Public Transportation Demand Model for Low Density Territories

HARLEY LARGO[1a], PAULO J. G. RIBEIRO[1b], GABRIEL DIAS[2c], CLAUDIA P. B.TRUJILLO[1d]
[1]Department of Civil Engineering
University of Minho
Campus de Azurém, 4800-058
GUIMARÃES, PORTUGAL
[2]School of Civil Engineering
University of Santander
COLOMBIA
harleylargoamado94@hotmail.com[a], pauloribeiro@civil.uminho.pt[b], gjcd1992@hotmail.com[c], cpbaezt@uis.edu[d]

Abstract: In the course of recent years, there has been a gradual progressive deterioration in some territories, which is caused by economic problems, lack of job opportunities, deficiencies in mobility, among other problems that affect these territories. This deterioration has been causing a reduction in its population and therefore an emigration to large cities where the needs in economic, social and welfare terms can be met in a less complex way. The territories that have been presenting these affectations are denominated territories of low population density. For the specific case analyzed in this study, continental Portugal owns 59% of the municipalities as low density territories (LDT) (this is equivalent to 165 municipalities) [1]. Knowing the importance of public transport to the population of LDT (since it may be the only option to be mobilized) and considering that degradation of the public transportation (PT) increases the problems of isolation of these populations. This study aims to determine which factors (variables) have more influence in the estimation of the demand of public bus transportation for the LDT in mainland Portugal. In addition, a model is proposed to estimate the demand for these low-density territories. The mathematical model of multiple linear regression (MLR) is used based on the most influential socioeconomic and demographic variables for LDT. The model was developed with the statistical tool SPSS (statistical package for the social sciences). The estimated model presented an adjustment of 87%, taking into account the variables of number of illiterate people, population density and purchasing power. In addition, an analysis by regions (NUTS II classification) was carried out to determine which region presents a lower error percentage in the estimations. From this analysis it can be concluded that the northern region with 88% of municipalities (equivalent to 36 LDT) presents an estimate of error lower than 50%.

Key-Words: Low-density territories; demand model; public transport; multiple linear regression; SPSS.

1 Introduction

Public transport has played a crucial role in the population, for its economic, environmental and social contributions, since it has the purpose of allowing the displacement to the different activities in a society, complying with certain safety, speed, comfort and cost standards [2]. These quality standards have been affected over the years by various problems and behavioral changes in the communities that have not been foreseen or modeled by the entities in charge of planning. This is why there has been a deterioration in the public transport service (specifically service in bus) causing users to look for other displacement options, what leads to unsustainable practices and some lack of resilience of the transportation systems [3].

The gap in public transportation offer and the changes presented in the demand for public transportation have led to problems not only for large cities with large numbers of residents but also, this has been affecting the low population density territories as it is affirmed by Fernandes [4], that the operators of public bus transport in the TBD have had difficulty in developing effective, flexible and innovative measures to be able to satisfy the demand for this service in these territories.

The designation of TBD comes not only from demographic but also economic aspects (high unemployment, scarcity and poor diversity of economic activities), urban (insufficient size of most of its urban centers, even the most important ones), institutional range of entities with attributions and

competences of proximity) and relational (weak networks of partnerships and deficient rates of participation and involvement of the population) [5[6]**.** That is why depending on the public policies of a region or a country the determination of TBD can change.

Knowing what the problems of TBD in continental Portugal are, including the knowledge in areas not only as mobility and accessibility, but also in the demographic, social and economic areas, allows to make a more assertive estimate of transport demand for TBD. According to Domingues [7], The difficulty that the inhabitants of the TBD have to obtain employment opportunities, services and other types of assets have a great influence on the current problems of accessibility of these territories. In addition, Carey Curtis [8] determined that the TBD generates trips with a greater distance and with a dependence on private vehicles as the main means of transport, instead of more healthy modes with evident social and economic benefits [9,[11].

The multiple linear regression method is one of the most used models for the estimation of transport demand, since it allows estimating the number of users based on variables (whose obtaining in less complex) that have a more significant influence in the real behavior, in addition, that their results have shown a better approximation to the real demand. [10]–[14]

In any model for estimating transport demand either with the multiple linear regression method or with another mathematical model, one of the most important or fundamental steps is the determination of the variables that will represent the situation to be modeled (real situation). That is why Juan de Dios Ortuzar [15]**,** called the "cornerstone" of the transport demand model to the exogenous variables since they are the ones that contain real-life socioeconomic and demographic information. Uraiporn [16]**,** points out that there is a differentiated behavior with influence on the demand of transport in young people between 20 and 24 years old, whose modal choice is influenced by the level of schooling, the stage of life and age.

Prior to this study, a research was developed by Harley & Ribeiro [17], where the demand for transport in the TBD is estimated, having as main objective the classification of these territories based on the population density parameter, trying to analyze if the municipalities present behavior in the demand of different trips because of this parameter. One of the difficulties that were most important for the estimation of these models was the amount of data with which each model was estimated.

This article is composed of 6 sections. Section 1 deals with the introduction, presenting a summary literature review about the subject being studied, but also the objectives and structure of the paper. In section 2 is presented the methodology used. Section 3 presents the analyses of the case study. Section 4 the validation of the demand model. Section 5 analyzes of the main results. Finally, in section 6 the main conclusions are described.

2 Methodology

The methodology applied in this study for the estimation of the transport demand model for the territories of low population density in continental Portugal is based on [17] and [18].

2.1 Definition of Variables

To determine which variables have a greater influence on the dynamic behavior of journeys in territories with low population density, a bibliographical search was carried out [18], where it was concluded that there are socioeconomic, demographic and specific characteristics of each family that affect the generation of trips not only in cities with standard population densities but also in the LDT. The variables determined as influential for the transport demand in the bus service for the municipalities categorized as low density for mainland Portugal are:

- People between the ages of 15 and 24 and 55 or over;
- Number of women's trips;
- Number of unemployed;
- Number of persons without driving license;
- Number of vehicles per dwelling;
- Average income per family (monthly);
- Number of trips for purchases or social purposes;
- Travel time;
- Level of education;
- Walking time;
- Number of dwellings with one person.

2.2 Data Collection

The data of the municipalities classified as TBD were obtained from the census conducted in 2011 in mainland Portugal. The importance of censuses to obtain information regarding economic, social and environmental behavior, among others, makes this

study the only available and reliable source of obtaining data on the variables that influence transportation demand. Therefore, the choice of variables to work in the future analyzes will depend on whether there are data for them. The information was obtained and is available through the National Institute of Statistics (INE).

2.2.1 Sample Selection

The selection of the sample to estimate the transport demand model was carried out through the "random" method. This selection was made to the total of the LDT, after the data analysis (section 2.3) only 85% of the municipalities were separated, the remaining 15% of the municipalities are used to validate the model.

2.3 Data Analysis

The preliminary analysis for the data of the variables presented in section 2.1 and with the restriction exposed in section 2.2, was carried out with statistical techniques that allowed to know how the behavior of the data of the variables is. This procedure was carried out by [18], where a descriptive statistical analysis, an analysis of extreme values (using box plot and stem-and-leaf plot) and the evaluation of normality for each variable were considered. The purpose of this procedure is to analyze the behavior of the data of the variables in order to identify an anticipated way possible for municipalities that may have an impact on the estimation of the transport demand model.

2.4 Multiple Linear Regression (MLR) Model

The regression model is one of the most popular statistical techniques due to its high explanatory power and the ease of its interpretation and use in computer programs. [14]

The multiple linear regression (MLR) model is composed of several exogenous (explanatory) variables, the model is usually represented as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_j X_{ji} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad (1)$$
$$i = 1,2,\ldots,n$$

- n: sample size;
- k: Number of observable exogenous variables added to the constant, where X's and Y are observable variables
- εi: Non-observable and random exogenous variable, which includes all influences in Y that are not explained by X's;
- βk: are parameters of the model, that is, quantities that always assume the same value in it.

The statistical models are based on assumptions that allow a better approach to the problems, including MLR. [19]. Thus, the assumptions considered in such analysis will be presented, considering only the most relevant aspects for MLR:

- Linearity;
- Independence;
- Homoscedasticity;
- Normality.

2.5 Model Estimation

According to what is mentioned in section 2.2, regarding the availability of information for the variables that were determined as influential in the behavior of trips in the LDT. Next, we present the general equation and the variables that may be taken into accounts for the estimation of the transport demand model for the bus service in the territories with low population density:

$$\begin{aligned} \mathrm{VGA} = {} & \hat{B}_o + \hat{B}_1(Dsgd) + \hat{B}_2 PdC + \hat{B}_3 DP \\ & + \hat{B}_4 IM + \hat{B}_5 GM + \hat{B}_6 SnE \\ & + \hat{B}_7 ES + \hat{B}_8 IP \end{aligned} \qquad (2)$$

VGA: Number of trips by bus;
Dsgd: Number of unemployed;
PdC: Purchasing power per capita;
DP: Population density;
IM: Number of women over 15 years of age;
IP: Number of people between the ages of 15 and 24 and over 50;
GM: Average monthly income;
SnE: Number of illiterate people;
ES: Number of people with higher education degree.

The estimation of the model was made through SPSS Software. The variables with a greater significance for the model are chosen by the Stepwise method. This method according to Pestana & Gageiro [20] allows solving problems of multicollinearity. The stepwise consists of entering the explanatory variable that presents the highest coefficient of correlation with the independent variable (number of trips by bus). Then, the partial correlation coefficients are calculated for all the variables that are not part of the first regression. Thus, the next variable that enters is the one with the highest partial correlation coefficient. The new regression (equation) is estimated and is analyzed

whether one of the two independent variables should be excluded from the model. At the end, if both variables have significant *t-values*, new correlation coefficients are calculated for the variables that did not enter. The process ends when you no longer have to enter variables into the equation.

According to Juan de Dios Ortúzar [15], a model that has many explanatory variables will have a higher level of difficulty of interpretation, since the effects that some variables will have are small. Therefore, having a model with fewer variables, but with a Stronger explanation capacity will be more effective, since its predictive error is smaller compared to the one that has more variables.

2.6 Statistical Diagnosis of the MLR Model

In this section will be present the techniques to be used to verify the hypothesis of the multiple linear regression model discussed in section 2.4.

2.6.1 Diagnosis of Homoscedasticity

The condition of homoscedasticity verifies the assumption that the errors must have constant variance. The variance is constant to suppose that there are no observations included in the residual variable whose influence is more intense in the dependent variable.

The technique used to verify this hypothesis is recommended by Pestana & Gageiro [21] and Rodrigues [14], which consists in verifying the graphs of the standardized residuals (Y = Zred) and the estimated values of Y (X = Pre_1) , maintain an approximately constant amplitude in relation to the zero horizontal axis.

2.6.2 Diagnosis of Independence

To verify the zero covariance or also known as the independence of the residues, it is carried out through the Durbin-Watson test (DW), the interpretation of the test values is presented by [21] and corresponds to:

- For values close to 2, there is no autocorrelation of two residues;
- For values close to 0 it means a positive autocorrelation;
- For values close to 4 there is a negative autocorrelation.

2.6.3 Diagnosis of Normality

In this diagnosis it is verified if the waste has a normal behavior, the techniques used to finish the fulfillment of the assumption are:

- Test of Kolmogorov- Smirnov;
- Test of Shapiro- Wilk
- Normal graph Q-Q plot;
- Graphic Detrended Normal Q-Q plot.

2.6.4 Diagnosis of Multicollinearity

The term multicollinearity is used to express the strong correlation between more than two independent variables. Rodrigues [14] defines multicollinearity as the existence of a linear relationship between one of the independent variables and the remaining variables.

The intensity of multicollinearity is analyzed essentially through the following points:

- Correlation between independent variables: the correlation matrix is used. When the correlation coefficients between the independent variables are high (superior in absolute terms to 0.9) indicate the possibility of multicollinearity.
- Tolerance and VIF: Tolerance measures the degree to which one variable X is explained by all other independent variables. Therefore, the tolerance of variable X measures the proportion of its variation that is not explained by the other independent variables. This tolerance varies between zero and one, the closer it is to zero, the greater the multicollinearity.

2.7 Outliers and Influential Observations

The outlier is termed as an extreme, atypical, or aberrant value. Atypical extreme values may represent cases that have particular characteristics in relation to the study variable or result from errors made during data entry. It is also considered extreme observations, according to Rodrigues [14], the ones that are so far removed from most data that doubts arise as to whether they may or may not have been generated by the proposed model to explain this majority of the data.

The aberrant observations are not detected exclusively by the analysis of the waste, and other statistics need to be used for this purpose, in particular [20].

Statistics used to identify the outliers among the independent variables:

- Standardized wastes;
- Student waste;
- Leverage;
- Estimated adjusted values

Statistics used to identify influential cases:

- Eliminated student waste;
- Cook's distance;
- Standardized DfBeta;

- Standardized DfFit;
- Covariance ratio.

2.8 Validation of the Model

The verification of the results provided by the transport demand estimation model for territories of low population density is carried out through validation. The purpose of this technique is to verify that the results of the model are reasonable and acceptable in the estimation of transport demand. The percentage of municipalities chosen to carry out the validation process was 15%, the choice of the LDT was carried out randomly.

For each of the validation municipalities, the estimated number of bus trips (Yestimado) will be calculated by using equation 2 of the RLM model, and then based on the actual value of the number of bus trips, to determine the error percentage for every PET. Likewise, the frequency of errors will be calculated for the municipalities of validation, allowing to define if the model is relevant for the estimation of the trips in the territories of low density.

3 Case Study

3.1 Analysis of Extreme Values

Extreme values, namely those that are called "outliers", are values of the variables that are generally very far from the mean values. For some authors, when an observation has a value equal to or greater than 3 standard deviations from the mean value, it can be considered an outlier. Ribeiro [22] emphasizes the importance of knowing whether or not the outliers are to be accounted for in the model, since these values may have a significant weight in the overall study results depending on their value.

In order to decide the observations (municipalities) that are aberrant (outliers) in the database of low density territories, it was necessary to make a previous analysis for each variable with the techniques that were described in section 2.3 with the objective of identifying which are considered to be severe, in order to determine the frequency with which these values (municipalities) are repeated in all the variables under study for the demand estimation model. This analysis is done with the chart called box plot which situates the quartiles of the distribution. At the extremes of each box plot are positioned the minimum and maximum observations, the values or observations outside this box are called aberrant or outliers (Figure 1). The number of municipalities that presented extreme values in the data of most variables were 14 and will not be taken into account for the estimation of the model.

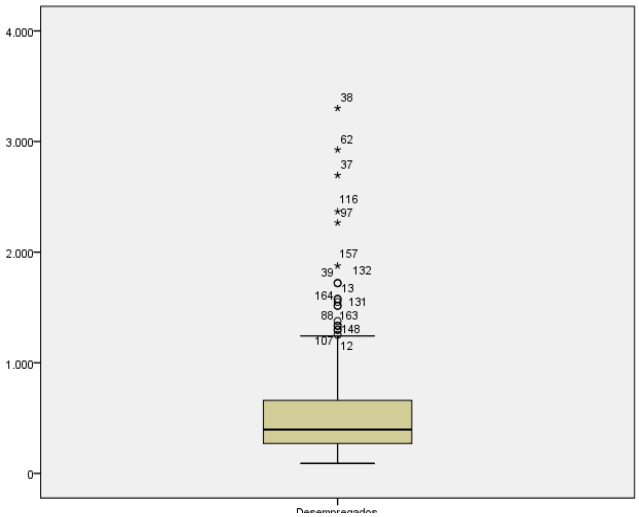


Fig.1: Example of a boxplot - Variable of the unemployed

3.2 Selection of Significant Variables

The choice of the variables that will enter in the analysis of the MLR model of bus transport demand for the LDT in Portugal were chosen based on the relationships between the variable dependents (number of bus trips) and the independent variables (presented in section 2.1 applying the condition 2.2 of the data availability), through the analysis of the level of linear correlation between the variables. To determine the intensity of association for all possible variables in integrating the estimation model, we used the statistical measures of Pearson's R and Spearman's Rho, which aims at an analysis of pairs of linear correlations.

The following are the variables that based on the previous analysis presented a correlation more adequate for its possible use in the prediction of the model. Besides, Table 1 presents the level of correlation for each of the chosen variables, it is important to note that although the correlation of the economic variables such as the average monthly gain and the purchasing power is not very strong with the other variables, they were not rejected from the analysis because they are the only ones with economic information of the LDT population. The variables used are as follows:

Table 1. Model summary

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	0,847[a]	0,718	0,716	209,506	0,718	321,006	1	126	0	
2	0,903[b]	0,815	0,812	170,602	0,096	65,019	1	125	0	
3	0,914[c]	0,836	0,832	161,105	0,021	16,171	1	124	0	1,513

a. Predictors: (Constant), Number of illiterate people
b. Predictors: (Constant), Number of illiterate people, Population density
c. Predictors: (Constant), Number of illiterate people, Population density,Purchasing power per capita
c. Dependent Variable: Number of trips by bus

- Population density;
- Number of people between the ages of 15 and 24 and over 50;
- Number of unemployed;
- Number of women over 15 years of age;
- Number of illiterate people;
- Number of people with higher education degree;
- Average monthly income;
- Purchasing power per capita.

3.3 Estimation of the Multiple Linear Regression Model

The municipalities classified as territories with low population density are 165 but taking into consideration the analysis of extreme values (section 3.1), it could be concluded that 14 municipalities demonstrate a behavior that would affect the estimation of the transport demand model. Therefore, the total of TBD that will be analyzed is 151 municipalities, based on section 2.2.1, where it is mentioned that 85% of TBDs are going to be considered in the estimation of the RLM, this percentage is equal to 128 municipalities. The number of dice will allow you to obtain a robust data base so that you can estimate model fosse or more trustworthy.

In principle, the estimation of the model begins, linking those variables that were considered significant according to the bivariate analysis of the correlations performed in the previous point. Table 1 presents the summary information of the model that was determined having as input the 8 variables considered significant. The model summary is presented, and it is possible to observe that the first variable that enters the model corresponds to people who do not have any level of education (illiterate), this is because it is the variable with the highest correlation coefficient. The model summary also presents the multiple correlation coefficient, R for the variable VGA - number of trips by bus. The R squared or the determination coefficient is a measure of the quality of fit and allows to conclude that approximately 84% of the behavior of the variable is taken into account by model 3 (definitive for the stepwise procedure).

The estimated model corresponds to the number 3, which presents a greater explanation of the variable depends (VGA) and has a greater number of variables.

$$VGA = 321{,}88 + 0{,}191 * SnE + 3{,}344 * DP - 6{,}435 * PdC \quad (3)$$

VGA: Number of trips by bus;
SnE: Number of illiterate people; [645;8221]
DP: Population density; [6;166]
PdC: Purchasing power per capita; [56,08;105,7]

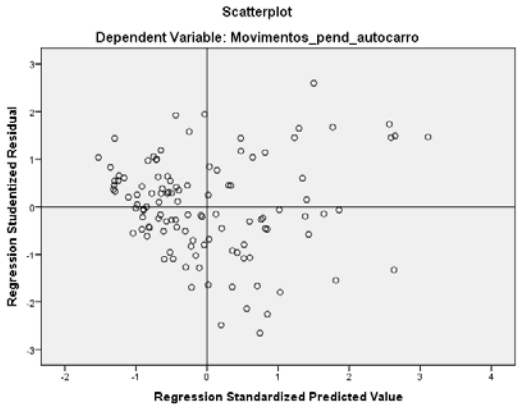
It is important to emphasize that the estimated model presented in equation (3) may suffer some alteration, especially in the coefficients, since at the moment of the analysis of the residues, the values considered outliers can be eliminated, given the potential deviation that can cause between the values and the modeling.

3.4 Diagnostics of MLR Model

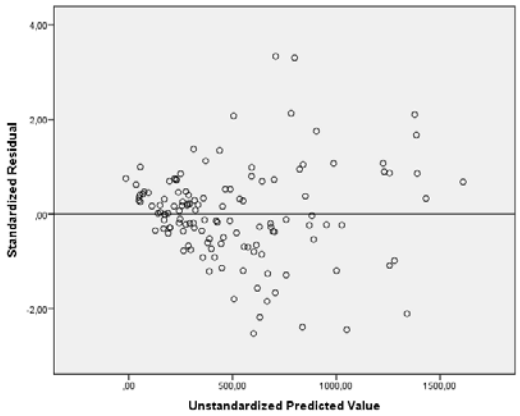
In this section the hypotheses of homoscedasticity, independence, normality of the residual random variables and the hypothesis of multicollinearity between the independent variables will be analyzed separately, with the techniques that were exposed in section 2.6.

3.4.1 Diagnostics of Homoscedasticity

The homoscedasticity verifies the assumption that the errors must have constant variance. The processes to analyze this hypothesis correspond to the analysis of graphs a) and b) in figure 2.



a)



b)

Fig.2: Graphs of analysis of variance

Both graphs show that the residuals maintain an approximately constant amplitude with respect to the zero-horizontal axis, that is, they do not show increasing or decreasing trends, reason why the hypothesis of homoscedasticity is not rejected.

3.4.2 Diagnostics of Independence

In order to determine if the residues comply with their independence, it is necessary to verify that the value of the Darwin-Watson test is within the region of non-rejection of the null hypothesis (Ho), that is, that the residues are independent of each other.

Table 2. Verification of independence

k	n	D-W
3	128	1,513
Não rejeitar a Ho		
[1,638;2,362]		

Based on table 2 the value of the D-W test lies in the non-rejection zone of the null hypothesis. Thus, we conclude that there is no autocorrelation between the residues. Gageiro & Pestana [21], make reference to the importance of non-violation of this hypothesis, since the model parameters estimates appear more accurate, because the standard error of the regression has a lower value when there is autocorrelation, giving confidence intervals for Bi smaller than they really are.

3.4.3 Diagnostics of Normality

For this analysis, the four procedures mentioned in section 2.6.3 of the methodology to evaluate the normality hypothesis will be used.

Table 3 presents the Kolmogorov-Smirnov test with the Lilliefors correction and the Shapiro-Wilk test. Since the sample has a size of 128 data, it is relevant to the analysis of the K-S test (the Shapiro-Wilk test is parsed samples less than 30 data according to [20]). Therefore, it is possible to conclude that for a significance level of 5%, the residues do not present a normal behavior, since their significance is inferior to 0,05. But it is important to emphasize that the value is close to 0,05, reason why it is possible to be assumed that the residues present a normal behavior. However, to be able to have a more correct decision, the remaining two procedures will be analyzed.

Figure 3 and figure 4 show that the observations are arranged around the oblique and horizontal straight lines, indicating the non-violation of normality. These graphs also allow us to analyze observations that deviate from normality. This is the case of the municipalities identified with code 609 and 613, which corresponds to Miranda do Corvo and Penacova, respectively, they present a behavior that is distant from normality, which makes them candidates for outliers, this analysis will be done in the next section. In addition, to complete the analysis of normality, the histogram (figure 5) of the residues is presented with its trend line of a normal distribution to help the visualization of the distribution of the values.

Table 3. Test of Normality

	Kolmogorov - Smirnov			Shapiro - Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	0,08	128	0,045	0,968	128	0,004

a. Lilliefors Signifiance Correction

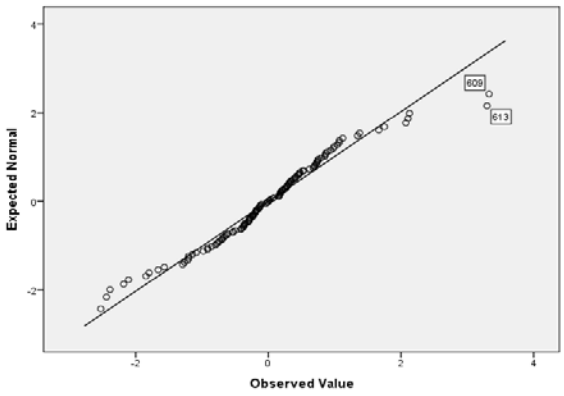


Fig.3: Normal Q-Q Plot of Standardized Residual

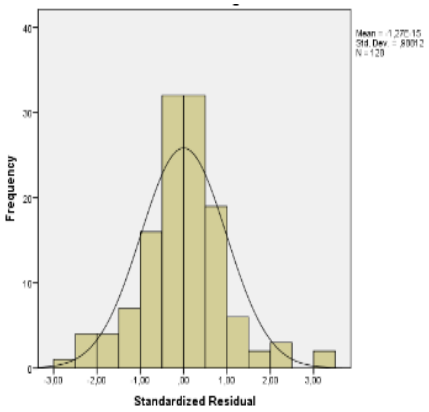


Fig.5: Histogram

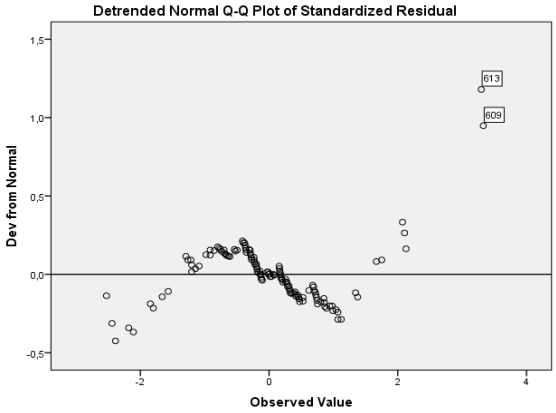


Fig.4: Detrended Normal Q-Q Plot of Standardized Residual

3.4.4 Diagnostics of Multicollinearity

The multiple linear regression model assumes that the independent variables are actually independent, which means they are not correlated. To verify this assumption, the correlation between the independent variables is analyzed through the matrix of model correlations and the tolerance/VIF.

The verification of multicollinearity through table 4 is a preliminary form of analysis. When the correlation coefficients between the independent variables are high (above 0,9), they indicate the possibility of multicollinearity.

Table 4 Correlations of model

		Number of trips by bus	Number of illiterate people	Population density	Purchasing power per capita
Pearson Correlation	Number of trips by bus	1,000	0,847	0,566	0,074
	Number of illiterate people	0,847	1,000	0,321	0,307
	Population density	0,566	0,321	1,000	-0,050
	Purchasing power per capita	0,074	0,307	-0,050	1,000
Sig. (1-tailed)	Number of trips by bus	-	0,000	0,000	0,202
	Number of illiterate people	0,000	-	0,000	0,000
	Population density	0,000	0,000	-	0,287
	Purchasing power per capita	0,202	0,000	0,287	-

Therefore, since the model does not present a correlation between independent variables with a value higher than 0,9, the highest correlation being 0,32, indicating that there should be no multicollinearity.

The second procedure is the analysis of table 5, whose values presented are tolerance and VIF (the two allow to define the same aspect), the tolerance measures the degree to which one variable X is explained by all other independent variables. If the tolerance value were very close to zero it would be

an indicator of the existence of multicollinearity. However, in the model under analysis there is no case with values close to zero.

Table 5 Collinearity Statistics

Coefficients			
Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	Number of illiterate people	1,000	1,000
2	(Constant)		
	Number of illiterate people	0,897	1,115
	Population density	0,897	1,115
3	(Constant)		
	Number of illiterate people	0,792	1,263
	Population density	0,872	1,146
	Purchasing power per capita	0,881	1,135

3.5 Analysis of Outliers and Influential Observations

Outliers are not detected exclusively by residue analysis but need other statistics for this purpose [14]. In this section, will identify the outliers between the independent variables and the cases that influence the estimated regression coefficients. Taking into account that the influential observations are those that individually, or together with the other observations, show to have more impact than the others in estimating estimators.

The statistics used to identify the outliers between the independent variables and those used to identify the influential cases were mentioned in the methodology section and will not be analyzed in this article, since they are statistical processes with a considerable extent that deserve another type of analyzes in other types of studies ([18] provides more information about this statistical analysis).

The main objective of waste analysis and influential values is to eliminate those cases, whose values are considered to be affecting the behavior of the estimated model with greater influence. Thus, it is necessary to carry out the analysis of each municipality that is identified and considered influential in the model, as well as the respective frequency in each part of the analysis.

In conclusion, soon after the implementation of the statistics to identify the outliers in the variables and the data with the greatest influence, we found 14 municipalities that will not be taken into account for the estimation of the new multiple linear regression model that aims to estimate the transport demand in the TBD in mainland Portugal

Table 6. Model summary of new model

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	0,860[a]	0,740	0,738	162,994	0,74	319,227	1	112	0,00	
2	0,916[b]	0,84	0,837	128,631	0,099	68,834	1	111	0,00	
3	0,933[c]	0,871	0,868	115,84	0,031	26,867	1	110	0,00	1,754

a. Predictors: (Constant), Number of illiterate people
b. Predictors: (Constant), Number of illiterate people, Population density
c. Predictors: (Constant), Number of illiterate people, Population density,Purchasing power per capita
c. Dependent Variable: Number of trips by bus

Table 7. New coefficients model

		Unstandardized		Standardized Coefficienst Beta			Correlations			Colinearity Statistics	
Model		B	Std. Error		t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-42,703	31,232		-1,367	0,174					
	Number of illiterate people	0,205	0,011	0,86	17,867	0	0,86	0,86	0,86	1,000	1,000
2	(Constant)	-112,325	26,037		-4,314	0					
	Number of illiterate people	0,173	0,01	0,727	17,62	0,000	0,86	0,858	0,67	0,848	1,179
	Population density	3,779	0,456	0,342	8,297	0,000	0,626	0,619	0,315	0,848	1,179
3	(Constant)	339,66	90,297		3,762	0,000					
	Number of illiterate people	0,185	0,009	0,78	20,239	0,000	0,86	0,888	0,693	0,789	1,268
	Population density	3,335	0,419	0,302	7,957	0,000	0,626	0,604	0,272	0,813	1,23
	Purchasing power per capita	-6,541	1,262	-0,185	-5,183	0,000	-0,06	-0,443	-0,177	0,919	1,088

a. Dependent Variable: Number of trips by bus

3.6 Estimation of the New Model

After having determined and eliminated those municipalities that were considered outliers or influential observations through the previous analysis, a new model of estimation of bus transportation demand for LDT will be presented, otherwise, it is a refinement of the model determined in equation (3) in section 3.3. The changes of the new model are based on the amount of data that was estimated (taking into account that the model of equation (3) was determined with 128 municipalities and the new model with 114, since the same variables will be maintained.

Tables 6 and 7 present the statistical information of the estimated model. It can be analyzed that there is a similarity with the old model (unrefined model), since the first variable to enter the model was "the number of people who do not have schooling ", This is because the correlation value of this variable with the number of trips by bus is 0,86, that is, the highest of all. The R-squared coefficient had a 3% increase over model 3 determined in section 3.3 (old model), that is to say, that 87% of the behavior of the variable is taken into account by the new model 3 ".

Largo [18], performs the verification of all the hypotheses underlying a multiple linear regression model for the new estimated model, having a conclusion that the exclusion of the 14 municipalities, does not affect the fulfillment of the hypotheses for the MLR, but the opposite made an increase in the quality of the fit of the model. Thus, the final model proposed for the estimation of bus travel to the low-density territories in mainland Portugal is given by:

$$VGA = 339{,}66 + 0{,}185 * SnE + 3.335 * DP - 6{,}541 * PdC \quad (4)$$

Where:
VGA: Number of trips by bus;
SnE: Number of illiterate people, with SnE varying between [645;6772];
DP: Population density, with DP varying between [6; 127];
PdC: Purchasing power per capita, with PdC varying between [56,08; 95,96].

4 Validation of the Demand Model

The validation consists of estimating the number of bus trips for each municipality using equation (4) and comparing these results with the actual number of bus trips obtained in Census 2011, determining for that purpose the percentage of error for each municipality.

Table 8 presents the distribution of the error percentage of the municipalities used for validation. It is important to note that 15% of the municipalities were used to evaluate the model, this percentage is equivalent to 23 LDT. The error intervals that were used to analyze the validation are defined with a 5% amplitude.

Based on the analysis of the behavior of the data presented in table 8, the estimated model presents an error of less than 40% in 83% of the municipalities, i.e. 19 TBD, and 91% of municipalities have an error of less than 55%. Based on the foregoing, it is possible to conclude that the MLR model presents a good fit for the territories of low population density in mainland Portugal.

Table 8. Frequencies by interval of error for validation sample (15% = 23 municipalities)

	Error Intervals (%)	Numbers of munipalities	Accumulated percentage (%)
Validation data	0 - 5	2	9%
	5 - 10	0	9%
	10 - 15	4	26%
	15 - 20	2	35%
	20 - 25	2	43%
	25 - 30	5	65%
	30 - 35	3	78%
	35 - 40	1	83%
	40 - 45	0	83%
	45 - 50	0	83%
	50 - 55	2	91%
	55 - 60	0	91%
	60 - 65	0	91%
	65 - 70	0	91%
	70 - 75	0	91%
	75 - 80	0	91%
	80 - 85	0	91%
	85 - 90	0	91%
	90 - 95	1	96%
	95 - 100	1	100%
	>100	0	100%

5 Results

Figure 6 shows the distribution of the number of municipalities by the error intervals. From the

figure, it can be seen that the majorities of the results show an error of less than 35%, i.e. a deviation from the value with respect to the actual value of less than 35%. However, some results show an error greater than 100%, and an isolated analysis is necessary to identify the possible cause of the error.

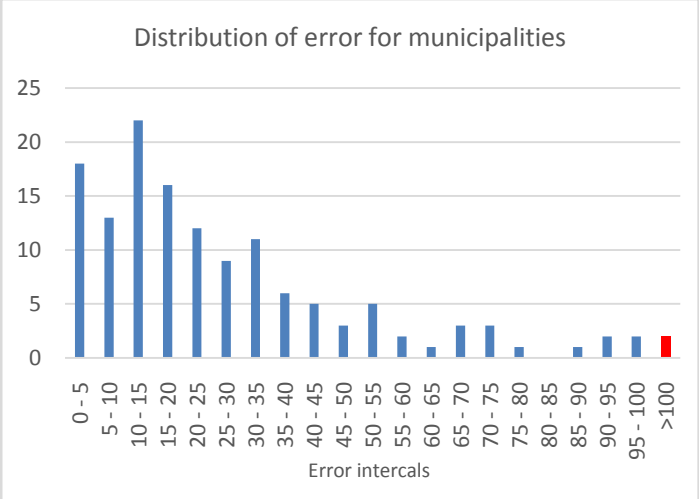


Fig.6: Distribution of the number of municipalities by interval of error

Table 9 allows analyzing in a more specific way the number of municipalities presented in the previous figure. Thus, 74% of the municipalities, which is equivalent to 101 municipalities, have a percentage error of less than 35%. This table also allows determining the critical line which corresponds to a percentage of error of less than 50% to which approximately 84% of the municipalities of the LDT, i.e., 115 municipalities.

Table 9. Frequencies by interval of error for modelled sample (85% = 137 municipalities)

	Error Intervals (%)	Numbers of munipalities	Accumulated percentage (%)
Portugal Continental	0 - 5	18	13%
	5 - 10	13	23%
	10 - 15	22	39%
	15 - 20	16	50%
	20 - 25	12	59%
	25 - 30	9	66%
	30 - 35	11	74%
	35 - 40	6	78%
	40 - 45	5	82%
	45 - 50	3	84%
	50 - 55	5	88%
	55 - 60	2	89%
	60 - 65	1	90%
	65 - 70	3	92%
	70 - 75	3	94%
	75 - 80	1	95%
	80 - 85	0	95%
	85 - 90	1	96%
	90 - 95	2	97%
	95 - 100	2	99%
	>100	2	100%

The aim of the present study is to perform a set of analyses of the results for the three regions of mainland Portugal (North, Center and South (Alentejo and Algarve)), in order to understand in which region the model presents the best estimates of the number of trips by bus, in order to be able to indicate in which region the MRLM could be used with greater reliability when developing transport planning.

Figure 7 was constructed based on the error distribution data by region. The analysis of the figure shows that the region with the highest amount of data with an error rate of less than 50% is the North region with 88% (ie, 36 municipalities), followed by the South region with 86% (equivalent to 38 municipalities) and the Central region with 79% of 41 municipalities. In addition, it can be observed that for an error of less than 25%, the Center region presents 63% of the municipalities, followed by the North with 61% and the South with 52%. In general terms, the region where the model of estimation of the number of journeys in buses presents a more adjusted estimate, that is, with less error is the North region.

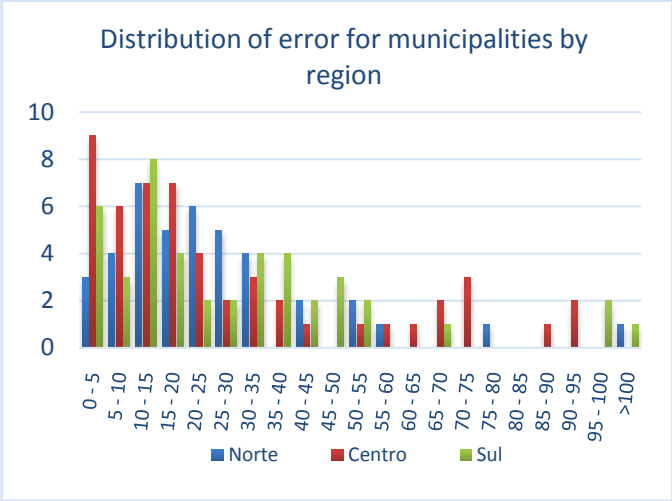


Fig.7: Distribution of error intervals by region

The box plot diagrams presented in figure 8 allow the comparison of the percent error distributions for each region, in addition to identifying the municipalities that are considered outliers (moderate - circles, severe - asterisk). As a

result, the Northern region has a more uniform behavior compared to the other regions, since 50% of the municipalities are in a much smaller error range than the other regions. On the other hand, the dispersion of the error in the North Region is much lower than in the other regions as it was perceived in the analysis to Figure 7. It should also be noted that the average error, considering the median, is very similar to the three regions with slight supremacy of the southern region.

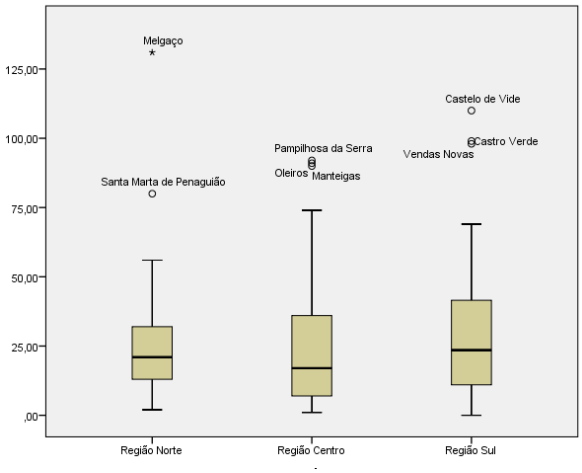


Fig.8: Box plot by regions

6 Conclusions

A transport demand model was estimated for the TBD of continental Portugal with data obtained in the 2011 censuses. This estimate was made at the level of municipalities counting on the classification of [1] determined to 165 municipalities with the characteristic of TBD. Based on [18], those variables that influence the demand for transport in the LDT were obtained.

The model was estimated with a total of 114 municipalities (equivalent to 69% of the total of the TBD), this number of data allows to ensure robustness to the model. The reduction in the number of municipalities was presented by the statistical analyses carried out throughout the case study, as how it was in the preliminary analysis of outliers for each variable (14 municipalities were discarded), the municipalities that were used for the validation process of the model (23 municipalities) and refinement of the model (14 municipalities).

With the variables, number of illiterate people, population density and purchasing power, the model presented an 87% adjustment. From the MLR model analysis it was possible to conclude that the variable with the greatest explanatory power, with the greatest influence, in the number of bus journeys in LDT is the number of people who do not have schooling, that is less educated, this population is usually related to lower rates of wealth and lower purchasing power using public transport as a means of transport in low-density territories. On the other hand, it should be noted that the variable purchasing power has a negative value for its coefficient in the MLR model, it is possible to conclude that an increase of this variable may lead to a decrease in the number of trips by bus in the municipalities of the LDT.

Analyzing the results of the MLR model by regions according to the NUTS II classification, it can be concluded that the region that presented the smallest deviations/error for the estimation of the number of people traveling by bus was the North region, since 84% of the municipalities presented an error of less than 50%, but in general terms the model adjusted well for the remaining regions where there are LDT, although in the Center region higher frequencies are found for error intervals greater than 50%.

References:

[1] M. Castro Almeida, “Alteração da deliberação relativa à classificação de territórios de baixa densidade para aplicação de medidas de diferenciação positiva dos territórios,” Lisboa, 2015.

[2] P.J.G. Ribeiro, F. Fonseca, and P. Santos, "Sustainability assessment of a bus system in a mid-sized municipality", Journal of Environmental Planning and Management, pp.1-21, doi:10.1080/09640568.2019.1577224, 2019.

[3] P.J.G. Ribeiro, L. A. P. J. Gonçalves, "Urban resilience: A conceptual framework", Sustainable Cities and Society, vol. 50, pp. 1 - 11, doi:10.1016/j.scs.2019.101625, 2019.

[4] A. Fernandes, J. F. Sousa, and M. Fonseca, “A Problemática da Mobilidade em Espaço Rural e Áreas de Baixa Densidade Urbana: o caso dos concelhos de Mértola e Ourique,” in Anais do I Congresso de Desenvolvimento Regional de Cabo Verde, 2009, pp. 2590–2617.

[5] L. Loureiro, E. Pereira, N. Costa, P. J. G. Ribeiro, and P. Arezes, "Global City: Index for Industry Sustainable Development", Advances in Intelligent Systems and Computing, pp. 294 - 302, doi:10.1007/978-3-319-60450-3_28, 2018.

[6] A. Silva, “Estratégias de eficiência colectiva em territórios de baixa densidade: reflexões a propósito do Minho-Lima e do Tâmega,” XII Colóquio Ibérico Geogr., pp. 3–8, 2010.
[7] A. E. D. Domingues, “Transporte Público em Territórios de Baixa Densidade-O Caso de Melgaço–Alto Minho.” 2009.
[8] C. Curtis and T. Perkins, “Travel Behaviour: A review of recent literature,” Perth, WA Urbanet, Curtin Univ. Technol., 2006.
[9] E. Arsenio, and P. J. G. Ribeiro, "The Economic Assessment of Health Benefits of Active Transport", Sustainable Urban Transport, vol. 7, pp. 1-22, doi:10.1108/S2044-994120150000007011, 2015.
[10] V. F. R. Teixeira, “Um modelo de procura de transporte público rodoviário de passageiros: aplicação a Portugal e à região do Algarve.” 2002.
[11] P.J.G. Ribeiro, D.S. Rodrigues, and E. Taniguchi, "Comparing standard and low-cost tools for gradient evaluation along potential cycling paths", WSEAS Transactions on Environment and Development, vol. 11, pp. 29-40, 2015.
[12] B. (Colombia). A. M. S. de T. y Transporte and W. F. C. Triana, Manual de planeación y diseño para la administración del tránsito y el transporte. Alcaldía Mayor, 2005.
[13] J. de Dios Ortuzar and L. G. Willumsen, Modelling transport. John Wiley & Sons, 2011.
[14] S. C. A. Rodrigues, “Modelo de regressão linear e suas aplicações.” Universidade da Beira Interior, 2012.
[15] J. de Dios Ortúzar, Modelos de demanda de transporte. Ediciones UC, 2012.
[16] U. Kattiyapornpong, “Understanding travel behavior using demographic and socioeconomic variables as travel constraints,” 2006.
[17] H. Largo and P. J. G. Ribeiro, “A bus demand model for Low-Density Territories in Continental Portugal,” Int. J. Transp. Syst., vol. 4, 2019.
[18] H. A. Amado, “Modelo de procura de transporte público em autocarro para territórios de baixa densidade,” University of Minho, 2018.
[19] J. A. Cavada Herrera, “Formulación y análisis de modelos de demanda agregada de validaciones y viajes de Transantiago,” 2014.
[20] M. H. Pestana and J. N. Gageiro, Descobrindo a regressão: com a complementaridade do SPSS. 2005.
[21] M. H. Pestana and J. N. Gageiro, “Análise de dados para ciências sociais: a complementaridade do SPSS,” 2008.
[22] P. J. G. Ribeiro, “Estudo de vias urbanas: Processo de selecção de indicadores ambientalmente sustentáveis de gestão de tráfego,” 2005.