

# Geostatistical Study for Fraud and Energy Losses

Eduardo de Rezende Francisco, Business Analyst, AES Eletropaulo – [eduardo.francisco@aes.com](mailto:eduardo.francisco@aes.com)

Eduardo Bortotti Fagundes, Electric Technician, AES Eletropaulo – [eduardo.bortotti@aes.com](mailto:eduardo.bortotti@aes.com)

## ABSTRACT

One of the biggest challenges faced by AES Eletropaulo and other big Electric Utilities is the understanding and reduction of loss rates, particularly in regard to commercial losses. To face this challenge, a lot of effort has been made aiming at understanding the generating factors of energy losses, such as fraud and illegal connections, which includes household income, cultural aspects and other social-economical variables – all of them geographically located and mutually influenced. In order to handle it, this study uses geographic analysis to create a energy losses indicator. We used ArcGIS Spatial Analyst and Geostatistical Analyst, the Spatial Dependence extension of R statistical package and a MapObjects geostatistical application called GeoDA. This work aims at establishing a geostatistical methodology for the identification of areas with greater potential of clients with energy losses (reducing operational costs), which also confirms the statistical relationship between social economical variables and energy losses.

**Index Terms:** Bad debts, Census Tracts, Consumption of Electric Energy, Cut-off Customers, Energy Losses, Geostatistics, Household Income, Power Losses, Spatial Analysis, Trend Analysis.

## I. INTRODUCTION

Every year, AES Eletropaulo, the biggest Brazilian electricity distribution company, does not bill 2,700 GWh derived from non-technical losses, an amount of energy that is greater than the residential consumption of the cities of São Bernardo, São Caetano, Santo André and Diadema which, altogether, have a population of 2 million people. According to ANEEL, the Brazilian Energy Regulator Office, the annual commercial loss in the Brazilian power distribution segment was about R\$ 5.1 billion in 2005.

Due to its relevance, loss management is a strategic area for power distribution companies and requires a joint effort for its understanding, which origins and motivations can be classified as frauds, measurement anomalies, failures in internal billing processes, irregular connections, cut-off clients directly connected to the grid or techniques, as well as for the definition of action strategies that are effective to mitigate their impact on the financial results of the company.

Spatial modeling and geostatistical analysis are key elements in planning the development, management and operation of power losses prevention. The effectiveness, economy and reliability depend on the correct choice of dimensions (categories of variables) used in its manipulation.

Previous studies [4] highlighted consumption of electric energy as a substitute of household income, in a set of census tract coverage area (about 20,000 ones in AES Eletropaulo's concession area). On the other hand, power losses could be exactly measured through the difference between bought (income) and sold (outcome – sum of consumption in the customers) energy. This information exists for each

Transforming Distribution Station (ETD) and its coverage area.

Handling diverse levels of geographical polygonal areas is a challenge that can be handled using geographical analysis. Evaluate and quantify geographical influence of this phenomena needs geostatistical analysis. This work uses these tools to support the understanding of power losses primary motivations.

## II. METHODOLOGY APPLIED

Understanding the spatial distribution of data derived from phenomena occurring over a certain geographical region constitutes a huge challenge for the clarification of core issues in several areas of public and corporate administration, such as health, urban dynamics, environment, marketing and asset management. Such studies have become more and more common, due to the availability of Geographical Information Systems (GIS), and also the need for clarification of the spatial distribution of common problems and social-economical interaction variables that the traditional and classical models usually do not address. Geostatistics is a field of regional science that approaches the theme of dependence and spatial heterogeneity, critical aspects in any study on regional economy. These characteristics can invalidate the use of conventional econometric techniques. Spatial quantitative methods are a special case of overall statistical focus and, therefore, require a set of different methods and techniques that those utilized in conventional statistics. Companies that have geographical information as essential to their business benefit dramatically from the application of those techniques.

Income is a factor traditionally adopted in studies on life conditions and poverty, since it is through income that people have access to goods and services required for their survival. We understand income as the total sum of monthly revenue from work plus any income derived from other sources [2]. However, the difficulty in obtaining accurate information about this variable, frequently changed by understatement, overstatement, forgetfulness, seasonableness of the income source and refusal, makes the use of this indicator in market research very difficult.

As a consequence, Research Institutes choose to capture Economic Class or Consumption Power of individuals through indicators based on the possession of durable goods and on the educational level of the head of the family. Such indicators may be used with a certain accuracy in place of income. Besides being an indirect measure of family income, the quantity of goods a family has indicates the level of comfort they reached along time.

The most recent indicator of this type is the Brazil Economic Classification Criterium (CCEB), or simply, Brazil Criterion, defined in 1996 by the National Association of

Research Companies (ANEP). This indicator is based on a scoring scale that may range from 0 (zero) to 34 (thirty-four) points, segmented into 7 economic classes [1].

However, the criterion presents important regional differences and is not adequate for the characterization of families positioned in the extremes of income distribution. The criterion is useful to segment large masses and is adequate for nationwide studies. For regions or specific segments, deeper studies require specialization or customization of the Brazil Criterion regarding the initial characterization of the population under analysis, often involving variables that best characterize the aptitude for consumption. The use of consumption indicators that have an overall scope and application can be very useful in this process of characterizing consumers.

Among the indicators of this type is the one regarding power consumption. Nationwide, power distribution reaches 97.0% of Brazilian households, and that index reaches 99.6% in the urban area, with 99.4% in the Southeastern region and 99.9% in the city of São Paulo [3]. It has more capilarity and coverage than services from other utility companies, such as fixed and mobile telephone, water and gas. And that is so because, basically, everybody has access to power, mainly when considering irregular connections and frauds, but not everyone has piped water or even telephone. Besides that, databases from power distribution companies have information on consumption of each of their clients, from higher to lower classes.

Due to the fact that it is an essential, encompassing and relatively democratic service compared to other utilities, its register and commercial information may offer subsidy for a comparative knowledge on social-economical and demographic characteristics of the families under study. Besides, deeper studies in certain classes such as, for example, the poorest classes, require mechanisms that enable detailing of specific ranges and a better understanding of the characteristic of classification sub-levels.

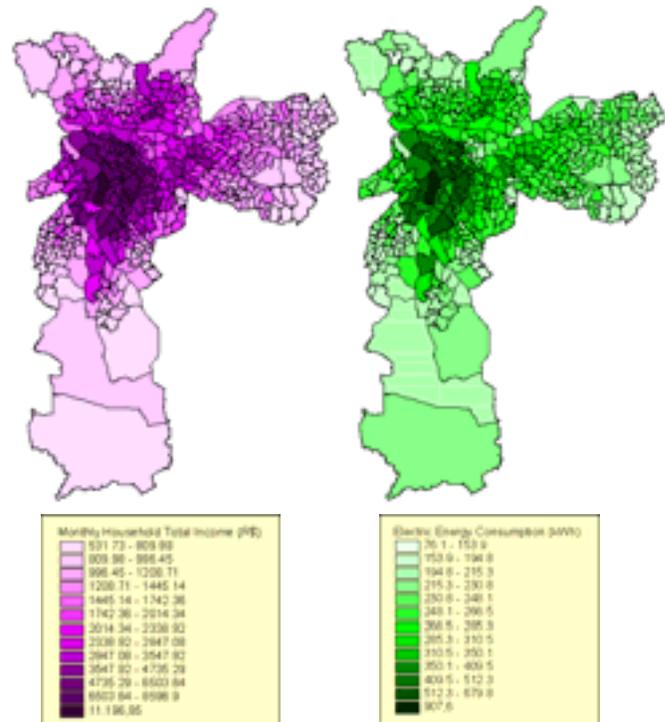
Together with the information on location, historical data and seasonableness, energy indicators can help to infer a better social-economical classification from consumption ranges, and thus contemplate a better definition of low-income consumer in regions of lower access and higher difficulty for gathering information in consumer goods.

For retail companies, the incorporation of power consumption may improve the identification of potential consumer families for specific offers. Consumers within the same economic classification according to the Brazil Criterion may be differentiated though the power consumption of their households. Since power consumption, to a certain extent, reflects the possession and use of electric durable goods, associated to the total number of rooms of the house and the number of people who live in it, the companies may improve their return rate in promotional materials, direct mail and associated costs, through a better segmentation of its market, the increase of reactivity of consumers due to their higher purchasing power, as well as the better identification of their target market.

Moreover, for power distribution companies, the application of a social-economical classification model that is partially or totally based on power consumption enables

higher efficacy in studies on market identification, segmentation and forecast, fraud and bad debt analysis and the elaboration of customer relations strategies, among other benefits.

The relationship between Income and Electricity Consumption are very strong, with coefficients of determination R-squared ranging from 0.912 to 0.960 . This result (illustrated by the similarities between two maps in Figure 1) supports the use of mean household electricity consumption, at a territorial aggregated level, as an excellent regional indicator of income concentration in the city of São Paulo, extend to whole AES Eletropaulo's concession area.



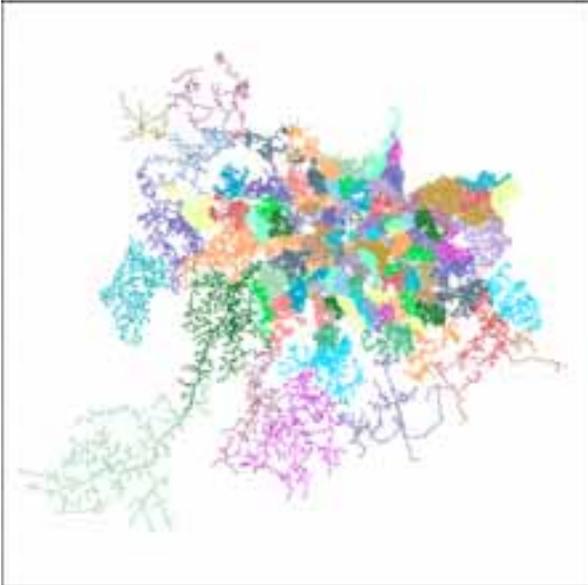
**Fig. 1.** Maps of the City of São Paulo representing: (i) Monthly Household Total Income, and (ii) Electric Energy Consumption per Weighted Areas (set of Census Tracts)

Source: Francisco, 2006, using ArcGIS ArcMap 8.3

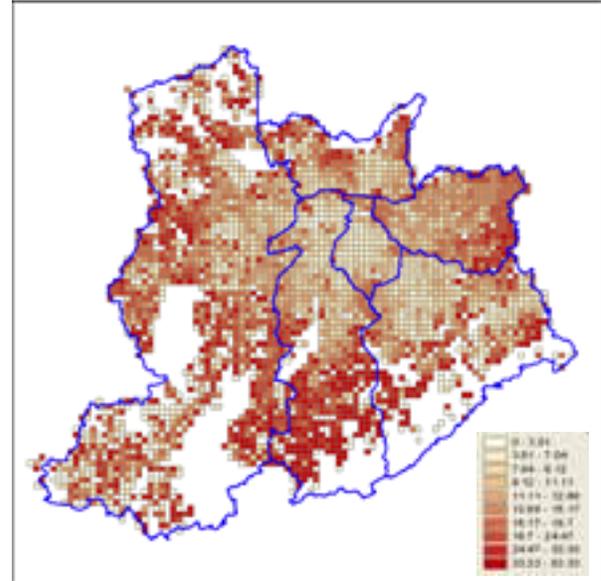
This work presents the development of a methodology for the identification of areas with higher potential of incidence of clients with power loss, which also confirms the statistical relationship between social-economical variables and power losses. Such methodology utilizes geostatistical techniques and integrates, for each geographical region under study, four basic pieces of information: (I) percentage of power loss (total), (II) average household income, (III) percentage of the area occupied by slums, and (IV) percentage of cut-off clients.

The methodology consisted, basically, of the following steps:

- (i) Mapping of the primary grid of all AES Eletropaulo circuits and association of each grid segment to the Transforming Distribution Station (ETD) that feeds it (Figure 2 shows the primary grid – each color represents the different ETDs);

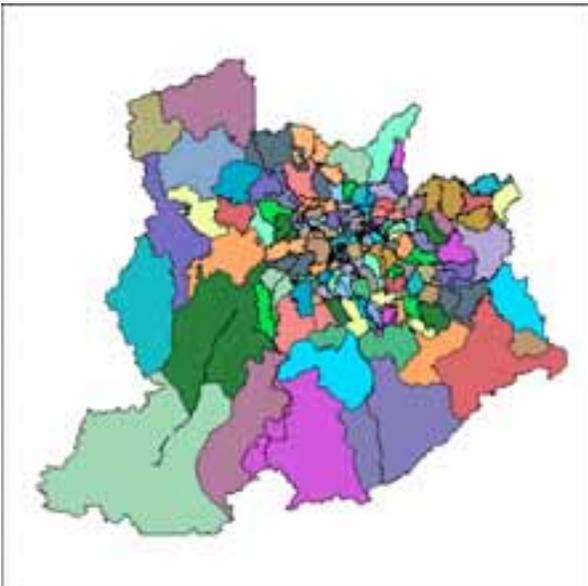


**Fig.2.** Mapping of the Primary grid  
Source: AES Eletropaulo GIS technical management system.



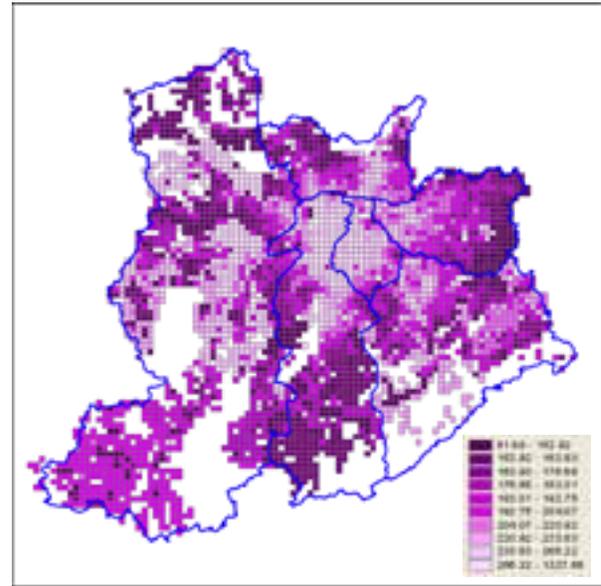
**Fig. 4.** Concentration of Cut-off Clients  
Source: AES Eletropaulo Commercial management system (SICON-B).

- (ii) Establishment of the coverage area of each ETD, through the geostatistical technique of Thiessen polygons, from the proximity of each geographical site to the grid segments (Figure 3 uses the same legend for each ETD's coverage area);



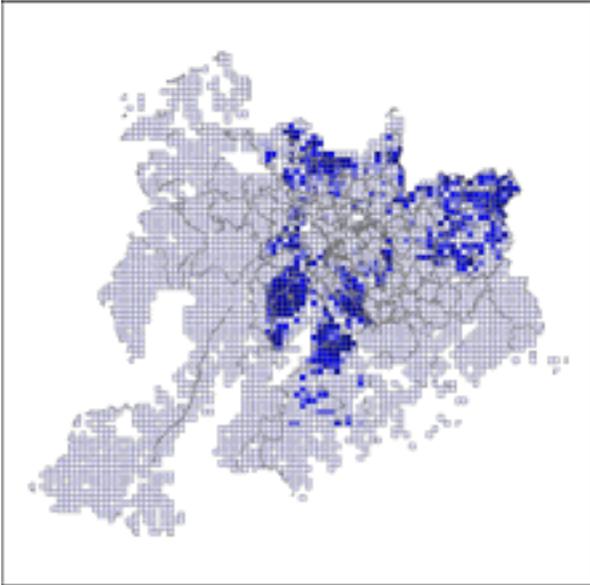
**Fig. 3.** Determination of the Coverage area of each ETD  
Source: The authors, using ArcGIS Spatial Analyst 8.3.

- (iv) Mapping of the Concentration of Power Residential Consumption, as *proxy* of Household Income and association to the same area units of the density of cut-off clients (square mesh of one kilometer each side) – Figure 5 uses dark colors for low income and clear colors for high income;



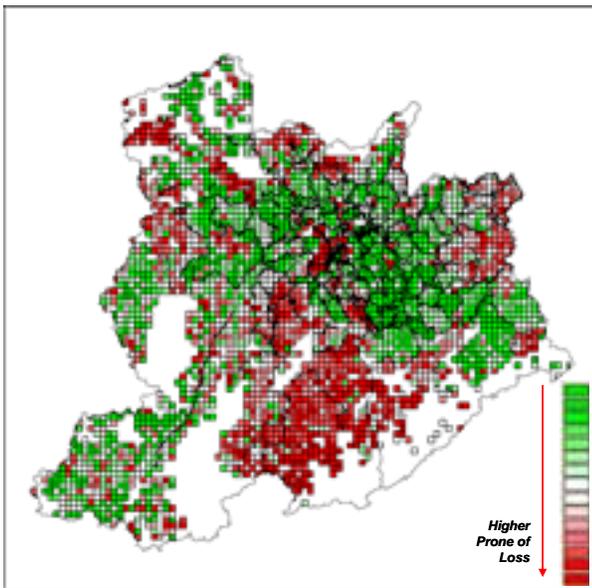
**Fig. 5.** Household Income (Electricity Consumption) Concentration  
Source: The authors, based on information on AES Eletropaulo Commercial and Technical management systems and on Francisco (2006)'s previous study.

- (iii) Mapping of the Geographic Density of Cut-off Customers (we used a square mesh of one kilometer each side) – Figure 4 shows the percentage of cut-off customers;
- (v) Identification of Slum areas and concentrations of subnormal gatherings for the determination of the percentage of slum areas of each unit of the density area of cut-off clients – Figure 6 shows these areas;

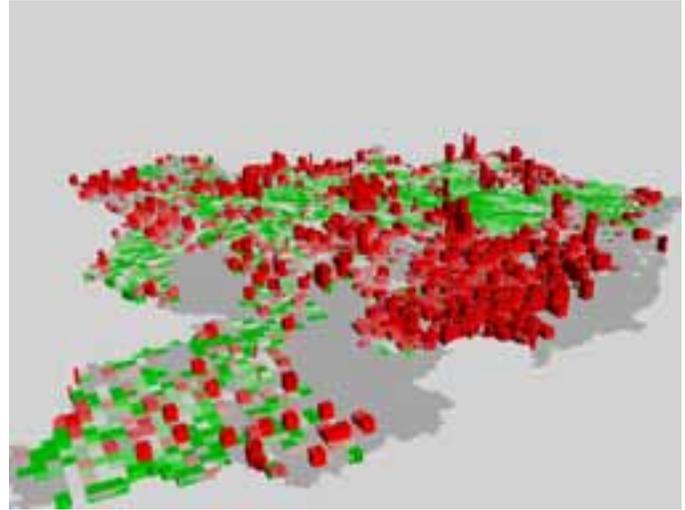


**Fig. 6.** Concentrations of subnormal gatherings  
Source: Municipality of São Paulo

- (vi) Spatial Overlay between geographical layers obtained in steps (ii), (iii), (iv) and (v) – using this technique we get an spatial layer in which each polygon contains information of the four dimensions (respectively, the four steps mentioned above) used in the study of power losses phenomena;
- (vii) Ponderation among the four information dimensions associated to each area unit, for the generation of the Commercial Loss Potential Indicator.



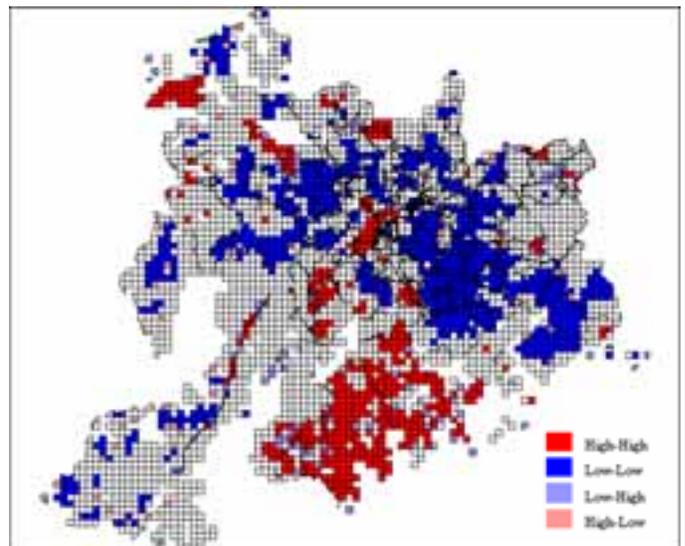
**Fig. 7.** Ponderation among information dimensions  
Source: The authors.



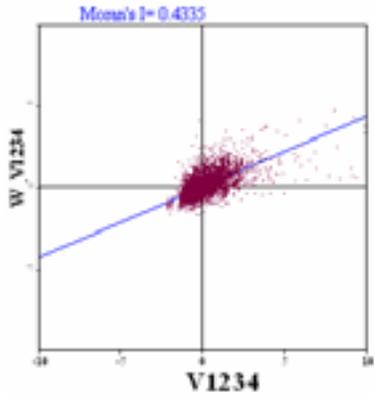
**Fig. 8.** 3D Visualization of Commercial Loss Potential Indicator  
Source: The authors, using ArcGIS 8.3 Spatial Analyst and 3D Analyst.

Figures 7 and 8 summarize the initial purpose of the project, and the integration between the domains utilized in the project for the investigation of the prospection of commercial losses: (i) losses per ETD coverage, (ii) income density (utilizing power consumption as *proxy*), (iii) proximity and concentration of slum areas, and (iv) percentage of cut-off clients.

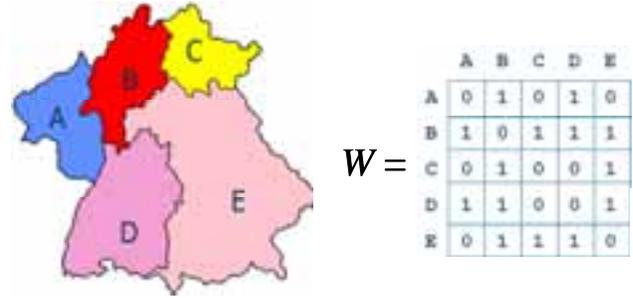
Based on Commercial Loss Potential Indicator (CLPI), multivariate regressions were performed, using step (2) methodology layer (total loss associated with each primary grid ETD) as a dependent variable, and the other three dimensions as independent ones. As one of the most important results of this study, the use of spatial regression method increases in 190% the power of comprehension of this phenomena (coefficient of determination R-squared moved from 11% to 32%). Figures 9 and 10 show LISA map (Local Indicator Spatial Association) and Moran's Spatial Auto-correlation index (a measure of spatial dependence). 43% of the Loss prone is correlated to its neighborhood loss.



**Fig. 9.** LISA map of Primary grid ETD coverage area's loss  
Source: The authors, using GeoDA 0.9.5-i5.



**Fig. 10.** LISA map of Primary grid ETD coverage area's loss  
Source: The authors, using GeoDA, using GeoDA 0.9.5-i5.



**Fig. 12.** Neighborhood Matrix  
Source: The authors.

Figure 11 shows the traditional and the spatial regression models.

**Traditional Multivariate Regression**

$$Loss = \beta_0 + \beta_1 Income + \beta_2 CutoffCustomers + \beta_3 SlumAreas$$

**REGRESSION**  
**SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION**

Data set	: fase1		
Dependent Variable	: Perda Percentual	Number of Observations:	7651
Mean dependent var	: 1.21553e-008	Number of Variables	: 4
S.D. dependent var	: 0.999935	Degrees of Freedom	: 7647
R-squared	: <b>0.111343</b>	F-statistic	: 319.373
Adjusted R-squared	: 0.110994	Prob(F-statistic)	: 0
Sum squared residual	: 6798.23	Log likelihood	: -10404.2
Sigma-square	: 0.889006	Akaike info criterion	: 20816.4
S.E. of regression	: 0.942871	Schwarz criterion	: 20844.2
Sigma-square ML	: 0.888541		
S.E of regression ML	: 0.942625		

**Spatial Multivariate Regression**

$$Loss = \beta_0 + \beta_1 Income + \beta_2 CutoffCustomers + \beta_3 SlumAreas + \beta_5 W Loss$$

**REGRESSION**  
**SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION**

Data set	: fase1		
Spatial Weight	: fase1.GAL		
Dependent Variable	: Perda Percentual	Number of Observations:	7651
Mean dependent var	: 1.21553e-008	Number of Variables	: 5
S.D. dependent var	: 0.999935	Degrees of Freedom	: 7646
Lag coeff. (Rho)	: 0.545362		
R-squared	: <b>0.320782</b>	Log likelihood	: -9571.13
Sq. Correlation	: -	Akaike info criterion	: 19152.3
Sigma-square	: 0.67913	Schwarz criterion	: 19187
S.E of regression	: 0.824093		

**Fig. 11.** Results of Regression Models  
Source: The authors, using GeoDA 0.9.5-i5, R 2.2.1 and SPDEP extension.

The *W* term above means the Neighborhood matrix, as shown in the example of Figure 12.

The purpose is to systematize this methodology so as to incorporate it to the prioritization management model. Geostatistics is a tool of great potential for the accomplishment of this objective.

Basically, the GIS part of this project consisted on ArcGIS ArcMap, Spatial Analyst and 3D Analyst extensions, used to create ETD coverage (based on Thiessen's polygons technique), establish spatial join action and show Commercial Losses Potential Indicator. Additionally, in the Geostatistical part we've used a MapObjects application called GeoDA 0.9.5-i5 (GeoData Analysis Software) and Spatial Dependence extension of statistical package R 2.2.1, for spatial multivariate regressions and LISA maps. All data handled is in shapefile format. Production version of AES Eletropaulo's electric geodatabase is stored in Oracle ArcSDE.

### III. CONCLUSIONS

Developed as a pilot project within the area corresponding to the Eastern Regional Division of AES Eletropaulo, and already expanded to the whole concession area of the company, this study is in its preliminary phase of supporting the definition of strategies for each region. Such support takes place vis-à-vis at each source of loss detected, such as: combat to frauds only at connected clients and in certain geographical areas, regularization of illegal nuclei, recovery of cut-off clients through commercial strategies, and it may help in the better orientation of future installations of anti-fraud grids and specific equipment, among other strategies.

Since the information utilized is internal and under AES Eletropaulo management, this methodology can be easily implemented and maintained. The innovation is in the integration of technical information, such as total loss per transforming distribution station, and commercial information, such as average billed power consumption and the status (cut-off versus non-cut-off) of each consumer unit. Such integration is enabled by the GIS, and the generation of the indicator is made by the determination of geostatistical techniques, still of very limited application, and thus, pioneer in the power distribution industry.

Another important aspect to be mentioned in this work is the identification of the representative correlation among the variables analyzed: Income, Bad Debt, Slum Concentration and Losses.

The present study is being enhanced in the sense of identifying and understanding the root cause of energy losses, and the next steps will be the evaluation of the relationship of the variables seen in this study and the client satisfaction regarding the power distribution company.

Disseminating GIS and Geostatistical potential uses is a additional objective of this project. GIS for technical and operational purposes is in a stage of excellence in AES Eletropaulo, but for commercial and customer relationship purposes it is in an embryonic stage and has to be improved.

### REFERENCES

#### Technical Reports:

- [1] ABEP – Associação Brasileira de Empresas de Pesquisa (2004, dezembro 22). *Critério de Classificação Econômica Brasil*. São Paulo, Brazil. Available in <http://www.abep.org.br>.
- [2] IBGE – Instituto Brasileiro de Geografia e Estatística (2003). *Censo Demográfico 2000*. Documentação dos Microdados da Amostra. Rio de Janeiro, Brazil: Author.
- [3] IBGE – Instituto Brasileiro de Geografia e Estatística (2005). *Pesquisa Nacional por Amostra de Domicílios 2004*. Rio de Janeiro, Brazil: Author.

#### Dissertations and Thesis:

- [4] FRANCISCO, E.R., *Relação entre o Consumo de Energia Elétrica, a Renda e a Caracterização Econômica de Famílias de Baixa Renda do Município de São Paulo*. Master Thesis, Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, São Paulo, Brazil, 2006.

### BIOGRAPHIES



**Eduardo de Rezende Francisco** is a PhD student and has a master degree in Business Administration by the FGV-EAESP and is graduated in Computer Sciences by IME-USP. He is a senior business analyst at AES Eletropaulo, working with Marketing Strategies and Market Research Planning; he is columnist of INFOGEO magazine and president of GITA Brasil (Brazilian Geospatial Information and Technology Association).  
**Email:** [eduardo.francisco@aes.com](mailto:eduardo.francisco@aes.com)  
**Phone:** 55 11 2195-1537



**Eduardo Bortotti Fagundes** is an Electric Technician and has been working with prevention losses since 2002. He works in the identification of the relationship between Losses and other external variables, identification of the profile of consumers with measurement irregularities through *Data Mining* systems, among other studies for the understanding and reduction of power loss levels.  
**Email:** [eduardo.bortotti@aes.com](mailto:eduardo.bortotti@aes.com)  
**Phone:** 55 11 2195-2771

### AUTHORS INFORMATION

**Name:** Eduardo de Rezende Francisco  
**Title:** Senior Business Analyst  
**Organization:** AES Eletropaulo  
**Address:** R. Tabatinguera, 164 – Sé – São Paulo – Brazil  
**Phone:** 55 11 2195 1537  
**Fax:** 55 11 2195 1098

**Name:** Eduardo Bortotti Fagundes  
**Title:** Electric Technician  
**Organization:** AES Eletropaulo  
**Address:** R. 25 de Janeiro, 320 – Luz – São Paulo – Brazil  
**Phone:** 55 11 2195 2771  
**Fax:** 55 11 2195 2687