A Consumer Income Predicting Model Based on Survey Data: an Analysis Using Geographically Weighted Regression (Gwr)

Eduardo Francisco, FGV-EAESP, Brazil
Peter Whigham, University of Otago, New Zealand
Francisco Aranha, FGV-EAESP, Brazil
Felipe Zambaldi, Universidade Metodista de Sao Paulo (UMESP), Brazil

Based on data of 662 households from 75 districts in the city of Sao Paulo, this paper investigates the relations between electricity consumption and household income with use of geographic weighted regressions (GWR). Findings reveal that electricity consumption is useful for characterizing household income, a frequently used proxy for purchasing power. The employed GWR were more effective to the studied task than traditional linear regressions. Also, alternatives for allocation of points were analyzed, because their exact locations were not available. The results may be useful for marketing professionals, policy makers, and credit agents who are committed to characterizing consumers socio-economically.

[to cite]:

[url]:
http://www.acrwebsite.org/volumes/14064/la/v2_pdf/LA-02

[copyright notice]:
This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
A Consumer Income Predicting Model Based on Survey Data: An Analysis Using Geographically Weighted Regression (GWR)

Eduardo Francisco, FGV-EAESP, Brazil
Peter Whigham, University of Otago, New Zealand
Francisco Aranha, FGV-EAESP, Brazil
Felipe Zambaldi, UMEP, Brazil
Mateus Ponchio, UNIMEP, Brazil

ABSTRACT

Based on data of 662 households from 75 districts in the city of São Paulo, this paper investigates the relations between electricity consumption and household income, with use of geographic weighted regressions (GWR). Findings reveal that electricity consumption is useful for characterizing household income, a frequently used proxy for purchasing power. The employed GWR were more effective to the studied task than traditional linear regressions. Also, alternatives for allocation of points were analyzed, because their exact locations were not available. The results may be useful for marketing professionals, policy makers, and credit agents who are committed to characterizing consumers socio-economically.

INTRODUCTION

A substantial portion of marketing research is still empirical and exploratory, a scenario which is similar to the one described by Sheth (1971) a few decades ago. Since then, marketing researchers have employed countless qualitative and quantitative techniques to analyze data, perhaps helped by the rapid development of fast computer processing and statistical packages.

Nowadays, it is very common to collect “spatially enabled” survey data. The motivation has come from several disciplines, including the universe of marketing and social sciences, and the subsequent improvements that spatial statistics allow in the interpretation, measurement of relationships, and prediction. As pointed out by Bradlow et al. (2005), by generalizing the notion of a map to include demographic and psychometric representations, spatial models can capture a variety of effects (spatial lags, spatial autocorrelation, and spatial drift) that affect firm or consumer decision behavior.

However, it is also common that the data related to these subjects is not quite suitable for spatial analysis. Many forms of data collection do not make available adequate information about location (in accuracy or precision)-these situations produce data for which the precise location of each observation is not known--making available just information about the region where the observation is located (like district, region, postal code, or municipality).

This situation could be viewed as the opposite of the known “Modifiable Areal Unit Problem” (Jelinski and Wu 1996), which describes the effect on the observed spatial relationships of data due to scaling and zonation. For the situation described in this paper, rather than having a set of point data that can be aggregated in a variety of ways, we are given a fixed zonation with associated data and intend to place these data within the zone in an effective way.

A helpful tool to address such issues is geographically weighted regression (GWR), which takes advantage of data from locations near the focal location, being more informative about the relationship between the independent and dependent variables in the focal location. When calculating the estimates for a focal location, GWR gives more weight to data from locations that are closer to it than to those that are distant (Mittal, Kamakura, and Govind, 2004).

ENERGY CONSUMPTION AND CONSUMER INCOME

Good examples of GWR applications in marketing are the studies of Lu et al. (2006) that modeled the relationship between area brand performance and related marketing phenomena taking the sales of an European car brand; and of Mittal et al. (2004), that used GWR to develop an approach that enables a firm to identify regional patterns in data about satisfaction, providing guidance in the implementation of a service strategy on a national basis.

This paper investigates approaches to point allocation inside polygons using an empirical study by applying GWR models on a specific survey of the Brazilian power distribution sector. The regression models constructed aim to predict household income with only one independent variable: the monthly billed residential electricity consumption (or simply energy or electricity consumption). Previous studies have related the consumption of electricity to income concentration in Brazil. Araújo (1979) characterized the domestic consumption of electricity by means of household income variability. Pompermayer and Charnet (1996) found statistically significant influences of social-economic and demographic factors on the consumption of electricity in the State of São Paulo.

Results reveal that the GWR models implemented in this context lead to superior fit to data when compared to traditional linear models (LM). They also reinforce the argument that this spatial-econometrics technique can effectively be applied to improve marketing research.

The next section describes the relevance of the study. In the sequence, the main concepts applied in the model, the GWR technique, and point allocation alternative methods are shown in the methodology chapter. Results are then described and, finally, concluding remarks are presented with suggestions for future work.

INCOME CONCENTRATION AND CONSUMER INCOME

Income can be understood as the summation of all earnings provided by work and other sources (IBGE 2003) and may be calculated for individuals, families or households. It includes the sum of gross income (before taxation) from work, pensions, government, and public social security programs (such as minimum income, school grants, or unemployment benefits). It is usually the adopted subject descriptor in studies of poverty and living conditions, since it provides access to basic goods and services.

However, accurate indicators of income are difficult to collect, as its declaration is frequently altered in interviews, and it is subject to seasonal changes, thus becoming an imprecise indicator in market researches.

Many research professionals prefer to capture indicators of socioeconomic classification and purchasing power based on sessions and educational levels as proxies for income and welfare. A recent example of such an indicator is the Brazilian Economic Classification Criterion (CCEB), or simply the Brazilian Criterion, created in 1996 by the National Research Enterprises Association (ANEPR). Based on possessions, the indicator results in a scale varying from zero to 34, and segments seven economic classes (ABEP 2004). The Brazilian Criterion, however, faces important
In a traditional linear regression, we assume that the relationship modeled holds everywhere in the study area, that is, the regression parameters are considered “whole-map” statistics. In many situations, this is not correct, as mapping residuals may reveal.

Many different solutions have been proposed for dealing with spatial variation in studied relationships. GWR permits parameter estimates to vary locally. Regression coefficients are determined by examining the set of points within a well-defined neighborhood of each point, using a weighting scheme (normally, bi-square or Gaussian).

A bandwidth that defines the neighborhood is the key factor; it may be defined manually or alternatively by some form of adaptive method such as the minimization of the Akaike Information Criterion (AIC) (Fotheringham, Charlton and Brunsdon 2002; De Smith, Goodchild and Longley 2007).

As the district of each interview is the only locational data recorded in this study, and GWR considers points as the basic spatial unit of observation, the simplest way to proceed is to associate the district’s centroid to each interview. Under that criterion, many interviews are associated with the same “location”.

Looking at this peculiarity, the unit of observation becomes an issue. The units are households, but we do not have any means of geo-coding their location except to the centroid of the district to which they are associated. This means that any households in the same polygon would effectively be stacked one on top of the other at the centroid. This would result in a weight of one for the local sample regressions for each of these stacked points, while any households in adjacent polygons would receive a smaller weight, but again, weights would all be equal.

Centroid GWR should not produce realistic results due to similar weights being given to data that may well be spatially dispersed and, therefore, have intrinsically different neighborhood influences. Since this would be the most naive approach point allocation, it can be viewed as the null model; it is expected that any model that can produce a more realistic spatial allocation of points within each polygon improves the overall spatial regression performance.

Two alternatives of point distribution are applied in this study: Alternative 1-Generation based on Density of Households; and Alternative 2-Generation based on Probability of Fitness for Energy Consumption. The alternative implementations were produced with the statistical environment R 2.6.1 (R Development Core Team 2007), using spatial packages (extensions) MAPTOOLS 0.7-4, SPLANCS 2.01-23 and SPATSTAT 1.12-5. The GWR models were implemented using SPGWR 0.4-7 extension. Both alternatives are described next.

For Alternative 1-Generation based on Density of Households, information about the density of households in the city of Sao Paulo was used as a surrogate for likelihood of survey location. This information was obtained from AES Eletropaulo, which is the unique power distribution company covering the studied area. Hence, every electrified household is a customer of this company.

Using a grid of 100 squared meters, we computed the number of residential customers (e.g. households) per cell per district. Based on this grid, we generated a random point pattern containing \( n \) independent and identically distributed random points, with the density of households’ grid as the specified distribution (common probability density), considering that you have \( n \) interviews in the district.

Alternative 2-Generation based on Probability of Fitness for Energy Consumption is based on the distribution of energy consumption in the city of Sao Paulo, according to the customers of AES Eletropaulo. Alternative 2 generates a grid of fitness for
electricity consumption, from the computation of the average of electricity consumption per cell in the grid. Similarly to Alternative 1, we generate a random point pattern of \( n \) independent points distributed according to this surface.

Figure 1 shows examples of the implementation of both alternatives. Notice that some points sampled in the example of Alternative 2 are located in cells with low density of population. This is because there is a small (but non-zero) probability for these points to be chosen. Of course, the alternatives could be improved by increasing the accuracy of the underlying distribution data. Moreover, this example shows that the densities of households and energy consumption seem to follow opposite patterns; it illustrates different approaches considering two important geographical distributions to support point allocation. A combination of information on both densities could be used to estimate a third model, where locations would be obtained by sampling from a bivariate distribution. The use of sample planning information (e.g. how census sectors and households are sampled and eventually substituted in the field survey) could support alternative distributions.

The alternative approaches to point allocation were tested by generating 1,000 iterations of GWR income-predicting model based on energy consumption for each alternative, considering the AIC minimisation local sample size suggested for each GWR model.

We compared the resulting performance for each alternative with the original GWR model (the null model that used the districts’ centroid for point placement) and with the traditional aspatial Linear Model (LM). We named these original models as “GWR centr” and “LM centr”, respectively.

In addition, we computed the average of income and of energy consumption per district, and used this data set (with just one observation per district) as a simplified aggregation model with each data point associated with the district centroid. The linear regression and GWR models were also applied to these aggregated datasets. These models are referred to as “LM aggr” and “GWR aggr”, respectively. Results and analyses are described in the next section.

RESULTS

The ABRADEE survey applied in the city of Sao Paulo in 2004 had 662 valid respondents. Income and Energy Consumption were collected as continuous variables—in R$ (“reais”, Brazilian currency) and in kilowatt-hour (kWh), respectively. Seventy-five (75) districts were sampled for this survey. Figure 2 shows the map of 96 districts of Sao Paulo (in gray), highlighting (in dark gray) the 75 districts sampled for the ABRADEE survey.

The following graphs (Figure 3) show the scatter-plot of electricity consumption and the household income for centroid
placement and the predicted values of income for “LM centr” and “GWR centr”, respectively. In addition, the same information is showed for “LM aggr” and “GWR aggr”. The aggregate version has 75 observations (one per district).

A significant improvement in the coefficient of determination values were obtained using GWR models (centroid placement varied from 0.19788 to .45426, and the aggregation version varied from .38533 to .54904).

The GWR implementation used an Adaptive Kernel-which employs a fixed number of observations in each neighbourhood instead of a fixed radius of bandwidth- and Gaussian weights Scheme. Figure 4 and Table 1 show the dispersion of the coefficient of determination for each alternative, by means of box-plots of 1,000 computed iterations, in comparison with the original GWR.

There is little difference between any of the alternative approaches, with the $R^2$ coefficient around 0.40. The dispersions—calculated by means of standard deviations—of the coefficients of determinations were also quite similar for all alternatives.

The variation of bandwidth between alternatives showed the same behaviour as the coefficient of determination; both were very highly correlated, what suggests a very important role for AIC minimization. Figure 5 shows the scatter plot and correlation of this relationship.

The important role of AIC minimization-in this context for local models-is shown by the variation in results when points are allocated with each model. The alternatives have been implemented so that bandwidth changes with each different point allocation in the space, and this potentially affects the smoothness of the global model. Surprisingly, the resulting alternatives had lower coefficients of determination than the simple centroid model, even though it was assumed that the point-allocation was adding additional information to the data and therefore should have resulted in improvements in model prediction.

This counter-intuitive result must be further explored to understand in what ways the model behaviour and contribution to variance explanation are interacting in a similar manner to previous research into cross-validation properties in GWR (Farber and Paez 2007).

**FINAL REMARKS**

This paper developed an income-predicting GWR model using electricity consumption as an independent variable, which was shown to be a useful indicator for predicting household income under geographic effects. The consumption of electricity can potentially enrich the economic characterization of households, which is traditionally measured by means of other indicators of consumption and purchasing power-many of which are not easily collected nor periodically stored, such as the possession of goods, for example.

In such a context, the current automated process of registering the amount of kilowatts consumed in a household employed by energy distributors represents a business opportunity for these firms to provide an indicator of consumer welfare to the market. Therefore, this study may inspire the creation of regional indicators of electricity supply, which can be useful for research institutes and organizations dealing with public and urban affairs, customer segmentation, credit policies, and regional models which capture consumers and households in a broad way.

Mainly, this work offers a methodological contribution to the issue of economic characterization of households by discussing and empirically testing the use of GWR in substitution to traditional linear regression models in the prediction of income. The GWR models applied to the ABRADEE survey led to a significant improvement to the explanation of variability for the income-predicting model based on energy consumption. The results of the coefficients of determination changed from 0.19788 to 0.45426 when the centroids were used as the point allocation of every interview conducted in the districts.

We have analyzed alternatives for point allocation based on realistic assumptions about the distribution of observations in the studied sample, in detraction of district centroid placement. The implementation of the alternatives suggests the minimization of AIC (Akaike Information Criterion). However, coefficients of
determination measured for two alternatives of reallocation of points (households interviewed) inside districts’ polygons were revealed to be of around 0.40. It suggests that the use of additional information of the polygon—in this case, the density of population and of energy consumption—has similar impact in the point-location allocation. It also suggests that, in the studied context, the most realistic coefficient of determination for income predicted by means of electricity consumption is around 0.40.

This work has presented an experimental study in the allocation of point patterns with secondary data. The results using spatial point pattern statistics and mixed models have been shown to improve predictions about the socioeconomic characteristics of a population, therefore representing a prolific field for future empirical studies in marketing and, more specifically, customer segmentation.

REFERENCES
FIGURE 4
COEFFICIENT OF DETERMINATION AND DISPERSION FOR GWR PER ALTERNATIVE FOR EACH REGRESSION

Note: Local Sample Size for GWR (Adaptive Kernel) was determined by AIC minimization for each iteration.

TABLE 1
DESCRIPTIVE STATISTICS OF ALTERNATIVE POINT ALLOCATION

<table>
<thead>
<tr>
<th>Model Results</th>
<th>Descriptive Statistics</th>
<th>Alternative 1</th>
<th>Alternative 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2 (coefficient of determination)</td>
<td>Mean</td>
<td>0.40373</td>
<td>0.40483</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.39813</td>
<td>0.39936</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.03268</td>
<td>0.03341</td>
</tr>
<tr>
<td>Bandwidth (in percentage of total observations)</td>
<td>Mean</td>
<td>0.10697</td>
<td>0.10792</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.10726</td>
<td>0.10727</td>
</tr>
<tr>
<td></td>
<td>Std. Dev.</td>
<td>0.02726</td>
<td>0.02845</td>
</tr>
</tbody>
</table>

FIGURE 5
SCATTER-PLOT AND CORRELATION OF BANDWIDTH AND COEFFICIENT OF DETERMINATION FOR EACH POINT DISTRIBUTION ALLOCATION


