Modeling the "average basket" of a retail store using Geographically Weighted Regression

Paula Duarte Guerra
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide
Lisboa, Portugal
g2016066@novaims.unl.pt

Ana Cristina Costa
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide
Lisboa, Portugal
ccosta@novaims.unl.pt

Abstract

The decision on defining the location of a store is of great importance to the retail companies because of the levels of commitment and investment that implies, so it is dependent on the billing expectation. Among the explanatory variables of the sales performance of a store is the average shopping bill of the customers ("average basket"). This study aimed at investigating a linear regression model to explain the variability of the "average basket" of a retail store located in the north of Portugal. Ordinary Least Squares (OLS) was applied to many alternative regression models, which were diagnosed for all OLS assumptions. The explanatory variables of the model that passed the diagnosis criteria were then included in a Geographically Weighted Regression (GWR) model, namely the "Number of family rented houses per km^2", the "Direct distance to the point of sale" and the "Number of main residences in buildings with two dwellings per km^2". Results show a distinct pattern of the predictors' parameters in the northwest area, which is mainly a fishing and bathing area.
Keywords: retail, spatial non-stationarity, spatial regression, OLS, GWR, Portugal.

1 Introduction

In the 'retail mix', location is often considered the most important element (Reigadinha 2015). The creation of a new store represents not only a substantial investment (almost always), but also a long-term commitment (Owen 1998). Therefore, both must be balanced against the expected return, which makes the decision to place a point of sale a microeconomic problem, dependent on the billing forecast of the store (Krause-Traudes 2008). In the area of applied economics, several models have been developed that aim at incorporating the space component in the billing forecast of a point of sale (Anderson 2004).

There are several mathematical models, more or less complex, applied to the location of new stores, with different perspectives: minimising the distances covered, maximising coverage, etc. (Cheng 2004). However, in practice, despite the existence of complex mathematical models for the choice of locations, the effort and knowledge required for their application make them out of reach for most retailers. Only a small number of retailers are using very complex or sophisticated approaches, with most of them using a combination of science and intuition in decision making (Hernandez, 2007).

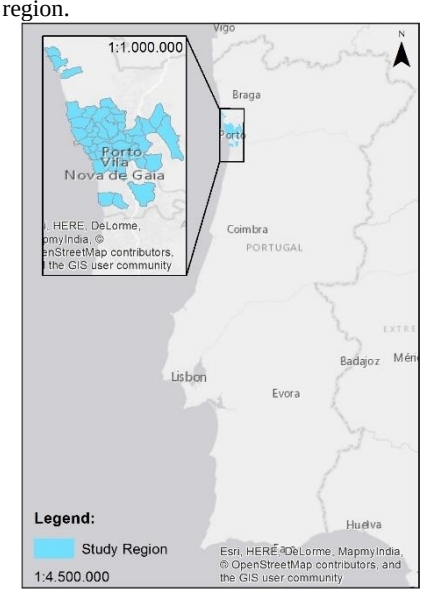
Mendes (2005) developed a study on small and medium-sized food retail shops, in Portugal, where he identified the main explanatory variables of the billing in these stores: sales area, depth of range, chain image, accessibility, visibility of the store, area of competitors, competition quality, dimension of the area of influence, demographics, average basket, consumer preferences and income classes.

This study aims to explain the behaviour of the average basket, which may also be influenced by several factors.

2 Study region and data

The study area of this work is the territory of continental Portugal covered by customers who have visited a particular retail point of sale in the municipality of Matosinhos in northern Portugal (Figure 1). The average basket data were obtained by collecting postal codes of the customers' dwellings of a retail point of sale of a large surface area. Additionally, 37 potential explanatory variables' data (described in the Appendix) were collected. The data from most of them corresponds to 2011 Census data at parish level collected by Statistics Portugal.

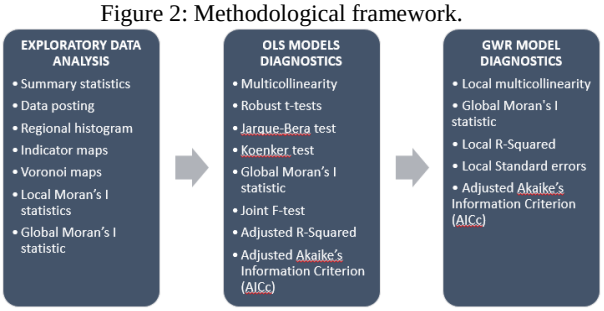
Figure 1: Municipalities in the north of Portugal that define the study region.



3 Methodology

The methodological framework included three major steps (Figure 2): 1) exploratory data analysis; 2) Ordinary Least Squares (OLS) modelling and diagnosis; 3) Geographically Weighted Regression (GWR) modelling and diagnosis. The GWR model used an adaptive kernel with a near-Gaussian weighting function given by $w_{ij} = exp[-(d_{ij}/b)^2]$, where d_{ij} is the distance between i and j, and b is the bandwidth, which is determined using the Adjusted Akaike's Information Criterion (AICc).

Figure 2: Methodological framework.



4 Results and discussion

4.1 Exploratory data analysis

The sample size was equal to 47 average baskets, which varied between 29,16€ and 83,75€. The average of sample values was 54,08€ and the median was 53,79€. The standard deviation was 9,99€. The lower values were found in the centre of the study area, around the location of the point of sale, but with values exhibiting an ellipse shape (west–east), while the highest values seem to form a ring (ellipsoid) around them, with a tendency to increase with distance.

Regarding the local spatial autocorrelation of the average basket, a cluster of low values was identified immediately to the north/northeast of the location of the point of sale, and a parish (in the north of the study area) was identified as an outlier of low values, which was excluded from the set of data to be used in the regression analysis. A clustered pattern was also identified in the average basket using the Global Moran's I statistic (Moran's Index = 0.23 with p-value = 0.0003).

4.2 OLS models and diagnostics

Considering the diagnostics results of the parameter estimates, residuals and overall goodness of fit, the best OLS model found (Table 1) was the one that used as predictors the "Number of family rented houses per km^2", the "Direct distance to the point of sale" and the "Number of main residences in buildings with two dwellings per km^2". These variables do not exhibit multicollinearity (small VIF value). Considering the 1% significance level, there is also evidence that the residuals are normally distributed (Jarque-Bera statistic) and homoscedastic (Koenker statistic), and that the spatial processes promoting the observed pattern of residual values is random chance (Global Moran's I statistic). The model only explains 44% (Adjusted R-Squared) of the variability of the average basket, but its global fit (Joint F-Statistic) is statistically significant at the 0.1% level.

Table 1: Final OLS model results and diagnostics.

Diagnostics statistics	
Adjusted R-Squared	0.44
AICc	319.88
Joint F-Statistic	12.74*
Variance Inflation Factor (VIF)	2.12
Jarque-Bera statistic	6.68
Koenker statistic	5.22
Global Moran's I statistic	0.11**
Model parameters	
Intercept	44.5916
Nr. of family rented houses per km^2	+0.0192*
Direct distance to the point of sale (meters)	+ 0.0009*
Nr. of main residences in buildings with two dwellings per km^2	–0.0931*
* Significant at the 0.1% level ** p-value	

4.3 GWR model

GWR was applied with the same variables of the described OLS model. The GWR model proved to have a better fit to the data, having a higher Adjusted R-Squared (58%) and a smaller AICc (308.08) than the OLS model, without evidencing problems through the possible diagnostic methods. It is important to point out that it is unclear what statistical tests can reliably diagnose GWR models' problems (Páez, Farber & Wheeler, 2011; Wheeler & Tiefelsdorf, 2005), and that the coefficients can be correlated even when there is no collinearity among explanatory variables (Griffith, 2008; Wheeler & Tiefelsdorf, 2005).

The spatial distribution of the residuals seems to be random (Figure 3), which was confirmed by the Global Moran's I statistic (p-value = 0.5978). All the parishes had a Condition Number less than 7, so there is no evidence of multicollinearity among the predictors. The Local R-Squared values vary between 0.45 and 0.68, with the higher values occurring in the northwest area and then decreasing towards the south (Figure 4).

The GWR approach provided additional insights about the regional variation of the explanatory variables (Figure 5), even though their coefficients can be unreliable because each local regression was based on few observations. In the extreme northwest of the study area, the average basket was better explained by the "Number of family rented houses per km^2" and the "Direct distance to the point of sale". This region corresponds to a fishing and bathing area, thus customers living there have distinct characteristics. Hence, it would be relevant to separately model that zone in future

studies. Likewise, it would be important to investigate the inclusion of other variables in the modelling process, particularly in the model of the other parishes of the study region.

Figure 3: Residuals of the GWR model.

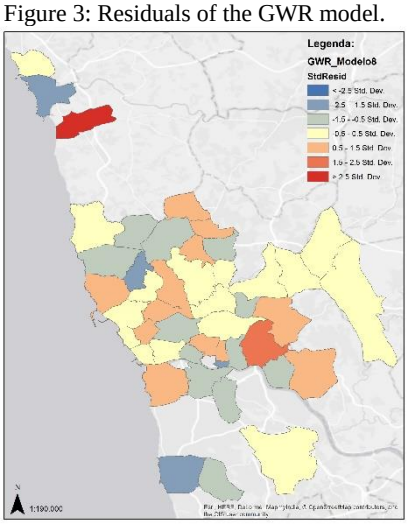


Figure 4: Local R-Squared of the GWR model.

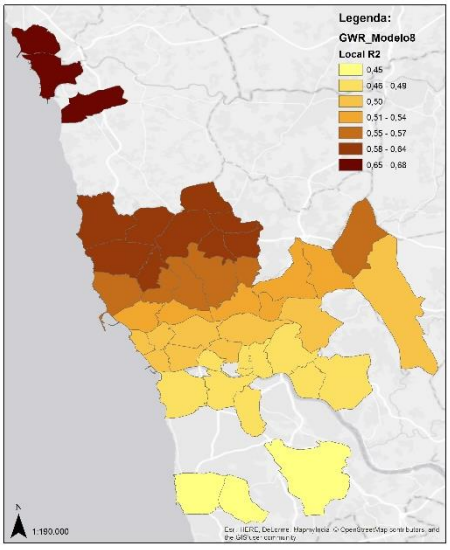
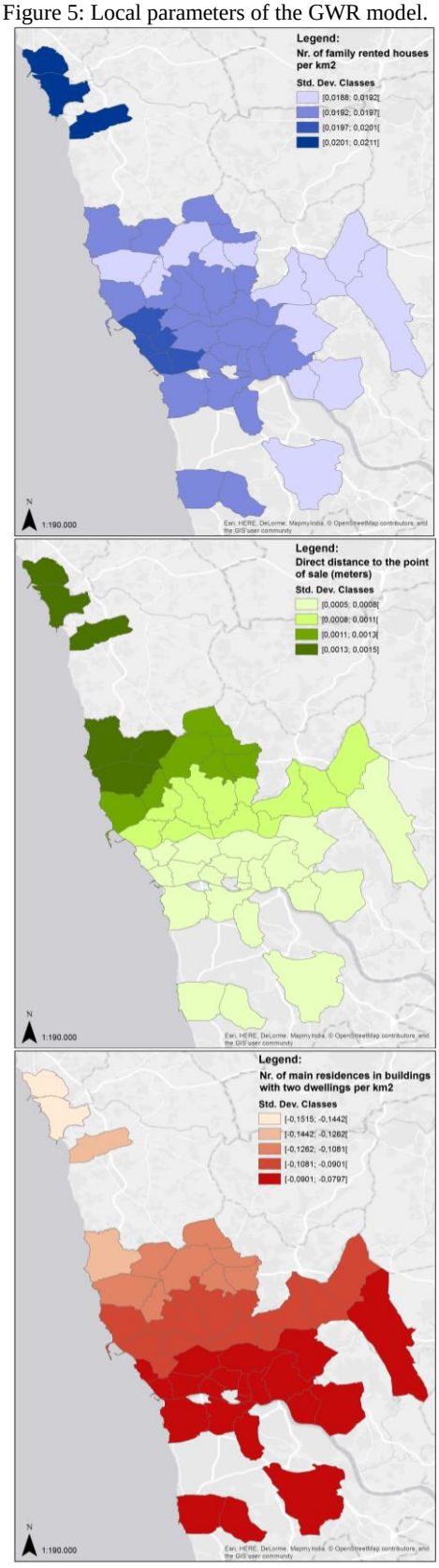


Figure 5: Local parameters of the GWR model.



References

Anderson, P. (2004) *Business Economics and Finance with MATLAB, GIS, and Simulation Models*. Chapman & Hall/CRC Press LLC, Florida (USA), Ch.13.

Cheng, E. W. L., & Li, H. (2004) Exploring quantitative methods for project location selection. *Building and Environment*, 39(12), 1467–1476.

Griffith, D. (2008) Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, 40(11), 2751–2769.

Hernandez, T. (2007) Enhancing retail location decision support: The development and application of geovisualization. *Journal of Retailing and Consumer Services*, 14(4), 249–258.

Krause-Traudes, M., Scheider, S., Rüping, S., Meßner, H. (2008) *Spatial data mining for retail sales forecasting*. 11th AGILE International Conference on Geographic Information Science 2008, University of Girona, Spain.

Mendes, A. (2005) *Modelação de vendas de novas superfícies comerciais*. PhD Thesis, Technical University of Lisbon, Portugal.

Owen, S. H., & Daskin, M. S. (1998) Strategic facility location: A review. *European Journal of Operational Research*, 111, 423–447.

Páez, A., Farber, S., & Wheeler, D. (2011) A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A*, 43(12), 2992–3010.

Reigadinha, T., Godinho, P., Dias, J. (2015) Portuguese food retailers – Exploring three classic theories of retail location. *Journal of Retailing and Consumer Services*, *34*, 102–116.

Wheeler, D., & Tiefelsdorf, M. (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2), 161–187.

Appendix

List of the 37 explanatory variables investigated:

- Nr. of buildings per km^2
- Nr. of buildings with one or two dwellings per km^2
- Nr. of buildings with three or more dwellings per km^2
- Nr. of main residences with an area below 50 m^2 per km^2
- Nr. of main residences with an area between 50 m^2 and 100 m^2 per km^2
- Nr. of main residences with an area between 100 m^2 and 200 m^2 per km^2
- Nr. of main residences with an area over 200 m^2 per km^2
- Nr. of main residences with 1 or 2 rooms per km^2
- Nr. of main residences with 3 or 4 rooms per km^2
- Nr. of family owned houses per km^2
- Nr. of family rented houses per km^2
- Resident individuals with the first cycle of elementary education per km^2
- Resident individuals with the second cycle of elementary education per km^2
- Resident individuals with the third cycle of elementary education per km^2
- Resident individuals with secondary education per km^2
- Resident individuals with higher education per km^2
- Direct distance to the point of sale (meters)
- Nr. of main residences per km^2
- Nr. of secondary residences per km^2
- Market Size in Euros per km^2
- Resident individuals per km^2
- Male resident individuals per km^2
- Female resident individuals per km^2
- Resident individuals with ages between 0 and 14 per km^2
- Resident individuals with ages between 15 and 24 per km^2
- Resident individuals with ages between 25 and 64 per km^2
- Resident individuals with ages over 64 per km^2
- Nr. of main residences in buildings with one dwelling per km^2
- Nr. of main residences in buildings with two dwellings per km^2
- Nr. of main residences in buildings with three or more dwellings per km^2
- Unemployment rate
- Male unemployment rate
- Female unemployment rate
- Resident individuals studying in the municipality of residence per km^2
- Resident individuals working in the municipality of residence per km^2
- Employed resident individuals per km^2
- Pensioners resident individuals per km^2