UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
DEPARTAMENTO DE CIÊNCIAS ADMINISTRATIVAS
PROGRAMA DE PÓS GRADUAÇÃO EM ADMINISTRAÇÃO

CARLA FREITAS SILVEIRA NETTO

FORECAST RECONCILIATION: METHODS, STRUCTURES, CRITERIA, AND A NEW APPROACH WITH SPATIAL DATA

Porto Alegre
2019

CARLA FREITAS SILVEIRA NETTO

FORECAST RECONCILIATION: METHODS, STRUCTURES, CRITERIA, AND A NEW APPROACH WITH SPATIAL 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

2019

CIP - Catalogação na Publicação

Silveira Netto, Carla Freitas
FORECAST RECONCILIATION: METHODS, STRUCTURES, CRITERIA, AND A NEW APPROACH WITH SPATIAL DATA / Carla Freitas Silveira Netto. -- 2019.
132 f.
Orientador: Vinícius Andrade Brei.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Escola de Administração, Programa de Pós-Graduação em Administração, Porto Alegre, BR-RS, 2019.

1. Time-series forecasting. 2. Gravitational models. 3. Reconciliation approaches. 4. Deep learning. I. Brei, Vinícius Andrade, orient. II. Título.

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

CARLA FREITAS SILVEIRA NETTO

FORECAST RECONCILIATION: METHODS, STRUCTURES, CRITERIA, AND A NEW APPROACH WITH SPATIAL 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.

PhD dissertation presented and approved on the 14th of August, 2019.

Examination board:

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

Prof. Dr. Luiz Antonio Slongo
UFRGS - Brazil

Prof. Dr. José Afonso Mazzon
USP - Brazil

Prof. Dr. Burçin Bozkaya
Sabanci University - Turkey

This work is dedicated to my mom. She taught me to face the world and not be ashamed of what I still do not know. She used to say that anything done with drive and love can and will be learnt.

ACKNOWLEDGEMENTS

First, I acknowledge the PhD itself. This, of course, means UFRGS, my CNPq scholarship, Monash University, Chair Tramontina Eletrik, and all people symbolized by those names. The universities' staff, professors, and business people. You are all amazing for letting me focus on my research and providing me with the resources for that.

Ancora Imparo (I am still learning) is the motto of Monash, and it is also mine. We never stop learning and this PhD dissertation does not end here. This represents only the current state of a story that has a beginning but does not have a due date (well, has a formal one. But not in spirit). It is bigger than this document than the years of the PhD and, I hope, bigger than me.

This story begins with the passing of my mom when I wished for a new direction in my life. Reviewed my priorities. Revisited a dream. At orientation week I remember how scared I got when I realized the commitment and the time it would take. If I only knew it would take what seems now a lifetime. And it would be the time of my life.

I met amazing people: Simoni Rohden, Natalia Englert, Marina Lugoch, Valentina Ubal, Tiago Carpenedo, Rodrigo Heldt. My dear colleagues that EMA in Belo Horizonte brought together and consolidated a friendship and partnership that, I hope, last forever. Many papers, posters, and presentations in conferences came from these partnerships. But also, some beers, pizzas, coffees, cakes, laughs, support, and hope. You enrich my cv and my life, guys. And it is just the beginning.

Thanks to Maria Alice for her support on the data set preparation and mostly for our friendship during the PhD.

The research groups. Thanks, GPMC. I learnt a lot, got a lot of support in moments I needed, and had the opportunity to contribute as well. Thanks also to GEMS and the cake meetings. Last but definitely not least, thanks to the market potential group: Mohsen Bahrami, Selim Balcisoy, Burçin Bozkaya, and Alex 'Sandy' Pentland.

329. Mixed feelings for a room that teaches about life. For good and bad.

I thank the PhD also for the trips, the conferences, the friends I met from all over the world. So many new stories to tell. I will never say thanks enough to the PhD for allowing me to do what I love the most: go to new places, meet people and learn with them. Thanks, especially, for the bowling masters. Thanks to this amazing group for all the support during the

ISF conferences (see you in Rio). Thanks to Oliver Schaer for your feedback and valuable insights on how to frame my contributions during the 2019 Marketing Conference.

Thanks for the opportunity to review for conferences and publications such as the International Journal of Forecasting. Thanks for the reviews I received as well, and the publications still to be derived from that.

Thanks for Australia, Melbourne. I felt at home as soon as I saw this land from the airplane window. How hard it was to leave. How hard it still is.

Thanks to Rob J. Hyndman for accepting me as a visiting Ph.D., for being patient with me, teaching me in a few months what would take me so much more time to learn by myself. Thanks for laughing with me of my mistakes, and for including me in everything, making me feel like a part of your team for a while. Thanks for giving me the opportunity to meet all the NUMBATS.

Thanks also to Yuejun Zhao, Madeleine Margaret, Lina Zhang, Steff Kobakian, Puwasala Gamakumara, Di Cook, Alex Cooper (my brother and teammate at Melbourne Datathon, on which we got 2nd place on the insights competition), and my “kids” Sophie New and Gabi Martins. Your support during my days in Melbourne was very important to me.

Thanks to Slongo, Mazzon, and Burçin for motivation and inspiration during this process. I am a fan. I admire your histories, personalities, wisdom. Thanks to Brei for a patience that can only match the one my dad has with me. Thanks for the partnership, for accepting with me every single challenge. For showing me the way when I was too anxious to see. For the motivation when I was almost losing hope or not finding a reason to keep going. Thanks for trying so hard to make me an optimist (and almost succeeding).

Thanks to my dad, for friendship and constant motivation. For being the happiest, proudest, person with my progress, and with every conference I had to attend. Thanks for experiencing all this with me, from the enrolment process to the very last moments. Thanks for raising me, with my mom, to be self-sufficient, free.

Thanks to me, for overcoming my personal challenges. Thanks for facing your fears, your limitations. For trying even when it seemed just a bit more than you could take. For giving yourself a chance, and allowing yourself to make mistakes, learn. You get out of the PhD not only with more knowledge about marketing. You leave cooking better, taking better care of yourself, knowing how to set boundaries, giving yourself value and respect. You leave a doctor. Not only in marketing, but of yourself. Congrats.

Thanks to God for allowing me to reach the end of this with health and (some) sanity, and for the opportunity to have this chapter in my biography.

ABSTRACT

This PhD dissertation is a collection of four papers that aim to explore, in the marketing field, the research on hierarchical and grouped time-series reconciliation approaches. Those approaches are necessary when different departments of an organization have different needs regarding forecast aggregations. This work focuses, besides reconciliation approaches, on time-series forecasting methods, and on the importance of geographical information to better forecast and plan marketing strategies. The first paper is theoretical and argues on the importance to marketing of having accurate forecasts. It explores the current state of marketing research on modelling in general, and on forecast specifically. It covers the classifications of methods, datasets explored on current research, the basic model studied, and existing gaps. The paper concludes that marketing focuses on explanation, living a gap on accuracy evidence and on the applicability of the models proposed. The second paper explores those gaps by applying two current topics of discussion on forecasting time-series literature: machine learning techniques and ensemble models. These methods are easy to implement and are reported in the literature to improve accuracy. The paper proposes an adaptation to portfolio optimization to calculate the weights of an ensemble based on each base model's accuracy and the covariance matrix of such accuracies. The proposed approach outperforms all 15 base models and the equal weights benchmark. The paper also provides evidence that, if single models are compared, statistical methods have better accuracy than the machine learning methods applied. The third paper uses a statistical method to forecast time-series (i.e. sales) combined with different structure and criteria of aggregation. The aim of the paper is to compare different criteria based on marketing mix variables. The empirical application presented in the paper indicates whether product category, channel type or region (geographic location) works best alone or combined. It also gives evidence of the importance of geographical considerations to improve forecast accuracy. The last paper furthers explore this finding by proposing a new reconciliation approach that distributes an aggregate forecast to lower levels of disaggregation using a gravitational model. This paper also contributes to the literature by comparing statistical, machine learning and deep learning methods (LSTM). All papers presented in this dissertation use open-source tools, combining proprietary data that is natural to the process of every organization and publicly available data. The focus is on methods and tools that are generalizable to all types of goods, can be easily applied by any organization, with relatively low investment. The contributions of the PhD dissertation are (1) to compare statistical, machine learning and deep learning methods to forecast sales on single and ensemble models; (2) to provide evidence on the criteria and structure of aggregation that improves forecast accuracy the most; and (3) to offer a new approach to distribute an aggregate forecast to new geographical regions when no historical data is available.

Keywords: Time-series forecasting. Gravitational models. Reconciliation approaches. Deep learning.

RESUMO

A presente tese de doutorado é uma coleção de quatro artigos científicos desenvolvidos com o objetivo de explorar, dentro da área de marketing, a pesquisa sobre reconciliação de previsão de séries temporais com estrutura hierárquica ou agrupada. Reconciliação de previsões é necessária quando diferentes áreas de uma organização necessitam de previsões em diferentes níveis de agregação. O presente conjunto de estudos foca, além da reconciliação de previsões, em métodos de previsão de series temporais e na importância de informações geográficas para melhor prever e planejar estratégias de marketing. O primeiro artigo apresentado é uma revisão da literatura atual em modelagem de marketing, focando nos estudos sobre previsão. O artigo argumenta sobre a importância para o marketing em ter previsões, nas diferentes classificações dos métodos, nos tipos de dados usados, no modelo básico estudado e nos potenciais para estudos futuros. O artigo conclui que marketing precisa de estudos que evidenciem acurácia e sejam fáceis de implementar na prática. O segundo artigo procurar preencher essas lacunas aplicando técnicas de *machine learning* e *ensemble*. Essas técnicas são discutidas atualmente na teoria sobre previsão de séries temporais por prometerem facilidade de aplicação e melhoria em acurácia. O artigo propõe uma adaptação da otimização de portfólio como estratégia para calcular os pesos dos diferentes modelos que compõe um ensemble. A proposta do artigo tem melhor acurácia, no teste realizado, que os 15 modelos únicos (estatísticos e de *machine learning*) e o *ensemble* usando pesos iguais para todos os modelos. O artigo contribui também para a discussão sobre *machine learning* para previsão de séries temporais, demonstrando, nesse caso, a superioridade dos modelos estatísticos. O terceiro artigo usa um método estatístico combinado com diferentes estruturas e critérios de agregação para prever séries temporais (vendas). O objetivo do artigo é comparar diferentes critérios baseados em variáveis de marketing. A aplicação empírica dá indícios de que informações sobre a localização das vendas aumenta a acurácia das previsões. O último artigo explora esse achado ao propor uma estratégia alternativa de reconciliação de previsões. Essa estratégia distribui uma previsão feita em um nível agregado para níveis desagregados usando um modelo gravitacional. O artigo também contribui para a literatura ao comparar métodos estatísticos e de *machine learning* com *long short-term memory* (LSTM), um método de *deep learning*. Todos os artigos usam ferramentas *open-source* e combinam dados públicos com dados proprietários que resultam naturalmente dos processos de qualquer organização. O foco dos estudos são métodos e ferramentas generalizáveis para todos os segmentos que possam ser facilmente implementados por qualquer empresa, com relativamente baixos investimentos. As contribuições dessa tese de doutorado são (1) comparar métodos estatísticos, de *machine learning* e *deep learning* para prever vendas em modelos únicos e combinados (*ensemble*); (2) prover evidências sobre os critérios e estruturas de agregação que melhoram a acurácia das previsões; e (3) oferecer uma nova estratégia para distribuir uma previsão agregada em novas regiões geográficas quando dados históricos não estão disponíveis.

Palavras-chave: Previsão de vendas. Séries de tempo. Modelo gravitacional. Reconciliação de previsões. *Deep learning*.

TABLE OF CONTENTS

1 INTRODUCTION

The papers presented in this PhD dissertation, combined, provide an approach to (1) more accurately forecast sales (statistical, machine learning and deep learning methods are compared); (2) select from the available criteria the one (or the set) that improves forecast accuracy the most when used to structure a forecasting system; and (3) distribute an aggregate forecast to new geographical regions when no historical data is available.

They also provide tools for the visualization of the results, in maps, helping to decide easily to allocate efforts in the most profitable geographical regions. According to Wedel and Kannan (2016) the development of dashboards – visualization tools allowing users to explore and analyze further the results – is one opportunity for future research in marketing.

However, before I proceed with the goals and contributions of this study, some distinctions between concepts need to be made. Forecasting is the estimation of certain value for a different or future situation (ARMSTRONG, 2001). Marketing responses include reactions to marketing instruments that can be translated into demand, market share, market potential, brand choice or sales, for example. This dissertation focus on the forecast of one of these responses, sales.

Demand, according to Armstrong (2001), is the need for a particular item and can be forecasted by each level of a distribution channel. Demand can be related to a category, product or brand (THE MASB COMMON LANGUAGE PROJECT, [*s.d.*]). The American Marketing Association dictionary defines demand as the number of customers willing to buy a product or service for a certain price, in a certain market, at a certain period of time. Another definition available on the same dictionary is the units of a product sold in a market in a certain period of time (AMA DICTIONARY, [*s.d.*]).

Mazzon and Toledo (1983) define demand by eight variables: a total volume (share, amount or monetary value); the way the product or service is obtained; if the product is new or pre-existing; a group of customers; a geographical region; period of time; macroeconomic and market variables; and marketing strategies (company and competitors).

The concept of market potential is associated with the level of saturation or the maximum number of individuals in a population that will eventually adopt an innovation (ARMSTRONG, 2001). However, this concept is not only connected to innovations, and can be also defined as the maximum total sales that can be obtained for a certain product (ARMSTRONG, 2001). According to AMA, the market potential is an estimation of the maximum volume of sales for a whole category in a certain period (AMA DICTIONARY, [*s.d.*]).

Finally, sales are related to the ability of the organization to satisfy a certain demand in a certain period. Sales forecasting is the prediction, projection or estimate of the volume of sales expected for a future period (AMA DICTIONARY, [*s.d.*]). Armstrong (2001) warns that it is not always possible to say that data regarding demand is equivalent to data regarding sales. Demand may not result in sales if the item is out of stock, for example.

This distinction between sales and demand is not only important in theory. According to Seaman (2018), to forecast sales or demand is a decision also based on the goal of the organization. The author argues that, for pricing, sales forecast information, price elasticity and competition information are enough. However, to set production and channel distribution, a demand forecast is also necessary. The author differentiates further these two needs of forecasting. For pricing, only aggregate, short-term, forecasts are sufficient. For distribution, the need to forecast is long-term and disaggregate. This implies that organizations have different needs of forecast aggregation, depending on the strategy that is going to be planned and for what purpose the forecast will be used as an input.

The reality of many organizations is of limited access to data, especially from competitors. This makes demand forecasting infeasible. Most companies still struggle to collect, store and analyze internal data, having to overcome challenges such as problems on data imputation, missing data, software integration, lack of channel collaboration to share information, security issues, to name a few. This is not, however, what can be concluded when reviewing the modelling literature. Academia has access to market data and makes the assumption that most companies have the same access.

Marketing literature builds choice models that will help companies allocate budget on marketing instruments. However, for these companies to implement such models, they will need to expend resources on data, acquiring it from Nielsen (or other companies that provide market data), if they work on a segment that has data collected by these companies. If not, they will need to spend even more resources (time and monetary) to collect data on their own. They will need to conduct intention surveys, ask employees to collect data from stores or find other sources and strategies to collect all information necessary.

Even after they have access to their own and competitors' data, they still will need to be able to estimate complex, computationally inefficient (slow to run) models, with no evidence of improved accuracy. For that reason, forecasting practice applies mostly judgmental and univariate models. The use of econometric methods is limited (FILDES; MA; KOLASSA, 2018) and so is marketing's contribution to practice, since models lack realism and are hard to implement (MEYER; HUTCHINSON, 2016; REISS, 2011). Another issue for practice is the limited computational resource, practitioners need to run algorithms for multiple stock-keeping units for several points of sale and models that take hours to run are not convenient.

In practice, the total sales of a firm and its competitors are considered demand, coinciding with AMA's definition. Fildes, Ma, and Kolassa (2018) state that sales data is not an observation of true demand when demand exceeds the available inventory. In this case, demand estimates would have a negative bias if only sales data is used. Sales data also do not portray latent demand without product availability information. Zero sales can have more than one interpretation, signaling either stock-outs or intermittent demand.

This definition includes demand still not supplied, overlapping with the concept of market potential. There is no theoretical distinction between the concepts unless the apparently incomplete AMA's demand definition is considered. Demand is the sum of the demand supplied, not supplied and latent. Market potential is the total of individuals interested in acquiring, in a certain period, a product or service.

In conclusion, a sales forecast is to estimate future sales of one particular firm; demand forecast is to estimate sales for a market (firm and its competitors); and, the market potential forecast is to estimate the total number of individuals that might be interested in a product or service, including demand not attended and latent.

According to Seaman (2018), to plan channel strategies would be necessary to forecast what I just defined as market potential (and that the author defines as demand). In the author's opinion, a strategy based solely on sales history could lead to lower stock levels than its potential sales. However, as mentioned before, knowledge of the real potential is not feasible for many organizations. Besides that, even Seaman (2018) states that, for some sectors such as convenience goods, sales and demand forecasts should be closely related, resulting in the same estimates.

Most organizations have different points of sale or business partners spread in different geographical locations. These organizations evaluate constantly to keep or close businesses in certain locations, and to open new businesses in others. To make all these decisions, they need an estimation of future sales. They need to know where their products and services will be needed and evaluate their profitability.

For a regional manager, this may be necessary at a state or city level. For a corporate manager, a total estimate (i.e. country) might be enough. However, those forecasts need to be coherent. All levels of hierarchy in an organization, all different departments, need to work with the same forecast. For that reason, there is a need for a unified, coherent system of forecasting that can be used with different levels of aggregation.

Marketing literature, so far, did not give as much attention to time-series data or to forecasting (DEKIMPE; HANSSENS, 2000; BEAL; WILSON, 2015) as to the discrete choice models. The focus of the field has been to build models to explain how marketing actions can influence market response, but not on how accurate or easy to implement in practice these models are. The field also focuses on disaggregate analysis using datasets from companies that have access to individual-level information (package goods companies with access to scanner data). This leaves unattended a large number of organizations that do not work in those sectors or do not have the resources to acquire data from providers, such as Nielsen, for example.

Despite the fact that these companies still struggle to gather, store and analyze data necessary to make better predictions, they still need to forecast with less uncertainty. Examples are small businesses and B2B companies that have small data and do not use science to base their decisions. They do not have access to data resources, do not invest in infrastructure to store and analyze data, and do not have organized associations that can provide data. At the same time, these are the companies that generate most of the jobs in the world.

On the other hand, companies that have access to data at an individual level and even big data, are being pressured by privacy regulations such as the general data protection regulation (GDPR). Some other companies are protecting consumer privacy as a strategy. These scenarios lead to increased emphasis on data minimization and data anonymization. And, lastly, some companies do not have access to online information, a great resource of data. For example, house painting or electrical components, products that are not bought or discussed online.

These companies have limited access to data resources, nonetheless, they face the same problems as a consequence of forecasting errors. Excess or lack of products on stores can impact on marketing metrics, such as imposing changes in pricing strategies, damaging brand reputation, or reducing quality perception, customer loyalty, repurchase intentions, and satisfaction.

There are at least three reasons that forecasting is important to marketing: to identify market opportunities; plan and allocate marketing efforts; and to control and evaluate the return of those efforts (MAZZON; TOLEDO, 1983). Forecasting is also important to marketing because it helps to plan and control all matters related to reaching customers more efficiently than the competition. It includes the selection of regions, new distribution channels, setting sales quotas, allocation of promotional efforts to certain products or regions, or the decision to invest in a new market (MAZZON; TOLEDO, 1983).

Combined with the lack of access to possible explanatory variables and the importance of forecasting to companies (and marketing), researchers have now access to new sources of time-series data. These data are more frequently available since dynamic data is the outcome of social network interactions, mobile phone use, online shopping and many other technology-based interactions with organizations. If before marketing had mostly access to market surveys to build models, now the field can access time-series data that are publicly available. There is also an increased capability of affordable data storage that makes it possible to access and analyze more information than before.

Time-series is also easily available as historical sales data is a natural outcome of the processes of all organizations. Metrics of the volume of use are recorded ever since the first operation of an organization is run. Time series data also gives researchers more information than the mere volume in time. It has information about seasonality, cycle, and trends, to name a few. It also allows predicting future behavior, using well established statistical methods and new state of the art methods such as deep learning algorithms.

Time-series are usually applied in models that are either univariate or that add explanatory variables as covariates. However, these models do not consider the structure of the company and the different needs that departments or hierarchical levels have. Marketing, for example, needs forecasts by product or by brand. If the marketing department run forecasts at these levels and share values with another department that estimates a total aggregated forecast, these numbers will not be the same. The problem is that the data naturally add up, following the hierarchical structure (from disaggregate to aggregate levels), but the forecasts may not. This can cause confusion when firms use the forecasts to plan their actions. For that reason, forecasts estimated in different levels of aggregation (hierarchical) or divided into different groups (grouped) must be reconciled to become "coherent", adding-up, like the data.

Different aggregations possibilities can be extracted from a standard dataset, using the natural structure of the company. From historical sales, forecasts can be estimated using different variables that will act as criteria, giving structure to these aggregations. The total sales of the company in a city can be divided by zones or by the district, for example. The sales can also be aggregated for the whole category of products, and then divided by different product types.

All these arguments and examples point to the necessity of marketing to work with aggregated data and distribute this to other levels of interest. Time-series forecasting literature works with this problem proposing different hierarchical and grouped time-series reconciliation approaches. This is the literature this PhD dissertation introduces to marketing and contributes to.

Regarding the methods applied in the papers that constitute this PhD dissertation, it does not focus on qualitative or judgmental methods. This class of methods include Delphi technique, salespeople opinion, and other employees estimates based on experience. These techniques are developed in fields such as operational research, that explores the combination of quantitative methods with judgmental adjustments (FILDES et al., 2008). However, those adjustments have been proven to be equally common in practice and susceptible to biases (FILDES; GOODWIN, 2007; FILDES; GOODWIN; ÖNKAL, 2019).

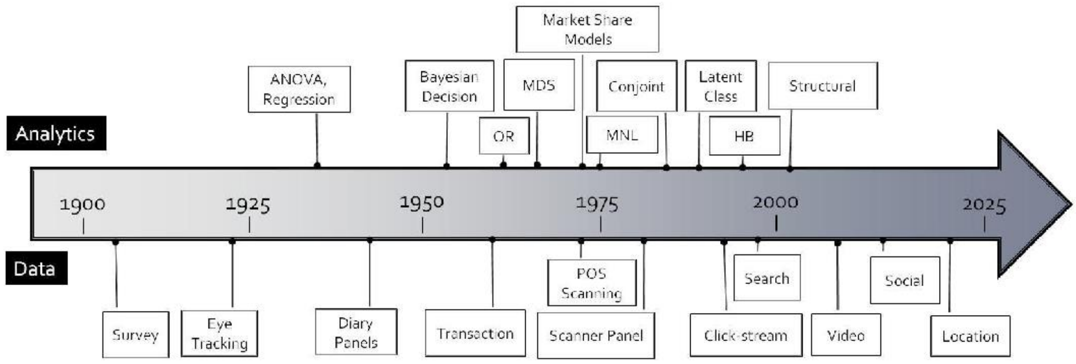
As a result of the scenario presented, I focused on exploring different quantitative methods to forecast sales considering different structures and aggregation criteria, based on marketing variables. I also explore different methods to forecast time-series, comparing accuracy performance of statistical, machine learning and deep learning methods.

The papers that constitute this dissertation also propose new criteria to choose forecast methods (chapter 2). It complements the literature on marketing models' classifications (ARMSTRONG, 2001; CHINTAGUNTA; NAIR, 2011; GENTRY; CALANTONE; CUI, 2006; ROBERTS, 1998), by providing criteria that combines the classifications based on forecasting and explanation goals. Another paper provides an adaptation to an approach to set the weights of ensemble models (chapter 3). It contributes to marketing literature by focusing on forecasting and time-series (DEKIMPE; HANSSENS, 2000; BEAL; WILSON, 2015), and to ensemble models literature by providing evidence that equal weights can be outperformed (COUSSEMENT; DE BOCK, 2013; DE BOCK; COUSSEMENT; VAN DEN POEL, 2010; LEMMENS; CROUX, 2006). Finally, it is provided a new reconciliation approach (chapter 5), contributing on the literature of reconciliation approaches (HYNDMAN *et al.*, 2011), and to agglomeration theory (BRONNENBERG; SISMEIRO, 2002; HUFF, 1964; LIU; STEENKAMP; ZHANG, 2018). The focus is on forecasting, time-series, and on the importance of geographical information to better forecast and plan marketing strategies (chapters 2, 4 and 5).

Geographical information despite being nowadays more available to marketing researchers and practitioners was not yet fully explored. Privacy concerns and regulation regarding individual-level data might be discouraging its use. This concern is valid since with only four spatial-temporal points researchers were able to uniquely identify 95% of the individuals in a dataset (DE MONTJOYE *et al.*, 2013). For that reason, Wedel and Kannan (2016) affirm that less individual data will be available and marketing will need to work with data in a more aggregated level. A challenge is, therefore, to develop tools to work with this type of data preserving anonymity, developing models and algorithms in an aggregated level, without losing predictive power.

Figure 1 shows Wedel and Kannan's timeline of data and methods applied by marketing research where is clear that location data is still a challenge and an opportunity for future research. It is also clear that the timeline misses showing important methods such as time-series and machine learning or deep learning techniques.

Figure 1 – marketing's timeline of data and methods



Source: Wedel and Kannan (2016)

Machine learning and deep learning techniques are still more popular in practice than in marketing research (CHINTAGUNTA; HANSSENS; HAUSER, 2016; CUI; CURRY, 2005; WEDEL; KANNAN, 2016). According to Wedel and Kannan (2016), this happens because marketing prefers methods that can provide support to marketing variables, which is not evidenced by these methods. Despite this, the application of these techniques is still an opportunity for future research (CHINTAGUNTA; HANSSENS; HAUSER, 2016; FILDES *et al.*, 2008; WEDEL; KANNAN, 2016). The possibility to use computer science methods, such as machine learning and deep learning, is important not only for their ability to handle greater volume of data but also for their performance and easier implementation, allowing for automated solutions (WEDEL; KANNAN, 2016).

One of the goals of the PhD dissertation, in line with what Fildes, Ma, and Kolassa (2018) ask, is to propose how to develop automatic scalable models that are robust to data limitations common in organization's operations. The authors also point out that issues of data hierarchies, where research solutions exist, have seen limited implementation. This PhD dissertation is an answer to this request of further implementation of hierarchies and it is also a contribution to this research stream. To achieve that, I (along with co-authors) answer different research questions, in each chapter of the dissertation (each a different paper), summarized in Table 1.

Table 1 – Research questions, chapters, methods and papers

Research questions	Chapter	Methods applied	Related papers and their current status
What is the current state of the art in marketing research regarding forecast methods? Which are the criteria proposed by literature to choose the most appropriate forecast method?	2	Theoretical paper	A previous version was presented at the 37th International Symposium on Forecasting (ISF) in 2017 and EnANPAD 2019. The current version is under review at a special issue of RAUSP. Co-author: Vinicius A. Brei
If ensembles are known to improve accuracy, is there a better way to set the weights of an ensemble than equal weights? Do ensemble and machine learning techniques have better accuracy than statistical methods?	3	Statistical methods: Additive nonlinear autoregressive model; ARIMA; Exponential smoothing; Linear Auto-Regressive model; Logistic Smooth Transition Autoregressive; Naïve and seasonal naïve; Random walk with drift; STL decomposition; TBATS; and Theta. Machine learning techniques: KNN and neural networks. Ensemble: equal weights and optimization approaches.	A previous version of this paper was presented at the 38th International Symposium on Forecasting (ISF), 2018. The current version, presented in the third chapter has substantial changes that were made by the author of this PhD dissertation since the presentation. Co-authors: Rodrigo Heldt and Cleo Silveira
Which marketing variables used alone or combined to the structure of a forecast system can improve accuracy?	4	Exponential smoothing with hierarchical and grouped structures. Different combinations of product, channels and geographical considerations were compared to create those structures. The reconciliation approach used was optimal, using MINT and WLS weights.	Paper presented at 39th International Symposium on Forecasting (ISF), 41st Annual ISMS Marketing Science Conference, 2019, and EnANPAD 2019. Co-authors: Rob J. Hyndman and Vinicius A. Brei

Since geographical considerations are important, could a gravitational model be used to reconcile a hierarchical time-series forecast? Will it have at least the same performance as other approaches even if it does not use historical data?	5	Deep learning method (LSTM) with a hierarchical structure based on geographical considerations. The reconciliation approaches used were optimal, top-down, and the new proposed approach, GSFR.	- A *working paper* by the author of this PhD dissertation and researchers from MIT. Co-authors: Vinicius Brei, Burçin Bozkaya, Mohsen Bahrami, Selim Balcisoy e Alex 'Sandy' Pentland.

Source: The author (2019)

Another goal and contribution of this dissertation are that all analyses are coded using open-source tools, and combine data natural to the process of every organization with publicly available data. Our goal is that all analyses proposed can be easily applied by organizations, despite its sector or size, with almost no additional investments.

Costumers location and information concerning how much is going to be sold or demanded in a certain geographical region are relevant information for retailers and manufactories that need to distribute their products and marketing efforts. Marketing research has explored location in many different aspects, however, not the predictive power of this variable. Our goal is to compare and contribute to forecasting methods, using data that is easily available, considering geographical regions and the problem of selecting and investing in profitable partners and regions.

The remainder of this dissertation is composed of a theoretical paper, on chapter 2, and 3 empirical papers exploring each contribution described in this introductory section. The last chapter is an overall conclusion regarding the findings of the papers, the limitations of each study and the possibilities of future research.

2 FORECASTING IN MARKETING: METHODS, TYPES OF DATA, AND FUTURE RESEARCH

Abstract: Forecasts are fundamental to plan and deliver products and services. Despite such relevance, marketers have difficulty to choose which forecast method is the best for their organizations. One possible explanation for this is that marketing literature focuses on the explanation of which variables impact on marketing responses rather than the accuracy of predictions of those responses. Consequently, the literature is not clear about forecasting methods' classifications, approaches, complexity, requirements, and efficiency. This theoretical paper tries to improve this scenario, reviewing the state of the art about forecasting in marketing. More specifically, we focus on: different classifications and approaches used to study marketing responses, especially the ones based on statistics/mathematics and computer-intensive methods; the types of data used; and suggestions of future research aimed at improving forecasting marketing response. Besides simpler, easier to implement models, further research is necessary to develop forecasting techniques that give evidence of accuracy, uses aggregated or anonymized data, or that incorporates publicly available data. Of foremost importance are datasets that manufacturers of durable goods can use in models for their businesses, the exploration of location data, and the combination of different models and datasets.

Keywords: forecasting, marketing, analytics, theoretical paper

2.1 Introduction

Marketing responses comprise the consequences of consumers' actions in reaction to companies' marketing mix strategies. These responses can be measured in different ways, such as sales, brand choice, demand, or market share. To be able to predict next purchases is a valuable thing to marketing more than for other fields in social sciences (CHINTAGUNTA; NAIR, 2011). However, the knowledge and the ability to forecast is not a skill of most marketing majors (BEAL; WILSON, 2015), and selecting the most appropriate forecasting technique is challenging. The choice of the method is usually based on familiarity and not on what is more appropriate to the market under investigation or the data available (CANITZ, 2016).

According to classical forecasting literature, the first criteria to select a method is related to the amount of objective data (ARMSTRONG, 2001). This will define if it is to follow a qualitative/judgmental approach or a quantitative one. Regarding quantitative methods, Singh (2016) divides the research on forecasting into four types: behavioral-focused (judgmental adjustments to statistical forecasts); business performance focused (impact of forecasting practices on performance); statistics/mathematics-focused (time-series and causal); and big-data-based or computer-intensive (the newest research stream).

In this theoretical paper, we consider that judgments or domain knowledge should be used to add structure to forecasting models, but not to override the estimations after they are done. Domain knowledge improves forecast accuracy and reduces the need to do such adjustments (CHASE, 2013), and it can and should come from economic and marketing literature. Business performance is not analyzed either since the goal is not to discuss the advantages or difficulties to implement the process of forecasting in companies. For those reasons, in the remaining of this theoretical paper, the focus will be on statistics/mathematics and computer-intensive techniques.

In this paper, we describe the classification of methods and types of data used in marketing response models. Their relevance increased with the growing access to data from different contexts, which allowed the development of many different types of models, for different purposes and with different properties (CHINTAGUNTA; NAIR, 2011). The manuscript unfolds as follows. First, we discuss the classification of marketing response methods, followed by a review of the techniques applied in practice and theory. This review is divided into two approaches, statistics/mathematics, and computer-intensive techniques. After that, we introduce the types of data used in marketing models. Finally, we present a methodological framework based on the literature reviewed, proposing some criteria to select forecasting methods. We also offer propositions of future research on forecasting.

2.2 Classifications of marketing response methods

The types of marketing response analysis can be divided based on their goals (CHINTAGUNTA; NAIR, 2011), which can be forecasting, measurement, and testing. Chintagunta and Nair (2011) subdivide such goals based on their respective models: descriptive models (for stable environments), structural models, and reduced-form causal effects. Descriptive models are the ones which focus on forecasting sales across time, based on variables that are available today (e.g. current marketing mix variables and sales). The emphasis is not on causality, given that these models cannot test theories about consumer or firm behavior. They can at the most show an econometric representation of the theory that may serve as the basis for such a causal test (REISS, 2011).

Structural models, on the other hand, use the theory to predict phenomena. These models combine theory and econometric specification to explain patterns in the data (CHINTAGUNTA; NAIR, 2011). They combine marketing models of behavior with statistical assumptions to derive empirical models that can be estimated (REISS, 2011). Discrete choice models are examples of structural models. They are among the most popular in marketing given that much of the data available consists of records of consumers making choices from a set of alternatives within a category. They can be causally interpreted but have limitations as well: (1) it is challenging to find the best combination of theory, data, and econometric specification; (2) they are time consuming; (3) it is difficult to build simple models that are realistic and can be estimated with the data available; (4) they usually present results that may be overly affected by strong assumptions; and (5) assumptions of distributions are made for computational convenience and most times do not have economic defense (CHINTAGUNTA; NAIR, 2011; REISS, 2011).

Finally, reduced-form causal models used for measurement and testing are diverse from structure models because they require fewer assumptions on distribution and specification (CHINTAGUNTA; NAIR, 2011). However, they share some similarities, since both structural and reduced-form models imply causality and require theory.

Another way to classify models is by demand systems. Using this criterion, they can be divided into characteristics space and in product space. Demand in characteristics space assumes that consumers choose products by groups of characteristics. These models are flexible and usually outperform the models of the product space system. However, an issue is the assumption that consumers choose no more than one good (NEVO, 2011). Demand in product space, on the other hand, considers that consumers first decide by categories, then by segments, and finally by brands. Therefore, products, not characteristics, are grouped into these models. Product space systems are simpler to estimate, generally using linear methods, which save computational time. On the other hand, the products need to be classified into segments that are frequently hard to justify. Product space also assumes that consumers buy a number of products of all brands, when they, in reality, may consume more than one brand, but not all of them (NEVO, 2011).

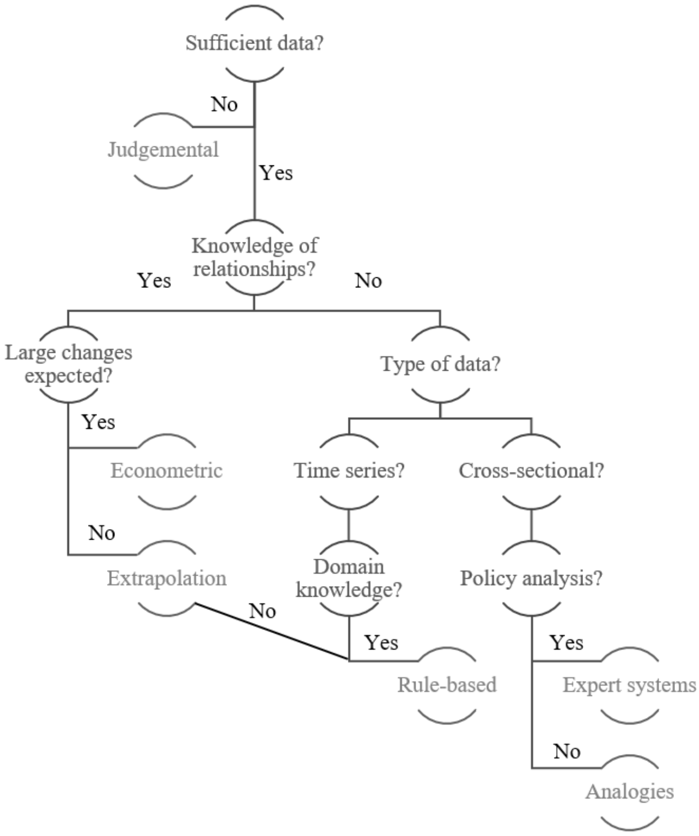
One common disadvantage of both systems of models is that they are static and for many markets the demand is dynamic. This means that they do not consider the possibility of consumers’ decisions in the present affecting the posterior decision, nor that the present decision is affected by expectations of the future (NEVO, 2011).

Another classification typology suggested by Roberts (1998) divides models by level (individual or aggregate) and by application (new or existing products). Models for new products at the individual level are used to predict or explain market share. Such models apply discrete choice analysis According to Roberts (1998), the focus of new products models has been on the aggregate-level through diffusion models. Since pre-launch forecasts are challenging, these studies are frequently done post-launch aimed at understanding the reasons that made the diffusion possible.

Post-launch models also apply discrete choice models at the individual level. The individual-level data comes from scanner data that is frequently used to analyze consumer preference and response to marketing instruments (ROBERTS, 1998). At the aggregate level, marketing has focused on the study of advertising effects and other marketing mix variables on sales (ROBERTS, 1998).

The classifications reviewed in this paper so far are broad, classifying methods that are applied with an aim to explain and to forecast. Armstrong's (2001) classification, on the other hand, focus on methods that have a forecasting goal and divides them based on knowledge of the relationships, type of data (time-series or cross-sectional), expectations of large changes, and domain knowledge. Armstrong (2001) also recommends the testing of different methods and, if they provide useful forecasts, the combination of those results. If they do not provide good results, one should simply use the best performing method. Figure 2 summarizes the techniques in a decision-tree format. However, the criteria proposed by the author misses to account the different goal of the techniques and does not include machine learning techniques.

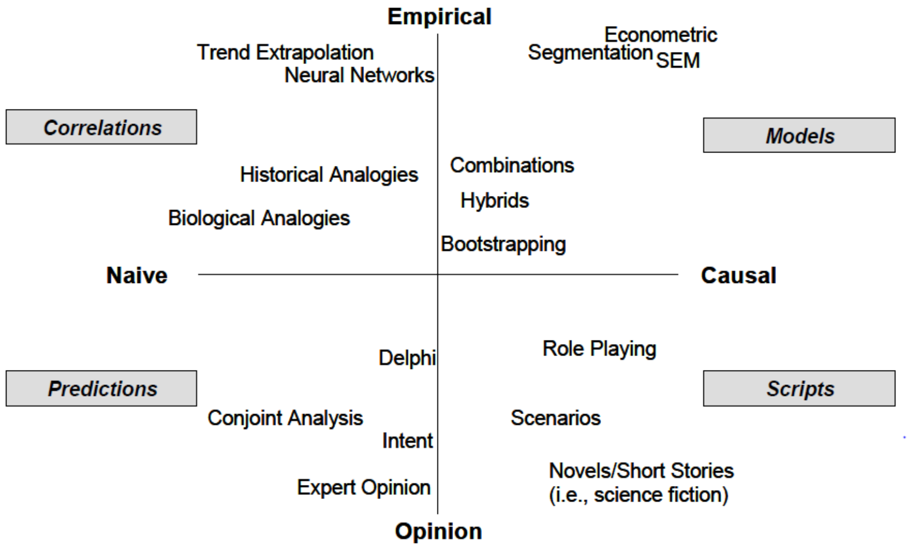
Figure 2. Criteria to select forecasting techniques.



Source: Adapted from Armstrong (2001)

Gentry, Calantone, and Cui (2006) suggest a different typology (Figure 3). According to the authors, although there are many different classifications, none is concise, exclusive, and exhaustive. They classify forecasting methods along two continuums: from casual to naïve, and from opinion based to empirical. This classification creates 4 different categories of techniques: (1) predictions that are based on opinions that do not have explicit assumptions; (2) scripts or scenarios based on casual assumptions; (3) correlations or techniques that give back predictions based on the performance of another factor, with no casual assumptions; and (4) models that return predictions based on casual assumptions.

Figure 3. Typology of methods



Source: Gentry *et al.* (2006)

Extrapolation methods and neural networks (just one of the machine learning techniques available) are classified by the authors as correlations. This can create confusion regarding the definitions of those techniques that apply more than correlations to return predictions. Extrapolation methods are easy to implement, highly accurate techniques that may (or may not) consider seasonality, trends, and cycles (ARMSTRONG, 2001). Gentry *et al.* (2006) justify this definition by the lack of explicit causal assumptions, however, the extrapolation methods do not lack an assumption. The assumption is of stability, or that the variables will continue to behave as they did in the past. The authors lastly call these methods “black boxes”, because one cannot interpret parameters the same way as in an econometric model. This statement misses informing that some machine learning techniques return the importance of variables, for example, giving the researcher more than the prediction and accuracy rate (which is not less than other models, just different in its goal).

2.3 Statistics/mathematics-focused methods

Statistics/mathematics-focused forecasting techniques used in practice can be divided into three streams: time-series (also called extrapolation methods), causal, and weighted combined forecasting (CHASE, 2013). Weighted combined forecasting combines methods (i.e., time series, causal, and/or judgmental) and creates a single forecast by giving each result an equal or different weight. This combination outperforms most single forecasts since biases among methods will compensate one another (CHASE, 2013).

Time-series techniques identify patterns (trend, seasonality, cyclical, and randomness) and make predictions for the future. They have higher predictive accuracy in stable markets. Some examples of time-series are moving average, simple exponential smoothing, Holt's two parameters, and Winters' three parameters. Time-series are simple to develop and require a minimal amount of data, however, they are unable to predict sudden changes in demand. ARIMA (Autoregressive integrated moving average) models are a more advanced class of time-series technique that combine regression elements. These modes are more accurate in long term predictions. ARIMA can model trend/cycle, seasonality, as well as other factors influencing demand or sales (explanatory variables) but require more data and are more complex to develop (CHASE, 2013).

Causal techniques assume that future sales are related to changes in other variables (price, promotions, among others). Examples are regression and ARIMAX (an extension of ARIMA). These techniques require more data and are more complicated to develop (CHASE, 2013). On the other hand, they can include intervention variables (using dummy variables).

Practitioners still practice simpler forecasting methods, favoring forecast accuracy than the ability to explain which variables impact the increase or decline of marketing response. One reason is that the models that result from marketing research are hard to implement in practice. This is a significant challenge for structural models, because "the large number of observations in practice, the large array of state and control variables, and the frequency of decision making can render the application of structural models infeasible" (MELA, 2011). Another reason for the choice of time-series models by practitioners is the availability of data. Most companies do not have access to detailed individual consumer information, nor market data regarding sales of competitors. In these contexts, it is very challenging to apply current marketing response models.

In marketing, structural models are the most used approach, using econometric techniques that focus on explanation, not on forecasting. There are also some models that apply a reduced form (e.g. BRIESCH; DILLON; FOX, 2013; CHUNG; DERDENGER; SRINIVASAN, 2013) or a combination of reduced form and structural model (e.g. CHUNG; DERDENGER; SRINIVASAN, 2013). Some studies apply structural models with a Bayesian approach (ARIBARG; ARORA; KANG, 2010; ARORA; HENDERSON; LIU, 2011; CHE; CHEN; CHEN, 2012; CHUNG; RAO, 2012; FEIT *et al.*, 2013; ROODERKERK; VAN HEERDE; BIJMOLT, 2011; ZHAO *et al.*, 2013; ZHAO; ZHAO; HELSEN, 2011).

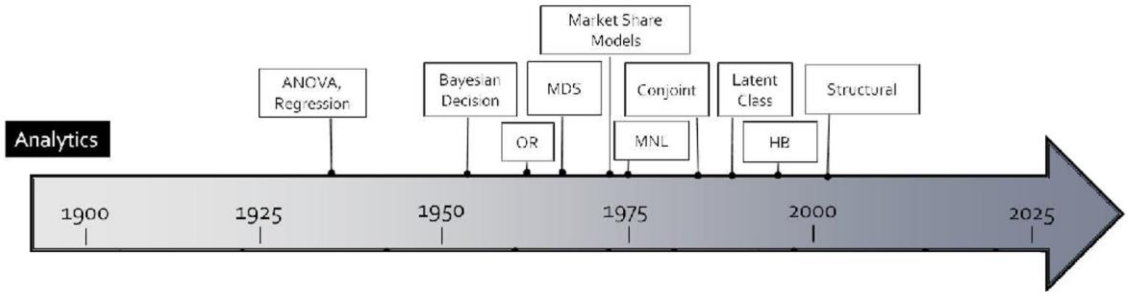
For that reason, another important distinction must be made between the use of classical and Bayesian statistics. This is important particularly in situations with limited information (ROSSI; ALLENBY, 2003). Bayesian statistics are commonly used in marketing, partially due to computing developments that have made it accessible (ALLENBY; BAKKEN; ROSSI, 2004). For example, Markov Chain Monte Carlo (MCMC) simulation made it easier to estimate complex models of behavior that would not be possible with other methods (ALLENBY; BAKKEN; ROSSI, 2004).

The Bayesian approach has some advantages, such as: it is able to reflect heterogeneity in consumer preferences; the developed models are more realistic; it allows disaggregate analysis; the Hierarchical Bayes methods have predictive superiority due to avoiding the restrictive analytic assumptions that alternative methods impose; it allows studies of high-dimensional data and complex relationships; and instead of point estimate of values for each respondent, it ends up with a distribution of estimates for each respondent (for a more comprehensive review see Allenby *et al.*, 2004; Rossi and Allenby, 2003). This distribution can be informative about uncertainty, but on the other hand, makes the analysis more complex.

One important difference between classical and Bayesian statistics is that the former says nothing about how to incorporate different sources of data, such as expert's information and other data sets (ROSSI; ALLENBY, 2003). That is important because merging information acquired from different data sets is challenging for marketing practitioners (ROSSI; ALLENBY, 2003).

Descriptive models can also be applied in marketing, however, the field does not have as many published papers that applied these methods, when compared to the other types of methods discussed in this paper (DEKIMPE; HANSSENS, 2000). Wedel and Kannan (2016) review the techniques applied in marketing (Figure 4) and ignore time-series models or machine learning techniques. They concentrate their analysis on econometric models explaining choices (and for segmentation, that is beyond the scope of this manuscript).

Figure 4. Techniques applied in marketing.



Source: Adapted from Wedel and Kannan (2016)[1]

Marketing has not given much attention to data-driven approaches as time-series (DEKIMPE; HANSSENS, 2000). This may have a connection with a lack of skilled researchers to apply these techniques (BEAL; WILSON, 2015). The same logic can explain the lack of publications on marketing journals using machine learning to forecast marketing responses. This is a symptom of the lack of emphasis in analytics skill on business, especially marketing academic curriculum. Therefore, the ability to work with big data is not common among marketing researchers (FEIT *et al.*, 2013), and marketing should strive to gain knowledge in computer science. Computer-intensive methods (i.e. machine learning techniques) will be discussed in the next section.

2.4 Computer-intensive methods

Methods applied to study marketing response are related to the characteristics of the variables and data sets that became available to marketing scientists. As scanner data enabled marketing to apply structured causal models from fields like transportation science and economics, now big data makes it possible to apply methods from machine learning (CHINTAGUNTA; HANSSENS; HAUSER, 2016).

Big-data is related to new types of variables, and the size of data sets used, but also to the different methods applied. With bigger data sets and a higher number of different attributes applied in marketing studies, analysis using conventional statistical methods became impractical or even infeasible because of computer constraints (FILDES *et al.*, 2008).

[1] On Figure 4, MDS stands for "multidimensional scaling"; OR, for "operations research"; MNL, for "multinomial logit"; and HB, for "Hierarchical Bayesian".

Computer-intensive methods apply statistical and machine learning algorithms for discovering valid, novel and potentially useful predictive information from large data sets in unstructured problems. Statistics is the intellectual base of these applications, given that finding useful patterns from data for prediction has long been a statistical challenge (FILDES *et al.*, 2008).

Most studies applying machine learning focus on segmenting or classifying customers, such as CRM related topics (up/cross-selling, churn analysis, credit scoring). One example of this application is the study of Sundsøy *et al.* (2014) that compared different machine learning algorithms to the judgment of the marketing team in identifying customers that were more likely to buy mobile internet service.

Ali *et al.* (2009) and Sun *et al.* (2008) are two examples of sales forecasting studies that used machine learning techniques. Although these studies are related to the marketing area, they were not published in marketing journals. Ali *et al.* (2009) forecast sales of a grocery store, in the presence of promotions. The authors compared different methods of forecasting, such as exponential smoothing, stepwise linear regression, support vector regression (the regression version of support vector machines), and CART (regression trees). They found that, for periods with no information about marketing instruments (in this case, the promotions), time series techniques performed better. In this case, machine learning techniques only matched the performance of other models (ALI *et al.*, 2009). The reason is that time-series techniques perform well if sales are stable, which corroborate to one of the advantages of these techniques mentioned previously. For periods with promotions, on the other hand, regression trees performed better. However, regression trees are a more complex technique that demand more intense data manipulation. Ali *et al.* (2009) propose a combination of these techniques in a forecasting system, applying time-series for periods when there is no promotion and regression tree model in periods with promotions.

Sun *et al.* (2008) applied a neural network technique called extreme learning machine to forecast sales in fashion retailing. The advantages of that technique are the higher generalization performance and that it avoids difficulties of other learning methods (stopping criteria and long computation time, for example). The disadvantage is that its results are different from time to time because the input weights and hidden biases are randomly chosen (SUN *et al.*, 2008).

More recently Liu, Singh, and Srinivasan (2016) used big data from different sources (Twitter, Google Search, IMDB reviews, among others) to predict TV shows ratings. Cloud computing was applied, so it would be possible to process a large amount of data and to prepare the unstructured data was used text mining (machine learning technique). Alternative machine learning models were tested, but they were outperformed by the structured model chosen by the authors.

Although some efforts were already made in marketing to apply computer-intensive methods on marketing response forecasting, there is still space for evolution (CUI; CURRY, 2005; WEDEL; KANNAN, 2016). To do so marketing researchers need to learn about (or associate with) other disciplines, like data science and machine learning (CHINTAGUNTA; HANSSENS; HAUSER, 2016).

Finally, is important to keep in mind that the method should be chosen after the research problem is defined and the data to solve is accessible. These two criteria should determine the model and not the opposite (REISS, 2011). The next section will review the types of data commonly used in marketing literature.

2.5 Types of data used

Choosing the data is related to the theory and the research problem wished to address. Researchers study the theory, the market, and develop a model based on assumptions. From this model, researchers know what data they need to gather. Mela (2011) outlines a process to data procurement for structural models that consist in determining the data necessary to the research problem, finding the source (or sources) of data and then negotiate, acquire, and check the data. Research problems can also be derived from access to new types of data that were not available before. New types of data can bring new opportunities to apply marketing theory.

In forecasting practice, according to Chase (2013), the most common data sources used are customer orders, customer shipments (or replenishment), points of sales data, promotions, price, Consumer Price Index, gross national product growth, and Consumer Confidence Index. The marketing literature, on the other hand, focuses on marketing instruments, such as promotions, price, and their relation to sales or market share. The basic model studied in marketing, according to Fildes *et al.* (2008), can be represented as:

$$\textit{Marketing response}_{ijt} = f\,(\textit{marketing instruments};\ \text{seasonality};\ \text{exogenous factors}) \qquad (1)$$

where i = brand or SKU, j = store, t = time, *Marketing response* = sales or market share, and *marketing instruments* = price, promotion, competition, among others.

Marketing theory has many potential explanatory variables to put on this basic model (FILDES *et al.*, 2008). Researches should caution that a greater number of variables in a model can explain better, but make the model lose forecasting performance (CHASE, 2013). For that reason, it is recommended to keep the model specification simple. That is not a preoccupation in marketing models since marketing has historically focused on explanation and not on forecasting performance. Consequently, models are very complex, hard to apply in practice, and there is still a lack of evidence concerning the impact of marketing variables to increase accuracy (FILDES *et al.*, 2008; FILDES; MA; KOLASSA, 2018).

Models that include marketing variables are econometrically endogenous. Therefore, the correction of such endogeneity is one of the contributions that authors seek. One way to address this issue is to use instrumental variables. But they are hard to choose, to justify, and do not correct fully the problem. Instrumental variables correct the intercept endogeneity, but not the slope endogeneity. This weakness has not been observed by marketing researchers (LUAN; SUDHIR, 2010). Luan and Sudhir (2010) define slope endogeneity as "private information possessed by managers about the heterogeneous effects of marketing-mix variables on sales".

Reiss (2011) explains an experimental approach in marketing responses studies to overcome problems with endogeneity bias. There is a bias when omitted variables (unobserved factors) are correlated with the error and one or more independent variables in a regression (ROSSI, 2014). The idea is to run experiments manipulating one variable (price, for example) independently of the others (e.g.; promotions), controlling this issue. This approach includes lab experiments, randomized field experiments, natural experiments, and instrumental variables (IV) methods.

Marketing researches prefer the more realistic field experiments that have external validity but is more difficult to control for confounds (REISS, 2011). Natural experiments are still rare in marketing but are commonly applied in economics (CHEN; WANG; XIE, 2011). These types of experiments investigate the effects of variables not in control of the researches, such as government intervention or policy changes, for example (CHEN; WANG; XIE, 2011; REISS, 2011). Instrumental variables are observables variables that are correlated with X variable but are not a part of the structural equation, not affecting Y (ROSSI, 2014). Instrumental Variables and experiments do not solve this problem without support from theory (REISS, 2011). The researcher should not use invalid or week IVs. If there is concern about an unobserved variable, it should be measured (ROSSI, 2014).

Another way to correct endogeneity is “to address the codetermination is to impose restrictions from an assumed model of supply […] into the demand estimation step” (CHINTAGUNTA; NAIR, 2011). Even if this is one way to address the issue, the authors consider that models in marketing literature are better at explaining demand than supply data and propose this as an area for future research. They also alert that supply-side models are harder to estimate with the current data and computing power (CHINTAGUNTA; NAIR, 2011).

In marketing literature, there is also an emphasis on disaggregate analysis. This is due to the fact that the area has access to consumer panels linked with data on marketing instruments (CHINTAGUNTA; NAIR, 2011). This means that the models in marketing deal with censored or truncated data (many zeros in the data). Also, it means that models on marketing are linked to economic theory, structural work, with emphasis on heterogeneity across consumers and considers that products are also differentiated (CHINTAGUNTA; NAIR, 2011).

Mela (2011) divide the types of data by firm property data, free public data, and commercially available market research. Commercially available data, such as scanner data, is often expensive, however, is much quicker to obtain and easier to use (MELA, 2011). Firm property data is of great importance not only because it is rich but also because it is rare to have access. Also having access to people directly involved in the decision making on the firms can bring insights into the “rules of the game” (MELA, 2011). Those are important to specify the model. Public data, such as word of mouth on social media or census data, and commercially available data share the disadvantage of being less customized to the research problem.

Some of the variables applied to the basic model (1) that can be found in the marketing literature are: price, discounts, sales, market share, demographics (income, household size), attributes of products and services, competition, periods (of stockouts, of release of products, or of product-harm crisis), promotion (advertising expenditures, earned media, celebrity endorsements), geographical distance (based on zip-codes), number of alternatives, number of points of sales, word-of-mouth data (volume, variance and valence of posts from experts, peers, and critics), among others.

As mentioned before, methods applied in marketing response studies are related to the characteristics of the variables and data sets that became available to marketing scientists. As developments are still occurring in big data, allowing access to structured and unstructured data, many recent studies apply it to study marketing response.

The bigger influx of information, though, is for those businesses that are already online. The connection between online and offline behavior is still a challenge (LIU; SINGH; SRINIVASAN, 2016), but there are some studies that attempt to find this relation. Chevalier and Mayzlin (2006) and Zhao *et al.* (2013) developed models to predict book sales with data from online book reviews. Dewan and Ramaprasad (2014), and Dhar and Chang (2009) studied the effect of blog posts and social network on sales of these music's singles and albums. Xiong and Bharadwaj (2014) used prerelease buzz to forecast video games sales. Godes and Mayzlin (2004) and Liu, Singh, and Srinivasan (2016) investigate the effect of user-generated content on TV ratings. Lastly, another relation investigated by authors is between word of mouth and sales in the movie industry (CHINTAGUNTA; GOPINATH; VENKATARAMAN, 2010; DELLAROCAS; ZHANG; AWAD, 2007; GOPINATH; CHINTAGUNTA; VENKATARAMAN, 2013; KARNIOUCHINA, 2011; LIU, 2006; MOON; BERGEY; IACOBUCCI, 2010; ONISHI; MANCHANDA, 2012).

In the next section, we propose a methodological framework based on the literature reviewed and develop a set of propositions of future research to contribute to the forecasting research.

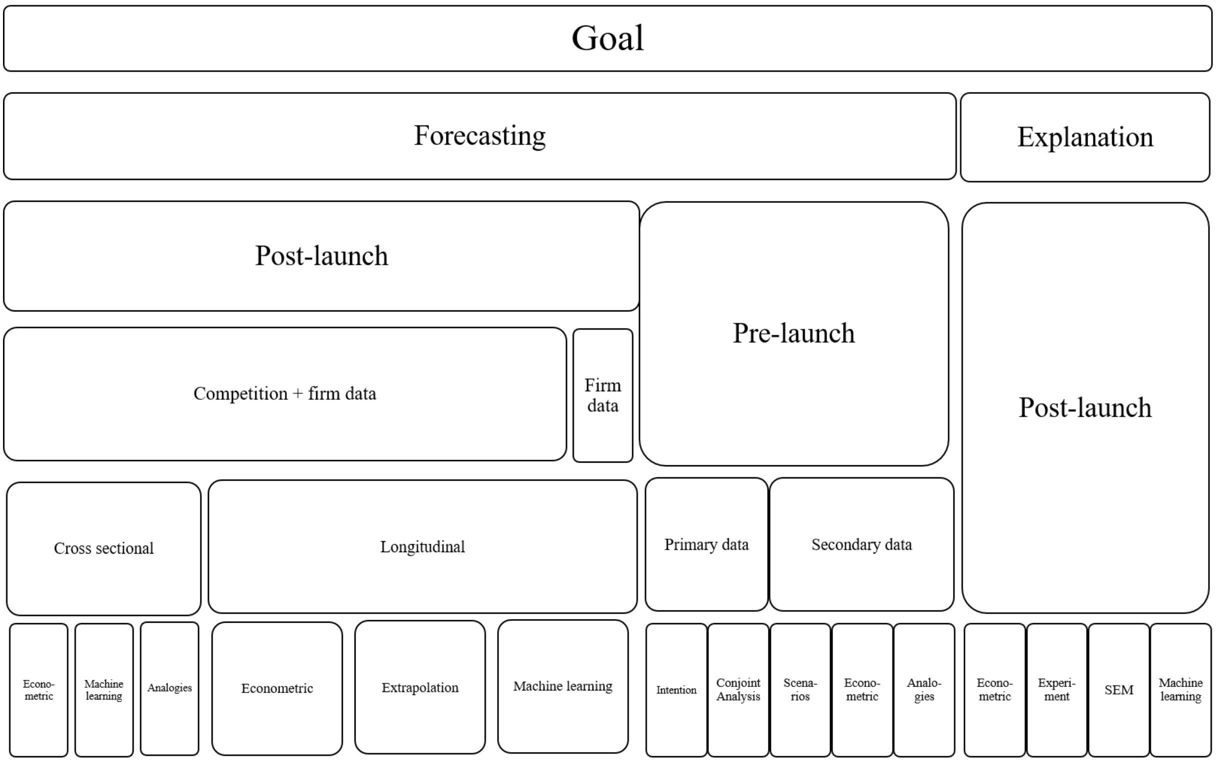
2.6 Future of forecasting in marketing

Some studies on marketing response have also contributed to improvements in methods. They have proposed different approaches (often Bayesian) or modified model versions to accommodate the data available or to solve problems of endogeneity in structural models (e.g. LUAN; SUDHIR, 2010; NARAYANAN; NAIR, 2013; PETRIN; TRAIN, 2010). Such studies also aspire for substantive contribution applying marketing theory to econometric models, such as brand awareness, choice complexity, satisfaction, word of mouth, conjoint purchase, or promotion expenditures. Still, there is a need for studies with a focus on forecasting that combines the advantage of descriptive models to be easy to apply with the capacity of structural models to give evidence of variables that impact on marketing response. Also, models and methods that ensure consumer privacy are, and will continue to be, on focus for companies and need to be addressed by researchers.

Most of the papers published study marketing response using structural models. These models have many benefits, but they also bring some challenges. The key challenge is to develop structural models that provide realistic descriptions of the environments in which firms market their products and services (REISS, 2011). Models like the basic marketing response (1) are not usually adopted by firms. As Fildes *et al.* (2008) sustain, the evidence of improved accuracy of such a model is lacking. This is because there is a need for simpler operational models that include marketing variables and that are downwardly compatible in the product hierarchy (e.g., from category to brand to SKU). In short, forecast accuracy and the level of complexity valuable in modelling the problem are under-researched (FILDES; MA; KOLASSA, 2018).

In Figure 5 we purpose a methodological framework based on our experience and the literature reviewed to help future researches and practitioners to select the methods to apply. To apply it, first, the researcher should choose the goal of the study. The choices are between forecasting or explanation, and post or pre-launch of a product or service. The following decisions are related to the data available. For post-launch forecasting, we can have internal data (firm) versus market data (firm and competition), and the final decision involves the availability of cross-sectional or longitudinal data. For post-launch forecasting, we can have primary or secondary data. The last row in Figure 5 shows the available methods for each combination of choices.

Figure 5. Proposed criteria to select market response methods



Source: The authors (2019)

Econometric models that appear in Figure 5 are the ones reviewed, mostly structural, using Bayesian or Classical inference. SEM stands for “structural equations modelling”. Combinations, such as ensembles or rule-based forecasting are not on the criteria since they imply the choice of more than just one method. Also, judgmental methods were not on the scope of the paper and were not included.

One of the opportunities for future researches on marketing response forecasting is the use of different computer-intensive methods, such as support vector machines and neural networks (FILDES *et al.*, 2008). However, in general, marketing academia is not providing professionals with the necessary analytical skills (KERWIN, ZMUDA, 2013; LEEFLANG *et al.*, 2014). Companies need to be able to predict and be ever more accurate and, because of the hype around machine learning accurate results, marketing practitioners are being pressured not only to understand what machine learning is, they also need to know how to work with it.

The combination of data from new sources is another opportunity. Especially real-time data that is already used in “areas as diverse as stock-price trading, electricity load forecasting […] and could be applied more widely” (FILDES *et al.*, 2008). These new sources of data need to be combined because databases do not have all the information about a customer’s behavior, even the best ones (CHEN; STECKEL, 2012). It remains an opportunity to combine data such as purchase, sensor, market, temporal (weather, traffic, etc.) and unstructured from social/mobile/digital/e-commerce (CHASE, 2013). This is important for products that are not frequently purchased (durables), for example, that do not have rich data sets like those from package consumer goods (CHEN; STECKEL, 2012).

However, ensuring anonymity is vital since companies are now facing new privacy regulations and discussions about leaks and unfair use of data are frequent on media. For example, it has been proved that metadata is able to uniquely identify individuals, and that “knowing four random spatiotemporal points […] is enough to uniquely reidentify 90% of the individuals and to uncover all of their records” (DE MONTJOYE *et al.*, 2013). This points to the importance to adapt marketing response models to work with aggregate data, maintaining the same predictive power (WEDEL; KANNAN, 2016).

Not only data but also methods tend to be combined to improve accuracy (i.e. ensemble models). The more the data and the methods differ, the greater the expected improvement in accuracy of the ensemble, if compared with the result of a single method forecast (ARMSTRONG, 2001; FILDES *et al.*, 2008). Such diversity can be achieved by combining the predictions obtained by methods applied in different samples (bagging is the most known example). Another possibility to increase diversity is to use different sets of variables (random forest or LASSO techniques can help select the set of variables), to use hybrid (stacking) or non-hybrid (bagging or boosting) methods, also ensemble with different combination rules, known as the weight functions (DE BOCK; COUSSEMENT; VAN DEN POEL, 2010).

Methods to be combined can be qualitative (judgmental), adding adjustments to the results of quantitative models, or by combining different expert and salesforce predictions. They can also be quantitative, with different machine learning or time series models, for example, combined. Ensemble methods are used in forecasting competitions, such as M4 competitions and Kaggle, with very good results.

Nevertheless, one important question that still needs to be answered is the real necessity of spending resources gathering and analyzing the vast amount of data that is now collected from consumers. With new sources and a bigger volume of data, researchers and companies face now an unfamiliar problem of selecting the variables to use and at what level of aggregation (DEKIMPE; HANSSENS, 2000). For theoretical purposes, it may be interesting to explore and explain the impact of any available variable on sales. However, for companies' managers that need to make better and faster decisions, it is more vital to know which variables will improve the predictive power of models and focus on them.

Concerning which variables are important to analyze, location is returning to the focus of researchers. This type of data is abundant nowadays because of mobile devices and applications (apps) that keep location metadata stored. They have been used, for example, to predict population movements (LU *et al.*, 2013). Also, data from satellite on the location of night lights have been used as an alternative to measure economic growth (HENDERSON; STOREYGARD; WEIL, 2012).

In marketing literature location data (although not metadata or satellite data, but zip codes and distances to stores/services) was used in marketing response models about hotels (ZHANG; KALRA, 2014), gas stations (CHAN; PADMANABHAN; SEETHARAMAN, 2007), fuel adoption (SHRIVER, 2015), drug prescriptions (STREMERSCH; LANDSMAN; VENKATARAMAN, 2013), solar panels (BOLLINGER; GILLINGHAM, 2012), organic products (SRIDHAR; BEZAWADA; TRIVEDI, 2012), and the car industry (ALBUQUERQUE; BRONNENBERG, 2012; BUCKLIN; SIDDARTH; SILVA-RISSO, 2008; NARAYANAN; NAIR, 2013).

There is still potential to explore location information such as studying similarities among customers preferences based on geographic location (CHUNG; RAO, 2012) or migration patterns influence on their purchases (BRONNENBERG; DUBÉ; GENTZKOW, 2012), for example. Wedel and Kannan (2016) state that location is a data source still to be explored by marketing.

A good example of the use of location in models is Bucklin, Siddarth and Silva-Risso (2008) study. They created a choice model with three measures of car dealers’ concentration, accessibility, and spread, based on the geographic locations of buyers and new car dealers. The authors found that these three measures were significantly related to new car choice, helping firms to decide the effects of opening or closing points of sale. The authors state that marketers need to be able to understand how changes in distribution (e.g., the size and structure of a dealer network) can affect the demand. They argue that only product, price, and promotion variables were incorporated as attributes in utility before their paper. However, the empirical relationship between market share (or sales) for a product and its level of distribution intensity is still an open question for the future research on consumer durables (BUCKLIN; SIDDARTH; SILVA-RISSO, 2008).

Even if new methods are not applied, there is still a chance for improvement in structural models that are the most commonly applied in marketing literature. Chintagunta and Nair (2011) suggest three new directions: dynamics, use of data on unobservable (primary data), and nonparametric approaches. Dynamics means models that consider purchase on present impacts purchase in the future, such as storability and durability. To model non-frequently purchased products such as durables is a challenge, not only because of its dynamics but also because scanner data are not so common for these products (ZHAO; ZHAO; HELSEN, 2011). The decisions concerning these types of products are also more sophisticated (involving more people and more time to decide), although empirical research on the subject is not frequent in marketing (NI; NESLIN; SUN, 2012). Therefore, a thorough understanding of consumer decisions about durables will not only help to develop and test both economic and consumer behavior theories but also will have important implications for managerial decisions.

Experience goods are also a dynamic problem because “purchase today provides a signal about quality, which updates the future information set” (CHINTAGUNTA; NAIR, 2011). To close the discussion of dynamics issues, complementarities refer to products that are purchased only after another one is, so the choice of the complementary product is dependent on the choice of the first product (common in technological products due to compatibility issues).

Another direction is related to improve models with primary data that reduce possible confounds, such as experiments. A final direction, also mentioned by Chintagunta and Nair (2011), is the use of nonparametric approaches due to the increased access to larger data sets. Some of these directions are overlapped, since dynamic problems such as intermittent demand (periods of no demand followed by periods of highly variable demand, which introduces lumpiness), can be improved by nonparametric approaches (FILDES *et al.*, 2008). Allenby, Garratt, and Rossi (2010) warn that current research is dominated by linear utility specifications and that it is necessary for structural models to use utility specifications that have more realistic assumptions. To sum up, the gaps to be explored in future research are:

• Development of simpler to implement and more realistic models;

• New approaches on correcting for endogeneity;

• The increasing integration of computer science knowledge;

• Incorporation of new sources of data;

• Development of new models with location/geographical variables;

• Combination data sets and methods;

• Improvement of models based on primary data;

• The consideration of the dynamic type of goods: durable, experience goods, and complementarities;

• Use nonparametric approaches and more realistic assumptions on utility specifications.

We showed that marketing literature focuses more on explanation and marketing practice has a higher accuracy / easy to implement orientation. One way of ending this contrasting focus is to use big data/computer-intensive methods. Marketing is the discipline responsible to understand and share customer knowledge with companies and society in general. Computer-intensive methods (combined or not with access to big data) unlocks possibilities not only to deepen such knowledge but also to predict it. Prediction of market response – i.e. sales, market share, demand – has developed in academia and in practice. Most of this development is now happening in other fields.

Marketing response models studied in marketing are mostly cross-sectional, hard to implement and since their focus is on explanation, they are less helpful when practitioners are planning future strategy. Companies rely on simpler models that give them a satisfactory level of accuracy and are easy to implement. Marketing should take the models that are used in practice and add marketing variables to improve it. For that, our field needs to integrate computer science knowledge to its reality. This even more important if the field wants to take advantage of big data. Marketing needs to change graduate and postgraduate curriculum. This will prepare market professional and researchers to the future of the field. It will provide them with the skills necessary to deal with new data sources and the methods most suitable to work with them. As for research, the focus on explanation should be divided or accompanied by a prediction application. In short, combining big data-based methods is still a major trend for future research in marketing forecasting.

3 IS EQUAL REALLY BEST? AN ALTERNATIVE APPROACH TO SET WEIGHTS OF ENSEMBLE FORECASTS

Abstract: Forecasting accuracy and easy to implement methods are not commonly found in marketing literature. The field has been focusing on the model's explanatory and not predictive power. However, sales predictions are important to define strategies and prevent unexpected changes in consequence of excess or lack of stock. We aim to contribute to marketing literature that focuses on prediction, by showing how different methods can improve forecasting accuracy, using easy to implement tools, with an empirical test case. We also show how methods used in practice, such ensemble techniques, can improve forecasts even further. The use of ensemble models is very popular in forecasting competitions such as Kaggle. They proved to be more accurate, by combining the results of more than one model and giving weights to each of them. We propose a method using optimization to calculate the weights based on each model's accuracy and the covariance matrix of such accuracies. The forecasting performance of this method is compared to equal weights methods for a database of sales from an electrical component manufacturer. The forecasts provided by the optimization method was more accurate than all 16 benchmark methods. In this way, the optimization method may be considered an alternative to current ensemble methods, using equal weights. Our paper also provides further evidence that ensembles have better performance than single models.

Keywords: marketing analytics; sales; forecasting; ensemble; weights

3.1 Introduction

Marketing modelling has given little attention to forecasting accuracy. This is partially due to the focus on explaining or describing consumer choice rather than to predict it. Marketing models also apply individual data and competition data, from consumer panels, usually obtained from companies such as Nielsen. These data are not available or affordable for most companies. Models are also very complex, using a large number of explanatory variables, making it computationally intensive, and slow to run. These models usually impose changes in the way companies collect, store and analyze data, making it difficult to implement.

All these factors have consequences for marketing as a research field and as a department. Not finding models that can be easily applied and that help decision making, practitioners turn to other fields to be able to plan their strategies. This also keeps marketing, as a department, distant from the strategic decision-making process (SILVEIRA NETTO; SLONGO, 2019). Prediction accuracy is important for companies, and for that reason should be also for marketing.

One type of data that is natural to every firm is the history of sales. Companies store which product was sold, when and for how much since the day they open for business, in most cases. This is a natural output of the sales process and it is accessible for all types of firms, from different sectors and sizes. Historical sales data has a times-series format and can be predicted using statistical methods. Researchers can predict using time-series many different marketing problems such as customer lifetime value, future buzz volume, stock value, however, one of the most intuitive and important ones is sales.

The methods applied to forecast specifically, and to marketing modelling, in general, are related to the types of data that became available to the researchers (CHINTAGUNTA; HANSSENS; HAUSER, 2016). Scanner data made possible to use structural models that were, before, applied in econometrics. Now access to bigger data sets has given researchers access to more explanatory variables, and bigger time-series, making it necessary to apply methods from computer science (i.e. machine learning techniques) to marketing problems. But these techniques do not only enable the use of bigger data sets, but they also improve accuracy and are easier to implement (WEDEL; KANNAN, 2016). These are, precisely, the gaps in marketing modelling.

Wedel and Kannan (2016) state that future studies and models in marketing should use computer science/machine learning approaches that are easier to implement, such as ensemble models. Some efforts to forecast sales using those techniques on marketing literature were made (e.g. ALI *et al.*, 2009; SUN *et al.*, 2008). However, marketing practice is far more advanced in its use then marketing research (CHINTAGUNTA; HANSSENS; HAUSER, 2016; CUI; CURRY, 2005; WEDEL; KANNAN, 2016).

As Silveira Netto and Slongo (2019) state, marketing needs to be ahead of the process of using data (big or small) and developing metrics. Marketing should be the department that provides knowledge about customers (i.e. data analyses) and that measures if they are satisfied (i.e. metrics). For that, it is important that the field proves to be accurate. More than explaining that is undoubtedly important, marketing needs to show that its models help to improve results, and to be more assertive on its actions. “Prediction”, “forecast”, “accuracy”, those are terms missing in marketing literature but are present in all executive board meetings.

Having that in mind, aggregate level sales forecasts are important to plan future actions and time-series models have been applied to forecast at this level (FILDES; MA; KOLASSA, 2018). However, unique models may perform well under some conditions but do not generalize. Literature compiles evidence that combining models have the best generalization performance, producing better forecasts across scenarios. The use of the combination of models (or ensemble models) to improve forecast is very popular in practice and in machine learning competitions, such as Kaggle or forecasting competitions, as the M4competition. They proved to be more accurate (COUSSEMENT; DE BOCK, 2013; DE BOCK; COUSSEMENT; VAN DEN POEL, 2010; LEMMENS; CROUX, 2006) and have much stronger generalization ability than single models (WANG *et al.*, 2014).

Ensembles combine the results of more than one base model by giving each of them weights. One common question concerns how to set those weights. The literature states that giving equal weights or averaging the results is the default and has not been outperformed yet (BARROW, 2016; GRAEFE *et al.*, 2015; KRAWCZYK, 2015). The present paper focuses on the use of ensemble learning and develops an alternative approach to stacking method and to equal weights, that focus on each base model's performance.

The aim is to achieve at least the same predictive power as the equal weights' technique when predicting sales, a regression problem. To accomplish that, the single base models used were combined in an ensemble model in which the base models' weights are defined using quadratic optimization. The dataset consists of sales from a manufacturer of electrical components. We compare two alternative techniques: equal weights, the default on literature so far; and quadratic optimization that optimizes both the error and the error variance of the different base models used, applied for subsamples of the data. The quadratic optimization alternative was inspired in the "vogging" method proposed by Derbeko, El-Yaniv and Meir (2002) in which the authors apply quadratic optimization to a classification problem to find weights to different applications of only one base model (support vector machine) for several subsamples.

The optimization technique proposed by Derbeko *et al.* (2002) was adapted from the finances field (portfolio optimization), and, in our proposition, was extended to take different base models into account. As far as we know, this still has not been applied to a time-series forecasting problem or in the marketing literature. This approach can contribute to establishing an alternative method to find an optimum combination of weights and improve the model's accuracy to predict consumer behavior, concerning time-series (regression problem).

We organize the remainder of this paper as follows: sections 2 and 3 comprises the literature review about the different base models, classifications of ensemble models, and the weights functions. We then describe the empirical application, results, and offer some concluding remarks regarding our approach, its limitations and future research.

3.2 Related literature

Ensemble learning makes predictions as a combination of single models' predictions. Wang *et al.* (2014) state that there are two conditions to attain a good ensemble: accuracy and diversity. That means that each base learner must have a different result (different pattern of errors) and be more accurate than a random guess. Ensemble models can also be considered along three dimensions (DE BOCK; COUSSEMENT; VAN DEN POEL, 2010): the training dataset (sampling, and variable selection); the algorithms of the single models chosen (hybrid or non-hybrid ensembles); and the weight function (combination rule).

The techniques of ensemble learning most commonly applied are bagging, boosting, and stacking. Bagging changes the sample, with the purpose to minimize variance (that can be caused by the overfitting of single models). Boosting uses the results of one model as a feature of the following model, changing the weights in a sequence for each instance in the training dataset (WANG *et al.*, 2014). The purpose of boosting is to improve predictions, reducing bias (that underfit can cause). Finally, stacking has the purpose to improve prediction and minimize variance by combining the results of several models, using another model (e.g. linear regression) to combine them.

Bagging, or bootstrap aggregating, applies any learning algorithm to bootstrap samples of the training set (DERBEKO; EL-YANIV; MEIR, 2002). It uses the same algorithm, changing the sample, getting different results independently from each other. After all the models predict the result, the majority will decide the result of the ensemble. Boosting is more complex to put in practice than bagging (LEMMENS; CROUX, 2006) and comprises estimating iteratively, with the next model focusing on the cases that the previous is not predicting correctly. This way, the result of each estimate is dependent on each other, unlike the bagging technique.

Armstrong (2001) claims that the result of combining different models will be more accurate the more the applied data and methods are distinct from each other. Stacking is the ensemble technique that follows this statement by using different algorithms and combining the results using another model. The results, however, may have the same advantages and disadvantages, simultaneously, of the methods used to generate them (HUNG; CHEN, 2009).

After deciding the sampling and variable selection (i.e. the ensemble method), the base models used need to be selected. Any method can be combined. Judgmental forecasts based on the opinion of frontline employees, expert opinions, and the results of statistical or machine learning methods. On our ensemble we focus on quantitative methods, combining the results of well-known time-series forecasting methods with machine learning techniques, in a hybrid approach.

From the range of models to forecast time-series the simplest is to repeat the last observation. This is called a naïve method or random walk and follows the notation (HYNDMAN; ATHANASOPOULOS, 2018):

$$\hat{y}_{T+h|T} = y_T \tag{2}$$

Or, for seasonal data, this can be adapted to repeat the last observation of the past most recent season available on the data:

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)} \tag{3}$$

Where m is the seasonal period and k is the number of complete periods on the data. Additionally, to seasonality, we can also allow the random walk or naïve method to account for trend (increase or decrease over time). This is called drift. These methods are well-known benchmarks when proposing new time series methods.

Linear and nonlinear (usually with cubic splines, or with smooth transition) autoregressive models can also forecast time-series. This means that the forecast will be the result of a model where the next value of the predicted variable is a function of the other values of the same variable (lagged values), present on the data.

Other well-known and most applied classes of models are Autoregressive Integrated Moving Average and Exponential smoothing (ETS). Exponential smoothing gives weights to past values, with bigger weights to the most recent. This class of models are reliable and computationally efficient. They have three components (Error, Trend, Seasonal) that, combined generate different models. Trend component can be none, additive, and additive damped. The seasonal component can be none, additive, and multiplicative. The final component, error, can be additive or multiplicative. Notations and detailed explanation of the different models obtained with these three components are provided by Hyndman *et al.* (2008).

Autoregressive Integrated Moving Average (ARIMA) models are complementary to ETS models and focus on the autocorrelations instead of trend and seasonality. The seasonal ARIMA is written (HYNDMAN; KHANDAKAR, 2008) as:

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d\, y_t = c + \Theta(B^m)\theta(B)\varepsilon_t \tag{4}$$

where m is the number of observations in the year, ε_t represents white noise $(0, \sigma^2)$, B or the lag is the effect of the last period. Polynomials of order p (the autoregressive part) and q (the moving average part) are represented, respectively, by $\phi(z)$ and $\theta(z)$, on the non-seasonal part of the model, while $\Phi(z)$ and Θ (z) have order P and Q, of the seasonal part. d is the degree of first differencing and D is the seasonal differencing. And c represents the changes (average) between consecutive values of y_t.

To forecast time-series, the most commonly implement machine learning technique is a neural network with a nonlinear autoregressive model. It is a multilayer feed-forward network with only one hidden layer, varying in the number of neurons, and with a linear output. In the time-series case, lagged values are used as inputs for the neural network. Algorithms implemented for time series forecasts fit a model that is similar to an AR(p) with nonlinear functions, for non-seasonal data. For seasonal, it is the equivalent of estimating an ARIMA(p,0,0)(P,0,0)[m] model, with nonlinear functions (HYNDMAN; ATHANASOPOULOS, 2018).

Ali *et al.* (2009) and Sun *et al.* (2008) are examples of forecasting applications that apply machine learning techniques with marketing variables in their models, even though those were not published in marketing journals. Despite the hype over machine learning techniques, the results of improved accuracy are mixed. Makridakis; Spiliotis and Assimakopoulos (2018) affirm that evidence of the accuracy of these techniques are still scarce and conclude from an empirical test that they do not outperform traditional statistical methods. The authors support that a fair comparison between statistical and machine learning techniques should evaluate longer forecast horizons (not only one-step ahead), benchmarks should also be used, and they should be applied to many different time-series and not just one.

After estimating the forecasts with any different set of single models, the results need to be combined in one, all the models need to reach an agreement on which is the final forecast. One of the most important questions that arise from the use of ensemble learning is how to design that combination of single models. This is called the weight function that will allow aggregating the outputs, creating a more accurate prediction.

In forecasting literature, the different approaches for combining forecasts are the median; performance-related; principal component; least squares (optimal) combination weights; projection on the mean; and shrinkage (Bayesian) combinations (for more on these approaches refer to Genre, Kenny, Meyler and Timmermann, 2013). However, the "forecast combination puzzle" refers to the fact that frequently simpler combinations perform best in practice, such as those that give equal weighs (GENRE *et al.*, 2013).

Bagging and boosting uses average (for regression) or a majority (for classification) vote, and stacking gives more weight to models that perform better (DERBEKO; EL-YANIV; MEIR, 2002; GENRE *et al.*, 2013; GRAEFE *et al.*, 2014; VARIAN, 2014; ZHANG *et al.*, 2014). The literature, as far as we know, agrees that majority voting or averaging was not outperformed yet (BARROW; CRONE, 2016; GRAEFE *et al.*, 2015; KRAWCZYK, 2015). So far, “no significant improvements have been found from applying more complex, weighted approaches to forecast combination” (BARROW; CRONE, 2016, p. 1113).

Graefe *et al.* (2014) claim that not always equal weights will be better, different approaches should be considered when we have a small number of forecasts that are very different in accuracy and access a larger sample (GRAEFE *et al.*, 2014). Some studies have applied alternatives to averaging and majority voting with some success, but with more computational costs (e.g. ZHANG *et al.*, 2014).

Derbeko, El-Yaniv and Meir (2002) developed “vogging” (variance optimized bagging), a technique that, according to them, “can consistently outperform bagging”. This technique was adapted from the theory of financial portfolios, more specifically the Markowitz Mean-Variance Portfolio Theory. The authors propose the construction of “optimized portfolio of bootstrapped classifiers” (DERBEKO; EL-YANIV; MEIR, 2002). The main idea is to produce an optimum linear combination of models that: has a better prediction accuracy (as in the stock return) and minimize the variance (as in the risk of a financial portfolio).

However, the authors apply this technique only to a classification problem and consider only one single base model (support vector machine). The purpose of the present paper is to fill those gaps. To do so, we apply the proposed method to a regression problem (sales forecasting), include several base models in the ensemble technique and compare its forecasting performance to equal weights ensemble.

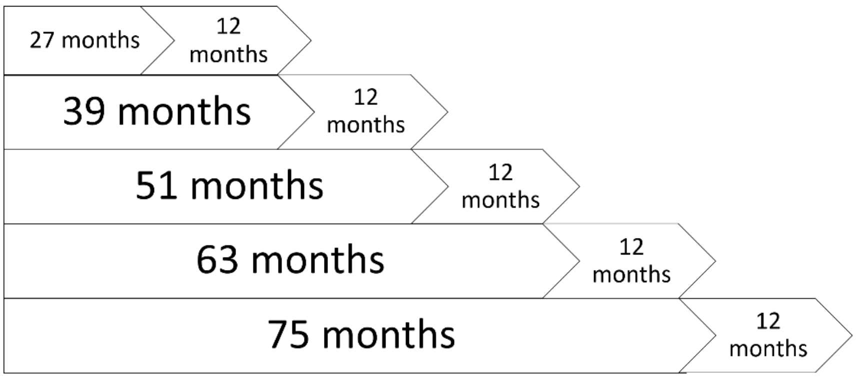
The concept behind these alternatives is to find a method to select weights for the models that are related to their predictive power and minimize variance at the same time. However, the method should also be easy to implement (i.e. automatized), in opposition to the propositions to choose, by trial and error, a combination of weights. This is the purpose of our approach. Next, we describe the dataset, the base models applied, and the pseudo-algorithm of the proposed approach.

3.3 Material and methods

In order to test our method to set the weights for the ensemble, we studied the algorithm proposed by Derbeko, El-Yaniv and Meir (2002) and adapt it to time-series forecasting. We first replicated their algorithm, then we proceed to adapt the steps to the problem we aim to contribute. For example, we were not able to apply bootstrapping to our dataset because time-series have an order on data that needs to be kept. For that reason, we separated different folds of the data using a rolling forecasting origin. This is a standard approach in time-series forecasting literature and ensures that the model has different training and test sets on every fold.

We ran the forecasts for the 5 different data sets (5 folds) on each of the 15 single (base) models. Figure 6 shows the size of each data set and how the rolling origin works.

Figure 6 – evaluation on a rolling forecasting origin



Source: The authors (2019)

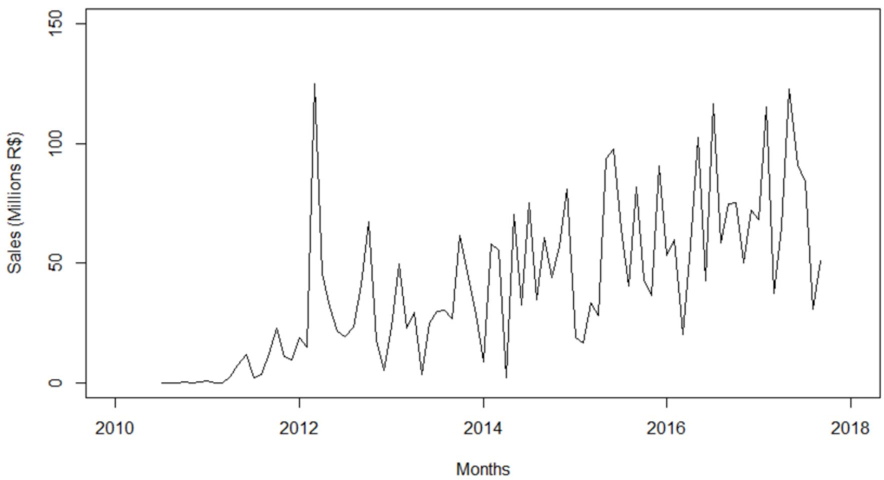
The four first data sets were used for the optimization part of our algorithm, and the final outcome of this phase are the weights of each base models to be combined. The last data set was used in the predictions' phase.

All data analysis, forecasting and output were conducted using R (R CORE TEAM, 2018). Single (base) models` algorithms were implemented using the R packages: forecast (HYNDMAN *et al.*, 2018; HYNDMAN; KHANDAKAR, 2008), tsibble (WANG; COOK; HYNDMAN, 2018), nnfor (KOURENTZES, 2019), tsfknn (MARTINEZ, 2018), NlinTS (HMAMOUCHE, 2019), tsDyn (NARZO; AZNARTE; STIGLER, 2009; STIGLER, 2010). The optimization used is quadratic, implemented by quadprog (WEINGESSEL, 2013). Descriptions of other packages used for more trivial tasks and are not presented here but are available at request, as well as the code for reproducing our analysis. The remaining subsections will describe the data, the 15 base models and each part of our pseudo-algorithm to set the weights of the ensemble.

3.3.1 Data

Sales to channels (B2B clients) located in São Paulo (Brazil) from a major manufacturer of plugs and light switches were used to run the models and the optimization approach proposed. We combined different datasets provided by the company. Sales made by the company's distributors was combined with sales made directly by the factory, and location data. Next, we filtered the data corresponding to the sales of electrical components made for the city of São Paulo. The data cleaning consisted of selecting the variables of interest (location and sales amount by time period) and aggregating sales in months (plot of the last fold is shown in Figure 7).

Figure 7 – Time-series' plot (fold #5)



Source: The authors - R output (2019)

Different time aggregations were tested to minimize the amount of missing data that would make forecasting infeasible. The final dataset used has the history of monthly sales from over seven years, from July 2010 to September 2017.

3.3.2 Base models

The criteria to choose the base models was that the model is well known in forecasting literature, and can be fitted automatically by algorithms available in CRAN (THE COMPREHENSIVE R ARCHIVE NETWORK, [*s.d.*]). The source for this choice was the network`s tasks views page, which presents a list of packages for time series forecasting. It served as a guide to establishing the methods that are well known among those available. The ones applied to our ensemble are presented in Table 2.

Table 2– Base models implemented

Model	Definition
AAR	Additive nonlinear autoregressive model implemented by package tsDyn
ARIMA	ARIMA model implemented by package forecast
ETS	Exponential smoothing state space model implemented by package forecast
KNN	KNN regression where the number of nearest neighbors and the lags are selected automatically implemented by package tsfknn
LINEAR	Linear Autoregressive model implemented by package tsDyn
LSTAR	Logistic Smooth Transition Autoregressive model implemented by package tsDyn
NAÏVE	Equivalent to an ARIMA (0,1,0) implemented by package forecast
NNETAR	Feed-forward neural networks with a single hidden layer and lagged inputs implemented by package forecast
NNETTS	Neural Network nonlinear autoregressive model implemented by tsDyn
RWF	Random walk with drift implemented by package forecast
sNAIVE	ARIMA (0,0,0)(0,1,0)m model where m is the seasonal period implemented by package forecast
STLM	applies an STL decomposition, implemented by package forecast
TBATS	Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components implemented by package forecast
THETAF	Equivalent to simple exponential smoothing with drift implemented by package forecast
VARMLP	Artificial Neural Network VAR (Vector Auto-Regressive) model using a MultiLayer Perceptron implemented by package NlinTS

Source: Adapted from R packages' manuals (THE COMPREHENSIVE R ARCHIVE NETWORK, [*s.d.*])

All parameters for the models are fitted automatically by the algorithms implemented by the R packages. The packages apply searches for the best model based on some criteria. For example, the auto.arima function of forecast package searches for the best ARIMA model according to information criteria (AIC or BIC value). All algorithms were applied using their default settings.

3.3.3 Pseudo algorithm to set ensemble`s weights

The ensemble was developed to predict the sales using two techniques: equal weights combination and our approach, inspired by “vogging” (DERBEKO; EL-YANIV; MEIR, 2002), that uses optimization to set the weights of each single (base) model. The pseudo algorithm of our proposed approach (optimization) was:

Input:

1) Clean data and aggregate it temporally (i.e. by month);
2) Choose the b base models that will be combined;
3) Set forecasting horizon (i.e. 12 months ahead).

Optimization:

1) Create n subsamples from the dataset, in a rolling forecast origin, dividing each fold into train and test sets;
2) Train each of the n subsamples on the b base models;
3) Test each b base model on each test set;
4) Measure the resulting relative accuracy rate (RMSE) of the predictions of each b base models on each n subsample (b x n accuracy measures);
5) Store b x n error statistics;
6) Weights definition:
   i. Calculate the mean accuracy measure of the n subsamples
   ii. Calculate the covariance matrix among the b x n accuracy measures
   iii. Set the constraint to have weights summing up to 1
   iv. Set the constraint to have weights non-negatives
   v. Run the quadratic optimization minimizing the error variance
7) Store the weights for each model.

Prediction:

1) Train the b base models on the train set;
2) Test the b base models on the validation set;
3) Store the prediction of each b base model for each observation in the validation set;
4) Multiply the predictions of each b base model by the weight vector;
5) Sum the weighted resulting values of each b model for each observation (linear combination) to find the final prediction vector.

The idea behind it is to solve a quadratic problem with linear constraints in order to optimize a linear combination of regression models' predictions to reduce error variance while attempting to preserve accuracy. The results of the empirical application are described in the next section.

3.4 Results

Table 3 shows the accuracy results of the last data set with each base model. ARIMA was the best performing method. All nonlinear methods and machine learning techniques were outperformed, in our empirical application, by a standard, well-known statistical time-series forecasting method. This gives further support to the recent findings that for the task of forecasting time-series, without additional features (or covariates), statistical methods have a better performance than machine learning techniques (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018).

Table 3– Average error measures of base models

	ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**	**Theil'sU**
ARIMA	3603499	25478211	19595122	-9.92	31.95	0.81	0.07	0.71
NAIVE	-3042604	27307277	21919910	-22.3	39.34	0.90	-0.03	0.78
THETAF	-1185096	27349169	21709841	-19.5	38.45	0.89	-0.03	0.78
ETS	4177159	27456850	21551034	-10.5	35.10	0.89	-0.03	0.74
TBATS	4436889	27497562	21594322	-10.0	35.03	0.89	-0.03	0.74
RWF	-9610396	29280951	23268805	-33.6	45.34	0.96	-0.01	0.92
sNAIVE	8884134	29394661	25975058	5.43	39.00	1.07	0.21	0.77
NNETAR	10628161	30393797	22439887	1.61	32.61	0.92	0.37	0.85
VARMLP	12477120	30876710	25223810	2.12	37.17	1.04	-0.02	0.83
KNN	14167350	33883950	28687890	8.11	44.05	1.18	0.43	0.88
LSTAR	21474130	37567460	29661910	17.03	38.57	1.22	-0.11	1.01
AAR	26357020	38121710	29848970	25.71	36.38	1.23	-0.01	1.02
LINEAR	27921130	39603620	30773630	28.61	36.67	1.26	0.02	1.08
STLM	30324807	39811905	30890133	35.08	36.59	1.27	0.14	1.14
NNETTS	34068020	44655840	35354870	37.91	41.35	1.45	-0.12	1.21

Source: The authors - R output (2019)

Table 4 shows the results of the best performing base model and the ensembles using equal weights and the weights calculated on the optimization phase of our approach.

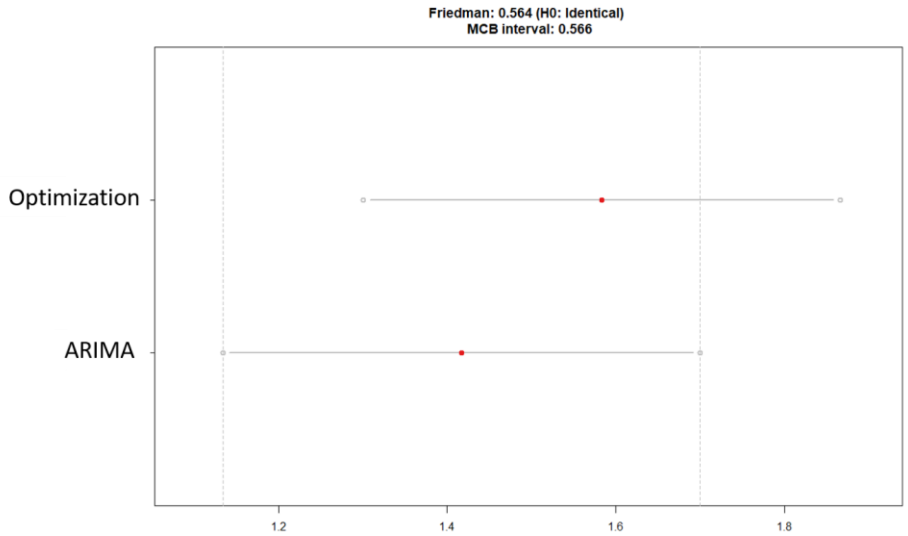
Table 4 – Average error measures of ensemble models and best performing base model

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil'sU
Optimization	4828497	24621082	19278772	-5.21	29.24	0.79	0.24	0.70
ARIMA	3603499	25478211	19595122	-9.92	31.95	0.81	0.07	0.71
Equal weights	12312088	28814389	22989908	3.72	32.79	0.94	0.09	0.77

Source: The authors - R output (2019)

The best performing method was the optimization ensemble in all the error measures estimated. The only exception is ME, that ARIMA has better performance, however not statistically different from our ensemble approach (Figure 8). The plot shows two lines on the edges of the best performing method and if the competing method has a statistically different forecast the mean (red dot) will be outside those lines.

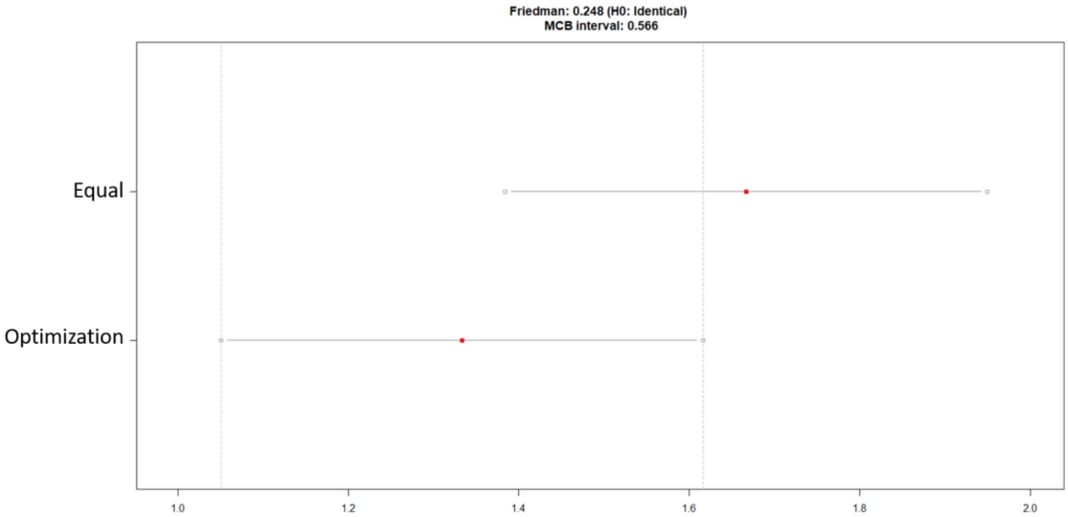
Figure 8 – Friedman and Nemenyi Test between ARIMA and optimization



Source: The authors - R output (2019)

When compared to the equal weights, our approach is not only outperforming the benchmark in all metrics, also these errors are statistically different, as the Friedman and Nemenyi Tests shows (Figure 9).

Figure 9 – Friedman and Nemenyi Test between ensemble approaches



Source: The authors - R output (2019)

We also analyzed the weights given by the optimization to every base model in Table 5. We can see that, apart from THETAF, the models with the best performance and lower standard deviation received weights different than 0.

Table 5 – Weights and standard deviation of each base method

	Weights	**sd**
AAR	0.00	5398183
ARIMA	0.26	3300355
ETS	0.00	5555206
KNN	0.00	6769904
LINEAR	0.00	9780009
LSTAR	0.00	9592754
NAIVE	0.15	4875240
NNETAR	0.08	5723460
NNETTS	0.00	11484189
RWF	0.06	4939575
sNAIVE	0.45	2513157
STLM	0.00	8781221
TBATS	0.00	5567844
THETAF	0.00	2424646

VARMLP	0.00	6118565

Source: the authors (2019), R output

The forecasts provided by the optimization method performed statistically different and better than the equal weights method. It is worth to notice that equal weights ensemble, in this case, did not improve accuracy when compared to more traditional methods such as ARIMA or ETS used alone. Such a result contradicts recent findings in the literature (BARROW; CRONE, 2016; KRAWCZYK, 2015). Notwithstanding, the forecasts provided by the optimization method performed better than all base models tested as well. The next section offers some conclusions and discusses limitations and future studies opportunities.

3.5 Discussion and conclusion

Ensemble models are known to be more accurate than any single models, by combining the results of more than one model and giving weights to each of them. The present paper discussed the question concerning the method to define those weights. We reviewed the literature on forecasting that states that equal weights or averaging the results have not been outperformed yet. We also described the types of ensemble and some of the most well-known statistical time-series forecasting methods.

We proposed and tested an optimization based on the finance literature to calculate the weights based on each model's accuracy and the covariance matrix of such accuracies. This approach proved, in our empirical test, to be a better performing alternative to equal weights approach and all 15 single (base) models tested.

This paper contributes to the literature by focusing on time-series forecasting; reporting accuracy evidence based on a longer forecasting horizon (12 months); comparing several benchmark models, including statistical and machine learning methods; and by adapting an alternative approach to set weights to ensemble methods. Marketing needs to gain familiarity and learn the skills to work with predictive analytics. For that reason, our goal is, mostly, to introduce to marketing to statistical and machine learning techniques applied to forecast time-series, allowing the field to analyze and advance on science based on this type of data that is becoming available with advances in big data and storage capabilities.

Our paper also serves as a guide to the methods most commonly applied to time-series forecasting that have algorithms already implemented in R programming language, an open-source tool, easy to implement. Our empirical application provides further evidence of the accuracy of statistical time-series forecasting methods. It also shows evidence that, contradicting what the literature reports to this day, equal weights can be outperformed by approaches that set the weights based on the performance of each base model.

Our empirical application has some limitations. The base models all gave highly correlated predictions and the models results are not diverse enough, so this could be interfering in the ensemble performance that could benefit even further from diversity. The optimization approach still needs to be validated in other datasets, to verify if this diversity will improve the technique and for generalization purposes. It also needs to be improved to allow a real "portfolio" optimization, letting us compare different sets of weights. Future research directions may also include the test for classification problems in marketing, such as churn prediction. Other studies might also compare the performance of other ensemble methods, besides equal weights, or to deep learning methods, such as long short-term memory (LSTM).

4 WHAT IS THE OPTIMAL STRATEGY OF AGGREGATION FOR FORECASTING SALES? TIME SERIES FORECAST RECONCILIATION BY REGION, PRODUCT CATEGORY, AND CHANNEL

Abstract: While some companies still struggle to gather, store and analyze data necessary to make better predictions, others are worried about increased requirements for data minimization and anonymization. This scenario raises questions about which variables are important to gather, and the resources necessary to do so. An important topic for academia and practice is how to improve forecasts, having limited access to data. In this paper, we consider such difficulties and propose forecasting strategies based on sales data in different aggregations criteria and structures. We compare aggregation criteria using both hierarchical and grouped time series structures, applying data that most companies already have access, marketing mix variables. Our paper indicates whether product category, channel type or region (geographic location) works best alone or combined when using the optimal reconciliation approach. This research suggests how to run sales forecasting more efficiently, using open-source tools. The method is also generalizable to all types of goods.

Keywords: sales; forecasting; marketing; hierarchical time series; grouped time series

4.1 Introduction

To plan and deliver products and services, it is necessary to know what the future might hold. Sales data are among the most important types of data related to future business performance, and sales forecasts are important to the most basic processes in any organization (ARMSTRONG, 2001; FILDES *et al.*, 2008; SEAMAN, 2018). Therefore, forecasting accuracy is not just an important topic for academics, to develop, test, and improve their methods. It is also a major concern for practitioners and for marketing strategy. For example, Seaman (2018) illustrates that an accurate forecast is important for retailer pricing and distribution strategies. Errors in sales forecasts can lead to products being out of stock (leading customers to buy from competitors) or be produced in excess (leading to holding costs and price promotions to increase sales). The consequences of out-of-stock products or excess price promotions are well-known, especially in the marketing literature. They include harms in brand reputation, quality perception, customer loyalty, repurchase intentions, and satisfaction. It is arguable that being able to forecast future purchase totals is more valuable to marketing than for other fields in social sciences (CHINTAGUNTA; NAIR, 2011).

Much of the marketing literature focuses on disaggregate analysis using data sets from companies that have access to individual-level information (CHINTAGUNTA; NAIR, 2011). However, most companies do not have easy access to this type of individual-level data. Gathering, storing, and analyzing this type of data is complex, expensive, and time-consuming. This difficulty is perceived by small and medium-sized enterprises (SMEs), that do not have the resources to handle such data. This is also true for offline business that does not have access to the end customer data, and B2B companies, such as low-involvement, durable goods producers (CHEN; STECKEL, 2012; HANSSENS, 1998). Not to mention the increasing concerns and costs associated with ensuring the privacy of consumer information, a task that is becoming a legal responsibility and a strategic asset in building relationships with customers (WEDEL; KANNAN, 2016). This scenario of lack of access, coupled with increased requirements for data minimization and anonymization, makes access to individual-level data increasingly difficult. A key theoretical and empirical question is how to make better plans and decisions, having limited access to resources and, possibly, to the data itself. By considering such difficulties, this paper proposes a sales forecasting strategy based on proprietary data to which most companies already have access, that is sales data aggregated, using different criteria and structures.

The decision of which strategy of data aggregation to use has been a subject of discussion for many years in different fields of knowledge, such as marketing (e.g. ABHISHEK; HOSANAGAR; FADER, 2015; RUSSELL; KAMAKURA, 1994; TELLIS; FRANSES, 2006), operations research (FLIEDNER; LAWRENCE, 1995), statistics (DUNN; WILLIAMS; DECHAINE, 1976), and economics (ZOTTERI; KALCHSCHMIDT; CANIATO, 2005). Sales can also be forecasted at many levels of aggregation, using different criteria to divide those levels, such as geographic considerations, product categories, types of channels, among others. This is necessary to plan organizational budgets more precisely (KREMER; SIEMSEN; THOMAS, 2015) and to help managers to diagnose forecast errors in more detail (DIVAKAR; RATCHFORD; SHANKAR, 2005). This will determine not only the level of production and distribution for each product, store, and location, but also the allocation of promotion resources, sales representatives, and even the amount of energy firms will spend on each business partner. As Fliedner (2001) and Seaman (2018) state, different departments and hierarchical levels in a firm have different, but related, interests when it comes to sales forecasts. For example, a retail chain store needs to access the forecasts of its own departments. On the other hand, its head office needs a more aggregated level forecast to plan the chains' strategy. The different forecasting needs must be delivered as part of a forecasting system. Further, they should be "coherent", meaning that disaggregated forecasts should add-up to the total forecasts, in the same way as the historical data (HYNDMAN *et al.*, 2011). These coherent forecasts, either by level or groups, can add accuracy to marketing efforts on those products, partners, departments, regions, channels, that are more likely to be profitable. Therefore, the theory should demonstrate which type of data and aggregation strategy produces the best sales forecasting results.

In this paper, we show which information (product category, channel type, or geographic location) combined with different structures (hierarchical or grouped time series) improves forecast accuracy the most. This, to the best of our knowledge, has not yet been addressed. Thus, the goal of this paper is to test which type of marketing variables works best when using the optimal reconciliation approach to forecast sales. The paper makes the following contributions. First, we advance the knowledge about sales forecasting by comparing different levels of aggregation and reconciling them through trace minimization (WICKRAMASURIYA; ATHANASOPOULOS; HYNDMAN, 2018). We propose an aggregation strategy that can lead to better decisions regarding budget allocation, leading to more precise actions at the point of sales and improved results. Second, we show that such a strategy can be used to improve sales forecast accuracy, by combining information that is easily available to firms. We demonstrate that it can be applied across many levels of aggregation, is easy to implement, and is relevant to practice. We also contribute to the management literature by providing a forecasting approach that uses what is one of the most basic types of variables existing in our field, that is, marketing mix variables. Third, we demonstrate our approach through a large-scale forecast reconciliation study, with several aggregation criteria using both hierarchical and grouped time series. Wedel and Kannan (2016) state that future studies should focus on models that are easy to implement in practice. Our last contribution follows this recommendation, by using automatic forecasting tools implemented in open-source software, based on data easily accessible to companies. Our strategy is generalizable to all types of goods.

This paper unfolds as follows. Section 4.2 presents the concepts and notations of grouped and hierarchical time series, the different aggregation criteria studied in the literature, and the methods for forecast reconciliation. Section 4.3 presents the data and details of the methods' implementation. Section 4.4 describes how the forecast accuracy of the different combinations of structures and aggregation criteria was evaluated and show the results. We conclude in section 4.5, by offering a discussion of the results and its implications to practice, followed by suggestions for future research.

4.2 Theory

Marketing research largely focuses on cross-sectional, structured, quantitative, and causal models that can help explain how the market responds to changes in the marketing mix. One of the most studied models in marketing is the discrete-choice class of models. They are popular because they can be causally interpreted, but also because much of microdata in marketing is about consumer choice from a fixed set of alternatives within a category (CHINTAGUNTA; NAIR, 2011).

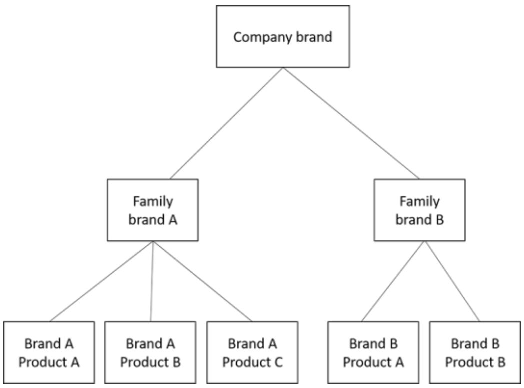
Even when research in marketing focuses on prediction rather than explanation, the results are challenging to marketing practice since their implementation is not easy. As stated by Fildes *et al.* (2008, p.1165), "evidence of improved accuracy is lacking [… and] there remains a need for simple operational models that include key marketing instruments and that are downwardly compatible in the product hierarchy (from category to brand to SKU)". Because forecasting accuracy is more important in practice than understanding the impact of each variable, practitioners choose forecasting methods that are simple to implement (FILDES *et al.*, 2008). For example, time-series models which identify patterns (trend, seasonality, cyclical, and randomness) and extrapolate them into the future. These techniques have higher predictive accuracy in stable markets, are simple to develop, and require a limited amount of data (CHASE, 2013). However, marketing has not given much attention to time-series. This is attributed, among other reasons, to the resistance of marketing scientists to data-driven approaches and to the lack of adequate data sources (DEKIMPE; HANSSENS, 2000).

The context of previous studies in marketing are, usually, companies that produce and distribute consumer packaged goods on a global scale (e.g., DIVAKAR; RATCHFORD; SHANKAR, 2005). This category of products is commonly used since it tends to provide richer data sets (CHEN; STECKEL, 2012). Nevertheless, one may face some challenges regarding the available data and the needs of a durable good manufacturer. Scanner data are uncommon for infrequently purchased (durable) items, so they rely on aggregated sales data. Also, it is more difficult to forecast sales of durable goods, leading to lower forecast accuracy than for other types of goods (WACKER; LUMMUS, 2002). Despite these challenges, our aim is to offer a tool that can be generalized for any company, regardless of the product or category, using aggregated data.

In forecasting literature, the issue of aggregation has been studied under two research streams (DEKKER; VAN DONSELAAR; OUWEHAND, 2004). One is the kind of aggregation, what we call in this paper "aggregation criteria", that is related to how the forecaster chooses to divide the levels of data. For example, one can divide the data by products or by similarity of seasonal patterns and then structure those levels as hierarchies or groups. The other is focused on how to adjust the forecasts done in different levels of aggregation until they add up. The later stream is called reconciliation (HYNDMAN; ATHANASOPOULOS, 2018).

Another issue to consider is the structure of the aggregation. Grouped time series are those that can be aggregated based on some criteria such as product characteristics, geographic regions, customer characteristics, and so on. When these criteria can also be represented in a tree structure, it is called a "hierarchical time series", as shown in Figure 10. In a marketing context, hierarchies commonly occur due to geography (where total sales are disaggregated by state, region, city, and store) and due to product classification (where brands are disaggregated into groups, sub-groups, and finally products). An example of grouped time series is when both geographic and product hierarchies are used simultaneously, such as when one wants to forecast for different products in different regions.

Figure 10 - Hierarchy example



Source: The authors (2019)

4.2.1 Aggregation criteria

There are several decisions to make when setting up a forecasting system with hierarchical or grouped time series, which can influence the system performance (FLIEDNER; LAWRENCE, 1995). One of the first decisions is what aggregation criteria to use. Coherent forecasts grouped by product or distribution channels can help determine companies' effort allocation. This can be done not only by product or channel, but also by partner, department, or region, giving information about which of those are more likely to be profitable.

Fliedner and Mabert (1992) studied the influence of different grouping criterion on the performance of hierarchical forecasts. They conclude that the criteria used to determine the groups for forecasting are determinant to the success of a forecasting system. However, the number or size of the groups does not have a significant impact. Series can also be aggregated based on their similarities using clustering methods (see DEKKER; VAN DONSELAAR; OUWEHAND, 2004; FLIEDNER; LAWRENCE, 1995; ZOTTERI; KALCHSCHMIDT; CANIATO, 2005) or by temporal aggregation (see KOURENTZES; ROSTAMI-TABAR; BARROW, 2017). However, when the groups do not correspond to market-related criteria, they are less useful for budget plans and strategy development.

In a later study comparing different groups based on volume but generated by cluster techniques, Fliedner and Lawrence (1995) found no evidence that the added sophistication improved forecast performance. For them, grouping items is responsible for improved forecast performance, but not the process of group formation. Divakar, Ratchford and Shankar (2005) on the other hand focused on forecasting for existing products by channels. Still, the comparison between different criteria based on marketing mix variables was not addressed, and the present paper intends to fill this gap.

4.2.2 Reconciliation approaches

Another decision in a hierarchical or grouped forecast system is related to the reconciliation approach. Following the notation of Hyndman and Athanasopoulos (2018), we denote the data at the most aggregate level at time t by y_t ($t = 1,\dots,T$). More disaggregated data are denoted by $y_{j,t}$, with j corresponding to the "node" of the observation. In this way, the time series of Figure 10 can be written as follows.

Bottom-level: $y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t}$ (5)

Middle level: $y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}$ (6)

$y_{B,t} = y_{BA,t} + y_{BB,t}$ (7)

Top-level: $y_t = y_{A,t} + y_{B,t}$. (8)

While the data will naturally add up appropriately, following the hierarchical structure, the forecasts may not. This can cause confusion when firms use the forecasts to plan their actions. For that reason, hierarchical forecasts must be "coherent" (WICKRAMASURIYA; ATHANASOPOULOS; HYNDMAN, 2018), that is, they must add up in the same way as the historical data.

We let $\hat{\boldsymbol{y}}_h$ denote the vector of forecasts for all nodes at horizon h, stacked in the same order as $\boldsymbol{y}_t$. These forecasts can come from any appropriate model. They are created independently for each node without regard for the hierarchical or grouped structure of the data. For example, the forecasts of the aggregate may be obtained from an ARIMA model, while forecasts for the most disaggregated series might come from a Delphi process for each sales division. We call these "base" forecasts.

Because we require coherent forecasts, these base forecasts must be reconciled; that is, they are adjusted to ensure they add up appropriately. Let $\tilde{\boldsymbol{y}}_h$ denotes the reconciled forecasts. These can be expressed as

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{R}\hat{\boldsymbol{y}}_h \tag{9}$$

where $\boldsymbol{R}$ is a reconciliation matrix that can be decomposed as $\boldsymbol{R} = \boldsymbol{SG}$, and $\boldsymbol{S}$ denotes a summing matrix representing the aggregation structure (groups or hierarchies) of the data. The matrix $\boldsymbol{G}$ depends on the reconciliation approach to be used.

The most common approach to obtaining coherent forecasts is known as "bottom-up" forecasting. It involves simply summing the most disaggregated forecasts to obtain forecasts for the other series of the structure. This corresponds to setting $\boldsymbol{G}$ equal to an identity matrix in the right-hand columns, and all zeros to the left. That is,

$$\begin{bmatrix} \tilde{y}_h \\ \tilde{y}_{A,h} \\ \tilde{y}_{B,h} \\ \tilde{y}_{AA,h} \\ \tilde{y}_{AB,h} \\ \tilde{y}_{AC,h} \\ \tilde{y}_{BA,h} \\ \tilde{y}_{BB,h} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{y}_h \\ \hat{y}_{A,h} \\ \hat{y}_{B,h} \\ \hat{y}_{AA,h} \\ \hat{y}_{AB,h} \\ \hat{y}_{AC,h} \\ \hat{y}_{BA,h} \\ \hat{y}_{BB,h} \end{bmatrix}.$$

While this approach has low computational costs, it leads to less accurate forecasts (HYNDMAN *et al.*, 2011), since the bottom-level series are typically noisy. Sales at the item level are more erratic, with more variation and may have insufficient data to construct reliable forecasts (DEKKER; VAN DONSELAAR; OUWEHAND, 2004).

Two other traditional approaches to obtain coherent forecasts apply only to hierarchical time series: top-down and middle-out. Top-down starts the forecast from the total and then divides it into different levels, commonly, by historical proportions. However, historical proportions might change over time, leading to less accurate forecasts. The middle-out approach forecasts some appropriate middle level of the hierarchy. It sums the forecasts to generate predictions for the higher levels and disaggregates the forecasts to obtain predictions for the lower levels.

Hyndman *et al.* (2011) introduced a fourth approach, the optimal reconciliation method, later refined by Wickramasuriya, Athanasopoulos and Hyndman (2018). Unlike the other methods, this approach considers the structure of the groups or hierarchies, using more information than the traditional methods, and therefore tends to be more accurate. In this approach, the matrix $\boldsymbol{G}$ is estimated by minimizing the forecast error variance of the coherent forecasts.

The weights that form the matrix $\boldsymbol{G}$ depend on the hierarchical structure and are the result of a linear regression problem. Hyndman *et al.* (2011) called it "optimal" because the difference between the reconciled forecasts and the incoherent base forecasts is minimized. Wickramasuriya *et al.* (2018) showed that the trace of the forecast error covariance matrix is minimized using optimal reconciliation, provided the base forecasts are unbiased, and so the method is now called "MinT" or minimum trace. A variation on MinT is WLS (weighted least squares) which uses only the diagonal of the covariance matrix, setting all other values to zero. In the present paper, we apply both MinT and WLS approaches to obtain reconciled forecasts.

4.3 Material and methods

The steps we followed in our empirical evaluation were: (1) selecting the grouping criteria; (2) setting the forecast horizon; and (3) selecting the time series forecast method.

All data analysis, forecasting and output was conducted using R (R CORE TEAM, 2018) with the following packages: tidyverse (WICKHAM, 2017), forecast (HYNDMAN *et al.*, 2018; HYNDMAN; KHANDAKAR, 2008), hts (HYNDMAN; LEE; WANG, 2017), lubridate (GROLEMUND; WICKHAM, 2011), and tsibble (WANG; COOK; HYNDMAN, 2018). The code for reproducing our analysis is available at request.

4.3.1 Data and selection of aggregation strategy

We used a dataset provided by a major manufacturer of electrical components that has 10 factories in Brazil and is present in more than 120 countries. Our intention (like DEKKER; VAN DONSELAAR; OUWEHAND, 2004; and NENOVA; MAY, 2016) is to test the accuracy of our proposal using complex real data and not merely on simulations that may not portray the challenges and circumstances that are incorporated on real data. The dataset consists of the history of sales from one company of electrical components (plugs and light switches). Each record is a stock-keeping unit (SKU) sold from the industry to the channels located in São Paulo, Brazil, comprising over seven years of sales records, from July 2010 to September 2017. No individual customer information and transactions were used. The database refers only to channels (points of sale) purchases and their characteristics (the type of channel, size, revenue, etc.). The total number of observations (channels purchases) is 13,719.

We constructed three different structures by aggregating product categories, channel characteristics, and geographic considerations (Table 6). The 339 products were categorized into three types: plugs, light switches, and others. The database comprised 220 points of sale, which were categorized into four types: distributor, retail, warehouse, and others. Finally, the geographic hierarchy comprised the city zones (center, east, west, north, and south), and 54 of the 96 city districts (the missing districts had no sales records during the database time frame). At the most disaggregated level, there were 339 x 220 = 74,580 product-store combinations. However, such disaggregated data are too noisy to be useful, so we do not consider them.

Table 6 - Geographic hierarchy, product categories, and channel categories

Levels	**Series**
Geography	
Total	1
Zones	5
Districts	54
Products	
Total	1
Product categories	3
Products	339
Channels	
Total	1
Channel categories	4
Stores	220

Source: The authors (2019)

The selection of such different hierarchies was based on both theoretical and managerial reasons. For the managers, being able to access forecasts of these levels and characteristics gives them better knowledge to plan budgets and strategies with partners. Theoretically, it can help to give insight on which marketing mix variables can be used as grouping criteria, improving the accuracy of sales forecasts. Our goal was to evaluate the forecast accuracy of reconciled forecasts generated from the hierarchies. To achieve such a goal, we used different marketing variables, combining them in a grouped time series forecast.

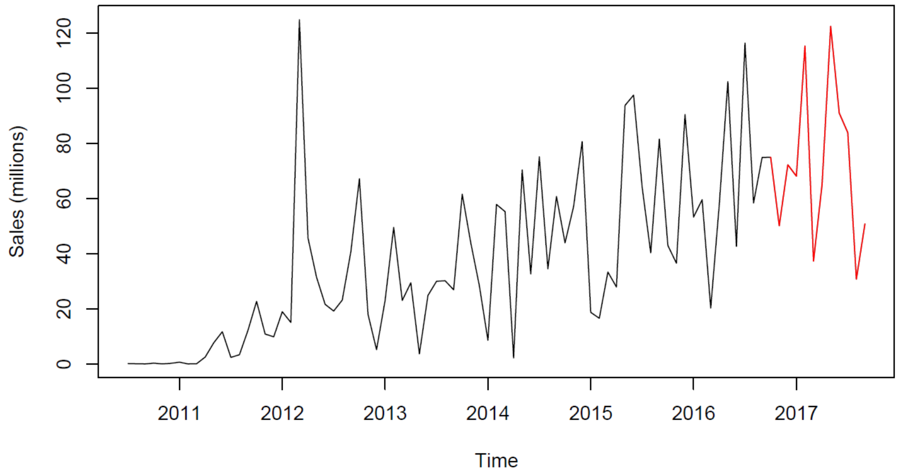
The first step was to combine different data sets provided by the company, including: (1) sales made by the company’s distribution center and direct by the factory; (2) information about the channels’ characteristics; and (3) location data of the channels. Next, we filtered the data corresponding to the sales of electrical components made for the city of São Paulo. The categories of products and channels were created based on the descriptions provided by the data sets.

Further aggregation of the products, channels and geographic regions was carried out so that the amount of missing data would not make the forecasting infeasible. According to Seaman (2018), handling missing data is one of the most important parts of the analysis of forecasting accuracy. While imputation methods are available, they do not allow the accuracy of the actual sales to be measured.

The literature on hierarchical forecasting suggests how to distribute the forecasts throughout different levels (HYNDMAN; ATHANASOPOULOS, 2018). Historically, the major theoretical question has been whether to forecast aggregated data and divide it among different levels or to produce disaggregated forecasts that are then added up. These two common strategies are known as "top-down" and "bottom-up", as reviewed in section 4.2.2. In this paper, we apply an alternative approach proposed in forecasting research, known as "optimal reconciliation" (HYNDMAN *et al.*, 2011), also reviewed in section 4.2.2.

We executed five different analyses. First, we estimated the forecasts for the aggregates based on the three grouping variables (geography, products, and channels) separately, as a hierarchical time series. Next, we combined the product and geographic hierarchies. Finally, we integrated all three hierarchies. In each case, we divided the time series into a training set of 75 months for model estimation, and a test set of 12 months for post-sample evaluation, as Figure 11 shows for the total series. Thus, the maximum forecast horizon we consider is 12 months.

Figure 11 - Total sales



Source: The authors (2019)[2]

4.3.2 Time series method

Exponential smoothing models were used (HYNDMAN *et al.*, 2008), and the selected model was ETS(A,N,N). The ETS(A,N,N) model is a non-stationary linear state-space model with additive homoscedastic errors, no trend, and no seasonality. The model is defined via an observation equation that establishes the relationship between unobserved states and the observations,

$$y_t = l_{t-1} + \varepsilon_t \ , \tag{10}$$

and a state equation, that establishes the evolution of states over time,

$$\mathrm{l}_t = \mathrm{l}_{t-1} + \alpha\varepsilon_t \ , \tag{11}$$

Where l_t represents the anticipated sales volume and $\varepsilon_t \sim \mathrm{NID}(0, \sigma^2)$, represents the unanticipated sales volume. The latter is usually assumed to be from a Gaussian white noise process with variance σ^2. The smoother parameter α denotes the degree of change in the sales' volume over time. The forecast can be expressed as a linear function of historical observations.

The algorithm implemented by the forecast package selects the best model according to Akaike's Information Criteria (AIC); optimizes the parameters using maximum likelihood

[2] Training data are shown in black and test data in red.

estimation; forecasts using the selected model for the nominated horizon; and calculates the associated prediction intervals (HYNDMAN; KHANDAKAR, 2008).

The name "exponential smoothing" is related to the fact that weights are given to observations in the forecast function decrease exponentially with time (the older the observation is, the less weight it will have). ETS models were fitted to the series at each node in each of the five grouping structures. These were then reconciled using the WLS and MinT methods as described in section 4.2.2.

When applied on a large scale, a natural concern is the computational costs. Because our approach separates the generation of forecasts from their reconciliation, it is easily parallelizable. Forecast for all nodes can be computed in parallel, and the reconciliation step can then be calculated using sparse matrix algebra. The computational time for generating the forecasts is much more substantial than the time required to reconcile them. We have successfully reconciled forecasts from 5 million nodes in less than 30 seconds using a standard laptop computer.

To reduce the computational time required to generate individual forecasts, data for similar products or regions can be clustered. This also helps when the time series contain numbers close to zero or short time series, that would make forecasts infeasible. Our approach also uses programming languages that are open and accessible for use by any organization. It also uses information that is easily available to companies and is generalizable and easy to implement.

4.4 Results

The hierarchical and grouped time series forecast accuracy measures are shown in Tables 7 - 9. The various grouping structures make little difference to the accuracy of total aggregate sales but can make a substantial difference to some of the disaggregated forecasts. To evaluate the different reconciliation methods and grouping criteria we used the scaled error proposed by Hyndman and Koehler (2006) in order to remove the effect of the scale of the series at each node.

MASE is an evaluation measurement that can be used for comparing forecasts with different horizons, time frames or even different time series (HYNDMAN; KOEHLER, 2006). Percentage errors are also unit-free but have the disadvantage of not being useful when observations are zero or close to zero. In our study, the horizon was kept fixed, but the different grouping criteria created different time series to be compared.

For seasonal time series, MASE is defined by Hyndman and Athanasopoulos (2018) as:

$$\text{MASE} = \text{mean}(|q_j|) \tag{12}$$

where $q_j = e_j/Q, e_j$ is a forecast error,

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|$$

is the scaling factor computed on the training data, and m is the number of observations per year. MASE will return a value smaller than one if the out-of-sample forecast error is smaller than the in-sample one-step forecast MAE of the seasonal naïve method and will be greater than one otherwise. The values of MASE for the different grouping criteria and methods in our study can be seen in Table 7. Each level was divided into a training set with 75 months of data for model estimation and a test set with 12 months, for post-sample evaluation.

Table 7 - MASE by level

Level	MASE
Products & geography	
Total	0.9
Products	0.83
Zones	1.01
Districts	1.27
All groups	
Total	0.9
Products	0.83
Zones	1
Districts	1.27
Channels	0.81
Geography	
Total	0.9
Zones	1.02
Districts	1.31
Channels	
Total	0.9
Channels	0.82
Products	
Total	0.91

Products	0.84

Source: The authors (2019)

Table 7 also gives more detailed information about the MASE at each level. At the aggregate level, "total", forecast accuracy is higher than in each of the disaggregated levels, only when geographic considerations are used. On the hierarchical and grouped structures, the level of products and channels has smaller errors than at the most aggregate level. If the interest is on forecasting by product category, either of the grouped structures has a slightly better performance. If the aim is to forecast only the total, aggregate level, adding product category to the hierarchical structure was not enough to reduce the error. However, if the interest is to forecast by channels or geographic areas, the grouped structure with all the marketing mix variables considered gives better performance for the level of the channel but requires additional information to be collected, stored and analyzed.

Table 8 - MASE by products, channels, and geography

Districts	**GTS**		**HTS**		
	Products & geography	All groups	Geography	Products	Channels
Total	0.9	0.9	0.9	0.91	0.9
Products					
I	1.07	1.07		1.07	
O	0.44	0.44		0.49	
T	0.84	0.84		0.85	
Channels					
DT		0.31			0.28
OT		1.82			1.69
RT		0.81			0.83
WS		1.43			1.44
Zones					
Center	0.47	0.47	0.49		
East	2.06	2.05	2.05		
North	1.15	1.15	1.16		
South	2.79	2.77	2.81		
West	1.29	1.25	1.32		

Source: The authors (2019)

Overall, the structure with the best MASE values is the one that combines all the groups. Tables 8 and 9 give the MASE values by the group at the lowest level of aggregation. Again, the structure with all the groups provides the best forecasts in most of the nodes. Of all the nodes, 65% (43 of 66) had a better performance with the structure that uses the information of all the groups. The grouped time series forecast with the combined information about product categories and geographic considerations had a better performance in 45% (28 of 62) of the nodes. The hierarchical forecast based on channel characteristics was the best in 3 of 5 nodes (60%). The forecasts based on geography and product categories performed better in 20% (12 of 59) and 25% (1 of 4) of the cases, respectively.

Table 9 - MASE by district

Districts	**GTS**		**HTS**
	Products & geography	All groups	Geography
Água Rasa	0.47	0.48	0.48
Aricanduva	1.31	1.27	1.33
Artur Alvim	1.63	1.63	1.64
Barra Funda	1.06	1	0.95
Belém	6.6	6.59	6.58
Brasilândia	1.75	1.76	1.8
Cachoeirinha	3.34	3.34	3.37
Campo Grande	0.87	0.67	0.8
Campo Limpo	4.27	4.3	4.34
Cangaíba	1.73	1.72	1.77
Capão Redondo	4.95	4.98	4.92
Casa Verde	1.1	1.09	1.1
Cidade Ademar	3.14	2.87	3.26
Cidade Dutra	5	4.94	5
Cidade Líder	2.08	2.07	2.08
Freguesia do Ó	0.51	0.54	0.55
Grajaú			
Iguatemi	0.87	0.84	0.93
Itaim Bibi	12.86	12.81	12.85
Itaquera	0.49	0.48	0.53
Jabaquara	0.98	1.06	1.36
Jaçanã	2.26	2.27	2.3
Jaraguá	0.9	0.91	0.94
Jardim Helena	2.24	2.21	2.41
Jardim São Luís	1.5	1.5	1.52
Mandaqui	1.72	1.72	1.74
Moema	0.67	0.65	0.98
Parelheiros	1.99	2.08	2.4
Pari	36.28	35.96	36.97

Pedreira	0.49	0.46	0.54
Penha	1.82	1.79	1.88
Perus	9.76	9.76	9.76
Pinheiros	4.63	5.08	7.01
Pirituba	1.25	1.25	1.25
Raposo Tavares	4.25	4.26	4.31
República	0.55	0.55	0.56
Rio Pequeno	11.26	11.65	10.76
Sacomã	1.71	1.75	1.68
Santa Cecília	0.51	0.51	0.53
Santana	0.61	0.61	0.65
Santo Amaro	12.48	12.45	12.4
São Domingos	2.9	2.87	3.08
São Mateus	1.6	1.55	1.59
São Rafael	2.08	2.04	2.24
Sapopemba	0.67	0.64	0.76
Tatuapé	3.1	3.67	4.37
Tremembé	1.22	1.23	1.28
Vila Curuçá	0.69	0.58	0.76
Vila Leopoldina	0.38	0.36	0.29
Vila Maria	2.04	2.03	2.04
Vila Mariana	20.96	20.93	20.84
Vila Matilde	1.67	1.67	1.66
Vila Prudente	1.1	1.09	1.12
Vila Sônia	0.27	0.19	0.26

Source: The authors (2019)

For one of the districts (Grajaú), the MASE was undefined because of all the historical observations on the training set were equal. That is the only circumstance under which a MASE is undefined. When it is necessary to disaggregate the forecast, our results provide evidence that it is best to add more information in a grouped structure. The results also suggest that, when forecasting sales, geographic considerations are important to improve accuracy.

4.5 Discussion and conclusion

Even though researchers have been studying the aggregation issue for a long time, there is still no clear consensus about the criteria to determine its levels (ABHISHEK; HOSANAGAR; FADER, 2015). The aggregation criteria studied to date are either based on the similarity of the time series using clustering methods (see FLIEDNER; LAWRENCE, 1995 and ZOTTERI; KALCHSCHMIDT; CANIATO, 2005), by temporal aggregation

(KOURENTZES; ROSTAMI-TABAR; BARROW, 2017), or product, using the volume of sales. However, comparing criteria based on different marketing mix variables is a key question that the theory should answer, due to its relevance for management and, more specifically, marketing practice. In this paper we have compared different structures of hierarchical and grouped forecasting, using different marketing variables.

This paper contributes to the literature in a few ways. First, we recommend the variables necessary to be collected and managed to allow predictive analysis. The results bring insights into which marketing mix variables are valuable to predict sales more accurately and subsequently plan actions. Second, our empirical application provides evidence that it is best to add more information about the marketing mix in a grouped structure, rather than in a hierarchical structure. Combining the information of product categories, channel characteristics, and geographic considerations led to more accurate forecasts than choosing only one criterion. The results also suggest that geographic considerations are most important to improve the accuracy of sales forecasts. Also, it suggests that grouping time series based on marketing variables improves accuracy more than considering the hierarchical structure.

We have demonstrated how forecast reconciliation can be used in a large marketing application involving several types of aggregation criteria. The resulting forecasts have the advantage of coherency and accuracy, thus providing retailers with substantial useful information. The channel characteristics and geographic considerations added useful information when considered in a grouped structure, improving the forecasts at the product level. The information about different product categories or channel characteristics alone were not enough to improve the accuracy of the forecasting models, but when added to other information regarding the grouping structure, it led to the better sales forecast.

The forecasting strategy we propose allows firms to communicate consistent information at all hierarchical levels and departments, leading the efforts in the same direction. The system scales easily and provides consistency for retailers and manufacturers. Many retailers have thousands of stores spread geographically, while manufacturers may have thousands of business partners in different locations selling their products and requiring tailored forecasts. Our strategy provides a coherent forecast with almost no human effort, by using an optimal reconciliation approach. It can be adapted for goals, horizons, update necessity and levels of aggregation.

One clear limitation of our research is the lack of information about other marketing variables, such as promotion or price strategies. This can be addressed in future research. Also, given the importance of geographic disaggregation in our analysis, future research can explore greater geographic disaggregation. Location data are now abundant due to mobile devices and applications that store location meta-data. It has been used, for example, in models of market response concerning hotels (ZHANG; KALRA, 2014), gas stations (CHAN; PADMANABHAN; SEETHARAMAN, 2007), fuel adoption (SHRIVER, 2015), drug prescriptions (STREMERSCH; LANDSMAN; VENKATARAMAN, 2013), solar panels (BOLLINGER; GILLINGHAM, 2012), organic products (SRIDHAR; BEZAWADA; TRIVEDI, 2012), and the car industry (ALBUQUERQUE; BRONNENBERG, 2012; BUCKLIN; SIDDARTH; SILVA-RISSO, 2008; NARAYANAN; NAIR, 2013).

Another research opportunity regarding geographic information is to explore the concentration of stores in a specific area. Sales of low involvement and low-cost durables such as plugs, and switches tend to be concentrated in areas of higher retail activity. This is present in our data. Sales are correlated (0.7) to the retail activity index (calculated by the Central Bank of Brazil) of each district of São Paulo. This is consistent with agglomeration theory (for a marketing application and revision see Liu, Steenkamp and Zhang (2018)). This theory states that for consumers it is more convenient when stores from the same category are concentrated on a specific geographic area. This way they can compare alternatives and get more information about products. That should be especially important for products that typically are not bought and do not have reviews shared online. Agglomeration of companies in a certain geographic area will, for that reason, have a positive impact on sales, explaining demand more than the agglomeration of consumers (LIU; STEENKAMP; ZHANG, 2018).

Marketing research efforts have focused on explaining the impact of the marketing mix on market response, and less attention has been given to forecasting accuracy and which marketing instruments can help to improve it. However, accurate forecasts can help plan budgets and production levels, and influence brand image, price perception, customer satisfaction, and many other areas of marketing interest. We hope that the forecast reconciliation tool will help address this issue and prove useful in many marketing analyses.

5 GRAVITATIONAL SALES FORECAST RECONCILIATION (GSFR) APPROACH– AN APPLICATION ON TURKEY AND BRAZIL

Abstract: When organizations plan to enter a new market or to expand their business to new locations, they might face limitations on access to disaggregated forecasts, necessary to make decisions. They might be arriving at a new territory with just a total demand estimation. It needs disaggregated forecasts to plan district-level activities. Or they might be planning to expand sales by evaluating new sales channels or partners at districts where it currently does not sell. Or a company might need to predict changes in future sales when a new channel opens at a district/region. Currently, you need to access actual sales data on a disaggregated level or at least have historical sales proportions to estimate these forecasts. We propose in this paper a new approach to disaggregate a total estimation in lower-level forecasts for geographical areas. It applies a gravitational model that it is estimated only with a total forecast and publicly available data. Our approach does not use disaggregate sales information and outperforms or have at least a comparable performance to benchmark approaches that have access to that information.

Keywords: sales forecasting; marketing; hierarchical time series; reconciliation approaches; gravitational model

5.1 Introduction

While forecasting methods are data-hungry, with improved accuracy if it estimates based on bigger time-series, historical sales data is scarce. Despite their availability are increasing, time-series are mostly recent (not many data points), depending on the time aggregation one works with. This is a major limitation in sales forecasting for business owners, especially those companies which are new in a market. Many companies usually do not have access to final consumer data, unless they engage in agreements with retailers or other channel members. Manufactures of products that are not from the packaged goods sector in advanced economies have an even harder task to predict sales accurately since most of them do not have access to scanner data (HANSSENS, 1998). In those cases, they need to make the best use of the limited available data to make decisions regarding budget allocation, selection of most profitable partners, and members of their channel strategy.

To address this limitation, we propose a new data-driven hybrid method, combining a deep learning technique adapted to time-series forecasting with a modified version of the Huff gravity model that works with longer (i.e. 7 years of monthly sales) or shorter (i.e. 3 years) time-series. We validate our model with two case studies, in two different countries (Brazil and Turkey) from two continents, two manufacturers with different product categories, and two different sizes of sales history datasets. Our research provides a forecasting tool suitable for, however not exclusively, manufactories. Our method can be embedded in production planning decision support systems and help companies devising their future market strategies.

The Huff gravity model is basically designed to predict the market share of facilities and it is incapable of forecasting future sales amount as the time series models do. To predict sales volume, our proposed model combines a deep learning technique, long short-term memory (LSTM), with a modified model inspired by the Huff gravity model.

The total (most aggregate level) monthly sales were forecasted and then reconciled to lower levels forecasts with a top-down approach. The difference with our strategy is that, instead of using historical proportions as the standard top-down approach, we distributed the total sales among regional units (districts) with a variation of Huff gravity market share model. We compare the results of our approach to the top-down approach using historical proportions and to the optimal reconciliation approach (HYNDMAN *et al.*, 2011) for a dataset with sales of a Brazilian manufacturer and then we apply the same approach to forecast sales of a Turkish manufacturer.

Our research contributes to companies that do not have points of sale historical information by using data that is easily/publicly available to any sector to forecast disaggregate sales. We also contribute to the forecast reconciliation theory by proposing a hybrid model that integrates a deep learning and a gravitational model. Our model is easy to implement, and since it relies on open source tools and publicly available data, it does not impose a large investment on companies. As in Dekker, Van Donselaar, and Ouwehand (2004), Hanssens (1998), and Williams and Waller (2011), we also contribute to the forecasting literature using easily accessible data, validating our approach with companies' proprietary data.

Using our proposed reconciliation approach, where the inputs are mostly publicly available datasets, business owners can (1) distribute the forecasted aggregated sales into smaller regional units such as districts, zip codes, etc. when there is no fine-grained sales information available; (2) predict the future regional sales/market-share based on the new POIs built across different regions in a city; and (3) estimate the market potential of the new regions where they are planning to enter the market and have no historical sales data.

This paper unfolds as follows. Section 5.2 briefly explores the literature to which we contribute. Section 5.3 documents the data collection process for this paper. Section 5.4 details our empirical strategy, gives the notations of the methods applied, and citation of the programming language and packages used. Section 5.5, describes the accuracy of the different methods applied and discusses the results. We conclude in Section 5.6 by describing some limitations, discussing implications of the results, and presenting suggestions for future research.

5.2 Related Literature

Most companies rely on models that are easier to implement than those proposed in the marketing literature (MEYER; HUTCHINSON, 2016; REISS, 2011). Forecasting is critical for decision making in production, sales, marketing, finance, and logistics. Even in such a vital tool for budget planning and allocation, companies still rely on their own experience and familiarity with methods (CANITZ, 2016; FILDES; MA; KOLASSA, 2018) rather than the most advanced models in the literature.

Models to study sales can be classified by demand system, aggregation level, previous or post product launch models, or by goal. Seminal literature on forecasting also relates the choice of method to the amount and nature of the data available, which defines if the method will be qualitative (or judgmental) or quantitative (ARMSTRONG, 2001). Our research focuses on quantitative methods, for post-launch products, with a forecasting goal, and at different aggregation levels.

Regarding the time series forecasting literature, aggregation is a key question. Fliedner and Mabert (1992) investigated two parameters for constrained forecasting: product group size and appropriate grouping criterion. The authors clustered the products using different criteria and found that the size of family groups (the number of component items) is not of great importance. However, they showed that the homogeneity of parent groups (correlation among group members) is important, although no indication of the direction of this effect was provided. They also observed that criteria used to form family groups which are based upon demand volume (e.g. unit volume and dollar volume) enhance hierarchical forecast system performance. Later, Fliedner and Lawrence (1995) expanded these findings by examining alternative forecast methodologies and techniques for forming family groups within hierarchical systems. The three factors of cluster technique, grouping criterion, and number of parent groups resulted in insignificant performance differences. The level of sophistication utilized in the group determination process of hierarchical forecasting systems was not responsible for improved forecast performance.

At a more aggregate level, it is easier to distinguish the seasonality from the randomness. Dekker, Van Donselaar, and Ouwehand (2004) use the aggregated data to determine the seasonal indices and then forecasts the non-aggregated level with these indices. The aim of the authors was to understand the added value of product aggregation and to keep the forecasting methods simple and generic so they would be easier to develop, implement, and maintain in practice. Besides product aggregation, Dekimpe and Hanssens (2000) expected further refinements in terms of temporal aggregation in the future. The authors stated that previous studies used quarterly or monthly data, and they perceived a recent movement for weekly aggregations. However, there is no clear consensus on which is the optimal level. Kourentzes, Rostami-Tabar, and Barrow (2017) discuss on their paper if it is better to use an optimal level or multiple temporal aggregations, favoring the last strategy.

Dekker, Van Donselaar, and Ouwehand (2004) suggested that a better way of forming product families can help to fully exploit the possibilities of criteria for aggregating the time series. Trying to answer this question, Silveira Netto, Hyndman, and Brei (2019) tested different strategies of forecast reconciliation by region, product category, and channel type. They concluded that a grouped time-series structure that combines all the information about the marketing mix variables is a better strategy. They also highlighted the importance of information about location (regions), that improved forecasting accuracy more than the other aggregation criteria. However, the authors did not analyze whether the concentration of retail activity in different neighborhoods may help to improve forecasting accuracy.

Chan, Padmanabhan; and Seetharaman (2007) state that location is determinant for retail performance and competition. While analyzing the impact of distribution intensity on car sales, Bucklin, Siddarth, and Silva-Risso (2008) demonstrate that store location and its distance to consumers' homes play an important role in sales. This, however, is not new. Lesage (1999) states that location and distance are important forces that influence human geography and market activity. In diffusion literature in marketing (e.g. GARBER *et al.*, 2004) the correlation between location and word of mouth spread is well established.

According to Bronnenberg and Sismeiro (2002), sales correlation between geographical markets is probably more common than literature leads to believe since very few studies explicitly consider spatial dependencies. The authors use this correlation with near territories to predict performance in markets where there is little or no data available.

Spatial models applied to marketing already demonstrated that spatial effects influence, for example, retail type choice (GONZALEZ-BENITO; MUNOZ-GALLEGO; KOPALLE, 2005). In the Japanese market, individual car choice was better explained by networks of customers based on geographical locations than by demographic information (YANG; ALLENBY, 2003). Spatial models were also applied to explain sales in different sectors, such as hotels (ZHANG; KALRA, 2014), gas stations (CHAN; PADMANABHAN; SEETHARAMAN, 2007); alternative fuel adoption (SHRIVER, 2015); medicine prescriptions (STREMERSCH; LANDSMAN; VENKATARAMAN, 2013); solar panels (BOLLINGER; GILLINGHAM, 2012); organic products (SRIDHAR; BEZAWADA; TRIVEDI, 2012); and car industry (ALBUQUERQUE; BRONNENBERG, 2012; BUCKLIN; SIDDARTH; SILVA-RISSO, 2008; NARAYANAN; NAIR, 2013).

By those studies, it is assumed that the answer of individuals near each other are more related and similar than those from individuals more distant (BRADLOW *et al.*, 2005). However, not only individual consumers have their choices influenced by location. Bronnenberg and Mahajan (2001) show that market-shares are different among regions and that it has a spatial structure due to the non-observed behavior of points of sales that are spatially spread, in territories. According to the authors, that happens because points of sale act locally, giving more attention to their own territories.

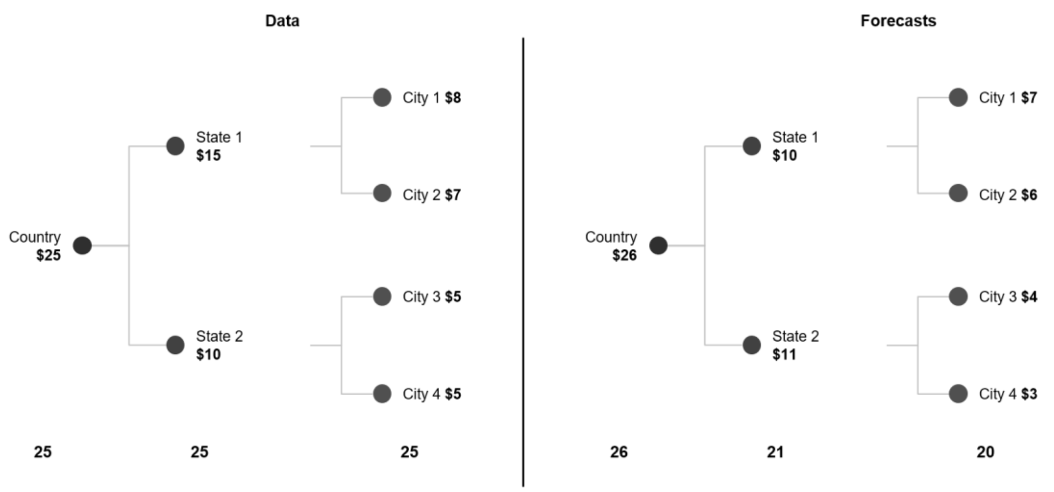
One of the most used models to estimate spatial attractivity is the Huff model (HUFF, 1964). This model uses two determining factors for the attractivity potential of a store: the size of the store (directly proportional); and the distance between stores and consumers (inversely proportional). Besides its original application, the model has been extended by several studies. Authors have included other variables to the model, such as price, service level, opening hours and brand image (e.g. TELLER; REUTTERER, 2008).

We do not study the diffusion of innovations or have access to individual location data. Yet, we focus on the spatial relationship that can be explained by a different theory, agglomeration, as in Liu, Steenkamp, and Zhang (2018). The concentration of stores from the same sector in the same neighborhood makes it easier for clients to compare the alternatives and make better-informed decisions. This is especially true for low-involvement products, those that are not purchased frequently (such as durable goods), and those not traditionally purchased online. Examples of that are home refurbishment materials, like light switches or house paint products. For that reason, we expect that the concentration of stores will have a correlation with sales (LIU; STEENKAMP; ZHANG, 2018).

Fildes, Ma, and Kolassa (2018) state that current location models are seeing their credibility diminishes because of changes in consumer behavior, and that so far space interaction models (like gravitational models) have better explanatory power than predictive performance. To build new models using location data that are relevant is a challenge. The authors also suggest that forecasting models could be used to identify stores to be closed. Consequently, to know whether the concentration of stores improves forecasting accuracy is relevant to theory and practice. This is especially true considering that manufacturers and retailers' practice is disaggregated, while their forecasts are usually done at the most aggregated level.

Geographical information is naturally structured in hierarchical levels, with different aggregation levels (i.e. country, states, cities). If forecasts are needed in lower levels of aggregation, data can be divided into those different levels and forecasts can be estimated. However, differently from the data that adds up following that structure, forecast do not (Figure 12).

Figure 12 – Data and forecasts in a hierarchical structure



Source: The authors (2019)

To solve this issue, forecasting literature proposes what is called reconciliation approaches. Forecasts can come from any appropriate model, created independently for each node ("base" forecasts), following the general notation (13).

$$\tilde{y}_h = R\hat{y}_h \tag{13}$$

where R is the reconciliation matrix, decomposed as R = SP. S is the summing matrix that represents the aggregation structure, it assigns each forecast to its respective group in the structure. P is a matrix with the weights of each forecast and depends on the reconciliation approach to be used.

There are four different established approaches to make the forecasts coherent. The top-down approach is done by forecasting the most aggregate level and dividing it to the lower levels using historical proportions. However, historical proportions might change over time, leading to less accurate forecasts (HYNDMAN *et al.*, 2011). The bottom-up approach is the opposite, forecasts the most disaggregate level and sum up, however, since bottom level series are typically noisy, they lead to less accurate forecasts. Middle-out approach forecasts any middle level of the structure and then sum it to the upper levels and divide it to the lower levels. It shares the same problems as the other two approaches.

Hyndman *et al.* (2011) propose a fourth approach, optimal, that forecasts all levels and reconcile them using a linear combination giving weights to each forecast unit (nodes of the hierarchy). It considers the structure of the groups or hierarchies and tends to be more accurate. It is called "optimal" because the difference between the reconciled forecasts and the incoherent base forecasts is minimized.

However, all of these approaches need the actual sales data or at least the historical proportions of each level. When using a geographical hierarchy, companies not always have this information available, only a total aggregated forecast. This study aims to propose a new reconciliation approach that will use a gravitational model to overcome this difficulty, disaggregating a total forecast to lower levels, without using actual data or historical proportions. Huff gravity model is designed to predict market share at the individual (merchant) level, but in our paper, we are using it at the aggregate level for each geographical region. These proportions can be used to distribute an aggregate sales forecast. In forecasting terms, this is a top-down reconciliation approach with a different strategy to distribute the forecast to lower levels, or, as we are calling it, a gravitational sales forecast reconciliation (GSFR). The next section will present the data we had available and later we detail our empirical strategy.

5.3 Data

In this study, we make use of various datasets including public and private datasets easily available to most organizations. In Brazil, we concentrated the study in the city of São Paulo, and for the case of Turkey, we focus on the Istanbul metropolitan area. We mainly utilized 5 datasets, 3 for Brazil and 2 for Turkey, besides demographical and geographical (i.e. shapefiles) information available on São Paulo's city hall website and retail activity index of São Paulo and Istanbul by district was also collected. Table 10 summarizes the data sources used.

Table 10 – Datasets

Dataset	Description
São Paulo's city hall website	Demographic (population, human develop index) and geographical (i.e. shapefiles) information (2017 estimation)
Brazil's central bank	Retail activity index at districts level (2016)
Brazil's sales time series	Sales records from a manufacturer of plugs and light switches (2010-2017)
Google Places API	~275,000 Points of interest (POIs) in São Paulo (in Jan. 2019), including users reviews
ANAMACO's data set	National Association of Construction Materials Business people: survey describing retailers and wholesalers in each SP district in 2018 (~13k stores)
Turkey's sales history	Sales records from a Turkish company that produces house paint products. Istanbul districts level sales (2015-2017)
Turkey's economic indicator	GDP proxy 2016 Insurance sales of a large insurance company at districts level (0.992 correlation with GDP at city level)
Here.com	Points of interest (POIs) in Istanbul 2016 Q1 updated POIs in 16 groups of POI types, such as Shops, Hospitals, Financial institutes, Entertainment, Education, etc. Includes more than 386,000 POIs
Distances	Distances are replaced with travel times using public transportation, collected from Google Distance Matrix API

Source: The authors (2019)

In Brazil, we used sales records from a manufacturer of plugs and light switches, Google Places data, and retailers and wholesaler's information from ANAMACO's data set. In Turkey, we used sales records from a Turkish company that produces house paint products, and Here.com location data.

Sales records from the Brazilian company consists of stock-keeping units (SKU) sold from the industry or the distribution center to stores located in the city of São Paulo. It comprises sales records from July 2010 to September 2017. No individual customer information was used.

The database refers only to stores' purchases and their characteristics (type of channel, size, revenue, etc.). Turkish company sales records are a smaller data set compared to the Brazilian, consisting of records of SKUs sold from 2015 to 2017, in Istanbul. The data set has information about each SKU sold to which store, from which brand, in which district, in a similar format to the Brazilian company.

The information about points of interest (POIs) for Istanbul was collected at Here.com, which was published quarterly from the first quarter of 2015 to the end of the first quarter of 2016. This dataset includes 16 types of points of interest, namely: financial institutes, educational institutes, business centers, entertainment places, shopping places, community service centers, restaurants, hospitals, parks, travel destinations, parking lots, auto services, transportation hubs, and level of access to railroads, sectional highways, and major highways.

For São Paulo case, we utilized Google Places API to collect the points of interest information. It represents about 271 thousand unique POI information which was collected in January 2019. Our goal was to extract all existing places in São Paulo that were available, at the time, on Google Places API.

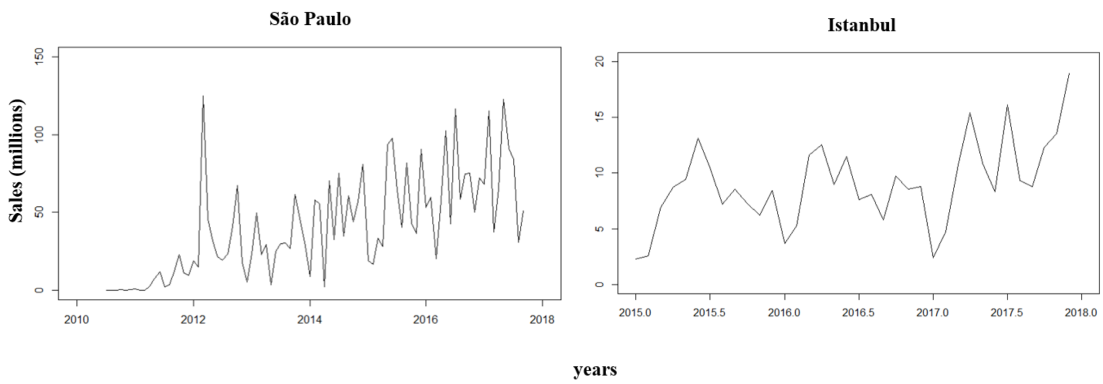
This data collection was performed in 3 steps: 1) we set the strategy to split the city; 2) wrote the code to collect the data; 3) and collected the data itself using the Google Places API. The collected data includes 89 unique place categories, where we used 13 types which were more relevant to our study topic and also consistent with the case of Istanbul. These types are shopping malls, tourism and travel agencies, entertainment places, financial institutes, parking lots, restaurants, courthouses, lawyers, local government offices, lodging, electronics store, police, beauty services. The code to collect and prepare the data was written in the R programming language, using googleway package (COOLEY, 2018). All codes written for this paper are available on a GitHub repository, that can be provided at request.

Finally, ANAMACO stands for National Association of Construction Materials Businesspeople, and it represents the companies of this economic sector. One of its activities is to carry out a nationwide survey about the construction materials sector. The data set we used is from the survey released in June of 2018, where each row represents a company of the sector. The results were bought by the Brazilian company and shared with the authors. Further details on the data sets and their preparation can be found in Appendix A and Appendix B. The next section will detail the empirical strategy followed to build and test GSFR (gravitational sales forecast reconciliation).

5.4 Empirical strategy

For our empirical strategy, first, we combined and cleaned data from various sources, as described in section 5.3. The data came in different files containing each year, or different sources (sales from distribution centers or directly from factory). These files needed to be combined and the variables names changed. We also checked the variables` formats and missing data. Finally, we aggregated the data using geographic considerations and its hierarchical structure and transformed it into time-series format. For São Paulo, Brazil, sales records were from July 2010 to September 2017. For Istanbul, from January 2015 to December 2017. Figure 13 shows the time-series after transformations were made to ensure we met the requirements of a non-disclosure agreement signed by the authors with the data providers.

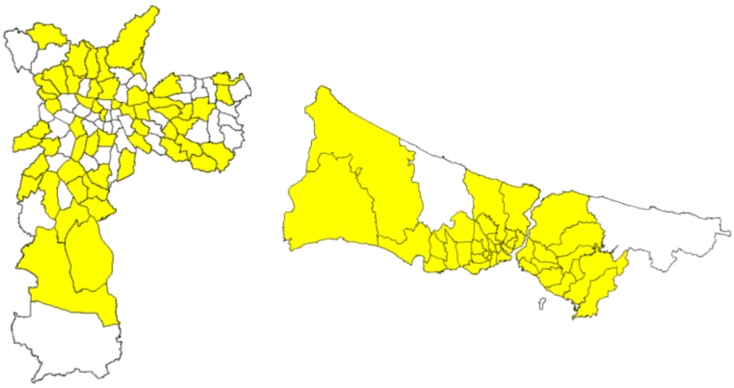
Figure 13 – São Paulo's and Istanbul's time series



Source: The authors – R output (2019)

The geographical hierarchy of the Brazilian data comprised of 3 levels. Total level with city sales, 5 zones (center, east, west, north, and south), and 54 of the 96 city districts of the city (the missing districts had no sales records during the database time frame). The Turkish data had a total level for the city and a lower level with 38 districts, however, we considered only 36 from which we had data to estimate the gravitational model. The districts considered are in yellow in Figure 14. We provide further explanation on Turkish data cleaning process on Appendix C.

Figure 14 – Districts considered in our analysis



Source: The authors – R output (2019)[3]

After that, we tested different methods and forecasting horizons before we proceed with the reconciliation approaches tested. This was necessary to decide which method to apply to create the independent base forecasts. We considered the as competing methods standard time-series methods, such as ARIMA, exponential smoothing, random walk, and trigonometric Exponential Smoothing. Long short-term memory (LSTM) was the selected method for both data sets after we ran the total (most aggregated level) forecast for each data set in different horizons of forecasts. For comparison, the current error rate for the Turkish data provider is of 50%, no better than a coin toss. Table 11 shows the results of different methods applied to the most aggregated level (total), for a forecast horizon of three months. Although we have the results of each method for different horizons, we are illustrating only the horizon selected, for concision.

Table 11 – Methods compared

		MAPE (Mean Absolute Percentage Error)		**RMSE ($ millions)**	
		Istanbul	São Paulo	Istanbul	São Paulo
LSTM	Long short-term memory (deep learning method)	12.8	26.5	3.0	11.3
ARIMA	Integrated autoregressive and moving average	29.6	77.6	5.8	34.1
STLF	Forecasting using decomposition	35.3	87.7	6.0	39.6
ETS	Exponential smoothing	36.9	78.5	6.5	35.3

[3] São Paulo's districts are on the map at the left-hand side and Istanbul's at the right-hand side

MAPA	Multiple Aggregation Prediction Algorithm	37.8	76.5	6.6	34.3
NAÏVE	Last month value repeated	38.9	94.6	6.8	42.0
TBATS	Trigonometric Exponential Smoothing	41.9	72.9	7.4	32.5

Source: The authors – R output (2019)[4]

Table 12 shows the results of different horizons, using LSTM, for both datasets. For the Brazilian case, the error metrics had very similar performance no matter the horizon. The difference in MAPE using 3 months horizon and 12 months horizon was of 4.7 p.p, However, for the Turkish data, this difference was of 21.4p.p. For that reason, we based our decision on the Turkish case. Three months ahead is the time horizon that has the smallest error for the Turkish data. Another reason to select this horizon is that, while discussing early results with the data providers, this horizon was stated as one of the feasible to plan and execute actions. Smaller horizons (one or two months) would not allow time to act based on the forecasts. The test set is from one to six months and then 12 months, in a rolling window (the rest of the data is used as a training set). Again, although we have the results of each horizon in different methods, we are illustrating only the model selected, for concision.

Table 12 – Forecast horizons

	MAPE (Mean Absolute Percentage Error)	
	Istanbul	São Paulo
1 month ahead	12.9	15.6
2	13.7	15.8
3	12.8	26.5
4	22.3	23.7
5	30.3	21.6
6	23.5	22.8
12	34.2	21.8

Source: The authors – R output (2019)

[4] Results for 3 months horizon. Arranged by the Istanbul column, from lower to higher error. 1 Brazilian Real = 0.26 American dollars. 1 Turkish Lira = 0.17 American dollars. Exchange rate of July 3rd, 2019.

After making the decision regarding the method and the sizes of training and test sets, we turned to model estimation. We divided the companies' time series into a training set of 32 months (Turkey's data) and 84 months (Brazil's data) for model estimation, and a test set of 3 months for post-sample evaluation.

A deep learning technique, long short-term memory or LSTM, for short, was the method selected. LSTM models were fitted to each time series (node or base forecasts) of each level of aggregation. Then the forecast of each of those aggregations was made and reconciled using different approaches.

The LSTM models were implemented using the R programming language and the algorithms of the keras package (JJ ALLAIRE; CHOLLET, 2019). For that, we first adapted an implementation of the same algorithm from a Python application and tested with another dataset. Once we made sure we were getting the same results in both programming languages, we applied the model to the datasets of the present study.

First, we normalize the data and prepared it before we inputted into the neural network. This preparation consists of creating labels and a 3D array with the number of time steps, number of samples and number of features. After training and test sets are in the right format, we set the model. We used a keras sequential model, with adam optimizer (KINGMA; BA, 2014) and a goal to minimize the mean square error (loss function). This model passes into the neural network a list of sequential data, their labels and information about the shape of the input.

We added four layers of LSTM with 50 units each, followed by four dropouts’ layers with 0.2 rate to avoid overfitting (GAL; GHAHRAMANI, 2016). A dropout rate is a number between 0 and 1 with the proportion of the units to give 0 weight, at each update during training. The first layer had the input shape and three of the four LSTM layers had a setting to return the full sequence. The final (output or activation) layer was a core keras layer of the dense type with one unit. The model was trained with 100 epochs, i.e. how many times the model will run the entire sample (training and labels). The batch size was 32, meaning that 32 samples were used in each gradient update, which is the default size.

The same model was applied to each node of the hierarchical structure. This means that we had 37 base forecasts to run with the Turkish data, and 60 with the Brazilian data. Each LSTM model takes approximately 44.72 seconds to run in a laptop computer (8GB RAM, core i7, 512GB) after the code is written and automated. This time does not take into account the time to write and test the code, or data cleaning, and preparation of the dataset to represent the hierarchical structure.

The next step after all the models are trained is to run the predictions, transform them back (denormalize) and transform to time-series format. After we have all predictions as time-series we can run the accuracy of each forecast without any reconciliation approach and with optimal, top-down, and GSFR.

For the optimal approach, we estimated independent LSTM models for all the nodes and then adjusted it with the weight calculated by the recursive algorithm (HYNDMAN *et al.*, 2011). To estimate these weights, we apply OLS. We chose to add the optimal reconciliation approach because it considers the hierarchical structure and tends to be more accurate than the top-down approach, configuring a benchmark that is harder to beat. If our model approximates or is more accurate than the optimal, this means we are proposing a strong alternative approach to reconcile forecasts.

To estimate the top-down approach we used only the upper-level forecast (total aggregated sales) and distributed this estimation to the lower levels using historical proportions. This historical proportion was estimated with data from the training set from the same months of the test set but in the previous year. For example, if the test set was from January to March of 2020, we would compute the proportion of sales per district or zones of the sum of the sales between January and March of 2019. That gives us a vector of proportions that sum one, that is multiplied by the total forecast to estimate the lower levels forecasts, following the notation of Hyndman *et al.*, (2011), adapted from the general notation (13) we already described,

$$\tilde{y}_n(h) = SP\hat{y}_n(h) \tag{14}$$

where $P = \left[p|0_{m_k\,X\,(m-1)}\right]$.

Top-down approach uses historical proportions to distribute the forecasts to the lower levels, which can change over time and make estimations highly unreliable. Top-down and optimal approaches also may not be feasible for some situations, since these approaches require either the actual sales of each level (optimal) or knowledge of the proportions of sales (top-down). However, some organization may not have enough data by disaggregate level to run forecasts in the district level. They might also not have historical proportions. This is a reality for many companies that are opening stores in a certain region or are in the early years of operations. To overcome these issues. and considering what agglomeration theory establishes, we applied a modified version of the Huff gravity model to distribute the forecasts to the different districts.

The Huff gravity model estimates facilities market-share by calculating the probability of customers patronization of a particular retail facility that depends on two factors: the facility attractiveness and the distance between the customer's location and the retail facility. Various attractiveness measures have been introduced in the literature based on relevant criteria such as facility size, price level, variety of goods and services offered in those facilities. These models approximate the market share of each facility based on the total number of visits or the total money customers spend.

Market share is calculated as follows:

$$U_{ij} = \frac{A_j^{\theta}}{D_{ij}^{\gamma}} \tag{15}$$

$$f_{ij} = \frac{U_{ij}}{\sum_{j'} U_{ij'}} \tag{16}$$

Where:

- U_{ij}: Utility of customers living in population center *i* from shopping in facility *j*
- A_j: Attractiveness of facility *j* (such as facility area, price level, etc.)
- f_{ij}: The fraction of total transactions (or amount spent) made by population center *i* to be spent in facility *j*
- D_{ij}: Euclidian distance between population center *i* and facility *j*
- $\theta\ and\ \gamma$: model fitting parameters

We took a step further and proposed using the model for region units (here districts), instead of individual merchants. In the Huff model and its variations "attractiveness" measure are mostly related to physical attributes of a merchant, also some studies have considered the variety and price level of goods and services offered in those facilities as measures of attractiveness. Inspired by the idea of Glaeser, Kolko, and Saiz (2001), we contend that the number of amenities, as well as their diversity, could be a region’s attractiveness for citizens. We propose a multiplicative model of POI count and their diversity in business type as a measure of attractiveness for each district. We suggest the following variation of the Huff gravity model for shopping district choice model:

$$U_{ij} = \frac{A_j^{\theta}}{D_{ij}^{\gamma}} \tag{17}$$

$$A_j = \left(\#POI_j\right)^{\alpha} * \left(D(POI_j)\right)^{\beta} \tag{18}$$

$$D(POI)_j \ = \ \sum_k -P_k^j \log log\left(P_k^j\right) \tag{19}$$

$$f_{ij} = \ \frac{U_{ij}}{\sum_{j'} U_{ij'}} \tag{20}$$

Where:

- U_{ij}: Utility of customers living in district *i* from shopping in district *j*
- A_j: Attractiveness of district *j*
- $\#POI_j$: Number of POIs in district *j*
- $D(POI_j)$: Diversity of POIs in district *j* based on their business type using Shannon's Entropy.
- P_k^j: The percentage of POIs of type *k* in district *j*
- f_{ij}: The fraction of total transactions made by residents of district *i* in district *j*
- D_{ij}: The travel time by public transportation between districts *i* and *j*
- *α, β,* and *γ*: Model fitting parameters

Using this proposed model, we aim to predict each regional unit's (district here) market-share from total retail activity by citizens. Data regarding transactions made by customers in the retail sector is scarce and mostly private. We address this problem by assuming that all population are potential customers and contribute to retail activity. For each district, we calculate the proportion of its population that are predicted to visit other districts (in-out flows) and the proportion of people that are predicted to shop inside their own district. Thus, for each district, we will have a new population contributing to retail activity consisting of its own residents that prefer their own district for shopping, plus the people in flowing from other districts for shopping in that particular district. This population is parametric based on the model fitting parameters. We assume the number of transactions is equal to the population volume after implementing the flow modeling. We consider the geographic center of each district as its population center. Therefore, distances between districts are calculated as geo-distance (Haversine distances) between two district centers, and the distance of residents from their home district is calculated as the radius of the largest circle possible inside that district divided by a parameter *t*. To achieve the final proportions to use in the reconciliation approach we apply the algorithm described in section 5.4.1.

5.4.1 GSFR Algorithm

GSFR is estimated by the following steps:

Step 1: Attractiveness measures (as in equation 19)

First, we calculate the number of POIs and their diversity based on the POI types using Shannon's Entropy. Since these two measures have different numerical ranges, we normalize them to the interval [1,10] using min-max normalization. The normalization makes the effect of the attractiveness measure comparable.

Step 2: Utility Calculation

$$U_{ij} = \frac{\left(\#POI_j\right)^{\alpha} * \left(D(POI_j)\right)^{\beta}}{D_{ij}^{\gamma}} \tag{21}$$

In this step, we calculate the utility for the customers from district *i* to make a purchase at district *j*. The utilities are parametric based on the still unknown model fitting parameters: α, β, γ, and *t*.

Step 3: Probability/Proportion Calculation

We turn utilities calculated in the second step into probabilities/proportions, using equation 20 to normalize and turn them into probabilities or proportion of transactions going from each district to other districts.

Step 4: Customer flow Calculation

$$flow_{ij} = f_{ij} * customers_i \tag{22}$$

Then we multiply calculated proportions for each district with the corresponding district’s population to find out the outflows from that district.

Step 5: “New” customers (Potential buyers) Calculation

$$New\_customers_j = \sum_{all\ districts\ i} flow_{ij} \tag{23}$$

Summing up all in-flows from other districts to each district plus the population predicted to visit their own district, the new population (potential buyers) for each district is achieved.

Step 6: Best model fitting parameters calculation

$$\boldsymbol{Predicted}\ Retail\ Activity\ Index = \beta_0 + \beta_1\ (new\ customers) \qquad (24)$$

We use a 4-D grid search on model fitting parameters to find the best combination of positive integer values. Using these numbers, we calculate the correlation between the predicted values for retail activity indicator and its actual values. β_0 and β_1 are obtained from regressing actual retail activity index over potential buyers. The parameter set that gives the highest correlation between the predicted values for retail activity indicator and its actual values is considered as an optimal combination for forecasting.

Step 7: Reconciliation of forecasts

Using equation 14, we use the vector of proportions that outcomes from our model as P, multiplying it by S and the total independent base forecast $\hat{y}_n(h)$, estimated by any model (LSTM in the present study), to distribute the estimate by district level.

All analyses were conducted using R (R CORE TEAM, 2018) and its packages. The main packages used were: keras (JJ ALLAIRE; CHOLLET, 2019), RSNNS (BERGMEIR; BENÍTEZ SÁNCHEZ, 2012), tidyverse (WICKHAM, 2017), forecast (HYNDMAN *et al.*, 2018), hts (HYNDMAN; LEE; WANG, 2017), lubridate (GROLEMUND; WICKHAM, 2011), and tsibble (WANG; COOK; HYNDMAN, 2018). The code for reproducing our analysis is available at GitHub by request to the authors. In the following section, providing the results we will explain how this model works for two cases, São Paulo and Istanbul.

5.5 Results

In this section, we present the results applying the reconciliation approaches tested, and with no reconciliation. Table 13 shows the actual test set data and the estimates of each reconciliation approach by level. These values are a sum of all nodes in each level and a sum of all three months of test data. It shows clearly why reconciliations approaches are necessary since the sum of forecasts in the lower levels of the geographical hierarchy have a substantial difference from the total level forecast if no reconciliation is applied.

Table 13 – Sum of forecasts by level over 3 months test set

BRASIL	**Total**	**Sum of zones**	**Sum of districts**
Actual data	$ 165,500,689	$ 165,500,689	$165,500,689
Optimal	$ 174,495,191	$ 174,495,190	$174,495,191
GSFR	$ 180,704,754	$ 180,704,753	$ 180,704,750
No reconciliation	$ 180,704,768	$ 150,869,858	$ 105,687,198
Top-down	$ 180,704,769	$ 180,704,769	$ 180,704,769
TURKEY			
Actual data	$ 44,799,617.39	NA	$ 44,799,617.39
Top-down	$ 40,933,030.00	NA	$ 40,933,029.25
No reconciliation	$ 40,933,029.27	NA	$ 37,344,040.29
GSFR	$ 40,932,620.00	NA	$ 40,932,619.94
Optimal	$ 40,744,136.00	NA	$ 40,744,135.10

Source: The authors – R output (2019)[5]

When we consider only the total level forecast for a three month ahead forecast, the results are in Table 14. Our approach has the same performance as the top-down approach. This is expected since both approaches consider the same total forecast estimation (same LSTM model`s output) and the adjustments are concentrated on the lower levels’ forecasts.

Table 14 – Error metrics of the sum of forecasts for 3 months on the total level

BRASIL	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**
Optimal	-$ 8.994.502,00	$ 8.994.502,00	$ 8.994.502,00	-5,43	5,43
GSFR	-$15.204.065,00	$ 15.204.065,00	$15.204.065,00	-9,19	9,19
No reconciliation	-$15.204.080,00	$ 15.204.080,00	$15.204.080,00	-9,19	9,19
Top-down	-$15.204.080,00	$ 15.204.080,00	$15.204.080,00	-9,19	9,19
TURKEY					
Top-down	$ 3.866.587,00	$ 3.866.587,00	$ 3.866.587,00	8,63	8,63
No reconciliation	$ 3.866.588,00	$ 3.866.588,00	$ 3.866.588,00	8,63	8,63
GSFR	$ 3.866.997,00	$ 3.866.997,00	$ 3.866.997,00	8,63	8,63
Optimal	$ 4.055.481,00	$ 4.055.481,00	$ 4.055.481,00	9,05	9,05

Source: The authors – R output (2019)

[5] The Turkish data set has a hierarchy with only 2 levels (total and by districts), no estimations by zone are available

For the lower levels, we will present RMSE and MASE measures, and we summarize the reasons in Table 15.

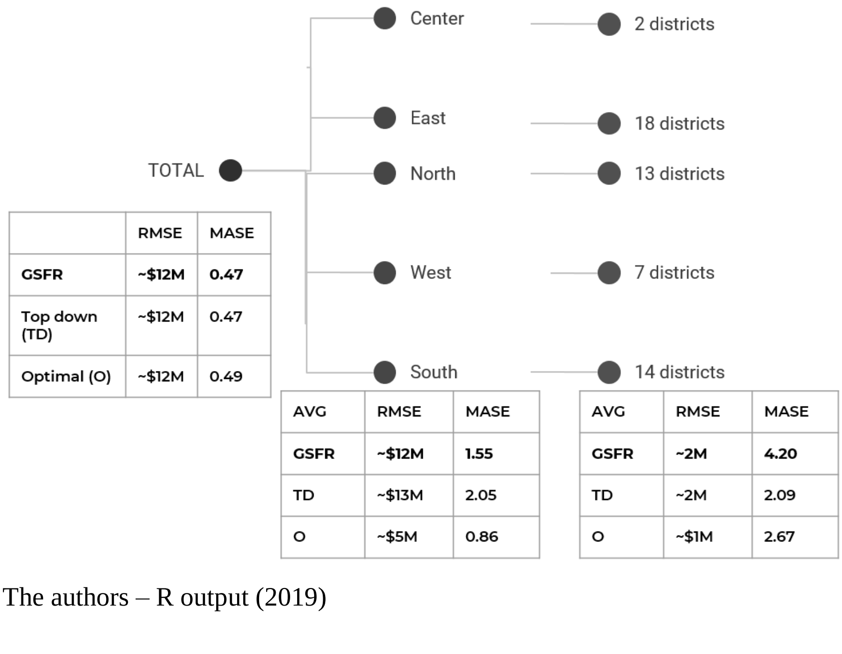
Table 15 – Error measures

RMSE	MASE
Scale-dependent error	Scaled error
Minimizing the RMSE will lead to forecasts of the mean	Less than 1 if it arises from a better forecast than the average naïve forecast computed on the training data
Forecast errors are on the same scale as the data	Scales the errors based on the training MAE from a seasonal random walk
Cannot be used to make comparisons between series that involve different units	Alternative to using percentage errors when comparing forecast accuracy across series with different units
Our goal to report this measure is to give an idea of the size of the error	Our goal is to compare different series

Source: Adapted from Hyndman and Athanasopoulos (2018)

If we consider the average error of each approach on each level, we achieve the results presented in Figures 15 and 16. The amounts of the 3 separate months are summed and all values are in local currency (Brazilian Real and Turkish Lira).

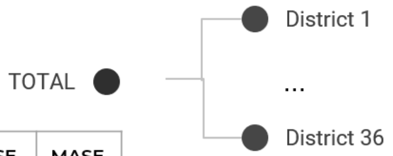
Figure 15 – Average error measures by level in Brazilian data



Source: The authors – R output (2019)

We can see in Figures 15 and 16 that on the most aggregate level, our approach and top-down have a similar performance than optimal. On the first level of disaggregation, on average, our approach is not as good as optimal, but it is better than top-down. Optimal, however, uses the actual data of every region, so it is expected to be more accurate. To the most disaggregated level, on average, our approach has a comparable performance on RMSE. However, MASE shows that also on average, it is performing worse.

Figure 16 – Average error measures by level in Turkish data



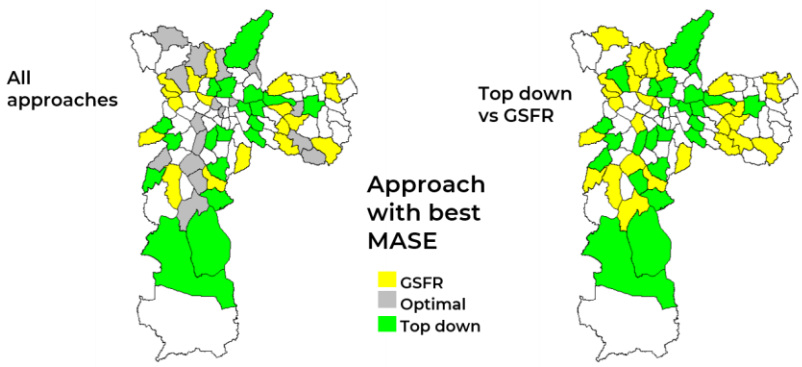
	RMSE	MASE
GSFR	**~$3M**	**0.91**
TD	~$3M	0.91
O	~$3M	0.90

AVG	RMSE	MASE
GSFR	**$396k**	**4.26**
TD	$217K	1.6
O	$153K	1.24

Source: The authors – R output (2019)

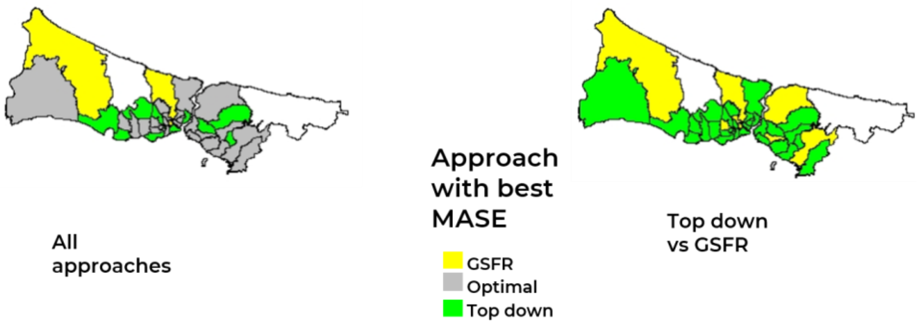
The same measure, MASE, on the other hand, shows that our approach can outperform the others ins some nodes (districts). When looking to each individual district, in a number of them our approach has better performance, even when the optimal approach (best performing in disaggregated levels) is added to the comparison, as we illustrate in Figures 17 and 18.

Figure 17 – Approach with best MASE by district in São Paulo - Brazil



Source: The authors – R output (2019)

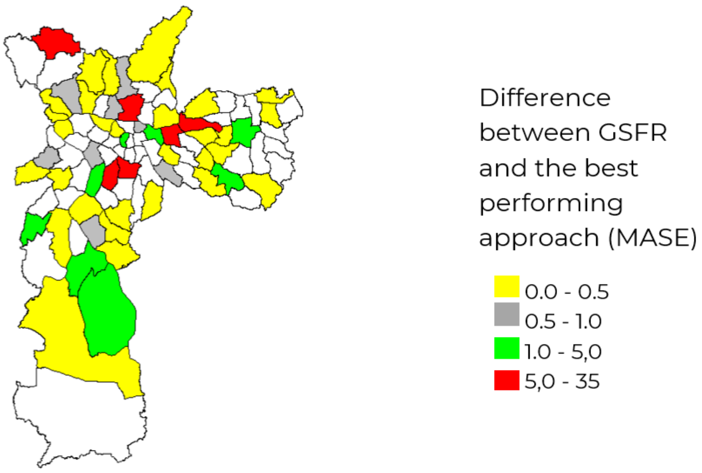
Figure 18 – Approach with best MASE by district in Istanbul – Turkey



Source: The authors – R output (2019)

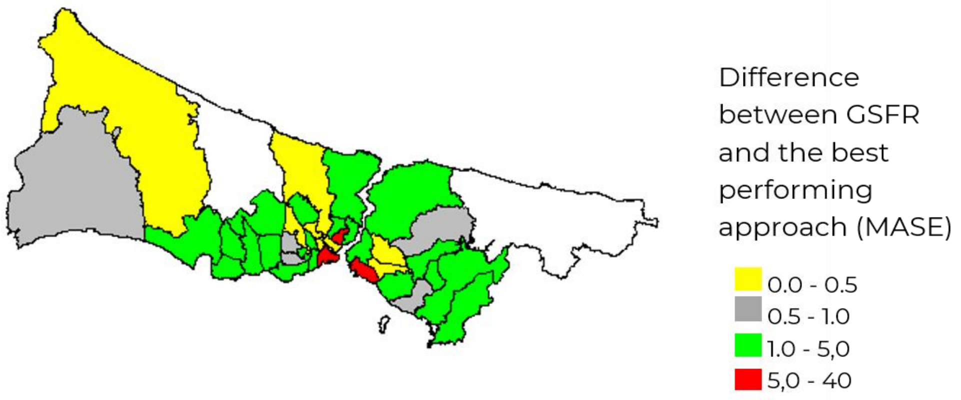
Finally, if we look at the differences between our approach and the best performing in each district, based on MASE, it is, in most cases, very small. 71% of the districts in São Paulo have a difference smaller than 1 (yellow and gray in Figures 19 and 20). In the Turkish case, GSFR has a small difference in a third of the districts.

Figure 19 – Differences in São Paulo – Brazil



Source: The authors – R output (2019)

Figure 20 – Differences in Istanbul – Turkey



Source: The authors – R output (2019)

On Appendix E we present further results in disaggregated level in other error metrics and, on Appendix D an outlier analysis. Summarizing our findings, at the aggregate level, all methods have similar performance. At the disaggregated level, the optimal approach has a better performance in more geographical regions than the other methods, in all metrics considered, because it uses the actual sales of each region. When compared to the benchmark (top-down), GSFR has comparable performance, or outperform it in some regions, even if the method does not have any information on actual sales or historical proportions.

5.6 Concluding remarks

In our study we aimed to, using agglomeration theory and an adapted gravitational model, contribute to literature and practice in several ways. First, our contribution is mostly methodological, by proposing a new forecast reconciliation approach that distributes and reconcile forecasts to lower levels of aggregation, when no actual data or historical proportions are available.

We also contribute to practice by (1) proposing an approach that can be applied to new territories, points of sales, or channel partners; (2) allows a coherent forecast system to be implemented; (3) it is cheap and easy to apply in any organization; finally, (4) uses data that is publicly available or natural to the process of most organizations.

Our contribution to theory is that we show that gravitational models can be used in combination with state-of-the-art forecasting methods (such as LSTM) to not only explain sales but also to predict it, with evidence of accuracy. The ability of the gravitational model based on attractiveness measures to distribute sales between regions gives further evidence of the importance of considering spatial effects on studies regarding sales, either explaining or forecasting it. We also introduce hierarchical structures and the need for reconciliation approaches to marketing literature.

In this paper we suggest an approach that allows companies to distribute forecasts in lower levels of aggregation, that also has lower costs for companies to implement, using open source tools. We provide evidence of the accuracy of such an approach in two empirical test cases, using data from Brazil and Turkey. We also provide a model that can be applied by companies of any sector, especially those that have a limited amount of resources to invest in the process to gather, storage and analyze data.

Our method has some limitations. Since we don’t have individual buyers’ information, we address this problem by assuming that all residents of a district are potential customers. For each district, we calculated the proportion of its population that are predicted to visit other districts (outflows), and the proportion that is predicted to shop inside their own district. So, for each district, we will have a new population consisting of its own residents which prefer their own district for shopping, and the people coming from other districts. This population is parametric based on the still unknown parameters. We assume the number of transactions is equal to the population volume after implementing the flow modeling.

Also, customers’ residences are not defined, as a result, we are not able to define a population center for each district. To address this limitation, we consider the district center coordinates as the residents’ home locations. Thus, distances between districts are calculated as geo-distance between two district centers, and the distance of residents from their home district is calculated as the radius of the largest circle possible inside that district divided by a parameter t. Finally, we are aware that there are socio-demographic differences between the population of different districts which will result in different behavioral patterns, however, we had to fit the model with same parameters for all the districts assuming that all population of different districts will show similar behavior.

In São Paulo’s data, we removed all the districts without sales from the calculations in all levels. However, those districts with no sales records have potential customers that can move to districts with sales records to buy. Considering all 96 districts as customers and 53 districts as sellers in the gravitational model could improve further the accuracy of GSFR and be a more realistic scenario. Future research can address these limitations.

Our aim was to suggest a general approach that can be used by any organization, so we did not focus on some particular districts' results. However, some districts had higher errors measures than the benchmark, that could be considered outlier districts, with unusual sales volume. We show in Appendix D an "outlier analysis" to address that. We checked for outliers in two variables: sales' history and districts. We cannot remove one month from the data because it seems to be an outlier. We could remove districts, however, not without a good reason. In time-series analyses we use real data and outliers would only happen if data is inputted incorrectly, which, in our case studies, is hard to prove. Otherwise, outlier data points are related to reality and need to be a part of the analyses. For those reasons, we kept all the results with all the districts with data available. Future research can explore further what we could call outliers and the reasons these districts are behaving differently from the rest.

Another strategy that could be implemented to improve accuracy is to test different temporal aggregations to smooth out fluctuations. This means that instead of using monthly data and model fitting, we could consider aggregated sales bi-monthly, quarterly, biannually, or annually. To be able to test it, we need to use a test set with 12 months (one year), this implies that a larger dataset is necessary, with enough data points to allow all possible temporal aggregations.

Even without applying these strategies to improve accuracy, our approach, GSFR, has a comparable or better performance than methods that use the actual historical sales. Our study contributes to marketing literature, providing further evidence on the importance of agglomeration theory to explain and predict sales. To practice, our approach gives a feasible alternative to disaggregate sales forecasts to geographical areas without information about historical proportions when only a total forecast (for the most aggregate level) is available.

6 CONCLUSIONS

The four papers that were presented in this PhD dissertation introduced the necessity of marketing field to give more attention to time-series forecasting, but also to (1) the use of spatial information and agglomeration theory, (2) the importance of reporting evidences of accuracy of the models proposed, and (3) to focus on models that are more feasible to implement in practice. They introduce to marketing field the research on hierarchical and grouped time-series reconciliation approaches. The studies show why it may be necessary to consider those structures and to apply reconciliation approaches when dealing with different levels of aggregation.

The first paper is a necessary introduction on forecasting to the marketing field, connecting it to the marketing modelling literature. It answers the questions about the current state of the art in marketing research regarding forecast methods. It also provides criteria to choose the most appropriate forecast method, combining forecasting and explanation goals, contributing to the literature on marketing models' classifications (ARMSTRONG, 2001; CHINTAGUNTA; NAIR, 2011; GENTRY; CALANTONE; CUI, 2006; ROBERTS, 1998). The paper also gives some guidance for future research, such as the use of machine learning techniques, ensemble models, explore location information, and the importance to adapt marketing response models to work with aggregate data, maintaining the same predictive power (WEDEL; KANNAN, 2016). The main conclusion is that marketing should focus on improving the methods applied in practice by adding marketing variables.

The second paper follows on the guidance to use machine learning and ensemble methods made by the previous paper. It was proposed an approach to set the weights of an ensemble using optimization from finance literature. This approach had a better performance than equal weights approach and all 15 single (base) models tested. It contributes to ensemble models literature (COUSSEMENT; DE BOCK, 2013; DE BOCK; COUSSEMENT; VAN DEN POEL, 2010; LEMMENS; CROUX, 2006) by answering the question regarding a better way to set the weights of an ensemble than equal weights. It also contributes to a popular question currently in forecasting literature regarding the accuracy performance of ensemble and machine learning techniques compared to statistical methods (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2018). For marketing literature, the contribution is achieved by focusing on forecasting and time-series (DEKIMPE; HANSSENS, 2000; BEAL; WILSON, 2015).

The third paper seeks to answer which marketing variables used, alone or combined, to the structure of a forecast system can improve accuracy. It follows the main suggestion of the first paper, using a statistical method applied in practice and adding marketing variables to the structure of a forecasting system. This paper introduces the concepts of structure and criteria of aggregation and forecast reconciliation approaches (HYNDMAN *et al.*, 2011). It also suggests that location information is important to improve accuracy.

The last paper explores further the importance of location information, using agglomeration theory (BRONNENBERG; SISMEIRO, 2002; HUFF, 1964; LIU; STEENKAMP; ZHANG, 2018) and an adapted gravitational model to reconcile a hierarchical time-series forecast, the gravitational sales forecast reconciliation (GSFR) approach. Forecasts are estimated using state-of-the-art forecasting methods, a deep learning technique (LSTM) that proved to be more accurate than the other methods tested. GSFR approach has comparable performance to other approaches that use actual data or historical proportions.

Despite the fact that I designed the papers to overcome the limitations of one another. For example, by using deep learning techniques as suggested on the second, and agglomeration theory as suggested on the third. Still, the papers have some limitations that can be explored in future research. First, replications of the studies in different datasets or for different marketing problems can give further evidence that our propositions can be easily applied to any organization, with different sizes or from different sectors. Different datasets might also provide information about other marketing variables, such as promotion or price strategies. This can help overcome the limitation of the third paper of this PhD dissertation. With more marketing variables, different combinations can be tested.

The third paper focused only on the optimal approach, and on the weights that were reported in forecasting literature to be the best performing, MinT and WLS. Since we already were comparing structures and criteria, to add further comparisons would make it complex to reach and communicate a clear conclusion. Despite forecasting literature has already established that the optimal approach using MinT tends to be more accurate, future research can compare the performance of different reconciliation approaches. Also, a known problem is that, for datasets with many repeated values or zeros, MinT becomes unfeasible. For that reason, we had to use WLS in some of the reconciliations estimated. Another opportunity for studies is to test or suggest different weights estimations to be used with the optimal approach.

The last paper, in addition to the limitations mentioned in its concluding remarks section, some public datasets may be discontinued by new governmental policies. This might impose difficulties applying GSFR in the future. Future research might test other data sources' ability to provide proxies of the variables collected from those sources. For example, satellite data has been studied as a measure of economic activity (HENDERSON; STOREYGARD; WEIL, 2012). However, these data are costly regarding economic and (skilled) human resources.

One feedback this research had on a marketing conference was that this topic is not related to marketing. Forecasting sales, in that particular marketing researcher's opinion, is a concern of supply channel, and not marketing. Our purpose is to change this mindset in marketing literature, following the path lead by Beal and Wilson (2015), showing why the field needs to gain knowledge not only on the explanation but also on the prediction of this variable. It seems odd that it is needed to justify why forecasting sales is important to marketing since it has been stated in marketing literature before (CHINTAGUNTA; NAIR, 2011). Also, it seems odd to justify why marketing needs to gain the ability to work with time-series data (DEKIMPE; HANSSENS, 2000).

If marketing researches have confidence in the theory developed that focus on explaining how marketing actions influence customers response (which include sales) they should be just as confident that these variables will help in prediction. Chapter 5 gives evidence of one of those theories, agglomeration theory (BRONNENBERG; SISMEIRO, 2002; HUFF, 1964; LIU; STEENKAMP; ZHANG, 2018). However, other theory and other marketing problems can benefit from the methods and approaches applied in this PhD dissertation. One most evident problem is to predict future customer lifetime value, giving guidance to those customers or segments that need to be activated or abandoned. Future studies can adapt our studies to those data and issues, easily.

Another reason to focus on forecasting is the impact that a wrong forecast has in marketing metrics. Excess or lack of products on stores can impose changes in pricing strategies. Companies might need to develop price promotions or, on the contrary, increase prices to avoid stock-outs. Excess of price promotions has a very well documented impact on brand reputation and quality perception. Stock-outs also impact on customer loyalty, repurchase intentions, and satisfaction.

Marketing is the department responsible to know how to satisfy customers, how to reach them, how to keep them engaged in a relationship with the organization. However, is not enough to show which variables explain those behaviors, our field needs to prove that these variables can impact on planning more efficiently, on predicting those behaviors more accurately. Marketing needs to give further evidence of the accuracy of its models and allow them to be implemented in practice more easily (FILDES; MA; KOLASSA, 2018).

In addition to that, as stated in the introduction, researchers have now access to new sources of time-series data. It is the outcome of technology-based interactions with organizations and data is being now collected by organizations longitudinally, allowing dynamic effects to be studied. However, if that data is still not available to researchers, time-series is a natural outcome of the processes of all organizations and has been studied by other fields with further improvements made recently by machine learning and deep learning techniques.

I (along with the co-authors of the papers) also challenge our field to build models that can be applied to organizations of all sizes and sectors. Marketing literature favors companies that have access to rich datasets, with individual-level information and collects promotional, pricing and other variables. This dataset structure is very similar to those obtained from companies like Nielsen. This dependency on private data providers is prejudicial to our field. Only companies that can afford to buy the service of these companies or invest in their own data collection could apply marketing models. Literature, however, shows that even those do not implement those complex models (FILDES; MA; KOLASSA, 2018; MEYER; HUTCHINSON, 2016; REISS, 2011).

These papers offer several contributions. Those are mostly methodological; however, they are also theoretical and practical. To practice, the contribution was to focus on evidence of the accuracy of all methods and approaches applied and suggested in our papers. All analysis uses open-source tools, proprietary data that is natural to the process of every organization, and publicly available data. The goal was to favor methods and tools that are generalizable to all types of goods and can be easily applied with minimal investment. Since one of the issues for practice is the limited computational resources, our code is computationally efficient, with each forecasting model running is less than a minute on a laptop. The estimation that needs more time to run is the 4-D grid search on model fitting parameters of GSFR, however, this can be improved by parallelizing the process, if necessary.

Methodologically it proposes a new approach to set the weights of ensemble models; compare accuracy performance of different forecasting methods (statistical, machine learning and deep learning techniques); compare different structures (hierarchies and groups) and criteria based in marketing variables; and develop a new approach to reconcile forecasts when no actual sales data or historical proportions are available, using an adapted gravitational model. Theoretically, the papers introduced to marketing literature the concepts of hierarchical and grouped structures and reconciliation approaches. They also provide evidence of the importance of location information to improve accuracy and of agglomeration theory to distribute forecast between regions. I hope to see future research in marketing applying the methods presented in this PhD dissertation (and others) to different marketing problems, and proving that other marketing variables and theories can improve not only the explanatory power of models but also its predictive ability.

REFERENCES

ABHISHEK, Vibhanshu; HOSANAGAR, Kartik; FADER, Peter S. Aggregation bias in sponsored search data: The curse and the cure. **Marketing Science**, [*s.l*], v. 34, n. 1, p. 59–77, 2015.

ALBUQUERQUE, Paulo; BRONNENBERG, Bart J. Measuring the impact of negative demand shocks on car dealer networks. **Marketing Science**, [*s.l*], v. 31, n. 1, p. 4–23, 2012.

ALI, Özden Gür *et al.* SKU demand forecasting in the presence of promotions. **Expert Systems with Applications**, [*s.l*], v. 36, n. 10, p. 12340–12348, 2009.

ALLENBY, Greg M.; BAKKEN, David G.; ROSSI, Peter E. The HB revolution how Bayesian methods have changed the face of marketing research. **Marketing Research**, [*s.l*], v. 16, p. 20–25, 2004.

ALLENBY, Greg M.; GARRATT, Mark J.; ROSSI, Peter E. A model for trade-up and change in considered brands. **Marketing Science**, [*s.l*], v. 29, n. 1, p. 40–56, 2010.

AMERICAN MARKETING ASSOCIATION. **AMA Dictionary**. [*s.d.*]. Available on: https://www.ama.org/resources/Pages/Dictionary.aspx?dLetter=M. Retrieved 10 Sept. 2016.

ARIBARG, Anocha; ARORA, Neeraj; KANG, Moon Young. Predicting joint choice using individual data. **Marketing Science**, [*s.l*], v. 29, n. 1, p. 139–157, 2010.

ARMSTRONG, Jon Scott. **Principles of forecasting: a handbook for researchers and practitioners**. [*s.l.*]: Springer Science & Business Media, 2001. v. 30.

ARORA, Neeraj; HENDERSON, Ty; LIU, Qing. Noncompensatory dyadic choices. **Marketing Science**, [*s.l*], v. 30, n. 6, p. 1028–1047, 2011.

BARROW, Devon K.; CRONE, Sven F. A comparison of AdaBoost algorithms for time series forecast combination. **International Journal of Forecasting**, [*s.l*], v. 32, n. 4, p. 1103–1119, 2016.

BARROW, Dk. Forecasting intraday call arrivals using the seasonal moving average method. **Journal of Business Research**, [*s.l*], v. 69, n. 12, p. 6088–6096, 2016.

BEAL, Mary; WILSON, J. Holton. Forecasting in the marketing curriculum. **International Journal of Business, Marketing, & Decision Science**, [*s.l*], v. 8, n. 1, 2015.

BERGMEIR, Christoph Norbert; BENÍTEZ SÁNCHEZ, José Manuel. Neural networks in R using the Stuttgart neural network simulator: RSNNS**. Journal of Statistical Software**, 46(7): online (2012). Available on: http://hdl.handle.net/10481/39548. American Statistical Association, 2012. Retrieved 28 May 2019.

BOLLINGER, Bryan; GILLINGHAM, Kenneth. Peer effects in the diffusion of solar photovoltaic panels. **Marketing Science**, [*s.l*], v. 31, n. 6, p. 900–912, 2012.

BRADLOW, Eric T. *et al.* Spatial models in marketing. **Marketing Letters**, [*s.l*], v. 16, n. 3–4, p. 267–278, 2005.

BRIESCH, Richard A.; DILLON, William R.; FOX, Edward J. Category positioning and store choice: The role of destination categories. **Marketing Science**, [*s.l*], v. 32, n. 3, p. 488–509, 2013.

BRONNENBERG, Bart J.; DUBÉ, Jean-Pierre H.; GENTZKOW, Matthew. The evolution of brand preferences: Evidence from consumer migration. **The American Economic Review**, [*s.l*], v. 102, n. 6, p. 2472–2508, 2012.

BRONNENBERG, Bart J.; MAHAJAN, Vijay. Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. **Marketing Science**, [*s.l*], v. 20, n. 3, p. 284–299, 2001.

BRONNENBERG, Bart J.; SISMEIRO, Catarina. Using multimarket data to predict brand performance in markets for which no or poor data exist. **Journal of Marketing Research**, [*s.l*], v. 39, n. 1, p. 1–17, 2002.

BUCKLIN, Randolph E.; SIDDARTH, Sivaramakrishnan; SILVA-RISSO, Jorge M. Distribution intensity and new car choice. **Journal of Marketing Research**, [*s.l*], v. 45, n. 4, p. 473–486, 2008.

CANITZ, Henry. Overcoming Barriers to Improving Forecast Capabilities. **Foresight: The International Journal of Applied Forecasting**, [*s.l*], n. 41, p. 26–34, 2016.

CHAN, Tat Y.; PADMANABHAN, V.; SEETHARAMAN, P. B. An econometric model of location and pricing in the gasoline market. **Journal of Marketing Research**, [*s.l*], v. 44, n. 4, p. 622–635, 2007.

CHASE, Charles W. **Demand-driven forecasting: a structured approach to forecasting**. [s.l.]: John Wiley & Sons, 2013.

CHE, Hai; CHEN, Xinlei; CHEN, Yuxin. Investigating effects of out-of-stock on consumer stockkeeping unit choice. **Journal of Marketing Research**, [*s.l*], v. 49, n. 4, p. 502–513, 2012.

CHEN, Yubo; WANG, Qi; XIE, Jinhong. Online social interactions: A natural experiment on word of mouth versus observational learning. **Journal of Marketing Research**, [*s.l*], v. 48, n. 2, p. 238–254, 2011.

CHEN, Yuxin; STECKEL, Joel H. Modeling credit card share of wallet: Solving the incomplete information problem. **Journal of Marketing Research**, [*s.l*], v. 49, n. 5, p. 655–669, 2012.

CHEVALIER, Judith A.; MAYZLIN, Dina. The effect of word of mouth on sales: Online book reviews. **Journal of Marketing Research**, [*s.l*], v. 43, n. 3, p. 345–354, 2006.

CHINTAGUNTA, Pradeep; HANSSENS, Dominique M.; HAUSER, John R. Editorial—Marketing Science and Big Data. **Marketing Science**, [*s.l*], v. 35, n. 3, p. 341–342, 2016.

CHINTAGUNTA, Pradeep K.; GOPINATH, Shyam; VENKATARAMAN, Sriram. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. **Marketing Science**, [*s.l*], v. 29, n. 5, p. 944–957, 2010.

CHINTAGUNTA, Pradeep K.; NAIR, Harikesh S. Structural workshop paper-discrete-choice models of consumer demand in marketing. **Marketing Science**, [*s.l*], v. 30, n. 6, p. 977–996, 2011.

CHUNG, Jaihak; RAO, Vithala R. A general consumer preference model for experience products: application to internet recommendation services. **Journal of Marketing Research**, [*s.l*], v. 49, n. 3, p. 289–305, 2012.

CHUNG, Kevin YC; DERDENGER, Timothy P.; SRINIVASAN, Kannan. Economic value of celebrity endorsements: Tiger Woods' impact on sales of Nike golf balls. **Marketing Science**, [*s.l*], v. 32, n. 2, p. 271–293, 2013.

COOLEY, David. **googleway: Accesses Google Maps APIs to Retrieve Data and Plot Maps**. [s.l: s.n.]. 2018. Available on: https://CRAN.R-project.org/package=googleway. Retrieved 12 Mar. 219

COUSSEMENT, Kristof; DE BOCK, Koen W. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. **Journal of Business Research**, [*s.l*], v. 66, n. 9, p. 1629–1636, 2013.

CUI, Dapeng; CURRY, David. Prediction in marketing using the support vector machine. **Marketing Science**, [*s.l*], v. 24, n. 4, p. 595–615, 2005.

DE BOCK, Koen W.; COUSSEMENT, Kristof; VAN DEN POEL, Dirk. Ensemble classification based on generalized additive models. **Computational Statistics & Data Analysis**, [*s.l*], v. 54, n. 6, p. 1535–1546, 2010.

DE MONTJOYE, Yves-Alexandre *et al.* Unique in the crowd: The privacy bounds of human mobility. **Scientific Reports**, [*s.l*], v. 3, p. 1376, 2013.

DEKIMPE, Marnik G.; HANSSENS, Dominique M. Time-series models in marketing: Past, present and future. **International Journal of Research in Marketing**, [*s.l*], v. 17, n. 2–3, p. 183–193, 2000.

DEKKER, Mark; VAN DONSELAAR, Karel; OUWEHAND, Pim. How to use aggregation and combined forecasting to improve seasonal demand forecasts. **International Journal of Production Economics**, [*s.l*], v. 90, n. 2, p. 151–167, 2004.

DELLAROCAS, Chrysanthos; ZHANG, Xiaoquan Michael; AWAD, Neveen F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. **Journal of Interactive Marketing**, [*s.l*], v. 21, n. 4, p. 23–45, 2007.

DERBEKO, Philip; EL-YANIV, Ran; MEIR, Ron. Variance optimized bagging. In: European Conference on Machine Learning. **Proceedings […]** Berlin, Heidelberg: Springer, 2002. p. 60-72.

DEWAN, Sanjeev; RAMAPRASAD, Jui. Social media, traditional media, and music sales. **Mis Quarterly**, [*s.l*], v. 38, n. 1, p. 101–121, 2014.

DHAR, Vasant; CHANG, Elaine A. Does chatter matter? The impact of user-generated content on music sales. **Journal of Interactive Marketing**, [*s.l*], v. 23, n. 4, p. 300–307, 2009.

DIVAKAR, Suresh; RATCHFORD, Brian T.; SHANKAR, Venkatesh. Practice Prize Article—CHAN4CAST: A Multichannel, Multiregion Sales Forecasting Model and Decision Support System for Consumer Packaged Goods. **Marketing Science**, [*s.l*], v. 24, n. 3, p. 334–350, 2005.

DUNN, DM; WILLIAMS, WH; DECHAINE, TL. Aggregate versus subaggregate models in local area forecasting. **Journal of the American Statistical Association**, [*s.l*], v. 71, n. 353, p. 68–71, 1976.

FEIT, Eleanor McDonnell *et al.* Fusing aggregate and disaggregate data with an application to multiplatform media consumption. **Journal of Marketing Research**, [*s.l*], v. 50, n. 3, p. 348–364, 2013.

FILDES, Robert *et al.* Forecasting and operational research: a review. **Journal of the Operational Research Society**, [*s.l*], v. 59, n. 9, p. 1150–1172, 2008.

FILDES, Robert Alan; MA, Shaohui; KOLASSA, Stephan. Retail forecasting: research and practice. [*s.l*], 2018. Available on: https://www.researchgate.net/publication/328095900_Retail_forecasting_research_and_practice. Retrieved 12 Mar. 2019.

FILDES, Robert; GOODWIN, Paul. Against your better judgment? How organizations can improve their use of management judgment in forecasting. **Interfaces**, [*s.l*], v. 37, n. 6, p. 570–576, 2007.

FILDES, Robert; GOODWIN, Paul; ÖNKAL, Dilek. Use and misuse of information in supply chain forecasting of promotion effects. **International Journal of Forecasting**, [*s.l*], v. 35, n. 1, p. 144–156, 2019.

FLIEDNER, Eugene B.; LAWRENCE, Barry. Forecasting system parent group formation: An empirical application of cluster analysis. **Journal of Operations Management**, [*s.l*], v. 12, n. 2, p. 119–130, 1995.

FLIEDNER, Eugene B.; MABERT, Vincent A. Constrained forecasting: some implementation guidelines. **Decision Sciences**, [*s.l*], v. 23, n. 5, p. 1143–1161, 1992.

FLIEDNER, Gene. Hierarchical forecasting: issues and use guidelines. **Industrial Management & Data Systems**, [*s.l*], v. 101, n. 1, p. 5–12, 2001.

GAL, Yarin; GHAHRAMANI, Zoubin. A theoretically grounded application of dropout in recurrent neural networks. In: Advances in Neural Information Processing Systems**.** 2016. Available on: https://arxiv.org/abs/1512.05287. Retrieved 28 May 2019.

GARBER, Tal *et al.* From density to destiny: Using spatial dimension of sales data for early prediction of new product success. **Marketing Science**, [*s.l*], v. 23, n. 3, p. 419–428, 2004.

GENRE, Véronique *et al.* Combining expert forecasts: Can anything beat the simple average? **International Journal of Forecasting**, [*s.l*], v. 29, n. 1, p. 108–121, 2013.

GENTRY, Lance; CALANTONE, Roger J.; CUI, Shaojie Anna. The forecasting classification grid: a typology for method selection. **Journal of Global Business Management**, [*s.l*], v. 2, n. 1, p. 48–60, 2006.

GLAESER, Edward L.; KOLKO, Jed; SAIZ, Albert. Consumer city. **Journal of Economic Geography**, [*s.l*], v. 1, n. 1, p. 27–50, 2001.

GODES, David; MAYZLIN, Dina. Using online conversations to study word-of-mouth communication. **Marketing Science**, [*s.l*], v. 23, n. 4, p. 545–560, 2004.

GONZALEZ-BENITO, Oscar; MUNOZ-GALLEGO, Pablo A.; KOPALLE, Praveen K. Asymmetric competition in retail store formats: Evaluating inter-and intra-format spatial effects. **Journal of Retailing**, [*s.l*], v. 81, n. 1, p. 59–73, 2005.

GOPINATH, Shyam; CHINTAGUNTA, Pradeep K.; VENKATARAMAN, Sriram. Blogs, advertising, and local-market movie box office performance. **Management Science**, [*s.l*], v. 59, n. 12, p. 2635–2654, 2013.

GRAEFE, Andreas *et al.* Combining forecasts: An application to elections. **International Journal of Forecasting**, [*s.l*], v. 30, n. 1, p. 43–54, 2014.

GRAEFE, Andreas *et al.* Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems. **International Journal of Forecasting**, [*s.l*], v. 31, n. 3, p. 943–951, 2015.

GROLEMUND, Garrett; WICKHAM, Hadley. Dates and Times Made Easy with lubridate. **Journal of Statistical Software**, [*s.l*], v. 40, n. 3, p. 1–25, 2011.

HANSSENS, Dominique M. Order forecasts, retail sales, and the marketing mix for consumer durables. **Journal of Forecasting**, [*s.l*], v. 17, n. 3–4, p. 327–346, 1998.

HENDERSON, J. Vernon; STOREYGARD, Adam; WEIL, David N. Measuring economic growth from outer space. **The American Economic Review**, [*s.l*], v. 102, n. 2, p. 994–1028, 2012.

HMAMOUCHE, Youssef. **NlinTS: Non Linear Time Series Analysis**. [s.l: s.n.]. 2019. Available on: https://CRAN.R-project.org/package=NlinTS. Retrieved 10 July 2019

HUFF, David L. Defining and estimating a trading area. **Journal of Marketing**, [*s.l*], v. 28, n. 3, p. 34–38, 1964.

HUNG, Chihli; CHEN, Jing-Hong. A selective ensemble based on expected probabilities for bankruptcy prediction. **Expert Systems with Applications**, [*s.l*], v. 36, n. 3, p. 5297–5303, 2009.

HYNDMAN, Rob J. *et al.* **Forecasting with exponential smoothing: the state space approach**. [s.l.]: Springer Science & Business Media, 2008.

HYNDMAN, Rob J. *et al.* Optimal combination forecasts for hierarchical time series. **Computational Statistics & Data Analysis**, [*s.l*], v. 55, n. 9, p. 2579–2589, 2011.

HYNDMAN, Rob J. *et al.* **forecast: Forecasting functions for time series and linear models**. [s.l: s.n.]. 2018. Available on: http://pkg.robjhyndman.com/forecast. Retrieved 10 July 2019

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting: principles and practice**. Melbourne, Australia: OTexts, 2018. Available on: https://OTexts.org/fpp2/. Retrieved 29 May 2019

HYNDMAN, Rob J.; KHANDAKAR, Yeasmin. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, [*s.l*], v. 26, n. 3, p. 1–22, 2008.

HYNDMAN, Rob J.; KOEHLER, Anne B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, [*s.l*], v. 22, n. 4, p. 679–688, 2006.

HYNDMAN, Rob J.; LEE, Alan; WANG, Earo. **hts: Hierarchical and Grouped Time Series**. [s.l: s.n.]. 2017. Available on: https://CRAN.R-project.org/package=hts. Retrieved 10 July 2019

JJ ALLAIRE; CHOLLET, François. **keras: R Interface to “Keras”**. [s.l: s.n.]. 2019. Available on: https://CRAN.R-project.org/package=keras. Retrieved 10 July 2019

KARNIOUCHINA, Ekaterina V. Impact of star and movie buzz on motion picture distribution and box office revenue. **International Journal of Research in Marketing**, [*s.l*], v. 28, n. 1, p. 62–74, 2011.

KERWIN, A.; ZMUDA, N. Ad-exec of the future? Big-Data draper. **Advertising Age**, [*s.l*], v. 84, n. 11, p. 32, 2013.

KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, [*s.l*], 2014. Available on: https://arxiv.org/abs/1412.6980. Retrieved 28 May 2019

KOURENTZES, Nikolaos. **nnfor: Time Series Forecasting with Neural Networks**. [s.l: s.n.]. 2019. Available on: https://CRAN.R-project.org/package=nnfor. Retrieved 10 July 2019

KOURENTZES, Nikolaos; ROSTAMI-TABAR, Bahman; BARROW, Devon K. Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? **Journal of Business Research**, [*s.l*], v. 78, p. 1–9, 2017.

KRAWCZYK, Bartosz. One-class classifier ensemble pruning and weighting with firefly algorithm. **Neurocomputing**, [*s.l*], v. 150, p. 490–500, 2015.

KREMER, Mirko; SIEMSEN, Enno; THOMAS, Douglas J. The sum and its parts: Judgmental hierarchical forecasting. **Management Science**, [*s.l*], v. 62, n. 9, p. 2745–2764, 2015.

LEEFLANG, Peter SH *et al.* Challenges and solutions for marketing in a digital era. **European Management Journal**, [*s.l*], v. 32, n. 1, p. 1–12, 2014.

LEMMENS, Aurélie; CROUX, Christophe. Bagging and boosting classification trees to predict churn. **Journal of Marketing Research**, [*s.l*], v. 43, n. 2, p. 276–286, 2006.

LESAGE, James P. The theory and practice of spatial econometrics. **University of Toledo. Toledo, Ohio**, [*s.l*], v. 28, p. 33, 1999.

LIU, Angela Xia; STEENKAMP, Jan-Benedict EM; ZHANG, Jurui. Agglomeration as a Driver of the Volume of Electronic Word of Mouth in the Restaurant Industry. **Journal of Marketing Research**, [*s.l*], 2018.

LIU, Xiao; SINGH, Param Vir; SRINIVASAN, Kannan. A structured analysis of unstructured big data by leveraging cloud computing. **Marketing Science**, [*s.l*], v. 35, n. 3, p. 363–388, 2016.

LIU, Yong. Word of mouth for movies: Its dynamics and impact on box office revenue. **Journal of Marketing**, [*s.l*], v. 70, n. 3, p. 74–89, 2006.

LU, Xin *et al.* Approaching the limit of predictability in human mobility. **Scientific Reports**, [*s.l*], v. 3, 2013.

LUAN, Y. Jackie; SUDHIR, K. Forecasting marketing-mix responsiveness for new products. **Journal of Marketing Research**, [*s.l*], v. 47, n. 3, p. 444–457, 2010.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. Statistical and Machine Learning forecasting methods: Concerns and ways forward. **PloS one**, [*s.l*], v. 13, n. 3, p. e0194889, 2018.

MARTINEZ, Francisco. **tsfknn: Time Series Forecasting Using Nearest Neighbors**. [*s.l: s.n.*]. 2018. Available on: https://CRAN.R-project.org/package=tsfknn. Retrieved 10 July 2019

MAZZON, José Afonso; TOLEDO, Geraldo Luciano De. Métodos de previsão em marketing. In: **Marketing: aplicações de métodos quantitativos**. São Paulo (SP): Atlas, 1983.

MELA, Carl F. Structural Workshop Paper-Data Selection and Procurement. **Marketing Science**, [*s.l.*], v. 30, n. 6, p. 965–976, 2011.

MEYER, Robert J.; HUTCHINSON, J. Wesley. (When) Are We Dynamically Optimal? A Psychological Field Guide for Marketing Modelers. **Journal of Marketing**, [*s.ls.l*], v. 80, n. 5, p. 20–33, 2016.

MOON, Sangkil; BERGEY, Paul K.; IACOBUCCI, Dawn. Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. **Journal of Marketing**, [*s.l*], v. 74, n. 1, p. 108–121, 2010.

NARAYANAN, Sridhar; NAIR, Harikesh S. Estimating causal installed-base effects: A bias-correction approach. **Journal of Marketing Research**, [*s.l*], v. 50, n. 1, p. 70–94, 2013.

NARZO, Antonio Fabio Di; AZNARTE, Jose Luis; STIGLER, Matthieu. **tsDyn: Time series analysis based on dynamical systems theory**. [s.l: s.n.]. 2009. Available on: https://cran.r-project.org/package=tsDyn/vignettes/tsDyn.pdf. Retrieved 10 July 2019

NENOVA, Zlatana D.; MAY, Jerrold H. Determining an optimal hierarchical forecasting model based on the characteristics of the data set. **Journal of Operations Management**, [*s.l*], v. 44, p. 62–68, 2016.

NEVO, Aviv. Empirical Models of Consumer Behavior. **Annual Review of Economics**, [*s.l*], v. 3, n. 1, p. 51–75, 2011.

NI, Jian; NESLIN, Scott A.; SUN, Baohong. Database submission-the isms durable goods data sets. **Marketing Science**, [*s.l*], v. 31, n. 6, p. 1008–1013, 2012.

ONISHI, Hiroshi; MANCHANDA, Puneet. Marketing activity, blogging and sales. **International Journal of Research in Marketing**, [*s.l*], v. 29, n. 3, p. 221–234, 2012.

PETRIN, Amil; TRAIN, Kenneth. A control function approach to endogeneity in consumer choice models. **Journal of Marketing Research**, [*s.l*], v. 47, n. 1, p. 3–13, 2010.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018. Available on: https://www.R-project.org/. Retrieved 10 July 2019

REISS, Peter C. Structural Workshop Paper-Descriptive, Structural, and Experimental Empirical Methods in Marketing Research. **Marketing Science**, [*s.l*], v. 30, n. 6, p. 950–964, 2011.

ROBERTS, John H. Marketing approaches to forecasting problems. **Journal of Forecasting**, [*s.l*], v. 17, n. 3–4, p. 169–174, 1998.

ROODERKERK, Robert P.; VAN HEERDE, Harald J.; BIJMOLT, Tammo HA. Incorporating context effects into a choice model. **Journal of Marketing Research**, [*s.l*], v. 48, n. 4, p. 767–780, 2011.

ROSSI, Peter E. Invited Paper—Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications. **Marketing Science**, [*s.l*], v. 33, n. 5, p. 655–672, 2014.

ROSSI, Peter E.; ALLENBY, Greg M. Bayesian statistics and marketing. **Marketing Science**, [*s.l*], v. 22, n. 3, p. 304–328, 2003.

RUSSELL, Gary J.; KAMAKURA, Wagner A. Understanding brand competition using micro and macro scanner data. **Journal of Marketing Research**, [*s.l*], p. 289–303, 1994.

SEAMAN, Brian. Considerations of a retail forecasting practitioner. **International Journal of Forecasting**, [*s.l*], 2018.

SHRIVER, Scott K. Network effects in alternative fuel adoption: Empirical analysis of the market for ethanol. **Marketing Science**, [*s.l*], v. 34, n. 1, p. 78–97, 2015.

SILVEIRA NETTO, Carla F.; HYNDMAN, Rob J.; BREI, Vinicius Andrade. What is the Optimal Strategy of Aggregation for Forecasting Sales? A Hierarchical Forecast Reconciliation by Region, Product Category, and Channel. In: EnANPAD 2019. **Proceedings […]** São Paulo: ANPAD, 2019.

SILVEIRA NETTO, Carla F.; SLONGO, Luiz A. Marketing metrics, Big Data and the role of the marketing department. **Brazilian Journal of Management/Revista de Administração da UFSM**, [*s.l*], v. 12, n. 3, 2019.

SINGH, Sujit. Forecasting: Academia versus Business. **Foresight: The International Journal of Applied Forecasting**, [*s.l*], n. 41, p. 46–47, 2016.

SRIDHAR, Karthik; BEZAWADA, Ram; TRIVEDI, Minakshi. Investigating the drivers of consumer cross-category learning for new products using multiple data sets. **Marketing Science**, [*s.l*], v. 31, n. 4, p. 668–688, 2012.

STIGLER, Matthieu. **Threshold cointegration: overview and implementation in R**. [*s.l: s.n.*]. 2010. Available on: https://cran.r-project.org/package=tsDyn/vignettes/ThCointOverview.pdf. Retrieved 10 July 2019

STREMERSCH, Stefan; LANDSMAN, Vardit; VENKATARAMAN, Sriram. The relationship between DTCA, drug requests, and prescriptions: Uncovering variation in specialty and space. **Marketing Science**, [*s.l*], v. 32, n. 1, p. 89–110, 2013.

SUN, Zhan-Li *et al.* Sales forecasting using extreme learning machine with applications in fashion retailing. **Decision Support Systems**, [*s.l*], v. 46, n. 1, p. 411–419, 2008.

SUNDSØY, Pål *et al.* Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling. In: International Conference on Social Computing, Behavioral-Cultural Modeling, And Prediction 2014. **Proceedings […]**. Springer, 2014. Available on: http://link.springer.com/chapter/10.1007/978-3-319-05579-4_45. Retrieved 1 Sept. 2016.

TELLER, Christoph; REUTTERER, Thomas. The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? **Journal of Retailing and Consumer Services**, [*s.l*], v. 15, n. 3, p. 127–143, 2008.

TELLIS, Gerard J.; FRANSES, Philip Hans. Optimal data interval for estimating advertising response. **Marketing Science**, [*s.l*], v. 25, n. 3, p. 217–229, 2006.

The Comprehensive R Archive Network. [*s.d.*]. Available on: https://cran.r-project.org/. Retrieved 3 May 2019.

The MASB Common Language Project. [*s.d.*]. Available on: http://www.marketing-dictionary.org/Sales+Potential+Forecast. Retrieved 10 Sept. 2016.

VARIAN, Hal R. Big data: New tricks for econometrics. **The Journal of Economic Perspectives**, [*s.l*], v. 28, n. 2, p. 3–27, 2014.

WACKER, John G.; LUMMUS, Rhonda R. Sales forecasting for strategic resource planning. **International Journal of Operations & Production Management**, [*s.l*], v. 22, n. 9, p. 1014–1031, 2002.

WANG, Earo; COOK, Di; HYNDMAN, Rob J. **tsibble: Tidy Temporal Data Frames and Tools**. [*s.l: s.n.*]. 2018. Available on: https://CRAN.R-project.org/package=tsibble. Retrieved 10 July 2019

WANG, Gang *et al.* Sentiment classification: The contribution of ensemble learning. **Decision Support Systems**, [*s.l*], v. 57, p. 77–93, 2014.

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

WEINGESSEL, Andreas. **quadprog: Functions to solve Quadratic Programming Problems.** [*s.l: s.n.*]. 2013. Available on: https://CRAN.R-project.org/package=quadprog. Retrieved 10 July 2019

WICKHAM, Hadley. **tidyverse: Easily Install and Load the "Tidyverse"**. [*s.l: s.n.*]. 2017. Available on: https://CRAN.R-project.org/package=tidyverse. Retrieved 10 July 2019

WICKRAMASURIYA, Shanika L.; ATHANASOPOULOS, George; HYNDMAN, Rob J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. **Journal of the American Statistical Association**, [*s.l*], p. 1–16, 2018.

WILLIAMS, Brent D.; WALLER, Matthew A. Top-down versus bottom-up demand forecasts: the value of shared point-of-sale data in the retail supply chain. **Journal of Business Logistics**, [*s.l*], v. 32, n. 1, p. 17–26, 2011.

XIONG, Guiyang; BHARADWAJ, Sundar. Prerelease buzz evolution patterns and new product performance. **Marketing Science**, [*s.l*], v. 33, n. 3, p. 401–421, 2014.

YANG, Sha; ALLENBY, Greg M. Modeling interdependent consumer preferences. **Journal of Marketing Research**, [*s.l*], v. 40, n. 3, p. 282–294, 2003.

ZHANG, Wei; KALRA, Ajay. A Joint Examination of Quality Choice and Satisfaction: The Impact of Circumstantial Variables. **Journal of Marketing Research**, [*s.l*], v. 51, n. 4, p. 448–462, 2014.

ZHANG, Yong *et al.* A weighted voting classifier based on differential evolution. In: **Abstract and Applied Analysis 2014**. Hindawi Publishing Corporation, 2014.

ZHAO, Yi *et al.* Modeling consumer learning from online product reviews. **Marketing Science**, [*s.l*], v. 32, n. 1, p. 153–169, 2013.

ZHAO, Yi; ZHAO, Ying; HELSEN, Kristiaan. Consumer learning in a turbulent market environment: Modeling consumer choice dynamics after a product-harm crisis. **Journal of Marketing Research**, [*s.l*], v. 48, n. 2, p. 255–267, 2011.

ZOTTERI, Giulio; KALCHSCHMIDT, Matteo; CANIATO, Federico. The impact of aggregation level on forecasting performance. **International Journal of Production Economics**, [*s.l*], v. 93, p. 479–491, 2005.

APPENDIX A – GOOGLE PLACES' DATA SET

To collect Google Places data set, we, first had to split the city of São Paulo for some reasons: (1) Google uses a relevance criterion to return places located in a certain area; (2) it limits the number of results per query to 60 observations/places. Thus, for every place type with more than 60 results, it was necessary to run the query more than once; and (3) each of the 90 different place types had a specific quantity of results in each area. Thus, there was not only one circle that could get all existing results.

We tried to find the optimal split of Sao Paulo city using zip codes (CEPs) and also the optimal distance (in radius) from each chosen split. Based on Google Maps API libraries for Python and googleway package (COOLEY, 2018), we found that it was not possible to split the city into squares, what would be the best option to save time and money. Thus, we decided to develop the code based on circles around a set of latitude and longitude coordinates and its corresponding radius.

Considering the tests and the constraints we described, we concluded that the best approach would be to split the city of Sao Paulo in circles with minimal intersections among them; center each circle by a latitude and longitude coordinate; and use the quantity (density) of ZIP codes (called "CEP" in Brazil) as a proxy of the number of places of a certain area. Since CEPS are defined by the Brazilian Postal Service, we concluded that this would be a good proxy for the number of people in each area. We used, for the great majority of places, a crescent number of circles, starting from the city center in direction to its borders. Finally, we calculated the radius of each circle following the same logic of CEP's density. Our goal was to have a reasonable amount of duplicated observations, to guarantee that we would collect all important places of each circle, but not too many duplicates that would make the data collection expensive. Based on trial and error, we decided that a radius that resulted in a range of 10% to 30% of duplicates was reasonable for the resources we had available.

We tested different queries, such as by address and CEPs (zip codes), with different radii size. We acquired from Cepaberto a database of all CEPs from Brazil. We filtered the CEPs that starts on 01000-xxx and finishes on 05894-xxx, corresponding to the city of Sao Paulo. Even with this filter, the total number of CEPs was substantial (51,753 unique CEPs) and would result in a very expensive data collection. To check if the selection of full CEPs (i.e., the ones that finish with -000) would work as a reasonable distance between 2 CEPs, we randomly checked many place types directly on the web, using full CEPs addresses. As it seemed to work, we considered the quantity of 2,800 full CEPs as our starting point to gather the latitude and longitude coordinates. However, many of the CEPs did not have accurate coordinates associated with it. For that reason, we decided to use full CEPs just as a starting point, but not as our final search criteria.

Before running the queries, we searched manually on Google Maps to have an approximation of expected sizes of the final datasets by place type. There were 90 different places types in Google Places, with surprisingly different number of results from what was expected. For example, a manual search for "electrician" returns around 220 observations. The same search made by the API query returned 1807 unique observations. For that reason, for some type of places, queries need to run more than once. We proceeded based on the expected sizes and split Sao Paulo into circles and radii. Following, we ran the corresponding queries. For the low frequent places (e.g., airport, casino, and zoo) just one circle with a 38,400m radius was enough. We applied an increasing number of circles and radii, as the number of expected results grew. We also applied different search criteria to cover the city, depending on the number of expected results: 4 circles of 14,500m; 11 smaller circles in Downtown, plus 3 larger circles in the rest of the city; 90 circles; 330 circles; and finally 4929 circles for place types (i.e. stores) with larger numbers of expected results. At the end of the data collection, the total number of observations (including duplicates) was 437,763, and the number of unique places 271,001.

We collected and stored Google Place data on 89 different csv files, one for each type of place. Subsequently, we combined all the datasets into one and excluded the duplicated places (rows) solely based on the place ID. Place IDs uniquely identify a place in Google Places API. The next step was to use the many types listed on the data to correctly classify each place. We created two new variables (type 2 and type 3) with the first two strings of Google's type classification. Google allows more than one type of place to be assigned to each place and returns it as a list when you collect data using the Google Places API. Then a third variable, "clean type" was created. The reasoning behind the "clean type" variable was that if the second string (Type 3) is a generic "point of interest" (POI), then the most correct classification would be on the first string (Type 2). However, if the second string was different from POI, the second string was, in general, more informative. This means that if a place is of type "zoo" and only this, usually the second string is "point of interest". If it is a store inside a zoo, the second string is "store" and it is much more informative to our purposes, so the code will keep that type (the second string or type 3). The same happens, for example, with travel agencies that also exchange currency, they are identified as "finance" with our code.

This, however, does not eliminate all misclassifications, for example, a tattoo studio was found classified as "Hindu Temple" (first string) and "place of worship" (second string), the same as the real temples. This would still be misclassified after running the code. Also, there are some user inputs that are full of mistakes. Some houses and buildings are classified as bank, museums or art galleries, probably as a user joke. To reduce mistakes like that, as much as possible, we made a qualitative analysis of the data by place type and wrote specific lines of code to clean them. This solution was not automated, took longer to code, and it is not mistake proof. The code written changed the type of specific rows based on its original type or place ID. Names and addresses were avoided as much as possible to allow anyone to run the code without running into errors over special Latin characters. We had to use strings, however, to detect the correct shopping malls, airports and museums, but those do not have special characters.

We identified more than 300 rows that had the same coordinates (latitude and longitude) but with different addresses assigned. This was identified by visual inspection when plotting the data on Sao Paulo's map. To correctly filter these rows, we created variables that counted rows with the same coordinates and addresses (this variable was also used to identify commercial centers). We could deduce that this happens for two reasons: (1) if the address of the entry is not completely identified, Google gives the coordinates of the nearest place it can find. Praça da Sé coordinates (latitude -23.5505199, longitude -46.6333094), the ground zero of the city, is given to the entry if it can detect that it belongs to the city of Sao Paulo. If only the state is clear, a general coordinate pointing to the state of Sao Paulo is given. However, if the address has some information about the neighborhood, Google will give the coordinates of some point near that neighborhood. For example, it gives latitude -23.59326 and longitude -46.60794 to the neighborhood of Ipiranga. It is not, however, the center of the neighborhood; and (2) all other cases are, as we suspected, commercial centers such as shopping malls, hospitals, bus stations, or airports. We decided to keep the rows with neighborhood coordinates. But the ones with ground zero or state coordinates were too inaccurate, so we excluded them.

Finally, we decided to delete from the data set helipads. They are not representing unique places since the buildings that they are located on were already on the data (for example, a helipad on the rooftop of a bank. The bank is more relevant for our research purposes than the helipad). Some places that were clearly duplicated, a mistake (such as a house classified as an airport or a cemetery classified as a shopping mall), or were not fit to any of the types (an NPO), were deleted as well. At the end of the preparation, the total number of unique observations was 270,450. The variables of the Google Places data set are described in Table 16.

Table 16 – Google Places' variables

Variable	Format	Description
lat	Numerical	Latitude
lon	Numerical	Longitude
g_code	Numerical	Google Place code, ranging from 1 to 90
g_type	Text	Google Place type
radius	Numerical	Size of the radius of the query that generated the result (in meters)
name	Text	Name of the place
address	Text	Address of the place
place_id	Text	Unique Google identity (id) of each place
rating	Numerical	Rating of the place (from 1 to 5)
user_ratings	Numerical	Number of ratings of the place given by users
type	List	Classification of the place. Each place may receive more than one classification

Source: The authors (2019)

APPENDIX B – ANAMACO'S DATA SET

To prepare the data we followed three steps (1) filtered only São Paulo observations, and make corrections (i.e. some wrong latitude and longitude data); (2) translated the variables and its values to English, standardizing the observations; and (3) checked for duplicates.

We have found no duplicates in "cnpj" variable (the unique identification from the Brazilian Government of each company), but found many duplicates in the variable "name_publc", what was expected because there were many branches of the same company in the data set. We decided to adopt the following criteria to eliminate a duplicate observation: to be deleted, an observation needed to have the same name and address.

The original Anamaco data set has 57 variables and 335,337 observations. Most of the variables are not useful for our purposes, so we filtered the ones we chose to test. At the end of the preparation, the total number of observations was 24,341. The final variables and observations of the Anamaco data set are described in Table 17.

Table 17 – Anamaco's variables

Variable	Description	Scale
cnpj	Unique identification from the Brazilian Government	Nominal
rais	Code of activity in the Ministry of Work	Nominal
type	Head Office of Branch	Categorical (head office or branch)
name_reg	The official name of the company	Nominal
name_publc	Public name of the company	Nominal
address	Full address	Nominal
neighb	Declared neighborhood	Nominal
neighb_cep	Neighborhood located by the cep	Nominal
cep	Postal code	Numeric
state	State of Brazil	Nominal
city	Name of the city	Nominal
lat	Latitude	Numeric
lon	Longitude	Numeric
buss_type	Type of business	Categorical
employees_store	Number of employees in the store	Numeric
employees_company	Number of employees in the company	Numeric

revenue	Declared revenue	Categorical (7 ranges)
year_foundation	Year of foundation	Numeric

Source: The authors (2019)

APPENDIX C – TURKISH DATA SET

We did not have available information of all Istanbul's districts to estimate the gravitational model. For that reason, we decided to use the same districts on the forecasts and the gravitational model, and not consider the sales information of 2 districts. Another issue at the cleaning process of the Turkish data was that we had to input the districts on the 2015 data set. This resulted in different values from the total data set available, for the following reasons:

1. We considered fewer districts than all the data;
2. We ended up with fewer clients than all the data - 488 of the 909 clients that made a purchase on 2015 were not on the 2016 and 2017 data sets and was not possible to input the district;
3. Negative numbers' effect - for 2017 the total number increases because the rows that were not considered (2 districts with no data for gravitational model) had some negative numbers.

Table 18 shows the differences in the total dataset available and the data used in the model.

Table 18 – Turkish data set

	All data available	**36 districts' data**
2015	167,805,490.50	91,181,759.40
2016	102,871,543.40	102,260,926.47
2017	129,594,964.70	131,584,900.00
Total	400,271,999.00	325,027,576.00

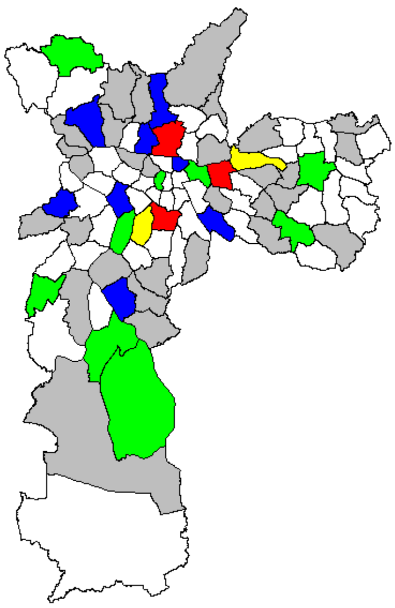
Source: The authors (2019)

APPENDIX D – OUTLIER ANALYSIS

On Brazil's case, we have 18 districts where our approach is the best performing and 20 where it is close to the best performing. If we consider that a difference in MASE up until 1 is good, our model could still be improved in 14 districts (green, yellow and red in the map). However, only 5 of them have more extreme differences (represented in yellow and red). Districts with differences lower than 1 (blue and grey in Figure 21) are not presented on the table.

Figure 21 – MASE by outlier district in São Paulo and map

ITAQUERA	1.11
GRAJAU	1.26
CID DUTRA	1.46
REPUBLICA	1.50
BELEM	1.59
SAO MATEUS	1.91
CAPAO REDONDO	2.18
ITAIM BIBI	2.39
PERUS	5.49
MOEMA	10.19
PENHA	14.36
SANTANA	25.91
TATUAPE	31.25
VILA MARIANA	34.04

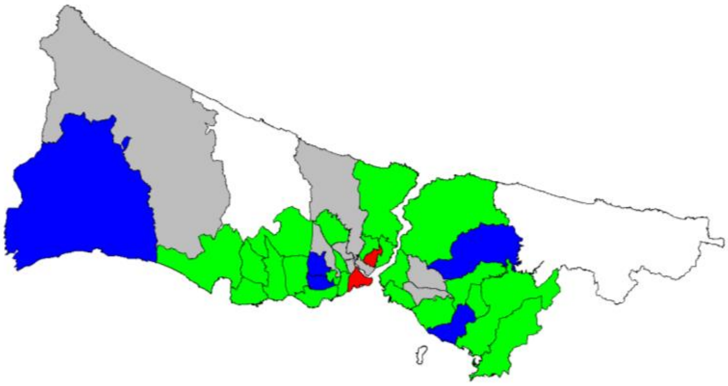


Source: The authors (2019)

On Turkey's case, we have 2 districts on which our approach is the best performing; 10 districts performing very similar to the best approach; 22 districts between 1 and 10; and only 2 extremely different districts, in red in Figure 22.

Figure 22 – MASE by outlier district in Istanbul and map

BUYUKCEKMECE	1,12
PENDIK	1,14
BAYRAMPASA	1,2
SARIYER	1,23
GUNGOREN	1,43
MALTEPE	1,44
TUZLA	1,53
BEYKOZ	1,62
ESENYURT	1,68
SULTANGAZI	1,83
USKUDAR	1,88
BEYLIKDUZU	2,36
BASAKSEHIR	2,47
ZEYTINBURNU	2,71
KAGITHANE	2,86
BESIKTAS	3,05
AVCILAR	3,07
SANCAKTEPE	3,51
BAKIRKOY	4,08
SULTANBEYLI	4,3
KUCUKCEKMECE	4,38
KADIKOY	5,84
SISLI	15,85
FATIH	38,3



APPENDIX E – RESULTS ON ALL LEVELS AND ERROR METRICS

Tables 19 to 24 present the error metrics of São Paulo and Istanbul reconciled forecasts with the three approaches compared (optimal, top-down and GSFR).

Table 19 – São Paulo's results with the optimal approach

Optimal	MASE	ME	RMSE	MAE
Total	0,49	-$2.998.167,00	$12.246.884,00	$11.894.903,00
Center	0,15	-$2.084.151,00	$2.458.335,00	$2.084.151,00
East	1,38	$2.868.209,00	$8.010.282,00	$6.092.325,00
North	0,45	-$4.189.218,00	$4.911.068,00	$4.189.218,00
West	0,36	-$2.144.804,00	$3.066.134,00	$2.451.619,00
South	1,99	$2.551.796,00	$7.329.308,00	$5.621.781,00
República	0,2	-$2.742.530,18	$2.747.987,70	$2.742.530,20
Santa Cecília	0,55	$658.379,50	$1.384.340,30	$1.046.932,10
Artur Alvim	1,8	-$288.852,21	$427.157,30	$356.043,90
Água Rasa	0,73	$169.020,39	$569.037,50	$483.046,60
Aricanduva	3,88	$4.074.372,51	$8.568.462,50	$5.744.538,20
Belém	6,47	-$367.333,89	$458.720,50	$367.333,90
Cangaíba	1,29	-$296.162,42	$544.424,70	$516.912,40
Cidade Líder	1,17	-$336.640,22	$458.288,40	$400.491,80
Iguatemi	1,07	-$211.824,33	$461.722,10	$389.924,80
Itaquera	1,02	-$449.802,94	$494.649,50	$449.802,90
Jardim Helena	2,99	-$427.556,67	$447.620,50	$427.556,70
Pari	2,81	$2.750.574,31	$4.568.068,00	$3.013.067,60
Penha	7,75	-$398.412,72	$459.408,70	$398.412,70
Sapopemba	4,64	-$464.439,43	$503.709,70	$464.439,40
São Mateus	1,08	$595.234,89	$646.082,00	$595.234,90
São Rafael	2,37	-$464.439,43	$503.709,70	$464.439,40
Tatuapé	6,01	-$465.268,20	$504.474,00	$465.268,20
Vila Curuçá	0,74	-$300.197,59	$441.469,90	$344.015,90
Vila Matilde	1,53	$34.899,12	$637.403,60	$578.777,00
Vila Prudente	2,16	-$284.961,73	$381.602,40	$293.399,60

Brasilândia	0,71	-$153.839,24	$313.603,40	$278.379,00
Cachoeirinha	4,63	-$384.840,92	$422.974,20	$384.840,90
Casa Verde	0,7	-$418.146,53	$475.152,80	$418.146,50
Freguesia do Ó	0,25	$290.559,78	$706.461,20	$509.683,70
Jaçanã	3,58	$5.072,90	$492.733,30	$446.615,10
Jaraguá	1,04	-$153.676,53	$395.450,40	$389.629,30
Mandaqui	2,01	$221.301,63	$679.748,50	$531.840,10
Perus	4,17	$325.139,91	$325.823,10	$325.139,90
Pirituba	0,38	-$2.871.667,79	$3.771.358,80	$3.244.654,20
Santana	2,03	-$361.303,47	$402.108,40	$361.303,50
São Domingos	0,9	-$384.840,92	$422.974,20	$384.840,90
Tremembé	1,32	-$384.840,92	$422.974,20	$384.840,90
Vila Maria	1,94	$81.864,27	$830.043,30	$747.067,30
Barra Funda	0,3	-$1.581.316,58	$2.103.596,40	$1.823.214,20
Itaim Bibi	0,87	$84.496,70	$297.336,20	$292.887,60
Pinheiros	0,88	$408.707,53	$858.378,10	$824.595,60
Rio Pequeno	3,17	-$256.925,56	$274.469,70	$256.925,60
Raposo Tavares	21,11	-$352.388,16	$377.589,30	$352.388,20
Vila Leopoldina	0,1	-$100.761,42	$199.558,70	$141.616,70
Vila Sônia	0,45	-$346.616,30	$376.150,80	$346.616,30
Cidade Ademar	3,43	-$388.461,19	$413.389,80	$388.461,20
Cidade Dutra	1,67	$231.291,09	$1.148.699,20	$983.720,30
Campo Grande	0,87	-$382.756,58	$420.624,60	$382.756,60
Campo Limpo	1,9	-$479.869,70	$513.608,50	$479.869,70
Capão Redondo	3,37	$6.020.539,29	$10.468.325,30	$6.533.021,30
Grajaú	6,69	-$362.431,44	$437.215,40	$362.431,40
Jabaquara	1,89	-$375.833,64	$418.054,70	$375.833,60
Jardim São Luís	1,11	-$394.051,19	$522.841,60	$433.776,80
Moema	2,5	-$340.323,28	$401.080,00	$340.323,30
Parelheiros	1,49	-$288.354,16	$400.328,70	$315.810,30
Pedreira	2,14	-$361.592,84	$394.533,40	$361.592,80
Sacomã	1,91	-$223.596,60	$296.219,40	$246.464,40
Santo Amaro	4,63	$43.877,51	$772.812,20	$698.736,30
Vila Mariana	9,95	-$146.641,63	$479.548,60	$450.211,50

Source: The authors (2019)

Table 20 – São Paulo's results with the top-down approach

TOP-DOWN	MASE	ME	RMSE	MAE
Total	0,47	-$ 5.068.027,00	$ 11.749.640,00	$ 11.456.406,00
Center	0,83	-$ 11.969.589,00	$ 14.410.649,00	$ 11.969.589,00
East	2,57	$ 11.349.161,00	$ 16.207.851,00	$ 11.349.161,00
North	0,79	-$ 3.194.681,00	$ 8.545.425,00	$ 7.300.546,00
West	1,38	-$ 5.824.347,00	$ 10.220.661,00	$ 9.438.725,00
South	4,68	$ 4.571.429,00	$ 16.315.996,00	$ 13.216.200,00
República	1,03	-$ 13.902.568,07	$ 16.325.772,25	$ 13.902.568,07
Santa Cecília	1,02	$ 1.932.978,60	$ 2.465.492,71	$ 1.932.978,60
Artur Alvim	2,01	$ 273.282,31	$ 530.503,98	$ 398.061,50
Água Rasa	0,65	$ 431.772,61	$ 593.522,68	$ 431.772,61
Aricanduva	3,85	$ 5.536.399,03	$ 9.285.679,73	$ 5.705.087,26
Belém	3,77	-$ 50.723,40	$ 249.519,55	$ 214.194,55
Cangaíba	0,94	-$ 216.673,87	$ 533.554,91	$ 377.941,34
Cidade Líder	1,28	$ 13.680,08	$ 515.880,82	$ 436.964,20
Iguatemi	0,83	$ 211.930,48	$ 362.220,19	$ 305.070,68
Itaquera	0,2	$ 88.832,13	$ 153.861,77	$ 88.832,13
Jardim Helena	2,55	-$ 364.115,68	$ 503.893,05	$ 364.115,68
Pari	3,25	$ 3.485.231,49	$ 5.244.164,85	$ 3.485.231,49
Penha	1,61	$ 82.982,47	$ 143.729,85	$ 82.982,47
Sapopemba	-	$ -	$ -	$ -
São Mateus	3,45	$ 1.901.154,19	$ 2.190.836,83	$ 1.901.154,19
São Rafael	-	$ -	$ -	$ -
Tatuapé	4,15	-$ 321.455,76	$ 556.777,71	$ 321.455,76
Vila Curuçá	0,44	$ 85.206,77	$ 272.702,73	$ 205.712,56
Vila Matilde	2,43	$ 112.735,81	$ 997.760,58	$ 922.728,30
Vila Prudente	1,21	$ 78.922,34	$ 175.304,35	$ 163.966,09
Brasilândia	1,22	$ 307.751,31	$ 697.985,13	$ 479.664,43
Cachoeirinha	2,93	-$ 243.278,33	$ 297.970,76	$ 243.278,33
Casa Verde	0,35	$ 209.169,74	$ 256.921,63	$ 209.169,74
Freguesia do Ó	0,52	$ 368.192,98	$ 1.159.496,44	$ 1.050.498,54
Jaçanã	3,61	$ 117.639,98	$ 569.871,65	$ 450.181,41
Jaraguá	0,77	$ 193.746,45	$ 353.638,71	$ 286.886,65
Mandaqui	3,25	$ 861.786,19	$ 1.198.214,63	$ 861.786,19
Perus	9,73	$ 759.144,82	$ 782.846,43	$ 759.144,82
Pirituba	0,78	-$ 5.730.956,66	$ 8.965.370,65	$ 6.667.903,10
Santana	0,14	$ 25.453,18	$ 35.904,70	$ 25.453,18

São Domingos	0,76	-$ 327.706,22	$ 412.955,85	$ 327.706,22
Tremembé	0,05	-$ 15.803,68	$ 19.517,11	$ 15.803,68
Vila Maria	1,74	$ 280.179,58	$ 879.943,20	$ 670.019,42
Barra Funda	1,13	-$ 4.416.998,63	$ 7.190.133,02	$ 6.819.899,41
Itaim Bibi	1,91	$ 380.273,65	$ 711.589,88	$ 641.972,98
Pinheiros	3,05	-$ 1.637.771,29	$ 3.958.738,05	$ 2.844.819,56
Rio Pequeno	1,41	-$ 114.486,43	$ 142.611,87	$ 114.486,43
Raposo Tavares	3,28	-$ 10.273,93	$ 64.992,83	$ 54.783,26
Vila Leopoldina	0,1	$ 42.676,66	$ 154.174,45	$ 120.233,02
Vila Sônia	0,09	-$ 67.766,96	$ 117.375,81	$ 67.766,96
Cidade Ademar	2,72	-$ 265.198,75	$ 358.500,86	$ 307.178,35
Cidade Dutra	5,49	$ 512.917,41	$ 3.813.851,60	$ 3.226.698,27
Campo Grande	1,13	-$ 499.380,38	$ 648.086,93	$ 499.380,38
Campo Limpo	2,32	-$ 587.237,14	$ 891.794,73	$ 587.237,14
Capão Redondo	6,15	$ 4.657.857,24	$ 15.372.674,35	$ 11.913.299,90
Grajaú	3,6	$ 195.271,84	$ 338.220,75	$ 195.271,84
Jabaquara	0,08	-$ 15.803,68	$ 19.517,11	$ 15.803,68
Jardim São Luís	1,31	-$ 123.074,77	$ 531.049,56	$ 512.521,81
Moema	0,28	$ 37.628,92	$ 65.175,20	$ 37.628,92
Parelheiros	0,15	$ 30.694,49	$ 53.164,42	$ 30.694,49
Pedreira	0,24	-$ 39.970,60	$ 69.231,11	$ 39.970,60
Sacomã	2,66	-$ 177.718,39	$ 464.696,75	$ 343.860,08
Santo Amaro	5,67	$ 591.487,94	$ 1.267.502,80	$ 856.646,96
Vila Mariana	5,61	$ 253.955,19	$ 439.863,29	$ 253.955,19

Source: The authors (2019)

Table 21 – São Paulo's results with GSFR approach

GSFR	MASE	ME	RMSE	MAE
Total	0,47	-$ 5.068.022,00	$ 11.749.638,00	$ 11.456.405,00
Center	1,58	-$22.670.712,00	$ 23.035.355,00	$ 22.670.712,00
East	1,72	$ 7.512.918,00	$ 11.696.156,00	$ 7.575.875,00
North	0,55	$ 4.667.210,00	$ 7.953.782,00	$ 5.049.172,00
West	0,45	-$ 2.823.177,00	$ 3.908.710,00	$ 3.090.739,00
South	3,44	$ 8.245.739,00	$ 13.557.865,00	$ 9.709.581,00
República	1,70	-$23.035.850,00	$ 23.543.400,00	$ 23.035.850,00
Santa Cecília	0,56	$ 365.139,70	$ 1.360.489,00	$ 1.065.841,00
Artur Alvim	1,82	-$ 39.605,03	$ 369.715,60	$ 359.610,80

Água Rasa	0,85	$ 559.562,30	$ 674.956,40	$ 559.562,30
Aricanduva	3,71	$ 5.370.684,00	$ 9.101.005,00	$ 5.497.701,00
Belém	5,36	-$ 285.244,90	$ 369.795,10	$ 304.290,00
Cangaíba	0,73	$ 254.586,90	$ 473.329,10	$ 291.341,70
Cidade Líder	0,66	$ 162.080,40	$ 292.856,90	$ 224.812,50
Iguatemi	0,83	$ 300.803,50	$ 383.830,50	$ 304.555,80
Itaquera	1,31	-$ 574.634,20	$ 601.605,00	$ 574.634,20
Jardim Helena	0,37	$ 53.233,69	$ 92.411,11	$ 53.473,33
Pari	3,55	$ 3.806.573,00	$ 5.326.760,00	$ 3.806.573,00
Penha	15,97	-$ 820.951,80	$ 847.101,40	$ 820.951,80
Sapopemba	4,42	-$ 442.357,10	$ 450.511,60	$ 442.357,10
São Mateus	2,99	$ 1.649.056,00	$ 1.685.145,00	$ 1.649.056,00
São Rafael	-	$ -	$ -	$ -
Tatuapé	35,40	-$ 2.742.962,00	$ 2.793.526,00	$ 2.742.962,00
Vila Curuçá	0,37	$ 153.120,10	$ 284.359,10	$ 174.849,60
Vila Matilde	1,34	$ 384.774,50	$ 729.116,60	$ 506.548,10
Vila Prudente	2,03	-$ 275.800,90	$ 311.722,70	$ 275.800,90
Brasilândia	1,08	$ 308.046,00	$ 636.338,90	$ 422.050,10
Cachoeirinha	1,44	-$ 119.992,40	$ 122.204,40	$ 119.992,40
Casa Verde	1,19	-$ 713.812,00	$ 777.929,40	$ 713.812,00
Freguesia do Ó	0,23	$ 53.045,22	$ 530.440,00	$ 476.678,50
Jaçanã	3,70	$ 249.664,70	$ 629.890,20	$ 461.236,60
Jaraguá	0,75	$ 256.045,80	$ 352.678,40	$ 281.665,90
Mandaqui	2,86	$ 756.442,90	$ 1.112.881,00	$ 756.442,90
Perus	9,66	$ 753.434,60	$ 777.124,30	$ 753.434,60
Pirituba	0,93	$ 7.980.394,00	$ 9.398.109,00	$ 7.980.394,00
Santana	26,05	-$ 4.628.009,00	$ 4.710.977,00	$ 4.628.009,00
São Domingos	0,12	-$ 50.702,69	$ 51.637,35	$ 50.702,69
Tremembé	0,16	-$ 47.050,52	$ 47.917,86	$ 47.050,52
Vila Maria	1,87	-$ 130.296,20	$ 732.369,50	$ 720.569,00
Barra Funda	0,24	-$ 120.153,60	$ 1.541.395,00	$ 1.450.294,00
Itaim Bibi	3,26	-$ 1.097.389,00	$ 1.133.922,00	$ 1.097.389,00
Pinheiros	1,75	-$ 1.630.904,00	$ 2.054.760,00	$ 1.630.904,00
Rio Pequeno	2,01	$ 146.556,10	$ 193.147,50	$ 163.157,20
Raposo Tavares	1,54	$ 25.634,68	$ 36.376,18	$ 25.634,68
Vila Leopoldina	0,10	$ 75.314,24	$ 171.678,10	$ 128.682,50
Vila Sônia	0,29	-$ 222.235,20	$ 228.275,50	$ 222.235,20
Cidade Ademar	1,29	-$ 145.969,40	$ 153.823,90	$ 145.969,40
Cidade Dutra	3,13	$ 1.539.868,00	$ 2.936.917,00	$ 1.840.241,00

Campo Grande	1,53	-$ 676.403,40	$ 687.722,30	$ 676.403,40
Campo Limpo	-	-$ 12,89	$ 13,13	$ 12,89
Capão Redondo	5,55	$ 10.382.880,00	$ 14.607.210,00	$ 10.757.350,00
Grajaú	4,86	-$ 10.827,97	$ 278.678,30	$ 263.462,50
Jabaquara	0,39	-$ 78.260,10	$ 79.702,77	$ 78.260,10
Jardim São Luís	0,66	-$ 46.595,65	$ 260.444,50	$ 256.478,70
Moema	10,47	-$ 1.425.904,00	$ 1.463.542,00	$ 1.425.904,00
Parelheiros	0,49	$ 57.838,69	$ 142.474,10	$ 103.322,30
Pedreira	0,25	$ 42.014,25	$ 73.639,19	$ 43.013,97
Sacomã	1,78	-$ 229.692,10	$ 243.285,00	$ 229.692,10
Santo Amaro	5,00	$ 631.483,90	$ 1.202.972,00	$ 754.699,80
Vila Mariana	39,65	-$ 1.794.680,00	$ 1.937.459,00	$ 1.794.680,00

Source: The authors (2019)

Table 22 – Istanbul's results with the optimal approach

Optimal	**MASE**	**ME**	**RMSE**	**MAE**
Total	0,90	$ 1.351.827,00	$ 3.073.935,00	$ 2.127.392,00
Atasehir	0,52	-$ 45.686,66	$ 72.273,89	$ 66.747,13
Avcilar	0,50	-$ 16.520,28	$ 75.261,29	$ 68.267,20
Bagcilar	2,19	$ 129.978,65	$ 304.882,23	$ 215.341,38
Bahcelievler	1,05	$ 21.074,40	$ 238.709,53	$ 212.734,68
Bakirkoy	3,81	-$ 15.907,19	$ 83.822,07	$ 82.229,28
Basaksehir	1,18	-$ 134,57	$ 70.984,83	$ 64.630,75
Bayrampasa	0,49	-$ 9.026,94	$ 24.891,46	$ 24.524,03
Besiktas	1,79	$ 64.070,35	$ 221.004,13	$ 175.741,40
Beykoz	0,72	-$ 68.598,85	$ 73.383,67	$ 68.598,85
Beylikduzu	2,29	$ 2.755,99	$ 126.226,86	$ 117.446,03
Beyoglu	0,98	$ 50.335,08	$ 100.368,07	$ 91.645,66
Buyukcekmece	1,11	-$ 27.495,06	$ 86.723,20	$ 86.647,26
Catalca	0,94	-$ 41.178,31	$ 41.386,66	$ 41.178,31
Cekmekoy	1,58	-$ 30.657,94	$ 43.725,08	$ 39.414,60
Esenler	0,12	$ 16.353,03	$ 36.626,33	$ 35.572,06
Esenyurt	1,18	$ 3.747,02	$ 103.557,81	$ 95.539,37
Eyup	0,52	$ 89.870,49	$ 102.934,53	$ 89.870,49
Fatih	0,78	-$ 9.078,60	$ 26.397,03	$ 20.284,64
Gaziosmanpasa	0,61	-$ 33.973,63	$ 39.103,01	$ 33.973,63
Gungoren	0,24	-$ 34.635,77	$ 35.019,52	$ 34.635,77

Kadikoy	0,85	-$ 11.746,37	$ 116.311,01	$ 105.170,21
Kagithane	1,25	$ 147.640,77	$ 370.971,49	$ 271.644,48
Kartal	1,27	$ 58.221,06	$ 265.282,81	$ 223.313,66
Kucukcekmece	1,20	$ 37.556,94	$ 96.315,65	$ 85.982,99
Maltepe	1,20	$ 170.220,11	$ 250.141,99	$ 228.926,70
Pendik	0,23	$ 32.836,04	$ 70.250,57	$ 47.604,80
Sancaktepe	0,54	-$ 9.909,75	$ 38.825,15	$ 38.693,27
Sariyer	1,33	-$ 9.031,17	$ 210.451,86	$ 186.832,16
Silivri	0,66	$ 6.137,48	$ 84.486,26	$ 76.792,29
Sisli	0,95	$ 14.291,44	$ 82.659,82	$ 81.209,82
Sultanbeyli	3,67	$ 633.755,03	$ 681.981,63	$ 633.755,03
Sultangazi	3,07	$ 59.205,92	$ 191.486,49	$ 151.828,79
Tuzla	0,73	-$ 41.643,29	$ 42.008,45	$ 41.643,29
Umraniye	1,89	-$ 191.304,49	$ 489.483,95	$ 487.222,04
Uskudar	0,55	$ 15.129,59	$ 66.447,75	$ 55.894,22
Zeytinburnu	2,51	$ 395.176,88	$ 574.019,54	$ 442.705,05

Source: The authors (2019)

Table 23 – Istanbul's results with the top-down approach

Top-down	**MASE**	**ME**	**RMSE**	**MAE**
Total	0,91	$ 1.288.863,00	$ 3.059.872,00	$ 2.146.912,00
Atasehir	1,56	-$ 200.816,15	$ 214.417,66	$ 200.816,15
Avcilar	1,74	$ 218.555,22	$ 312.225,65	$ 236.450,04
Bagcilar	3,35	$ 301.700,80	$ 443.959,66	$ 329.622,09
Bahcelievler	1,5	$ 272.232,59	$ 437.379,00	$ 301.658,07
Bakirkoy	3,45	$ 74.482,69	$ 117.553,11	$ 74.482,69
Basaksehir	1,16	$ 61.323,04	$ 99.019,01	$ 63.304,57
Bayrampasa	0,66	$ 33.554,02	$ 39.049,76	$ 33.554,02
Besiktas	0,83	$ 81.229,16	$ 83.479,61	$ 81.229,16
Beykoz	2,61	-$ 248.505,19	$ 280.879,40	$ 248.505,19
Beylikduzu	2,01	$ 102.211,23	$ 172.140,40	$ 102.958,06
Beyoglu	1,23	-$ 31.809,71	$ 122.921,04	$ 115.538,47
Buyukcekmece	0,39	-$ 12.736,04	$ 35.073,98	$ 30.222,88
Catalca	0,85	-$ 34.568,07	$ 58.379,43	$ 37.206,98
Cekmekoy	1,08	$ 8.064,71	$ 32.335,90	$ 26.832,85
Esenler	0,11	$ 2.178,66	$ 34.981,79	$ 32.190,84
Esenyurt	1,83	$ 89.830,64	$ 209.931,47	$ 147.933,24

Eyup	0,89	$ 128.111,41	$ 187.939,62	$ 154.527,53
Fatih	0,71	-$ 18.527,93	$ 20.737,41	$ 18.527,93
Gaziosmanpasa	0,69	-$ 38.114,00	$ 47.749,04	$ 38.114,00
Gungoren	0,88	$ 127.426,24	$ 139.547,29	$ 127.426,24
Kadikoy	1,32	-$ 105.241,78	$ 178.285,69	$ 163.669,86
Kagithane	1,37	$ 298.828,92	$ 439.857,80	$ 298.828,92
Kartal	1,37	-$ 166.214,11	$ 331.823,04	$ 241.027,82
Kucukcekmece	2,65	$ 142.815,12	$ 222.060,00	$ 189.312,10
Maltepe	1,58	-$ 250.011,50	$ 345.277,80	$ 302.509,91
Pendik	2,17	-$ 453.361,72	$ 458.799,74	$ 453.361,72
Sancaktepe	1,19	$ 72.469,65	$ 124.095,39	$ 85.956,23
Sariyer	2,13	-$ 4.910,05	$ 319.057,69	$ 299.836,61
Silivri	0,85	-$ 7.363,24	$ 113.198,55	$ 99.489,74
Sisli	2,09	$ 513,45	$ 199.944,66	$ 179.051,40
Sultanbeyli	2,75	$ 475.016,43	$ 506.399,34	$ 475.016,43
Sultangazi	4,46	$ 220.302,81	$ 319.647,04	$ 220.302,81
Tuzla	0,9	-$ 51.690,64	$ 73.146,13	$ 51.690,64
Umraniye	1,08	-$ 277.033,30	$ 339.354,42	$ 277.033,30
Uskudar	1,57	$ 118.720,80	$ 191.764,02	$ 159.802,13
Zeytinburnu	2,8	$ 360.198,58	$ 579.838,47	$ 493.003,87

Source: The authors (2019)

Table 24 – Istanbul’s results with GSFR approach

GSFR	MASE	ME	RMSE	MAE
Total	0,91	$ 1.288.999,00	$ 3.059.931,00	$ 2.146.867,00
Atasehir	0,81	-$ 104.333,70	$ 122.548,47	$ 104.333,70
Avcilar	3,57	$ 484.562,60	$ 555.533,29	$ 484.562,60
Bagcilar	3,12	-$ 230.980,42	$ 337.143,09	$ 307.257,21
Bahcelievler	1,62	$ 179.802,27	$ 453.714,43	$ 326.346,08
Bakirkoy	7,53	-$ 162.686,59	$ 181.001,17	$ 162.686,59
Basaksehir	3,63	-$ 198.849,76	$ 208.801,31	$ 198.849,76
Bayrampasa	1,69	$ 85.594,39	$ 90.638,96	$ 85.594,39
Besiktas	3,88	-$ 380.322,74	$ 428.057,91	$ 380.322,74
Beykoz	2,34	$ 222.601,02	$ 225.176,25	$ 222.601,02
Beylikduzu	4,37	$ 224.025,71	$ 273.303,13	$ 224.025,71
Beyoglu	1,28	$ 84.635,57	$ 134.206,48	$ 120.158,53
Buyukcekmece	1,51	-$ 84.354,77	$ 126.741,04	$ 117.163,52

Catalca	0,11	$ 4.732,19	$ 5.929,25	$ 4.732,19
Cekmekoy	1,87	$ 46.633,36	$ 53.213,26	$ 46.633,36
Esenler	0,59	$ 170.620,99	$ 174.981,67	$ 170.620,99
Esenyurt	2,86	-$ 231.468,20	$ 250.508,49	$ 231.468,20
Eyup	0,40	$ 37.898,31	$ 72.193,13	$ 69.838,58
Fatih	39,01	-$ 1.014.264,91	$ 1.014.550,70	$ 1.014.264,91
Gaziosmanpasa	0,67	$ 37.087,04	$ 39.453,09	$ 37.087,04
Gungoren	1,67	$ 242.215,21	$ 242.759,98	$ 242.215,21
Kadikoy	6,69	-$ 826.934,61	$ 834.583,11	$ 826.934,61
Kagithane	4,11	$ 893.436,99	$ 975.772,58	$ 893.436,99
Kartal	2,22	$ 389.574,73	$ 463.135,41	$ 389.574,73
Kucukcekmece	5,58	-$ 399.351,21	$ 419.138,96	$ 399.351,21
Maltepe	2,64	$ 503.686,14	$ 535.452,82	$ 503.686,14
Pendik	1,37	$ 286.640,38	$ 290.708,14	$ 286.640,38
Sancaktepe	4,05	$ 291.915,11	$ 295.128,55	$ 291.915,11
Sariyer	2,56	-$ 359.998,08	$ 427.318,35	$ 359.998,08
Silivri	1,38	$ 161.350,17	$ 185.843,37	$ 161.350,17
Sisli	16,80	-$ 1.440.057,35	$ 1.444.428,90	$ 1.440.057,35
Sultanbeyli	7,05	$ 1.218.013,68	$ 1.241.895,26	$ 1.218.013,68
Sultangazi	4,90	$ 242.084,47	$ 329.191,79	$ 242.084,47
Tuzla	2,26	$ 129.163,58	$ 129.201,16	$ 129.163,58
Umraniye	1,53	$ 114.698,22	$ 473.590,15	$ 393.579,57
Uskudar	2,43	-$ 247.074,20	$ 252.500,19	$ 247.074,20
Zeytinburnu	5,22	$ 918.703,58	$ 968.887,86	$ 918.703,58

Source: The authors (2019)