UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

GRAZIELE CAMARGO KEMMERICH

CONSUMER SEARCH AND PURCHASE: GRAVITATIONAL SPATIAL SALES MODEL USING SEARCH ENGINE QUERY DATA

Porto Alegre
2023

Graziele Camargo Kemmerich

CONSUMER SEARCH AND PURCHASE: GRAVITATIONAL SPATIAL SALES MODEL USING SEARCH ENGINE QUERY DATA

PhD dissertation presented to the graduate business administration program of the Federal University of Rio Grande do Sul as a final requirement to obtain the title of PhD in Business Administration, with an emphasis in marketing

Advisor: Vinícius Andrade Brei

Porto Alegre
2023

CIP - Catalogação na Publicação

Kemmerich, Graziele Camargo

Consumer Search and Purchase: Gravitational Spatial Sales Model Using Search Engine Query Data / Graziele Camargo Kemmerich. -- 2023.

68 f.

Orientador: Vinicius Andrade Brei.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Escola de Administração, Programa de Pós-Graduação em Administração, Porto Alegre, BR-RS, 2023.

1. Search Engine. 2. Prediction. 3. Regression Model. 4. Huff Model. 5. Digital Marketing. I. Brei, Vinicius Andrade, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Graziele Camargo Kemmerich

CONSUMER SEARCH AND PURCHASE: GRAVITATIONAL SPATIAL SALES MODEL USING SEARCH ENGINE QUERY DATA

PhD dissertation presented to the graduate business administration program of the Federal University of Rio Grande do Sul as a final requirement to obtain the title of PhD in Business Administration, with an emphasis in marketing

Examination board:

____

Prof. Dr. Vinícius Andrade Brei
Advisor (UFRGS - Brazil)

____

Prof. Dr. Fernando Bins Luce
UFRGS - Brazil

____

Prof. Dra. Soraia Raupp Musse
PUCRS - Brazil

____

Prof. Dr. Lelis Balestrin Espartel
IADE - Portugal

ABSTRACT

People usually get information from different sources to search for products and services in their purchase journey. Online communication channels, such as search engine platforms, are one of the sources that consumers can obtain information. The search engine platform is a powerful online tool that acts as an intermediary between consumer interest related to a product (organic search) and the interests of the firms that provide this product (sponsored search). Although the Marketing literature has brought important advances in understanding how the search engine platform works and, above all, in understanding the factors that influence a consumer to click or not on an online ad, there is still an important research gap regarding the potential that this type of data can offer. Research in other areas of knowledge has already used the volume of online searches as a proxy to measure consumer interest and, therefore, predict events in the real world, taking advantage of the fact that search engine data provide the geographic location of those seeking information online. Therefore, this dissertation aims to propose a forecast model capable of estimating sales in a given region based on the consumer's online search behavior and the geographic location of these searches. To achieve this goal, four regression models were developed, in which the influence on sales of variables that express online search behavior (Xg) and consumer spatial behavior (Xt and Xu) was tested. Consumer online search behavior was measured based on the volume of searches performed on Google. The spatial behavior was calculated using an adapted version of Huff's probability equation, in which the territorial extension and the urbanized area of cities defined as the origin and destination of displacement were used as attractiveness metrics. The product chosen to get Google data and estimate sales was the light switch. The Google database and the electrical material company's sales base were filtered to cover the period from July 2018 to November 2019. The metropolitan cities located in the Brazilian states of Bahia (BA), Espírito Santo (ES), Goiás (GO), Pará (PA), Pernambuco (PE), Minas Gerais (MG), Rio Grande do Sul (RS), Santa Catarina (SC), São Paulo (SP) and Tocantins (TO) were the regions chosen for model application. The results showed that the consumer's online search behavior is a good predictor of sales in regions where information about the selected product was searched on the internet. However, the performance of the models did not have a strong performance in the case of state capitals. In addition, in terms of predictive power, the use of the variable that uses the value of the urbanized area as an attractiveness metric presented a performance equivalent to that whose metric used was the value of the territorial extension of the city. The company's lack of sales reach in relation to the selected product and the lack and/or little return of online search volume in the region was the main limitation of the research. In addition, it was not possible to precisely identify the location of the online search, just as it was not possible to identify the physical location of the product's sale. Among the suggestions for future research is the expansion of the time period for applying the models and the combination of online search data with other forms of variables that represent the spatial behavior of the consumer.

Keywords: Search engine. Prediction. Regression model. Huff model.

RESUMO

As pessoas costumam obter informações de diversos tipos de fontes para buscar produtos e serviços em sua jornada de compra. Os canais de comunicação online, como as plataformas de motores de busca, são uma das fontes através das quais os consumidores podem obter informação. A plataforma do motor de busca é uma poderosa ferramenta online que atua como intermediária entre o interesse do consumidor relacionado a um produto (pesquisa orgânica) e os interesses das empresas que fornecem esse produto (pesquisa patrocinada). Embora a literatura de Marketing tenha trazido importantes avanços no entendimento de como a plataforma de search engine funciona e, sobretudo, no entendimento dos fatores que influenciam um consumidor a clicar ou não em um anúncio online, ainda existe uma lacuna importante de pesquisa no tocante às potencialidades que esse tipo de dado pode oferecer. Pesquisas em outras áreas do conhecimento já utilizaram o volume de buscas online como proxy para mensurar o interesse do consumidor e, com isso, prever eventos no mundo real, aproveitando o fato de que os dados de search engine fornecem a localização geográfica de quem busca informação online. Sendo assim, o objetivo dessa tese é propor um modelo de previsão que seja capaz de estimar as vendas em uma determinada região a partir do comportamento de busca online do consumidor e da localização geográfica dessas buscas. Para atingir esse objetivo, foram desenvolvidos 4 modelos de regressão, em que foi testada a influência nas vendas de variáveis que expressam o comportamento de busca online (Xg) e espacial do consumidor (Xt e Xu). O comportamento de busca online do consumidor foi mensurado a partir do volume de buscas realizadas no Google. Já o comportamento espacial foi calculado através de uma versão adaptada da equação de probabilidade de Huff, em que foram usadas como métricas de atratividade a extensão territorial e a área urbanizada de cidades definidas como origem e destino do deslocamento. O produto escolhido para obter os dados do Google e estimar as vendas foi o interruptor de luz. A base de dados do Google e a base de vendas da empresa de materiais elétricos foi filtrada de modo a abranger o período de julho de 2018 a novembro de 2019. As cidades metropolitanas localizadas nos Estados brasileiros da Bahia (BA), Espírito Santo (ES), Goiás (GO), Pará (PA), Pernambuco (PE), Minas Gerais (MG), Rio Grande do Sul (RS), Santa Catarina (SC), São Paulo (SP) e Tocantins (TO) foram as regiões escolhidas para aplicação dos modelos. Os resultados demonstraram que o comportamento de busca online do consumidor é um bom preditor de vendas nas regiões em que a informação sobre o produto selecionado foi buscada na internet. Contudo, o desempenho dos modelos não tiveram um desempenho acentuado no caso das capitais dos Estados. Além disso, em termos de poder preditivo, o uso da variável que utiliza como métrica de atratividade o valor da área urbanizada apresentou desempenho equivalente àquele cuja métrica empregada foi o valor da extensão territorial da cidade. A ausência de alcance de vendas da empresa em relação ao produto selecionado e a falta e/ou pouco retorno de volume de buscas online na região foi a principal limitação da pesquisa. Além disso, não foi possível identificar com precisão o local da busca online, assim como não foi possível identificar o local físico da venda do produto. Entre as sugestões de pesquisas futuras está a expansão do período de tempo para aplicação dos modelos e a combinação dos dados de busca online com outras formas de variáveis que representam o comportamento espacial do consumidor.

Palavras-chave: Plataforma de busca. Previsão. Modelo de regressão. Modelo de Huff.

LISTS OF TABLES

LISTS OF FIGURES

SUMMARY

1 INTRODUCTION

People seek to achieve certain goals in their lives. Depending on the intended goal, there are paths to be taken, with different types of decisions to be made. The consumer literature defines this path as the consumer journey, a process that encompasses several actions, interactions and experiences that take place during certain stages (Hamilton; Price, 2019; Lemon; Verhoef, 2016). In the beginning of this journey, the consumer evaluates different sources of information about the product or service that he/she wants to purchase, and this information can come from social sources (e.g. friends, colleagues, social media users) or it can come from online communication channels (Hamilton *et al*., 2021; Li *et al*., 2020). One of these communication channels is the search engine.

Search engine platforms are tools in which consumers can obtain information about products (organic search), while firms can advertise the sale of these products to consumers (sponsored search). As highlighted by Li and colleagues (2020), this type of online information channel allows a quick and efficient search and it is the preferred choice when the consumer wants to acquire certain types of purchases, like in the case of utilitarian products. Currently, people have many options of search engine platforms (e.g. Bing from Microsoft, Google Search, Yahoo! Search, DuckDuckGo, among others), with Google being considered the most utilized among them (Hu; Du; Damangir, 2014). This greater utilization of Google Search also applies in academic settings, because many studies use data from the platform (through search queries) as a proxy to measure consumer interest (e.g. Choi; Varian, 2009; Hu; Du; Damangir, 2014).

Considering a linear online search process, the consumer first seeks for information about a product or service and then he/she decides to accept or not the ad shown in the page results. In this path, it is important to understand the reasons that lead consumers to click or not on advertising, because this consumer action may contribute to increased relevant metrics related to profitability of firms and also to the profitability of search engine platforms, such as click-through rate (CTR) and conversion rates (Méndez-Suárez; Monfort, 2020; Park; Agarwal, 2018). Several past studies have already identified relevant factors that explain consumer click behavior, such as the characteristics of keywords (e.g. rutz; Bucklin, 2011; Rutz; Bucklin; Sonnier, 2012), the quality of the advertiser (e.g. Pallant *et al*., 2017; Yang *et al*., 2015) and the position in which the ad appears in the ranking of search results (e.g. agarwal; Hosanagar; Smith, 2011; Ghose; Ipeirotis; Li, 2012; Ghose; Yang, 2009; Jerath *et al*., 2011).

However, although past research has revealed interesting findings, there still are some unanswered questions. The greater focus to understand the consumer response to paid search advertising has neglected the consumer initiative, that is, the starting point of the consumer search process related to the moment before click behavior. The literature refers to this initiative as "consumer search behavior" or "consumer keyword search" (Park; Agarwal, 2018). This behavior is related to the consumer's interests and it can be investigated through the search queries, a type of user-generated content in the form of textual data, that consumers use on the search engine platform (Berger *et al*., 2020; Bradlow *et al*., 2017; Humphreys; Wang, 2018).

Recently, some studies have brought important findings about this initial stage of the consumer journey. Humphreys and colleagues (2021) identified an association between the language used by the consumer in the phrases of their search queries and the mindset that he/she presents in the purchase path. Li and Ma (2020) found a connection between the semantic information contained in the search phrases and the consumers' purchase decisions. Other aspects can be explored in order to understand the reasons that lead someone to put a term in the search engine tool box and to investigate what impacts result from this behavior.

In other areas of knowledge, such as epidemiology, studies are using search engine query data in predictive models, trying to anticipate events in the real world, like the incidence of disease outbreaks. In particular, recent research has used the volume of consumers' search queries associating them with spatial information. In other words, the focus is not just on predicting the incidence of an event, but is also to predict where this event might happen geographically. This is the case of the study of Lampos and colleagues (2021) that used the volume of search queries to map the cases of coronavirus in different countries, as well as that of Aiken and colleagues (2020), who applied three types of predictive models to anticipate outbreaks of certain diseases in specific regions of the world.

Although the marketing literature have some examples of studies that used consumer search queries to predict consumption (e.g. Du; Kamakura, 2011; Hu; Du; Damangir, 2014), to the best of my knowledge, the association between sales prediction, search engine query data and spatial information is still an unexplored gap.

Therefore, this dissertation aims to address the following research question:

How can search engine query data be used to predict sales geographically?

The remainder of this dissertation is organized as follows. In Section 2, I provide an overview of two literature streams: search engine marketing and spatial marketing. Then, in Section 3, I describe methodological procedures, such as data collection, pre-processing and description of the models. Section 4 presents the empirical results. Finally, Section 5 brings the main contributions, limitations and suggestions for future research.

2. THEORETICAL BACKGROUND

2.1 SEARCH ENGINE MARKETING

Before the purchase decision, the consumer collects different sources of information and evaluates the alternatives (Li *et al.*, 2020). One source for individuals to get information about a product or service in their purchase journey is through social contacts, that is, through interactions with people who belong to their social network, such as family, friends and colleagues, but this information can also arise from proximity with other consumers in the store environment (Hamilton *et al.*, 2021; Lemon; Verhoef, 2016).

Lemon and Verhoef (2016) consider this kind of contact an external contact, a social touchpoint of the consumer journey that influences the purchase decision process and the customer experience. According to these authors, other consumers who join the same store environment may influence actions of individuals through extra role behavior or simply through proximity. In the same way, Grewal and Roggeveen (2020) comment that social influence can occur through an active form, with verbal or physical interactions between the customer and others (e.g. verbal communication interactions), or a passive form through the physical presence of others, and this presence can be represented by other shoppers or employees (e.g. nonverbal social information, social appearance, to name just a few).

Another way to obtain information during the journey is through online searches. According to Bradlow and colleagues (2017), the web-search results expose the consumer to a large variety of need-solutions (especially in terms of products) in a more effective way compared to traditional advertising. Currently, there are many options of online communication channels, such as search engines, social media channels, third-party reviews and deal sites, to name just a few. Different moments of the journey reveal different preferences regarding the type of channel chosen and different channels are associated with different purchase purposes, as demonstrated by Li and colleagues (2020). The authors found that when the purpose of purchase is utilitarian, channels like search engines and third-party review sites tend to be more effective in this type of search. In contrast, when consumers want to make a hedonic purchase, they tend to search through channels that involve an emotional experience, such as social media.

Li and colleagues (2020) point out that search engines are a resource that allows an easy, fast and efficient search for products of interest through the use of search queries. Hu and colleagues (2014) mention that consumers who use the Internet to obtain product information have relied increasingly on search engines, because the abundance of information available in this online communication channel helps them to find the most relevant information about their consumption needs.

The search engine can be understood as a platform that acts as an intermediary between the consumer's interests and the firm's interests. When a consumer seeks information about a specific subject (e.g a brand, product or service), he/she inserts this term into the platform's search box (this term, called a "search query", can be a word, an expression or a combination of words), and the search engine algorithms provide results related to that term for the consumer. At the moment that this natural consumer search (called an "organic search") is carried out, the search engine also provides a list of sponsored results on the screen. Companies that sell the searched items partake in real time and continuous auctions for specific keywords. Companies are interested in those terms that the consumer is likely to click on. The order (also called "ranking") in which the sponsored results appear is rendered based on a combination of the bids submitted and the performance of past clicks. The rank of organic links is related to the relevance of the search query and can be measured by the daily number of buyers that search for products or sellers.

Currently, a considerable portion of marketing articles about search engines focus on understanding how the search engine platform works, because the knowledge about these aspects contributes to profitability of firms (e.g. Katona; Sarvary, 2010; Yao; Mela, 2011). It is important to note that the online search process is costly, both for the consumer and for the firms. Therefore, the creation of ways to improve and optimize this process through the development of models or the application of techniques that allow personalization of search results has aroused great interest in marketing research (e.g. Yoganarasimhan, 2020).

Understanding aspects related to auctions and bids on search engine platforms also helps to understand the consumers' click behavior. Park and Agarwal (2018) mention that the profitability of the search engine and the online advertiser is correlated, that is, the fact that the consumer clicks on sponsored ads found in the results page is positive for both sides. The sponsored ads are classified by Lemon and Verhoef (2016) as a brand-owned touchpoint of customer experience, meaning that this touchpoint is an action that can be more easily managed by the company. Therefore, it is not surprising that there are many marketing studies about search engines to investigate the factors that contribute to the consumer clicking on the ad. Some of the main findings about these factors are the specific location (or placement) occupied by the ad in the search results (e.g. Agarwal; Hosanagar; Smith, 2011; Ghose;

Ipeirotis; Li, 2012; Ghose; Yang, 2009; Jerath Et Al., 2011) and the keyword selection strategy (e.g. Rutz; Bucklin, 2011; Rutz; Bucklin; Sonnier, 2012).

Regarding consumer click behavior, Jeziorski and Moortly (2018) found that there is an interaction between the position of the advertisement on the search page and the brand identity. One of the important findings of the research is that companies that are starting in the market would be more favored with an ad positioned at the top of the page, compared to companies already consolidated in the market. In other words, a higher ad position is more valuable to a less prominent search advertiser than to a more prominent search advertiser (Jeziorski; Moorthy, 2018).

As highlighted by Berman and Katona (2013), consumers tend to rely more on organic links and this finding makes advertisers look for increasingly sophisticated ways to insert their ads in the online consumer search process. However, this sophisticated strategy requires the selection of the most appropriate keyword. In this way, the keyword is another element explored in marketing articles on search engines. Search ads have very limited space, a few lines that must include all important information about the product, brand or service (headline, description and display URL). In this context, it is important to have a match between the keyword that appears in a consumer's search query and the keyword that the advertiser bid to display in the search engine (Klapdor *et al*., 2014; YANG et al., 2015). So, selection of the right keywords by the firm is an important initial step for companies (Klapdor *et al*., 2014).

Klapdor and colleagues (2014) investigated the role of keyword characteristics on the effectiveness of paid search campaigns. Keyword performance was measured by metrics called click-through rate (CTR) and conversion rate (CR). The authors confirmed the importance of keywords containing an advertiser's name. Yang and colleagues (2014) analyzed whether and how market competition impacts what firms advertise in their search ads. The authors found that different competitive settings lead to different content and advertisement types. Market competition leads to more price advertising than brand advertising. In addition to this, intermediaries are likely to engage in more ads with price content when there are more intermediaries in a market compared to the situation where there are more brand suppliers in the market.

According to Humphreys and colleagues (2020), the language used by the consumer in search queries indicates the mindset (or also called "level of mental construal") that she/he presents in the shopping journey. The authors found that consumers tend to generate phrases that are more abstract when their mindset is more abstract. These findings are important because, as the authors point out, when there is a convergence (match) between a consumer's goal at a certain stage of the purchase journey (symbolized by the consumers' mindset), there is a greater chance that the content displayed on search platforms (such as paid search advertising) will receive positive responses from the consumer (e.g. consumer click behavior, visits on the firm's website, purchase intention).

In the same way, the language used by the consumer was also studied by Li and Ma (2020), who investigated the connection between the semantic information contained in search phrases and the consumers' purchasing decisions. The research findings showed that there is a relationship between the stage of the journey and the searched topic, with consumers at the beginning of the journey prioritizing topics related to convenience and loyalty aspects. This finding about the initial stage of the journey is important to understand the search process. As Humphreys and colleagues (2020) highlighted, the early stages of the consumer decision journey has not been widely investigated in the literature, although it is very valuable in terms of insights about consumer behavior.

2.2 SPATIAL MARKETING

Attractiveness is one of the most important conditions for the survival of a store (Cliquet, 1995). According to Douard and colleagues (2015), retail attraction is characterized by a product or service that exerts attraction from a point of sale with the status of a pole of attraction. In this case, the attracted entity is represented by the consumer. Therefore, commercial dynamics of territories, business performance of areas and consumers' spatial behavior are important factors to be incorporated into firms' marketing decisions (Douard; Heitz; Cliquet, 2015). The choice of the right location for a new retail business could maximize profitability (Suhara et al., 2021).

Spatial models are often used to measure the attractiveness of a store over certain periods of time (Cliquet, 1995). The literature distinguishes two model approaches: deterministic and probabilistic. The deterministic models, supported by authors, such as Converse (1949) and Keane (1989), assume that consumers are attracted to visit a city or a particular store according to a determined utility function (Douard; Heitz; Cliquet, 2015). This type of model performs better when applied in rural contexts and does not have a very high-quality predictive power (Cliquet, 2013).

The probabilistic model, in turn, assumes that consumers are attracted to a particular store according to a function that considers the past behavior of individuals, related to the probability that a consumer will frequent this particular store (Cliquet, 2013; Douard; Heitz; Cliquet, 2015). The application of this type of model is more suitable in urban contexts, and this type of model has a very high-quality predictive power compared to deterministic models (Cliquet, 2013).

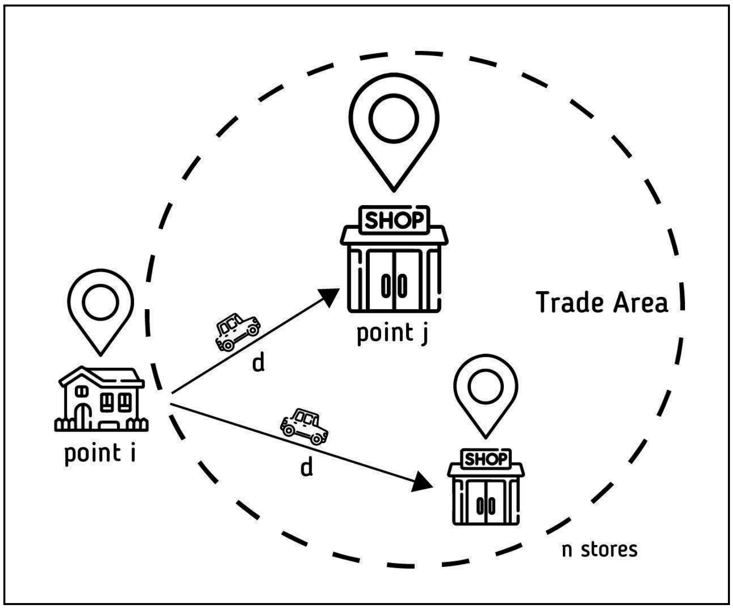
According to Douard (2015), probabilistic models, also defined as gravity-type models, are spatial applications that better capture consumer choices related to the attractiveness of a point of sale. Among the models that follow the probabilistic approach, the model developed by Huff (1963) is the most used, mainly due to its flexibility and adaptability (Suhara *et al.*, 2021). From the beginning until now, the Huff model has been applied for tasks such as predicting consumer spatial behavior, delineating trade areas, locating retail and service facilities, analyzing market performance, simulating different market scenarios, and forecasting sales (HUFF, 2003).

The Huff model, as an analogy of Newton’s law of universal gravitation, is essentially a gravity-based spatial interaction model, because it uses the notions of distance and mass (Liang *et al.*, 2020). The geographic or temporal distance and the mass (represented by the sales surface area of the store) are the two major factors that influence the likelihood that a consumer will choose a store (Cliquet, 2013; Douard; Heitz; Cliquet, 2015; LIANG et al., 2020; Suhara *et al.*, 2021; Wang *et al.*, 2016). The Huff model is calculated as follows (Equation 1):

$$P_{ij} = \frac{A_j / D_{ij}^{\lambda}}{\sum_{j=1}^{n} \left(A_j / D_{ij}^{\lambda} \right)} \qquad (1)$$

where *Pij* is the probability of customers located in region *i* go to visit commercial facility or retail agglomeration *j*; *A*j is the attractiveness, represented by the size of store (or commercial facility or retail agglomeration), measured in square meters of sales surface area; *Dij* is the distance between the consumer located in region *i* and commercial facility or retail agglomeration *j;* λ is a parameter that reflects the effect of distance decay estimated from empirical observations, and n is the total number of commercial facilities or retail agglomerations, including the store *j.* Figure 1 illustrates the main elements of Huff's probability equation in a hypothetical situation:

Figure 1. The original Huff model



Source: prepared by the author.

As highlighted by Liang and colleagues (2020), Huff's original model also associated store attractiveness with its merchandise offerings, which is the ability of the store to fulfill the customers' needs. When a store has a large number of items, it has a greater chance of attracting more consumers to visit, even if these consumers are located in other regions (Cliquet, 1995). The visitation possibilities, distances and attractiveness are the necessary attributes to apply as inputs to the Huff model (Wang et al., 2016). These attributes are capable of reflecting the purchase flow of consumers, identified through the delimitation of the trade area (Douard; Heitz; Cliquet, 2015).

According to Huff (Huff, 1964), the "trade area" is defined as a geographically delineated region containing potential customers to purchase products or services offered by a particular firm or by a particular agglomeration of firms (Huff, 1964; Liang *et al*., 2020). Surveys are commonly used to obtain the spatial information to delimit the trade area, such as how often consumers make purchases (e.g. how many times per week), and which places (commercial centers or stores) they typically visit to shop or make their purchases (Wang *et al*., 2016).

In the same way, Douard and colleagues (2015) comment that the starting point to analyze consumers' purchase flow is to divide up a territory into basic geographical sub-areas, which may be neighborhoods for cities, or municipalities or groups of municipalities. The origin of the purchase flow is commonly defined as the consumer's area of the residence, and the destination of the purchase flow is defined as the retail agglomeration belonging to this trade area (Douard; Heitz; Cliquet, 2015). The preference for the consumer's place of residence as the origin of the flow is due to the fact that people tend to visit stores and make their purchases closer to where they live or work, because the shorter distances between these two points means less resources (e.g. time and money) employed, and this increased accessibility positively impacts the willingness to visit the store (Liang *et al*., 2020; Suhara et al., 2021).

Previous research demonstrated the importance of delimiting the geographical context to analyze the spatial behavior of consumers. Cliquet (1995), for example, defined the political subdivisions of departments in France (called cantons) as the geographic space to apply the Multiplicative Competitive Interaction Model (MCI model) and to introduce in this model consumer judgments as subjective variables related to store attraction. Douard and colleagues (2015) conducted their studies about consumer spatial behavior on a geographical area straddling Val d'Oise and Yvelines in France to analyze the commercial attractiveness of five nearby towns. Suhara and colleagues (2021) split their dataset into 17 regions based on the administrative districts to apply the Huff model to different merchant categories. And, recently, Wang and colleagues (2016) delimited trade areas using social media data. These authors used a clustering algorithm to extract the activity center for Chinese social media users.

2.2.1 Adaptations of Huff's model over the years

Compared to other methods that map retail poles of attraction, the factors considered by the Huff model are relatively complete, as they are able to identify with good accuracy the probability of consumers moving from their origin area to a store or retail agglomeration (WANG et al., 2016). However, despite the good performance of the Huff model, other studies have emerged in the literature, over the years, to make improvements to the original model and adapt it to different contexts, by including more variables in addition to store size and distance. The consumer's choice process is complex, and there is a multiplicity of variables that influence this choice (Cliquet, 1995; Douard; Heitz; Cliquet, 2015).

One of the first adaptations with the goal of improving the Huff model were proposed by Nakanishi and Cooper (1974) in their MCI (Multiplicative Competitive Interaction) model. The MCI model takes into account the relationship between market share and marketing actions, and allows a combination of multiple dimensions of attractiveness into a single measure (Cliquet, 1995; Douard; Heitz; Cliquet, 2015; Suhara *et al*., 2021). Cliquet (1995) comments that, although the MCI model incorporates new variables to Huff's original model, these variables tend to be objective, such as the number of customer visits, the number of employees or the number of checkout lanes. The inclusion of these variables encouraged the emergence of model updates.

The so-called "MCI-type models" are models that still have the multiplicative essence of the model developed by Nakanishi and Cooper (1974), however, allow the inclusion of subjective variables related to perception of the offering and marketing actions, as well as store images (Douard; Heitz; Cliquet, 2015). Cliquet (1995), for example, conducted surveys to collect the personal opinions regarding subjective factors of store attractiveness, such as the perception that consumers have about the product quality and sales competence.

The evolution related to data collection also favored the emergence of new extensions of Huff's model. Liang and colleagues (2020) developed a more dynamic Huff model to estimate hourly store visits. The authors incorporated granular spatio-temporal mobility data to analyze the factors that influence consumer store visits across categories and brands in the 10 most populated US cities. Suhara and colleagues (2021) developed a model to estimate customer retail patronage through real transaction data collected from customers’ credit card activities. Sevtsuk and Kalvo (2018) proposed a variant of the Huff model that uses street network-based distance. And, more recently, Busu and colleagues (2020) demonstrated that financial attributes, such as current asset, fixed asset, and the number of employees can be good attractiveness measures in a model to predict the net profit of a store.

3. METHOD

3.1 THE CHOICE OF SEARCH ENGINE PLATFORM (GOOGLE)

The decision to use Google in this research is because it is the most well-known and widely used search engine platform in the world (Hu, Du, and Damangir 2014). This more widespread use of the platform and the fact that data from Google is obtained in a more accessible and less invasive way compared to other methods traditionally used to collect information from consumers aroused interest for its use also in the academic setting. Search queries entered by consumers in the platform's search box can be used as a proxy to measure consumer interest (e.g. Hu; Du; Damangir, 2014; choi; Varian, 2009),

Another advantage of using Google data is the fact that this platform provides searchable information that can be customized, as highlighted by Hu and colleagues (2014). Search terms can be filtered by different textual forms (e.g. single word, a combination of words or phrases) and also by geographic areas and time ranges.

It is important to highlight that the tool used to collect user search data was the "Google Keyword Planner". This tool is accessible to users who have an administrative panel to manage the creation of online advertisements. Both Google Trends (resource most commonly cited in academic publications) and Google Keyword Planner refer to the same database (that is, search information from platform users). However, the presentation of data takes place in different ways. Google trends provides search volume as a normalized index from 0 to 100 (relative value), while Google Keyword Planner provides the actual value. This last data return option is more suitable for the present research. The period of time was defined to encompass the data returned by search engine platform and the sales data available. The time period was from July 2018 to November 2019.

3.2 THE PRODUCT CHOICE (LIGHT SWITCHES)

After defining the search engine platform that would be used to obtain the data, another important decision in the development of the present research was the choice of a product. The decision related to product choice is important because it influences the term that will be used in the search box. It is on the basis of this term that the data representing the consumer's interest in their online search process will be collected.

The firm has a variety of products for the electrical materials market, both for residential and industrial purposes. The light switch was chosen for two main reasons: 1) the product has the highest sales volume in the company, considering the electrical materials segment; and, 2) the product is familiar to the final consumer, because it is commonly purchased for use in homes. Consumers are more likely to perform internet searches for a product that is familiar. Since the predictive models will be applied in Brazilian cities, the product name that will be used as a search query will be written in Portuguese ("interruptor").

It is also important to highlight the utilitarian aspect of the light switch. According to Li and colleagues (2020), search engine platforms are often the preferred choice when the motivation for a purchase is utilitarian, mainly because this online channel is efficient for consumers to make comparisons for product offerings, product attributes and their prices before completing the purchase transaction. Pallant and colleagues (2017) comment that consumers who go to a retailer's website by clicking on a search engine link are more likely to make visits that are associated with a purchase goal. This means that people looking for utilitarian products on search engine platforms are often not merely browsing.

The utilitarian bias of the chosen product (light switches) also favors the application of gravitational models. Recently, a study that applied a gravitational model to 04 merchant categories (Grocery, Gas Station, Clothing and Restaurant) showed that the performance of the model for the restaurant category was not satisfactory. Suhara and colleagues (2021) found that Pearson's correlation metric performed worse for restaurants compared to the other categories studied. The authors' interpretation was that restaurant customers' patronage behaviors do not fully follow the Huff model's assumption. It is expected, for example, that consumers who are going to go to the grocery or gas station look for a place close to where they live or where they work, due to the accessibility of the place. This behavior is not necessarily the same for consumers who go to restaurants, since the experience in this type of place is more hedonic, as it involves curiosity, novelty and a variety of tastes and expectations that encourage consumers to frequent places far from where they are living or working (Suhara *et al.*, 2021).

3.3 THE CHOICE OF GEOGRAPHICAL AREA

Brazil was the country chosen to be the study area. More specifically, the metropolitan cities located on the following States: Bahia (BA), Espírito Santo (ES), Goiás (GO), Pará (PA), Pernambuco (PE), Minas Gerais (MG), Rio Grande do Sul (RS), Santa

Catarina (SC), São Paulo (SP) and Tocantins (TO). The metropolitan areas consist of arrangements of bordering municipalities and are established by a complementary state law, pursuant to Article 25, paragraph 3 of the 1988 Federal Constitution (IBGE 2023). Although Brazil is a Federative Republic composed of 26 states plus the Federal District (Brasília), it was decided to select only some regions of the country to better understand the development and application of the proposed models. The criteria for this selection were: 1) the urban centers proximity and the equivalence of information between search engine query data and product sales data provided by the firm.

Metropolitan regions, in particular, the metropolitan region where the capital of the State are located tend to favor the movement of people in the region. According to Douard and colleagues (2015), the division of geographical areas to analyze the consumer purchase flow must take into account areas likely to display a certain homogeneity in purchase behavior, in view of access to infrastructure serving retail outlets. The proximity between cities and the accessibility to these places are attributes that make the metropolitan area an appropriate choice to delimitate the geographical area to apply the models. In addition, urban contexts tend to respond better to probabilistic gravitational models, such as the Huff model whose calculation was used as a predictor variable in the forecast model (Alternative Model 03).

The other criterion for choosing the cities was the equivalence between the information collected on the Google platform about the product and the sales that the firm provided of this product (light switch) from July 2018 to November 2019. It was not possible to include all cities of these metropolitan regions, because some of them did not place orders for the product (light switch) in the selected period. Table 1 shows the States and metropolitan cities that were included in this research.

Table 1: Metropolitan regions

State	Metropolitan region	Cities
BA	Salvador	Camaçari; Dias d'Ávila; Lauro de Freitas; Salvador; Simões Filho
ES	Vitória	Cariacica; Serra; Vila Velha; Vitória
GO	Goiânia	Aparecida de Goiânia; Senador Canedo; Nerópolis, Goiânia, Goianápolis
MG	Belo Horizonte	Belo Horizonte; Sete Lagoas; Pará de Minas; Itaúna; Contagem; Betim
PA	Belém	Ananindeua; Belém; Castanhal
PE	Recife	Jaboatao Dos Guarara; Recife; Paulista; Olinda
RS	Porto Alegre	Porto Alegre, Cachoeirinha; São Leopoldo; Gravataí; Campo Bom; Montenegro; Esteio; Sapiranga; Novo Hamburgo; Igrejinha; Ivoti; Guaíba; Charqueadas; Dois Irmãos; Sapucaia do Sul; Taquara; Viamão;
SC	Florianópolis	Florianópolis; Palhoça; São José
SP	São Paulo	Mairiporã, taboão da Serra; Suzano; São Paulo; São Caetano do Sul; São Bernardo do Campo; Santo André; Santana de Parnaíba; Osasco; Mogi das Cruzes; Mauá
TO	Palmas	Palmas, Porto Nacional, Paraíso do Tocantins

Source: Prepared by the author

3.4 DATA PREPROCESSING

This dissertation will use two main datasets: the search engine query data provided by Google and the sales record of a major Brazilian electrical manufacturer.

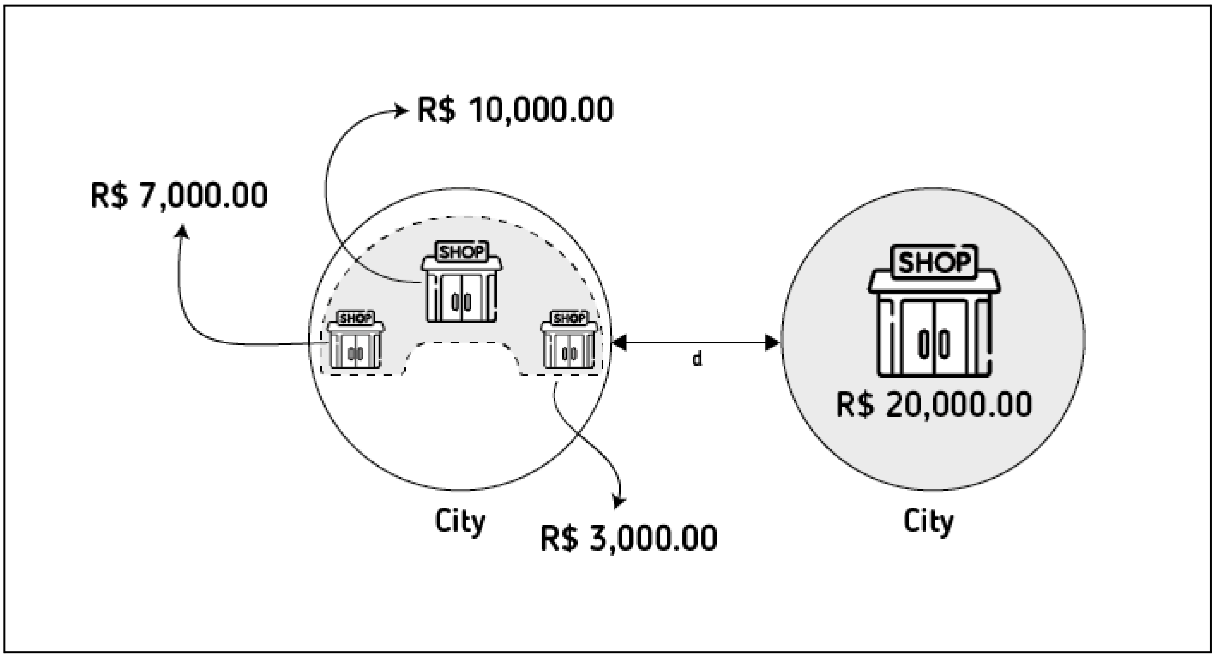
The first step was to organize the firm's dataset. The sales data do not come directly from the final consumer, so it was necessary to make some adaptations for this research. Each city has different points of sale that offer the selected product to the final consumer. However, all these points of sale were grouped and considered as if they were just one. So, all orders placed by the various stores were added, and the total result of these orders was considered to be the total sales made in that specific city. For example, assuming that a hypothetical city X has 03 stores (Store A, Store B and Store C) that sell the selected product (light switch), and each of these stores sells, respectively, R$ 10,000.00, R$ 5,000 and R$ 7.000 of this product, the sales value of each of these stores will be combined, and this total will be considered the sales value of the product for the hypothetical city X. Figure 2 illustrates this procedure.

To filter the products on the firm's dataset, it was checked in the company's catalog which products would be classified as switches for residential use. Using R and Python programming language commands, only the lines containing the names and terms specific to this type of product were selected. Although each product of the company has its code, it was decided to filter the data based on the names that the product "switch" has in the catalog. The option for the name and not for the number (code) is important in the case of this research, considering that the names that represent the product have a greater proximity to the way in which consumers would use it to search for information on the search engine platform.

In the company's database only kept information relevant to the development of the models, such as the location of the companies, type of product, quantity of items sold, value of orders (which would be the purchases that each store makes/orders for the factory or branches of the electrical materials company, as this is a B2B basis) and period of time. In view of this information, a filter was made in the resulting base so that: 1) the cities of the metropolitan regions of the Brazilian states remained; 2) the sales period from July 2018 to November 2019, and 3) products classified as "switches".

The dataset from the search engine platform also was organized. All columns related to the account's overall performances were excluded from the base (e.g. ad impressions, top of page, yoy change, concept), because the goal of this dissertation is not to evaluate the organic reach versus the sponsored reach of the keyword used by the firm in its ad. Therefore, the dataset was filtered in order to keep only the necessary information for the application of the proposed models. This information is the keywords directly related to the search term ("interruptor") and the search volume for these keywords for each of the months of the period from July 2018 to November 2019. After these initial procedures, the list of keywords was filtered so that only the terms classified as being "highly competitive" by the platform remained in the dataset. According to information from Google, highly competitive keywords are the most disputed and the ones that most convert into sales, precisely because they "match" with the queries that people search for on the platform. All keywords classified in this way were grouped and the number of searches for each one of them was added.

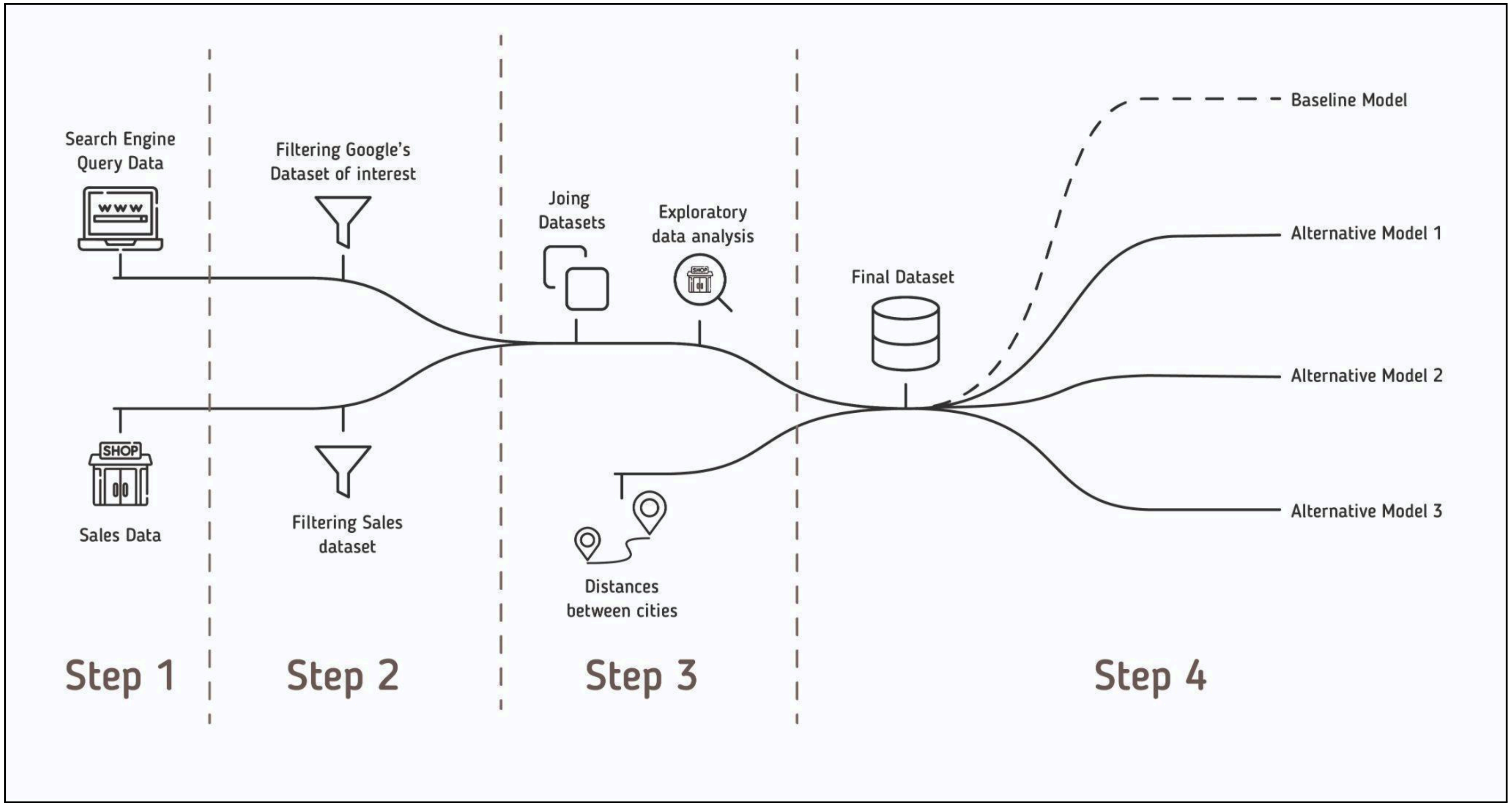
Figure 2. The modified Huff model



Source: prepared by the author.

The figure 3 illustrates the stages that were developed during the development of the research.

Figure 3. Workflow



Source: elaborated by the author.

3.5 VARIABLES DEFINITION

The goal of this research is to analyze if search engine query data are good predictors of models involving gravitational behavior. Based on this goal, it is necessary to define, first, which variables will be applied in the models of the thesis.

3.5.1 The product' sales (*Yi)*

The variable "sales" is indicated in forecast models by the symbol *Yi.* It is the target variable, that is, the variable that the model seeks to predict the value based on the impact of other variables. In the specific case of this thesis, the models try to predict the value of sales considering the impact of the variables amount of orders (*Xi*), search volume provided by Google (*Xg*), and the probability of the consumer moving from his city of origin to the point of sale located in another city with a certain territorial extension (*Xt*) or urbanized area (*Xu*). It is important to point out that the sales are related to the type of product previously selected (light switches), and the cities mentioned belong to the previously defined geographic region.

For a company, predicting the sales of a product based on consumer interest in each region is essential for the business. It is necessary to emphasize that this interest that consumers express when looking for product information on the search engine platform should not be understood as demand. Although "demand" and "sales" are often used by researchers as a single concept to select studies involving the consumer's propensity to acquire a good or service in a given market, these concepts encompass different elements (Cadavid; Lamouri; Grabot, 2018; Fildes; Ma; Kolassa, 2022). Therefore, demand forecasting and sales forecasting are different proposals for estimating future events, each requiring appropriate approaches (Cadavid; Lamouri; Grabot, 2018).

According to Cadavid and colleagues (2018), demand forecasting is usually guided by a more current panorama of what is happening in the market, and it uses data that reflects the effect of prices and promotions made by the company. On the other hand, sales forecasting seeks to make a projection based on the company's recent history, and uses data related to the effect of promotions or stock shortages collected directly from the point of sale (Cadavid; Lamouri; Grabot, 2018).

Fildes and colleagues (2022) comment that estimating demand is crucial for the company to make operational and strategic decisions, such as pricing, space allocation, ordering and inventory management, distribution, and labor force. However, the accuracy of demand forecast models is often a challenge, given the complexity of the data. Demand is influenced by a series of exogenous factors, especially when it is inserted in an unstable market or when the product is impacted by several variables, such as seasonality or weather conditions (Cadavid; Lamouri; Grabot, 2018; Fildes; Ma; Kolassa, 2022).

This complexity involving demand data prevents demand observations made in previous periods from being sufficient to create a model with good predictive performance (Fildes; Ma; Kolassa, 2022). Sales forecasting, in turn, can be developed with reasonable accuracy from a historical series that shows past sales behavior. Particularly concerning this research, the other variables that will be incorporated into the models (e.g., consumer online search behavior) and the nature of the product chosen to test the prediction (light switch) do not suffer sudden and direct fluctuations from the market in which they are inserted.

3.5.2 The amount of orders (*Xi*)

The variable "orders" is indicated in forecasting models by the symbol Xi. It represents the quantity of orders for the product requested from the company that manufactures that product. The number of orders is a variable that directly influences the company's sales, because the greater the volume of orders in relation to the product, the greater the sales volume of this product by the company. As with sales, the number of orders refers only to the selected product (light switch), in cities belonging to pre-established metropolitan regions. Thus, the number of orders for the light switch product placed in the metropolitan city of Florianópolis influences the sales of light switches in the city of Florianópolis, the number of orders for light switches in the city of Porto Alegre interferes with the sales of switches in the city Porto Alegre, and so on.

3.5.3 Search engine query data (*Xg*)

The variable "search volume" is indicated in the forecast models by the symbol Xg, and represents the consumer's online search behavior. The inclusion of this independent variable in the models tests the power that the volume of searches on Google on the product exerts in predicting the sales volume of that product in the regions where the events occur.

In general, predictive models that use search engine query data are associated with other types of sources, such as public or government data. Medical reports, historical epidemiological data and archival data from institutions, like the Center for Disease Control

(CDC) have already been used to detect outbreaks of diseases such as influenza (e.g. Ginsberg *et al*., 2009; Santillana *et al*., 2014), and more recently, the coronavirus outbreak (e.g. Lampos *et al*., 2020; Aiken *et al*., 2020; Galoni; Carpenter; Rao 2020). Other examples of the combined use of search engine data with public source information is the case of the research conducted by Choi and Varian (2009) that predicted periods of economic crisis based on the number of people who applied for unemployment assistance.

3.5.4 Consumer spatial behavior (Xt e Xu)

The purpose of this research is to propose a sales predictive model that combines online search behavior and consumer spatial behavior. As seen previously, the online search behavior was defined based on the volume of searches that the consumer performs on the search engine platform in relation to the product. The spatial behavior of the consumer, in turn, will be expressed in this research through the probability of the consumer moving from his place of origin to the place where there is a point of sale that supplies the product that was researched on the search engine platform.

In the forecast models, the spatial behavior of the consumer will be represented by the variables Xt and Xu. The calculation to obtain the value of each of the variables (Xt and Xu) will be done using the Huff probability formula. However, the original Huff model was adapted to better fit the needs of this research, as well as other authors have done in the past (e.g. Nakanishi and Cooper 1974; Cliquet, 1995; Sevtsuk; Kalvo, 2018; Liang *et al*., 2020; Suhara *et al*., 2021).

The main elements that allow the Huff probability equation to be applied refer to the consumer's place of origin, the store where the visit or purchase of the product is made, the attractiveness of the destination and the distance between the place of origin and the chosen destination. For each of these elements, it was necessary to make some adaptations so that the definition of the variables Xt and Xu could be made and included in the forecast models.

In Huff's original model, the origin was defined as being the consumers' place of residence. In the composition of the variables Xt and Xu, in turn, the origin will be considered as the place where searches are carried out on the search engine platform, identified through the geographical location of the search query. Both the origin and the destination are cities located in the metropolitan region.

Attractiveness is closely related to destination, because the attractiveness of the store, commercial facility or retail agglomeration that motivates the consumer's mobility behavior.

In the original Huff model, attractiveness was measured through the size of the store (square meters of sales surface area). In the composition of the variables that represent the spatial behavior of the consumer, the attractiveness metric will be measured through two types of information: the measure of the territorial extension area of the city and the measure of the urbanized area of the city. Both measures were collected by the Brazilian Institute of Geography and Statistics (IBGE) website.

These two measures of attractiveness were chosen by the author because they are more in line with the proposal established by the research of considering metropolitan cities as the point of origin and destination of the purchase flow. If each metropolitan city plays the role of an individual store, the value of the territorial extension and the urbanized area of the city plays the role that the size of the store plays in the measure of attractiveness of Huff's original model. Therefore, when the variable that represents the spatial behavior of the consumer has in its composition the attractiveness metric referring to the value of the territorial extension of the destination city, it will be symbolized in the forecast model with the symbol Xt. When the variable that represents the spatial behavior of the consumer is defined from the attractiveness metric referring to the value of the urbanized area of the destination city, it will be symbolized in the forecast model with the symbol Xu.

The distance between the city of origin and the city of destination will be defined by the travel time in minutes provided by Google Maps. This platform is a location tool that allows users to search specific geographical locations and calculate routes in an optimized way (Google, 2023). This tool offers the possibility of selecting the route by different kinds of transportation, like car, on foot or by public transport. In the case of this thesis, the distance was defined by the travel time by car, because this type of transportation was used in previous research (e.g. Cliquet, 1995). It is important to highlight that the travel distance between cities was collected during periods of the day when traffic conditions are normal. So, it was avoided to collect information on weekends, holidays and in the early morning and late afternoon during the weeks, because in these periods, the flow of cars is more intense on the highways.

The original Huff's probability model (Equation 1) also includes the parameter λ . It is the exponent responsible for calibrating the effects of the distance (Dij) between the point of origin and the local destination of the consumer (Newing; Clarke; Clarke, 2015). The distance-decay parameter λ is the result of empirical observations and was elaborated by geographers as an effect whose intensity of interaction between two places decreases as the distance between these two places increases (Huff, 1984).

Berman and colleagues (2009) highlight that the distance-decay function is a metric difficult to define. For this reason, the determination of what would be the ideal value raises discussions on the part of geographers and scholars of spatial models (Cliquet, 1995; Douard; Heitz; Cliquet, 2015; Li; Liu, 2012; Newing; Clarke; Clarke, 2015). Because this thesis does not have as its main objective the mathematical deepening of the elements of the Huff equation, but the development of a sales forecast model, it was decided to use a parameter value already validated in the past literature (Li; Liu, 2012; Douard; Heitz; Cliquet, 2015). Therefore, the value of the parameter λ was fixed as 2.

3.5.5 The development of predictive models

In order to develop the baseline model and the alternative models, this thesis follows the procedures and the logic described in previous research that used search engine query data to predict outcomes (e.g. Vosen; Schmidt, 2011; Bulut, 2017). At first, the baseline model of the research was established based on the use of an explanatory variable that was capable to express certain influence on the variable that is intended to predict the future behavior (target variable). In the case of this thesis, it was decided to use the amount of orders as a variable that influences the amount of sales.

The regression equation for the baseline model is describe below, where Yi is represented by the sales of the firm in a given city in the metropolitan region, and Xi is represented by the amount of orders (related to selected products) realized in that city located in the metropolitan region:

$$Yi = \beta o + \beta 1\, Xi + ei \qquad (02)$$

Relevant studies that used search engine query data to predict, such as the research conducted by Vasen and Schmidt (2011), defined their baseline model based on a leading indicator that represented economic information (e.g. private household spending). At first, the author intended to do the same through the application of an index that measures the inflation rate of products from retail trade relative to household expenditure, called Extended National Consumer Price Index (IPCA), provided by The National System of Consumer Price Indexes (IBGE, 2023). However, this index covers families living in certain urban areas of Brazil, and these areas do not include all the metropolitan regions analyzed in this thesis. So, it was decided to use the product orders as the explanatory variable of the baseline model, since the preliminary findings demonstrated a positive correlation between the amount of orders and the sales.

To analyze the impact of search engine query data on sales, the Alternative model 01 considers the sales of each city (Yi) as dependent on the amount of orders *(Xi)* and, also, dependent of the volume of searches on the search engine platform (*Xg)* realized by users located in that city:

$$Yi = \beta o + \beta 1Xi + \beta 2Xg + ei \tag{03}$$

To analyze the impact of the gravitational behavior on the sales, the Alternative Model 02 considers the sales of each city (Yi) as dependent on the amount of orders *(Xi)*, the volume of searches on the search engine platform realized by users located in that city (*Xg)* and, also, dependent on the probability of consumers located in a given city of the metropolitan area to go to visit another city to make their purchases (*Xt*). In this alternative model 02, the attractiveness measure used in the Huff equation will be the value of territorial extension of the city destination:

$$Yi = \beta o + \beta 1Xi + \beta 2Xg + \beta 3Xt + ei \tag{04}$$

The Alternative Model 03 is similar than the model above, however, the attractiveness measure that will be apply in the adapted Huff model will be the value of urbanized area of the city of destination (*Xu*):

$$Yi = \beta o + \beta 1Xi + \beta 2Xg + \beta 3Xu + ei \tag{05}$$

The models described above seek to contemplate situations in which the sales of a product can be estimated based on different, but complementary, behaviors of the consumer: online search behavior for information and the possibility of moving this consumer to the place where the product is sold whose information was obtained from the internet.

4 EMPIRICAL RESULTS

4.1 CORRELATION VALUES

The first point that draws attention in the preliminary analysis of the data is the low correlation between the volume of online searches and the volume of sales in the capitals of the States analyzed. The cities of Belo Horizonte (MG) and Palmas (TO) were the capitals that presented the lowest correlation values, with a Pearson coefficient of -0.004 and 0.0068, respectively. The highest correlation between search volume on the search engine platform and product sales volume was found in the city of Campo Bom (RS), with a Pearson coefficient value of 0.64, followed by the city of Porto Nacional (TO) which showed an association in the opposite direction between sales and online searches, with a coefficient of -0.5. The cities of Sapiranga (RS) and Simões Filho (BA) also showed slightly more pronounced correlation values compared to the rest of the sample, 0.42 and -0.41, respectively. Although these values indicate a weak to moderate association between the variables (sales and searches), it is an observation that deserves to be highlighted, as it stands out from the others. The values of the correlations are contained in Table 5 of Appendix 1.

The possible association between the amount of orders (understood as items that the commercial establishments ordered from the company to restore their stocks) and the volume of searches performed on the product in the search engine platform was also investigated. Contrary to the situation described above, some capitals stood out with more pronounced correlation values, when compared to the other cities in the sample. This was the case of the capitals of the states of Santa Catarina, Espírito Santo and Pará. The highest correlation value was found in the capital Florianópolis (SC), with a Pearson coefficient of 0.54. Other cities that had interesting correlation indices were the cities of Campo Bom (RS) and Lauro de Freitas (BA), with a Pearson coefficient of 0.47 in both cities. Other places that presented similar correlation indexes (above 0.4) were the cities of Sapiranga (RS), Taquara (RS), Ivoti (RS) and Taboão da Serra (SP). The values of the correlations are contained in Table 5 of Appendix 1.

4.2 BASELINE MODEL RESULTS

The baseline model aims to forecast sales of a product in a given city related to the amount of product orders in that city. About the overall significance of the baseline model (F-statistic), all cities belonging to the metropolitan areas analyzed present a p-value less than 0.05, which means that the null hypothesis can be discarded in all situations in which the models were applied. In this model, the cities Simões Filho (BA) and Nerópolis (GO) presented the highest performance compared to the rest of the analyzed sample, with R-square values of 0.994 and 0.986, respectively. The values of the baseline models are contained in Table 6 to Table 15 of Appendix 1.

4.3. ALTERNATIVE MODEL COMPARISONS

Table 2 presents the performance results of the proposed models compared to baseline model performance of metropolitan cities located in the following Brazilian States: Bahia (BA), Espírito Santo (ES), Goiás (GO), Pará (PA), Pernambuco (PE), Minas Gerais (MG), Rio Grande do Sul (RS), Santa Catarina (SC), São Paulo (SP) and Tocantins (TO). The value of R-squared was chosen as the metric to indicate which model offers better prediction.

Regarding alternative model 01, the inclusion of search engine query data as one more independent variable in the model increased its performance for the greatest number of cities. In particular, the city of Osasco (SP) presented the largest increase: the value of R-squared verified in alternative model 01 is 0.908, which is better compared to the value verified in the baseline model (R-squared = 0.658). Some cities did not present an increase in model performance, with the value of R-squared remaining the same as that verified in the base model.

Alternative model 02, which included an adapted version of the Huff model equation to represent the probability of consumers located in a given city of the metropolitan area to go to visit another city to make their purchases, did not present the same good results as found in alternative model 01. However, the results observed in some cities deserve to be highlighted. Senador Canedo (BA) presented the highest value of R-squared in this model (0.964), when compared to the other models. The performance of this model for this city was better than that of the baseline model and alternative model 01. This was also the case for some other cities, for example, Jaboatão dos Guararapes (R-squared=0.925) and Itauna (R-squared=0.837). The highest value of R-squared was found in the capitals Florianópolis (R-squared=0.842), Porto Alegre (R-squared= 0.801) and Palmas (R-squared= 0.983).

Alternative model 03 follows the same logic proposed by the alternative model 02, however, in this model, the attractiveness measure used in the Huff equation (adapted version) is the value of the territorial extension of the city. The predictive power to estimate product sales in each metropolitan city is very similar between alternative model 02 and alternative model 03. In most cities, the value of R-squared is the same (Table 2).

Table 2: Model Comparisons

		Baseline Model	**Alternative Model 01**	**Alternative Model 02**	**Alternative Model 03**
Cities	**State**	**R-squared**	**R-squared**	**R-squared**	**R-squared**
Ananindeua	PA	0.757	0.758	0.935	0.932
Aparecida De Goiania	GO	0.905	0.924	0.966	0.965
Belem	PA	0.945	0.945	0.772	0.774
Belo Horizonte	MG	0.952	0.957	0.753	0.734
Betim	MG	0.743	0.754	0.846	0.839
Cachoeirinha	RS	0.931	0.938	0.727	0.727
Camacari	BA	0.65	0.657	0.967	0.967
Campo Bom	RS	0.72	0.754	0.74	0.74
Canoas	RS	0.956	0.956	0.735	0.735
Cariacica	ES	0.615	0.643	0.712	0.729
Castanhal	PA	0.775	0.785	0.943	0.944
Charqueadas	RS	0.938	0.947	0.74	0.74
Contagem	MG	0.726	0.737	0.957	0.956
Dias D Avila	BA	0.951	0.951	0.964	0.964
Dois Irmaos	RS	0.61	0.622	0.744	0.744
Esteio	RS	0.809	0.811	0.74	0.74
Florianopolis	SC	0.477	0.549	0.842	0.844
Goianapolis	GO	0.961	0.962	0.964	0.963
Goiania	GO	0.963	0.963	0.914	0.911

Table 2: Model Comparisons - Continues

		Baseline Model	Alternative Model 01	Alternative Model 02	Alternative Model 03
Cities	State	R-squared	R-squared	R-squared	R-squared
Gravatai	RS	0.734	0.843	0.74	0.74
Guaiba	RS	0.937	0.967	0.738	0.738
Igrejinha	RS	0.872	0.874	0.739	0.739
Itauna	MG	0.391	0.394	0.837	0.836
Ivoti	RS	0.955	0.958	0.74	0.74
Jaboatao Dos Guararapes	PE	0.492	0.524	0.925	0.924
Lauro De Freitas	BA	0.917	0.917	0.967	0.966
Mairipora	SP	0.919	0.935	0.719	0.714
Maua	SP	0.896	0.902	0.709	0.706
Mogi Das Cruzes	SP	0.921	0.928	0.705	0.704
Montenegro	RS	0.886	0.887	0.74	0.739
Neropolis	GO	0.986	0.988	0.965	0.964
Novo Hamburgo	RS	0.536	0.557	0.746	0.746
Olinda	PE	0.92	0.921	0.825	0.824
Osasco	SP	0.658	0.908	0.796	0.795
Palhoca	SC	0.659	0.675	0.76	0.745
Palmas	TO	0.832	0.836	0.983	0.983
Para De Minas	MG	0.845	0.852	0.837	0.836
Paraiso Do Tocantins	TO	0.985	0.987	0.851	0.84
Paulista	SP	0.929	0.938	0.824	0.824
Porto Alegre	RS	0.736	0.744	0.801	0.802
Porto Nacional	TO	0.958	0.958	0.843	0.849
Recife	PE	0.918	0.923	0.521	0.531
S Bernardo Do Campo	SP	0.867	0.97	0.707	0.706

Table 2: Model Comparisons - Ends

		Baseline Model	**Alternative Model 01**	**Alternative Model 02**	**Alternative Model 03**
Cities	**State**	**R-squared**	**R-squared**	**R-squared**	**R-squared**
Salvador	BA	0.923	0.955	0.989	0.989
Santana De Parnaiba	SP	0.645	0.689	0.73	0.721
Santo Andre	SP	0.843	0.86	0.705	0.704
Sao Caetano Do Sul	SP	0.339	0.342	0.717	0.712
Sao Jose	SC	0.939	0.94	0.612	0.61
Sao Leopoldo	RS	0.898	0.91	0.739	0.738
Sao Paulo	SP	0.772	0.772	0.583	0.585
Sapiranga	RS	0.732	0.764	0.741	0.741
Sapucaia Do Sul	RS	0.937	0.938	0.74	0.739
Senador Canedo	GO	0.387	0.417	0.964	0.964
Serra	ES	0.709	0.719	0.742	0.738
Sete Lagoas	MG	0.982	0.982	0.844	0.838
Simoes Filho	BA	0.994	0.994	0.937	0.936
Suzano	SP	0.945	0.945	0.706	0.704
Taboao Da Serra	SP	0.978	0.978	0.726	0.72
Taquara	RS	0.844	0.849	0.74	0.739
Viamao	RS	0.884	0.903	0.74	0.74
Vila Velha	ES	0.872	0.898	0.706	0.704
Vitoria	ES	0.731	0.749	0.715	0.714

4.4 APPLICATION OF MODELS WITH ANOTHER PARAMETER VALUE

As previously explained, the spatial dynamics of the consumer were calculated through Huff's probability model (Equation 1). The application of this model requires the inclusion of certain elements, and one of these elements is the parameter λ. The first empirical results shown in this dissertation compares the performance of the models using the value two as the exponent value for the parameter λ (Table 3). However, as a way of investigating the possibility of improving the performance of the proposed models, a new value to set parameter λ was tested, contained in Alternative Model 02 and Alternative Model 03. The parameter's value (λ) has been the subject of several studies and different estimation procedures have been proposed, most of them leading to an approximate value of 2 (Douard, Heitz, and Cliquet 2015; Cliquet 1995). However, as highlighted by Li and Liu (2012), studies involving models of spatial interactions present a range of values that can be used to calibrate the Huff equation related to the effects of distance. Therefore, considering this range, the new value for the parameter λ was set 1.5.

The results for Alternative Model 02, which accepts the inclusion of parameters in its Xt variable, showed performance values in the analyzed cities very close to those initial results in which the exponent value had been set at 2 (Table 3). It is worth highlighting the performance of the cities of Cariacica (ES) and Palhoça (SC). By setting the parameters at 1.5, Alternative Model 02 explains 70% of the impact on product sales in the city of Cariacica compared to the previous version, whose parameter had been set at 2 (R square = 0.693). For the city of Palhoça, however, the situation was the opposite: the variance explained by Alternative Model 02 was smaller when using this new exponent value.

As occurred in the Alternative 02 model, the definition of a new value for the Alpha and Beta parameters of the Huff probability equation also presented similar performance for the Alternative 03 model in the analyzed Brazilian cities (Table 3). However, contrary to what happened in the previous model, by setting the value 1.5 as a component element of the variable Xu, the difference in the performance of some cities was more expressive than that verified in the previous model. This is the case, for example, of some state capitals, such as Belém (PA), Recife (PE) and São Paulo (SP). Another highlight worth mentioning is the city of Vilha Velha (ES).

Table 3: Comparing parameters for gravitational model

		Model 2		Model 3	
Cities	State	Adj. R-squared Parameter = 2	Adj. R-squared Parameter = 1.5	Adj. R-squared Parameter = 2	Adj. R-squared Parameter = 1.5
Ananindeua	PA	0.929	0.929	0.925	0.924
Aparecida De Goiania	GO	0.964	0.964	0.963	0.963
Belem	PA	0.75	0.75	0.752	0.749
Belo Horizonte	MG	0.744	0.742	0.724	0.722
Betim	MG	0.84	0.841	0.833	0.832
Cachoeirinha	RS	0.724	0.724	0.724	0.724
Camacari	BA	0.966	0.966	0.966	0.966
Campo Bom	RS	0.737	0.737	0.737	0.737
Canoas	RS	0.732	0.733	0.733	0.733
Cariacica	ES	0.693	0.701	0.712	0.712
Castanhal	PA	0.937	0.938	0.939	0.939
Charqueadas	RS	0.737	0.737	0.737	0.737
Contagem	MG	0.955	0.955	0.954	0.954
Dias D Avila	BA	0.963	0.963	0.963	0.963
Dois Irmaos	RS	0.742	0.742	0.742	0.742
Esteio	RS	0.737	0.737	0.737	0.737

Table 3: Comparing parameters for gravitational model - Continues

		Model 2		**Model 3**	
Cities	**State**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**
Florianopolis	SC	0.826	0.827	0.829	0.830
Goianapolis	GO	0.962	0.962	0.961	0.961
Goiania	GO	0.91	0.91	0.906	0.906
Gravatai	RS	0.738	0.738	0.738	0.738
Guaiba	RS	0.735	0.735	0.735	0.735
Igrejinha	RS	0.737	0.737	0.737	0.737
Itauna	MG	0.831	0.831	0.830	0.830
Ivoti	RS	0.737	0.737	0.737	0.737
Jaboatao Dos Guarara	PE	0.92	0.921	0.920	0.920
Lauro De Freitas	BA	0.965	0.965	0.964	0.964
Mairipora	SP	0.714	0.716	0.709	0.707
Maua	SP	0.704	0.707	0.701	0.701
Mogi Das Cruzes	SP	0.7	0.702	0.699	0.700
Montenegro	RS	0.737	0.737	0.737	0.737
Neropolis	GO	0.963	0.963	0.962	0.962
Novo Hamburgo	RS	0.743	0.743	0.743	0.743
Olinda	PE	0.814	0.814	0.813	0.813

Table 3: Comparing parameters for gravitational model - Continues

		Model 2		Model 3	
Cities	**State**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**
Osasco	SP	0.792	0.793	0.791	0.789
Palhoca	SC	0.737	0.724	0.721	0.719
Palmas	TO	0.981	0.981	0.981	0.981
Para De Minas	MG	0.831	0.832	0.830	0.830
Paraiso Do Tocantins	TO	0.842	0.841	0.830	0.830
Paulista	SP	0.813	0.813	0.813	0.812
Porto Alegre	RS	0.799	0.798	0.80	0.800
Porto Nacional	TO	0.833	0.834	0.840	0.841
Recife	PE	0.491	0.491	0.502	0.508
S Bernardo Do Campo	SP	0.702	0.704	0.701	0.702
Salvador	BA	0.989	0.989	0.988	0.988
Santana De Parnaiba	SP	0.725	0.725	0.716	0.711
Santo Andre	SP	0.7	0.701	0.699	0.699
Sao Caetano Do Sul	SP	0.712	0.713	0.707	0.707
Sao Jose	SC	0.575	0.58	0.572	0.576
Sao Leopoldo	RS	0.736	0.736	0.736	0.736
Sao Paulo	SP	0.576	0.572	0.578	0.572

Table 3: Comparing parameters for gravitational model - Ends

		Model 2		Model 3	
Cities	**State**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**	**Adj. R-squared Parameter = 2**	**Adj. R-squared Parameter = 1.5**
Sapiranga	RS	0.739	0.739	0.739	0.739
Sapucaia Do Sul	RS	0.737	0.737	0.737	0.737
Senador Canedo	GO	0.963	0.962	0.962	0.962
Serra	ES	0.725	0.726	0.722	0.722
Sete Lagoas	MG	0.837	0.837	0.830	0.830
Simoes Filho	BA	0.934	0.934	0.933	0.934
Suzano	SP	0.701	0.703	0.699	0.699
Taboao Da Serra	SP	0.721	0.721	0.715	0.712
Taquara	RS	0.737	0.737	0.737	0.737
Viamao	RS	0.737	0.737	0.737	0.737
Vila Velha	ES	0.688	0.682	0.686	0.699
Vitoria	ES	0.698	0.696	0.696	0.696

4.5 TESTING THE PREDICTIVE POWER OF MODELS WITH MACHINE LEARNING

As a way of investigating the predictive power of the variables analyzed in estimating product sales, it was decided to add a new form of evaluation, through the use of machine learning. In this case, supervised Learning algorithms, such as linear regression and logistic regression are the most appropriate choice when the purpose involves predicting new or future observations (Shmueli, 2010). Machine learning is a subfield of Artificial Intelligence that is very successful in dealing with large and complex datasets (Wedel; Kannan, 2016).

In the case of this dissertation, the sample containing the metropolitan cities was divided into training and test, in the proportion of 80% and 20%, respectively. It was decided to divide the sample only into Models Alternative 01 and Alternative 03, considering that these models can express the main goal of the research. Multiple linear regression is the machine learning algorithm applied to the models. This type of algorithm is usually used when there is the presence of more than one explanatory variable to predict the response of the target variable (Lee *et al*., 2021). The Scikit-learn from the Python library was used in the data samples. As in the previous results, the evaluation of the performance of the selected models after applying the machine learning algorithm was measured using the R square value.

Regarding Alternative Model 01, the new results obtained with the application of the machine learning algorithm showed a worse performance of the predictive power of the search volume in estimating the sales of light switches in the analyzed regions. The application of machine learning in Alternative Model 03, which used the spatial behavior of the composite consumer in the urbanized area as a predictor variable, also presented lower performance results than those initially verified. However, this worst performance in the predictive power was not so accentuated. In some cities, the model's explanatory power values remained high. This is the case, for example, of the metropolitan cities of Aparecida de Goiânia (GO) and Osasco (SP).

Table 4: Models Comparison with Machine learning

		Baseline Model	Model 01	Model 01 with machine learning	Model 02	Model 03	Model 03 with machine learning
Cities	**State**	**R-squared**	**R-squared**	**R-squared**	**R-squared**	**R-squared**	**R-squared**
Ananindeua	PA	0.757	0.758	0.189	0.935	0.932	0.720
Aparecida De Goiania	GO	0.905	0.924	0.414	0.966	0.965	0.920
Belem	PA	0.945	0.945	0.062	0.772	0.774	0.161
Belo Horizonte	MG	0.952	0.957	0.009	0.753	0.734	0.547
Betim	MG	0.743	0.754	0.057	0.846	0.839	0.658
Cachoeirinha	RS	0.931	0.938	0.111	0.727	0.727	0.859
Camacari	BA	0.65	0.657	0.004	0.967	0.967	0.073
Campo Bom	RS	0.72	0.754	0.011	0.74	0.74	0.586
Canoas	RS	0.956	0.956	0.025	0.735	0.735	0.583
Cariacica	ES	0.615	0.643	0.223	0.712	0.729	0.146
Castanhal	PA	0.775	0.785	0.012	0.943	0.944	0.810
Charqueadas	RS	0.938	0.947	0.001	0.74	0.74	0.561
Contagem	MG	0.726	0.737	0.001	0.957	0.956	0.766
Dias D Avila	BA	0.951	0.951	0.021	0.964	0.964	0.173
Dois Irmaos	RS	0.61	0.622	0.011	0.744	0.744	0.582

Table 4: Models Comparison with Machine learning - Continues

Cities	State	Baseline Model R-squared	Model 01 R-squared	Model 01 with machine learning R-squared	Model 02 R-squared	Model 03 R-squared	Model 03 with machine learning R-squared
Esteio	RS	0.809	0.811	0.031	0.74	0.74	0.567
Florianopolis	SC	0.477	0.549	0.017	0.842	0.844	0.103
Goianapolis	GO	0.961	0.962	0.411	0.964	0.963	0.862
Goiania	GO	0.963	0.963	0.01	0.914	0.911	0.255
Gravatai	RS	0.734	0.843	0.059	0.74	0.74	0.578
Guaiba	RS	0.937	0.967	0.003	0.738	0.738	0.577
Igrejinha	RS	0.872	0.874	0.135	0.739	0.739	0.565
Itauna	MG	0.391	0.394	0.011	0.837	0.836	0.346
Ivoti	RS	0.955	0.958	0.116	0.74	0.74	0.571
Jaboatao Dos Guararapes	PE	0.492	0.524	0.05	0.925	0.924	0.676
Lauro De Freitas	BA	0.917	0.917	0.184	0.967	0.966	0.036
Mairipora	SP	0.919	0.935	0.002	0.719	0.714	0.698
Maua	SP	0.896	0.902	0.003	0.709	0.706	0.464
Mogi Das Cruzes	SP	0.921	0.928	0.087	0.705	0.704	0.604
Montenegro	RS	0.886	0.887	0.017	0.74	0.739	0.587
Neropolis	GO	0.986	0.988	0.09	0.965	0.964	0.893

Table 4: Models Comparison with Machine learning - Continues

Cities	State	Baseline Model R-squared	Model 01 R-squared	Model 01 with machine learning R-squared	Model 02 R-squared	Model 03 R-squared	Model 03 with machine learning R-squared
Novo Hamburgo	RS	0.536	0.557	0.02	0.746	0.746	0.572
Olinda	PE	0.92	0.921	0.005	0.825	0.824	0.268
Osasco	SP	0.658	0.908	0.016	0.796	0.795	0.923
Palhoca	SC	0.659	0.675	0.033	0.76	0.745	0.069
Palmas	TO	0.832	0.836	0.002	0.983	0.983	0.188
Para De Minas	MG	0.845	0.852	0.016	0.837	0.836	0.348
Paraiso Do Tocantins	TO	0.985	0.987	0.019	0.851	0.84	0.526
Paulista	SP	0.929	0.938	0.008	0.824	0.824	0.290
Porto Alegre	RS	0.736	0.744	0.081	0.801	0.802	0.161
Porto Nacional	TO	0.958	0.958	0.155	0.843	0.849	0.431
Recife	PE	0.918	0.923	0.039	0.521	0.531	0.061
S Bernardo Do Campo	SP	0.867	0.97	0.000136	0.707	0.706	0.612
Salvador	BA	0.923	0.955	0.007	0.989	0.989	0.683
Santana De Parnaiba	SP	0.645	0.689	0.028	0.73	0.721	0.719
Santo Andre	SP	0.843	0.86	0.181	0.705	0.704	0.429

Table 4: Models Comparison with Machine learning - Ends

		Baseline Model	Model 01	Model 01 with machine learning	Model 02	Model 03	Model 03 with machine learning
Cities	State	R-squared	R-squared	R-squared	R-squared	R-squared	R-squared
Sao Caetano Do Sul	SP	0.339	0.342	0.171	0.717	0.712	0.681
Sao Jose	SC	0.939	0.94	0.057	0.612	0.61	0.013
Sao Leopoldo	RS	0.898	0.91	0.205	0.739	0.738	0.569
Sao Paulo	SP	0.772	0.772	0.098	0.583	0.585	0.017
Sapiranga	RS	0.732	0.764	0.01	0.741	0.741	0.569
Sapucaia Do Sul	RS	0.937	0.938	0.00000852	0.74	0.739	0.570
Senador Canedo	GO	0.387	0.417	0.063	0.964	0.964	0.872
Serra	ES	0.709	0.719	0.001	0.742	0.738	0.208
Sete Lagoas	MG	0.982	0.982	0.046	0.844	0.838	0.631
Simoes Filho	BA	0.994	0.994	0.17	0.937	0.936	0.574
Suzano	SP	0.945	0.945	0.002	0.706	0.704	0.614
Taboao Da Serra	SP	0.978	0.978	0.04	0.726	0.72	0.713
Taquara	RS	0.844	0.849	0.432	0.74	0.739	0.572
Viamao	RS	0.884	0.903	0.083	0.74	0.74	0.568
Vila Velha	ES	0.872	0.898	0.043	0.706	0.704	0.119
Vitoria	ES	0.731	0.749	0.065	0.715	0.714	0.222

4.6 DISCUSSION

The consumer purchase process is complex, because it involves a number of factors that impact the decision (Li *et al.*, 2020; Hamilton *et al.*, 2020; Lemon; Verhoef, 2016). Trying to synthesize this complexity in a sales prediction model that aggregates information about online search behavior and geographic location is not a simple task. Through the development and application of four regression models, the goal of this dissertation was to identify in which situations the use of search engine query data combined with the spatial behavior of the consumer prove to be valid to estimate sales in a certain region.

The alternative model 01 tested the predictive power of search engine query data on sales prediction, considering that the online searches and sales related to the product occurs in the same city. According to previous research, search queries are able to reveal consumers' interests and what they are looking for in some specific moments of the journey, such as their purchase intentions (Liu; Toubia, 2018; Li; Ma, 2020). Then, researching the product ("interruptor") on the search engine platform can be a good indication that this product will have a return on sales for the firm that supplies it in a set physical space.

The results of alternative model 1 showed that the search query data helped to better explain the demand forecast for the searched product (light switch). The model's performance in most cities was greater than the results verified in the baseline model, which only considered product orders as an explanatory variable. This finding further contributes to establishing this type of data (search engine query) as a valid mechanism to apply in forecasting models. Relevant publications from other areas of knowledge already use online consumer behavior to predict events in the so-called "offline world". Lampos and colleagues (2020), for example, used the queries search volume as a way to map the incidence of covid-19 in different countries and Aiken and colleagues (2020) demonstrated that the combination of a model that uses historical epidemiological data with volume of Google search data has a superior predictive power to anticipate outbreaks of diseases in certain countries.

The application of alternative model 01, which incorporated the variable related to google searches, did not lead to an increase in the R-square values when compared to the R-square values of the baseline model, in the case of state capitals. In some capitals, such as Florianópolis (SC) and Porto Alegre (RS), it was possible to observe a variation in values from one model to the other, however, this variation was not as pronounced as seen in other non-capital cities.

One of the reasons that might explain these results for capital cities is the fact that this kind of city has a more complete and robust structure than the other cities, in particular, in terms of accessibility, resources and growth opportunities. These characteristics also influence the consolidation and emergence of many businesses in the area. This multiplicity of retail locations spread across the capitals allows the consumer to have easy access to the products they need. In other words, it is very likely that the consumer will spend more time researching the product on the internet than finding the product physically to buy in a location in the city.

The alternative models 2 and 3 use an adaptation of the Huff probability equation, based on the replacement of the two main elements of the equation: distance and attractiveness. As mentioned in the method section, this replacement was necessary due to the dataset that the author has access to, related to a business-to-business context. When Huff's original model was created, the attractiveness measure was represented by the size of the store and distance was measured from the average time spent by the consumers from their residence to the store that was considered by them the most attractive to make their purchase. Over the years, other studies emerged bringing modifications to the original model, many largely focused on the aspect of the store's attractiveness (Cliquet, 1995; Douard; Heitz; Cliquet, 2015).

In this dissertation, the urbanized area and the territorial extension area of each city were used as attractiveness metrics in alternative models 3 and 4, respectively. Both metrics, provided by a relevant Brazilian public institute, were included by the author due to the greatest internal diversity of the studied area chosen. Brazil is a country whose federative units (States) present a large degree of diversity and these differences are also reflected in the characteristics of the cities themselves. When a city has a higher value for territorial extension, this does not mean that the local area is being occupied homogeneously or that it is heavily populated. For this reason, the territorial extension and the urbanized area were elements incorporated in the different models.

Probabilistic models, such as the Huff model, are more suitable in urban contexts (Cliquet, 2013). Therefore, it was expected that Alternative model 3, whose attractiveness metric was the value of the urbanized area, would present a better performance than alternative model 4, in which the metric was territorial extension area. However, this was not the case. The R-squared values verified in the two models were very close, and in most cities, the value of R-squared was the same.

5 CONCLUSIONS

In this research, I propose a sales prediction model integrating search engine query data with spatial consumer behavior. From a theoretical point of view, this dissertation expands the marketing literature by incorporating the geographic location of the online search in the sales forecast. Variables that refer to the physical space in which the consumer is located are still little explored in the consumer literature. In a bibliographic review carried out by myself, 68 publications used information related to search platforms in the main Marketing journals in the last decade and no forecast models were found that incorporate this type of explanatory variable. Forecasts of real-world events that incorporate online search data with the aim of mapping the geography of the place where these events are likely to occur are more present in research focused on health.

As previously described in the theoretical framework, Huff's gravitational model has undergone several adaptations over time. Although many of them have brought significant results for the evolution of gravitational models and for understanding the spatial dynamics of the consumer, no model has yet been developed in a context considering the cities of metropolitan regions as protagonists of the points of origin and destination of the individuals who belong to them. This new proposal of the Huff model leads to other forms of adaptation of the model which contribute even more to the evolution of the spatial marketing literature. Among the other modifications are the use of new attractiveness metrics, represented in this study by the territorial extension of the city and the urbanized area, both based on public data, collected and made available by Brazilian research institutions (IBGE).

From a methodological viewpoint, the use of more modern geolocation tools that use GPS data, such as google maps, also contributed to updating the model. The information provided by this type of tool allows the researcher to have access to real traffic conditions, road conditions and duration of trips using different means of transport (eg: car, bus, subway, etc.). In addition, it was possible to identify particularities of access to some cities analyzed, such as the use of ferries or the need to use private roads. The use of public data or data from companies that provide free access and, therefore, do not necessarily require direct access to the consumer, which is often costly, time-consuming and invasive, contributed to the model by providing a more impartial and innovative method of mapping the spatial behavior of the consumer.

In managerial terms, the geographic identification of consumers who are interested in the product supplied by the company is a great starting point for managers to be able to draw a map of attraction to expand or even reallocate their commercial activities, optimizing their inventories in each region and managing logistics of resources to meet this demand.

In the initial stages of the research, it was possible to verify that there are certain places in the country that have a large volume of online searches for the product, however, the company does not sell there. In the State of Amazonas (AM), for example, in the metropolitan cities of Manacapuru and Itacoatiara there were online searches for the product in the analyzed period, but the company did not make sales in the same period in these cities. This lack of sales prevented the databases from being merged and, consequently, from being analyzed with more robust data. A similar situation occurred in other metropolitan cities in other states. In São Paulo, an economic hub of great importance for the country, the search engine platform provided online search information for the product in the cities of São Lourenço da Serra, Pirapora do Bom Jesus and Juquitiba, but no sales of the researched product were found in the company dataset.

5.1 LIMITATIONS AND RESEARCH OPPORTUNITIES

The present research brought several contributions to the Marketing literature. However, as it occurs in every scientific study, research has limitations, regardless of how well it is conducted and constructed.

5.1.1 Lack of data available for analysis

The first limitation is related to the impossibility of making some associations between online information searches and product sales. This occurred in geographical regions where the search engine platform did not return search information about the product in the analyzed period (July 2018 to November 2019). This situation occurred in the metropolitan regions of the states of Acre (AC) and Mato Grosso do Sul (MS). This gap in online consumer search information is a fact that is not congruent with the reality of the firm's dataset. Several cities from the metropolitan areas of these Brazilian states placed orders with the company. The lack of data from Google may indicate that people who live in these regions do not usually use online search platforms as a source of information in their purchase journey related to the selected product. Or, if they use this platform, this use is not relevant enough for Google to recognize a substantial volume of searches in that region.

Another situation that occurred due to a gap in the google dataset was when the search engine platform provided information about few metropolitan areas, or when the sample of metropolitan cities presented a difference when compared to the sample of these same cities in the sales dataset. For example, regarding the State of Amazonas (AM), Google returned information for the metropolitan cities of Manacapuru and Itacoatiara. However, these cities did not place orders for the firm in the period that was analyzed. For other geographical areas, such as Rondonia State (RO), there was only data for one city, which was the capital. Limitations of this nature also occurred in the States of Alagoas (AL), Amapá (AP), Roraima (RR), Piauí (PI), Mato Grosso (MT), Sergipe (SE), Rio Grande do Norte (RN), Ceará (CE) and Maranhão (AM).

In both situations described above, the remaining sample of cities was small, and the lack of data made it difficult to forecast sales based on spatial behavior and online consumer searches. In order to calculate the Huff probability, it is important that consumers have, at least, two alternative destinations to visit or make purchases, since the equation shows the store that a consumer is more likely to visit, comparing the attractiveness and accessibility between two destination points. The adaptation proposed in this study considered each metropolitan city as the origin and destination of the model. Therefore, it is important to have a minimum of three cities to define one of them as "origin" and the other two cities as alternative destinations for the consumer to visit.

Future research could select more months and/or years to apply the models and to analyze the data. A more extensive period of time means more search information of online behavior. It is important to highlight that the period used to apply the models was from July 2018 to November 2019, a period that preceded the incidence of the coronavirus pandemic. It would be interesting, therefore, to compare consumer online search behavior and product sales in pre- and post-pandemic periods.

5.1.2 Lack of accuracy about search query location and retail destination

The use of search engine query data brings many opportunities for research. Previous studies have already used this type of data to predict real-world events. However, this type of platform also generates some challenges for the researcher and consequent research limitations. One of these challenges is related to the level of data aggregation. The more aggregated the data is, the more difficult to obtain individualized information. Regarding our context, this lack of individualization means less familiar and personalized consumer information, both in terms of their characteristics and their motivations for individual purchases.

Despite the interesting findings, it was impossible to define exactly who the individuals searching for products were and whether the same individuals who carried out the search also went to the store to purchase the product. Likewise, it was not possible to define exactly where the purchase was to be made. In order to be able to apply the models in a B2B context, the individual points of stores were replaced by cities. This type of adaptation of Huff's model is an important contribution of the dissertation, as it became a viable alternative for sales forecasting purposes.

In the past, researchers who investigated the spatial behavior of consumers through Huff's gravitational model collected consumption information through surveys carried out, normally, in people's homes (Douard, Heitz, and Cliquet 2015; Cliquet 1995). In those surveys, consumers were asked about aspects such as their purchase history and the variables that make them attracted to a store (e.g. promotions, seller competence, product quality, etc.). However, this type of collection method has the disadvantage of being costly and time-consuming. Today, with the advancement of technology and the widespread use of cell phones, it is possible to collect more personalized information from consumers, with the possibility, even, of tracing the consumption journey in real time. Future research can use the combination of search engine data with mobile data to better understand consumer spatial behavior and, therefore, predict

REFERENCES

AGARWAL, A.; HOSANAGAR, K.; SMITH, M. D. Location, Location, location: an analysis of profitability of position in online advertising markets. **Journal of Marketing Research**, s.l., v. 48, n. 6, p. 1057-1073, 2011.

AIKEN, E. L.; MCGOUGH, S. F.; MAJUMBER, M. S.; WACHTEL, G.; NGUYEN, A. T.; VIBOUD, C.; SANTILLANA, M. Real-time estimation of disease activity in emerging outbreaks using internet search information. **PLOS Computational Biology**, s.l., v. 16, n. 8, p. e1008117, 17 ago. 2020.

BERGER, J.; HUMPHREYS, A.; LUDWIG, S.; MOE, W. M.; NETZER, O.; SCHWEIDEL, D. A. Uniting the tribes: using text for marketing insight. **Journal of Marketing**, v. 84, n. 1, p. 1-25, 2020.

BERMAN, O.; DREZNER, T.; DREZNER, Z.; KRASS, D. Modeling competitive facility location problems: New approaches and results. **Decision technologies and applications**, s.l., p. 156-181, set., 2009.

BRADLOW, E. T.; GANGWAR, M.; KOPALLE, P. VOLETI, S. The Role of Big Data and Predictive Analytics in Retailing. **Journal of Retailing**, s.l., v. 93, n. 1, p. 79-95, 1 mar. 2017.

CADAVID, J. P. U.; LAMOURI, S.; GRABOT, B. Trends in machine learning applied to demand & sales forecasting: A review. In: **7th International Conference on Information Systems, Logistics and Supply Chain ILS Conference 2018**, July 8-11, Lyon, France, 2018.

CHOI, H.; VARIAN, H. Predicting initial claims for unemployment benefits. **Google Inc**, v. 1, n. 2009, p. 1-5, 2009.

CLIQUET, G. Implementing a subjective MCI model: An application to the furniture market. **European Journal of Operational Research**, s.l., v. 84, n. 2, p. 279-291, jul., 1995.

CLIQUET, G. **Geomarketing**: Methods and strategies in spatial marketing. s.l.: John Wiley & Sons, 2013.

DOUARD, J.-P.; HEITZ, M.; CLIQUET, G. Retail attraction revisited: From gravitation to purchase flows, a geomarketing application. **Recherche et Applications en Marketing (English Edition)**, v. 30, n. 1, p. 110-129, 2015.

DU, R. Y.; KAMAKURA, W. A. Measuring contagion in the diffusion of consumer packaged goods. **Journal of Marketing Research**, s.l., v. 48, n. 1, p. 28–47, 2011.

FILDES, R; MA, S; KOLASSA, S. Retail forecasting: research and practice. **International Journal of Forecasting**, s.l., v. 38, n. 4, p. 1283-1318, 2022.

GHOSE, A.; IPEIROTIS, P. G.; LI, B. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. **Marketing Science**, v. 31, n. 3, p. 493–520, 2012.

GHOSE, A.; YANG, S. An empirical analysis of search engine advertising: Sponsored search in electronic markets. **Management science**, v. 55, n. 10, p. 1605-1622, 2009.

HAMILTON, R. et al. Traveling with companions: The social customer journey. **Journal of Marketing**, v. 85, n. 1, p. 68-92, 2021.

HAMILTON, R.; PRICE, L. L. Consumer journeys: Developing consumer-based strategy. **Journal of the Academy of Marketing Science**, Springer, v. 47, p. 187-191, 2019.

HU, Y.; DU, R. Y.; DAMANGIR, S. Decomposing the impact of advertising: Augmenting sales with online search data. **Journal of marketing research**, s.l., v. 51, n. 3, p. 300–319, 2014.

HUFF, D. Defining and Estimating a Trading Area. **Journal of Marketing**, s.l., v. 28, n. 3, p. 34-38, jul., 1964.

HUFF, J. O. Distance-decay models of residential search. In: **Spatial Statistics and models**. Dordrecht: Springer Netherlands, 1984. p. 345-366.

HUFF, D. L. Parameter estimation in the Huff model. **Esri, ArcUser**, p. 34–36, 2003.

HUMPHREYS, A.; ISAAC, M. S.; WANG, R. J.-H. Construal matching in online search: applying text analysis to illuminate the consumer decision journey. **Journal of Marketing Research**, v. 58, n. 6, p. 1101–1119, 2021.

JERATH, K. et al. A “position paradox” in sponsored search auctions. **Marketing Science**, s.l., v. 30, n. 4, p. 612–627, 2011.

JEZIORSKI, P.; MOORTHY, S. Advertiser prominence effects in search advertising. **Management science**, s.l., v. 64, n. 3, p. 1365-1383, 2018.

KATONA, Z.; SARVARY, M. The race for sponsored links: Bidding patterns for search advertising. **Marketing Science**, s.l., v. 29, n. 2, p. 199-215, 2010.

KLAPDOR, S. et al. Finding the right words: The influence of keyword characteristics on performance of paid search campaigns. **Journal of Interactive Marketing**, s.l, v. 28, n. 4, p. 285-301, 2014.

LAMPOS, V.; MAJUMDER, M. S.; YOM-TOV, E.; EDELSTEIN, M.; MOURA, S.; HAMADA, Y.; RANGAKA, M. X.; MCKENDRY, R. A.; COX, I. J. Tracking covid-19 using online search. **NPJ digital medicine**, s.l, v. 4, n. 1, p. 17, 2021.

LEMON, K. N.; VERHOEF, P. C. Understanding customer experience throughout the customer journey. **Journal of marketing**, s.l., v. 80, n. 6, p. 69–96, 2016.

LEE, Jungwon et al. A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. **Journal of Theoretical and Applied Electronic Commerce Research**, s.l., v. 16, n. 5, p. 1472-1491, 2021.

LI, H.; MA, L. Charting the path to purchase using topic models. **Journal of Marketing**

Research, s.l., v. 57, n. 6, p. 1019–1036, 2020.

LI, J. et al. Path to purpose? How online customer journeys differ for hedonic versus utilitarian purchases. **Journal of Marketing**, s.l., v. 84, n. 4, p. 127–146, 2020.

LIANG, Y.; GAO, S.; CAI, Y.; FOUTZ, N. Z.; WU, L. Calibrating the dynamic Huff model for business analysis using location big data. **Transactions in GIS**, s.l., v. 24, n. 3, p. 681–703, 2020.

MÉNDEZ-SUÁREZ, M.; MONFORT, A. The amplifying effect of branded queries on advertising in multi-channel retailing. **Journal of Business Research**, s.l., v. 112, p. 254–260, 1 maio 2020.

NEWING, Andy; CLARKE, Graham P.; CLARKE, Martin. Developing and applying a disaggregated retail location model with extended retail demand estimations. **Geographical Analysis**, s.l., v. 47, n. 3, p. 219-239, 2015.

PALLANT, J. I. et al. An empirical analysis of factors that influence retail website visit types. **Journal of Retailing and Consumer Services**, s.l., v. 39, p. 62–70, 1 nov. 2017.

PARK, C. H.; AGARWAL, M. K. The order effect of advertisers on consumer search behavior in sponsored search markets. **Journal of Business Research**, s.l., v. 84, p. 24–33, 1 mar. 2018.

RUTZ, O. J.; BUCKLIN, R. E. From Generic to Branded: A Model of Spillover in Paid Search Advertising. **Journal of Marketing Research**, s.l., v. 48, n. 1, p. 87–102, 2011.

RUTZ, O. J.; BUCKLIN, R. E.; SONNIER, G. P. A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising. **Journal of Marketing Research**, s.l., v. 49, n. 3, p. 306–319, 2012.

SHMUELI, G. To explain or to predict? **Statistical Science**, s.l., v. 25, n. 3, p. 298-310, 2010

SUHARA, Y.; BAHMARI, M.; BOZKAYA, B.; PENTLAND, A. S. Validating Gravity-Based Market Share Models Using Large-Scale Transactional Data. **Big Data**, s.l., v. 9, n. 3, p. 188–202, jun. 2021.

WANG, Y.; JIANG, W.; LIU, S.; YE, X.; WANG, T. Evaluating trade areas using social media data with a calibrated huff model. **ISPRS International Journal of Geo-Information**, s.l., v. 5, n. 7, p. 638868205, jul. 2016.

WEDEL, Michel; KANNAN, P. K. Marketing analytics for data-rich environments. **Journal of marketing**, s.l., v. 80, n. 6, p. 97-121, 2016.

YANG, Y.; LU, O.; TANG, G.; PEI, J. The Impact of Market Competition on Search Advertising. **Journal of Interactive Marketing**, s.l., v. 30, n. C, p. 46–55, 2015.

YAO, S.; MELA, C. F. A Dynamic Model of Sponsored Search Advertising. **Marketing Science**, s.l., v. 30, n. 3, p. 447–468, 2011.

YOGANARASIMHAN, H. Search Personalization Using Machine Learning. **Management Science**, s.l., v. 66, n. 3, p. 1045–1070, mar. 2020.

APÊNDICE 1

Table 5: Correlations

Cities	State	Sales and Orders	Sales and Google Search	Cities	State	Sales and Orders	Sales and Google Search
Ananindeua	PA	0.511	-0.298	Novo Hamburgo	RS	0.669	0.366
Aparecida De Goiania	GO	0.942	-0.17	Olinda	PE	0.883	0.041
Belem	PA	0.588	0.185	Osasco	SP	0.965	-0.185
Belo Horizonte	MG	0.796	-0.004	Palhoca	SC	0.518	-0.251
Betim	MG	0.702	-0.27	Palmas	TO	0.538	0.007
Cachoeirinha	RS	0.868	0.339	Para De Minas	MG	0.647	0.338
Camacari	BA	0.604	-0.14	Paraiso Do Tocantins	TO	0.985	0.211
Campo Bom	RS	0.848	0.643	Paulista	PE	0.884	-0.221
Canoas	RS	0.893	0.233	Porto Alegre	RS	0.479	0.193
Cariacica	ES	0.289	-0.205	Porto Nacional	TO	0.882	-0.502
Castanhal	PA	0.631	0.049	Recife	PE	0.758	-0.15
Charqueadas	RS	0.697	0.076	S Bernardo Do Campo	SP	0.979	-0.293
Contagem	MG	0.506	0.014	Salvador	BA	0.906	-0.076
Dias D Avila	BA	0.917	0.129	Santana De Parnaiba	SP	0.59	-0.269

Table 5: Correlations - Continues

Cities	State	Sales and Orders	Sales and Google Search	Cities	State	Sales and Orders	Sales and Google Search
Esteio	RS	0.626	-0.157	Sao Caetano Do Sul	SP	0.334	-0.074
Florianopolis	SC	0.503	0.152	Sao Jose	SC	0.921	0.099
Goianapolis	GO	0.949	0.283	Sao Leopoldo	RS	0.825	-0.162
Goiania	GO	0.606	-0.089	Sao Paulo	SP	0.265	-0.256
Gravatai	RS	0.899	0.198	Sapiranga	RS	0.794	0.417
Guaiba	RS	0.92	0.236	Sapucaia Do Sul	RS	0.909	0.015
Igrejinha	RS	0.656	0.193	Senador Canedo	GO	0.434	-0.288
Itauna	MG	0.271	-0.096	Serra	ES	0.499	-0.118
Ivoti	RS	0.929	0.351	Sete Lagoas	MG	0.966	0.196
Jaboatao Dos Guarara	PE	0.54	-0.155	Simoes Filho	BA	0.939	-0.407
Lauro De Freitas	BA	0.587	0.037	Suzano	SP	0.804	0.028
Mairipora	SP	0.794	-0.299	Taboao Da Serra	SP	0.928	0.306
Maua	SP	0.895	-0.196	Taquara	RS	0.826	0.314
Mogi Das Cruzes	SP	0.815	-0.34	Viamao	RS	0.837	0.26

Table 5: Correlations - Ends

Cities	State	Sales and Orders	Sales and Google Search	Cities	State	Sales and Orders	Sales and Google Search
Montenegro	RS	0.837	0.283	Vila Velha	ES	0.882	-0.181
Neropolis	GO	0.981	0.05	Vitoria	ES	0.613	0.221

Table 6: Baseline Model for Bahia (BA)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Camacari	0.992	0.991	1872.828	p < 0.001	4.386	0.101	43.276
Dias D Avila	0.969	0.967	497.799	p < 0.001	3.994	0.179	22.311
Lauro De Freitas	0.996	0.996	4471.322	p < 0.001	4.088	0.061	66.868
Salvador	0.997	0.996	4810.894	p < 0.001	4.241	0.061	69.361
Simoes Filho	0.997	0.997	5592.297	p < 0.001	2.775	0.037	74.782

Table 7: Baseline Model for Espírito Santo (ES)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Cariacica	0.99	0.989	1536.584	p < 0.001	3.437	0.088	39.199
Serra	0.983	0.982	942.479	p < 0.001	2.794	0.091	30.7
Vila Velha	0.974	0.972	595.086	p < 0.001	3.204	0.131	24.394
Vitoria	0.985	0.984	1043.711	p < 0.001	3.435	0.106	32.307

Table 8: Baseline Model for Baseline Model (GO)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Aparecida De Goiania	0.99	0.989	1588.273	p < 0.001	2.541	0.064	39.853
Senador Canedo	0.964	0.961	422.368	p < 0.001	3.374	0.164	20.552
Neropolis	0.982	0.981	887.488	p < 0.001	3.682	0.124	29.791
Goiania	0.997	0.997	5798.074	p < 0.001	2.854	0.037	76.145
Goianapolis	0.985	0.984	1080.275	p < 0.001	4.147	0.126	32.868

Table 9: Baseline Model for Minas Gerais (MG)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Belo Horizonte	0.995	0.995	3296.511	p < 0.001	2.921	0.051	57.415
Sete Lagoas	0.989	0.988	1383.797	p < 0.001	2.647	0.071	37.199
Para De Minas	0.945	0.941	273.642	p < 0.001	2.802	0.169	16.542
Itauna	0.997	0.997	6133.521	p < 0.001	2.859	0.037	78.317
Contagem	0.999	0.999	23706.671	p < 0.001	3.431	0.022	153.97
Betim	0.995	0.994	2978.078	p < 0.001	3.489	0.064	54.572

Table 10: Baseline Model for Pará (PA)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Ananindeua	0.984	0.983	1009.048	p < 0.001	4.844	0.152	31.766
Belem	0.996	0.996	4171.081	p < 0.001	3.595	0.056	64.584
Castanhal	0.998	0.998	8305.125	p < 0.001	4.733	0.052	91.132

Table 11: Baseline Model for Pernambuco (PE)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Jaboatao Dos Guarara	0.993	0.992	2127.227	p < 0.001	4.826	0.105	46.122
Recife	0.993	0.993	2291.708	p < 0.001	4.106	0.086	47.872
Paulista	0.998	0.998	10353.549	p < 0.001	4.611	0.045	101.752
Olinda	0.998	0.998	8048.307	p < 0.001	4.569	0.051	89.712

Table 12: Baseline Model for Rio Grande do Sul (RS)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Cachoeirinha	0.99	0.99	1644.886	$p < 0.001$	3.072	0.076	40.557
Sao Leopoldo	0.992	0.991	1981.773	$p < 0.001$	3.071	0.069	44.517
Gravatai	0.992	0.992	2108.844	$p < 0.001$	4.828	0.105	45.922
Campo Bom	0.989	0.988	1453.46	$p < 0.001$	3.233	0.085	38.124
Montenegro	0.964	0.962	426.804	$p < 0.001$	3.52	0.17	20.659
Esteio	0.979	0.978	751.254	$p < 0.001$	3.688	0.135	27.409
Sapiranga	0.995	0.994	3024.029	$p < 0.001$	5.473	0.1	54.991
Novo Hamburgo	0.985	0.984	1071.978	$p < 0.001$	3.128	0.096	32.741
Igrejinha	0.981	0.98	839.447	$p < 0.001$	3.146	0.109	28.973
Ivoti	0.985	0.984	1032.409	$p < 0.001$	3.318	0.103	32.131
Guaiba	0.988	0.987	1337.184	$p < 0.001$	3.365	0.092	36.568
Canoas	0.997	0.996	4679.72	$p < 0.001$	4.627	0.068	68.408
Porto Alegre	0.993	0.993	2380.686	$p < 0.001$	4.109	0.084	48.792
Charqueadas	0.966	0.964	459.282	$p < 0.001$	5.31	0.248	21.431
Dois Irmaos	0.994	0.994	2645.793	$p < 0.001$	4.871	0.095	51.437
Sapucaia Do Sul	0.956	0.953	346.932	$p < 0.001$	3.885	0.209	18.626
Taquara	0.972	0.97	558.769	$p < 0.001$	4.166	0.176	23.638

Viamao	0.994	0.994	2603.783	p < 0.001	4.951	0.097	51.027

Table 13: Baseline Model for Santa Catarina (SC)

Cities	**R-squared**	**Adj. R-squared**	**F-statistic**	**Prob (F-statistic)**	**Coef (orders)**	**Std error**	**t**
Florianopolis	0.996	0.996	3847.831	p < 0.001	4.19	0.068	62.031
Palhoca	0.952	0.949	318.081	p < 0.001	1.487	0.083	17.835
Sao Jose	0.992	0.991	1883.154	p < 0.001	4.735	0.109	43.395

Table 14: Baseline Model for Tocantins (TO)

Cities	**R-squared**	**Adj. R-squared**	**F-statistic**	**Prob (F-statistic)**	**Coef (orders)**	**Std error**	**t**
Palmas	0.992	0.991	3581.046	p < 0.001	4.297	0.072	59.842
Porto Nacional	0.987	0.986	1215.36	p < 0.001	4.803	0.138	34.862
Paraiso Do Tocantins	0.99	0.99	1658.902	p < 0.001	4.268	0.105	40.73

Table 15: Baseline Model for São Paulo (SP)

Cities	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Coef (orders)	Std error	t
Mairipora	0.996	0.995	3635.22	p < 0.001	3.416	0.057	60.293
Taboao Da Serra	0.985	0.984	1077.522	p < 0.001	3.366	0.103	32.826
Suzano	0.979	0.978	745.983	p < 0.001	2.968	0.109	27.313
Sao Paulo	0.997	0.997	5521.81	p < 0.001	3.016	0.041	74.309
Sao Caetano Do Sul	0.988	0.988	1354.603	p < 0.001	2.117	0.058	36.805
S Bernardo Do Campo	0.996	0.996	4334.572	p < 0.001	3.394	0.052	65.837
Santo Andre	0.957	0.954	357.19	p < 0.001	4.581	0.242	18.899
Santana De Parnaiba	0.961	0.959	394.846	p < 0.001	4.539	0.228	19.871
Osasco	0.994	0.993	2445.608	p < 0.001	3.56	0.072	49.453
Mogi Das Cruzes	0.969	0.967	501.043	p < 0.001	2.517	0.112	22.384
Maua	0.971	0.969	528.133	p < 0.001	3.044	0.132	22.981