 POF, products related to fine chocolate were selected. Next, we identified which individuals consumed at least one of the items selected, obtaining the total expenditure of each respondent with chocolate, as well as their income, gender, age and location information. From this information, the classifier was able to establish a standardized consumption of the product, having, as a prerequisite for the forecast, the characteristics obtained in the POF - 2008-2009. After Machine Learning, the application stage occurred in the 2010 IBGE Census database, with the final result mapping demand for fine chocolates throughout the Federal District.

From the analysis of the results and the successful mapping of business opportunities to the fine chocolate market in the Federal District, it can be said, with a practical foundation, that Geomarketing, combined with machine learning techniques, represents a tool with valuable resources for the delimitation of organizational strategies, either in the aid of Marketing actions or in matters of locational choice and business expansion, with possibility of focus and consumption forecasts for a specific target audience.

The quantitative tool developed in the present work has shown to be able to help the locational decision process of a franchise network, successfully succeeding in proposing a solution to the research problem, given that the tool's potential to generate improvements in locational choice strategies a network of franchises were evidenced

through the elaboration of the map as final result, containing the demand forecasts for the chocolates market in all the Federal District, and the location of direct competition of the chosen network.

One of the greatest successes of this study was its intellectual contribution to a knowledge area that is still undeveloped, mainly in Brazil, and which can bring many benefits both to the private sector and to the public sphere. The differential of this methodology is the possibility of making estimates of demand, or acceptance of a given product, to new markets, that is, regions that do not yet have a history of consumption in relation to the object of analysis. In addition, the final result of the analysis, as shown in Figure 3, offers the manager georeferenced maps containing detailed information about the functioning of the market in relation to its sector of activity. An example of this is the classification of each sub-district by Standard Expenditure (GP) bands and the location of the competition. An end product such as this reduces the risk involved in the decision and allows the manager, having the report at hand, to trace their strategies quickly and efficiently.

5.2 Limitations and recommendations

Regarding the limitations, some observations and recommendations were identified for future work in the area of Geomarketing and Learning and Machine. The objective of this study was, in an objective way, the elaboration of a quantitative tool capable of identifying business opportunities and contributing to the strategic process of locational choice of a network of franchises. However, it would be interesting if, after identifying the points classified as business opportunities, a “ranking” based on the estimated sales of each business point was made. The intention would be to find the most advantageous place for opening the business. In addition, one of the limitations identified during the study was the fact that the database used in the application - IBGE Demographic Census - contains information collected in the year 2010. Certainly the demographic characteristics of the Federal District have undergone changes during these six years, that ends up compromising, in a way, the prediction of the classifier. The same criticism is made for training and validation data - Family Budgets Survey - considering that they were collected during the years of 2008 and 2009 and may not reliably reflect consumer behavior in 2016.

Another point that requires special attention is the fact that the POF 2008-2009 can not include, depending on the object chosen for forecast, some products or services. An example of this is the luxury markets, which might not be able to use this dataset to map their market, considering that POF 2008-2009, for which it was observed, does not have product specifications for this type of demand . The market for fine chocolates, used in the application of the method, is an area that fits, in a way, in this luxury market. Who consumes a Cocoa-Show chocolate is not necessarily the same person who buys a chocolate bar in the supermarket, which ends up compromising the prediction of the classifier. Despite these limitations, which were basically focused on the database used, the method proved to be extremely robust and proved to be useful both for the private sector and for the public sphere.

The results of this application in the chocolate market of the Federal District provide a window of prospects that have the potential to sediment a more detailed analysis for a great variety of circumstances, since the methodology is robust for versatile applications. In this way, the results of this article are of great value to investors, entrepreneurs, managers, decision makers, policy makers, development institutions

and governments. Therefore, it is recommended the replication of the present study for different market segments for different locations and different time cuts.

References

Aranha, F. (1996). Sistemas de informação geográfica: uma arma estratégica para o Database Marketing. *RAE - Revista de Administração de Empresas*, 36(2), 12-16. http://dx.doi.org/10.1590/S0034-75901996000200003.

Baviera-Puig, A., Buitrago, J. M., Escriba, C., & Clemente, J. S. (2009). Geomarketing: aplicación de los sistemas de información geográfica al marketing. In *Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática*. Winter Garden: International Institute of Informatics and Systemics.

Briozo, R. A., & Musetti, A. (2015). Método multicritério de tomada de decisão: aplicação ao caso da localização espacial de uma Unidade de Pronto Atendimento – UPA 24 h. *Revista Gestão & Produção*, 22(4), 805-819. http://dx.doi.org/10.1590/0104-530X975-13.

Carnasciali, A. M. S., & Delazari, L. S. (2011). A localização geográfica como recurso organizacional. *Utilização de Sistemas Especialistas para Subsidiar a Tomada de Decisão Locacional do Setor Bancário*., 15(1), 103-125.

Cliquet, G. (2006). *Geomarketing: methods and strategies in spacial marketing*. London: ISTE Limited.

Fan, F., & Collischonn, W. (2014). Integração do Modelo MGB-IPH com Sistema de Informação Geográfica. *Revista Brasileira de Recursos Hídricos*, 19(1), 243-254. http://dx.doi.org/10.21168/rbrh.v19n1.p243-254.

Furlan, A. A. (2011). Geoprocessamento: estudos de Geomarketing e as possibilidades de sua aplicação no planejamento do desenvolvimento socioeconômico. *Espaço e Tempo*, 29, 97-105.

Genaro, S. (1986). *Sistema especialista: o conhecimento artificial*. Rio de Janeiro: Livros Técnicos e Científicos Editora S. A.

Harmon, P., & King, D. (1988). *Sistemas especialistas* (A. F. Carpinteiro, Trad.). Rio de Janeiro: Campus.

Instituto Brasileiro de Geografia e Estatística – IBGE. (2010). *Pesquisa de Orçamentos Familiares 2008-2009: despesas, rendimentos e condições de vida*. Rio de Janeiro: IBGE.

Machado, H. P. V. (2016). Crescimento de pequenas empresas: revisão de literatura e perspectivas de estudos. *Revista Gestão & Produção*, 1(2), 419-432. http://dx.doi.org/10.1590/0104-530x1759-14.

Mangini, E. R., Rossini, F. H. B., Santos, A., & Urdan, A. T. (2017). Análise de localização de estações ferroviárias e uso de Geomarketing. *Revista Brasileira de Gestão e Desenvolvimento Regional*, 13(2), 129-152.

Monard, M. C., & Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In S. O. Rezende (Ed.). *Sistemas inteligentes-fundamentos e aplicações* (pp. 89-114). Barueri: Manole.

Shepard, D. (1993). *Database marketing: o novo marketing direto*. São Paulo: Makron Books.

Waterman, D. A. (1983). *Building expert systems*. Canadá: Addison-Wesley Publishing Company.

Waterman, D. A. (1986). *A guide to expert systems*. Canadá: Addison-Wesley Publishing Company.