Calculating economic indexes per household and censal section from official Spanish databases

Sonia Frutos[(1)], Ernestina Menasalvas[(1)], Cesar Montes[(1)], Javier Segovia[(1)]

[(1)]Facultad de Informática, Campus de Montegancedo, UPM. 28660 Boadilla del Monte, Madrid, Spain

Abstract.

In the competitive environments, in which all sorts of organisations move it is of utmost importance to have information about clients. Public databases offer information about households and families. However, the non-crossed and non-georeferenced format of these databases often makes it difficult to extract typologies and information.

There are only two public databases from which to get information at the household or family level in Spain: Population and Housing Censuses, which provide aggregated and georeferenced information, and the Family Expenditure Surveys, which provide information on household consumption, both published by the National Statistics Institute. The two databases cannot be directly cross-referenced, because the Family Expenditure Surveys offer a detailed description of the families, whereas the Census provides the same data but aggregated without cross-references.

In this paper, we define a procedure for cross-referencing these DBs and calculating the economic household indexes for Spanish censal sections that define the average quarterly economic behaviour of the households located in each censal section. The necessary *Symbolic Data Analysis* procedure is based on neural networks and provides an estimate of the trend in these indexes over a series of years. The procedure can be easily extrapolated to similar problems with official data sources from other countries.

1. Introduction

It is of utmost importance in the competitive environments in which all shorts of organisation operate to have geographical, social and economic information about their customers. The National Statistics Institute's public databases offer information about households and families generally (second-order objects or

macro data) from which behaviours can be extrapolated that can be transferred to real customers, but the format and content of these databases is limited by the data protection laws, which often makes it difficult to extract, for example, household typologies and information by geographic location.

In Spain there are only two public databases from which information at the household and family level can be obtained:

- The Population and Housing Censuses (PHC), conducted approximately every 10 years, which provides aggregated information at the censal section level. A **Censal Section** is the smallest administrative unit of information about which there is National Statistics Institute's censal information and is composed of about 400 households over a total of approximately 32000 sections into which the country is divided.
- The quarterly Family Expenditure Surveys (FES) with information on consumption by household of over 300 products.

Both databases are published by the National Statistics Institute (INE) and would yield some economic household indexes by Spanish censal sections of utmost importance today in customer relationship management (CRM) applications. There are many applications that can be obtained from these indexes, for example:

- Evaluate what censal sections in the country are more predisposed (higher index) to expenditure on a given consumer product, so that a company engaged in the sale of this product can easily locate its best points of sale.
- Use income or financial interest indexes for a bank to do a mailing to the households of the censal sections with highest income and likelihood to use financial services.
- Study the trend and periodicity of expenditure on a product to help to plan marketing strategies depending not only on the censal section but also the quarter and year in which it is launched.

The two above-mentioned databases cannot be directly crossed because whereas the Family Expenditure Surveys (FES) offer a detailed description of the families/households surveyed about their expenditure and income, explaining for each one of these the socio-economic condition of the principal earner, the household type, etc., the Census provides the same data but aggregated, without cross-referencing, indicating, for example, how many families there are in a censal section of a given socio-economic level and how many of a type of household, but not their cross-reference, that is, it does not indicate how many there are of a given socio-economic level and at the same time of a type of household. One possible shortcut for overcoming this problem would be to situate the families surveyed in the FES by censal section and then extrapolate, but the FESs do not indicate the source relative to the censal sections of each surveyed family, which means that it is impossible to directly locate the results of a FES in the censal sections.

In this paper, we provide a neural-network based procedure for estimating quarterly household economic indexes for Spanish Censal Sections, based on the

above-mentioned DBs, as well as a linear forecasting model for estimating the trend of each of the defined indexes.

The remainder of the article is organised as follows. Section 2 summarises other approaches taken in the same direction. Section 3 presents the definition of the household economic indexes to be calculated and section 4 details a procedure for estimating these values, as well as their trend over this set of years. Section 5 presents a practical application of the approach presented to validate the technique. Finally, section 6 presents the conclusions and future lines of work.

2. Related Work

The first standard micromarketing tool, called MOSAIC [1, 2, 3, 4] and launched by the British company Experian, appeared in Spain in the early 90s. Others, like the German Bertelsman's REGIO or Equifax's Microvision [7], soon followed. These were first-generation micromarketing tools, oriented at new customer conquest strategies.

These tools are based on having classified the Spanish population by life style typologies, using for this purpose statistical classification techniques. The classification criteria used are descriptive socio-demographic variables, which means that each group or typology is very similar with regard to the variables studied (e.g., age, socio-economic condition, educational level, …) and, at the same time, very different from the members of the other groups or typologies. By means of this typification process, each of the roughly 32,000 censal sections in Spain is associated with a given life style typology.

This type of tools are characterised by:

- They do not explain behaviours; they only describe the socio-demographic characteristics of each typology.
- They use statistical classification techniques (cluster analysis)
- Only one typology is associated with a censal section.

In the late 90s, Bertelsmann-Direct published its Family Expenditure Items Indexes within the Habits® [8] product. These indexes are different from other tools like Mosaic or Regio and even from the Life Style and Consumption indexes of Habits® itself, in that what they express is a distribution of family consumption in a series of products for each censal section, by means of indexes whose philosophy follows the idea to be explained in section 3. However, the statistical technique used to find the model that explains the household variable is Generalized Linear Models, which very much limits the predictive capability of the indexes, as the underlying models are clearly non-linear. Moreover, a measure or application of these techniques as a benchmark with which to compare the technique proposed in this paper, unlike what we do in section 5, is not publicly available.

3. Definition of Household Economic Indexes by Spanish Censal Sections

The household economic indexes by Spanish censal sections to be calculated define the average quarterly economic behaviour of the households located in each Spanish censal section. These indexes are expressed in absolute (euros or pesetas) and relative (percentage) terms with respect to the national, autonomous community and municipal mean.

These economic indexes are divided into three categories, expenditure, earnings and financial:

- **Expenditure indexes per quarter and household**: these are 259 indexes that indicate the expenditure on different budget items that a family living in a Spanish censal section would have and are divided into:
  1. Food, Drink and Tobacco (86 items)
  2. Clothing and Footwear (24 items)
  3. Housing, Heating and Lighting (26 items)
  4. Furniture, Furnishings, Fittings and Current Household Maintenance Costs (29 items)
  5. Medical Services and Health Expenditure (11 items)
  6. Transport and Communications (18 items)
  7. Leisure, Entertainment, Education and Culture (35 items)
  8. Other Goods and Services (20 items)
  9. Other Expenditure (10 items)

  The budget items cover the Household Expenditure totalling 259 products. Their definition coincides with the description in the Continuous Family Expenditure Survey (ECPF'92), Base = 1985, published by the Spanish National Statistics Institute, up to the variety level of (Cod.Ecpf).
- **Quarterly Income by Household** in each censal section. Specifically, the following 7 indexes are defined according to the definition of these income found in the ECPF'92:
  1. Income from employment: monetary and non-monetary
  2. Income from self-employment: monetary and non-monetary
  3. Income from capital and property: monetary and non-monetary
  4. Income from pensions
  5. Income from unemployment benefits
  6. Income from other regular transfers
  7. Other income: monetary and non-monetary
- **Quarterly Financial Indexes per Household** in each censal section of financial interest, the following are derived from the above:
  1. Income: the estimated average income from the sum of the 7 income concepts per quarter and household.
  2. Expenditure: the estimated average consumption of the sum of the 259 items per quarter and household.
  3. Debt: the average quarterly debt per household generated in the censal section.

4. Savings: the average quarterly saving per household generated in the censal section.

4. Data Analysis Procedure for Calculating Indexes and Trend

The procedure followed to estimate the average value of each of the household economic indexes by Spanish censal sections, as well as their trend, is not simple as it involves cross-referencing information from the two DBs, one with georeferenced information -the PHC- and the other with consumption information –the FES-, whose content and fields are different. On the one hand, the Population and Housing Census (PHC) offers an aggregated description of the socio-economic composition of a censal section with parameters like:

- Sex (Male, Female)
- Age (5 ranges)
- ...

But with the constraint that the data are not cross-referenced. For example, the PHC indicates how many households are headed by males, how many by females, how many by people aged from 25 to 30 years, etc., but does not indicate how many there are headed by males who are **also** aged from 25 to 30 years. Why would cross-referencing be useful? Because the FES offers information on consumption by household, also providing information on the socio-economic composition of the household surveyed. If we knew the composition of the households by censal section in the PHC, all we would have to do is to go to the FES and look for surveyed families with a similar composition, look at their expenditure and extrapolate to the families of a similar profile of the censal section. Unfortunately, this is not possible and we have to follow a procedure like the following:

1. Group families surveyed in the FES. The aim is to form groups with surveyed families who are known to live close to each other. These family groups may or may not belong to different censal sections. This is one of the critical points of the methodology. Fortunately, the FES usually include a field that indicates such closeness, either because it is so explicitly, or for other reasons such as indicating that they have been surveyed by the same interviewer on the same day, which leads to think that they are not very far away to prevent he interviewer wasting time travelling.
2. Calculate the socio-economic composition of each family group. The aim is to get the socio-economic description of the non-cross-referenced composition of each family group (percentage of families whose principal earner is male, percentage of families whose principal earner is female, percentage of 1-, 2-, 3-member families, etc., percentage of families whose principal earner has higher education,…) The content of this description must be obtained from the

information available in the survey used, must be as broad as possible and, above all, must be information also available in the PHC.

3. Get the average indexes per family group. For each family group, the average value of each index must be calculated.
   3.1. The income, expenditure, investment and property indexes are calculated by summing their components at the surveyed family level and then calculating the average per group.
   3.2. The saving index is calculated by first getting the savings at the level of surveyed family, where saving is equal to income minus expenditure. The goal of this index is to find out the average amount of saving of the group families, that is, the amount of money the families of the group in question may have available to purchase new goods. For this purpose, the index must be adjusting in its calculation using the value 0 (no saving) in families whose saving is negative. All the saving indexes at family level must be 0 or positive. After this adjustment, the average per group is calculated.
   3.3. The debt index is calculated by first getting the debt at the level of surveyed family, where debt is equal to income minus expenditure. This index is adjusted using the value 0 (no debt) for families whose debt is positive. All the debt indexes at family level must be 0 or negative. After this adjustment, the average is calculated for the group.

It is important to note that the calculated indexes are an average per family and quarter estimated for the period of years covered by the Family Expenditure Survey used, as the family groups used have been selected by geographic proximity irrespective of the time at which the survey was conducted.

4. Get estimation models for each index. The aim is to get estimation models that would take the non-cross-referenced socio-economic description of the composition of a family group as an input and would output an estimate of the value of the average index for all the group families obtained in the above point. These models will then apply to other family groups whose composition is in principle unknown, which means that it is important that these models satisfactorily generalise the solution. The description of the models used here is found in section 4.1 and 4.2.
5. Calculate the socio-economic composition of each censal section. For each censal section of the PHC, the same socio-economic description as calculated in step 2 has to be calculated. This is why it was specified then that the socio-economic description used should be available in both databases.
6. Get indexes per censal section applying the models. Apply the non-linear models calculated in point 4 to each censal section, taking the descriptors obtained in step 5 as input. We would get then the 270 expenditure, income and financial interest indexes for each censal section.
7. To get the indexes relative to nation, autonomous community and municipality, first we calculate the indexes for each censal section, which are averaged out for the nation, autonomous community and municipality, and then the ratio is established with respect to these measures.

8. Get the temporal evolution of the indexes. For each of the 272 indexes, their evolution must be calculated separately for a set of years that cover at least the years during which the Family Expenditure Survey used was conducted.

4.1 Method for estimating each index

Due to the required characteristics, the model to be used must be a **non-linear** technique, which is why a neural net was chosen.

The socio-economic description of the household composition of each group will be the neural net input x_i and the output s_j will be the values of the indexes estimated, by means of a non-linear transformation.

The chosen non-linear transformation is characterised by the use of the transformation function

$$f(a) = 2.5ae^{-a^2}$$

For this purpose, one model per index is created rather than a single model for all the indexes to be estimated. All the models are identical expect for a series of constants that are later adjusted. The non-linear transformation model suited for the procedure follows the following mathematical formulation: le xi be the inputs and I their number and let s be the output to be calculated. The output or index will be calculated using the following non-linear formula,

$$y_j = f(\sum_{i=1}^{I} w_{ji} x_i); \quad s = \sum_{j=1}^{20} \varpi_j y_j \quad where\ f(a) = 2.5ae^{-a^2}$$

The constants to be adjusted for the model to achieve a correct estimation of each index are wji and ϖj. These constants will be adjusted using the well-known back propagation method described in [9]. This method calls for a measure of square error that has to be calculated for all the family groups created. The error will be measured as the square root of the difference between the value of the real index and the value of the estimated index.

4.2 Method for estimating temporal evolution

In a first step, the surveys are grouped according to a time unit, years, quarter, month or web, and the average of the index for each unit of all the surveys conducted during these periods is calculated. This grouping is totally independent and different from the family groupings carried out in point 1, as whereas the latter were based on a criterion of geographic neighbourhood, these are based on a criterion of temporal and geographic proximity jointly, as the data obtained are not statistically significant.

The second step involves adjusting a time forecasting model for each index and applying it in the period of time to be studied. For this purpose, we propose the

following linear prediction model of the index s where Y is an integer that indicates the year, Q the quarter, M the month and W the week in question:

$$s(Y,Q,M,W) = \alpha Y + \beta Q + \chi M + \delta W + \varepsilon$$

The constants α, β, χ, δ, and ε are adjusted by techniques of linear regression techniques on the data obtained in the previous sep. Varying the values of year, quarter, month or week, we will be able to get predictions on the indexes in time periods outside what are included in the survey used.

5. An example of methodology validation: calculation of the income per household in the Autonomous Community of Madrid

The symbolic data analysis methodology has been tested on the 1991 Population and Housing Census (PHC) and the Continuous Family Expenditure Survey (CFES), Base 1985, published by the Spanish INE. The CFES covers 100% of the variables, shares with the PHC many descriptive variables of the households and includes a field with geographic information. The CFES is quarterly and covers the years from 1992 to 1996.

The composition of the household groups has been done using the following 5 variables

- Socio-economic category of the principal earner (7 levels)
- Educational level of the principal earner (7 levels)
- Sex (2 types)
- Age (5 levels)
- Municipality size (4 levels)

Totalling 25 descriptors per family group.

In the CFES, we managed to form 584 family groups that are known to live close geographically. The CFES surveys the same number of families several times during the quarters of the years 1992 to 1996, leading to approximately 106 different families. That is, the 584 groups correspond to grouping of these 106 families surveyed at different times.

Using the database of 584 groups of the CFES, with their 25 socio-economic descriptors and their 270 indicators of income and expenditure, as a training set, the neural models are calculated and applied to the PHC. Additionally, the temporal evolution of the indexes is obtained on the temporal basis of the quarter and calculating its prediction for the years 1997 to 2000.

The model has been extensively validated at private companies, but here we are going to present a validation with an official index. In particular, the Statistics Institute of the Autonomous Community of Madrid annually produces a Municipal

Household Income (MHI indicator). The Autonomous Community of Madrid includes about 1,400,000 households distributed in 148 municipalities.

This indicator is produced mainly on the basis of the Income Tax Returns, particularly, on the basis of the assessment basis of this tax, corrected by other indirect indicators that can be used to estimate this variable and the available family income. The aim is to estimate this important macromagnitude that measure the real expenditure capability of families through the family income obtained in net terms (that is, having removed taxes, deductions and withholdings). The important thing for our purpose is that it is an index produced by direct means and absolutely unrelated to the DBs used in our methodology.

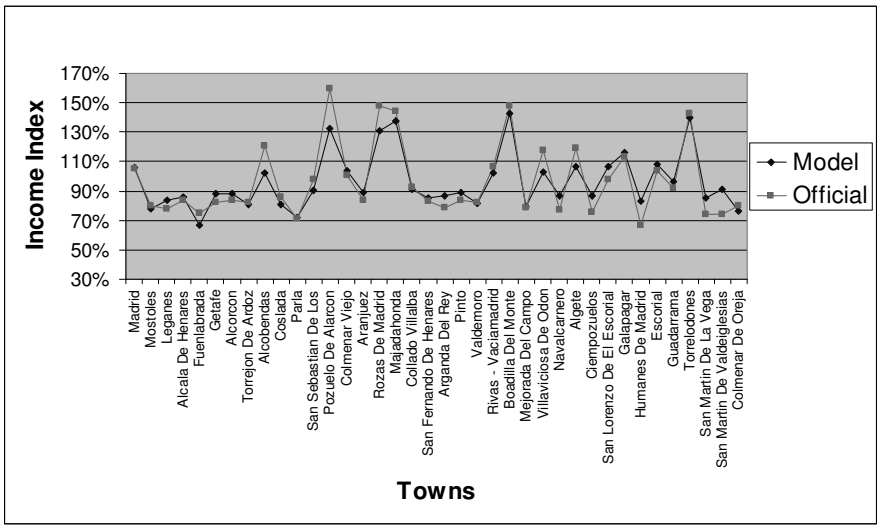

Figure 1. Comparison between the official income index for 1997 and the predicted by the model.

The MHI index used is the 1997 index and is provided at municipality level, which means that index calculated with our methodology must be aggregated to this level, the municipality level. Figure 1 shows a comparison of the indexes for the biggest municipalities in the Autonomous Community of Madrid. The index is a percentage, where 100% represents the average expenditure per household of the Autonomous Community of Madrid. The correlation level of the indexes is high, 0.92.

Deviations can be appreciated in the figure, which we put down to two problems:

- The PHC provides the distribution of the households in 1991, a distribution that has varied considerably in municipalities in expansion during the 1991-1997 period, like Alcobendas and Pozuelo de Alarcon.

- The income index of the model has been calculated by means of the CFES, averaged out for the 1992-1996 period.

It is to be assumed that the degree of adjustment would be greater with more updated databases.

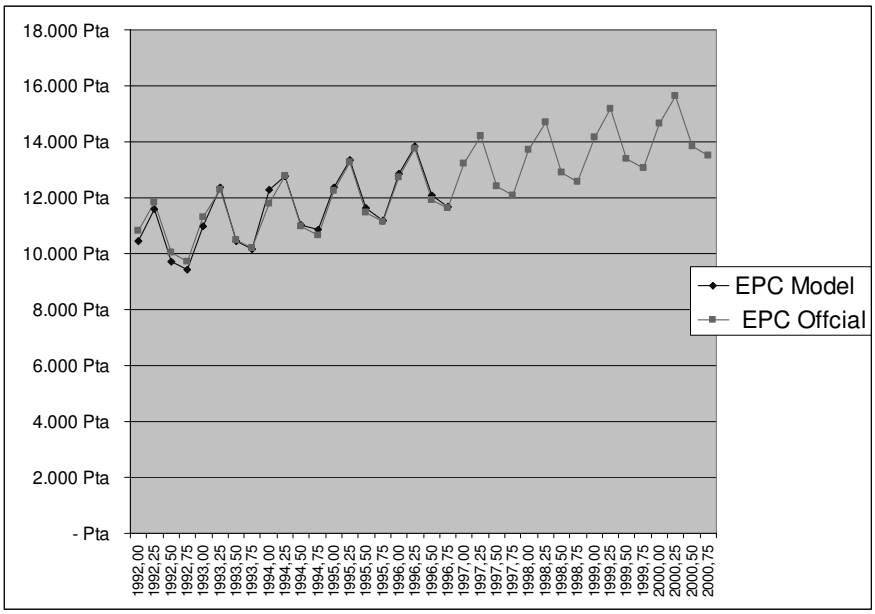

Figure 2. Averaged Electric Power Consumption per home (EPC), in pesetas. Comparison between official and model data. Model includes prediction for period 1997-2000

On the other hand, it is impossible to validate the estimate of the evolution of the indexes, but some results are encouraging. Figure 2 shows the comparison between the estimated expenditure and real consumption of electric power in principal residence per household (EPC). The calculation was done as a means value per family. The official EPC, calculated on families surveyed in the CFES includes the period of 1992 to 1996, whereas the model estimates in the range 1992 to 2000. It is clear how well the model grasps the trend and seasoning features of the index.

6. Conclusions

In this paper, we have presented the definition of the household indexes by Spanish censal sections characterised because they define the average quarterly behaviour of households located in each of the Spanish censal sections, and are expressed in absolute terms (euros or pesetas) and relative terms (percentage) with respect to the national, autonomous community or municipal average.
Additionally, we have established a procedure for estimating the average value of each of the economic indexes of the household by censal section within a set of years preventing the problems of having to work with aggregated data based on official statistics.
The results are promising and easily transferable to DBs in other countries.

References

[1] MOSAIC http://www.micromarketing-online.com/play.htm

[2] Gabbot, M.and Sutherland, E. (1993). 'Marketing information systems in universities', Marketing Intelligence & Planning, Vol.11 (7).

[3] Halsley, A. (1992), 'Opening Wide the Doors of Higher Education', NCE briefing (6), National Commission on Education, London.

[4] Mitchell, V. W., McGoldrick, P., (1993) 'The Role of Geodemographics in segmenting and targeting consumer markets: A Delphi study' European Journal of marketing, Vol 28 No. 5, 1994, pp 54-72, MCB University Press

[5] Sleight, P., (1997), 'Targeting customers, Second edition – How to use Geodemographic and Lifestyle data in your Business', NTC Publications, Oxford.

[6] Tonks, D. and Farr, M. W. (1995) 'Market Segments for higher education', Market Intelligence and planning, Volume 13 (4).

[7] MICROVISION http://www.asnefequifax.es/micromkt.htm

[8] HABITS http://www.bertelsmann-direct.com/Webbertelsmann/

[9] D. Rumelhart, G. Hinton y R. Williams, "Learning representations by back-propagating errors". Nature, 323 pp. 533-536, 1986