Geomarketing

Raphaël Pinault, Camille Pousse, Cindy Puech

Contents

1 Introduction and definition of the study area

Our market analysis is centered around the Provence-Alpes-Côte d'Azur region, also known as PACA (Code region 93 in the data base IRIS), which exhibits a continuous coverage without any holes, as depicted in the map below. The PACA region encompasses six shops identified by the following numbers 17, 18, 25, 34, 47, and 53, and has a resident population of 2,309,979 inhabitants. Although some stores are located at the extremities of the region, we have carefully checked that their locations do not extend into adjacent areas not included in our analysis. It's worth noting that the PACA region is located in a coastal zone, so the extremities of the region are bordered solely by the ocean, thus excluding any other population not taken into account in our study.

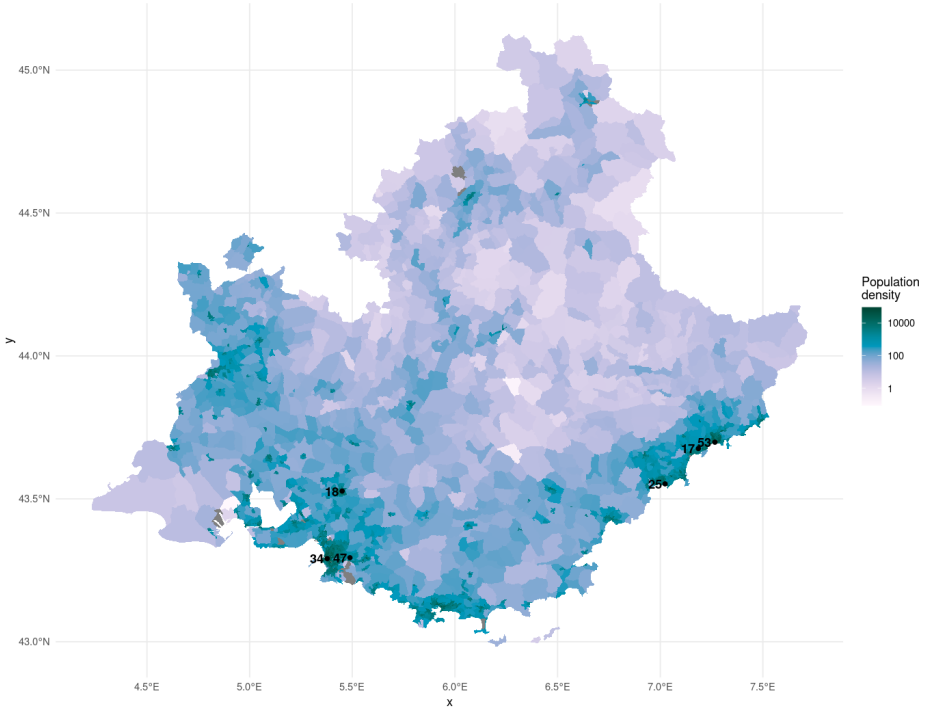

Figure 1: Population density in the PACA region

The aim of this repport is to delineate principal trading areas of existing stores and advise on optimal locations for new establishments in PACA region. To achieve these purposes, we used the following methodology. Once defined, we preprocess marketing data, aggregating at IRIS and shop levels and incorporating additional measures like overall sales, competitor presence, and area attractiveness proxies. Statistical models are then estimated to explain sales volumes and market shares, drawing on explanatory variables from open-source data on the INSEE website [1]. Subsequent phase involve identifying primary trading zones around client stores. Finally, we evaluated potential new shop locations based on competitors and sales model.

2 Preparation of the data and Descriptives statistics

After having defined our study area in the PACA region, our focus in this section will be on data preparation for our geomarketing analysis in this region. This is subdivided into three sections, aimed at improving data at shops, IRIS and IRIS/store pair levels.

2.1 Shop level

In our shop analysis, we incorporate the following details for each shop: the sum of sales, average basket size and number of visits from all customers for each store in PACA region. These measures give us an overview of the commercial activity in each store, and are essential for assessing their performance.

Subsequently, we introduce an additional metric per shop that calculates the number of competitors in the vicinity. To achieve this, we opt to count the number of competitors within a 25 km radius. We justify this choice on the belief that customers are unlikely to travel more than 25km to visit a competing establishment. In addition, we justify our choice to assess competition within a 25 km radius on the basis of Michaud-Trévinal's research [2], showing that the vast majority of daily trips are within this radius. This study indicates that the average distance traveled to visit an establishment is 23 km, with fewer than 4% of trips exceeding 80 km. Thus, we hypothesize that customers restrict their travels to visit competing establishments and a 25 km radius seems appropriate for our study in the PACA region. We observe that implementing a 25 km buffer zone around each of our stores, which can be seen as the trade area, overlaps with other stores. Consequently, we identify cannibalization zones, which occur at the intersection of two trade areas.

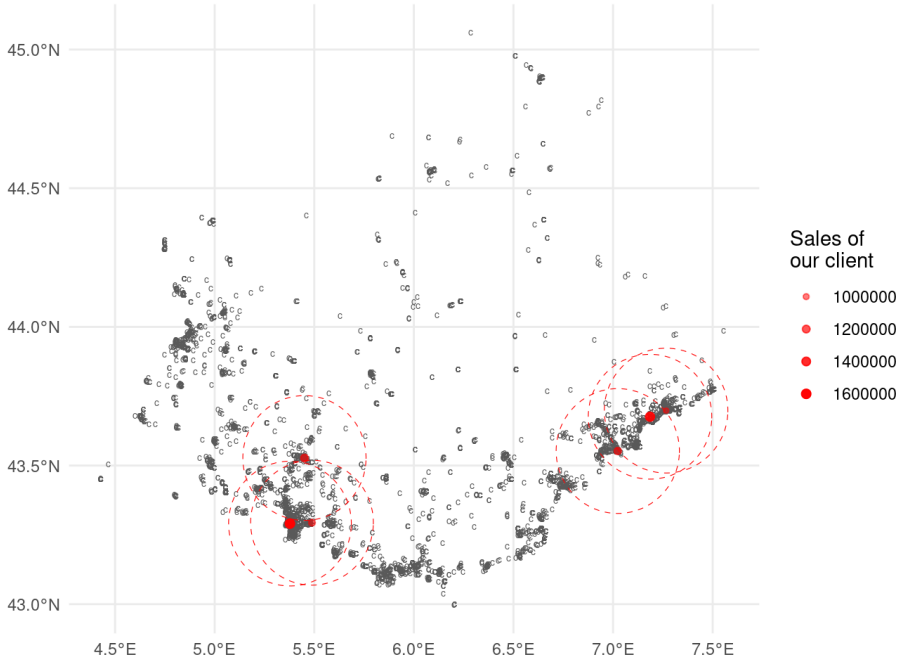

Figure 2: 25km radius around each store

Regarding the attractiveness of the 25km buffer zone, we have dervied two metrics : the first one is the total population within the area because we assume that a large number of people in close proximity to the shop can create a potentially significant market, promoting business success. The second one is the number of individuals in the buffer zone with a high socio-professional status, defined as those holding a bachelor's degree (or equivalent) or higher because we assume that attractive for businesses, an environment with a concentration of individuals with high socio-professional status can contribute to a higher number of sales and larger financial transactions.

Below is a summary table of all the metrics we have for each shop.

Shop number	Sales	Visits	Basket value	Competitors	Population	CSP
17	1 506 316	1 076	1 400	2 086	1 007 227	81 576
18	1 174 042	1 546	759	1 114	1 044 871	84 324
25	1 114 020	1 311	850	1 491	750 868	61 842
34	1 613 373	1 438	1 122	2 107	1 358 387	110 783
47	1 050 105	1 385	758	2 092	1 356 638	113 543
53	987 683	836	1 181	1 549	886 467	72 999

Table 1: Metrics by shop

This initial analysis provides us with a preliminary overview of these shops. Upon examining sales figures, it is evident that store number 34 stands out significantly with total revenue reaching 1 613 373. This particular store appears to attract a high number of visitors, suggesting either strong brand recognition or a strategically advantageous location. Additionally, it presents a high average basket value of 1 122 units, indicating that customers tend to spend more during their visits to this store.

Regarding competition, store number 17 faces a considerable number of competitors, with 2 086 stores in its catchment area.

In terms of the population served, store number 34 and 47 appears to have a wide reach, potentially serving a population of approximately 1 350 000 inhabitants. This implies significant potential customer base for the store, likely contributing to its high sales performance.

2.2 IRIS level

In this section, we aggregate the data at IRIS level to get an overview of sales performance and competition in each geographical area. As in the shop level analysis, we include the total sales, the number of visits, and the count of competitors by IRIS. Below is a summary table of the metrics we have aggregated over all IRIS.

Statistic	Minimum	Q1	Median	Mean	Q3	Maximum	Variance	Standard deviation
Visits	0.0	0.0	2.0	3.1	5.0	28.0	15.9	4.0
Sales	0	0	475	3 059	3 166	188 753	81 155 499	9 008.6
Competitors	0.0	0.0	0.0	3.5	2.0	261.0	182.1	13.5

Table 2: Metrics by IRIS

We notice that on average, the number of visits amounts is 3.1, with a considerable dispersion ranging from 0 to 28. The average sales reach 3 059, though with significant dispersion reflected by a high standard deviation of 81 155. The average number of competitors is 3.5 with a notable dispersion measure by standard deviation of 13.5.

From there and through maps, we can explore where the majority of customers come from and determine if there is competition in certain areas. To begin, we are interested in the distribution of sales by IRIS. It is clear that sales are significantly higher in areas close to the stores. This observation suggests a direct correlation between store proximity and sales volume. More precisely, as one moves further away from the stores, the value of sales tends to decrease noticeably. This gradual decline underlines the importance of store location.

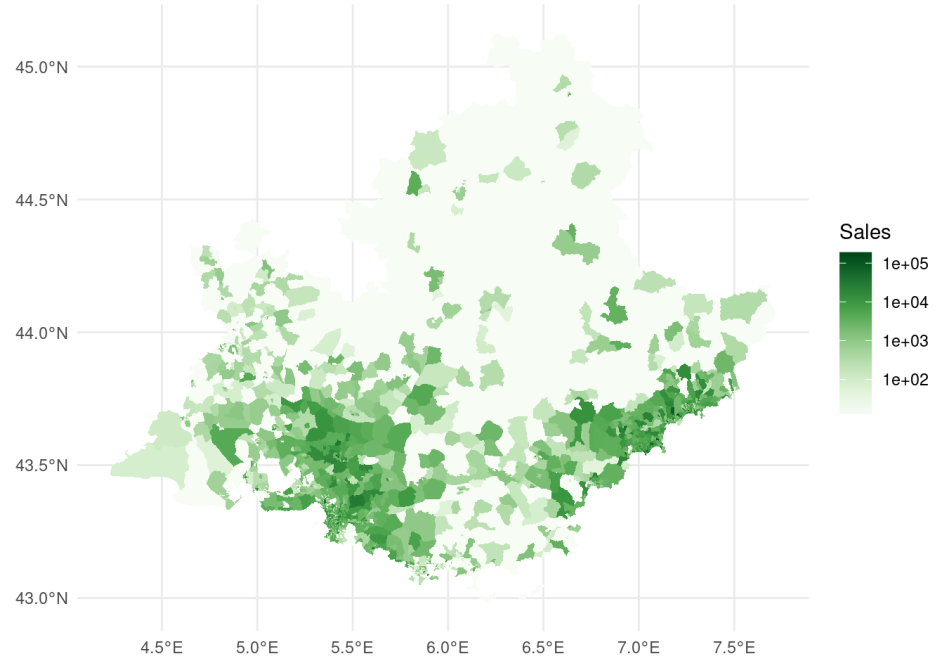

Figure 3: Repartition of sales by IRIS

Then, we are interesting in the number of competitors by IRIS which reveals a certain heterogeneity in the number of competitors in the PACA region. A slightly more pronounced competitiveness appears to be evident at the geographical boundaries of our study area as well as in certain inland iris zones, distributed rather randomly. This observation may indicate a specific economic and social dynamic in these areas.

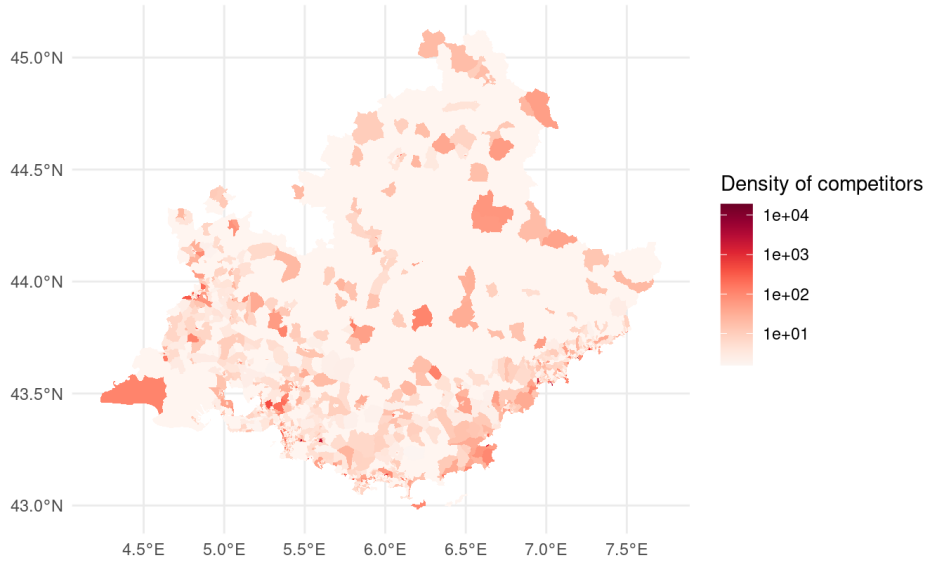

Figure 4: Density of competitors by IRIS

Finally, the last graph at IRIS level below shows the number of visits per IRIS. We can see a similar trend to that observed in the first graph: areas close to stores record a higher number of visits, while this number gradually decreases as you move further away from the stores. This correlation between store proximity and number of visits underlines the importance of spatial accessibility in attracting customers. Store location can therefore have a significant impact.

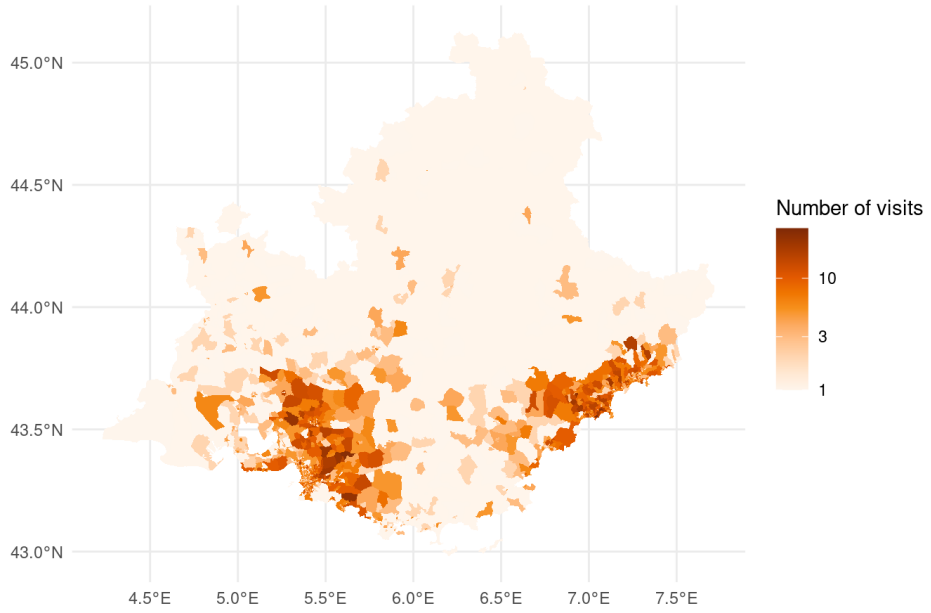

Figure 5: Repartition of the number of visits by IRIS

In this section, we aggregated the data at IRIS level to analyze sales performance and competitiveness in each geographical area. In conclusion, the graphs obtained suggest a correlation between store proximity and sales, as well as a general instability of competitiveness in the region, with a slight accentuation near geographical borders. These results suggest that the store's location influences its success, highlighting the relevance of use of spatial statistic tools.

2.3 Pair level data

In this part, we prepare data describing pairs of IRIS and stores. This involves aggregating customer data with respect to both indexes (IRIS and store). Given that we have 6 shops and 2 434 different IRIS in the PACA region, we obtain a dataset of 14 604 (=6 $\times$ 2 434) rows corresponding to each combination of IRIS x shop.

Then, we added two types of distances : the great circle distance and OSRM server-based travel time in minutes to the data. The great circle distance refers to the shortest distance between two points on the surface of a sphere, following the arc of a great circle. This provides us with measures of both direct distance and estimated travel time between each pair of store and IRIS, where the IRIS point is its centroids.

Regarding the calculation of travel distances in minutes, we encountered difficulties using OSRM on the server, so we decided to save the data locally, then export it to run OSRM on our own computers. Once the process was complete, we re-imported the results. The script used locally to answer this question is located in the following file: "/data/cg2024/group3/calculating_distance/calcul_distance.R". In this same file, you will find the databases used, named "dest.RData" and "origin.RData", as well as the results incorporated into the base "data.RData". This script was used to calculate distances and travel times using the data provided and running OSRM locally.

Figure 6: Scatterplot of drive times vs circle distance

The scatter plot above illustrates the relationship between direct distances and travel minutes.

Direct distances are computed using the great circle distance formula, while travel minutes represent the estimated time required to traverse these distances via the actual road network. Upon examining the plot, we observe a seemingly linear relationship between the two variables. Consequently, we proceeded to calculate the correlation coefficient between direct distances and travel minutes. The correlation coefficient obtained was approximately 0.87, indicating a strong positive correlation between the variables. This result aligns with our expectations, as an increase in direct distance typically corresponds to an increase in travel time.

3 Models of the sales volumes and market shares

To build our models, we chose to use variables at the IRIS level available on open data on the INSEE website [1].

First, we add the variable **P18_POP15P**, which represents the number of people aged 15 or over. This variable was selected as it is indicative of the customer potential in the region studied. Indeed, it provides an estimate of the size of the local market, which is crucial for assessing sales opportunities. In addition, from the age of 15 onwards, individuals tend to become more mobile and often have disposable income for consumption, which reinforces the importance of this variable in our models.

We also included the variable **P18_NSCOL15P_SUP5** which gives us information on the number of people aged 15 or over who hold a higher education diploma of Master's degree level or above. These individuals tend to have higher purchasing power and are likely to spend more on quality products.

We added the variable **P18_RP_PROP**, which indicates the number of owner-occupied primary residences. This provides information on the stability of the local population and is often correlated with loyal and committed customers. These owners, having invested in their homes, are likely to contribute significantly to local sales due to their attachment to the region and their commitment to local businesses. Furthermore, owning a house reflects a good standard of living and thus may be associated with higher expenses

In addition, household composition, represented by the variable **C18_MENCOUPSENF**, was considered an important factor. Households without children may have different purchasing behaviors, with potentially more resources available to spend on non-essential products. This variable therefore enriches our model by providing insights into the consumption preferences of different household types.

We also include the variable **P18_SCOL1824**, which indicate the number of schooled people between 18 and 24 years old. The presence of the student population in our model can help us understand whether they are targeted by the brand under study or not. On the contrary, we might expect this population to consume less in the stores due to a lower purchasing power.

We also use the variable **P18_RP_VOIT2P**, which provides information about the number of households with at least two cars. This variable allows us to consider the mobility of populations. Indeed, a family with at least 2 cars will have greater flexibility in traveling to do their shopping.

Finally, we include the variable **minutes** that we computed using OSMR, which provides us the travel time, in minutes, between an IRIS, identified by its centroid, and a shop. We opt for travel time instead of distance because it considers various factors such as traffic conditions and road speed limits, providing a more accurate representation of the actual travel duration between locations. We decide to reduce our data by keeping only the observations where the travel duration in minutes is less than the 80th percentile. This step removes approximately 5 000 observations and halves the maximum travel time, decreasing from a maximum of 243 minutes (4 hours) to 124 minutes (2 hours). We do this to avoid considering rare customers who only consume when they are on vacation in the PACA region in order to focus solely on regular customers.

3.1 Sales model

Of the three possible gravitation models for estimating sales, namely the Reilly model, the logarithmic spatial linear interaction model and the non-linear spatial interaction model, the choice of the logarithmic spatial linear interaction model seems the most appropriate for several reasons. Firstly, this model offers a direct and intuitive way of interpreting the estimated coefficients, notably through the elasticities. By taking the logarithm of the explanatory variables, we linearize the relationships between the variables, making it easier to interpret the effects of each variable on sales. In addition, this linearization reduces distortions caused by extreme values and large variations in the data. Finally, this model is easy to implement, since we can estimate it using the lm function on R.

In addition, we have chosen to add one to the logarithmic transformation of the explanatory variables, a practice adopted to avoid problems associated with zero or near-zero values. By adding a unit to the values before taking the logarithm, we ensure that zero values do not generate undefined or infinite results in the log transformation, which could distort the model estimates. In this way, the transformation helps to stabilize the estimates and guarantee the model's robustness to such situations.

The equation of sales model is therefore as follows :

$$\log(\text{sales}_{i,j}) = \beta_0 + \beta_1 \times \log(\text{distance}_{i,j}) + \beta_2 \times \mathbf{X}_i^T \mathbf{W}_i + \epsilon_{i,j} \tag{1}$$

Where

- i denotes an IRIS in PACA and j denotes a shop
- $\mathbf{X}_i = [\log(\text{P18\_RP\_VOIT2P}_i + 1), \log(\text{P18\_SCOL1824}_i + 1), \log(\text{C18\_MENCOUPSENF}_i + 1), \log(\text{P18\_RP\_PROP}_i + 1), \log(\text{P18\_NSCOL15P\_SUP5}_i + 1), \log(\text{P18\_POP15P}_i + 1)]$
- $\mathbf{W}_i$ represents the coefficient vectors associated with the variables $\mathbf{X}_i$.
- $\epsilon_{i,j}$ is the error term of location(i, j)

The model outcomes are detailed in the appendices, while the R code for model implementation is presented below :

```
log_sales_model <- lm(log(sales + 1) ~ log(minutes + 1) + log(P18_RP_VOIT2P
    +1) + log(P18_SCOL1824 +1) + log(C18_MENCOUPSENF+1) + log(P18_RP_PROP+1) +
     log(P18_NSCOL15P_SUP5+1) + log(P18_POP15P+1), data = data)
```

Listing 1: Sales model code

Regarding the model performance, the adjusted R-squared value is 48.1, indicating that approximately 48.1% of the variability in sales is explained by the independent variables in the model.

All coefficients of the model are significant at the 1% level, except for the variable **P18_RP_PROP**, which is not significant. This suggests that the number of owner-occupied primary residences have not a significant impact on sales. All others coefficients are interpreted in elasticity since both the explanatory variables and the dependent variable are expressed in logarithm in the model.

- The estimated coefficient for the variable **minutes** is -2.519. This implies that, holding all other variables constant, an increase in the travel time (in minutes) to a shop by 1% results in a decrease of 2.519% in sales. This means that the farther a shop is from the residential area, the fewer consumers are likely to visit.

- Regarding **P18_RP_VOIT2P**, the coefficient is estimated to be 0.140. This suggests that an increase of 1% in the number of households with at least two cars in an IRIS implies an increase of 0.14% in sales. This is explained by the fact that a family with at least 2 cars makes them more mobile for shopping and indicating a higher purchasing power as well.

- For **P18_SCOL1824**, the coefficient is estimated to be -0.094. This implies that a 1% increase in the number of 18-24 year olds attending school results in a decrease of 0.094% in sales. We can therefore assume that the brand doesn't necessarily target the 15-24 age group, or simply that students may not have enough purchasing power to consume in these stores.

- The estimated coefficient for **C18_MENCOUPSENF** is 0.280 which indicates that an increase of 1% in the number of couples without children results in an increase of 0.280% in sales. We can thus assume that this brand is not intended for children.

- In the case of **P18_NSCOL15P_SUP5**, the coefficient is estimated to be 0.187. This implies that a 1% increase in the number of people higher than 15 year olds with at least 5 years of higher education is associated with an increase of 0.187% in sales. Individuals with a Bachelor's degree or higher tend to spend more at the store, which can be attributed to their likely higher income and consequently greater purchasing power.

- Finally, for **P18_POP15P**, the coefficient is estimated to be -0.155 which indicates that an increase of 1% in the population aged 15 and over results in a decrease of 0.155% in sales. One

possible explanation for this could be that an increase in the population aged 15 and over might lead to higher competition among businesses, resulting in a fragmentation of market share for the specific store in consideration. Moving forward, it's important to delve into market share modeling to better grasp how these demographic changes may influence competition and the store's position in the market.

3.2 Market share model

For market share modeling, we have several options : two of the most commonly used being the Huff model and the MCI (Multiplicative Competitive Interaction) model. On one hand, the Huff model is a probabilistic approach to gravitation that aims to estimate the probability that an individual chooses a particular store among all the stores available in an area. It takes into account both the attraction of each store and the distance between the potential consumer and each store to estimate the probability that a consumer visits a specific store. On the other hand, the MCI model which is a generalization of the Huff model, extends the concept of modeling market shares to situations with multiple choices, making it more flexible for modeling consumption behaviors in environments where consumers have multiple options to choose from. Unlike the Huff model, which is primarily designed for single-choice situations, the MCI model can be used to estimate market shares when consumers have multiple available alternatives. The MCI model also considers the relative attraction of each option in the market, but it allows for estimating market share for each option while taking into account competition between them.

For the market share model in PACA region, our aim is to model the ratio between the sales made in a particular store by consumers from a specific IRIS in the PACA region, and the market potential of that IRIS. For this purpose, we decided to use a Huff model, which predicts the probability that a consumer chooses a specific shop for making a purchase, taking into account various factors such as distance, attractiveness of the location, and other characteristics. Our initial approach was to establish a logistic regression model because the concept behind calculating market shares is to compute a probability of capturing a market, which naturally falls between 0 and 1, making it suitable for logistic models. However, given that we have data on the sales of each shop per IRIS in PACA, we encountered many instances where there were no sales recorded. Consequently, we had a mass at zero, which hindered the proper fitting of a probit model. Therefore, we opted for an Ordinary Least Squares (OLS) model.

The equation of market share model is as follows :

$$\text{market\_share}_{i,j} = \frac{\text{sales}_{i,j}}{\text{market\_potential}_i} = \beta_0 + \beta_1 \times \log(\text{distance}_{i,j}) + \beta_2 \times \mathbf{X}_i^T \mathbf{W}_i + \epsilon_{i,j} \qquad (2)$$

Where

- i denotes an IRIS in PACA and j denotes a shop

- $\mathbf{X}_i = [\log(\text{P18\_RP\_VOIT2P}_i + 1), \log(\text{P18\_SCOL1824}_i + 1), \log(\text{C18\_MENCOUPSENF}_i + 1), \log(\text{P18\_RP\_PROP}_i + 1), \log(\text{P18\_NSCOL15P\_SUP5}_i + 1), \log(\text{P18\_POP15P}_i + 1)]$
- $\mathbf{W}_i$ represents the coefficient vectors associated with the variables $\mathbf{X}_i$.
- $\epsilon_{i,j}$ is the error term of location(i, j)

The model implemented in R is provided below, and the detailed results of the model are presented in the appendices.

```
ols_market_share <- lm(market_share ~ log(minutes + 1) + log(P18_RP_VOIT2P +1)
    + log(P18_SCOL1824 +1) + log(C18_MENCOUPSENF+1) + log(P18_RP_PROP+1) +
   log(P18_NSCOL15P_SUP5+1) + log(P18_POP15P+1), data = data)
```

Listing 2: Market share model code

The adjusted R-squared value is 0.09215, indicating that approximately 9.215% of the variability in market share is explained by the independent variables in the model which is much lower than in the sales model. The explanatory variables of shop sales do not seem to be as relevant in our market share model. We can justify a lower adjusted R-squared by the fact that estimating market shares is much more complex due to addressing much smaller variations, given that we are dealing with percentages rather than absolute values, as observed in sales.

Concerning the significance of the variables, we have all the coefficients that are significant at the 1% level, except for the variable **P18_POP15P**. This means that the number of people aged 15 and over in an IRIS does not impact the market share of the stores in that IRIS. This variable had a negative coefficient in the sales model, and we justified this by the fact that the higher the population in an IRIS, the greater the competition. This is confirmed in the market share model, which indicates no impact of this variable. Regarding the other coefficients, we also interpret them in elasticity since the market share model is also in log-log form.

- The coefficient for the variable **minutes** is -0.001 which mean that an increase of 1% in travel time to a results in a decrease of 0.001% in market share. We obtain the same sign of the coefficient as in the sales model, which seems consistent: the farther the distance between a consumer and a shop, the fewer market shares the shop is likely to gain from that consumer.
- For the variable **P18_RP_VOIT2P**, the coefficient is estimated to be -0.0002. Thus, an increase of 1% in the number of households with at least two cars in an IRIS implies an decrease of 0.0002% in market share in this IRIS. In the sales model, we observed a positive coefficient for this variable, but it appears negative in the market share model. This could be explained by the fact that households with multiple cars are more mobile and therefore have easier access to competing stores. Additionally, it's possible that these households are more affluent and the type of stores offered may not align with their expectations or preferences.

- Regarding the variable **P18_SCOL1824**, the estimated coefficient is -0.0001 which mean that a 1% increase of the number of schooled people between 18 and 24 years old in an IRIS will decreases the market share by 0.0001% in this IRIS. We also had a negative coefficient in the sales model, which we justified by assuming that individuals aged 18 to 24 were not the target demographic for the brand under study.

- The coefficient for the variable **C18_MENCOUPSENF** is positive at 0.0003 which indicates that an increase of 1% in the number of households without children results in an increase of 0.0003% of market share. This is in line with the results of the sales model. We thus assume that the brand is not aimed at children.

- Concerning the variable **P18_RP_PROP** which was not significant in the last model, its coefficient is -0.0002. An increase of 1% in the number of owner-occupied primary residences will results in a decrease of 0.0002 of market share. We can further justify these results by considering that individuals residing in owner-occupied primary residences, typically associated with higher socioeconomic status, may not exhibit a strong affinity towards this particular brand.

- For the variable **P18_NSCOL15P_SUP5**, the estimated coefficient is 0.0002. This implies that the market share will increase by 0.0002% after a 1% increase in the number of people aged over 15 with at least 5 years of higher education. This is still consistent with our previous model and is justified by higher purchasing power.

4 Principal trading areas of your shops

In this section, we focus on the areas around each store from which the majority of customers originate. These areas are called core catchment areas, and they generally cover around 80% of the sales generated by the store. We need to create these zones for all stores in the PACA region. Before defining the main catchment areas, we plotted the distribution of distances between customers and stores.

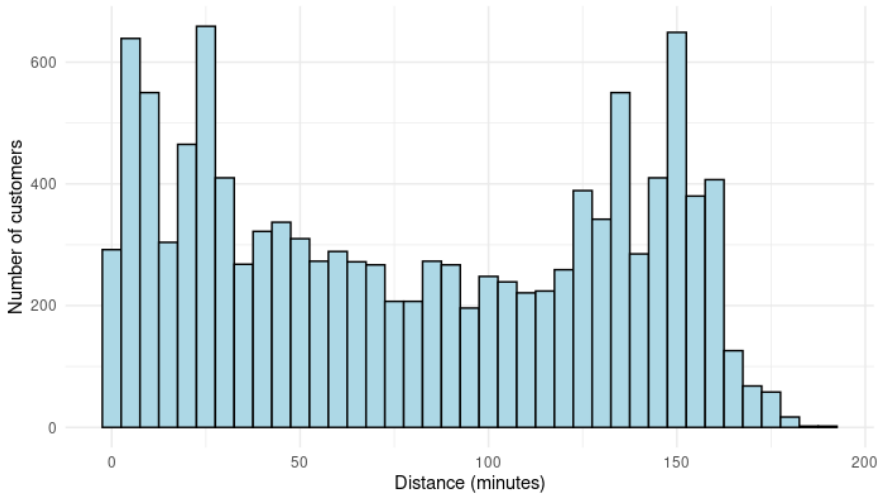

Figure 7: Distribution of distances between customers and shops

The distribution of travel distances in minutes between consumers and shops ranges from 0 to 200 minutes (equivalent to 2 hours and 30 minutes) and exhibits two peaks: one around 30 minutes and another around 150 minutes.

Then, we have defined the main trading areas using the cumulative turnover method : we sort the data in ascending order of travel time and derive the cumulative turnover per shop. Then we select for each shop, the IRIS areas where there are the most consumers per shop, while retaining those IRIS areas with a cumulative sale below the 80th percentile. We apply the same method to obtain the trade areas using predicted sales instead of actual sales. This allows us to determine the maximum distance traveled by the top 80% of consumers and to define the corresponding IRIS areas as the main catchment area. The maximum distance obtained with actual sales data is 135 minutes, compared to 146 minutes with predicted sales data.

For the purpose of graphic representation, we then focus on the store with the most customers. In the PACA region, the store with the highest number of customers is the shop number 18, with 1 546 customers. To graphically compare the catchment areas generated by the two methods, we create a map for each method, highlighting the main catchment area identified for each method.

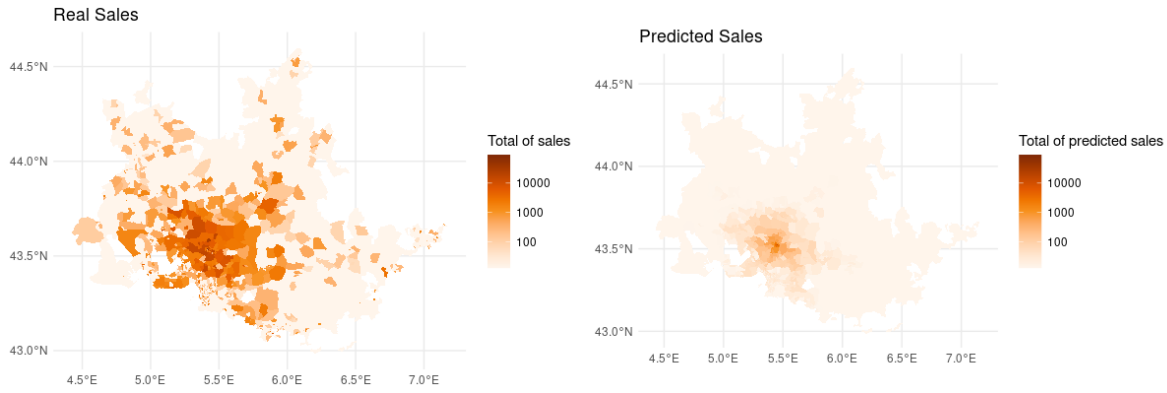

Figure 8: Comparison of trading areas between real and predited sales

We observe that the principal trading areas differ between actual sales data and predicted ones. This discrepancy can be attributed to our sales prediction model, which heavily relies on the **minutes** variable, measuring the travel time between consumers and shops. Consequently, we obtain a nuanced map based on the location of shop 18, where lighter colors indicate greater distance and thus lower sales predictions.

5 Evaluating potential new shops

In order to evaluate potential new shops, we randomly select 10 competitors shops from the PACA region. We obtained 1 shop in the Var department (83), 3 shops in the Alpes-Maritimes department (06), and 6 shops in the Bouches-du-Rhône department (13). We have plotted these new shops in red on the map below, adjacent to the existing shops plotted in yellow.

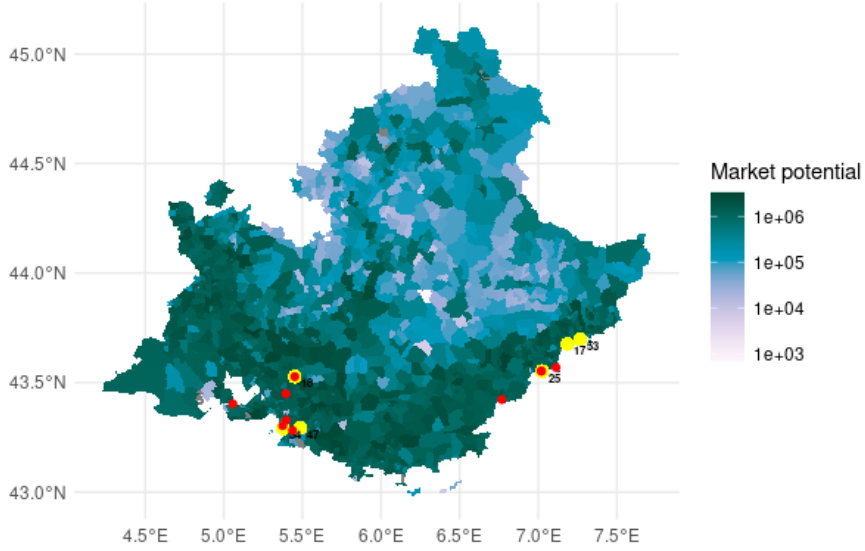

Figure 9: Distribution of new shops and market potential in the PACA region

We then add to these data all the explanatory variables used in our models to be able to apply them to these new shops. For this, we had to use ORSM again locally to calculate the travel times in minutes between consumers and shops. Then, we created a database containing all pairs of IRIS in the PACA region and shops in order to apply our models.

We estimate predicted sales and predicted market shares per store per IRIS from the previous models using the following R commands. We replace negative values predicted by the models with 0 for consistency. Additionally, as we modeled the logarithm of sales to obtain sales, we used the exponential function to reverse the logarithmic transformation. Furthermore, we subtracted 1 to retrieve the exact sales level.

```r
# Sales predicted
ns_data$predicted_sales <- exp(predict(log_sales_model, newdata = ns_data)) -
    1
ns_data$predicted_sales[ns_data$predicted_sales < 0] <- 0

# Market share predicted
ns_data$predicted_market_share <- predict(ols_market_share, newdata = ns_data)
ns_data$predicted_market_share[ns_data$predicted_market_share < 0] <- 0
```

Listing 3: Sales model code

From the estimated sales, we can once again calculate the main trading areas as we did previously for the shops in PACA. Similarly, we only retain the IRIS representing 80% of the turnover for each shop. You will find the graphs of these areas for the 10 new shops in the appendices.

Below is the table for each new shop summarizing their total sales, their total market share, as well as the number of IRIS overlapping with the IRIS of the shops already installed in PACA sort in decresing order of the total sales.

Table 3: Summary of new shops

New shop id	Sales	Market share (%)	IRIS coverage
ns2	567 193.85	2.15	971
ns1	405 367.34	2.12	980
ns8	197 086.70	1.19	1045
ns9	156 435.14	1.20	980
ns6	92 381.46	0.91	798
ns3	59 855.30	0.35	64
ns10	35 183.46	0.79	519
ns4	34 524.40	0.85	808
ns7	11 741.79	0.41	697
ns5	10 921.39	0.66	788

The new shop 2 stands out with an impressive revenue of 567 193, giving it a market share of 2.15%. It competes in 971 IRIS, indicating a widespread presence in the region. Similarly, shop 1 shows significant sales of 405 367, with a market share of 2.12% and coverage of 980 IRIS. However, the performance of other new shops is relatively modest in comparison. For instance, shops 8 and 9 achieve sales of 197 086 and 156 435 respectively, with market shares of 1.19% and 1.20%, but have a higher coverage of IRIS, 1 045 and 980 respectively. On the other hand, shops 7 and 5 display more modest figures in terms of sales and market share, with sales of 11 741 and 10 921 respectively, and market shares of 0.41% and 0.66%. These shops cover 697 and 788 IRIS respectively. Thus, the new shop number 2 appears to be the most promising location.

6 Conclusion

In conclusion, our geomarketing analysis of the Provence-Alpes-Côte d'Azur (PACA) region has provided valuable insights into the market, consumer dynamics particularly in trades zones and strategic perspectives for new store locations by leveraging various statistical models and spatial analysis techniques. Our study began with a thorough examination of the PACA region, defining its geographical boundaries and outlining its population distribution.

Then, before performing sales and market share models, we prepared the data available to us, by consolidating information at various levels, including individual shops, IRIS units, and IRIS/store pairs, to facilitate comprehensive analysis. Our analysis revealed the significance of factors such as travel time, population demographics, and competition density in shaping consumer behavior and market dynamics. We found that proximity to stores and accessibility play crucial roles in driving sales, with customers generally exhibiting a preference for nearby establishments. Additionally, variables such as household composition, education levels, and car ownership emerged as relevant determinants of consumer spending patterns.

After that, by applying our predictive models to assess potential new shop locations, we identified promising opportunities for expansion, with some locations showing strong sales potential and market share. However, we also observed variations in performance across different locations, highlighting the importance of considering a range of factors, including competition and demographic characteristics, when evaluating new site opportunities.

Finally, it is essential to acknowledge the limitations of our study. Notably, the absence of detailed information regarding the nature of shops and their product offerings restricted our ability to tailor variables accurately to consumption patterns, potentially influencing the reliability of our models. In addition, our findings underscore the importance of considering factors such as proximity, however, as highlighted by Michaud-Trévinal's [2] research cited earlier in the report: the majority of customers are willing to travel further for specialty stores like Ikea and on the contrary will prefer to go closer for everyday purchases. In short, knowing the type of store we're talking about could enable us to go further in our analysis. Furthermore, it is conceivable that our analysis did not encompass all crucial variables, which could impact the accuracy of our results.

Appendices

Table 4: Results of the sales model

	Dependent variable:
	log(sales + 1)
log(minutes + 1)	-2.519^{***}
	(0.026)
log(P18_RP_VOIT2P + 1)	0.140^{***}
	(0.037)
log(P18_SCOL1824 + 1)	-0.094^{***}
	(0.034)
log(C18_MENCOUPSENF + 1)	0.280^{***}
	(0.078)
log(P18_RP_PROP + 1)	0.003
	(0.042)
log(P18_NSCOL15P_SUP5 + 1)	0.187^{***}
	(0.034)
log(P18_POP15P + 1)	-0.155^{**}
	(0.066)
Constant	10.588^{***}
	(0.216)
Observations	11,683
R^2	0.482
Adjusted R^2	0.481
Residual Std. Error	2.208 (df = 11675)
F Statistic	$1{,}549.479^{***}$ (df = 7; 11675)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Results of the market shares model

	Dependent variable:
	market_share
log(minutes + 1)	-0.001^{***}
	(0.00004)
log(P18_RP_VOIT2P + 1)	-0.0002^{***}
	(0.0001)
log(P18_SCOL1824 + 1)	-0.0001^{**}
	(0.0001)
log(C18_MENCOUPSENF + 1)	0.0003^{**}
	(0.0001)
log(P18_RP_PROP + 1)	-0.0002^{***}
	(0.0001)
log(P18_NSCOL15P_SUP5 + 1)	0.0002^{***}
	(0.0001)
log(P18_POP15P + 1)	-0.00004
	(0.0001)
Constant	0.007^{***}
	(0.0004)
Observations	11,683
R^2	0.093
Adjusted R^2	0.092
Residual Std. Error	0.004 (df = 11675)
F Statistic	170.387^{***} (df = 7; 11675)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 10: Predicted sales for new shops

References

[1] INSEE. Institut national de la statistique et des études économiques. https://www.insee.fr/fr/accueil, 2024.

[2] Aurélia Michaud-Trévinal. Le comportement spatial des consommateurs : conceptualisation et exploration des parcours piétonniers de magasinage : le cas de l'équipement de la personne. https://theses.hal.science/CREM-MM/tel-02401349v1, 2004.