ISBN 978-989-9121-31-7
2nd International Conference on Challenges in Engineering, Medical, Economics and Education: Research & Solutions (CEMEERS-23b)
Istanbul (Turkiye) Nov. 23-24, 2023

Explanatory Model for Bus Demand in The Municipality of Cascais in Portugal

Paulo Matos Martins[1], Stehanie Mendes[2]

[1]ISEL/Instituto Politécnico de Lisboa, Portugal, pauloj.matosmartins@isel.pt
ISEL/Instituto Politécnico de Lisboa, Portugal

Abstract: *Demand on the bus network serving the municipality of Cascais has undergone major variations in recent years due to Covid, the introduction of the Navegante and Viver Cascais passes, and other social and physical factors. The objective of this research is to help interpret and explain those changes. To this end, a tool was built that, based on data from the Cascais Próxima information system, semi-automatically allowed testing of a very large set of linear regressions with the aim of finding the best explanatory variables for variations in demand. Several limitations to be overcome in the future were identified, but very satisfactory results were also obtained that allow us to answer some of the questions raised: Covid brought an estimated loss of around 250 thousand passengers per month in the period between March 2020 and February 2023 or that school periods bring an increase of 200 thousand passengers, probably students. Many other results were found*

Keywords: *Public Transport Demand Analysis, Linear Regression, Data Analysis, Categorical Variables, Robustness;*

1. Introduction

This paper describes the research work developed in the second author's final master's thesis under the guidance of the first author. The research included a 4-month in-person internship at the municipal company Cascais Próxima, in the municipality of Cascais in the Lisbon region in Portugal. The internship took place in the mobility department at Cascais Próxima, focusing on the public bus network under the responsibility of the company and the municipality.

The objective of the research was to help interpret and explain the variations in demand for Public Transport in the municipality of Cascais, namely, the variations that existed during the periods affected by Covid-19, but not only.

The municipality of Cascais has around 215 thousand inhabitants, and its public transport network is made up of the municipality's municipal bus network and the urban trains that connect Cascais to Lisbon and Sintra, and the synergy between these modes is fundamental to guaranteeing efficiency and coverage of the transport network in the municipality. The bus network is made up of 44 lines. In 2022, around 560 thousand trips were made with an offer of around 410 million seatsxkm and 106.5 million passengersxkm transported, corresponding to an average occupancy rate of 26% (Cascais Próxima data).

Public transport services are available with appropriate frequencies, at peak times, and between peak times, and the fleet is equipped with modern and comfortable buses, including air conditioning, Wi-Fi, and good accessibility for people with reduced mobility.

https://doi.org/10.17758/DIRPUB15.F1123110

More than 9 million passengers are transported annually, with different types of transport tickets, including the Navegante pass and the local Viver Cascais pass, which is currently free for residents and employees in the municipality of Cascais.

Fig. 1 shows the geographical distribution of the 44 bus lines managed by Cascais Próxima.

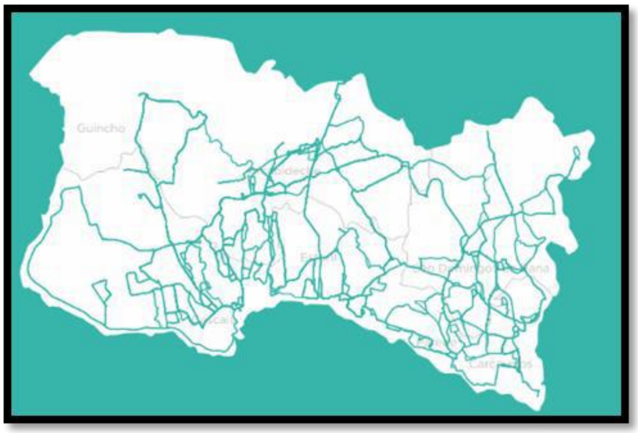


Fig. 1 Bus lines under the supervision of Cascais Próxima

Mobility in Cascais is based on the use of road modes, but these modes have to compete with the high-quality public transport offered by Cascais Próxima, which is complemented through sustainable parking management strategies and the provision of soft modes, such as bicycles and standing scooters for shorter routes.

Cascais Próxima has an advanced information system that allows the collection of numerous operational data from buses and other transport assets. They also have advanced descriptive analysis capabilities by using a multiple dashboard system developed in Power BI for the company. The origin of the data is diverse, but the main sources are:

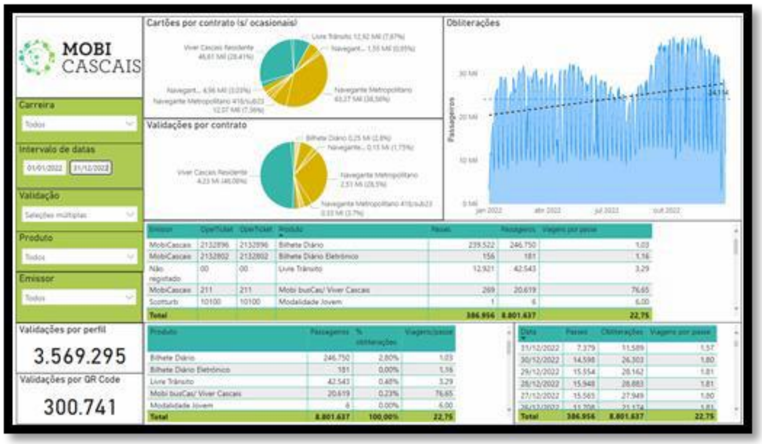
- Ticketing system control data (sales and users' validations)
- Geographic location data (according to the GTFS standard - General Transit Feed Specification)
- Fuel consumption control data per vehicle

The biggest gap in available information is the lack of data about the end of passenger journeys. The beginning of trips is referenced with the obliteration of the transport title (ticket, pass, etc.), but the bus exit is not, because the bus network is not a closed system.

The one-stop shop multiple dashboard system developed to support decision-making by the company is based on the following analysis dashboards:

- Passengers
- Obliterations
- Trips
- Monitoring of free transit
- Consumptions
- Maps
- Other indicators related to deviations and KPIs

As an example, Fig. 2 shows the obliterations dashboard for the year 2022.


MOBI CASCAIS
Cartões por contrato (s/ ocasionais)
Livre Trânsito 12,92 Mil (7,87%)
Navegant... 1,55 Mil (0,95%)
Viver Cascais Residente 46,61 Mil (28,41%)
Navegant... 4,96 Mil (3,03%)
Navegante Metropolitano 4|18/sub23 12,07 Mil (7,36%)
Navegante Metropolitano 63,27 Mil (38,56%)
Validações por contrato
Bilhete Diário 0,25 Mi (2,8%)
Navegante... 0,15 Mi (1,75%)
Viver Cascais Residente 4,23 Mi (48,00%)
Navegante Metropolitano 2,51 Mi (28,5%)
Navegante Metropolitano 4|18/sub23 0,33 Mi (3,7%)
Obliterações
Passageiros
30 Mil
20 Mil
10 Mil
0 Mil
jan 2022
abr 2022
jul 2022
out 2022

Carreira: Todos

Intervalo de datas: 01/01/2022 – 31/12/2022

Validação: Seleções múltiplas

Produto: Todos

Emissor: Todos

Validações por perfil: 3.569.295

Validações por QR Code: 300.741

Emissor	OperTicket	OperTicket	Produto	Passes	Passageiros	Viagens por passe
MobiCascais	2132896	2132896	Bilhete Diário	239.522	246.750	1,03
MobiCascais	2132802	2132802	Bilhete Diário Eletrónico	156	181	1,16
Não registado	00	00	Livre Trânsito	12.921	42.543	3,29
MobiCascais	211	211	Mobi busCas/ Viver Cascais	269	20.619	76,65
Scotturb	10100	10100	Modalidade Jovem	1	6	6,00
Total				386.956	8.801.637	22,75

Produto	Passageiros	% obliterações	Viagens/passe
Bilhete Diário	246.750	2,80%	1,03
Bilhete Diário Eletrónico	181	0,00%	1,16
Livre Trânsito	42.543	0,48%	3,29
Mobi busCas/ Viver Cascais	20.619	0,23%	76,65
Modalidade Jovem	6	0,00%	6,00
Total	8.801.637	100,00%	22,75

Data	Passes	Obliterações	Viagens por passe
31/12/2022	7.379	11.589	1,57
30/12/2022	14.598	26.303	1,80
29/12/2022	15.554	28.162	1,81
28/12/2022	15.948	28.883	1,81
27/12/2022	15.565	27.949	1,80
26/12/2022	11.708	21.174	1,81
Total	386.956	8.801.637	22,75

Fig. 2 Obliterations Dashboard - Cascais Próxima information system

The one-stop shop dashboard system designed by Cascais Próxima is an excellent system, but like all descriptive systems is based only on the analysis and synthesis of operational information. Despite allowing adequate operational support, it is limited when it is necessary to go deeper into analysis to identify upstream explanatory factors that are not directly related to the operation or when it is necessary to make predictions to support insights for the tactical and strategic planning of the PT.

In fact, these systems do not allow answering questions such as:

- What are the social, economic, seasonal, and physical factors that affect the demand for PTin Cascais?
- What real impact did COVID-19 have on reducing demand for PT?
- What influence did the introduction of the Navegante pass by AML and the Viver Cascais pass by the municipality have on demand?
- What is the effect of annual seasonality on demand variation?
- What is the effect of school seasonality on demand variation?

The aim of this research work was to contribute to answering these questions, using as a starting point the information contained in the Cascais Próxima system, by applying several (more aggregated or more disaggregated) regression models using the PT demand data available since 2019. It was the first test of applying this type of analysis to demand data in Cascais and therefore the regression models tested are not complex and are based on multiple dummy variables (1/0) that represent the presence or absence of the several induction effects studied.

To carry out the analyses in a timely manner, the second author used a tool for automating linear regression models developed by the first author in Excel using the VBA environment to be able to carry out multiple analyses including very different sets of lines on the fly, as well as carry out the stepwise process of refining the models (by choosing variables) inherent to the development of this type of model

2. Methods

2.1. Exploratory Analysis

Exploratory data analysis was originally introduced by American statistician John Turkey in the 1970s with the publication of his book Explanatory Data Analysis. In this publication, the author aimed to encourage statisticians to explore the data and formulate hypotheses, even before applying statistical tests, as he believed that through this prior analysis of the data, one could better understand the data and even induce the formulation of more hypotheses [1].

According to Bruce&Bruce [2], exploratory data analysis aims to ensure that the data scientist has prior knowledge of the data, by using visual tools, summary measures, and the formulation of other tests of hypotheses that can be considered during graphical analysis.

In the present work, in the first phase, still during the in-person internship, the descriptive statistics that Cascais Próxima's information system provided were studied. Subsequently, additional descriptive analysis was carried out to identify the relevant variables to be included in the regression models and to be able to characterize and cluster lines whose demand had similar or disparate behaviors.

As part of the exploratory data analysis, the following procedures were carried out:

1. Identification of potentially relevant independent variables.
2. Identification and treatment of outliers.
3. Identification of relevant representative measures.

According to Morettin & Bussab [3] variables can be classified into two groups, qualitative (categorical) and quantitative (metric) variables. Quantitative variables are those that can be measured on a numerical scale and express a quantity. Qualitative variables are those that present some researched attribute, that is, a characteristic.

According to Fávero & Belfiore [4], these variables can be divided into distinct groups or categories, being classified according to the number of categories they present. Dichotomous or dummy variables are binary variables that assume two characteristics. This is the case of the regression model used in which the qualitative variables represent the presence or absence of an attribute. For example, if we are in school time or not. For these variables to be included in a statistical calculation, they must be encoded in a numerical format.

The need for this numerical transformation is explained by Missio & Jacobi [5] as a method of quantifying attributes so that they can be analyzed statistically and is a technique used in linear regression models to ensure that categorical variables can be used in the model. Typically, dummy coding involves creating a binary variable that takes the value of 1 when the condition is present and the value of 0 when the condition is absent.

2.2. Linear Regression

Regression models are mathematical representations used to describe the relationship between two or more quantitative variables. These models aim to determine the equation that best represents the relationship between the variables, and their development process can be divided into three main phases [6]:

- Obtaining coefficient estimates to adjust the equation.
- Application of significance tests to regression and coefficients.
- Calculation of confidence intervals.

In this work, the focus will be on the multiple linear regression model that, as defined by Makridakis & Wheelwright [7], is a special form of regression in which the variable to be predicted depends in a linear way on two or more explanatory variables.

$$Y = \sum \alpha_i \cdot X_i(0/1) + \beta \tag{1}$$

For Morettin & Bussab [3], parameter estimates can be explanatory or predictive in nature. The ones of an explanatory nature are those that demonstrate a strong mathematical relationship but do not optimize the prediction of the cause-effect relationship. Predictive models are focused on obtaining the most reliable predictions possible for the dependent variable, based on the value of future observations of X without trying to optimize the explanatory character of each variable.

Explanatory estimates aim to understand the relationship between the independent variable (Y) and the predictor variables (Xi), aiming to determine how the variation in the variable (Y) is explained by the predictor variables. These parameter estimates are used to explain the behavior of the response variable in terms of the predictor variables and to understand which of these variables have the greatest impact in explaining the independent variable.

In our study, the main objective is to identify the factors inherent to the variation in demand for road public transport, allowing a better perception of how they influence demand, so the focus will be on explanatory estimates. However, the model can also be used for some predictive purposes, despite not being optimized for such purposes.

To check the quality of the adjustment of the coefficients to the regression equation, some significance tests are used to validate the results as to whether they are statistically significant. Typically, the following tests are performed:

- Student's t-test.
- F test.
- Adjustment coefficient.

Additionally, all variables may or may not be interrelated, so it is necessary to always quantify the degree of association between them in advance, to guarantee a result with the highest level of reliability [6]. This test, also known as Pearson's correlation coefficient, has the function of measuring the degree of correlation as well as the direction of variation between variables. In simple regression, this analysis measures the degree of linear relationship between the two variables, in the case of multiple regression, it measures the degree of collinearity between the independent variable and the set of all other variables.

2.3. Automation Tool

The regression automation tool was built by the first author purposely to support the second author's research work, who was focused on obtaining concrete answers to the research questions already mentioned.

Although there are several pieces of software on the market (some quite sophisticated) that allow the identification and execution of the stepwise process of refining multiple regressions, the authors opted for this solution for three main reasons:

- None of the authors usually work with statistical analysis software, so it would be necessary to foresee a learning curve that would condition the proposed objectives.
- Most importantly, the need to be able to have a tool that could very easily generate dozens of regressions with discretionary variation, not only of the variables but also of the data sets used (sets of PT lines).
- And finally, the fact that this experimental tool can be used to leverage new, more sophisticated research from an analytical point of view, being always available to be customized for the development of future projects.

The development of the regression automation tool was preceded by the creation of a data model developed in Excel in which the raw data exported from the Cascais Próxima information system was worked on in Power Query and Power Pivot (see Fig. 3) to be later made available to the automation tool in the form of a Pivot Table with monthly demand on all TP lines available.

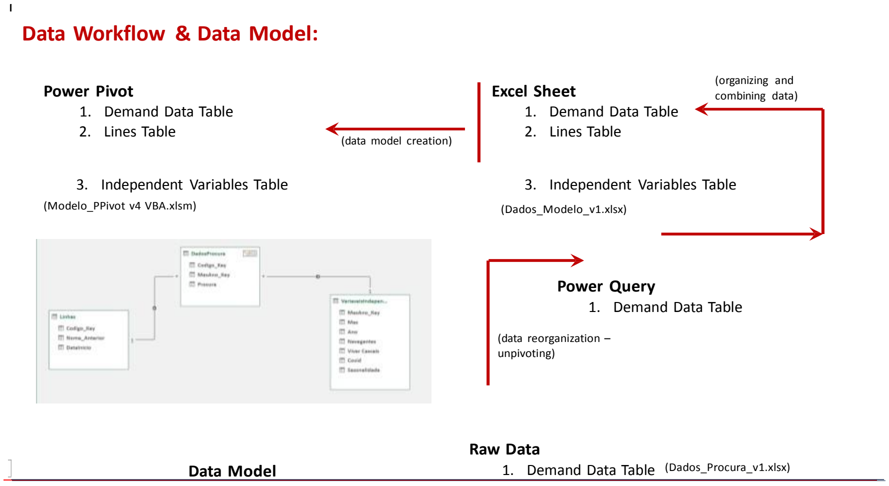


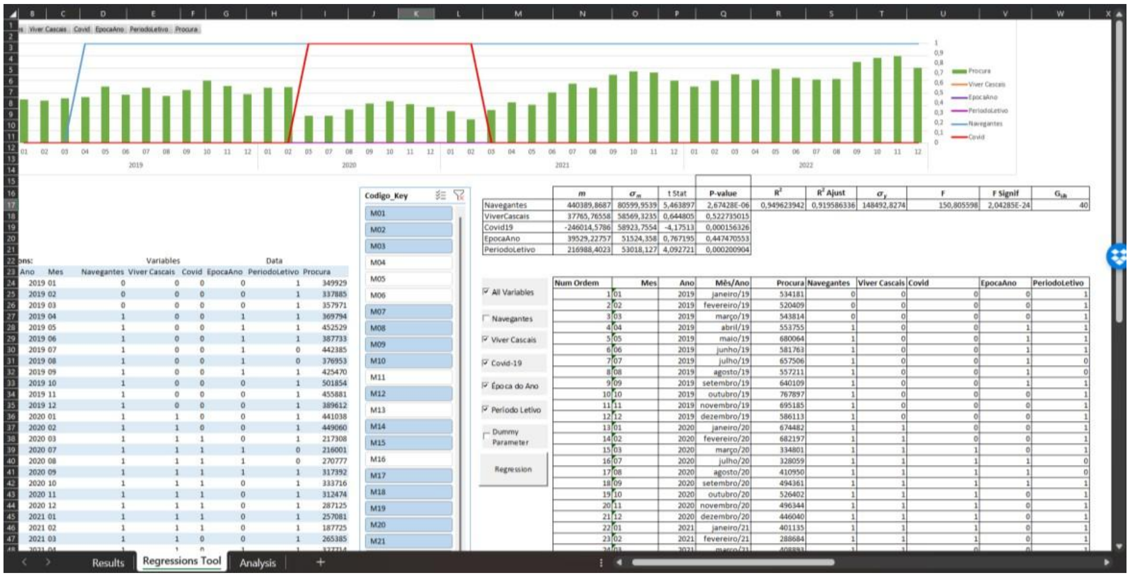
Fig. 3 Data Workflow and Data Model for the regression analysis.

The great advantage of this data model for processing and analysis using Excel is that it is relatively simple to execute and is always functional, as long as the data is updated. There are two large groups of data to update. The demand data generated by the Cascais Próxima PT information system and the data associated with the categorical variables which must be included/updated by the analyst in a separate table.

As can be seen on the left side of Fig. 4, the data generated by the model is represented. On the right side, the analyst can choose the variables he wants to test and just press the “Regression” button. Results, including all significance tests, appear instantly.

A graph is available at the top of the screen, which is essential for the analyst to be able to follow and understand the logic of the correlation between the independent variable (demand) and the categorical variables and between these explanatory variables.

The tool presented is an experimental prototype that may evolve by expanding its capabilities, namely, to include a dynamic number of variables to be inspected, which will allow the analyses to be expanded. Non-categorical variables may also be included, allowing quantification, for example, of growth in demand over time, without being associated with categorical factors.



	Ano	Mes	Navegantes	Viver Cascais	Covid	EpocaAno	PeriodoLetivo	Procura
24	2019	01	0	0	0	0	1	349929
25	2019	02	0	0	0	0	1	337885
26	2019	03	0	0	0	0	1	357971
27	2019	04	1	0	0	1	1	369794
28	2019	05	1	0	0	1	1	452529
29	2019	06	1	0	0	1	1	387733
30	2019	07	1	0	0	1	0	442385
31	2019	08	1	0	0	1	0	376953
32	2019	09	1	0	0	1	1	425470
33	2019	10	1	0	0	0	1	501854
34	2019	11	1	0	0	0	1	455881
35	2019	12	1	0	0	0	1	389612
36	2020	01	1	1	0	0	1	441038
37	2020	02	1	1	0	0	1	449060
38	2020	03	1	1	1	0	1	217308
39	2020	07	1	1	1	1	0	216001
40	2020	08	1	1	1	1	0	270777
41	2020	09	1	1	1	1	1	317392
42	2020	10	1	1	1	0	1	333716
43	2020	11	1	1	1	0	1	312474
44	2020	12	1	1	1	0	1	287125
45	2021	01	1	1	1	0	1	257081
46	2021	02	1	1	1	0	1	187725
47	2021	03	1	1	0	0	1	265385

	m	σ_m	t Stat	P-value	R^2	R^2 Ajust	σ_y	F	F Signif	G_{lib}
Navegantes	440389,8687	80599,9539	5,463897	2,67428E-06	0,949623942	0,919586336	148492,8274	150,805598	2,04285E-24	40
ViverCascais	37765,76558	58569,3235	0,644805	0,522735015						
Covid19	-246014,5786	58923,7554	-4,17513	0,000156326						
EpocaAno	39529,22757	51524,358	0,767195	0,447470553						
PeriodoLetivo	216988,4023	53018,127	4,092721	0,000200904						

Num Ordem	Mes	Ano	Mês/Ano	Procura	Navegantes	Viver Cascais	Covid	EpocaAno	PeriodoLetivo
1	01	2019	janeiro/19	534181	0	0	0	0	1
2	02	2019	fevereiro/19	520409	0	0	0	0	1
3	03	2019	março/19	543814	0	0	0	0	1
4	04	2019	abril/19	553755	1	0	0	1	1
5	05	2019	maio/19	680064	1	0	0	1	1
6	06	2019	junho/19	581763	1	0	0	1	1
7	07	2019	julho/19	657506	1	0	0	1	0
8	08	2019	agosto/19	557211	1	0	0	1	0
9	09	2019	setembro/19	640109	1	0	0	1	1
10	10	2019	outubro/19	767897	1	0	0	0	1
11	11	2019	novembro/19	695185	1	0	0	0	1
12	12	2019	dezembro/19	586113	1	0	0	0	1
13	01	2020	janeiro/20	674482	1	1	0	0	1
14	02	2020	fevereiro/20	682197	1	1	0	0	1
15	03	2020	março/20	334801	1	1	1	0	1
16	07	2020	julho/20	328059	1	1	1	1	0
17	08	2020	agosto/20	410950	1	1	1	1	0
18	09	2020	setembro/20	494361	1	1	1	1	1
19	10	2020	outubro/20	526402	1	1	1	0	1
20	11	2020	novembro/20	496344	1	1	1	0	1
21	12	2020	dezembro/20	446040	1	1	1	0	1
22	01	2021	janeiro/21	401135	1	1	1	0	1
23	02	2021	fevereiro/21	288684	1	1	1	0	1

Fig. 4 Front end of the Automation Tool for Linear Regressions with Multiple Categorical Variables.

3. Results

3.1. Exploratory Analysis

Initially, a systematic analysis of peaks and outliers was carried out. The main objective of identifying outliers was to detect data anomalies that would lead to their exclusion from regression analyses, or to correct the data by eliminating them when this was possible. The peaks, which we can define as “small outliers with a logical explanation” were also identified and may or may not be excluded from the regressions (the peaks may also give rise to new explanatory variables in a second phase of refinement of the regression models). An example of an outlier and a peak is presented in the appendix.

A careful analysis was then carried out to look for the existence of various types of seasonality in the data. This analysis resulted in the expected observation of a reduction in demand on weekends, but it also verified the existence of an annual seasonality associated with the four seasons of the year and seasonality associated with school periods. These seasonalities are not similar for all PT lines. There are lines in which school seasonality is felt more strongly than others.

Additionally, the exploratory analysis made it possible to identify a broad set of categorical independent variables that could explain demand behavior. The analysis of some of these variables had been previously suggested by Cascais Próxima itself, others were identified in the exploratory analysis. Of the variables identified, some were not used due to the lack of information for their characterization (such as the existence of rainy days, or cultural or sporting events), but others were modeled. The fact that the tool developed does not allow dynamic variation in the number of variables to be tested in the regressions was also a limiting factor in the choices made.

As a result of the exploratory analysis, it was decided to use the following independent variables in the building process of the regression model:

- Covid – in which the months were divided into Covid months and non-Covid months.
- Navegante – a variable that characterizes the inclusion of the Navegante pass in Cascais PT lines.
- Viver Cascais – a variable that characterizes the inclusion of the Viver Cascais pass in the PT lines.
- School Period – a variable that characterizes the months in which there are classes or no classes (holidays).
- Time of Year – a variable that characterizes the year into two periods, the first being associated with Summer and good weather and the second associated with Winter and cold weather.

https://doi.org/10.17758/DIRPUB15.F1123110

- New Lines – a variable that emerged during the creation of the regression models and that characterizes the lines that underwent changes in their layout (expansions or contractions) in the middle of the analysis period, due to a restructuring of the PT offer that was carried out at Cascais.

3.2. Linear Regression

Based on the previous variables, various types of regressions were carried out, depending on the groupings of services included in them. The regressions were grouped by "families", namely:

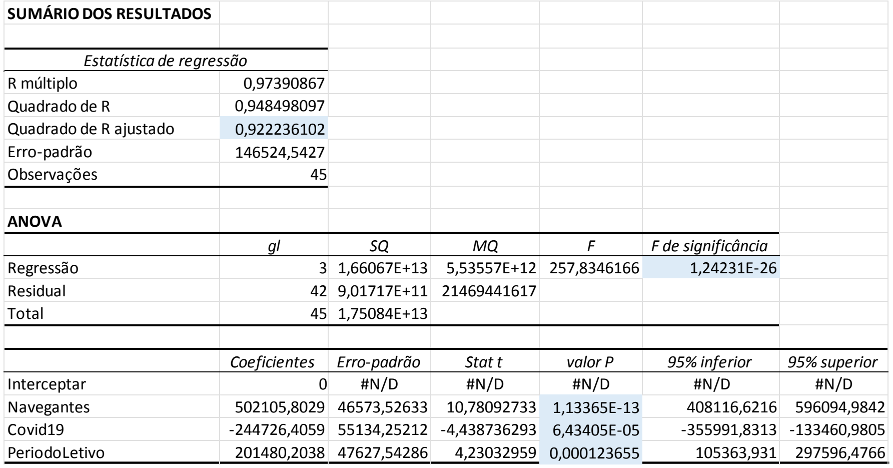
- Regressions A: the most comprehensive regressions possible, including the 44 PT lines or almost.
- Regressions B: developed according to the geographic region in Cascais.
- Regression C: depending on the typology of the lines.
- Regression D: a specific case of lines close to the NOVA SBE college, due to the strong impact that the student population of this university has in Cascais.

Almost all the analyses carried out allowed us to obtain robust and significant results, with the relevant variables (significant) not always being the same. The results of regression A including all TP lines, and a more specific case, of type C regression, to study the inclusion of the new lines M10 and M17, from May 2021, are presented. These lines are in some way competing with the previously existing line M03. This means they remove demand from it and all three lines must be analyzed together to understand the global impact. A more detailed analysis of all the results obtained can be found in the second author's final master's work [8].

The analysis procedure for all regressions will imply the following steps:

- Linear regression with all variables, for the chosen lines sample.
- Analysis of the regression parameters, such as the multiple correlation coefficient, analysis of variance, and significance analysis for each variable.
- Repeating the linear regression only with the variables that showed significance in this first analysis.
- Confirm analytical values with the graphical analysis.
- Obtain and assess the results.

The final results and graphs of the significant variables for the case of regression A are presented in the next paragraphs. Of the various variables analyzed, the regression obtained in the stepwise process with greater explanatory capacity includes the Navegante pass, the effect of Covid 19, and School Periods, or that is school seasonality. The adjusted R2 has a value of 0.922 and the F test of significance presents a p=1.242^-26. All variables are statistically significant.



SUMÁRIO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,97390867
Quadrado de R	0,948498097
Quadrado de R ajustado	0,922236102
Erro-padrão	146524,5427
Observações	45

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	3	1,66067E+13	5,53557E+12	257,8346166	1,24231E-26
Residual	42	9,01717E+11	21469441617		
Total	45	1,75084E+13			

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	0	#N/D	#N/D	#N/D	#N/D	#N/D
Navegantes	502105,8029	46573,52633	10,78092733	1,13365E-13	408116,6216	596094,9842
Covid19	-244726,4059	55134,25212	-4,438736293	6,43405E-05	-355991,8313	-133460,9805
PeriodoLetivo	201480,2038	47627,54286	4,23032959	0,000123655	105363,931	297596,4766

Fig. 5 Regression results for the 44 PT lines for Cascais

The graphical analysis of the evolution of demand and its "contrast" with categorical variables allows us to see that their explanatory power is quite relevant.

Although the Navegante variable has a good statistical performance, as we can see in Fig. 6, its explanatory effect is not clear, since the data existing before its inclusion only relates to 2 months and is therefore not representative.

https://doi.org/10.17758/DIRPUB15.F1123110

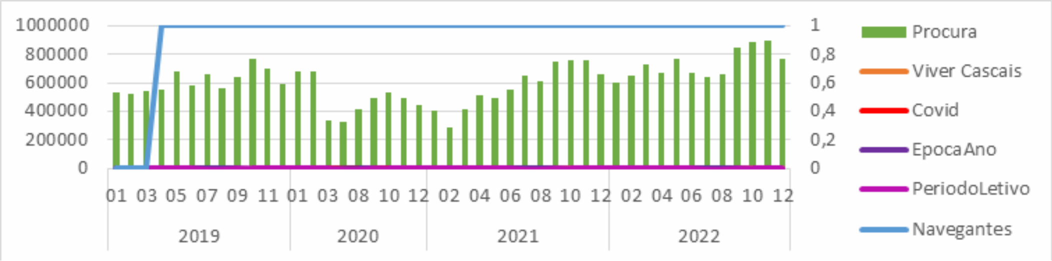


Fig. 6 Comparison of Demand versus Navegante variable

The Covid effect is shown in the respective graph (Fig. 7), so the quantification obtained by the model could be relevant.

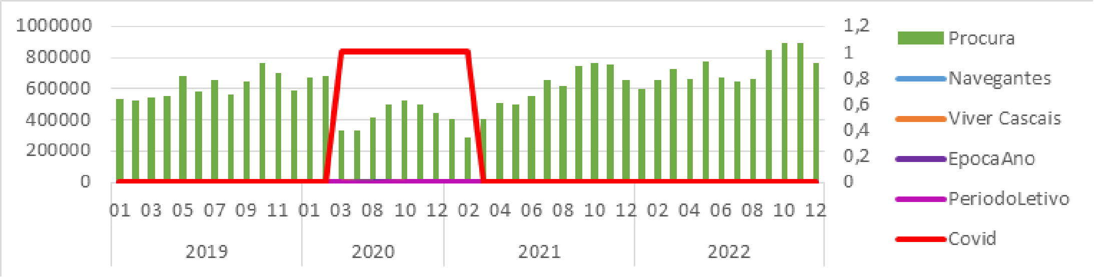


Fig. 7 Comparison of Demand versus Covid variable

The same applies to the variable associated with the functioning of the school period. As we can see, there is some relationship between this and the data, as shown in Fig. 8.

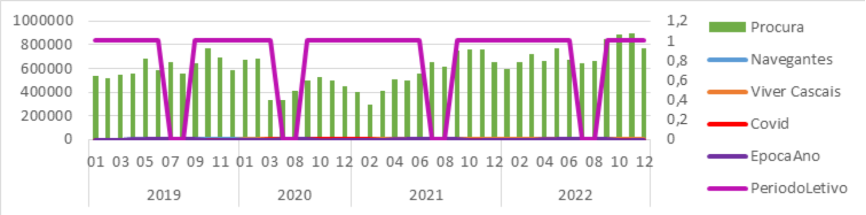


Fig. 8 Comparison of Demand versus Navegante variable

We also present the final results and graphs of the significant variables for the case of the previously existing line M03 and the new lines M10 and M17, whose exploration began in May 2021.

This example illustrates a disaggregated simulation with just three lines. It was a typical case in which the malfunction of the model's global specification led to the introduction of a new explanatory variable related to the new lines. So, it was necessary to introduce the variable “New Lines” to allow quantifying the increase in demand across the three lines together. The map of the lines is shown in Fig. 9.

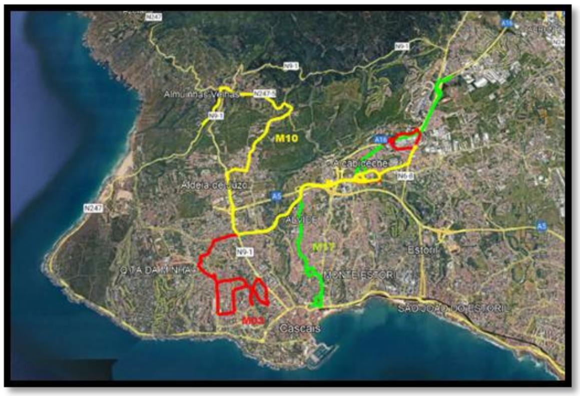
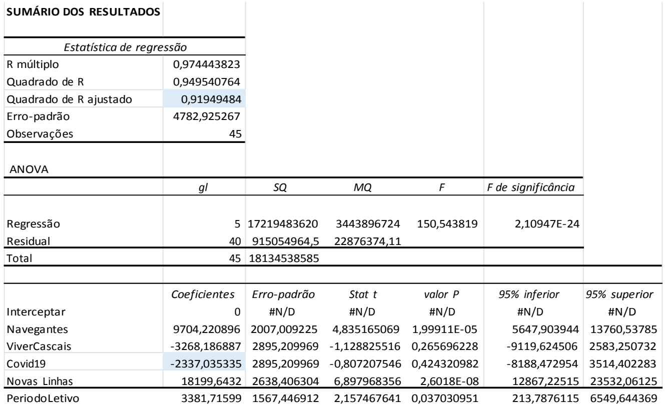


Fig. 9 Comparison of Demand versus Navegante variable

Fig. 10 shows the result of the first regression (with all the variables) carried out with the inclusion of the variable “New Lines”. This regression was later refined by excluding ViverCascais variable.



SUMÁRIO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,974443823
Quadrado de R	0,949540764
Quadrado de R ajustado	0,91949484
Erro-padrão	4782,925267
Observações	45

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	5	17219483620	3443896724	150,543819	2,10947E-24
Residual	40	915054964,5	22876374,11		
Total	45	18134538585			

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	0	#N/D	#N/D	#N/D	#N/D	#N/D
Navegantes	9704,220896	2007,009225	4,835165069	1,99911E-05	5647,903944	13760,53785
ViverCascais	-3268,186887	2895,209969	-1,128825516	0,265696228	-9119,624506	2583,250732
Covid19	-2337,035335	2895,209969	-0,807207546	0,424320982	-8188,472954	3514,402283
Novas Linhas	18199,6432	2638,406304	6,897968356	2,6018E-08	12867,22515	23532,06125
PeriodoLetivo	3381,71599	1567,446912	2,157467641	0,037030951	213,7876115	6549,644369

Fig. 10 Regression results for the M03, M10 and M17 analysis

The graph in Fig. 11 shows the increase in demand generated with the inclusion of the two new lines M10 and M17. This increase in demand involves the effect of transferring passengers from line M03 to the new lines. The explanatory value of the coefficient corresponds to an increase of around 18,200 monthly passengers on all three lines.

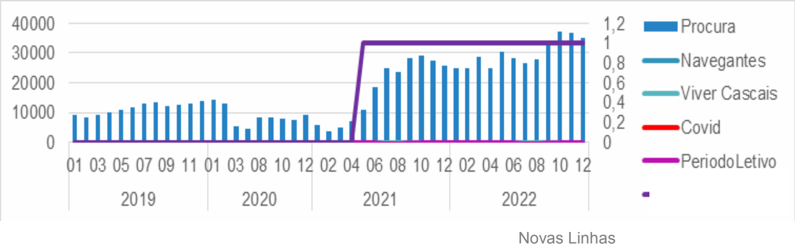


Fig. 11 Comparison of Demand versus NovasLinhas variable

4. Discussion

The results obtained in this research are very satisfactory and are the first evidence of the success of the proposed methodology. This approach can bring additional insights to demand analysis in relation to those obtained from the Cascais Próxima information system dashboards.

The approach now proposed allows to explore the in-depth aspects of the analysis and to understand and quantify more deeply the social and physical drivers of bus demand in Cascais.

The proposed analysis can be done globally, or locally, as the results presented illustrate. It is not possible to present and discuss all the results and evidence obtained during the research in this reduced format, so we will carry out a summary analysis of the results presented here, giving some answers to the questions initially formulated.

The global model identified a monthly loss of around 250 thousand passengers per month in the period between March 2020 and February 2023, due to the presence of Covid. Schools working in the municipality, with classes, imply an overall increase in monthly demand of around 200 thousand passengers, probably students. Finally, the introduction of the Navegante pass explains, in the regression, a share of around 500 thousand monthly users. However, the authors have doubts about the validity of this variable because the period analyzed began in January 2019 and the Navegante pass was introduced in March 2019. There is no previous historical data that allowing to validate the demand changes. What the variable seems to be explaining is the existing base demand with the new demand for the Navegante added. Trying to separate these values would only be possible with data from at least 2018.

https://doi.org/10.17758/DIRPUB15.F1123110

In relation to the disaggregated analysis presented for lines M03+(M10+M17), it suggests that the Navegante variable is responsible for around 10 thousand passengers per month (maintaining the effect explained before), the Covid variable was responsible for the loss of around 2.5 thousand passengers on the lines (in the Covid period), the schools explain around 3.5 thousand new passengers per month (in school periods) and the variable that represents the introduction of the new lines represents an increase in demand of around 18 thousand new passengers.

All analyses and variables presented were significant. There were some serious correlation problems between some variables, such as the Navegante and Viver Cascais passes (whose analysis had been suggested by Cascais Próxima) which did not allow conclusions to be drawn about the relationship between them.

Finally, there must be interest in continuing to deepen the exploratory analyses presented in this first work, namely through improving the collections of available raw data, but also through refining and greater automation of the analysis tool developed.

5. Acknowledgment

The authors would like to thank the municipal company Cascais Próxima for all the support provided during the completion of S. M. Authors' final master's thesis, particularly for providing a 4-month in-person internship, which was very enriching for this author, and for giving the authors access to the data necessary for this research work.

6. Appendix

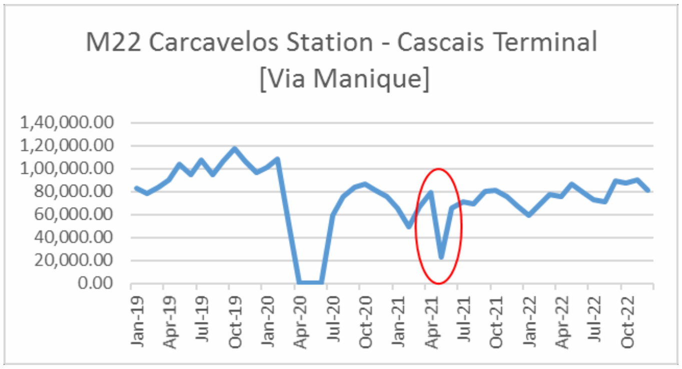


Fig. 12 Example of an outlier detected on Line M22 which was corrected.

The drop in demand due to Covid-19 was not considered an outlier because it was simulated by including a specific variable for that purpose.

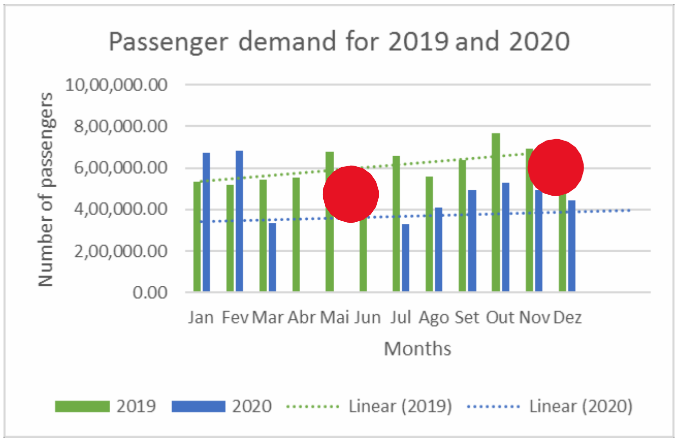


Fig. 13 Monthly passenger demand in 2019 and 2020.

7. References

[1] Turkey, J. (1977). Explanatory Data Analysis

[2] Bruce, P., & Bruce, A. (2019). Estatística Prática para Cientistas de Dados.

[3] Morettin, P. A., & Bussab, W. O. (2017). Estatística Básica

[4] Fávero, L. P., & Belfiore, P. (2017). Manual de Análise de Dados -Estatística e Modelagem Multivariada com Excel, SPSS e Stata. In Elsevier. http://dergipark.gov.tr/cumusosbil/issue/4345/59412

[5] Missio, F., & Jacobi, L. F. (2007). Variáveis dummy: especificações de modelos com parâmetros variáveis. Ciência e Natura, 29(1), 111–135.

[6] Daniels, L., & Minot, N. (2018). Introduction to Statistics and Data Analysis. SAGE. https://doi.org/10.31399/asm.hb.v08.a0009212

[7] Makridakis S, Wheelwright SC, H. R (1997). Forecasting methods and applications.

[8] Mendes, S., (2023). Modelo explicativo da procura de autocarros no Concelho de Cascais. Trabalho Final de Mestrado, ISEL!IPL.