GRAVITATIONAL FORECAST RECONCILIATION

Carla F. Silveira Netto, Mohsen Bahrami, Vinicius Brei, Burçin Bozkaya, Selim Balcısoy, and Alex 'Sandy' Pentland

Abstract: When organizations plan to enter a new market or to expand their business to new locations, they need sales time-series or final consumers scanner data in time and geographical disaggregation to implement their desired marketing strategy. However, such data are hardly available, what may seriously undermine the organizations' disaggregated forecasting efforts. Based on agglomeration and gravitational theories, we propose a new forecast reconciliation approach that distributes and reconciles forecasts to lower levels of aggregation, when no actual disaggregated sales data or historical sales proportions are available. We combine the deep learning technique Long Short-Term Memory (LSTM) applied to time-series forecasting with a new gravitational model approach and validate our method with real sales data of two different companies from two different countries. The results show that our proposed reconciliation approach based on easily or freely available data has a similar or better performance than the benchmark approaches using proprietary sales data at all levels of the companies' channel hierarchy.

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

1 Introduction

Forecasting is a key component of any marketing strategy. To achieve good forecasting results, companies need time-series of sales data, in time and geographical disaggregation corresponding to their marketing strategy. When proprietary time-series data is not available, they can still rely on final consumers' scanner data to predict sales. However, in many situations companies do not have the necessary time-series nor end-consumers scanner data, either in the historical time frame and in the granularity or geographical level they need. Such shortcomings may seriously compromise any forecasting effort. Even more challenging, if those companies want to enter a new market or to expand their business to new locations, they will probably arrive at a new territory having just a total sales estimation for the new country or city. This may be a problem because marketing activities are usually executed at a more local, disaggregated level. This explains why forecasting granularity is strategic to the success of the new operations, either to evaluate their performance, but also because the sales forecasting figures of the new operations will need to be summed up or reconciled with the forecasting of the already established operations of those companies. In short, to predict sales of either established or new operations one needs to access actual sales data on a disaggregated level or at least have historical sales proportions to estimate these forecasts. The problem is that such data may not be available.

In this paper, we solve this problem by proposing a new forecasting reconciliation method. Our approach combines a deep learning technique -- Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) -- adapted to time-series forecasting with a new gravitational model approached inspired in Huff's gravity model (Huff, 1964). This model was originally designed to predict the market share of service facilities but is not capable of forecasting future sales volume, as the time series models do. Our gravitational forecast reconciliation (GFR) method can be used to either forecast sales and to reconcile the granular predictions using gravitational data. We forecast the total (most aggregated level) sales and then reconcile them to lower levels forecasts with a top-down approach.

The major advantage of the GFR method is that, instead of using historical proportions as the standard top-down reconciliation approach, we distribute the total sales (e.g. city level) among more granular levels (e.g., districts) with a gravitational model built with easily/publicly available and/or low-cost geographical POI data., We test and validate our approach with historical sales proprietary data of two different companies from two different countries, a Brazilian electrical products manufacturer and a Turkish paint company, both major players in their respective countries/markets. We compare the performance of our method with the top-down and with the optimal reconciliation approach (Hyndman, Ahmed, Athanasopoulos, & Shang, 2011) using real sales records. The results show that our proposed reconciliation approach has a similar or better performance than the benchmark approaches at all levels of the companies' channel hierarchy.

Our GFR approach brings contributions to the marketing forecasting theory, method, and practice. The key theoretical contribution is to demonstrate how the combination of gravitational models based on agglomeration theory (Liu, Steenkamp, & Zhang (2018) and state-of-the-art forecasting methods (such as LSTM) can be used to predict sales accurately. Our methodological contribution enables companies to distribute and reconcile forecasts to lower levels of aggregation when no actual disaggregated sales data or historical sales proportions are available. For marketing practice, the GFR approach offers several advantages. First, companies can apply it to new territories, points of sales, or marketing channels, distributing the forecasted aggregated sales into smaller granular spatial units. Second, GFR also allows the development of a coherent forecast system by predicting granular sales and market share based on location, quantity, and diversity of points of interest (POIs) of a specific region, according to the desired company strategy. Our method is also robust to estimate forecasts using cross-sectional data from different time periods because it uses relatively stable geographical information (i.e., POIs) that change slowly in most situations. Third, it is easy and relatively cheap to implement in any organization, from all sizes and economic sectors, helping them to estimate the market potential of the new regions or markets where the company plans to direct the market and have no historical sales data. Finally, GFR can be embedded in production planning decision support systems and help companies to devise their future marketing strategies.

The remainder of this paper unfolds as follows. Section 2 positions our contribution via-à-vis the forecasting and gravitational models literature. Section 3 documents the data collection process, followed by Section 4, which details our proposed reconciliation approach. In Section 5, we present our empirical applications to test the GFR accuracy along with its performance comparison with different forecasting methods. Finally, in Section 6 we conclude the paper by summarising key contributions, but also describing some limitations, implications of our method, and presenting suggestions for future research.

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, based on econometrics, such as choice models. In this paper we focus on a solution for sales forecasting optimization.

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.

2.1 Forecasts aggregation criteria

Aggregation is a key aspect in the time series forecasting literature. Fliedner & Mabert (1992) investigated two parameters for constrained forecasting: product group size and appropriate grouping criterion. The authors found that the size of family groups is not of great importance. However, they showed that the homogeneity is important. They also observed that criteria used to form family groups which are based upon demand volume enhance hierarchical forecast system performance. Later, Fliedner & Lawrence (1995) expanded these findings by examining alternative forecast methodologies and techniques for forming family groups within hierarchical systems. 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 et al. (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.

Dekker et al. (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, & 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 regions may help to improve forecasting accuracy.

2.2 Spatial models and forecasting

Chan, Padmanabhan, & Seetharaman (2007) state that location is determinant for retail performance and competition. For instance, while analyzing the impact of distribution intensity (i.e. the number of retailers in a given area) on car sales, Bucklin, Siddarth, & Silva-Risso (2008) demonstrate that store location and its distance to consumers' homes play an important role in sales. However, this is not new. LeSage (1999) states that location and distance are important forces that influence human geography and market activity. According to the diffusion literature in marketing (e.g. Garber, Goldenberg, Libai, & Muller, 2004) the correlation between location and word of mouth spread is well established.

As Bronnenberg & Sismeiro (2002) claim, sales correlation between geographical markets is probably more common than what is believed in the literature since very few studies explicitly consider spatial dependencies. The authors use this correlation with nearby territories to predict performance in markets where there is little or no data available.

Spatial models applied to marketing already demonstrate that spatial factors 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 et al., 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 et al., 2008; Narayanan & Nair, 2013).

In these studies, a general premise is that the behavior of individuals in close proximity are more related and similar than the behavior of those who are more distant (Bradlow et al., 2005). However, it is not only the individual consumers who have their choices influenced by location. Bronnenberg & Mahajan (2001) show that market-shares are different among regions and that this is due to the implicit spatial structure and the spread of points of sales over various territories. According to the authors, this happens because points of sale act locally, giving more attention to their own territories.

One of the most widely used models to estimate spatial attractiveness and hence store patronization is the Huff model (Huff, 1964). This seminal model uses two determining factors for the attractiveness potential of a store: the size of the store (directly proportional); and the distance between stores and consumers (inversely proportional). After the introduction of the original model and its application, the model has been extended in several studies (Nakanishi and Cooper, 1974; Berman and Krass, 2002; Drezner, 2006; Aboolian et al., 2007; Bozkaya et al., 2010). Other researchers have included other variables in the model, such as price, service level, opening hours and brand image (Teller & Reutterer, 2008).

Most studies based on the Huff model use individual location data since it is necessary to estimate the distance between stores and consumers. Our focus, however, is on the spatial relationship that can be explained by a different theory, namely agglomeration, as in Liu, Steenkamp, & Zhang (2018). The concentration of stores from the same sector in the same region 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 et al., 2018).

Fildes et al. (2018) state that many of the current location models see their credibility diminish because of recent changes in consumer behavior, and that the space interaction models (like the Huff-based gravitational models) have now better explanatory power than ad-hoc predictive approaches. To build new location-based models that are highly relevant and effective 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 the manufacturers' and retailers' practiced are disaggregated, while their forecasts are usually done at the most aggregated level.

2.3 Forecast's reconciliation approaches

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 add up following that structure, forecast do not (Figure 1).

Figure 1 – Data and forecasts in a hierarchical structure (example)

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 (1).

$$\tilde{y}_h = R\hat{y}_h \tag{1}$$

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: it forecasts the most disaggregate level and sums 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.

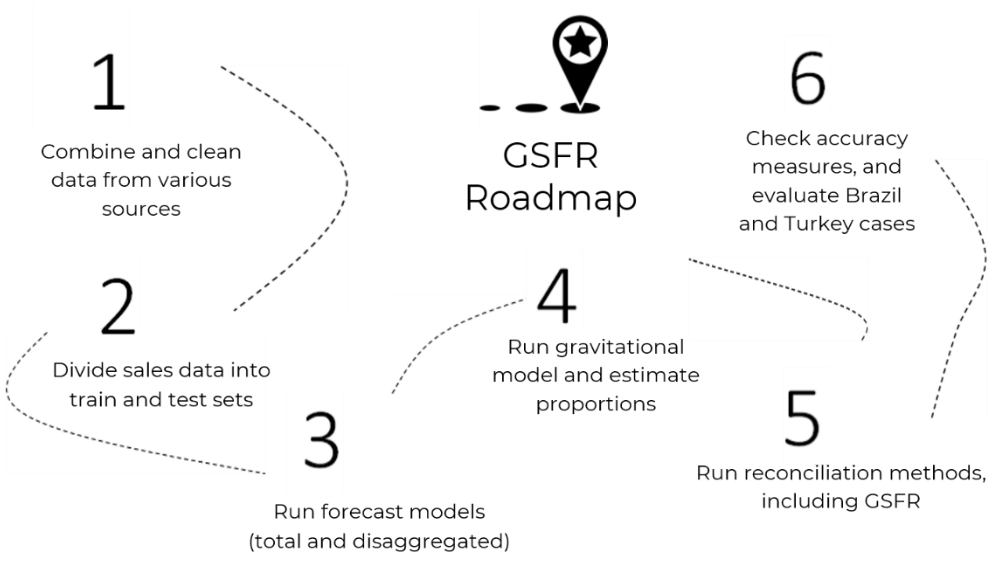
Hyndman *et al.* (2011) propose a fourth approach, named *optimal*, that forecasts all levels and reconciles them using a linear combination giving weights to each forecast unit (nodes of the hierarchy). It considers the structure of the groups or hierarchies, and tends to be more accurate. It is called "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 may not always have this information available, only a total aggregated forecast. In this study, we aim to propose a new reconciliation approach that will use a gravitational model to overcome this difficulty, disaggregating a total forecast to lower levels, without no actual sales 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 as an inspiration to develop a model that will predict the proportion of retail activity in a 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, that we named as gravitational forecast reconciliation (GFR) approach. The next section will present the data we had available and later we detail our empirical strategy.

3. Materials and methods

For our empirical application we followed the roadmap in Figure 2.

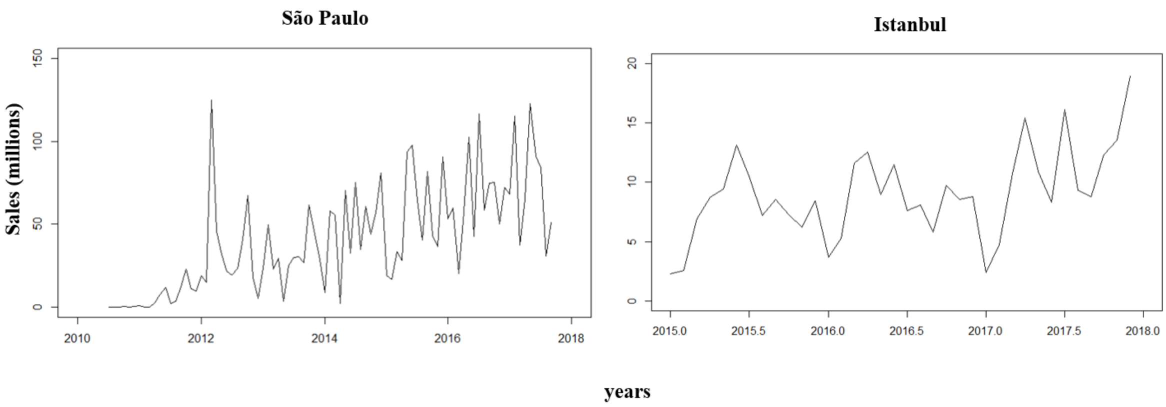
Figure 2 – GFR Roadmap



First, we collected, cleaned, and combined data from various sources in different formats, structures, and timeframes (further details in section 3.1). For example a part of sales data came from distribution centers and remaining came directly from the factory. These files needed to be combined and the variable names changed to be matched. We also solve the problem of differences between similar variables in different files and missing data. Finally, we aggregated the data using geographic considerations and its hierarchical structure, and transformed it into time-series format. Our study cases are two manufacturers based in two different countries in two different continents.

The first case is an electrical components producer in São Paulo, Brazil, and the second case is a paint producer in Istanbul, Turkey. Both metropolitans are the biggest cities of their countries with populations of 12.2 and 15 million habitants respectively. For São Paulo, Brazil, sales records were from July 2010 to September 2017, whereas for Istanbul, from January 2015 to December 2017. Figure 3 shows the time-series after rescaling, to ensure we meet the requirements of a non-disclosure agreement signed by the authors with the data providers.

Figure 3 – São Paulo's and Istanbul's time series



The geographical hierarchy of the Brazilian data comprised of three levels. The top level consists of the total sales for the city of São Paulo, followed by the sales divided in 5 zones (center, east, west, north, and south), and, at the most disaggregate level, sales are divided in 54 of the 96 districts of the city (the missing districts had no sales records during the data time frame). The Turkish data had two levels: the top level is the entire city and a lower level of 39 districts. However, we considered only 36 from which we had data to fit the gravitational model.

3.1 Data

In this study, we make use of various data sets including public and private data sets easily available to most organizations. In Brazil, we concentrated the study in the city of São Paulo (1,521 km^2), and for the case of Turkey, we focus on the Istanbul metropolitan area (1,539 km^2,). We utilized 5 data sets, 3 for Brazil and 2 for Turkey, besides demographic and geographical (i.e. shapefiles) information available on São Paulo's city hall website (http://dados.prefeitura.sp.gov.br/dataset?res_format=SHP, retrieved at 21, October, 2019) and the retail activity index of São Paulo (https://www.bcb.gov.br/estatisticas/indicadoresselecionados, retrieved 10, June, 2018) by district, which we also collected. Table 1 summarizes the data sources used.

Table 1 – Data sets

Datasources	Description of data sets
São Paulo’s city hall website	Demographic (population, human develop index) and geographical (i.e. shapefiles) information (2017 estimation)
Brazil central bank	Retail activity index at district level (2016)
Brazilian company	Sales records from a major manufacturer of plugs and light switches (2010-2017)
Google Places API	more than 275,000 Points of interest (POIs) in São Paulo (collected in January 2019), including type of POI, user reviews and ratings
Turkish manufacturer	District level sales records from a major Turkish company that produces house paint in Istanbul (2015-2017)
Turkey-Istanbul Economic Productivity Index	Istanbul district level economic indicator - GDP proxy at from 2014 to 2016 based on asset type insurance contract values of a major insurance company.
Istanbul POI data	Includes more than 386,000 POIs data for the first quarter of 2016 in 16 different types of POIs, such as Shops, Hospitals, Financial institutes, Entertainment, Education, etc.
Google Distance Matrix API	For the case of São Paulo we use Haversine distance between pairs of district centroids. However, Haversine distance is not relevant in the case of Istanbul as it is divided into two parts by the straight of bosphorus. To alleviate this problem, distances are replaced with travel times using public transportation, collected from Google Distance Matrix API

In Brazil, we used sales records from a manufacturer of electronic components such as plugs and light switches, and Google Places data. In Turkey, we used sales records from a Turkish company that produces house paint products, and Here.com point of interest location data.

Sales records from the Brazilian company consists of stock-keeping units (SKU) sold from the manufacturer or the distribution centers 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 much smaller in comparison to the Brazilian company sales size and length. The data set consists of records of SKUs sold from 2015 to 2017, in Istanbul. It has information about each SKU sold to which store, for which brand, in which district, in a similar format to the Brazilian company.

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

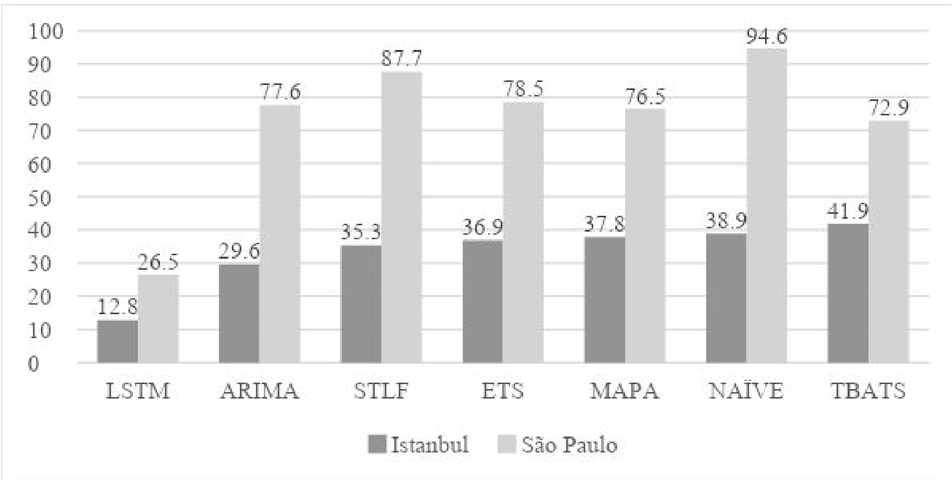
For the São Paulo case, we utilized Google Places API to collect the points of interest information. The data set includes more than 270,000 unique POI's, which were collected in January 2019. Our goal was to extract all existing places in São Paulo that were available, at the time, on Google Places API. The data collection over Google Places API 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 include 89 unique place categories, where we used 13 types which were to be consistent with the case of Istanbul. Examples of these places include shopping malls, tourism and travel agencies, entertainment places, financial institutions, parking lots, restaurants, courthouses, lawyers, local government offices, lodging, electronics store, police, and 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 upon request.

Further details on the data sets and their preparation can be found in supplementary material. The next section will detail the empirical strategy developed to build and test our proposed GFR (gravitational forecast reconciliation) approach.

3.2. Forecasting methods

We tested different methods, temporal aggregations, and forecasting horizons before we proceed with the reconciliation approaches. This was necessary to decide which method to apply in order to create the independent base forecasts.We tested the most used standard time-series methods, such as ARIMA, exponential smoothing, random walk, and trigonometric exponential smoothing (see Armstrong (2001) or Hyndman & Athanasopoulos (2018) for a more detailed view about traditional forecasting approaches). We also tested the more recently developed RNN methods, such as long short-term memory (LSTM). LSTM was the selected method for both data sets after we made the total (most aggregated level) forecast for each data set in different time horizons, as detailed in the next paragraphs. Figure 4 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, for concision we illustrate only the horizon selected.

Figure 4 – Methods performance comparison [1]



Different temporal aggregations were also investigated (Figure 5). To be able to test and compare all temporal aggregations, we used a fixed test set with 12 months (one year), always the same months, aggregated differently. We tested using ARIMA and LSTM, the best performing methods (see Figure 5), for the aggregated level only. Our aim was to apply regularization, i.e. instead of monthly data and model fitting, different temporal aggregations of the data (to smooth out fluctuations) and make predictions accordingly. LSTM results for quarterly (in the brazillian case) and annually (both Brazil and Turkey cases) aggregations were suffering from overfitting and were not further considered. Forecasting accuracies in four-monthly or biannual aggregations had

[1] Results of 3 months forecast horizon.

similar performance in both ARIMA and LSTM, with LSTM achieving more accurate results. Bi-monthly and monthly had different results in the Brazilian (LSTM performs better) and Turkish (ARIMA performs better) cases. However, we chose to proceed with monthly aggregated data to be able to work with maximum data points available that would not make forecasting infeasible. Another reason is that a different temporal aggregation might improve forecast accuracy at the most aggregate level, but that will not impact reconciliation approaches performance. The better the total forecast, the better the accuracy for all levels, however, the ability of the reconciliation approach to divide the forecasts will remain the same.

Figure 5 - Forecasts results with temporal aggregations of the historical data

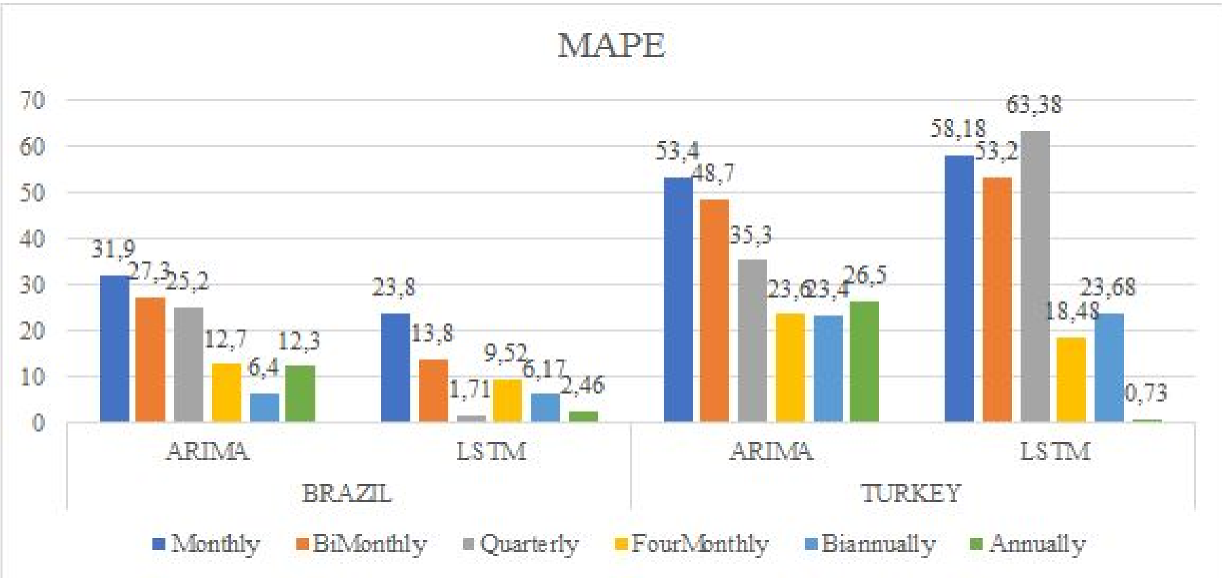
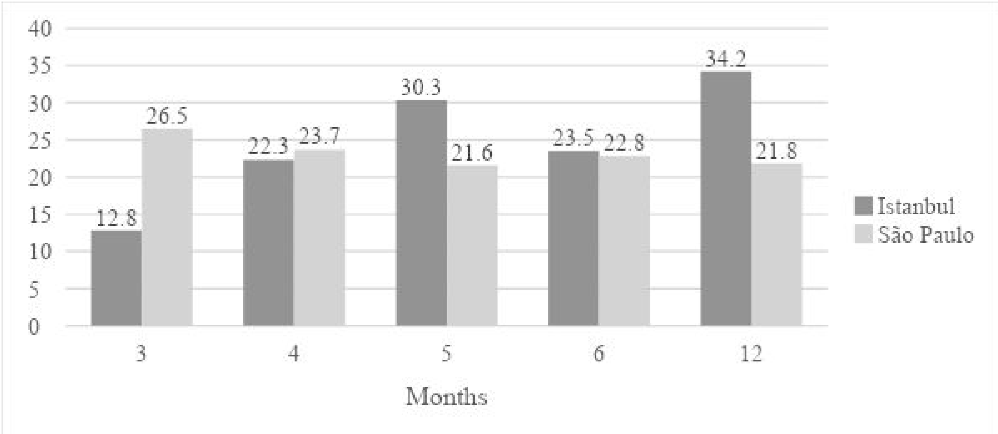


Figure 6 shows the results comparing horizons from 3 to 12 months. We considered that smaller horizons would not be useful for practice, since they do not allow time to plan based on the forecasts. Thus, we decided to run the forecast with a three months ahead horizon. Due to its superior performance in most cases, we selected LSTM as the baseline comparison method. For the Brazilian case, the performance was similar independent of the horizon selected. 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 21.4 p.p.. For that reason, and to be conservative, we based our decision on the best time horizon on the Turkish case.

Figure 6 – Forecasts results with different horizons



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, for concision we present only the selected method (LSTM).

After deciding 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 using the deep learning LSTM technique. These 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 same model was applied to each node of the hierarchical structure. That is, we ran 37 baseline forecasts on the Turkish data set, and 60 on the Brazilian data. The next step after all was to train the models to run the predictions, transform them back (denormalize) and transform to time-series format.

3.3 Reconciliation approaches

After we had all predictions as time-series we computed the accuracy of each forecast without any reconciliation approach, with optimal, top-down, and GFR. For the optimal approach, we estimated independent LSTM models for all the nodes and then adjusted them with the weights calculated by the recursive algorithm (Hyndman et.al, 2011). To estimate these weights, we applied ordinary least square (OLS). We chose to include the optimal reconciliation approach because it considers the hierarchical structure and tends to be more accurate than the top-down approach. If our approach approximates (or is more accurate than the optimal), this means our solution is an alternative approach as strong (or better) as reconcile forecasts than the optimal benchmark approach.

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. The historical proportions were estimated with data from the training set for 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 up to 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 (1) we already described,

$$\tilde{y}_n(h) = SP\hat{y}_n(h) \tag{2}$$

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

The top-down approach uses historical proportions to distribute the forecasts to the lower levels, which can change over time and make estimations highly unreliable. The 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 organizations may not have enough disaggregated data to run forecasts in the district level. They might also not have historical proportions. This is a reality for many companies when they open stores in a certain region or when they are in the early years of operations. Our proposed GFR model overcomes these issues based on agglomeration theory. Before introducing our GFR approach to distribute the forecasts to the different districts, we briefly review the Huff market share model and related literature.

3.4 Gravity Market Share Models

The Huff gravity model predicts merchant market share by calculating the probability of customers' patronization of those merchants based on two factors: merchant attractiveness, and the distance between the customer's location and each merchant. Various attractiveness measures have been introduced in previous studies based on relevant criteria such as facility size, price level, variety of goods and services offered in those facilities (Bozkaya et al 2010, Berman and Krass 2002, Aboolian et al 2007, Hodgson 2007, Suhara et al 2019). These models approximate the market share of each merchant based on the total number of visits or the total money customers spend.

Market share is calculated as follows:

$$U_{ij} = \frac{A_j^{\theta}}{D_{ij}^{\gamma}} \qquad (3)$$

$$f_{ij} = \frac{U_{ij}}{\sum_{j'} U_{ij'}} \qquad (4)$$

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) by population center i at facility j
- D_{ij} : Euclidean distance between population center i and facility j
- θ *and* γ : model fitting parameters

The Huff gravity model and its variation or well known and widely used in marketing and facility location studies (e.g. Bozkaya et al., 2010). Here we introduce a variation of this gravitational approach in order to distribute the total (aggregated) forecast to smaller regions based on the citizens purchase behavior, modeled by our proposed approach.

4. The proposed reconciliation approach - GFR

In our implementation, we make use of the gravity model that is proposed for regional units (in our case, districts), instead of individual merchants. In the Huff model and its variations, "attractiveness" measures are mostly related to physical attributes of a merchant, while some studies have also 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, & Saiz (2001), we hypothesize that the number of amenities, as well as their diversity, can be a proxy of a region's attractiveness. We use a multiplicative model of POI count and their diversity in business type as a measure of attractiveness for each district. This model has been shown to be valid in customer flow modeling (Chong et al., 2018). We utilize the following variation of the Huff gravity model for shopping district choice model:

$$U_{ij} = \frac{A_j^{\theta}}{D_{ij}^{\gamma}} \tag{5}$$

$$A_j = \left(\#POI_j\right)^{\alpha} * \left(Div(POI_j)\right)^{\beta} \tag{6}$$

$$Div(POI)_j \;=\; \sum_k - P_k^j \log log\,\left(P_k^j\right) \tag{7}$$

$$\widehat{f}_{ij} = \frac{U_{ij}}{\sum_{j'} U_{ij'}} \tag{8}$$

Where:

- U_{ij} : Utility of customers living in district *i* from shopping in district *j*
- A_j : Attractiveness of district *j*
- $\#POI_j$: Count of POIs in district *j*
- $Div(POI_j)$: Diversity of POIs in district *j* with respect to their business type using Shannon Entropy (Shannon 1948).
- P_k^j : The percentage of POIs of type *k* in district *j*
- $\widehat{f}_{ij}$: The fraction of total transactions predicted to be made by residents of district *i* in district *j*
- D_{ij} : The travel time by public transportation between districts *i* and *j*
- α, β, γ, and θ: Model fitting parameters

Using this proposed model, we aim to predict each regional unit's (i.e. district) market-share from total retail activity. In most cases, data regarding transactions made by customers in the retail sector is scarce and mostly private. We address this problem by assuming that all the population could be potential customers, and that they may contribute to retail activity. Using the proposed gravity model, for each district, we calculated the proportion of its population that is predicted to visit other districts (i.e. in and out flows of districts) and the proportion of people that are predicted to shop inside their own district. Thus, for each district, we estimate a new "population" contributing to the district's retail activity. This new population or potential customers, consists of its own residents that prefer their own district for shopping plus the people flowing in, from other districts for shopping in that particular district. The volume of potential customers is parametric based on the model fitting parameters. We assume the number of transactions is equal to the potential customers volume after fitting the gravity-based flow model. We consider the geographic center of each district as its population center. Therefore, distances between districts were calculated as geo-distance (Haversine distances) between their respective centers, and the distance of residents from their home district was 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 applied the algorithm described in section 4.1.

4.1 GFR Algorithm

GFR is estimated by the following steps:

Step 1: Attractiveness measures (as in equation 6)

First we calculate, for each district, the number of POIs and their diversity based on the POI types using Shannon 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 two attractiveness measures comparable, and it assumes values larger than or equal to 1, to achieve the increasing effect as the parameters in the exponent are increased.

Step 2: Utility Calculation

$$U_{ij} = \frac{(\#POI_j)^{\alpha} * (D(POI_j))^{\beta}}{D_{ij}^{\gamma}} \tag{9}$$

In this step, we calculate the utility for the potential customers from district *i* from visiting district *j* for shopping. The utilities are parametric based on the still unknown model fitting parameters: α, β, and γ.

Step 3: Probability/Proportion Calculation

We turn utilities calculated in the second step into probabilities/proportions, using equation 8 to normalize and turn them into probabilities or proportion of transactions going from each district to other districts.

Step 4: Customer flow Calculation

$$flow_{ij} = \widehat{f}_{ij} * customers_i \tag{10}$$

Then we multiply calculated proportions for each district with the corresponding district's population (i.e. potential customers originating from district *i*) to find out the outflows from that district.

Step 5: "New" customers (Potential buyers) calculation

$$New\_customers_j = \sum_{all\ districts\ i} flow_{ij} \tag{11}$$

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

Step 6: Best model fitting parameters calculation

$$Predicted\ Retail\ Activity\ Index = \beta_0 + \beta_1\ (new\ customers) \tag{12}$$

We used a 4-dimensional grid search on model fitting parameters α, β, γ, and θ to find the best combination of positive integer values. Using these values, 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 the local retail activity indicator and its actual values was considered as the optimal combination for forecasting and used as a general set of parameters for all districts.

Step 7: Reconciliation of forecasts

Using equation 2, we used the vector of proportions produced 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 the packages 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 upon request. In the following section we test our method in two cities of different countries: São Paulo and Istanbul.

5. Results

In this section, we present the results of applying the aforementioned reconciliation approaches, and compare our method results with a forecast with no reconciliation. Table 2 shows the actual test set data as an index: 1 for the total value in the data set and the proportion of that value that was predicted by each approach. Values closer to 1 are closer to the real value. and the estimates of each reconciliation approach by geographical level. These values are the sums of all nodes in each level and the 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 2 – Sum of forecasts by level over 3 months test set[2]

	SÃO PAULO			ISTANBUL	
	Total	**Sum of zones**	**Sum of districts**	**Total**	**Sum of districts**
Actual data	1,00	1,00	1,00	1,00	1,00
Optimal	1,05	1,05	1,05	0,91	0,91
GFR	**1,09**	**1,09**	**1,09**	**0,91**	**0,91**

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

No reconciliation	**1,09**	**0,91**	**0,64**	**0,91**	**0,83**
Top-down	1,09	1,09	1,09	0,91	0,91

The results considering only the total forecast for three months are presented in Table 3. 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 3 – MAPE of the sum of forecasts for 3 months on the total level

	MAPE	
	SÃO PAULO	ISTANBUL
Optimal	5,43	9,05
GFR	9,19	8,63
No reconciliation	9,19	8,63
Top-down	9,19	8,63

For the lower levels, we will present MASE measures, since we are comparing different time-series and MAPE would not be informative. This measure is an alternative to using percentage errors when comparing forecast accuracy across series with different units. It scales the errors based on the training MAE from a seasonal random walk and gives back a number that is less than 1 if it arises from a better forecast than the average naïve forecast computed on the training data (Hyndman & Athanasopoulos, 2018).

If we consider the average error of each approach on each level, we achieve the results presented in Figures 7 and 8.

Figure 7 – Average error measures by level (Sao Paulo)

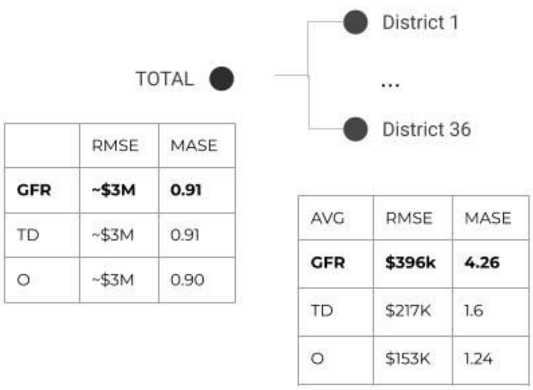
	RMSE	MASE
GFR	**~$12M**	**0.47**
Top down (TD)	~$12M	0.47
Optimal (O)	~$12M	0.49

AVG	RMSE	MASE
GFR	**~$12M**	**1.55**
TD	~$13M	2.05
O	~$5M	0.86

AVG	RMSE	MASE
GFR	**~2M**	**4.02**
TD	~2M	2.09
O	~$1M	2.67

We can see in Figures 7 and 8 that at the most aggregate level, our GFR approach and the Top-down have a similar performance compared to the Optimal approach. With the first level of disaggregation, on average, our approach is not as good as the 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. At the most disaggregated level, on average, our approach has a comparable performance with respect to RMSE. However, MASE shows that, on average, it is performing worse. Since this is an average, some districts with higher error are influencing this results. The same pattern of results were also observed in Istanbul (Figure 8).

Figure 8 – Average error measures by level (Istanbul)



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

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

The same measure, MASE, on the other hand, shows that our approach can outperform the others in some nodes (districts). When looking into each individual districts (figures 9 and 10), 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.

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 and about one third of the districts in Istanbul have a better performance than the other approaches or the difference is smaller than 1.0 (yellow and gray in Figures 9 and 10).

Figure 9 – Difference between GFR and best performing approach (São Paulo)

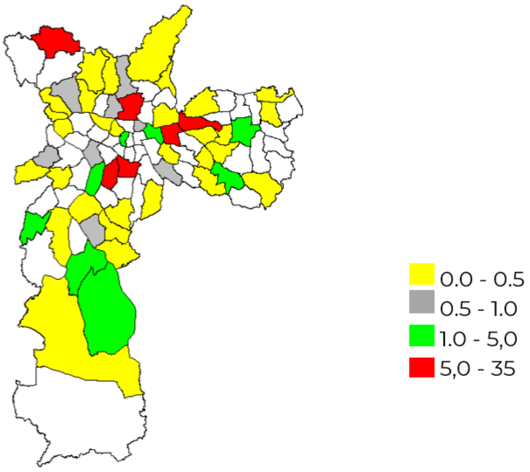
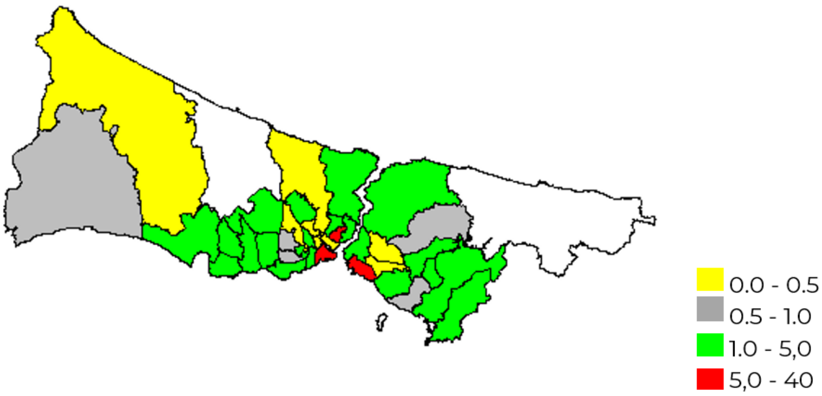


Figure 10 – Difference between GFR and best performing approach (Istanbul)



In supplementary material C, we present further results for disaggregated levels, and an outlier analysis that includes the results of an alternative approach to estimate the proportions aimed to overcome this issue. Since the alternative approach achieved different results for São Paulo and

Istanbul, we maintained the results of our original model, and chose to report these results in supplementary material C.

Summarizing our findings, at the aggregate level all methods have similar performance. At the disaggregated levels, the optimal approach has a better performance in more geographical regions than the other methods, in all metrics considered, mainly due to the fact that it uses the actual disaggregated sales of each region. When compared to the benchmark (top-down), GFR has comparable performance, or outperform it in many regions, even though the method does not have any information on actual sales or historical proportions.

6. Discussion and Conclusion

In our study, we used the agglomeration theory and an adapted gravitational model to contribute to the literature and practice on sales forecasting in several ways. Our first contribution is methodological in that we propose a new forecast reconciliation approach that distributes and reconciles forecasts to lower levels of aggregation, when no actual disaggregated sales data or historical sales proportions are available.

Secondly, we contribute to practice by proposing an approach that (1) can be applied to new territories, points of sales, or channel partners; (2) allows a coherent forecast system to be implemented; (3) is easy and inexpensive to implement in any organization; finally, (4) uses data that is publicly available or natural to the internal processes of most organizations. We contend that these dimensions of practicality make our proposed approach highly acceptable and applicable in real implementations.

Our key contribution to theory is that we show how gravitational models from the marketing (Nakanishi & Cooper, 1974; Drezner, 2006;) and economics (Berman & Krass, 2002; Aboolian et al., 2007) literature can be used in combination with the state-of-the-art forecasting methods (such as LSTM) to not only explain sales but also to predict it with an evidence of high accuracy. The ability of the gravitational model based on attractiveness measures to distribute sales amongst geographical regions highlights the importance of considering spatial effects and factors on studies involving sales data, either describing it or forecasting it.

Our GFR approach allows companies to distribute forecasts to lower levels of aggregation with zero or low costs using open source tools and publicly available data sets. We provide evidence of the accuracy of such an approach in two empirical test cases using real sales data from two companies in Brazil and Turkey. Our results show that the proposed approach is equal or better than the benchmark top-down approaches that use disaggregated sales data. Our deeper analysis, as reported in Supplementary Material C, suggest that regional behavioral patterns can be addressed by fitting model parameters at the district level using advanced optimization techniques, which is shown to reduce total forecast errors made, at least in the case of Brazil. In summary, our proposed modeling approach can be applied by companies of any sector, especially those that have a limited amount of resources to invest in the process of gathering, storing and analyzing data.

Our method has some limitations. Since we don't have access to individual buyers' information, we assume that all residents of a district are potential customers. For each district, we calculate the proportion of its population that are predicted to visit other districts (outflows), and the proportion that is predicted to shop within their own district. So, in essence, we calculate a new "population" for each district that represents the potential buyers in that district. This "population" is dependent on the attractiveness criteria and their respective parameters, and the parameters are derived from correlation-maximizing values in relation to economic indicators such as retail indexes, GDP, etc. Clearly, there might exist better proxies for district attractiveness and/or approaches to determine the best parameters for the model.

Another limitation is that we are unable to define population centers for districts since we do not have detailed information on the potential customers' residences. This, however, is by design given the fact that one should not have access to high resolution data at the individual or disaggregate level to make reasonably accurate forecasts. To alleviate this limitation, we consider each district's geographic center of gravity as the home locations of residents in that district. Thus, distances between districts are calculated as geo-distances between their respective centers, and the distance of residents to and from their home district is calculated as the radius of the largest possible circle inside that district divided by a parameter t.

Finally, we are aware that possibly there are socio-demographic differences between districts, which may result in different behavioral patterns. We attempted to address these differences by fine tuning model parameters at the district level, yet we found mixed results as we reported in Supplementary Material C. This is perhaps an area of further investigation.

We also note that some public data sets may be discontinued by new governmental policies. This might impose difficulties for applying GFR in the future. Future research might test the ability of other data sources in providing 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 may be more expensive to acquire than Google Places data and they may require professionals and researchers with different skills to analyze them.

Our aim was to suggest a general approach that can be used by any organization, applicable uniformly over a given geographical region at varying levels of aggregation. Yet, we observe varying levels of forecast errors, for instance over the set of geographical regions (ie. districts) considered. We have included an "outlier analysis" in Supplementary Material C to address this issue. We checked for outliers in two variables: sales history and districts. We cannot simply 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 with real data, outliers may be a reality and need to be a part of the analyses. For these reasons, we kept all the results for all the districts with available data. Future research can explore further what one could call an "outlier" and the reasons some districts behave differently from the rest.

Another strategy that could be implemented to improve accuracy is to test the different temporal aggregations to smooth out fluctuations that we showed in Table 3. This means that instead of using monthly data and model fitting, one could consider aggregated sales bi-monthly, quarterly, biannually, or annually. To be able to test it, one needs test data sets with longer time horizons (e.g. longer than 12 months), so that a sufficient number of data points can be made subject to various possible temporal aggregations while avoiding overfitting.

Even without applying these strategies to improve accuracy, our approach, GFR, 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. From the practice point of view, our approach gives a feasible alternative for disaggregating sales forecasts to geographical areas without information about historical proportions when only a total forecast (for the most aggregate level) is available.

References

Aboolian, R., Berman, O., & Krass, D. (2007) Competitive facility location and design problem. *European Journal of Operational Research*, 182(1): 40–62.

Albuquerque, P., & Bronnenberg, B. J. (2012). Measuring the impact of negative demand shocks on car dealer networks. *Marketing Science*, 31(1), 4–23.

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners* (Vol. 30). Springer Science & Business Media.

Bergmeir, C. N., & Benítez Sánchez, J. M. (2012). *Neural networks in R using the Stuttgart neural network simulator: RSNNS*. American Statistical Association.

Berman, O., & Krass, D. (2002) Locating multiple competitive facilities: Spatial interaction models with variable expenditures. *Annals of Operations Research*, 111(1-4), 197-225.

Bollinger, B., & Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, *31*(6), 900–912.

Bozkaya, B., Yanik, S., & Balcisoy, S. (2010) A GIS-based optimization framework for competitive multi-facility location-routing problem. *Networks and Spatial Economics 10*(3):297–320.

Bradlow, E. T., Bronnenberg, B., Russell, G. J., Arora, N., Bell, D. R., Duvvuri, S. D., … Yang, S. (2005). Spatial models in marketing. *Marketing Letters*, *16*(3–4), 267–278.

Bronnenberg, B. J., & Mahajan, V. (2001). Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science*, *20*(3), 284–299.

Bronnenberg, B. J., & Sismeiro, C. (2002). Using multimarket data to predict brand performance in markets for which no or poor data exist. *Journal of Marketing Research*, *39*(1), 1–17.

Bucklin, R. E., Siddarth, S., & Silva-Risso, J. M. (2008). Distribution intensity and new car choice. *Journal of Marketing Research*, *45*(4), 473–486.

Canitz, H. (2016). Overcoming Barriers to Improving Forecast Capabilities. *Foresight: The International Journal of Applied Forecasting*, (41), 26–34.

Chan, T. Y., Padmanabhan, V., & Seetharaman, P. (2007). An econometric model of location and pricing in the gasoline market. *Journal of Marketing Research*, *44*(4), 622–635.

Chong, S.K., Bahrami, M., Chen, H., Bozkaya, B., Balcisoy, S., & Pentland, A. (2019) Good cities are goods. Working paper.

Cooley, D. (2018). *googleway: Accesses Google Maps APIs to Retrieve Data and Plot Maps*. Retrieved from https://CRAN.R-project.org/package=googleway

Dekimpe, M. G., & Hanssens, D. M. (2000). Time-series models in marketing:: Past, present and future. *International Journal of Research in Marketing*, *17*(2–3), 183–193.

Dekker, M., van Donselaar, K., & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, *90*(2), 151–167.

Drezner, T. (2006) Derived attractiveness of shopping malls. *IMA Journal of Management Mathematics*, 17(4): 349-358.

Fildes, R. A., Ma, S., & Kolassa, S. (2018). *Retail forecasting: research and practice*.

Fliedner, E. B., & Lawrence, B. (1995). Forecasting system parent group formation: An empirical application of cluster analysis. *Journal of Operations Management*, *12*(2), 119–130.

Fliedner, E. B., & Mabert, V. A. (1992). Constrained forecasting: some implementation guidelines. *Decision Sciences*, *23*(5), 1143–1161.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in Neural Information Processing Systems*, 1019–1027.

Garber, T., Goldenberg, J., Libai, B., & Muller, E. (2004). From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, *23*(3), 419–428.

Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. *Journal of Economic Geography*, *1*(1), 27–50.

Gonzalez-Benito, O., Munoz-Gallego, P. A., & Kopalle, P. K. (2005). Asymmetric competition in retail store formats: Evaluating inter-and intra-format spatial effects. *Journal of Retailing*, *81*(1), 59–73.

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25.

Hanssens, D. M. (1998). Order forecasts, retail sales, and the marketing mix for consumer durables. *Journal of Forecasting*, *17*(3–4), 327–346.

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *The American Economic Review*, *102*(2), 994–1028.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hodgson, MJ (2007) A location—allocation model maximizing consumers' welfare. *Regional Studies 15*(6):493–506

Huff, D. L. (1964). Defining and estimating a trading area. *Journal of Marketing*, *28*(3), 34–38.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Retrieved from https://OTexts.org/fpp2/

Hyndman, R. J., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., … Yasmeen, F. (2018). *forecast: Forecasting functions for time series and linear models*. Retrieved from http://pkg.robjhyndman.com/forecast

Hyndman, R. J., Lee, A., & Wang, E. (2017). *hts: Hierarchical and Grouped Time Series*. Retrieved from https://CRAN.R-project.org/package=hts

JJ Allaire, & Chollet, F. (2019). keras: R Interface to “Keras” (Version R package version 2.2.4.1) [R]. Retrieved from https://CRAN.R-project.org/package=keras

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9.

LeSage, J. P. (1999). The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio*, *28*(11).

Liu, A. X., Steenkamp, J.-B. E., & Zhang, J. (2018). Agglomeration as a Driver of the Volume of Electronic Word of Mouth in the Restaurant Industry. *Journal of Marketing Research*.

Meyer, R. J., & Hutchinson, J. W. (2016). (When) Are We Dynamically Optimal? A Psychological Field Guide for Marketing Modelers. *Journal of Marketing*, *80*(5), 20–33.

Nakanishi, M., & Cooper, L.G. (1974) Parameter estimation for a multiplicative competitive interaction model: Least squares approach. *Journal of Marketing Research*, 11(3), 303.

Narayanan, S., & Nair, H. S. (2013). Estimating causal installed-base effects: A bias-correction approach. *Journal of Marketing Research*, *50*(1), 70–94.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Reiss, P. C. (2011). Structural Workshop Paper-Descriptive, Structural, and Experimental Empirical Methods in Marketing Research. *Marketing Science*, *30*(6), 950–964.

Shannon, CE. (1948) A mathematical theory of communication. *The Bell System Technical 443 Journal 27*(July 1928):379–423.

Shriver, S. K. (2015). Network effects in alternative fuel adoption: Empirical analysis of the market for ethanol. *Marketing Science*, *34*(1), 78–97.

Silveira Netto, C. F., Hyndman, R. J., & Brei, V. A. (2019). *What is the Optimal Strategy of Aggregation for Forecasting Sales? A Hierarchical Forecast Reconciliation by Region, Product Category, and Channel*. Presented at the Marketing Science Conference.

Sridhar, K., Bezawada, R., & Trivedi, M. (2012). Investigating the drivers of consumer cross-category learning for new products using multiple data sets. *Marketing Science*, *31*(4), 668–688.

Stremersch, S., Landsman, V., & Venkataraman, S. (2013). The relationship between DTCA, drug requests, and prescriptions: Uncovering variation in specialty and space. *Marketing Science*, *32*(1), 89–110.

Suhara, Y., Bahrami, M., Bozkaya, B., & Pentland, A. (2019). Validating Gravity-Based Market Share Models Using Large-Scale Transactional Data. *arXiv preprint arXiv:1902.03488*.

Teller, C., & Reutterer, T. (2008). The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services*, *15*(3), 127–143.

Wang, E., Cook, D., & Hyndman, R. J. (2018). *tsibble: Tidy Temporal Data Frames and Tools*. Retrieved from https://CRAN.R-project.org/package=tsibble

Wickham, H. (2017). *tidyverse: Easily Install and Load the "Tidyverse."* Retrieved from https://CRAN.R-project.org/package=tidyverse

Williams, B. D., & Waller, M. A. (2011). Top-down versus bottom-up demand forecasts: the value of shared point-of-sale data in the retail supply chain. *Journal of Business Logistics*, *32*(1), 17–26.

Yang, S., & Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research*, *40*(3), 282–294.

Zhang, W., & Kalra, A. (2014). A Joint Examination of Quality Choice and Satisfaction: The Impact of Circumstantial Variables. *Journal of Marketing Research*, *51*(4), 448–462.

SUPPLEMENTARY MATERIAL

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 São 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 São 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 São 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 data sets 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 São 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 data sets 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 São 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 São Paulo. If only the state is clear, a general coordinate pointing to the state of São 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 4.

Table 4 – 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

B – 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, since 488 of the 909 clients that made a purchase on 2015 were not on the 2016 and 2017 data sets. It was not possible to input the district on this clients purchases;
3. After deleting the rows of 2 districts with no data for gravitational model, the total sales of 2017 increases because it contained negative values.

C – OUTLIER ANALYSIS

In the brazilian case, our approach has the best performance in 8 nodes and in 34 it is equal or close to the best performing. Only 5 of them have higher differences (10 or more). Those were considered outliers, influencing the averages of our approach to be higher than the others. For that reason we developed a different model estimation to correct those outliers.

As explained in the methodology section, we used a set of parameters for each city found by a grid search over all parameters. This means we consider the same parameters for all districts, meaning that we assume all districts' population behave similarly. However, this may not be true and the parameters might differ from region to region based on the residents' behavior. As a result, using the same set of parameters for all districts could be a reason for some districts to become outliers. To alleviate this problem, we performed another step of optimization on the parameters but this time for each individual district. We used a Stochastic Gradient Descent method to minimize the squared percentage error by adjusting the parameters for each district individually. We took the general parameters as starting solution and at each iteration the parameters were adjusted for the randomly chosen district.

The results of our approach reported at the results and of the different estimation are on Table 5.

Table 5 – MASE by node in São Paulo

	Optimal	Top Down	GFR	Difference	Outlier corrected	Difference
Total	0,49	0,47	0,47	0	0,45	BEST
Centre	0,15	0,83	1,58	1,43	0,66	0,51
East	1,38	2,57	1,72	0,34	1,44	0,06
North	0,45	0,79	0,55	0,1	0,63	0,18
West	0,36	1,38	0,45	0,09	0,94	0,58
South	1,99	4,68	3,44	1,45	3,43	1,44
República	0,2	1,03	1,7	1,5	0,68	0,48
Santa Cecília	0,55	1,02	0,56	0,01	0,62	0,07

Artur Alvim	1,8	2,01	1,82	0,02	1,69	BEST
Água Rasa	0,73	0,65	0,85	0,2	0,64	BEST
Aricanduva	3,88	3,85	3,71	BEST	3,79	BEST
Belém	6,47	3,77	5,36	1,59	26,18	22,41
Cangaíba	1,29	0,94	0,73	BEST	0,7	BEST
Cidade Líder	1,17	1,28	0,66	BEST	1,03	BEST
Iguatemi	1,07	0,83	0,83	0	0,78	BEST
Itaquera	1,02	0,2	1,31	1,11	2,74	2,54
Jardim Helena	2,99	2,55	0,37	BEST	0,5	BEST
Pari	2,81	3,25	3,55	0,74	3,55	0,74
Penha	7,75	1,61	15,97	14,36	9,97	8,36
Sapopemba	4,64		4,42	0,22	0,01	BEST
São Mateus	1,08	3,45	2,99	1,91	1,67	0,59
São Rafael						
Tatuapé	6,01	4,15	35,4	31,25	43,97	39,82
Vila Curuçá	0,74	0,44	0,37	0,07	0,61	0,17
Vila Matilde	1,53	2,43	1,34	0,19	1,65	0,12
Vila Prudente	2,16	1,21	2,03	0,82	3,91	2,7
Brasilândia	0,71	1,22	1,08	0,37	1,03	0,32
Cachoeirinha	4,63	2,93	1,44	BEST	1,76	BEST
Casa Verde	0,7	0,35	1,19	0,84	10,9	10,55
Freguesia do Ó	0,25	0,52	0,23	0,02	0,27	0,02
Jaçanã	3,58	3,61	3,7	0,12	3,56	BEST
Jaraguá	1,04	0,77	0,75	BEST	0,7	BEST
Mandaqui	2,01	3,25	2,86	0,85	4,6	2,59
Perus	4,17	9,73	9,66	5,49	8,96	4,79
Pirituba	0,38	0,78	0,93	0,55	0,71	0,33
Santana	2,03	0,14	26,05	25,91	5,48	5,34

São Domingos	0,9	0,76	0,12	0,64	1,04	0,28
Tremembé	1,32	0,05	0,16	0,11	1,5	1,45
Vila Maria	1,94	1,74	1,87	0,13	2,03	0,29
Barra Funda	0,3	1,13	0,24	0,06	0,86	0,56
Itaim Bibi	0,87	1,91	3,26	2,39	0,92	0,05
Pinheiros	0,88	3,05	1,75	0,87	1,01	0,13
Rio Pequeno	3,17	1,41	2,01	0,6	2,82	1,41
Raposo Tavares	21,11	3,28	1,54	1,74	4,4	1,12
Vila Leopoldina	0,1	0,1	0,1	0	0,47	0,37
Vila Sônia	0,45	0,09	0,29	0,2	0,78	0,69
Cidade Ademar	3,43	2,72	1,29	BEST	0,48	BEST
Cidade Dutra	1,67	5,49	3,13	1,46	3,4	1,73
Campo Grande	0,87	1,13	1,53	0,66	0,02	BEST
Campo Limpo	1,9	2,32			1,29	BEST
Capão Redondo	3,37	6,15	5,55	2,18	5,65	2,28
Grajaú	6,69	3,6	4,86	1,26	5,82	2,22
Jabaquara	1,89	0,08	0,39	0,31	0,04	BEST
Jardim São Luís	1,11	1,31	0,66	BEST	1,1	BEST
Moema	2,5	0,28	10,47	10,19	2,78	2,5
Parelheiros	1,49	0,15	0,49	0,34	0,72	0,57
Pedreira	2,14	0,24	0,25	0,01	0,25	0,01
Sacomã	1,91	2,66	1,78	0,13	4,89	2,98
Santo Amaro	4,63	5,67	5	0,37	6,53	1,9
Vila Mariana	9,95	5,61	39,65	34,04	29,93	24,32

With the new estimation of proportions, we achieve a better performance on 17 nodes and 23 are equal or close to the best performing, however, we still have 5 outliers. The overall performance was improved with the new estimation, however only in the brazilian case. For the turkish forecasts, the new estimation made the overall performance worse. (Table 6).

Table 6 – Overall performance of the approaches (MAPE)[3]

	São Paulo	Istanbul
Optimal	5.43	9.05
Top Down	9.19	8.63
GFR	9.19	8.63
Outlier corrected	3.60	10.20

On the turkish case, we have 2 nodes on which our approach is the best performing; 11 districts performing equally or very similar to the best approach; and only 2 that we considered outliers districts (10 or more). After the new estimation of proportions, we still have 2 nodes performing better, and 17 that are perform equally or similar to the best approach. However, we have one more outlier than before (3).

Table 7 – MASE by node in Istanbul

	Optimal	Top Down	GFR	Difference	Outlier corrected	Difference
Total	0,9	0,91	0,91	0,01	0,92	0,02
Atasehir	0,52	1,56	0,81	0,29	0,96	0,44
Avcilar	0,5	1,74	3,57	3,07	2,09	1,59
Bagcilar	2,19	3,35	3,12	0,93	2,65	0,46
Bahcelievler	1,05	1,5	1,62	0,57	1,71	0,66
Bakirkoy	3,81	3,45	7,53	4,08	13,81	10,36
Basaksehir	1,18	1,16	3,63	2,47	1,18	0,02
Bayrampasa	0,49	0,66	1,69	1,2	1,63	1,14

[3] Error metric of the sum of 3 months forecasts on total level.

Besiktas	1,79	0,83	3,88	3,05	2,41	1,58
Beykoz	0,72	2,61	2,34	1,62	1,46	0,74
Beylikduzu	2,29	2,01	4,37	2,36	2,57	0,56
Beyoglu	0,98	1,23	1,28	0,3	2,55	1,57
Buyukcekmece	1,11	0,39	1,51	1,12	1,5	1,11
Catalca	0,94	0,85	0,11	BEST	0,11	BEST
Cekmekoy	1,58	1,08	1,87	0,29	1,87	0,29
Esenler	0,12	0,11	0,59	0,48	0,18	0,07
Esenyurt	1,18	1,83	2,86	1,68	0,95	BEST
Eyup	0,52	0,89	0,4	BEST	0,67	0,15
Fatih	0,78	0,71	39,01	38,3	23,77	23,06
Gaziosmanpasa	0,61	0,69	0,67	0,06	3,12	2,51
Gungoren	0,24	0,88	1,67	1,43	0,84	0,6
Kadikoy	0,85	1,32	6,69	5,84	5,91	5,06
Kagithane	1,25	1,37	4,11	2,86	1,85	0,6
Kartal	1,27	1,37	2,22	0,95	3,04	1,77
Kucukcekmece	1,2	2,65	5,58	4,38	1,66	0,46
Maltepe	1,2	1,58	2,64	1,44	2,11	0,91
Pendik	0,23	2,17	1,37	1,14	3,09	2,86
Sancaktepe	0,54	1,19	4,05	3,51	3,73	3,19
Sariyer	1,33	2,13	2,56	1,23	1,87	0,54
Silivri	0,66	0,85	1,38	0,72	2,12	1,46
Sisli	0,95	2,09	16,8	15,85	24,83	23,88
Sultanbeyli	3,67	2,75	7,05	4,3	6,82	4,07
Sultangazi	3,07	4,46	4,9	1,83	4,29	1,22
Tuzla	0,73	0,9	2,26	1,53	1,81	1,08
Umraniye	1,89	1,08	1,53	0,45	1,35	0,27
Uskudar	0,55	1,57	2,43	1,88	1,28	0,73

Zeytinburnu	2,51	2,8	5,22	2,71	3,71	1,2

We believe there are some reasons that make the new approach work better with the Brazilian forecasts. One reason is that sales of the Brazilian company have a better correlation to the indicator used on model estimation (retail index). For the Turkish model GDP was used and we found a small correlation between this indicator and sales. That leads us to conclude that the choice of the indicator to be used on our approach will impact its accuracy. The Turkish time series is also smaller and more volatile, which are known to lead to lower levels of accuracy. The last but not the least could be the difference in potential customers' behavior in those two cities. To elucidate, the potential customers of different districts in Istanbul behave more homogeneously, while the behavior in the case of Sao Paulo is more heterogeneous among the districts. This leads to better results for Istanbul when using a general set of parameters, but the best results are achieved for the case of Sao Paulo when each district uses a different set of parameters optimized individually.