Effect of cadastral update on land taxation in Colombia

36-707 Fall 2001, Project 2

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
e-mail *@stat.cmu.edu*

Abstract

The effect of cadastral update on land taxation participation in the total income, of 893 Colombian municipalities for the year 1998 was studied using multiple regression analysis. This study took into account if the cadastral update has been done in the urban or rural zones or both after the year 1994.

Land taxation participation was related to cadastral update after controlling by variables of social and industrial investment: Development Index, and taxation income: Total transfers from the central Government, Property value and Tributary income per capita, with a coefficient of correlation of 79%.

The findings in this study confirm the theorical importance of the updating process in order to build a more descentralized government with more fortified municipal treasuries,and capable of more social investment. In this manner, the criticisms from fiscal enteties of the National Goverment to the cadastral update process have been answered showing that this process has gotten its goals until now.

1 Introduction

At the beginning of the Eighties with the new strategies on decentralization presented by the National Government, Law 14 of 1983 or "Law of fortification of the municipal state treasuries" arose, placing as its main target the reinforcement of the municipal finances by means of instruments that allowed the administrations to improve their fiscal capacity with the purpose of catching resources and guiding them to the social investment. The basic idea for the fortification of the finances of each municipalities was to decentralize the principal sources of taxes, that is, land, industrial and trading taxation.

With Law 14, dispositions were established with the intention of forming, updating and conserving physical, legal, economic and fiscal cadastral information. For the time of its establishment, this information was not updated at all and it was not an optimal tool for collecting economical resources in the municipalities. This law also establised cadastral update periods of five years, and the authority was conferred to the Municipal Councils to establish the tariff regime according to the ranks established by the Law.The federal institution in charge of updating the information is the Agustin Codazzi Geographic Institute (*IGAC*) which a piece of the federal budget to do this task.

Recently, some fiscal entities start to ask themselves if this piece of federal budget has been used in the right way, that is, if the money that the goverment has spent in cadatral updating is recovered throught taxes, and the goverment had had to transfer less money to the municipalities. A manner to find out this, is studying if there is a relationship between cadastral update and land taxation participation over total income of the municipality. Hence, the main objective of the present study is to analyze, using multiple regression analysis, a possible effect of the cadastral update on land taxation participation in the Colombian municipalities for the year 1998[1] in order to evaluate actual budget policies related with the updating process.

This study is organized as follow: section 2 gives a brief description of the data collection, section 3 describes the methods applied in study as well as the results, in section 4 a discussion is presented. Finally, section 5 presents the final conclusions of this study as well as some recommendation.

2 Description of data

The data for 893[2] municipalities was collected by the following government agencies: Territorial Development Section[3] (SDTD), Social Mision (MS), IGAC, Republic Bank (BR) [4], and the National Administrative Department of Statistics (*DANE*). Each of this entities collect or maintain certain information, for instance, the *DANE* provides population projection based on the last population census[5] and number of lands. It is worth mentioning that some of these entities share some their information and most of the times they make some adjustment based on particular criteria, for example, SDTD may change the projections of population if they consider that the path the DANE followed to construct these projections lacks of some factor (i.e, violence). This situation makes quite difficult verify the information and most of the times difficult to use other's information in order to make imputations of missing data we may have. Table 1. specifies the variables in our dataset and the principal source[6] for the construction of them.

[1] There are projections of population and some taxation information for years 1999 and 2000, but 1998 is the last year with cadastral information and land taxation for each municipality

[2] Look at section 6

[3] Part of the National Planning Department (*DNP*)

[4] This is the bank of the Colombian government and one of its functions is to organize all the government finances

[5] The last census was in 1993

[6] We refer to the principal source because some of the data may come from another source

Table 1. Variables included in the analysis.

Variable	Description	Institution
update	**Cadastral Update** : It refers to the year in which the update was done. It is one, if the rural and urban zones of the municipality have not been updated or it was done before 1994. It is two if only the urban zone has been updated. It is three if only the rural zone been updated. It is four if all the municipality has been updated.	IGAC
nlan	**Number of lands**: real states that belong to a natural or juridical person, or a community, located in only one municipality and not separated by other public or private land.	IGAC
pv98	**Property value** calculated using the real state market.	BR
tinc	**Total Income**: total sum of current and capital taxes.	BR
ltax	**Land taxation**: municipality tribute over all the landed properties in rural and urban areas.	BR
trib	**Tributary income**: municipality tribute over alcohol beverages, cigarretes, lottery, gasoline, and others.	BR
tran	**Total Transfers**: budget execution related to resources transferred by the National Government to the municipality.	BR
popu	Projections of **population** based on census 93.	DANE
devi	**Development index**: index that measures the welfare and development for each municipality. This variable takes values between 1 and 8. The greater the index, the more developed the municipality.	SDTD
ICV	**Quality of life Index**: Index that measures the welfare and development for each municipality in terms of education, human capital, housing quality, quality and accessibility of public utilities, size and composition of the families. This variable varies between 0 and 100. The greater the ICV, the better the quality of life.	DNP

This dataset also contains the name of the municipality, the name and the number that identify the state which each municipality belongs to.

Based on this ”original” data set the following variables were computed:

Name	Description	Formula
Tparltot	land taxation participation in the total income, log-scaled Here we are interested in the participation of land taxation as measure of the level of decentralization of the municipality and success of the cadastral update.	$log(ltax/tinc)$
Tparltrib	Tributary income per capita in log-scale	$log(trib/popu)$
Tnlpopu	Number of lands per capita in log-scale Since a municipality with more lands will pay more land taxation, the population is used to standarize this variable	$log(nland/popu)$
Tpv98	Property value in log-scale	$log(pv98)$
Ttran	Total transfers in log-scale	$log(tran)$

As can be seen in this description, all these variables have been transformed in logarithmic scale. This was necesary since all of them are highly skewed as displayed in figure 1. More specificly, the logarithm was used because it is a most common transformation for this income data.

Note that our principal objective in this study is to identify if there is a relationship between *land taxation participation* and *cadastral update*, therefore an analysis of variance may see enough, but in this situation we need to study the effect of some covariates before finding the relationship we are interested on. To see this need, we can use a graphical device like boxplot, as is shown in figure 2. It is clear that there is not difference between the medians of the four categories of the variable *cadastral update.*

Finally, it is important to mention that in the present study the three biggest cities of the country: Bogota, Medellin and Cali are not include because they have a different cadastral institution in charge of all the information and management of their own cadastral update. Therefore, they are no relevant in the present study.

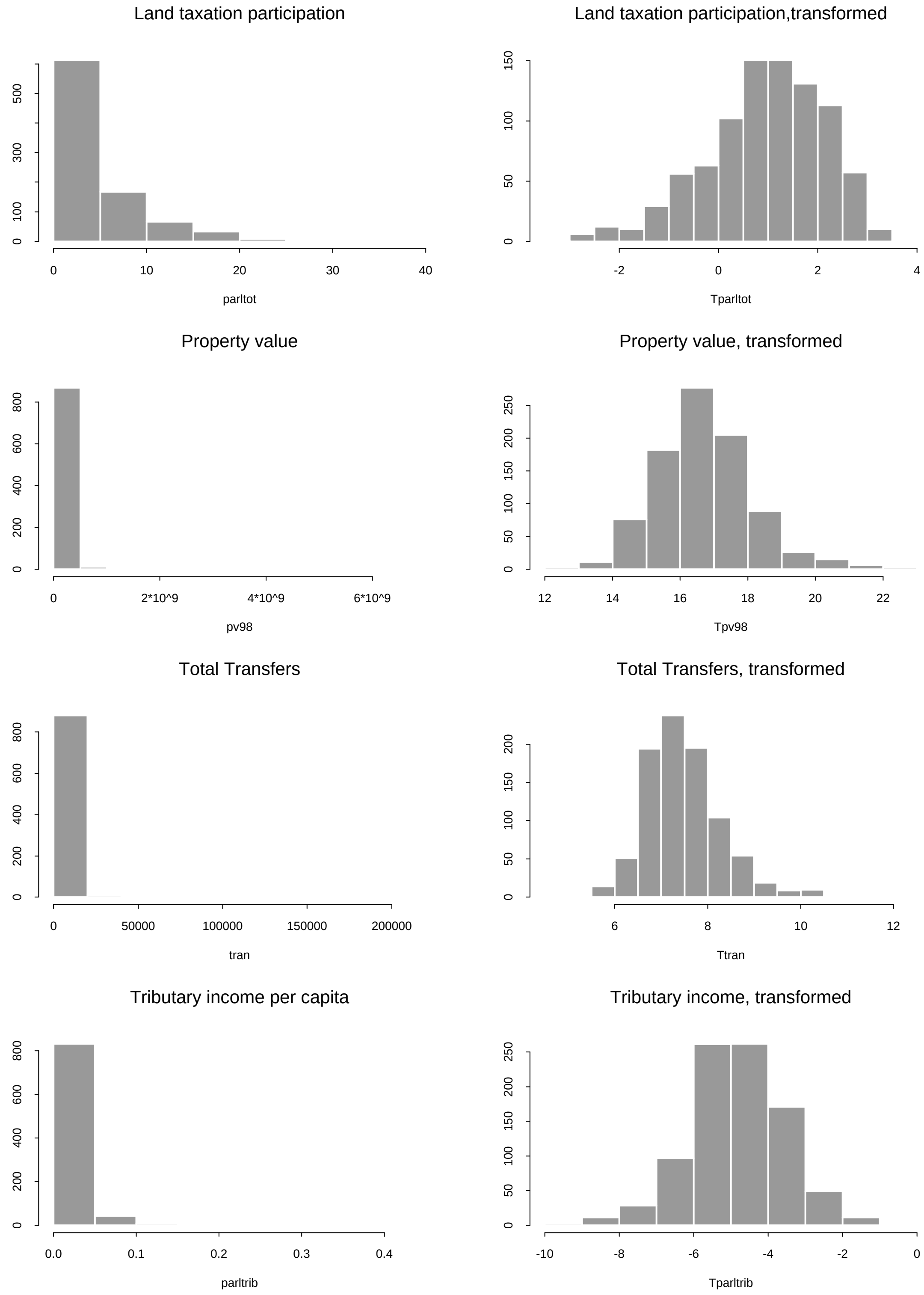

Figure 1: Histogram for *Land taxation participation*, *Total transfers*, *Tributary income per capita*, and *Property value* in their original scale and log-transformed

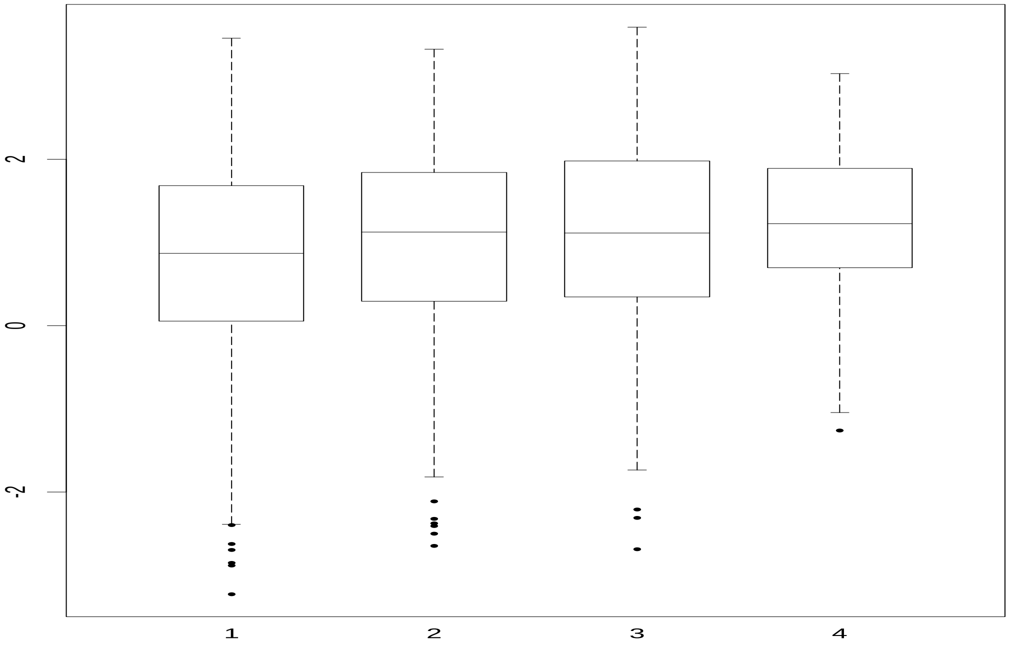

Figure 2: One-Way ANOVA using boxplot

3 Methods and Results

The statistical package Splus was used for EDA and multiple regression analysis. All the statistical tests were performed with a level of significance of $\alpha = 0.05$ unless otherwise stated.

An exploratory data analysis was carried out for all the variables registered in the study in order to detect some characteristics of the variables or subjects that later helped to address the study of the relations we are interested of.

From this analysis we concluded, as mentioned in section 2, that the variables *land taxation participation*, *tributary income per capita*, *number of lands per capita*, *Property value* and *total transfers* should be transformed, the logarithmic transformation was chosen.

Note: All transformed variables will be referred as their original name since the original variables do not appear anymore in the analysis.

Another important conclusion from this analysis is the possibility of collinearity problems. This is drawn from the fact that in the correlation matrix (shown below) there are many high correlations and it indicates a near-linear dependency.

```
Correlation Matrix

          Tparltot Tparltrib  devi  Tpv98  Ttran   ICV  Tnlpopu
 Tparltot    1.000     0.783 0.658  0.628  0.138 0.497    0.253
Tparltrib    0.783     1.000 0.714  0.585  0.263 0.546    0.259
     devi    0.658     0.714 1.000  0.571  0.196 0.777    0.283
    Tpv98    0.628     0.585 0.571  1.000  0.710 0.413   -0.063
    Ttran    0.138     0.263 0.196  0.710  1.000 0.110   -0.361
      ICV    0.497     0.546 0.777  0.413  0.110 1.000    0.290
  Tnlpopu    0.253     0.259 0.283 -0.063 -0.361 0.290    1.000
```

Multiple regression analysis was used to study the relation between *land taxation participation* and *catastral update* after adjusted by the variables *Development index*, *Tributary income per capita*, *property value*, *total transfers*, *number of lands per capita* and *ICV*. This initial regression is summarized in the next Splus output:

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.1446   0.3705     0.3903   0.6964
      Ttran  -0.6982   0.0371   -18.8454   0.0000
       devi   0.1100   0.0270     4.0811   0.0000
      Tpv98   0.4926   0.0253    19.4418   0.0000
        ICV  -0.0021   0.0028    -0.7432   0.4575
  Tparltrib   0.4430   0.0234    18.9491   0.0000
    Tnlpopu  -0.0186   0.0415    -0.4478   0.6544
    UPDATE1  -0.6627   0.0773    -8.5727   0.0000
    UPDATE2  -0.4087   0.0725    -5.6349   0.0000
    UPDATE3  -0.2521   0.0746    -3.3810   0.0008

Residual standard error: 0.5648 on 883 degrees of freedom
Multiple R-Squared: 0.777
F-statistic: 341.9 on 9 and 883 degrees of freedom, the p-value is 0
```

This analysis shows us that the variables *ICV* and *number of lands per capita* are not relevant in order to explain *land taxation participation*. Looking at the residuals for this model and doing some test (see Technical report), observations 57 and 872 are outliers and they were eliminated in further studies. There is another good reason to eliminate this two municipalities, that is, these two municipalities are really close to big cities, in such way that maybe in a near future they will disapear as independent municipalities and will be part of the big cities.

Automatic variable selection with BIC, AIC, R^2, R^2-adjusted and extra-sum-of-squares as criteria of selection, was used to reduce our initial model. The model obtained using automatic selection was further studied because it seemed that collinearity was a problem, but after some analysis using ridge regression [3] and principal components [2] (see Technical report), we concluded that we can use this model in order to answer the main question of the present project. The final model obtained was

$$\begin{aligned} log(\text{land Taxation participation}) &= 0.3920 - 0.6532 \cdot Ttran + 0.0914 \cdot devi + 0.4642 \cdot Tpv98 + \\ &\quad 0.4708 \cdot Tpaeltrib - 0.6371 \cdot UPDATE1 - 0.3960 \cdot UPDATE2 - \\ &\quad 0.2280 \cdot UPDATE3 \end{aligned}$$

Where *UPDATE1, UPDATE2 and UPDATE3* are dummy variables for the variable *cadastral update.*

In this model all the variables are significant and 79% of the variability of the *land taxation particpation* can be associated with the variation of the predictors *catastral update*, *Development index*, *Tributary income per capita*, *property value* and *total transfers.*

4 Discussion

The predictors in this study show near-linear dependencies because the economical constrains of the taxation system. Therefore, collinearity problems were of some concern in order to find the model that better explain the relationship between *land taxation participation* and *cadastral update.* Fortunally, after all the usual multiple regression analysis was good enough to explain their relation and more complicated model were not needed. This situation creates an advantage since this kind of study should be repeated for other years because the natural movility of the economy, and we may be aware that classical regression models are good enough to solve the question that may arise.

Analyzing further our model

$$\begin{aligned} log(\text{land Taxation participation}) \; = \; & 0.3920 - 0.6532 \cdot Ttran + 0.0914 \cdot devi + 0.4642 \cdot Tpv98 + \\ & 0.4708 \cdot Tpaeltrib - 0.6371 \cdot UPDATE1 - 0.3960 \cdot UPDATE2 - \\ & 0.2280 \cdot UPDATE3 \end{aligned}$$

many expected characteristics were found:

- ⊙ *land taxation participation* is negativelly related with *total transfers* form the Goverment. This an important feature of this model because it reflects that the process of decentralization proposed by the National Goverment, some years ago, is working, since less money transfers encourages the local goverment to recolect more efficiently land taxes.
- ⊙ The level of development of the municipalities is related with the *land taxation participation* since more developed municipalities have more control over the lands that belong to the municipality, and their citizens are more conscious of the obligation of paying taxes because they are more benefited with social improvement (more health centers, more schools, etc.).
- ⊙ Note that the principal factor that causes low cost of the property in Colombia is the violence. Therefore, the fact that *Property value* is positively related with *land taxation participation* is important because it shows that the National goverment have not successed and will not success in its descentralization process if violence is not eliminated of the country.
- ⊙ After controlling by the variables *Total transfers*, *Development index*, *Property value* and *Tributary income per capita*, *cadastral update* is positively correlated with *land taxation particpation* (the dummy variables are coded in opposite way).

With respect to our principal question, the *cadastral update* process had had the wanted consequences because it has increased the *land taxation participation* in the total income of each municipality. This increase depends also of the zones that the municipality has updated, that is, a municipality which updates its rural zone but not its urban zone is expected to increase its *land taxation participation* more than: **(a)** a municipality which updates its urban zone but not its rural zone, and **(b)** a municipality which does not update neither its urban zone nor its rural zone.

Finally, it is important to mention that the most important variable to control before studying the relationship between *land taxation participation* and *cadastral update* is *development index* (see Technical report). This a interesting finding since the money obtained by land taxation is "all[7]" used in social investment, one of the factors included in the *development index.*

5 Conclusions

We may summarize the finding in this project and recommendation for further analyses as follow:

- ⊙ The cadastral update have been effective increasing the land taxation participation in the total income of the Colombian municipalities. Hence, the descentralization process and fortification of the municipal state treasures proposed by Law 14 and financed by the National Government has had success at least until year 1998.
- ⊙ The Quality of Life Index were created to measure development and welfare of each Colombian family, therefore this varible should be used in further analysis, even though in this study was not important, but it should be recalculated with more recent information[8].
- ⊙ In this study we were just intereted on finding if there is a positive effect in the *land taxation participation* of each municipality when the cadastral update is done, further analysis looking for exact changes in the participation of land taxation will have to included more precise information for many years, for instance, we should know if the increment in land taxation participation in the total income for each municipality is due to an increment in the land taxation and not an decrement in the total income of the municipality.

[7] Maybe no all since there is a lot of corruption in the local and the national government

[8] The ICV used in this study is based on the last census, that is now eight years old.

6 Technical Appendices

1. Data set.

   Even though the information was kindly offered by the *DNP*, this was not completely ready to analyze because all the income variables, we are interested in, were disaggregated as in the "National Fiscal source" (form F-400[9]) for all the federal entities of the Colombian government. Since we are interested in municipalities, and most of their income comes from taxes, we just extracted income (by taxation) for each of the municipalities.

   Each municipality collects a total of 280 different taxes that are grouped in ten different categories. Therefore the longest process in order to obtain the final dataset was to find which taxes (of the 280) were collected for each municipality in 1998 and which category they belong to.

   As mentioned before, since there are different sources of information, a large process of rectification was needed. From this process, a total of 95 municipalities (out of 988) were dropped since the information was completely different from one source to another one, or it was not possible to obtain accurate information, or the dependent variable was missing.

2. Exploratory Data Analysis (EDA).

   After a careful inspection of the data, our next step in this data analysis was the recognition of patterns leading to simplify our problem.

   Figure 3 displays a histogram for the dependent variable *land taxation participation* using the original scale and the log-scale. As commonly happen with income variables, the distribution of *land taxation participation* is highly skewed to the right and the log transformation does a fairly good job making the data symmetric. Note that we may use another transformation that makes more symmetric the distribution of *land taxation participation*, for instance $T(x) = X^{-7}$, but we will be in trouble explaining our results, in other words, we better talk the same "language" that the economist.

   In the case of the other income variables, *Total transfers*, *Tributary income per capita*, and *Property value*, the same logarithmic transformation was done. As can be seen in Figure 4, the result of this transformation is again satisfactory, therefore no further transformation will be done. Nevertheless, we have to be aware in further analysis of some extrem observations. These observations are mostly the principal cities of Colombia, more precisely the capitals of the states with most population (without include the three biggest cities of the country: Bogota, Cali and Medellin).

   In all further analyses the transformed variables are used and refered with their original name in the text.

   Another variable in our study is the ICV. In this case, its distribution is fairly symmetric without need of transformation. But, as figure 6 shows, it is leptokurtic. This fact is due to the nature of this variable because it is an average over all the families in the municipality. This situation, therefore, may implie that *ICV* would not help us to characterize the *land taxation participation.*

[9] Form used to control the finances of the government entities

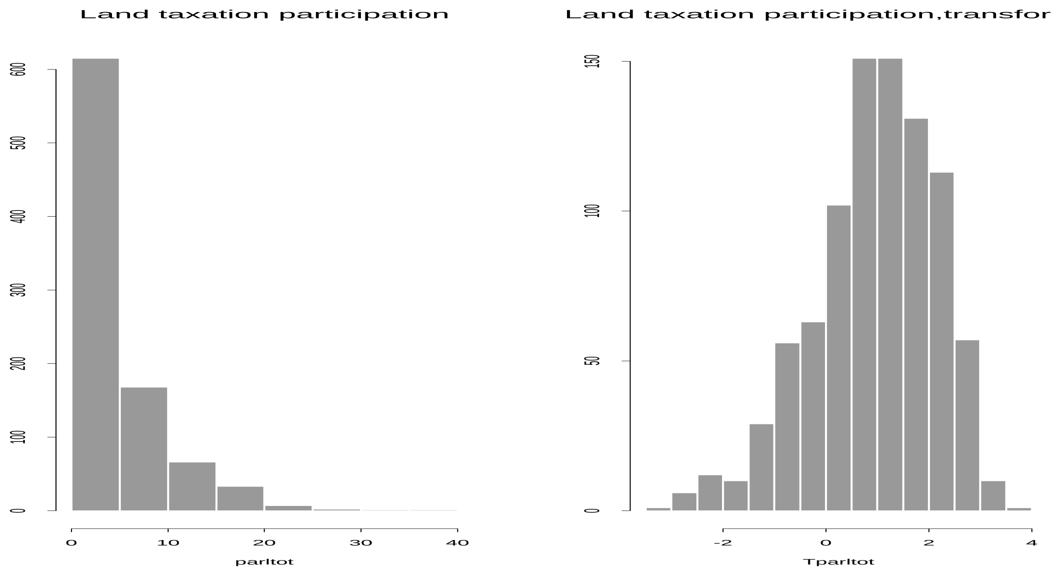

Figure 3: Histogram for *land taxation participation* in its original scale and log-transformed

In order to corroborate graphically if the variables in our data set are related, a scatterplot matrix for all the variables in this study is used. This graph is shown in figure 5. As revealed by this graph:

- there is a clear increasing linear relationship between the variables: (a) *Property value* and Total transfers; (b) *land taxation participation* and *tributary income per capita*, *Property value.*
- the variable *Development index* has a increasing relationship, with variables such as *ICV* and *tributary income per capita.*
- It seems that the principal variables in this study *Cadastral update* and *land taxation participation* are not directly correlated.
- The remaining pairs of variables give the impression of being weaker related, but at this point of our study may not be clear enough.

As a complement of the scatterplot matrix, the matrix of correlations for the continuous[10] variables is computed:

```
OUTPUT 1.
          Tparltot Tparltrib  devi Tpv98 Ttran   ICV
 Tparltot    1.000     0.783 0.658 0.628 0.138 0.497
Tparltrib    0.783     1.000 0.714 0.585 0.263 0.546
     devi    0.658     0.714 1.000 0.571 0.196 0.777
    Tpv98    0.628     0.585 0.571 1.000 0.710 0.413
    Ttran    0.138     0.263 0.196 0.710 1.000 0.110
      ICV    0.497     0.546 0.777 0.413 0.110 1.000
```

[10] *Development index* is ordinal but in this study it will be taken as continuous since working with 7 dummy variables is tedious

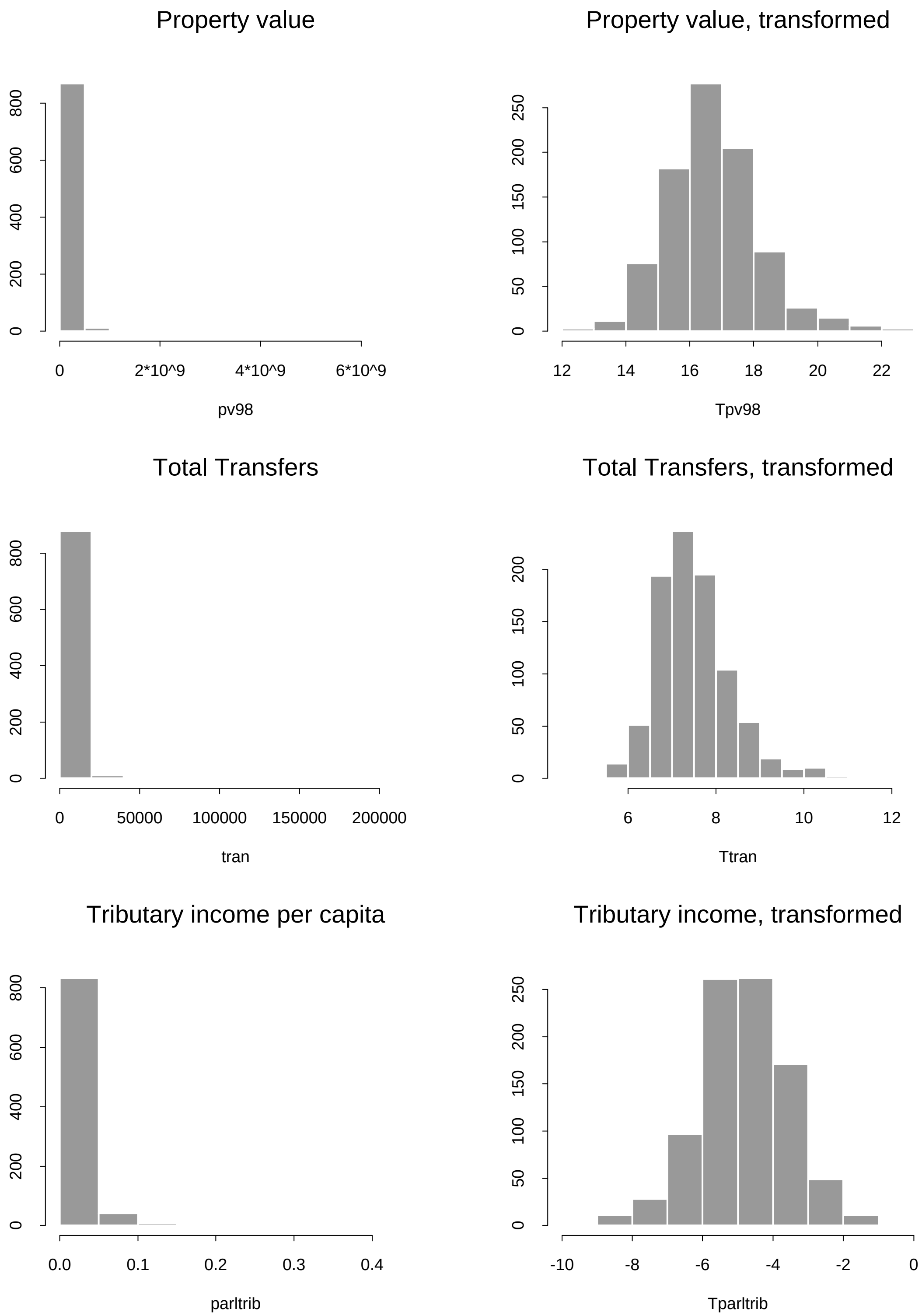

Figure 4: Histogram for *Total transfers*, *Tributary income per capita*, and *Property value* in their original scale and log-transformed

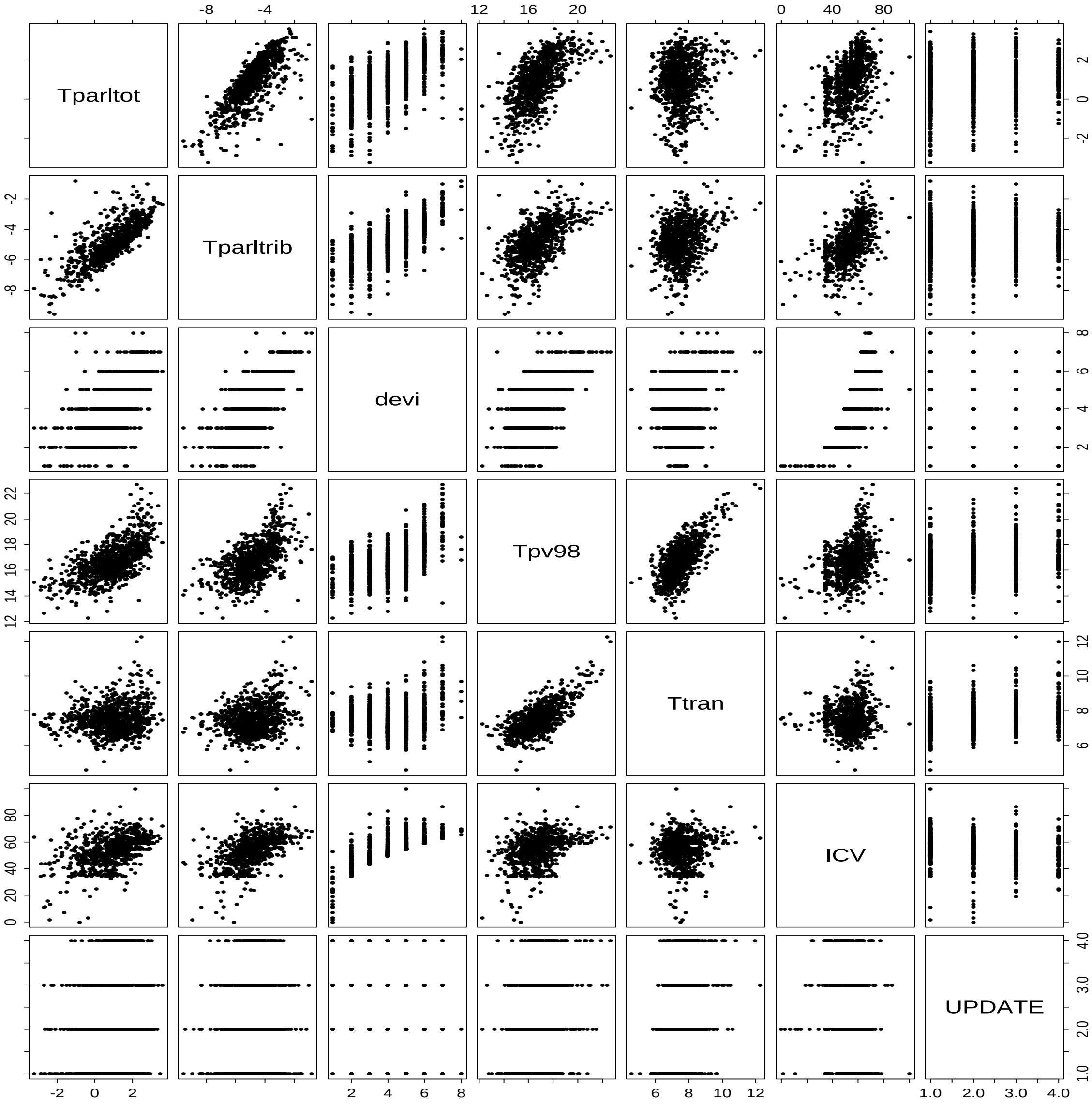

Figure 5: Scatterplot matrix for the variables *land taxation participation*, *Tributary income per capita*, *Development index*, *Property value*, *Total transfers*, *ICV* and *Update*

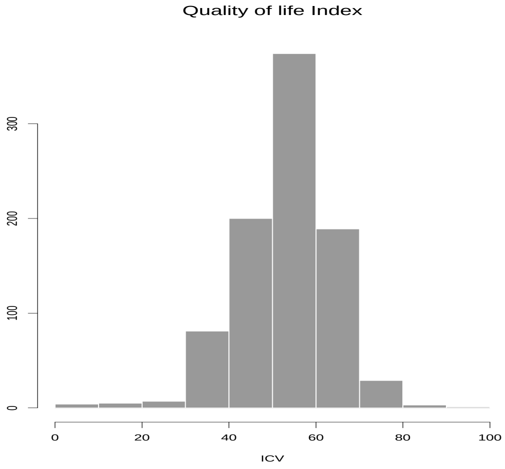

Figure 6: Boxplot for *ICV*

These correlation confirm what we already said about the relationship between the values in the study and also alert us the possibility of collinearity problems in futher regression analyses.

Finally, we take a look of one-way ANOVA's using boxplots with dependent variable *land taxation participation*

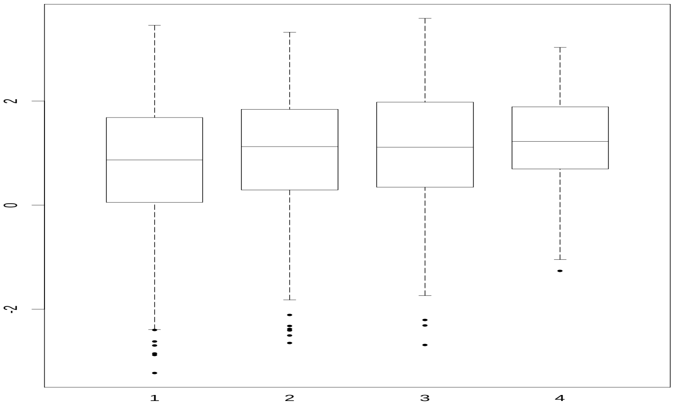

Figure 7: One-Way ANOVA using boxplot

All boxplots in Figure 7 show that there is not a clear separation between the medians of *land taxation participation* for each of the categories of *Cadastral update*. This situation may not give us too much hope in order to answer the questions of this study, but we should wait to look how this variable is related with another ones.

3. Multiple Regression Analysis

We start our analysis regressing (log-transformed) *land taxation parcipation* on all the independent variables, using three dummy variables instead of the nominal variable *cadastral update*:

```
UPDATE1<-ifelse(update==1,1,0)
UPDATE2<-ifelse(update==2,1,0)
UPDATE3<-ifelse(update==3,1,0)
```

The results of this multiple regression are summarized in output 2 and standart residual plots can be seen in figure 8. The

```
OUTPUT 2.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.1288  0.3686      0.3495   0.7268
      Ttran  -0.6925  0.0348    -19.9235   0.0000
       devi   0.1094  0.0269      4.0662   0.0001
      Tpv98   0.4923  0.0253     19.4455   0.0000
        ICV  -0.0022  0.0028     -0.7917   0.4287
  Tparltrib   0.4412  0.0230     19.1704   0.0000
    UPDATE1  -0.6624  0.0773     -8.5729   0.0000
    UPDATE2  -0.4080  0.0725     -5.6286   0.0000
    UPDATE3  -0.2508  0.0745     -3.3677   0.0008

Residual standard error: 0.5646 on 884 degrees of freedom
Multiple R-Squared: 0.777
F-statistic: 385 on 8 and 884 degrees of freedom, the p-value is 0
```

The test of the composite hypothesis that all eight regression coefficients are zero is highly significant. The coefficient of determination is 0.777, thus 78 % of the variability of the log-transformed *land taxation parcipation* can be associated with the variation of these eight predictors.

The t test of the partial regression coefficients $H_0 : \beta_j = 0$ (using the Bonferroni correction) seems to suggest that only *ICV* is unimportant and could be dropped from the model. To see if this is possible, a F-test is performed comparing the model with all predictors and the model with all predictors except *ICV*. It can be seen in output 3.

```
OUTPUT 3.

Analysis of Variance Table

Response: Tparltot

                                                                 Terms
1       UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib
2 UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + ICV + Tparltrib

  Resid. Df      RSS Test Df Sum of Sq   F Value     Pr(F)
1       885 281.9650
2       884 281.7652 +ICV  1 0.1997958 0.6268322 0.4287321
```

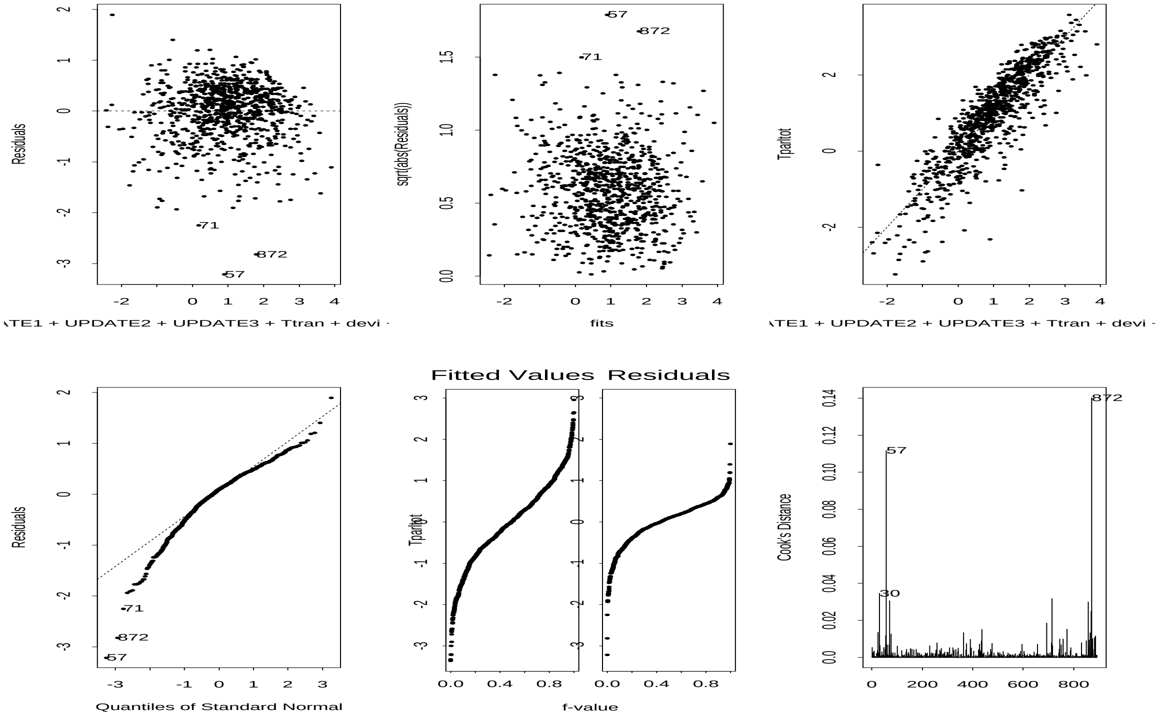

Figure 8: Splus's standard residuals plot for the model with all the predictors

Since the p-value is greater than $\alpha = 0.05$, we can drop *ICV* from the model.

We could now follow this process in order to find our best model, but it is easier if we use another criteria such as AIC and BIC. It process was performed using the function *stepwise*() and the results were displayed using Seltman's function [1] *sum.step*(), as it is shown in output 4.

```
OUTPUT 4.

          BIC     AIC    R2   R2a Ttrib devi Tpv98 Ttran ICV UPD1 UPD2 UPD3
7(#1)  -779.6  -760.4 0.777 0.775     1    1     1     1   0    1    1    1
6(#1)  -781.9  -765.1 0.774 0.772     1    1     1     1   0    1    1    0
8(#1)  -782.7  -761.1 0.777 0.775     1    1     1     1   1    1    1    1
7(#2)  -785.0  -765.8 0.774 0.772     1    1     1     1   1    1    1    0
6(#2)  -786.1  -769.3 0.772 0.770     1    0     1     1   0    1    1    1
5(#1)  -786.1  -771.7 0.770 0.769     1    0     1     1   0    1    1    0
7(#3)  -787.5  -768.4 0.773 0.771     1    0     1     1   1    1    1    1
6(#3)  -788.1  -771.3 0.771 0.769     1    0     1     1   1    1    1    0
5(#2)  -788.9  -774.5 0.769 0.767     1    1     1     1   0    1    0    0
4(#1)  -789.4  -777.4 0.767 0.766     1    0     1     1   0    1    0    0
5(#3)  -792.1  -777.7 0.767 0.766     1    0     1     1   1    1    0    0
4(#2)  -807.9  -795.9 0.757 0.756     1    0     1     1   0    0    0    1
3(#1)  -809.0  -799.4 0.754 0.753     1    0     1     1   0    0    0    0
4(#3)  -812.0  -800.0 0.754 0.753     1    0     1     1   0    0    1    0
```

```
3(#2)  -947.4  -937.8 0.665 0.664     1    1     1     0   0    0    0    0
3(#3)  -953.8  -944.2 0.660 0.659     1    0     1     0   1    0    0    0
2(#1)  -954.2  -947.1 0.657 0.656     1    0     1     0   0    0    0    0
2(#2)  -984.4  -977.2 0.633 0.632     1    1     0     0   0    0    0    0
2(#3)  -999.1  -991.9 0.621 0.620     1    0     0     0   0    1    0    0
1(#1) -1004.6  -999.8 0.613 0.613     1    0     0     0   0    0    0    0
1(#2) -1175.3 -1170.5 0.433 0.433     0    1     0     0   0    0    0    0
1(#3) -1204.7 -1200.0 0.395 0.394     0    0     1     0   0    0    0    0
0(#3) -1425.5 -1423.1 0.000 0.000     0    0     0     0   0    0    0    0
1(#2) -1427.0 -1422.2 0.004 0.003     0    0     0     0   0    0    0    1
```

As can be seen in output 4, the best model is the model with all the predictor variables except *ICV*. Note how the four criteria displayed in output 6 point out this model as the best. Output 5 and 6 show the stepwise model using AIC and BIC with the *stepAIC*() function. We obtain the same model using the BIC, but using AIC, the predictor *ICV* seems to be important.
Since most of the criteria indicate us to eliminate *ICV* and work with all other predictors, we will continue working without *ICV*.

```
OUTPUT 5.

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Tparltot ~ UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + ICV + Tparltrib

Final Model:
Tparltot ~ UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + ICV + Tparltrib

  Step Df Deviance Resid. Df Resid. Dev       AIC
1                        884   280.7305 -1015.372

OUTPUT 6.

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Tparltot ~ UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + ICV + Tparltrib

Final Model:
Tparltot ~ UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib


   Step Df Deviance Resid. Df Resid. Dev       AIC
1                         884   280.7305 -972.2208
2 - ICV  1 1.234563       885   281.9650 -975.0969
```

The results for the multiple regression analysis without *ICV* are summarized in output 7. The coefficient of determination in this case is $R^2 = 0.7768$, just a little bit smaller than the R^2 including *ICV*. Other important feature that can be seen in output 7 is the fact that all the predictors are significant, as we should have expected since the automatic algorithms for selection of variables found this model as the best.

```
OUTPUT 7.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.0512   0.3553     0.1442   0.8854
    UPDATE1  -0.6633   0.0772    -8.5881   0.0000
    UPDATE2  -0.4069   0.0725    -5.6162   0.0000
    UPDATE3  -0.2518   0.0745    -3.3819   0.0008
      Ttran  -0.6915   0.0347   -19.9120   0.0000
       devi   0.0968   0.0217     4.4659   0.0000
      Tpv98   0.4925   0.0253    19.4575   0.0000
  Tparltrib   0.4412   0.0230    19.1761   0.0000

Residual standard error: 0.5645 on 885 degrees of freedom
Multiple R-Squared: 0.7768
F-statistic: 440.1 on 7 and 885 degrees of freedom, the p-value is 0
```

Now let us take a look of the residuals. As displayed in figure 8, the observations 57 and 872 seems to be outliers. One way of testing if they really are outliers is to construct two dummy variables that take the value 1 if the observation is the 57th and 872nd, respectively. The results of the multiple regression incluiding each of these dummy variables is summatized in output 8 and 9 respectively. These outputs also include the F test for comparing the full model with the model without each possible outlier.

```
OUTPUT 8.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.1842   0.3512     0.5246   0.6000
      Ttran  -0.6633   0.0347   -19.1326   0.0000
       devi   0.1020   0.0214     4.7671   0.0000
      Tpv98   0.4739   0.0252    18.7974   0.0000
  Tparltrib   0.4527   0.0228    19.8634   0.0000
       I872  -2.9380   0.5683    -5.1697   0.0000
    UPDATE1  -0.6520   0.0762    -8.5590   0.0000
    UPDATE2  -0.4059   0.0714    -5.6829   0.0000
    UPDATE3  -0.2528   0.0734    -3.4438   0.0006

Residual standard error: 0.5564 on 884 degrees of freedom
Multiple R-Squared: 0.7834
F-statistic: 399.6 on 8 and 884 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: Tparltot

                                                                     Terms
1 UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib + I872
```

```
2        UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib

  Resid. Df      RSS  Test Df Sum of Sq  F Value         Pr(F)
1       884 273.6907
2       885 281.9650 -I872 -1 -8.274289 26.72531 2.901111e-007
```

```
OUTPUT 9.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.2540   0.3505     0.7248   0.4688
      Ttran  -0.6820   0.0341   -19.9828   0.0000
       devi   0.0862   0.0214     4.0380   0.0001
      Tpv98   0.4833   0.0249    19.4104   0.0000
  Tparltrib   0.4589   0.0228    20.1389   0.0000
        I57  -3.2934   0.5615    -5.8654   0.0000
    UPDATE1  -0.6489   0.0759    -8.5532   0.0000
    UPDATE2  -0.3971   0.0711    -5.5820   0.0000
    UPDATE3  -0.2273   0.0732    -3.1051   0.0020

Residual standard error: 0.5541 on 884 degrees of freedom
Multiple R-Squared: 0.7852
F-statistic: 403.9 on 8 and 884 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: Tparltot

                                                                 Terms
1 UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib + I57
2       UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib

  Resid. Df      RSS Test Df Sum of Sq  F Value         Pr(F)
1       884 271.4029
2       885 281.9650 -I57 -1 -10.56214 34.40247 6.328724e-009
```

As shown in output 8 and 9, these two dummy variables are important in the model, situation that indicates that they are outliers, therefore they should be study separately of the rest of the data. Before to exclude them, another regression including both of the dummy variables was performed to establish if both at the same time should be dropped of our analysis. Looking the p-values for t test and for the F-test in output 10, we conclude that these two variables may be dropped of our analysis.

```
OUTPUT 10.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.3920   0.3461     1.1328   0.2576
      Ttran  -0.6532   0.0340   -19.1913   0.0000
       devi   0.0914   0.0211     4.3398   0.0000
      Tpv98   0.4642   0.0248    18.7392   0.0000
  Tparltrib   0.4708   0.0225    20.8808   0.0000
       I872  -2.9882   0.5573    -5.3617   0.0000
        I57  -3.3377   0.5530    -6.0361   0.0000
    UPDATE1  -0.6371   0.0747    -8.5253   0.0000
    UPDATE2  -0.3960   0.0701    -5.6523   0.0000
    UPDATE3  -0.2280   0.0721    -3.1625   0.0016

Residual standard error: 0.5456 on 883 degrees of freedom
Multiple R-Squared: 0.792
F-statistic: 373.5 on 9 and 883 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: Tparltot

                                                                        Terms
1 UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib + I872 + I57
2             UPDATE1 + UPDATE2 + UPDATE3 + Ttran + devi + Tpv98 + Tparltrib

  Resid. Df      RSS      Test Df Sum of Sq  F Value         Pr(F)
1       883 262.8453
2       885 281.9650 -I872-I57 -2 -19.11973 32.11531 3.441691e-014
```

The results for the multiple regression analysis without *ICV* and without the observations 57 and 872 is summarized in output 11. The coefficient of determination is $R^2 = 0.7896$, which has increased with respect to our precious models. As before all the predictors are significant.

Figure 9 shows the Splus's standard residuals plot for our actual model. As is revealed by the first two plots, the residuals are they are uncorrelated, but they are not normal distributed (see QQ-plot). Some other observation seems to be outliers, but after performed the same analyses like for observation 57 and 872, the dummy variables are not significant, then we do not have vaild reasons to drop them out of the model.

```
OUTPUT 11.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.3920   0.3461     1.1328   0.2576
      Ttran  -0.6532   0.0340   -19.1913   0.0000
       devi   0.0914   0.0211     4.3398   0.0000
      Tpv98   0.4642   0.0248    18.7392   0.0000
  Tparltrib   0.4708   0.0225    20.8808   0.0000
    UPDATE1  -0.6371   0.0747    -8.5253   0.0000
    UPDATE2  -0.3960   0.0701    -5.6523   0.0000
    UPDATE3  -0.2280   0.0721    -3.1625   0.0016

Residual standard error: 0.5456 on 883 degrees of freedom
Multiple R-Squared: 0.7896
F-statistic: 473.3 on 7 and 883 degrees of freedom, the p-value is 0
```

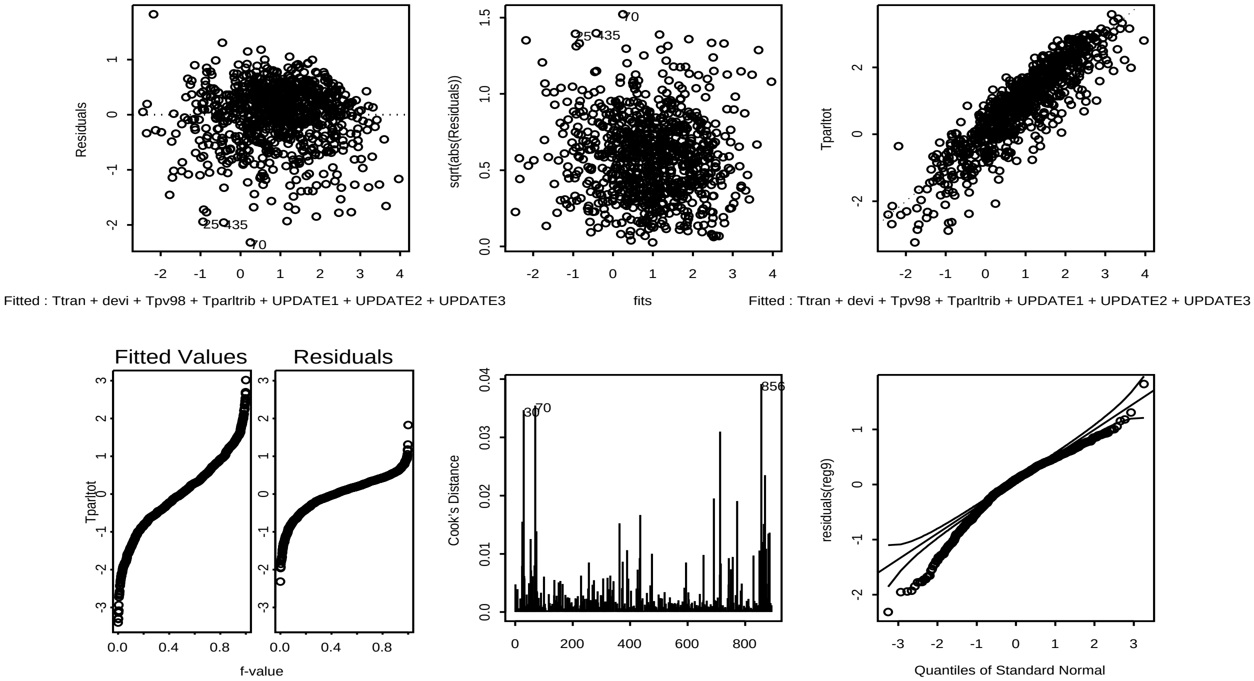

Figure 9: Splus's standard residuals plot for the final model

4. Collinearity?

   Based on the previous EDA, we are aware of the possibility of a collinearity problem in our model. Therefore this section was designed to check out if this is really a problem or we can just work with these variables applying ordinary least-squares.

   For this part of our analisys the three dummy variables are not used because they are categorical variables and most of the tecniques that I am aware of to study collinearity are for continuous variables. Moreover, it seems that the collinearity comes from the income variables and not from the em cadastral update.

   Output 13 shows the VIF for each of our variables and the condition number. The VIF were computed using Seltman's function *collin()* [1], and the condition number

$$\kappa = \left(\frac{\text{largest eigenvalue}}{\text{smallest eigenvalue}}\right)^2$$

```
OUTPUT 13.
VIF
          Tparltrib       2.298851
          devi            2.369668
          Tpv98           3.649635
          Ttran           2.375297
```

```
k = 40.28514
```

Following the proposal of declaring collinearity a problem if $\kappa \geq 30$ by Weisberg[2], we may be facing a problem with our predictors. But note that the maximun VIF is just 3.64 which is less than 5, a common cutoff for collinearity problems, therefore collinearity may not be a problem after all. Since we have two cirteria and they tell us different answers, principal components and ridge regression will be tried.

First, we work with principal component analysis. In these case we have four principal components, if we used the screeplot (Figure 10) to decide how many principal components we should work with, the selection will be two, but in this analysis we would rather include the four component and test if we should discard some of them using the t test and extra-sum-of squares F -test. These test can be seen in output 14.

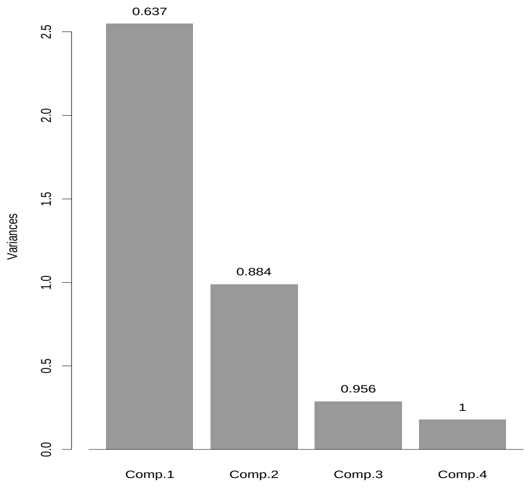

Figure 10: Screeplot

```
OUTPUT 14.

Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)   1.3615   0.0625    21.7933   0.0000
      Comp1   0.5256   0.0116    45.3644   0.0000
      Comp2  -0.5247   0.0203   -25.8667   0.0000
      Comp3  -0.2304   0.0368    -6.2621   0.0000
      Comp4  -0.7149   0.0444   -16.0888   0.0000
    UPDATE1  -0.6371   0.0747    -8.5253   0.0000
    UPDATE2  -0.3960   0.0701    -5.6523   0.0000
    UPDATE3  -0.2280   0.0721    -3.1625   0.0016

Residual standard error: 0.5456 on 883 degrees of freedom
Multiple R-Squared: 0.7896
```

```
F-statistic: 473.3 on 7 and 883 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: Tparltot

                                                         Terms Resid. Df      RSS
1 UPDATE1 + UPDATE2 + UPDATE3 + Comp1 + Comp2 + Comp3 + Comp4       883 262.8453
2                 UPDATE1 + UPDATE2 + UPDATE3 + Comp1 + Comp2       885 349.5635

          Test Df Sum of Sq  F Value Pr(F)
1
2 -Comp3-Comp4 -2 -86.71825 145.6602     0
```

NOte that the p-values for the t-test and the F-Test are much smaller than $\alpha = 0.05$, therefore we conclude that all the principal components are needed in our model. This situation may implie that the collinearity is not a severe problem in our model.

Another technique commonly used when there is collinearity problems is ridge regression analysis (see Hoerl and Kennard [3], [4], Weisberg [2]). The data have been standarized, as commonly is done in ridge regression. The difficulty using ridge regression is to find the value k in the equation $(X'X + kI)^{-1}XtY$, one way of determine the value of k, is to look at the ridge trace, that is, the value of each parameter for different values of k, and choose the value of k such as the trace of all parameters is constant. Figure 11 shows the ridge trace in our data, it is constant for most of the parameters after 0.3, therefore $k = 0.3$.

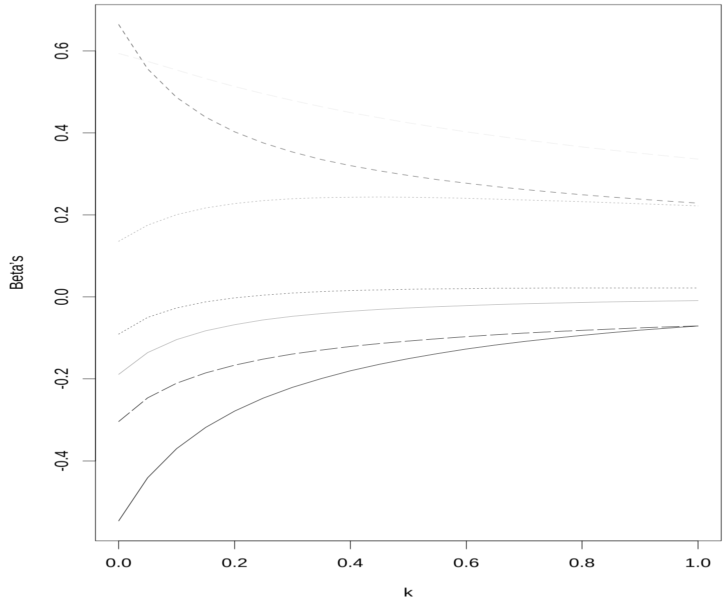

Figure 11: Ridege trace

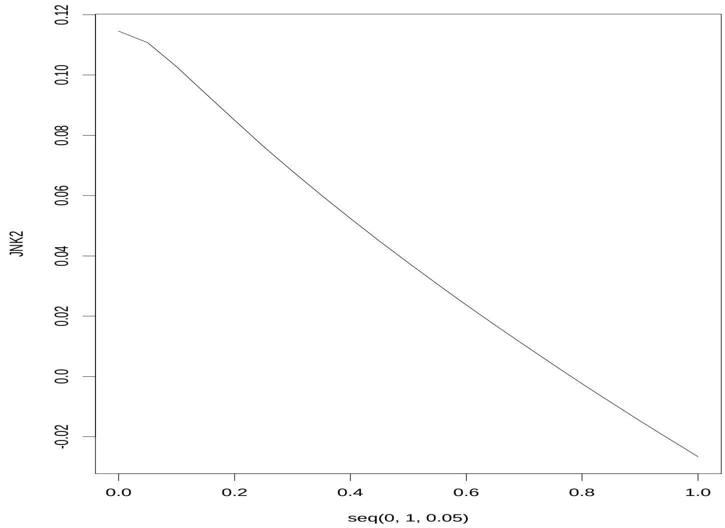

Figure 12: Cross-validation index for least-squares ridge regression

There is another less heuristic approach to find our ridge regression parameter k, the idea of this is to choose the value of k that maximize the cross-validation index defined as $I_{CV} = 1 - PRESS/PRESS_0$[5], where $PRESS_0$ is the PRESS-value for the ”null” predictor, $\hat{y} = \overline{y}$. Figure 12 shows the cross-validation index as function of k. This method suggests that k should be zero, that is we can just apply ordinary least-squares and the collinearity is not a problem.

The histogram of residuals for the model with all predictors but *ICV* without observations 57 and 872 appears in figure 13. This figure reveals that the residuals are not exactly normal as we saw in figure 9. Taking into account this situation other transformation were used, but a similar behavior was found.

We already found that after controling with the variables *Development index*, *Tributary income per capita*, *property value*, *total transfers*, we can observed the relationship between *land taxation participation* and *cadastral update*. As a way to find out, which of these variables contributes more, the *cadastral update* was dropped from the model, and the regression results were:

```
OUTPUT 15.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -0.4204   0.3484    -1.2068   0.2278
      Ttran  -0.6076   0.0353   -17.2164   0.0000
       devi   0.0050   0.0198     0.2527   0.8006
      Tpv98   0.5029   0.0256    19.6541   0.0000
  Tparltrib   0.5176   0.0230    22.4620   0.0000

Residual standard error: 0.5717 on 886 degrees of freedom
Multiple R-Squared: 0.7681
F-statistic: 733.8 on 4 and 886 degrees of freedom, the p-value is 0
```

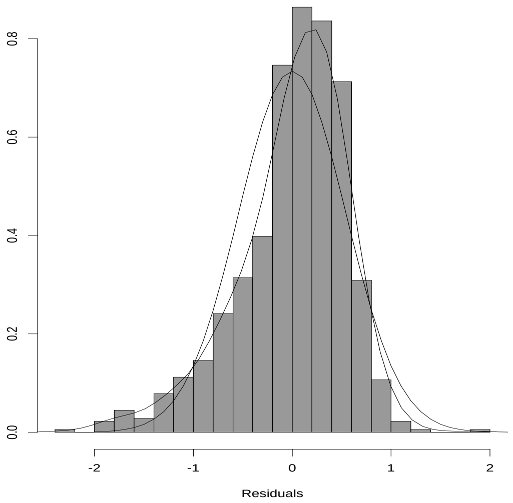

Figure 13: Histogram of residuals with the density estimation using kernels and the normal distribution with ML estimator of μ and σ^2

It is inmediate clear that *development index* is not needed anymore in the model. This is clear a consequence of drooping *cadastral update*. Now the model including *development index* and *cadastral update* is fitted:

```
OUTPUT 16.

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -0.7031   0.1057    -6.6502   0.0000
       devi   0.5874   0.0188    31.2062   0.0000
    UPDATE1  -1.1642   0.0996   -11.6863   0.0000
    UPDATE2  -0.7622   0.0984    -7.7454   0.0000
    UPDATE3  -0.4682   0.1057    -4.4297   0.0000

Residual standard error: 0.8117 on 886 degrees of freedom
Multiple R-Squared: 0.5326
F-statistic: 252.4 on 4 and 886 degrees of freedom, the p-value is 0
```

As we may have expected, all the variables are significant. Then, after adjusting by *development index* the basic relationships between *land taxation participation* and *cadastral update* are already clear. Nevertheless, the other three predictors also provide us information so it is important to keep them in the model.

References

[1] Seltman, H. (1999). http://lib.stat.cmu.edu/ hseltman/.

[2] Weisberg, s. (1985). Applied Linear Regression.John Wiley & Sons, New York.

[3] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal probmes. Technometrics, 12, 55-67.

[4] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. Technometrics, 12, 69-82.

[5] Björkström A. (2001). Ridge regression and inverse problems. http://www.matematik.su.se/ bjorks/eget/rraip.pdf.