INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Civil

MODELO EXPLICATIVO DO COMPORTAMENTO DA PROCURA DO TRANSPORTE PÚBLICO NO MUNICÍPIO DE CASCAIS ENTRE 2000 E 2022

STEPHANIE LORRAINE CARVALHO MENDES

(Bacharel em Engenharia Civil)

Relatório de Estágio para obtenção do grau de Mestre em Engenharia Civil na Área de Especialização em Vias de Comunicação e Transportes.

Orientador:

Doutor Paulo Matos Martins

Júri:

Presidente:

Doutora Paula Raquel Pires Da Cunha Lamego

Vogais:

Doutora Carmem De Jesus Geraldo Carvalheira

Doutor Paulo Matos Martins

Dezembro de 2023

Agradecimentos

Em primeiro lugar eu agradeço a todas as boas energias que me guiaram até aqui. Agradeço por me terem guiado e protegido neste momento tão importante em minha vida.

Agradeço também a minha amada e querida família, pois sem o apoio deles nada disso seria possível, agradeço ao apoio financeiro e emocional que sempre me deram.

Agradeço ao ISEL, sendo o meio que me possibilitou chegar até aqui, agradeço imensamente o Coordenador do curso Paulo Mendes, por ser tão solicito e exercer de maneira muito ética e responsável sua profissão.

Agradeço ao meu querido orientador que foi o responsável por me abrir os caminhos para trilhar esse novo caminho dos transportes, por ter paciência durante a orientação, por ser sempre muito solicito, educado e o parabenizo também por sua inteligência ímpar.

Agradeço a Cascais Próxima e ao Eng.º João tojal por terem me dado essa incrível oportunidade de experienciar o mundo dos transportes de uma forma tão leve acolhedora, em um ambiente organizacional muito agradável de se trabalhar.

E por último, mas não menos importante, muito pelo contrário, agradeço a Stephanie Mendes, que conseguiu concluir uma etapa tão importante em sua vida, atravessou altos e baixos em outro país, com pessoas de culturas diferentes e longe da família, agradeço sua resiliência, esforço e dedicação.

RESUMO

Este documento contextualiza e descreve os principais desenvolvimentos efetuados durante os 4 meses de estágio que decorreu na empresa municipal Cascais Próxima no âmbito do TFM da autora, bem como os trabalhos de investigação subsequentes para o desenvolvimento de um modelo explicativo para a procura de transporte público rodoviário no município de Cascais.

Foi aplicada uma ferramenta estatística de regressão linear múltipla com variáveis categóricas que visa identificar as variáveis explicativas que mais afetam a procura pelo transporte público em Cascais e fazer a quantificação das mesmas. Nessa pesquisa buscou-se identificar as relações de causa e efeito entre as variáveis independentes e a procura (variável dependente), além da interpretação dos coeficientes e significância de cada uma delas.

A metodologia para a realização desse trabalho envolveu um período de integração na empresa, pesquisa científicas sobre como são feitas as análises estatísticas contextualizado a mobilidade dos transportes. Após essa etapa de preparação, sucedeu-se todo o processo que envolve a análise de dados, tais como recolha, organização e tratamento dos mesmos, seguida de sucessivos testes de significância a fim de atingir os melhores resultados.

Os resultados obtidos na presente pesquisa são muito satisfatórios e demonstram o sucesso dos desenvolvimentos propostos. Os autores conseguiram ir bastante mais além do que as sínteses de estatísticas descritivas atuais que os *dashboards* do sistema de informação da empresa Cascais Próxima permitem.

Claro que estes sistemas são fundamentais para o apoio corrente à operação e à tomada de decisão, mas a abordagem agora proposta permite explorar uma vertente de aprofundamento da análise que permita conhecer e quantificar mais profundamente os drivers sociais e físicos da procura de autocarros em Cascais, permitindo antever os efeitos na procura da sua variação.

Palavras-chave: Transporte Público, Regressão Linear, Modelo Explicativo, Análise de Dados, Cascais.

ABSTRACT

This document contextualizes and describes the main developments achieved during the 4-month internship at the municipal company Cascais Próxima as part of the author's Master's thesis, as well as the subsequent research work for the development of an explanatory model for public bus transportation demand in the municipality of Cascais.

A statistical tool of multiple linear regression with categorical variables was applied to identify the explanatory variables that most affect public transportation demand in Cascais and to quantify them. This research aimed to identify the cause-and-effect relationships between independent variables and demand (dependent variable), as well as the interpretation of the coefficients and their significance.

The methodology for conducting this work involved an immersion period in the company, scientific research on how statistical analyses are conducted in the context of transportation mobility. After this preparation phase, the entire process involving data analysis, including data collection, organization, and processing, followed by successive significance tests, was carried out to achieve the best results.

The results obtained in this research are highly satisfactory and demonstrate the success of the proposed developments. The authors could go much further than the current descriptive statistics summaries provided by the information system dashboards of the Cascais Próxima company.

While these systems are essential for ongoing operational support and decision-making, the approach now proposed allows for a deeper analysis that enables a more profound understanding and quantification of the social and physical drivers of bus demand in Cascais, thereby anticipating the effects of its variations on demand.

.

Key Words: Public Transportation, Linear Regression, Explanatory Model, Data Analysis, Cascais.

INDICE GERAL

IN**DICE DE GRÁFICOS**

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

AML – *Área Metropolitana de Lisboa*

AMT – *Autoridade da Mobilidade e dos Transportes*

AED – *Análise Exploratória de Dados*

CCDRN – *Comissão de Coordenação e Desenvolvimento Regional do Norte*

DAX - *Data Analysis Expressions*

KPI – *Key Performance Indicator*

OMS – *Organização Mundial da Saúde*

TP – *Transporte Público*

CP – *Comboios de Portugal*

1 Introdução

1.1 Enquadramento

Este documento relata os principais desenvolvimentos efetuados no Trabalho Final de Mestrado (TFM) da aluna Stephanie Lorraine Carvalho Mendes, número 47860, do Mestrado de Engenharia Civil na especialização no ramo de Vias de Comunicação e Transportes, realizado sob a orientação do Professor Paulo Matos Martins.

O TFM em questão enquadra-se na modalidade de estágio e foi realizado na Empresa Municipal Cascais Próxima, no Departamento de Gestão de Mobilidade, Espaços Urbanos e Energias, E.M., S.A, teve a duração de quatro meses, e foi desenvolvido sob a tutoria do Engº João Pedro Tojal Silva, chefe da Divisão de Gestão de Operações no Departamento de Transportes da empresa.

1.2 Objetivos

Tratando-se de um relatório de estágio realizado numa empresa responsável pela gestão do transporte público no município de Cascais, o objetivo geral deste trabalho é o desenvolvimento de uma ferramenta inovadora que permita, usando técnicas de regressão linear múltipla, determinar quais são as variáveis relevantes para o estudo da variabilidade da procura do TP e suas influências.

Além do objetivo geral, este trabalho possui objetivos específicos ligados ao interesse da autora na exploração de conhecimentos matemáticos, relacionados a área de estatística e a análise de dados no âmbito da gestão da mobilidade.

Nesse aspecto, destacam-se os seguintes objetivos específicos:

- Aprofundar o conhecimento dos procedimentos e metodologias atualmente utilizados para análise de dados;
- Desenvolver e aprimorar as capacidades analíticas e lógicas;
- Entender como se dá o processo de mineração e tratamento de dados;
- Estudar algumas das principais ferramentas de análise de dados;

- Desenvolver a capacidade de expandir qualquer tipo de análise através de conhecimentos estatísticos.

Por fim, a autora buscou aplicar os conceitos teóricos do seu mestrado em um contexto corporativo, trabalhando com profissionais experientes na área dos transportes e da mobilidade, visando o aprimoramento de *networking* e capacidade de resolver problemas.

1.3 Metodologia

A metodologia para desenvolvimento deste trabalho, primeiramente centrou-se num período dedicado a **pesquisas académicas** sobre os conteúdos relacionados com a gestão da mobilidade e com a estatística, nomeadamente os indicadores e as métricas que são utilizadas para a análise dos sistemas de transportes e as metodologias de cálculo para análise explicativa de variáveis. Neste período de pesquisa foram analisadas diversas fontes como livros e artigos científicos sobre ambas as temáticas.

Em seguida, com o início do estágio e consequente o processo de **integração na empresa**, iniciou-se um período dedicado ao conhecimento das técnicas utilizadas pela empresa para gestão da mobilidade, quais seus principais desafios, conquistas e os desenvolvimentos futuros previstos na empresa. Nesse período buscou-se uma relação bem próxima com os colaboradores e chefes da divisão para que se pudesse colher o máximo de informações possíveis, para começo das verificações.

A terceira fase consistiu no **estudo dos relatórios** já produzidos pela empresa nos últimos anos. Essa etapa consistiu na análise minuciosa dos gráficos e medidas resumo, de forma a identificar padrões. Além disso também foi estudado como são feitos os procedimentos internos de recolha e tratamento de dados.

A quarta fase consistiu na construção do modelo **regressão linear em Excel**, que além do processo de construção matemática do modelo, envolveu previamente toda a mineração e tratamento de dados, englobando a classificação das variáveis, determinação da tipologia dos dados, identificação de *outliers* para garantir a integridade e robustez dos resultados.

Na sequência, a quinta fase baseou-se no **cálculo das regressões** com a realização de vários testes de significância, testes com diferentes combinações de variáveis e construção de

gráficos para análise paralelas, permitindo uma extrapolação abrangente das relações Inter correlacionadas, até que se atingisse o resultado com maior confiabilidade para o modelo.

E por fim, a sexta a fase foi a **produção do relatório, com as principais conclusões da análise efetuada e a identificação de desenvolvimentos futuros a efetuar.**

1.4 Motivação de escolha do tema

Os motivos que levaram à escolha deste tema são de duas naturezas, a primeira é a afinidade com o tema, e a segunda devido a percepção de que a área da tecnologia, automação e análise dos dados está em grande crescimento em diversos campos profissionais em todo o mundo.

Para tanto, estudar técnicas que modelem os dados presentes e passados para gerar informações relevantes para a gestão de uma empresa ou de um setor, com possibilidade de se realizar análises profundas e comprovadas matematicamente pareceu muito interessante, promissora e de extrema relevância.

Através da unidade curricular Modelação e Análise dos Sistemas de Transportes, lecionadas pelo orientador deste trabalho pode-se ter uma base do que é estimar situações relacionadas com a mobilidade e os transportes, transformando dados em informações e valor acrescentado.

Além disso, a relevância internacional do tema também foi um fator fundamental para a escolha do mesmo. A possibilidade de poder expandir esse conhecimento em outras partes do mundo é outro fator de satisfação.

1.5 Importância do tema

Por meio da vivência dos utilizadores, e pelo que se nota na sociedade em geral, o a mobilidade proporcionada pelo transporte público (TP) enquanto procura derivada, facilita e contribui para o lazer, acesso a equipamentos de saúde, centros culturais e aperfeiçoamento profissional as pessoas. De um modo geral, o TP contribui para as atividades mais básicas do ser humano, sendo um meio totalmente indispensável para a vida em sociedade.

Baseado nessa observação, fica nítido que a utilização dos TPs é essencial para a garantir a vivência em sociedade, onde as deslocações por parte das pessoas é algo inevitável e na grande maioria dos casos, são realizadas em uma frequência muito elevada. Por esse motivo é necessário um TP de alta qualidade que garanta uma adequada qualidade de vida do utilizador nos seus processos de mobilidade.

Dessa maneira, uma análise do sistema, com o intuito de se obter informações sobre os padrões de procura dos passageiros nas mais variadas situações é de grande importância para assegurar a eficiência do transporte e consequentemente proporcionar qualidade de vida aos seus utilizadores. A título de exemplo apresentam-se algumas das premissas básicas que os transportes públicos têm que cumprir para que se garanta a qualidade de vida do passageiro, que foram listadas em conformidade com a (CCDRN, 2008)

- Garantia da frequência de passagem dos autocarros;
- Cumprimento de ciclos e horários;
- Conforto dentro dos veículos;
- Taxa de passageiros por Km confortável para os utilizadores;
- Rota bem elaborada de maneira a servir todas as áreas necessitadas da cidade;
- Títulos de transporte a preços acessíveis;
- Localização das paragens de autocarros;
- Oferta que atenda a procura.

Para que todos esses fatores sejam cumpridos são necessários estudos e análises minuciosas de todo o sistema de transporte, que em parte são obtidos através de análise de dados. Com essa análise e a consequente produção dos relatórios das informações obtidas, pode-se prever as necessidades da população em relação aos TPs, e promover mudanças no âmbito social que possam impactar de maneira positiva a procura pelo transporte público.

Portanto, com o desenvolvimento deste trabalho pretende-se dar um pequeno contributo para o desenvolvimento de ferramentas inovadoras para a análise dos dados da procura, que sejam de grande utilidade para a melhoria do sistema de transportes públicos, neste caso, de Cascais. O objetivo final será o de caminhar para uma melhoria geral da mobilidade e das sociedades, considerando que um transporte público eficiente faz com que os utilizadores possam optar pelo uso desse sistema com maior frequência, gerando a redução gradual do uso de transportes individuais e indiretamente aumentando a sustentabilidade ambiental e social das nossas sociedades.

1.6 Estrutura do trabalho

Apresenta-se a estrutura do relatório e faz-se uma breve descrição da temática abordada em cada um dos capítulos:

- **Capítulo 1 - Introdução:** neste capítulo é apresentado o enquadramento geral do tema, abordando quais são os objetivos a serem alcançados, qual a metodologia utilizada no desenvolvimento do trabalho, a motivação da autora na escolha no tema e qual a sua importância.
- **Capítulo 2 – O transporte público em Cascais:** Serão apresentadas as considerações básicas a respeito da gestão da mobilidade, juntamente com a apresentação da empresa e as atividades que foram desenvolvidas durante o período de estágio.
- **Capítulo 3 – Revisão bibliográfica:** Revisão bibliográfica sobre estatística aplicada e técnicas de análise de dados que servem como base teórica para desenvolvimento das análises apresentadas.
- **Capítulo 4 – Análise do comportamento da procura**: Nesse capítulo são apresentados todos os desenvolvimentos efetuados: os *insights obtidos*, as análises estatísticas, análises gráficas e toda a envolvente do tema enfatizando todas as fases do processo.
- **Capítulo 5 - Conclusões:** Nesse capítulo são descritas todas as conclusões do trabalho, e apresentadas ideias para possíveis trabalhos futuros.

2 O Transporte Público em Cascais

2.1 A Cascais Próxima

A **Cascais Próxima** é uma empresa municipal responsável pelo gerenciamento da mobilidade e sistemas de transportes no concelho de Cascais. A organização utiliza um modelo de mobilidade integrada com abrangência em diversos modos de transportes, tendo sob sua tutela a rede de autocarros, os parques de estacionamento, as estações de carregamento de veículos elétricos, e o aluguer de bicicletas.

A história da Cascais Próxima remonta a 2009, quando a Câmara Municipal de Cascais iniciou um projeto piloto de gestão de estacionamento. Em 2012, a empresa foi oficialmente criada para gerir o estacionamento público em todo o concelho de Cascais, desde então, a Cascais Próxima tem trabalhado para melhorar a mobilidade urbana em Cascais, através da oferta de serviços inovadores e sustentáveis. (Câmara Municipal de Cascais, n.d.).

A empresa é 100% detida pela Câmara Municipal de Cascais, e sua estrutura organizacional inclui uma equipa de gestão, operações e manutenção, e administrativa. A Cascais Próxima trabalha em colaboração com outras entidades do setor de transportes para melhorar a eficiência e a sustentabilidade da frota de autocarros.

No âmbito do desenvolvimento sustentável, a empresa ganhou prestígio por ser a pioneira em Portugal na detenção de autocarros movidos a hidrogénio, como pode ser observado na Figura 2-1, contribuindo para uma mobilidade sustentável, descarbonização e redução da emissão de gases provocadores do efeito estufa. (Câmara Municipal de Cascais, n.d.-b)

Figura 2-1 Autocarro movido a hidrogênio (Câmara Municipal de Cascais, n.d.-a)

Outro pioneirismo de Cascais foi a instauração do tarifário gratuito nos títulos de transporte para residentes, trabalhadores e estudantes no Concelho através da aquisição do cartão Viver Cascais, estratégia essa que incentiva a utilização do TP, gera economia para os utilizadores e uma redução do tráfego automóvel.

Além disso, a Cascais Próxima tem investido em tecnologias avançadas para melhorar a experiência dos passageiros, como a disponibilização de informações em tempo real sobre o serviço de transporte público, bilhética eletrónica e sistemas de acompanhamento e monitorização dos autocarros em tempo real, através do MobiCascais. (Câmara Municipal de Cascais, 2020)

2.1.1 Atividades desenvolvidas no âmbito do estágio

Durante o período de execução do estágio, como forma de integração na empresa, foram realizadas várias atividades de maneira sequencial para que a estagiária pudesse obter uma compreensão mais analítica e aprofundada sobre o sistema de gestão do transporte público da empresa.

A primeira tarefa efetuada consistiu na análise cuidadosa e detalhada dos relatórios de caracterização do transporte público produzidos pela empresa, os dashboards. Através dessa análise foi possível uma compreensão prévia de quais as principais métricas utilizadas na gestão do transporte público e observar as flutuações que ocorrem na procura do transporte em determinadas a épocas do ano e a implementação de medidas administrativas.

A segunda etapa consistiu na análise do sistema de tratamento de dados, o que envolve a utilização de software específico, detecção das fontes de dados e quais os tipos de dados

disponíveis, identificação das associações que já foram feitas com os dados, e qual o tipo de informação que essas associações trazem.

Depois de ter estudado o funcionamento da empresa e fazer uma primeira análise dos dados, iniciou-se o processo de busca pelo melhor método estatístico que pudesse explicar algumas variações identificadas e dar resposta às questões de investigação suscitadas. Durante os meses de estágio foram também efetuadas diversas tarefas de pesquisa bibliográfica, de modo a recolher o melhor tipo de informação que pudesse auxiliar no desenvolvimento da pesquisa.

2.2 Caracterização do Transporte Público

2.2.1 Cobertura de rede e entidades

A rede de transportes públicos em Cascais, é basicamente alimentada pela rede de autocarros municipais e comboios urbanos que ligam Cascais a Lisboa e Sintra, sendo que a sinergia entre esses modos é fundamental para garantir a eficiência e cobertura da rede de transportes no concelho. Na Figura 2-2 pode-se verificar o traçado das linhas de autocarro geridas pela empresa.

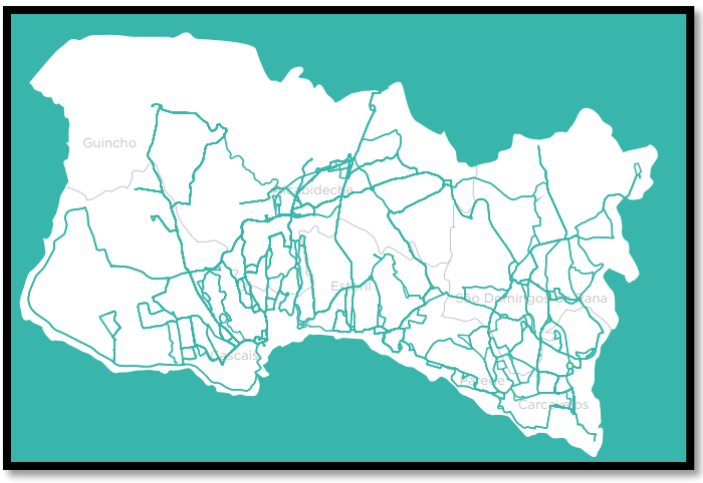


Figura 2-2 Traçado da rede de autocarros Cascais Próxima (Câmara Municipal de Cascais, 2021)

A linha de comboio de Cascais é gerida pela entidade pública Comboios de Portugal (CP) pertencente ao Estado Português e tem como função efetuar a ligação entre Cascais e as restantes localidades ao longo da margem norte do rio Tejo, com estação terminal em Lisboa (ver Figura 2-3), transportando diariamente cerca de 50mil pessoas, (André, 2022). Essa linha é composta por dezessete apiadeiros, sendo que sete delas são dentro do concelho de Cascais, dando ao utilizador também a opção para se deslocar dentro do concelho.

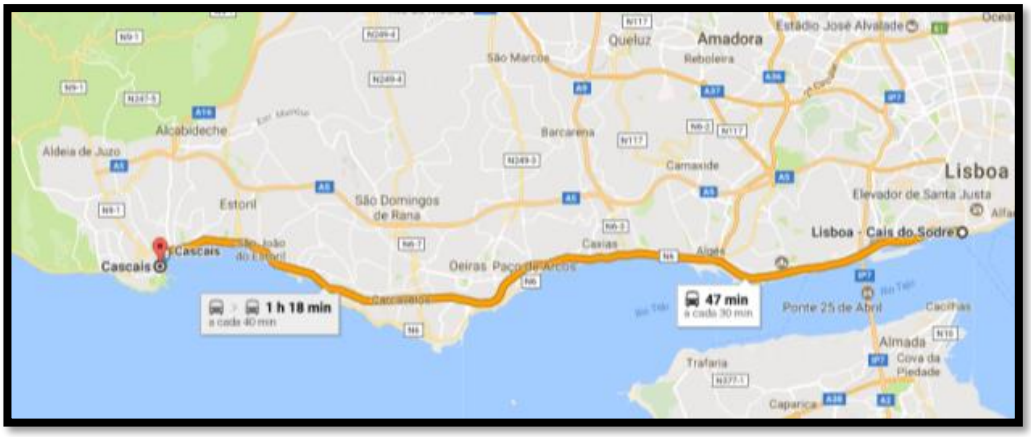


Figura 2-3 Linha de Cascais (Cascais - A Riviera Portuguesa, 2017)

A rede de autocarros é composta por 44 linhas que servem os principais bairros da cidade. Os serviços de transporte público estão disponíveis com frequência, em horários de pico e não pico. A frota é dotada de autocarros modernos e confortáveis, sendo equipados com ar-condicionado, Wi-Fi e acessibilidade para pessoas com mobilidade reduzida.

2.2.2 Classificação das linhas

A literatura de referência traz uma série de classificações para as tipologias de linhas, que podem ser, conforme traçado e funcionalidade.

Conforme o traçado podem ser: (Ferraz & Torres, 2004)

- Radial: Linha que conecta a área central (maior concentração de bens e serviços) a outra região da cidade (outros bairros) mais periféricos;
- Diametral: Conecta duas regiões passando pelo centro;

- Circular: Liga várias regiões através de um circuito fechado, podendo ou não passar pela zona central;
- Interbairros: Liga duas ou mais regiões, sem passar pelo centro;
- Local: Linha que se encontra totalmente dentro de uma área da cidade, para atender viagens diretas para pontos importantes;

Conforme sua função podem ser:

- Convencional: captação de passageiros na zona de origem, transporte até ao destino com distribuição na região;
- Alimentadora: atua na recolha de passageiros em uma zona para os levar até uma estação terminal de uma linha troncal e transportando passageiros na estação e distribuindo-os na região de cobertura;
- Troncal: linha que serve um corredor no qual há grande volume de procura, por exemplo, no transporte de passageiros de uma zona para outra da cidade;
- Expressa: linha que opera com poucas paradas para reduzir o tempo de viagem.

De acordo com as convencionais classificações das linhas de redes de autocarros, é comum encontrar um padrão em que as cidades possuem um centro definido e outras regiões periféricas. Contudo, o cenário em Cascais destaca-se por sua singularidade. Nessa localidade, não se observa a presença de um centro único, mas sim a existência de vários, resultado da formação da cidade que promoveu um desenvolvimento urbano policêntrico.

A configuração urbana diversificada de Cascais desafia as convenções estabelecidas, proporcionando uma dinâmica peculiar às suas redes de transporte público, reflexo da expansão equitativa por distintas áreas centrais ao longo do tempo. (Marques, 2017)

2.2.3 Indicadores de Transportes Públicos

De acordo com (Cruz & Carvalho 2004, in Almeida et al. 2015), indicadores são métricas quantitativas que servem como instrumento de análise de características de certo grupo de operação, através da utilização de atributos qualitativos do grupo observado. Também denominados de *key performance indicator* (KPI), esses indicadores fornecem uma visão mais precisa e completa dos sistemas, permitindo que gestores e planeadores identifiquem áreas de

melhoria e implementem estratégias para melhorar o desempenho do sistema, o que nos transportes públicos, se traduz na melhoria da qualidade no serviço oferecido.

Para (Coca & Torres 2004, in Almeida et al. 2015), a qualidade sentida pelos utilizadores depende entre outros fatores de:

- **Fiabilidade do sistema**: garantia ao utilizador que os horários e rotas previstas pelo fornecedor do serviço de transporte sejam cumpridos, garantindo uma frequência de atendimento de modo a gerar confiabilidade neste.
- **Acesso ao serviço**: a acessibilidade deve ser cumprida de modo a garantir o acesso ao serviço para pessoas com deficiências ou não, e no valor da tarifa, já que esta deve ser condizente com o nível socioeconómico dos utilizadores.
- **Nível de Ocupação das Viaturas**: as viaturas devem transportar passageiros com um nível de lotação confortável para o utilizador, tanto sentados, como em pé.
- **Oferta de serviço**: a oferta dos serviços de transporte deve ser adequada e baseada em rotas e horários que permitam uma boa experiencia de mobilidade ao utilizador.

Considerando que esses fatores servem como indicativos do funcionamento dos TP, a (CCDRN, 2008), apresenta alguns dos principais indicadores que auxiliam na avaliação de redes de transportes públicos, que abrangem volume de transporte, desempenho da rede, produção de transporte, produtividade e eficiência do transporte, como pode ser observado no Quadro 2-1

Quadro 2-1 Indicadores de Transportes Públicos

Indicadores de desempenho conforme divisão de categorias	
Tipo	**Representação**
Volume da rede	
Capacidade da frota	Capacidade de todos os veículos que compõem a frota
Número de linhas e comprimentos de rede	Quilometragem das rotas e da rede total
Número de paragens/ estações	Soma algébrica de todas as estações e paragens de autocarros refletindo a cobertura espacial dos serviços.
Volume anual de passageiros	Volume de passageiros total transportado no período de um ano
Desempenho da rede	
Intensidade do serviço oferecida	Razão entre a produção diária de transporte pelo comprimento total da rede
Produção de transporte	
Veic x Km (anual)	Soma das distâncias realizadas por cada um dos veículos da frota durante um ano
Lugares x Km (anual)	Produto de veic x km anual e a capacidade média dos veículos
Pas x Km (anual)	Produto do número de passageiros transportados e o comprimento médio da viagem
Produtividade	
Produtividade da linha	Representa a distância total percorrida por unidade de tempo, dada pelo produto do número de veículos (passageiros ou lugares) operando numa linha e sua velocidade média.
Eficiência do transporte	
Veic x Km / veic	É a razão do total de Veic x Km realizado num ano pela dimensão da frota
Pas / (Veic x Km)	É o quociente entre o número de passageiros e o veic x km para o período do ano, sendo indicador da intensidade do serviço, quanto maior seu valor, mais eficiente em termos econômicos
Pas / Veic	É a razão entre o volume anual de passageiros e a dimensão da frota, representando a utilização dos veículos em relação ao número de viagens
Pas x Km / Veic	Indica a produção de cada veículos

Fonte: Adaptado de CCDR, 2008

A análise desses indicadores permite o acompanhamento do desempenho do transporte urbano e a observância de variações, como crescimentos ou quedas na procura, que normalmente carecem de alterações na gestão operacional do sistema. Existem vários tipos de análises que podem ser feitas periodicamente. Um exemplo de análise feita mensalmente é o da evolução dos lugares.km, que permitem ao analista uma prévia leitura de aumento ou decréscimo da procura, antecipando as tomadas de decisão da empresa.

Finalmente, o principal objetivo das empresas públicas de transporte público é o de proporcionar um transporte com a maior qualidade possível para os utilizadores, visando atingir o que é essencial para a qualidade nos transportes públicos. Os indicadores de desempenho dos TP podem ser muito eficientes para o incremento da produtividade, qualidade e estratégia das empresas, além de constituírem uma importante ferramenta de monitorização dos sistemas.

2.3 Descrição do sistema de análise da empresa

Até o presente momento, foram apresentadas as funções e responsabilidades da empresa e ilustrada a relevância da boa gestão de mobilidade urbana como uma das principais atividades da mesma. Neste sentido, cabe agora efetuar uma descrição, ainda que sumária, sobre os **dados** as **ferramentas** utilizadas para construção dos modelos de análise da Cascais Próxima.

2.3.1 Fonte de dados

Existem diversas fontes de dados, em diversos formatos que alimentam o *software* de análise, como está listado abaixo:

- Folhas de cálculo para controle de combustível (.xls)
- Dados de localização geográfica (GTFS - *General Transit Feed Specification*)
- Dados de controlo originados na bilhética (.xls)

Relativamente ao terceiro item, o sistema de bilhética, este é responsável pelas vendas e recarga de bilhetes, monitorização e gestão da operação dos sistemas de transporte e é responsável pela garantia da qualidade e segurança do sistema de bilhética utilizado pela Cascais Próxima.

Figura 2-4: Dispositivo eletrônico de bilhética. (Câmara Municipal de Cascais, 2023)

Esses dados de bilhetica são então transmitidos ao sistema central da empresa que os armazena em formato tabular e os transfere para a Cascais Próxima, onde são tratados. Para que se entenda as informações que estão contidas nessas tabelas, pode-se pensar no exemplo de uma obliteração como se repara na Figura 2-4. Em uma validação do cartão, registra-se o tipo de título, o tipo de passe, a paragem de autocarro em que o utilizador entrou, horário, viatura, a rota, condutor e etc.

Todos esses dados, incluindo informações da bilhetagem, localização geográfica e controle de combustível, são reunidos e processados para a criação de relatórios no Power BI. Este processo envolve a coleta de dados, seguida pela implementação de procedimentos de organização e transformação. A partir dessa análise, são gerados relatórios no Power BI que fornecem uma visão clara e estruturada, facilitando a compreensão e tomada de decisões.

2.3.2 O Power BI

De acordo com (Microsoft, n.d.), o Power BI é uma plataforma de análise de dados e *business intelligence* que permite a recolha, análise e visualização de dados de diversas fontes com formatos diferentes, como bases de dados, folhas de cálculo, armazenamento em *cloud*, entre outros, funcionando por meio da conexão automática desses dados. Essa conexão gera informações numéricas e em formato de gráfico, que são dispostos e organizadas nos relatórios síntese em formato de *dashboard de dados.* Esse esquema de organização e etapas de trabalho na ferramenta de BI está ilustrado na Figura 2-5

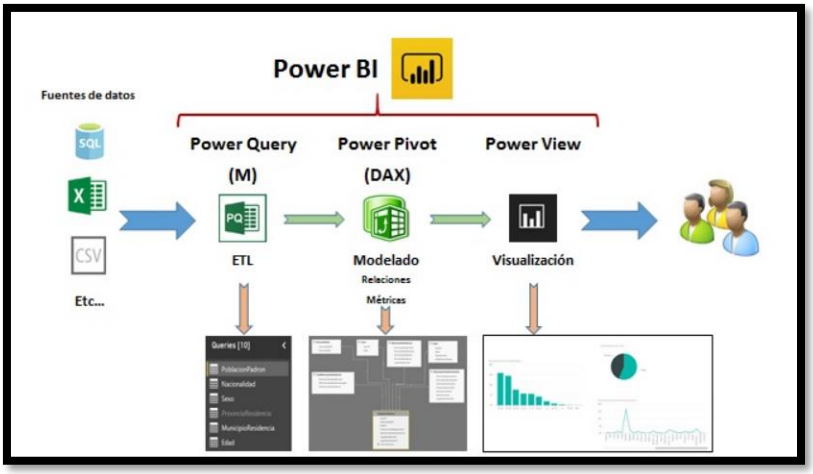


Figura 2-5: Esquema PowerBI, (Cuesta, 2021).

Sendo que:

- **Power Query**: é uma ferramenta de conexão, transformação e modelação de dados. O Power Query permite que os utilizadores se conectem a várias fontes de dados e realizem operações de transformação desses mesmos dados, como remoção de colunas, renomeação de colunas, filtragem, entre outras;
- **Power Pivot**: é uma ferramenta de modelação de dados que permite criar modelos de dados avançados. Com o Power Pivot, os utilizadores podem criar relações entre tabelas, gerando modelos de dados e criar cálculos complexos usando fórmulas DAX (*Data Analysis Expressions*);
- **Power View**: é uma ferramenta de visualização de dados que permite criar relatórios interativos e visualizações de dados avançadas.

No caso da Cascais Próxima, o Power BI é utilizado para a modelação de dados e apresentação de relatórios de desempenho do transporte público. Com esses relatórios torna-se possível o acesso e análise de várias informações relevante. Na sequência, serão ilustrados alguns dos principais relatórios personalizados da empresa.

2.3.2.1 Passageiros

Iniciando pelo relatório do número de passageiros, foi escolhido para demonstração o ano de 2022, com dados sobre todos os veículos, para todos os títulos de transportes.

Pela Figura 2-6, pode-se perceber de forma detalhada as informações que o *dashboard* traz. De cima para baixo, tem-se as oscilações que ocorrem entre os meses, os números totais de viagens, o histograma de passageiros por dia de semana, número de passageiros por hora do dia, e alguns KPIs como a relação lugares por quilómetro e passageiros por quilómetros.

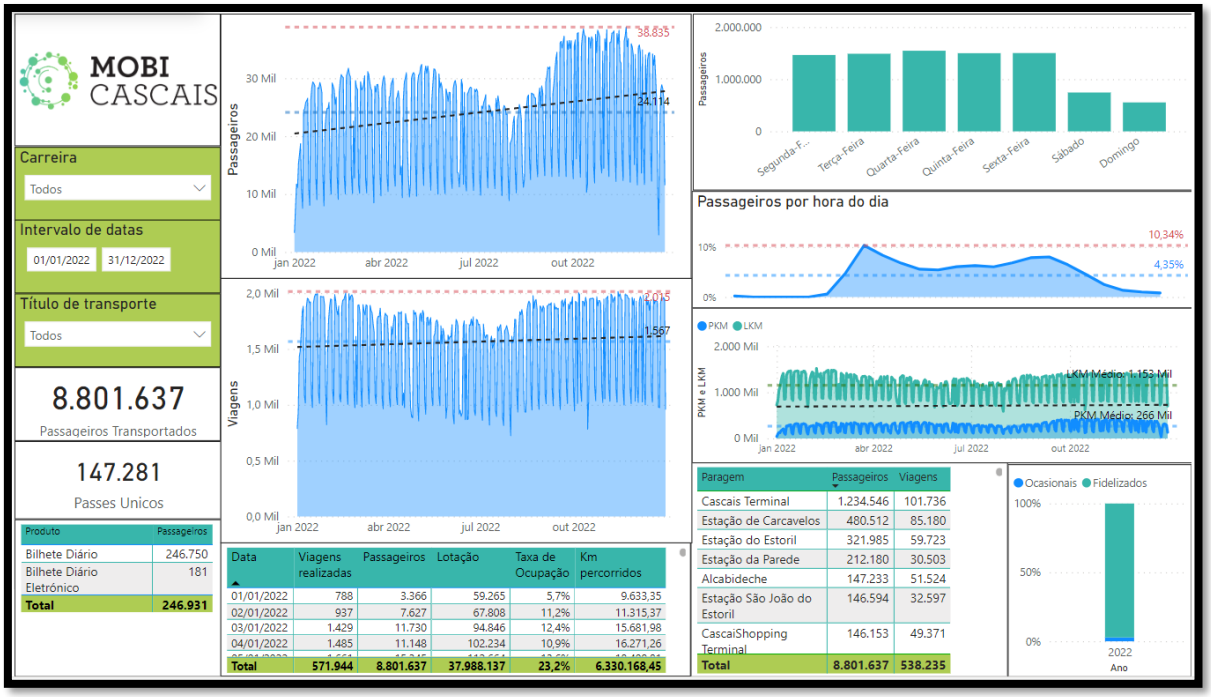


Figura 2-6 Dashboard passageiros, (Cascais Próxima, 2022).

Através deste *dashboard* pode-se fazer uma monitoração da procura. Os gráficos permitem a visualização dos padrões da procura ao longo da semana e durante os períodos do dia. Essas informações permitem a identificação dos horários de pico de passageiros e dias com maior e menor procura, contribuindo para a otimização da operação dos veículos, evitando sobrecargas ou subutilização.

Com base nessas informações também é possível fazer o planeamento de rotas e horários, com os dados do número de passageiros por paragem de autocarro pode-se identificar as paragens com mais movimentos e as rotas mais populares, o que permite a alocação de veículos para atender à procura em cada localidade. Como último exemplo, pode ser referida a contagem de passageiros por paragem, o que poderá ajudar à minimização do tempo de viagem e melhoria do tempo de serviço.

Por último, mas não menos importante, este relatório permite a avaliação da efetividade das estratégias de fidelização, através da distinção entre passageiros fidelizados ou ocasionais pode-se avaliar se as ações implementadas para atrair e manter os passageiros estão gerando impactos positivos, o que auxilia no desenvolvimento de novas estratégias de fidelização.

2.3.2.2 Obliterações

Quanto ao relatório das obliterações, este está mais focado na questão da gestão dos títulos de transporte. As informações que fornece estão ligadas às validações por título, produto e emissor, percentagem de utilização de cada título de transporte em relação ao total dos títulos e o número de viagens por passe.

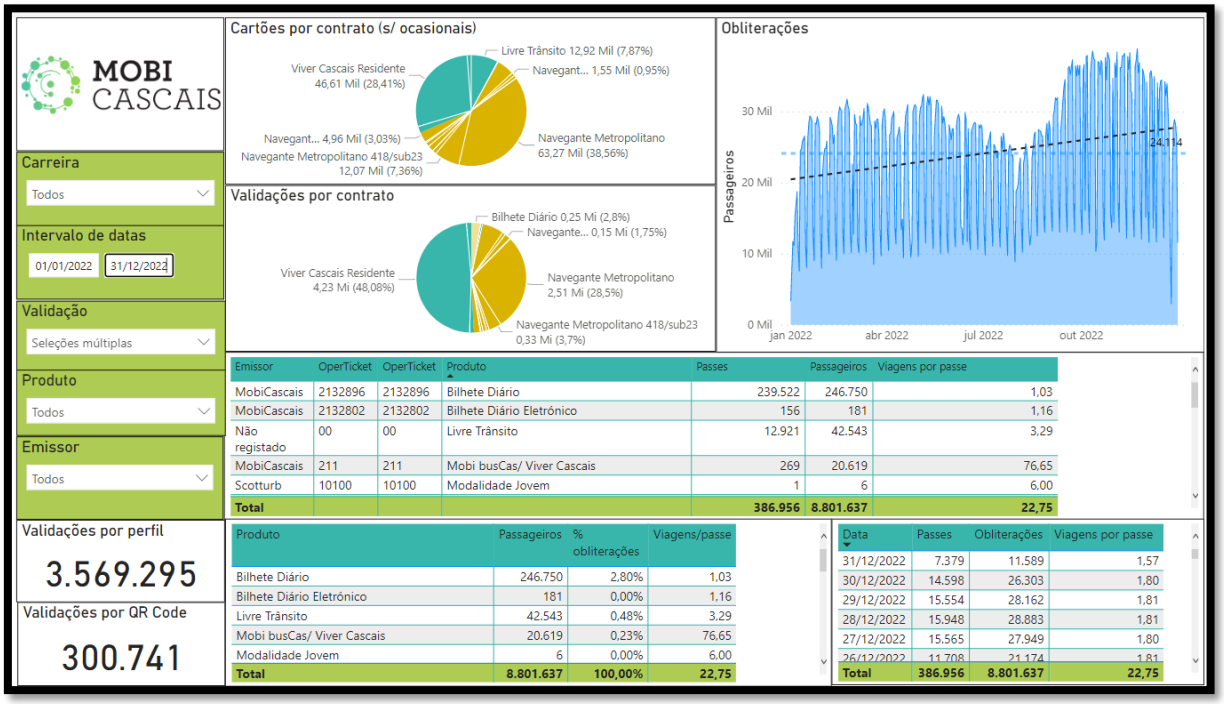


Figura 2-7 Dashboard obliterações, (Cascais Próxima, 2022).

Esse relatório permite a monitorização do uso dos passes. Através da análise do número de viagens com passe pode-se entender quais os padrões de utilização dos passes e identificar aqueles que são mais utilizados, bem como avaliar os que estão em baixo uso, o que permite um ajuste da oferta dos passes e revisão de preços.

Analisar a percentagem de cada tipo de passe em relação ao total é fundamental para entender como os diferentes títulos de transporte estão distribuídos. Essa análise fornece informações valiosas sobre as preferências dos usuários, revelando, por exemplo, quais tipos de passes são mais populares.

Esses *insights* são essenciais para direcionar estrategicamente os esforços de marketing, permitindo promover de forma eficaz o uso de um título específico. Em termos práticos, essa abordagem visa maximizar a adesão dos passageiros aos diferentes títulos de transporte, contribuindo para uma gestão mais eficiente e atendendo às demandas específicas dos usuários.

O vínculo entre a emissão do passe e a quantidade de passageiros que efetivamente utiliza cada linha proporciona uma ferramenta para monitorar o desempenho operacional. Essa correlação permite uma análise detalhada das linhas, revelando quais são as mais populares e aquelas com menor demanda.

Esse discernimento possibilita ajustes estratégicos, como a otimização da frequência dos veículos, realocação eficiente de recursos e até mesmo a consideração de iniciativas como a criação de novas linhas ou a modificação das existentes. Essa abordagem baseada em dados é crucial para uma gestão eficaz do sistema de transporte, visando atender de forma mais precisa às necessidades dos usuários e melhorar continuamente a qualidade do serviço oferecido.

2.3.2.3 Mapas

O relatório que apresenta a distribuição geográfica com a disposição das paragens de autocarro, apresenta a quantidade de passageiros que entram no veículo, o número de viagens realizadas a partir de cada paragem, os horários de partida dos autocarros com seu respectivo número de passageiros e a taxa de ocupação do veículo.

A quantidade de passageiros que entram no veículo e o número de viagens realizadas a partir de cada paragem fornecem informações importantes sobre a procura em diferentes locais e horários, o que permite uma análise das paragens que exigem maiores frequências tornando possível um ajuste na oferta de transporte nesses locais.

A monitoração da taxa de ocupação é um dado importante para avaliar a eficiência e a qualidade do serviço de transporte, regiões e horários específicos onde a taxa de ocupação é muito elevada ou muito baixa, merecem a atenção dos gestores, de modo a melhorar a experiência do usuário. Através dos horários de partida pode-se verificar se os veículos estão

em cumprimento de acordo com o planeamento das rotas e assim verificar sua eficiência e poder implementar medidas que possam melhorar o fluxo de passageiros ou até criação de novas paragens, como por exemplo a capacitação e treinamento de motoristas e sistema de monitoramento em tempo real.

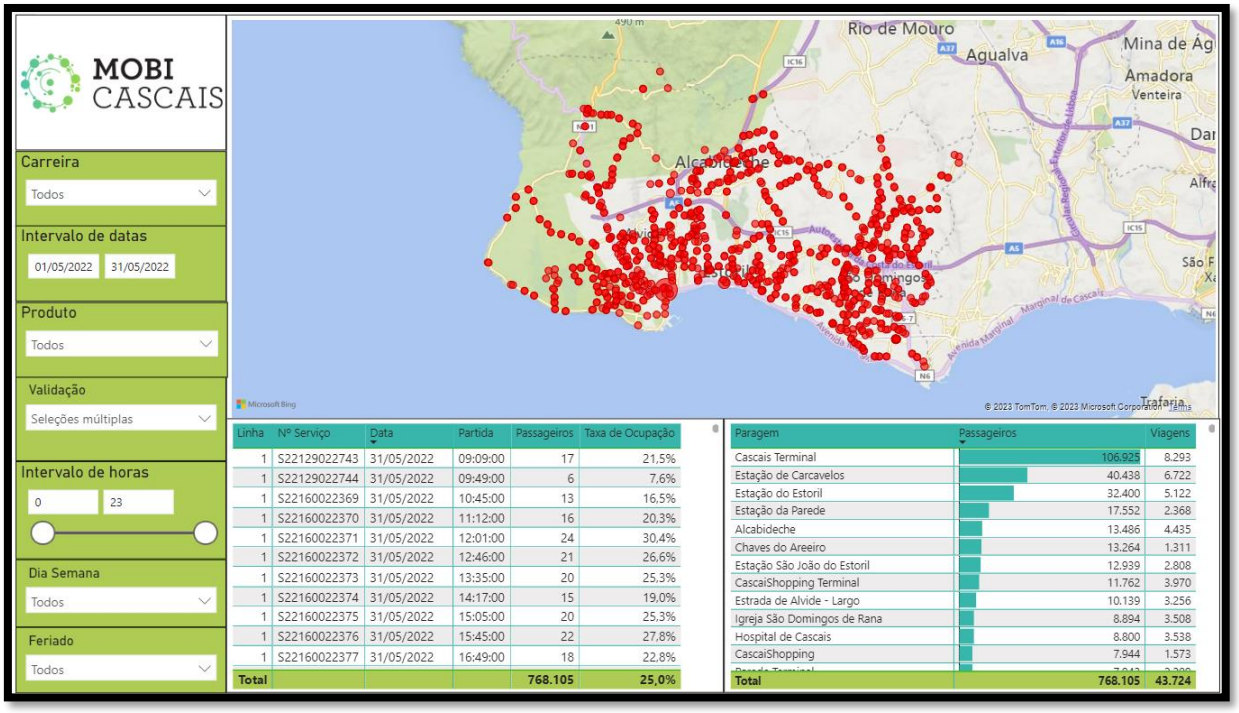


Figura 2-8: Dashboard do mapa, (Cascais Próxima, 2022).

2.3.2.4 Indicadores de desempenho

Por último, mas não menos importante, também deve ser destacado o relatório de indicadores de cumprimento, que indica a relação do que foi planeado, com o que foi executado. Este relatório relaciona as viagens, lugares por quilómetro (representa a média de passageiros transportados por quilômetro percorrido pelo veículo), quilómetros percorridos, viagens e quilômetros vazios (refere-se à distância total percorrida pelo veículo sem transportar passageiros).

Linha	Viagens planeadas	LKM planeados	Km planeados	KM Vazio planeados	Viagens realizadas	Km percorridos	KM Vazio	LKM	PKM	Taxa de Ocupação	Cumprimento	ID Máq.
1	14.235	10.360 Mil	163 Mil	1 Mil	12.453	151,87 Mil	1,47 Mil	11.995 Mil	2.124 Mil	17,7%	87,48%	113
2	6.858	9.005 Mil	141 Mil	3 Mil	6.037	110,92 Mil	7,78 Mil	8.740 Mil	3.505 Mil	40,1%	88,03%	113
3	8.944	11.726 Mil	184 Mil	1 Mil	7.721	167,62 Mil	1,79 Mil	12.530 Mil	3.032 Mil	24,2%	86,33%	113
4	11.372	5.427 Mil	85 Mil	3 Mil	9.661	70,68 Mil	5,45 Mil	5.580 Mil	2.015 Mil	36,1%	84,95%	15
5	4.518	6.151 Mil	97 Mil	3 Mil	4.020	90,51 Mil	6,02 Mil	7.138 Mil	3.114 Mil	43,6%	88,98%	113
6	16.974	17.583 Mil	276 Mil	1 Mil	14.705	235,09 Mil	3,61 Mil	18.111 Mil	7.240 Mil	40,0%	86,63%	113
7	13.029	7.925 Mil	125 Mil	2 Mil	11.190	102,74 Mil	3,93 Mil	8.110 Mil	1.835 Mil	22,6%	85,89%	113
8	10.873	4.508 Mil	71 Mil	3 Mil	9.645	54,37 Mil	5,52 Mil	4.283 Mil	1.459 Mil	34,1%	88,71%	113
9	13.508	7.411 Mil	116 Mil	3 Mil	12.115	97,58 Mil	2,10 Mil	2.960 Mil	1.775 Mil	60,0%	89,69%	114
10	11.997	9.546 Mil	150 Mil	1 Mil	9.915	128,21 Mil	0,67 Mil	3.880 Mil	604 Mil	15,6%	82,65%	114
11	15.132	15.546 Mil	244 Mil	2 Mil	13.318	208,78 Mil	7,72 Mil	16.036 Mil	3.342 Mil	20,8%	88,01%	113
12	10.585	5.596 Mil	88 Mil	2 Mil	9.175	74,10 Mil	5,45 Mil	5.851 Mil	1.226 Mil	20,9%	86,68%	113
13	26.965	32.311 Mil	508 Mil	3 Mil	23.526	411,83 Mil	12,55 Mil	32.476 Mil	15.650 Mil	48,2%	87,25%	113
Total	**961.164**	**732.568 Mil**	**11.509 Mil**	**320 Mil**	**559.256**	**6.173,93 Mil**	**39,37 Mil**	**409.648 Mil**	**106.522 Mil**	**26,0%**	**58,19%**	**1**

Figura 2-9 Indicadores de desempenho, (Cascais Próxima, 2022).

Por meio desses indicadores, conforme mostrado na Figura 2-9, permite a monitorização de metas e objetivos através de taxa de atendimento da procura de passageiros, sendo possível fazer análises para verificar se o sistema está a alcançar as metas estabelecidas, e identificar quais as áreas que precisam de ser melhoradas.

Concluindo, neste segundo capítulo do TFM apresentou-se uma síntese das atividades que foram desenvolvidas no estágio e que são desenvolvidas pela Cascais Próxima, bem como uma análise dos conceitos fundamentais de mobilidade e a caracterização da rede de autocarros de Cascais.

No próximo capítulo será apresentada a revisão bibliográfica efetuada e que serviu de base para o desenvolvimento das análises explicativas implementadas a partir dos modelos de regressão e que serão apresentadas no capítulo 4.

3 Revisão Bibliográfica

A bibliografia utilizada para desenvolvimento deste trabalho focou-se principalmente nos procedimentos e técnicas necessários ao estudo da ciência da análise de dados. A análise de dados é uma prática essencial em diversas áreas de conhecimento, e consiste em usar técnicas e métodos estatísticos para extrair informações relevantes a partir de um conjunto de dados, a partir dos quais se busca identificar padrões, tendências e relações que possam ser úteis para fornecer informações e resolver problemas específicos (Han & kamber, 2002)

Validando este conceito, (Kapel, 2020) afirma que o termo "análise de dados" pode ser definido como o processo de extração de conhecimento a partir de um grande conjunto de dados, o qual requer serviços analíticos sofisticados e escaláveis, bem como ferramentas de programação e aplicações especiais.

Para (Provost & Fawcett, 2013), a análise divide-se entre as seguintes etapas:

1) Entendimento do problema;
2) Entendimento dos dados:
3) Preparação dos dados;
4) Modelagem;
5) Avaliação.

Esses procedimentos, são realizados por meio de ferramentas estatísticas, que em conformidade com (Morettin & Bussab, 2017), é uma ferramenta para tomada de decisões baseadas em dados, permitindo que sejam avaliadas as evidências quantitativas e sejam feitas interferências a partir de amostras de dados.

Um procedimento estatístico normalmente é repartido em três grandes etapas que se complementam, dando início pela estatística descritiva, passando pela estatística probabilista e tendo um fim com a estatística inferencial. Para que se entenda de maneira mais clara, no Quadro 3-1encontra-se a definição e um maior detalhamento dessas etapas.

Quadro 3-1 Áreas da estatística

Áreas da Estatística	
Tipologia	Descrição
Estatística Descritiva	Se preocupa em descrever e sumarizar dados através de medidas descritivas, tabelas e gráficos, permitindo a visualização e compreensão das informações presentes nos dados. Tem como objetivo principal fornecer uma síntese dos dados recolhidos, sem inferir sobre a população.
Estatística Probabilística	Se preocupa em modelar fenómenos aleatórios através de distribuições de probabilidade, permitindo a realização de inferências sobre a população a partir de uma amostra. Tem como objetivo principal o fornecimento de uma base teórica para as técnicas estatísticas inferenciais.
Estatística Inferencial	É a parte da Estatística que se preocupa em fazer inferências sobre uma população a partir de uma amostra, utilizando técnicas de probabilidade e testes de hipóteses. Tem como objetivo principal generalizar as informações encontradas numa amostra para a população.

Fonte: Adaptado de (Morettin & Bussab, 2017; Triola, 2017)

A figura 3-1 mostra como essas áreas se inter-relacionam e alguns dos elementos básicos que são resultados de cada etapa.

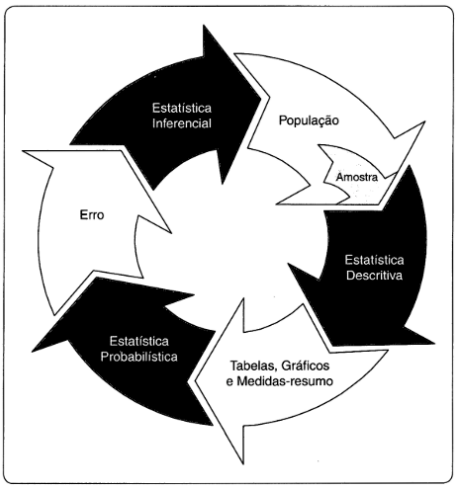


Figura 3-1 - Áreas da Estatística (Fávero & Belfiore, 2017)

- A estatística descritiva resulta na elaboração e análise de gráficos e tabelas do resumo de dados;

- Na estatística probabilística, são quantificadas as incertezas, através da estimação dos erros, e isso é feito com dados amostrais tirados da população;
    - A população é definida como o conjunto que possui todos os elementos a serem estudados, que possuem uma característica em comum, também podendo ser chamado de universo;
    - A amostra é definida como um subconjunto extraído da população para fins de análise considerando que esse subconjunto seja representativo da população em questão.

3.1 Análise exploratória dos dados

A análise exploratória de dados foi introduzida originalmente pelo estatístico americano John Turkey na década de 70 com a publicação de seu livro *Explanatory Data Analysis* (Análise Exploratória de Dados, em tradução livre). Nesta publicação, o autor visou incentivar os estatísticos a explorarem os dados e formular hipóteses, antes mesmo da aplicação de testes estatísticos, pois acreditava que através dessa análise prévia era possível conhecer melhor os dados e até induzir a formulação de mais hipóteses. (Turkey, 1977).

De acordo com (Bruce & Bruce, 2019), a análise exploratória de dados (AED), visa garantir que o cientista de dados tenha um conhecimento prévio dos mesmos, através da utilização de ferramentas visuais, medidas resumo, e da formulação de outros testes de hipóteses que podem ser considerados durante a análise gráfica.

Com essa descrição dos dados obtém-se então um resumo para visualização eficiente, clara e concisa, tornando-se possível a identificação de *outliers*, o estabelecimento de outras possíveis relações entre as variáveis e a comparação dos dados, resultados estes, que facilitam a organização dos dados e garantem uma boa escolha da ferramenta estatística a ser utilizado.

Uma das grandes vantagens de se aplicar a AED num estudo estatístico é que se pode colher muitas informações nessa etapa inicial, sem o peso das suposições probabilísticas, simplificando a forma como a base de dados pode ser compreendida seguindo uma estratégia de análise prévia para posterior construção do modelo.

.

3.1.1 Variáveis

Em conformidade com (Fávero & Belfiore, 2017), uma variável captura uma característica da observação, que pode ser medida e contada, ou ser um atributo dos indivíduos pesquisados, que são a matéria-prima das análises estatísticas.

Segundo (Morettin & Bussab, 2017) as variáveis podem ser classificadas em dois grupos, as variáveis qualitativas (categóricas) e quantitativas (métricas). As variáveis quantitativas são aquelas que podem ser medidas numa escala numérica, e exprimem uma quantidade. As variáveis qualitativas são aquelas que apresentam algum atributo do indivíduo pesquisado, ou seja, uma característica que o represente.

Esses grupos ainda podem ser subdivididos em mais dois grupos cada, como pode ser observado na Figura 3-2.

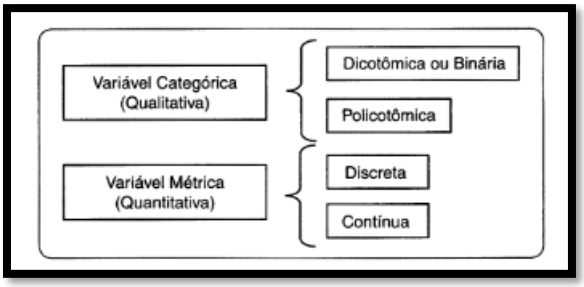


Figura 3-2 Tipos de variáveis. (Fávero & Belfiore, 2017).

3.1.1.1 Variáveis qualitativas

Variáveis qualitativas são variáveis que representam características (atributos) que não podem ser expressos numericamente de forma contínua. De acordo com (Fávero & Belfiore, 2017), essas variáveis podem ser divididas em grupos distintos ou categorias, sendo classificadas conforme o número de categorias que apresentam.

As que assumem duas categorias são denominadas dicotômicas e aquelas que assumem mais de duas categorias são denominadas policotômicas. As variáveis dicotômicas ou *dummy*, são também conhecidas como variáveis binárias. Para que essas variáveis entrem num cálculo estatístico, é necessário efetuar a sua codificação em um formato numérico.

A necessidade dessa transformação numérica é explicada por (Missio & Jacobi, 2007) como um método de quantificação dos atributos, para que possam ser analisados estatisticamente e é uma técnica utilizada em modelos de regressão linear para garantir que as variáveis categóricas possam ser utilizadas neste tipo de modelos. Usualmente, a **codificação *dummy*** envolve a criação de uma variável binária que assume o valor de 1 quando a condição está presente e o valor de 0 quando a condição está ausente.

Para o desenvolvimento da codificação *dummy* assume-se que se a variável qualitativa nominal apresentar "k" categorias distintas, então serão criadas "k-1" categorias. A categoria de referência é representada por 0 em todas as variáveis *dummy*, enquanto as outras categorias são representadas por 1 numa única variável *dummy* correspondente. (Fávero & Belfiore, 2017)

Outra classificação que as variáveis assumem refere-se à sua escala: variáveis nominais ou ordinais. As nominais são aquelas que não possuem uma ordem lógica entre as categorias e são expressas em termos de categorias exclusivas, sem uma relação entre elas. As variáveis ordinais têm ordem lógica. Estas variáveis representam atributos que possuem uma ordenação e são expressos numa escala ordinal.

3.1.1.2 Variáveis quantitativas

Assim como as variáveis qualitativas, as variáveis quantitativas também possuem a sua subclassificação, que neste caso é em função da sua natureza, dividindo-se em dois grupos, sendo as variáveis discretas e contínuas. As variáveis discretas são aquelas cujos valores formam um conjunto finito de números normalmente resultante de contagem, e as variáveis contínuas são resultantes de uma medição e pertencem sempre a um intervalo contínuo de números.

As variáveis quantitativas, segundo (Fávero & Belfiore, 2017) , por serem métricas são representadas na forma gráfica (histograma, gráfico de linhas, gráfico de dispersão e ramo-e-folhas), por meio de medidas de posição ou localização (média, mediana, moda, quartis e percentis), medidas de dispersão (amplitude, desvio-médio, variância, erro-padrão, desvio-padrão e coeficiente de variação).

A tipologia das variáveis é crucial no cálculo das estatísticas descritivas pois a utilização de cada método, e as interpretações das medidas descritivas, varriam conforme o tipo de

variável que está sendo utilizado. Cada tipologia apresenta características distintas, e as estatísticas descritivas devem ser aplicadas de maneira apropriada em cada tipo.

3.1.1.3 Estrutura dos dados

A estrutura dos dados refere-se à estrutura em que os dados estão armazenados e apresentados, e de acordo com (Han et al., 2023) podem ser estruturas transacionais, multidimensionais ou séries temporais. Para o desenvolvimento deste trabalho interessa debruçar-nos sobre as estruturas de dados organizadas em séries temporais, em que os dados são organizados numa sequência temporal, com observações realizadas num intervalo de tempo.

De acordo com (Machado, 2012), entende-se por série temporal as observações feitas de maneira sequencial ao longo do tempo sendo que, são dependentes umas das outras, fazendo com que sejam estatisticamente relacionadas. Isso quer dizer que as observações no momento presente são influenciadas pelas observações passadas e podem afetar as observações futuras, o que significa que os valores próximos tendem a estar correlacionados.

Ainda de acordo com o mesmo autor ao se realizar uma análise de uma série temporal existem algumas variações importantes que podem ser detetadas e observadas tais como:

- Tendência: variações da variável ao longo do tempo, que podem ser ascendentes ou descendentes, que permitem uma descrição do comportamento.
- Sazonalidade: é um movimento sistemático que acontece durante o ano, podendo ser ou não regular, causado por mudanças no clima, calendário e decisões da oferta que afetam diretamente a produção e consumo.
- Efeito Cíclico: o efeito cíclico é caracterizado por oscilações recorrentes em uma série temporal, que se repetem ao longo do tempo sem uma regularidade fixa e é resultado de fatores endógenos e exógenos que afetam o sistema em questão.
- Ruído aleatório: são componentes das séries temporais que contém elementos aleatórios que podem dificultar a identificação dos padrões.

3.1.2 Identificação de Outliers

Os *outleiers* são valores extremos ou discrepantes em um conjunto de dados que se afastam significativamente da maioria das outras observações (Martins, 2019). Em outros termos, são valores que se encontram muito longe dos demais e que podem afetar negativamente a análise estatística, principalmente em relação à média e variância dos dados.

A importância de se identificar e tratar os *outleiers* é fundamental para a garantia de precisão dos resultados, considerando que quando não são tratados, podem distorcer a interpretação dos mesmos e comprometer a fiabilidade das conclusões.

Uma das maneiras eficazes de identificar *outleiers* é através da utilização de ferramentas gráficas, isso porque estas ferramentas permitem visualizar rapidamente a distribuição dos dados e identificar quais pontos estão afastados da maioria das observações

3.1.3 Medidas representativas

É possível resumir as informações presentes em um conjunto de dados utilizando medidas numéricas apropriadas, conhecidas como medidas-resumo. Essas medidas são determinadas em conformidade com a quantidade de variáveis em questão.

Quando se tem apenas uma variável para analisar, algumas das medidas representativas mais tradicionais são médias, variância e desvio padrão. Essas medidas são utilizadas para descrever as características centrais de um conjunto de dados permitindo uma compreensão mais fácil e rápida das propriedades de distribuição de dados. (Fávero & Belfiore, 2017)

Quando o estudo é direcionado para duas variáveis, as medidas que traduzem informação dos dados são de correlação, sendo que as mais utilizadas são a covariância e o coeficiente de Pearson. A covariância mede a variação conjunta entre duas variáveis, e o coeficiente de correlação de Pearson é uma medida que indica o grau de associação entre as duas variáveis, tais medidas de correlação serão desenvolvidas com mais detalhes na secção 3.3.4.2.4.

Por fim, a análise exploratória dos dados busca a recolha das primeiras informações que o conjunto de dados pode oferecer, através de representações gráficas e medidas resumo, o que facilita e contribui de maneira significativa para o bom desempenho do estudo, informando o

cientista de dados sobre as informações mais “simples” de serem descobertas, podendo ajudar no desenvolvimento do estudo através da melhor compreensão dos dados.

3.2 Probabilidade

A probabilidade é uma área fundamental na estatística, que representa uma medida numérica capaz de descrever a chance de ocorrência de determinado evento, sendo também uma ferramenta de quantificação da aleatoriedade e da incerteza.

Esse conceito foi introduzido pelo matemático francês Laplace, e em conformidade com essas teorias Afonso & Nunes, (2019) afirmam que esta ferramenta se baseia na ideia de que em um espaço amostral finito e equiprovável, a probabilidade de um evento ocorrer é dada pela razão entre o número de casos favoráveis e o número de casos possíveis.

Quanto a sua finalidade, (Fávero & Belfiore, 2017) afirmam que a probabilidade visa explicar a frequência de ocorrência de determinados eventos incertos, visando a estimativa ou previsão de ocorrência de eventos futuros.

Na análise de dados, a probabilidade é utilizada para descrever e modelar a distribuição dos dados, de modo a atribuir uma possibilidade a cada resultado possível de uma experiência. Com o resultado da distribuição, estima-se o **intervalo de confiança** que corresponde a faixa de valores dentro da qual um parâmetro populacional é provavelmente encontrado com uma certa probabilidade.

A análise de probabilidades corresponde a uma abordagem utilizada ao longo de todo o estudo estatístico. Na estatística descritiva ela descreve a chance dos eventos ocorrerem, na estatística probabilística, ela modela a análise de fenómenos aleatórios, quantificando a incerteza associada aos mesmos e na inferência estatística visa, através do estudo da incerteza, diminuir o máximo possível os erros contidos nas informações de inferência, utilizando como principal ferramenta a distribuição de probabilidade

3.2.1 Distribuições de Probabilidade

As distribuições de probabilidade são funções matemáticas que descrevem a probabilidade de ocorrência de diferentes valores de uma variável aleatória e fornecem uma

representação visual quantitativa dos possíveis resultados e das suas probabilidades associadas. As distribuições de probabilidade são definidas por (Piana et al., 2009) como um modelo de descrição probabilística de uma população, sendo a população todo o conjunto de valores de uma variável aleatória.

De seguida apresenta-se uma descrição síntese das distribuições de probabilidades mais relevantes para o desenvolvimento deste trabalho.

3.2.1.1 Distribuição normal

As distribuições são desenvolvidas conforme a tipologia de suas variáveis quanto a aleatoriedade e métrica. Atualmente existem inúmeras metodologias que podem ser utilizadas para cálculo de probabilidade estatística. Entre as várias distribuições estatísticas, a **distribuição normal** é a mais comum e mais utilizada em estudos estatísticos, por ser mais representativa.

De acordo com (Piana et al., 2009), a distribuição normal é definida como uma distribuição teórica de frequências, em que grande parte das observações estão localizadas em torno da média (centro da distribuição) e vão diminuindo gradual e simetricamente nas extremidades. Essa distribuição é representada graficamente pela curva normal, ou curva de Gauss, que possui forma de sino, sendo simétrica em relação ao centro, onde se localiza a média μ como pode ser observado na Figura 3-3

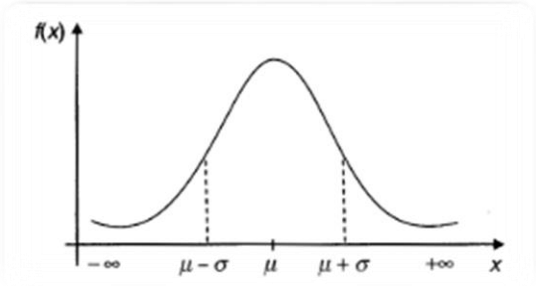


Figura 3-3 Gráfico de distribuição normal. (Fávero & Belfiore, 2017)

A função densidade de probabilidade descreve como a probabilidade está distribuída ao longo do eixo x e seus parâmetros são a média, que determina o centro da distribuição e a variância que determina a dispersão da distribuição. Graficamente a média representa o ponto

máximo da função, resultando numa distribuição simétrica em relação ao centro e os pontos de inflexão são exatamente $\mu-\sigma$ e $\mu+\sigma$. (Fávero & Belfiore, 2017)

Para este tipo de distribuição atenta-se ainda que mesmo que os dados não sigam uma distribuição normal, mas possua uma amostra com um número grande de dados, a distribuição da média amostral é uma distribuição aparentemente normal, o que pode ser comprovado pelo Teorema do Limite Central. Esse teorema infere que de acordo com o aumento do tamanho da amostra, a distribuição da média aproxima-se cada vez mais da distribuição normal (Henrique & Molin, 2016).

Além de descrever a forma e a variabilidade dos dados, a distribuição normal também é utilizada para a determinação dos intervalos de confiança, com base na propriedade de que a média amostral segue uma distribuição normal

3.2.1.2 Distribuição t de Student

Em conformidade com (Henrique & Molin, 2016), diferentemente do que se vê na distribuição normal, em que se conhecem os dados da população, a distribuição t de *Student* é calculada quando não se tem acesso a esses dados, ou seja, o desvio padrão da população é desconhecido. Dessa maneira, estimam-se os intervalos de confiança tendo como base a variação amostral.

Quanto a sua formatação, Bruce & Bruce, (2019) explicam que a distribuição t de *Student* é formatada como a distribuição normal, diferindo-se por ser mais espessa e mais longa nas caudas, o que significa que há mais probabilidade de se obter valores extremos. A parametrização, em conformidade com (Fávero & Belfiore (2017), é dada através do número de graus de liberdade (ν), calculado pelo número total de amostras menos 1, que define e caracteriza a forma da distribuição. Quanto maior o valor de ν, mais próxima da distribuição normal padrão, como pode ser observado na Figura 3-4.

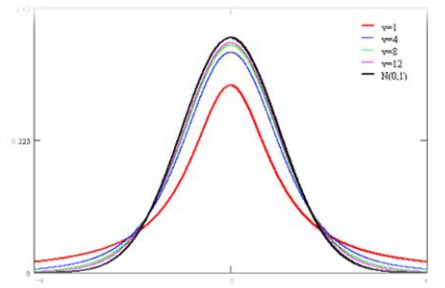


Figura 3-4 Distribuição t-Student com n-1 graus de liberdade. (Henrique & Molin, 2016)

3.2.1.3 Distribuição F de Snedecor

Em conformidade com (Triola, 2017), a distribuição de Snedecor, também conhecida como distribuição F, é uma distribuição de probabilidade contínua que surge na análise das variâncias (ANOVA), sendo utilizada para testar a igualdade de variâncias entre dois ou mais grupos.

De acordo com (Magalhães & Lima, 2018), a distribuição de Snedecor assume a forma assimétrica e pode ter valores positivos ou negativos. Ela é caracterizada por dois parâmetros: o número de graus de liberdade do numerador, que representa a variância entre grupos ν1 e o número de graus de liberdade no denominador, que representa a variância dentro do grupo ν2.

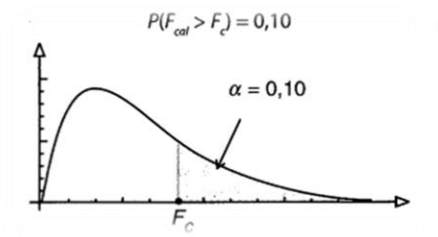


Figura 3-5 Distribuição F de Snedecor para α =0.1. (Fávero & Belfiore, 2017)

Nesta etapa, foram apresentados gráficos como uma representação visual didática das distribuições. No decorrer do trabalho, optou-se por não gerar gráficos para os resultados obtidos a partir da análise de variância (ANOVA). Em vez disso, os resultados foram diretamente expressos em formato numérico para facilitar a análise.

3.3 Inferência Estatística

A inferência estatística, é definida por Morettin & Bussab (2017) como a etapa em que são inferidos os possíveis resultados de uma população com base nos cálculos realizados a partir das suas amostras. Em sua metodologia, a inferência faz uso da **estimativa de parâmetros** e de realização dos **testes de hipóteses** para suposição de determinado parâmetro, de modo a induzir uma tomada de decisão.

Para (Henrique & Molin, 2016), como o processo de inferência decorre da indução de factos, ele pode não ser exato e está sujeito a erros, mas como se trata de uma análise científica, a inferência estatística~~,~~ informa o cientista de dados sobre **até que ponto ele pode estar "errando"** e com que probabilidade, com a execução **dos testes de significância do modelo e das variáveis**.

Os **testes de significância**, num modelo de regressão linear são utilizados para avaliar a importância das variáveis independentes em explicar a variação da variável dependente. No geral, estes testes são utilizados para determinar se o modelo de regressão é estatisticamente significativo e se as variáveis independentes contribuem significativamente para explicar a variância da variável dependente.

Entre os principais aspetos da inferência estatística está o **processo de amostragem**, que deve ser representativa da população, e a determinação do tamanho da amostra, que deve ser grande o suficiente para garantir a confiabilidade dos resultados.

De entre as técnicas e procedimentos existentes para inferir informações a partir de análises estatísticas de dados, uma das mais utilizada ainda vem sendo a **regressão linear**, usando como ferramenta de testes de hipótese a **análise de variância**.

3.3.1 Amostragem

O processo de amostragem é a técnica utilizada para selecionar uma parte da população para ser estudada, de forma a obter informações sobre a população como um todo. As tipologias de amostragem e suas respetivas descrições podem ser analisadas no Quadro 3-2

Quadro 3-2 Conceituação dos processos de amostragem

Processos de amostragem	
Tipo	Descrição
Amostragem aleatória simples	Consiste na seleção aleatória dos indivíduos da população, de forma que todos tenham a mesma chance de serem escolhidos.
Amostragem estratificada	Consiste na divisão da população em estratos (subgrupos) e na seleção de indivíduos aleatoriamente dentro de cada estrato. Usado para população com características heterogêneas.
Amostragem por conglomerados	Consiste na divisão da população em conglomerados (grupos) e na seleção aleatória de alguns grupos para serem estudados. Usado para populações muito grandes.
Amostragem por quotas	Consiste na seleção de indivíduos de acordo com determinadas características.

Fonte: Adaptado de (Magalhães & Lima, 2018)

O critério de escolha da tipologia de amostragem vai depender do tipo de estudo em que o cientista de dados estiver envolvido, e de quais os resultados que pretende obter, visto que para cada tipo de amostragem, o resultado da regressão varia.

3.3.2 Estimativa dos parâmetros

Esta etapa do processo consiste no cálculo de estimadores dos valores desconhecidos dos parâmetros populacionais, a partir dos dados das amostras. Estes estimadores são uma aproximação dos parâmetros da população e têm como função descrever de forma aproximada a mesma.

Os parâmetros, em conformidade com (Piana et al., 2009), são valores calculados diretamente da população, tais como a média e desvio padrão. Um estimador é uma função estatística utilizada para estimar o valor de um parâmetro desconhecido com base nos dados amostrais. Os estimadores geram estimativas do valor numérico do parâmetro, calculadas com base nesses dados.

Figura 3-6: Esquema parâmetros e estimadores. (Campos, 2019)

Os estimadores, para que sejam escolhidos, precisam atender alguns critérios para que sejam representativos da população, o que inclui:

- **Imparcialidade ou não tendenciosidade**: um estimador é um estimador imparcial do parâmetro θ se o valor esperado de E(θ) for igual a θ. Isto quer dizer que a função matemática escolhida para estimar o parâmetro não contém vieses, vícios ou tendências, de modo que retorne um resultado imparcial. $E(\theta) \rightarrow \theta$;
- **Eficiência ou variância mínima**: a eficiência serve para comparar estimadores e verificar qual é mais eficiente. Se dois estimadores de um mesmo parâmetro são imparciais, é mais eficiente aquele que possui menor variância;
- **Consistência**: um estimador é consistente se à medida que o tamanho da amostra aumenta, o valor do estimador se aproxima do parâmetro e a variância converge para zero.

Quanto aos métodos de estimação, Piana et al., (2009), afirmam que os parâmetros podem ser estimados por intervalo ou por pontos. A estimação por intervalo é um procedimento no qual se obtém um intervalo onde, com determinada probabilidade (nível de confiança), se pode encontrar o valor do parâmetro. A metodologia por pontos refere-se ao processo de obter apenas um único ponto e a estimação por intervalo mais comumente utilizada é a do intervalo de confiança, que será discutida no próximo tópico.

3.3.3 Teste de hipóteses

Num estudo estatístico, onde se busca através do tratamento de dados a modelação da incerteza, a proposição de algumas questões que direcionem o estudo são parte fundamental do procedimento. Antes mesmo que os cálculos se iniciem, o desenvolvimento do estudo é moldado através de questões que os invetigadores desejam esclarecer sobre determinados parâmetros. Em termos estatísticos, essas perguntas são denominadas hipóteses e precisam ser testadas para se verificar a sua veracidade ou não.

Piana et al., (2009), definem o teste de hipóteses como um procedimento estatístico através do qual se busca verificar uma hipótese a respeito da população, a partir de dados amostrais, tendo por base a teoria das probabilidades.

Os autores (Fávero & Belfiore, 2017) definem uma hipótese estatística como sendo uma suposição sobre um determinado parâmetro da população, em que a hipótese nula representa a afirmação de que não há diferença, e a hipótese alternativa representa a afirmação que há alguma diferença.

Ainda segundo estes autores, os testes podem ser classificados em paramétricos e não paramétricos. Os paramétricos são aplicados aos dados quantitativos e os testes não paramétricos são apropriados para os dados de natureza qualitativa. Neste estudo só serão abordados os testes do tipo paramétrico.

As etapas da execução de um teste de hipóteses incluem:

- formulação das hipóteses;
- definição do nível de significância;
- seleção do teste estatístico adequado;
- execução de testes de significância do modelo e dos parâmetros;
- interpretação dos resultados.

3.3.3.1 Formulação das hipóteses

Em conformidade com (Henrique & Molin, 2016), no teste paramétrico~~,~~ podem analisadas uma ou duas hipóteses: a hipótese nula (H_0) e a hipótese alternativa (H_α), sendo essa formulação para um teste bilateral. Noutros casos pode haver testes de hipóteses unilaterais,

que servem para testar se o parâmetro é maior ou menor do que a população. Em baixo estão ilustrados, respeetivamente, os testes unilateral e bilateral.

$$T: \begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu \neq \mu_0. \end{cases}$$

$$T: \begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu < \mu_0 \end{cases} \quad T: \begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$$

3.3.3.2 Definição nível de significância

Em conformidade com (Bruce & Bruce, 2019), ao inferir características de uma população através de testes de hipóteses, com os dados amostrais, deve-se levar em consideração que, para além das variações de regressão, também existem as variações dos resíduos que podem incidir sobre os valores de Y. O nível de significância, também denominado como taxa de erro, mede, através **de valores críticos,** o **intervalo de confiança** da distribuição.

De acordo com (Henrique & Molin, 2016), os níveis de significância (α) mais usuais são 1%, 5% ou 10%, sendo fixados pelo investigador com base em critérios empíricos. Nesses termos, caso seja escolhido um nível de significância 5%, que é mais usual, pode-se afirmar que existe um risco de 5% de concluir que há uma diferença entre os parâmetros, , quando na realidade não existe.

Ainda em conformidade com estes autores, ao conduzir um teste de hipóteses, o cientista de dados está exposto a dois tipos de erros:

- Erro tipo I (α): concluir erroneamente que um efeito é real, quando na verdade ele ocorre por acaso. É o caso de se rejeitar uma hipótese nula quando ela é verdadeira. Sendo que α representa a probabilidade de ocorrência desse erro.
- Erro tipo II (β): concluir erroneamente que um efeito não é real, quando na realidade ele é real. É o caso de se rejeitar a hipótese nula quando ela for falsa. Sendo que β representa a probabilidade de ocorrência desse erro.

Para (Bruce & Bruce, 2019), uma das maneiras de se reduzir essas duas taxas de erro (α e β) será com o aumento da amostra. No entanto isto nem sempre isso é possível, cabendo então ao cientista de dados a perceção de qual dos erros é mais grave, e procurar a sua diminuição.

Normalmente a probabilidade de ocorrência do erro tipo I é considerada mais grave, e é chamada de **nível de significância do teste**.

3.3.3.3 Estatística de teste

Ainda de acordo com (Bruce & Bruce, 2019), s estatísticas de teste visam a introdução de uma forma sistemática e objetiva de medidas de avaliação relativas às hipóteses propostas sobre os dados, para verificar se são suportadas pelos resultados observados, sendo de fundamental importância para garantir que as conclusões tiradas a partir dos mesmos sejam confiáveis e objetivas.

Ao escolher a estatística de teste para análise de hipóteses, é importante considerar critérios como: o tipo de dados, a sua distribuição, tipo de hipótese, entre outros. De acordo com (Oiseth et al., 2022), os testes estatísticos dividem-se em três grupos:

- Testes de Regressão;
- Testes de Comparação;
- Testes de Correlação.

Os testes de regressão, são aqueles que avaliam as relações de causa e efeito. Para (Martins, 2019), um modelo de regressão é um modelo matemático que descreve a relação entre duas ou mais variáveis do tipo quantitativo. Caso o estudo seja sobre duas variáveis e estas mantenham uma relação eminentemente linear, este é denominado regressão linear simples, caso o estudo seja sobre mais de duas variáveis, este então é denominado regressão linear múltipla.

A regressão linear simples testa como uma variável independente altera a variável dependente com a utilização de variáveis contínuas e a regressão linear múltipla testa como a combinação de duas ou mais variáveis independentes alteram a variável dependente, com a utilização de variáveis contínuas

Os testes de comparação são aqueles que comparam médias entre os diferentes grupos, e os mais representativos são o Teste t e a Análise de Variância (ANOVA). O teste t é utilizado para determinar se há diferenças significativas entre os valores médios de dois grupos de dados, e a ANOVA é uma técnica de comparação para mais de dois grupos.

E por fim, **os testes de correlação** que procuram associações entre diferentes variáveis. Para (Fávero & Belfiore, 2017), essa tipologia de teste não expressa se a variável X é significante e se é causa verdadeira para alterações em Y. Além disso não oferecem condições para afirmar se as variáveis X são adequadas, portanto deve ser analisado com cautela e sem demasiada importância. Como exemplo deste tipo de teste para variáveis contínuas, tem-se o teste r de Pearson, que testa a força de associação entre duas variáveis.

3.3.4 Regressão Linear

Os modelos de regressão são representações matemáticas utilizados para descrever a relação entre duas ou mais variáveis quantitativas. Esses modelos objetivam a determinação da equação que melhor representa a relação existente entre as variáveis, sendo seu processo executivo dividido em três etapas conforme listagem abaixo. (Daniels & Minot, 2018)

- obtenção das estimativas dos coeficientes para ajustar a equação;
- aplicação de testes de significância;
- cálculo dos intervalos de confiança.

Neste trabalho, o foco será no modelo de regressão múltipla, que, como definido por (Makridakis S, Wheelwright SC, 1997), é uma forma especial de regressão em que a variável a ser prevista depende de duas ou mais variáveis explicativas.

3.3.4.1 Estimativa dos parâmetros

Para (Morettin & Bussab, 2017), as estimativas dos parâmetros podem ser de caráter explicativo ou preditivo. As de caráter explicativo são aquelas que demonstram uma relação matemática, mas não provam uma relação de causa e efeito. Já as de caráter preditivo buscam obter uma relação que permita prever o valor de Y em futuras observações de X.

$$Y = aX + b \tag{3.1}$$

As estimativas de caráter explicativo buscam entender a relação entre a variável resposta (Y) e as variáveis preditoras (X), visando determinar como a variação na variável resposta é

explicada pelas variáveis preditoras. Essas estimativas de parâmetros são usadas para explicar o comportamento da variável resposta em termos das variáveis preditoras e para entender quais variáveis têm maior impacto na variável resposta.

No estudo apresentado no capítulo 4, como o objetivo é a caracterização das variáveis e como elas influenciam na procura, o foco será a estimativa de caráter explicativo.

Na estimativa de parâmetros da regressão linear, busca-se determinar quais os valores dos coeficientes **a** e **b,** e como nesse conjunto de observações existe uma parcela de erro, (Morettin & Bussab, 2017) afirmam que a estimativa é realizada através da soma dos mínimos quadrados dos desvios, que é denominada **método dos mínimos quadrados**.

O método dos mínimos quadrados busca a aproximação da função que melhor se ajusta ao conjunto de dados através na minimização da soma dos quadrados das diferenças entre os valores esperados e os valores previstos pela função. Neste caso, os pontos que relacionam as variáveis explicativas e a variável resposta são ajustados através da minimização.

A Figura 3-7 resume a reta de regressão, demonstrando as observações e as estimações, os erros provenientes do ajuste da reta e as novas variáveis resposta estimadas. O exemplo da figura mostra uma reta de regressão linear simples, no entanto no caso do TP de Cascais, apesar de os princípios serem os mesmos, a reta de regressão é substituída por um hiperplano de regressão que ocorre em várias dimensões (correspondentes ao número de variáveis independentes) e a minimização ocorre sobre esse hiperplano ao invés da reta.

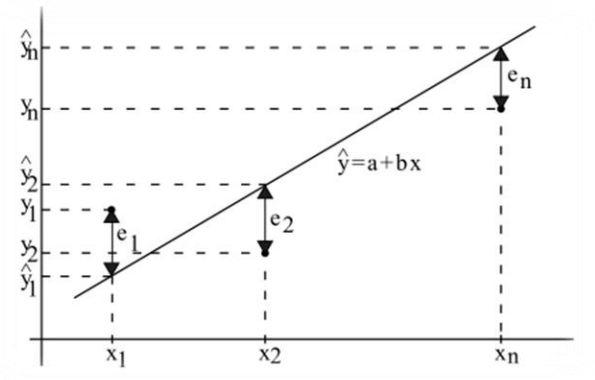


Figura 3-7 – Reta de regressão (Martins, 2019)

Sendo que:

- y: variável dependente observada;
- x: variáveis independentes;

- ŷ: variável dependente estimada;
- e: erro;
- a: intercepto onde x=0;
- b: parâmetro estimado que corresponde ao declive da reta.

O erro, também denominado de erro quadrático, que é uma medida da diferença entre os valores observados e os valores previstos pela regressão linear. Por outras palavras, ele representa a diferença entre a variável dependente real e a variável dependente prevista pela equação da reta de regressão.

Outro fator sobre os erros que deve ser levado em consideração é que existem também aquelas variáveis que podem estar implícitas no estudo e não são contabilizadas no modelo, nesse caso, os erros têm como função captar esses eventos das demais variáveis não presentes, portanto, para (Fávero & Belfiore, 2017), a estimativa da equação que melhor ajuste o ponto deve estabelecer condições para os resíduos (ou erros).

As condições estabelecem que o somatório dos resíduos deve ser zero e o somatório dos resíduos ao quadrado é a mínima possível, sendo que a segunda condição é um indicador da qualidade do ajuste do modelo.

Quanto à determinação dos estimadores de **a** e **b**, a expressão pode ser conferida abaixo:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{e} \quad a = \bar{y} - b\bar{x} \tag{3.2}$$

Onde x e y são as variáveis independentes e dependentes, respetivamente, $\bar{x}$ e $\bar{y}$ são as médias amostrais de x e y, e Σ representa a soma dos valores ao longo de toda a amostra. A estimativa de **b** é obtida pela divisão da covariância entre x e y pela variância de x, ou seja, é uma medida de quanto y varia em relação a x. Já a estimativa de **a** é a média amostral de y subtraída do produto de b pela média amostral de x, ou seja, é o ponto de interceção da linha de regressão com o eixo y, (Morettin & Bussab, 2017).

E por fim, de acordo com os estudos de (Montgomery et al., 2012), é possível listar alguns dos principais cuidados para garantir a qualidade dos resultados obtidos:

- **Identificar possíveis outliers**, que pode ser feita através da análise de resíduos;
- **Verificação de multicolinearidade**: verificar se há multicolinearidade entre as variáveis independentes;
- **Considerar a influência das variáveis explicativas**: algumas variáveis independentes têm mais impacto na variável dependente que outras.

3.3.4.2 Avaliação da Qualidade do ajuste

Para que se verifique a qualidade do ajustamento dos coeficientes à equação de regressão recorre-se a alguns testes de significância, nos quais se objetiva~~m~~ a avaliação dos resultados, no sentido de serem ou não estatisticamente significativos. Normalmente, são realizados os seguintes testes:

- teste t de *student;*
- teste F;
- coeficiente de ajuste;

3.3.4.2.1 Teste t de Student

Para (Fávero & Belfiore, 2017), o **teste t de Student** é um método estatístico utilizado para comparar a média de duas amostras independentes a partir de variabilidade das amostras. Quanto maior for a diferença entre as médias amostrais em relação à variabilidade das amostras, maior será o valor do teste t e mais provável será que a diferença seja estatisticamente significante. A equação que expressa o valor de T, pode ser verificada abaixo:

$$T_{cal} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \tag{3.3}$$

Em que:

- X: média da amostra;
- μ0: média da população;
- S: desvio padrão da amostra;

- n: número de elementos da amostra.

Tendo sido calculado o valor da estatística de T, recorre-se à tabela da distribuição t de *Student*, onde se busca o valor de probabilidade associada à cauda superior, o t-crítico, tendo em conta na escolha da tabela o número de grau de liberdade e o nível de significância definido. A partir de então, compara-se o t-crítico com o valor de T, caso este valor pertença a região crítica, a hipótese nula não é rejeitada.

A Figura 3-8 mostra graficamente como é feita essa repartição na distribuição de probabilidade. No gráfico nota-se a presença de uma região crítica (RC), e uma região de não rejeição (RN), sendo elas delimitadas pelo valor de t-crítico. De acordo com (Henrique & Molin, 2016), a região crítica corresponde a zona de rejeição de hipótese nula, que é determinada de acordo com a taxa de erro do modelo. A linha tracejada representa o valor da hipótese nula.

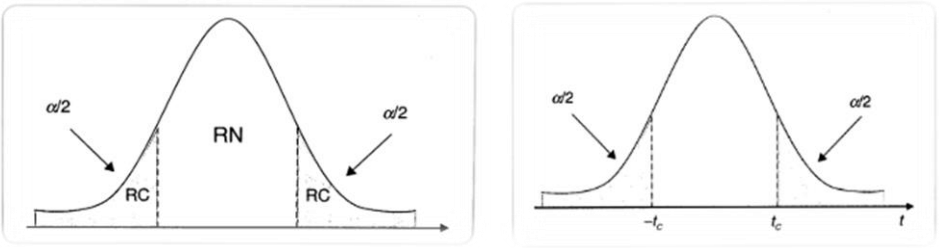


Figura 3-8 - Distribuição de Probabilidade em um teste bicaudal.(Fávero & Belfiore, 2017)

O teste t de Student é utilizado em análises para testar a significância estatística dos coeficientes individuais, se cada coeficiente é significativo ou não. Como este texto é uma revisão bibliográfica, optou-se por uma abordagem mais didática, entretanto no desenvolvimento do trabalho os dados serão tratados de maneira mais automática, através de cálculos feitos nos modelos desenvolvidos em Excel, gerando resultados de forma mais direta, sem a representação gráfica pormenorizada.

Quando o intuito é o teste de significância de cada variável independente, recorre-se ao teste t, por outro lado quando o objetivo é o teste de significância global do modelo, verificação se todas as variáveis em conjunto têm efeito significativo, recorre-se ao teste F.

3.3.4.2.2 Teste F

O teste F é uma medida de significância frequentemente utilizado na análise de variância (ANOVA), conforme apontado por (Henrique & Molin, 2016). Essa análise divide a variabilidade total dos dados em duas ou mais componentes. Quando a divisão é em duas componentes, divide-se entre as causas conhecidas (controláveis) e as causas desconhecidas (erro ou resíduo). Por outras palavras, a ANOVA representa a decomposição da soma total dos quadrados (SQT) em **variância residual e variância de regressão.**

A soma dos quadrados totais representa a variação dos valores médios dos grupos em relação à média geral, e indica o quanto as médias dos grupos diferem entre si, caso a variação de regressão seja grande em relação a variância dos resíduos. Isso significa que existe uma diferença significativa entre as médias dos grupos. Esse tipo de análise permite determinar se as diferenças observadas nas médias são estatisticamente significativas ou se podem ser atribuídas ao acaso.

O próximo passo da análise de variância consiste no cálculo da média dos quadrados, em que se faz necessário a definição dos graus de liberdade para cada componente.

Em conformidade (Fávero & Belfiore, 2017) os graus de liberdade, que representam a liberdade de variação dos dados, são calculados de maneira diferente para cada tido de fonte de variação. Para as variações entre os grupos, o grau de liberdade é dado por (k-1) e para variações dentro dos grupos é dado por (n-k). Sendo assim, a média dos quadrados é calculada pela razão entre a soma dos quadrados e o grau de liberdade respetivo para cada componente.

Levando em consideração que os testes de hipóteses buscam a validação de uma hipótese, por meio de um valor que meça essa confiabilidade, a ANOVA tem como *output* o valor de F, ou significância de F.

O valor do teste F é usado para determinar um valor de p (tabelado de acordo com a distribuição F e nível de significância), que é a probabilidade de se obter um resultado igual ou mais extremo do que o observado, assumindo-se que as amostras são realmente iguais. Se o valor de p for menor que o nível de significância, pode-se rejeitar a hipótese nula.

Esse teste, para (Fávero & Belfiore, 2017), é um teste que possibilita a verificação da existência do modelo, tendo em vista que se todos os parâmetros forem iguais a zero, a variável dependente não alterará em nada com a variação das variáveis independentes, ou seja, aqui é necessário a verificação dos testes de hipótese.

Com a verificação das variações dos erros e de regressão torna-se possível verificar se realmente as populações possuem diferenças significativas ou não. É o que afirmam (Fávero & Belfiore, 2017), quando dizem que a ANOVA tem o objetivo de verificar se as diferenças que existem entre as medias amostrais são significativas em relação as médias populacionais, ou então se essas variações são apenas variabilidade implícita da amostra (erro).

Na Figura 3-9 pode-se encontrar um resumo da anova

fonte	soma dos quadrados	graus de liberdade	média dos quadrados	v.a. F
tratamentos	SQT	$k-1$	$MQT = \frac{SQT}{k-1}$	$F = \frac{MQT}{MQR}$
erros	SQR	$\sum_{j=1}^{k} n_j - k$	$MQR = \frac{SQR}{\sum_{j=1}^{k} n_j - k}$	
TOTAL	STQ	$\sum_{j=1}^{k} n_j - 1$		

Figura 3-9 Resumo ANOVA, (Fávero & Belfiore, 2017)

Em que:

- k: número de parâmetros do modelo;
- n: número de elementos da amostra;
- SQR: soma do quadrado da regressão;
- SQU: soma do quadrado dos resíduos.

Se a variação entre as amostras for grande o suficiente em comparação com a variação dentro de cada amostra, então há evidências de que as amostras têm médias diferentes.

3.3.4.2.3 Coeficiente de ajustamento

Um dos métodos de cálculo do coeficiente de ajustamento é através do coeficiente de determinação (R^2). Este coeficiente é uma medida estatística capaz de medir o poder explicativo do modelo proposto ou então indicar o percentual de variabilidade da variável dependente que é explicado pelo comportamento das variáveis independentes.

Como já foi visto a reta de regressão é determinada através do método dos mínimos quadrados. Para (Fávero & Belfiore, 2017), a soma total dos quadrados (SQT), como é assim determinada, é composta por duas somas de quadrados, sendo elas a soma dos quadrados da

regressão (SQR), que indica a variação de Y em torno da própria média e a soma doa quadrados dos resíduos (SQU) que apresenta a variação de Y que não é explicada pelo modelo.

Dessa maneira o coeficiente de ajuste pode ser calculado através da seguinte expressão:

$$R^2 = \frac{SQR}{SQR + SQU} = \frac{SQR}{SQT}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(u_i)^2} \quad (3.4)$$

O coeficiente de ajustamento mais comumente utilizado é o coeficiente de determinação (R^2). Ele varia de 0 a 1 (0 a 100%) e indica a proporção da variação total a variável dependente que é explicada pelo modelo. Um valor de R^2 próximo a 1 indica que o modelo se ajusta bem aos dados, enquanto um valor próximo a 0 indica que o modelo linear não se ajusta nada bem à distribuição daqueles dados.

Ao comparar modelos de regressão com números diferentes de variáveis independentes, o R^2 é valioso para determinar quão bem cada modelo explica a variabilidade na variável dependente, permitindo escolher o modelo que oferece uma melhor explicação.

3.3.4.2.4 Coeficiente de correlação

Para (Daniels & Minot, 2018), por se tratar de variáveis que podem ou não estar interrelacionadas, é necessário quantificar o grau de associação entre elas, de modo a garantir um resultado com o maior nível de confiabilidade.

Também denominado como coeficiente de correlação de Pearson, este coeficiente tem a função de medir o grau de correlação, bem como a direção da variação entre as variáveis. Na regressão simples, esta análise mede o grau de relacionamento linear entre duas variáveis, no caso da regressão múltipla, faz a medição entre uma variável independente e um conjunto de outras variáveis.

O grau de associação é medido entre o intervalo $-1 \leq r \leq 1$, sendo que 0 representa uma "não correlação", -1 uma correlação negativa, e 1 uma correlação positiva. Esse cálculo é feito através da determinação do coeficiente de correlação que está representado na equação abaixo:

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3.5)$$

Em que:

- r é o coeficiente de correlação;
- x e y são os valores observados;
- n o número de amostras;
- x e y valores médios das observações;

Com o valor de r determinado, seguem as condições de análise:

- Quanto mais próximo de +1, maior o grau de relacionamento linear positivo, ou seja, as variáveis estão positivamente associadas, quando uma aumenta, a outra também aumenta;
- Quanto mais próximo de -1, maior o grau de relacionamento linear negativo, ou seja, as variáveis estão negativamente associadas, quando uma aumenta, a outra diminui;
- Quanto mais próximo de zero, menor o relacionamento linear entre as variáveis;

No contexto de análise de regressão, é fundamental que a variável resposta apresente uma correlação considerável com as variáveis independentes significativas. Isso garante que as variáveis independentes selecionadas têm um impacto relevante na explicação da variabilidade da variável resposta. Por outro lado, é muito importante que as variáveis independentes sejam pouco correlacionadas entre si, isso evita a multicolinearidade, o que poderia dificultar a interpretação dos coeficientes de regressão, uma vez que as variáveis independentes estariam explicando partes semelhantes da variabilidade da variável resposta.

4 Análise do comportamento da procura

Tendo sido abordado nos capítulos anteriores os objetivos deste trabalho e as metodologias a serem utilizadas, neste capítulo será abordado o estudo de caso. A primeira parte do trabalho consistiu na análise detalhada dos relatórios que são normalmente produzidos pela empresa, os *dashboards* em Power BI. Para além desses relatórios descritivos, foram elaborados pela autora alguns gráficos para análise complementar de acordo com os objetivos do estágio.

A análise dos dados da procura foi realizada em um período que compreende janeiro de 2019 a dezembro de 2022, já que os anos anteriores não existia uma quantidade significativa de observações que permita o estudo estatístico. Inicialmente foram verificados todos os *dashboards* com a intenção de se obter uma visão de maior abrangência no estudo, entretanto como o objetivo era estudar o comportamento da procura e detectar suas possíveis causas de variação, o *dashboard* de maior peso na análise foi o de passageiros.

4.1 Descrição dos dados

Durante a análise dos gráficos de números de passageiros por mês, o primeiro aspecto a chamar a atenção foi **a presença da sazonalidade**. O gráfico do número de passageiros anual segue um padrão muito parecido em todos os anos, podendo ser descrito com uma queda na procura entre os meses de janeiro a abril, segue para dois picos, um no mês de maio com queda em junho, e outro em julho com queda em agosto seguindo para uma alta em outubro, diminuindo até dezembro, como mostra o Gráfico 4-1, em que a linha dos quatro anos segue um padrão muito parecido, com oscilações quase que semelhantes nos mesmos meses, diferindo-se apenas em algumas épocas pontuais.

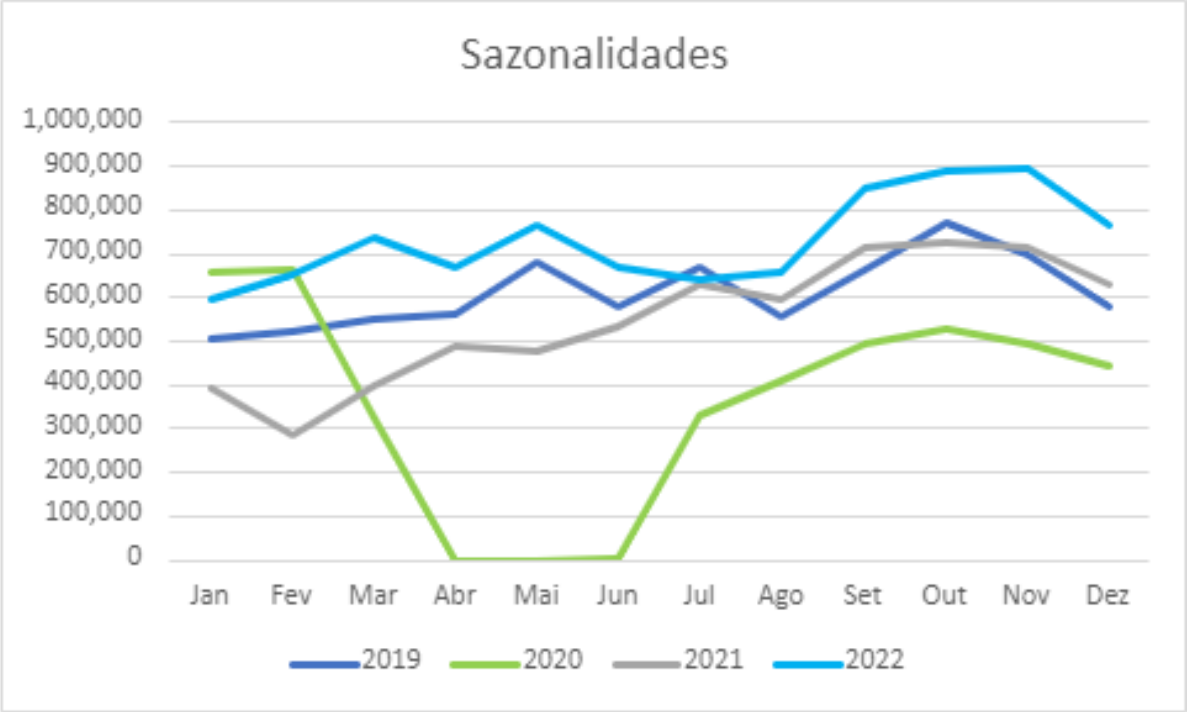


Gráfico 4-1 Comportamento sazonal da procura anual.

Além do efeito da sazonalidade, ainda no Gráfico 4-1 pode-se notar um vazio nos dados no segundo trimestre de 2020, que se deve ao período de isolamento social mais crítico devido a pandemia do **Corona vírus**. De acordo com (ONU News, 2020), em 11/03/2020 a Organização Mundial da Saúde declarou o covid-19 uma pandemia, o que implicou o isolamento social como principal medida de contenção do vírus. Dessa maneira, as pessoas começaram a sair de casa só por motivos essenciais e começaram a realizar todas as suas atividades de casa. Esse isolamento pode ser nitidamente verificado no gráfico, com uma queda e posteriormente uma recuperação no segundo semestre.

O ano de 2020 começou com um aumento no número de passageiros em relação ao ano anterior, sentiu os impactos da covid, e começou a apresentar sinais de recuperação a partir de julho de 2020, com uma pequena tendencia de alta. Essa tendência estendeu-se para o ano de 2021 que ainda no primeiro semestre passou pela diminuição das restrições e seguiu em tendencia de alta até o fim do ano. Um dos pontos a ser analisado na regressão é a quantificação do **impacto da corona vírus na procura do TP em Cascais.**

4.1.1 Investigação da sazonalidade

A partir do relatório da sazonalidade procurou-se encontrar os principais motivos para a ocorrência dessas variações, o que levou a uma caracterização dos meses, no que empiricamente pudesse ser o aumento ou variação da procura, como por exemplo a diminuição do número de passageiros em dia não úteis e o aumento em época ao início do ano letivo.

Quanto à presença de feriados, nota-se que os meses de abril, junho e dezembro são os meses com a maior incidência como demonstrado no Quadro 4-1

Quadro 4-1 - Lista dos feriados em Portugal em 2022.

Data	Dia	Feriados
1 de janeiro	Sábado	Ano Novo
15 de abril	Sexta-feira	Sexta-Feira Santa
17 de abril	Domingo	Páscoa
25 de abril	Segunda-feira	Dia da Liberdade
1 de maio	Domingo	Dia do Trabalhador
10 de junho	Sexta-feira	Dia de Portugal
16 de junho	Quinta-feira	Corpo de Deus
15 de agosto	Segunda-feira	Assunção de Nossa Senhora
5 de outubro	Quarta-feira	Implantação da República
1 de novembro	Terça-feira	Dia de Todos os Santos
1 de dezembro	Quinta-feira	Restauração da Independência
8 de dezembro	Quinta-feira	Imaculada Conceição
25 de dezembro	Domingo	Natal

Fonte: Adaptado de Calendarr, 2022

Quando esses feriados caem próximos aos finais de semana como se vê em junho e dezembro, nesse ano específico, as universidades e algumas empresas costumam “fazer ponte” com o feriado, o que faz com que grande parte das pessoas tiram esses dias de férias para desfrutar de um final de semana prolongado.

Quanto ao aumento da procura, o ano letivo em Portugal tem início em setembro. O primeiro semestre decorre de setembro a janeiro, e o segundo semestre de fevereiro a junho. Os meses de julho e agosto são os meses de férias escolares e cada semestre conta com mais duas semanas de férias. No primeiro semestre, existem duas semanas de folga: as semanas de Natal

e ano novo, e no segundo semestre uma semana de férias na Páscoa. Facto este que contribui em parte para a queda nos meses de abril e dezembro.

A partir da observação destes dados ficou decidido que uma das potenciais variáveis explicativas da variação da procura com a sazonalidade seria o **período letivo.**

Quando se pensa na componente sazonal, pensa-se na divisão por períodos, além da análise do período letivo, o estudo dos efeitos decorrentes das estações do ano pode ser também significativo, levando em consideração que é notável a mudança do comportamento das pessoas ao longo das estações do ano. Nos meses mais frios as pessoas têm tendência a efetuar menos atividades ao ar livre e ficam mais em suas casas. O que é oposto no verão, onde as pessoas saem mais pelo tempo, e pela extensão de horas claras durante o dia.

Com base nestas constatações foi decidido que a variação da procura ao longo do ano poderia ser também uma das potenciais variáveis explicativos a ser analisada.

4.1.2 Análise de picos

Ao analisar os gráficos da procura anual pode-se notar alguns picos que contribuíram para uma tendencia de alta, o que deve ser levado em consideração. Quando se analisa o ano de 2019, nota-se a partir de abril uma grande alta na procura de passageiros que se manteve, seguindo as variações de sazonalidade, uma tendencia de alta até os primeiros meses de 2020 antes das restrições do isolamento social devido à covid, e um outro pico no mês de outubro que como já foi mencionado, pode estar relacionado ao período letivo. O Gráfico 4-2 mostra esse pico em vermelho.

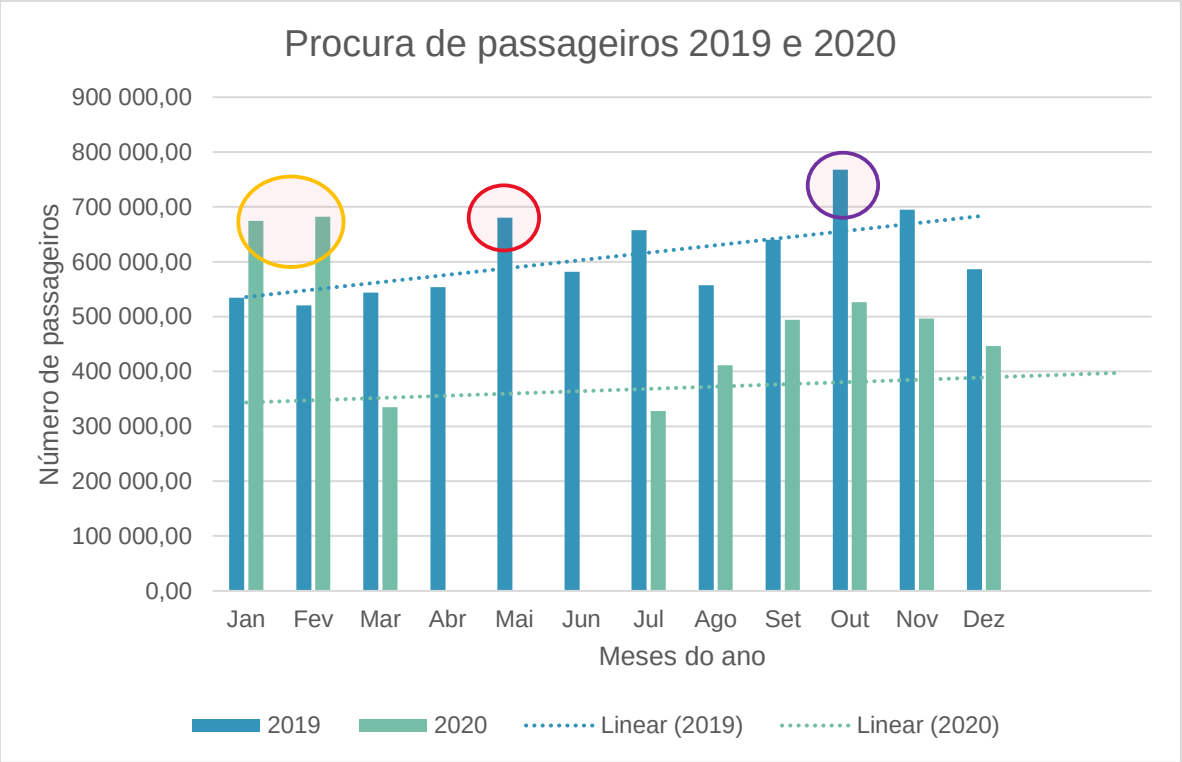


Gráfico 4-2 Procura mensal de passageiros nos anos de 2019 e 2020.

Uma possível causa para este aumento foi o aparecimento a partir de 1 de abril de 2019 na Área Metropolitana de Lisboa (AML) do **título de transporte Navegantes**. Esse título, foi uma medida coordenada adotada por todos os operadores de transporte da AML, e permite ao utilizador o uso de um de um passe mensal no valor 30€ ou 40€, que pode ser utilizado de forma ilimitada no período de 30 dias a contar do início do mês (Portal Viva, n.d.).

Esse título representa uma enorme redução no preço do transporte público comparada a situação anterior na qual os utentes precisavam de adquirir passes individuais a um preço elevado, ou títulos de viagem que podiam rondar os 2€ por viagem, contabilizando um alto despendimento financeiro no fim no mês. Sendo assim, os efeitos da introdução desse passe serão estudados.

Ainda nessa análise gráfica, agora incluindo todos os anos com dados, verificou-se um aumento na procura que se estendeu para o início de 2020. É sabido que no dia 01 de janeiro de 2020 passou a haver um novo título de transporte **Viver Cascais**, onde estudantes, trabalhadores e residentes de Cascais podem usufruir do transporte público de forma gratuita. A disponibilidade desse título leva a pensar que muito possivelmente houve um aumento na procura por transportes públicos, e o Gráfico 4-3 mostra esse aumento (retângulo amarelo). Sendo assim, uma das potenciais variáveis a ser analisada foi a introdução desse título.

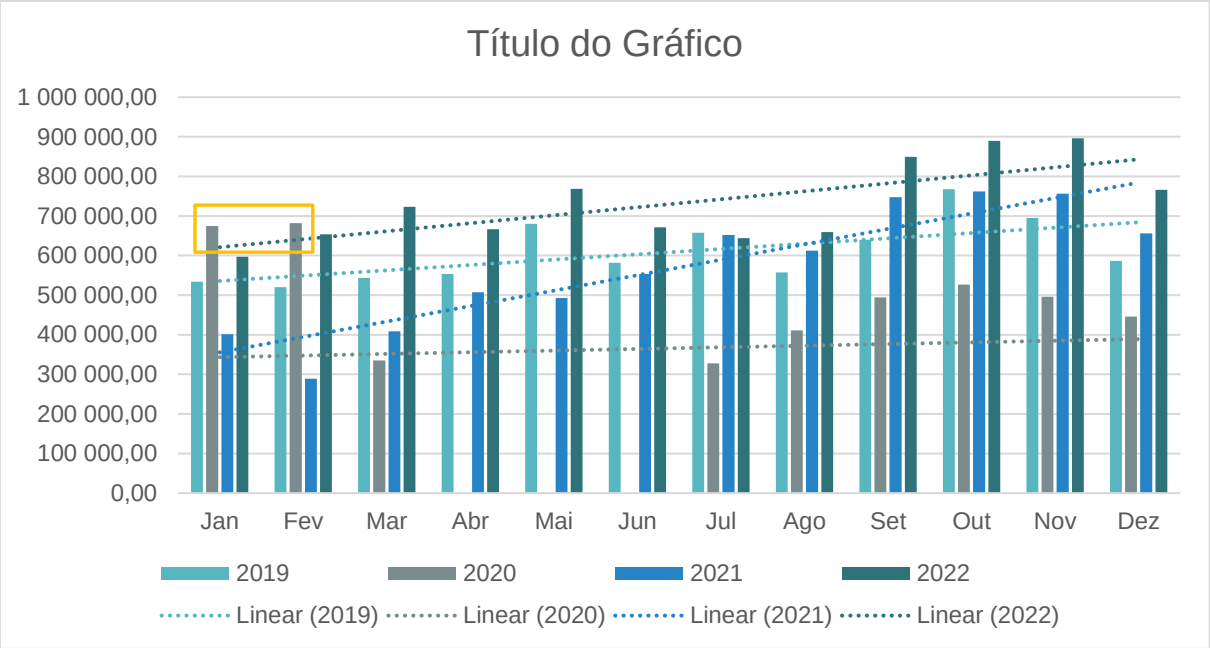
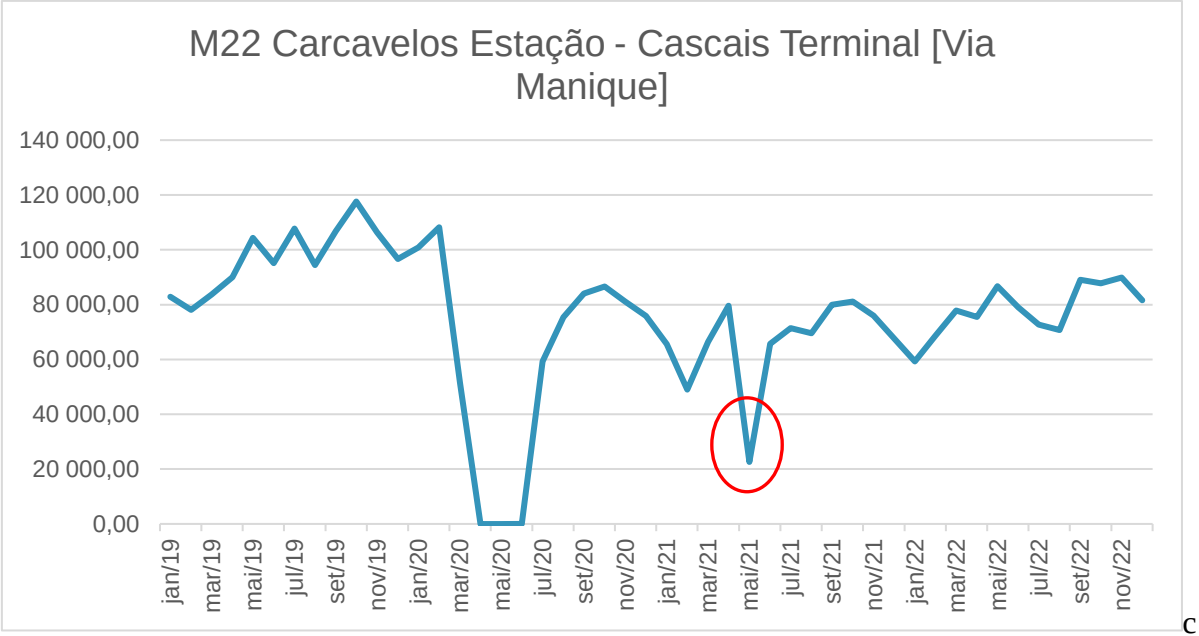


Gráfico 4-3 Histograma das procuras anuais.

Em resumo, para uma primeira análise exploratória da procura do transporte público serão consideradas os seguintes fatores: Covid, Título Navegantes, Título Viver Cascais, Período letivo e Época do ano. Na sequência da determinação das variáveis potenciais, iniciou-se o processo de identificação de *outleiers* e caracterização dessas variáveis.

4.1.3 Identificação de Outliers

A identificação dos outleiers foi efetuada a partir da **construção dos gráficos**, onde as maiores discrepâncias foram analisadas, verificadas e conferidas novamente e através do cruzamento de informações em que os números e a lógica se deveriam sempre manter equilibrados. Por exemplo, a linha M22 apresentava uma queda de procura muito "inexplicada" em maio de 2021. Sendo assim recorreu-se à tabela de base dados para verificar que havia um erro de digitação nos dados.



c

Gráfico 4-4 Representação de *outlier* em abril de 2021

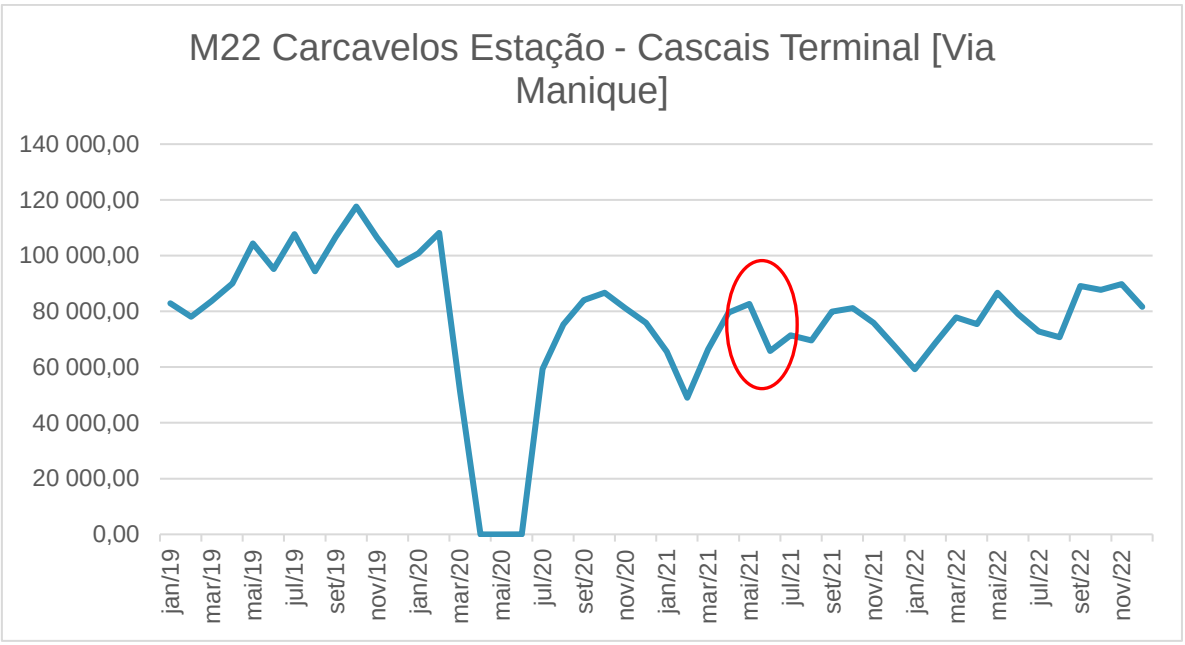


Gráfico 4-5 Representação da correção do outlier.

Além disso notou-se que como os **dados que foram coletados no período de isolamento social**, para algumas linhas poderão não ser expressivos ou até mesmo verdadeiros, porque devido às regras do isolamento, as portas dianteiras dos veículos não eram abertas para a recolha de passageiros, para segurança do motorista e deles mesmos. Por vezes o número de

obliterações diária de uma carreira contabilizavam uma dezena, ou menos, de registros. Nesses casos esses dados foram retirados da base de dados, sendo considerados *outleiers*.

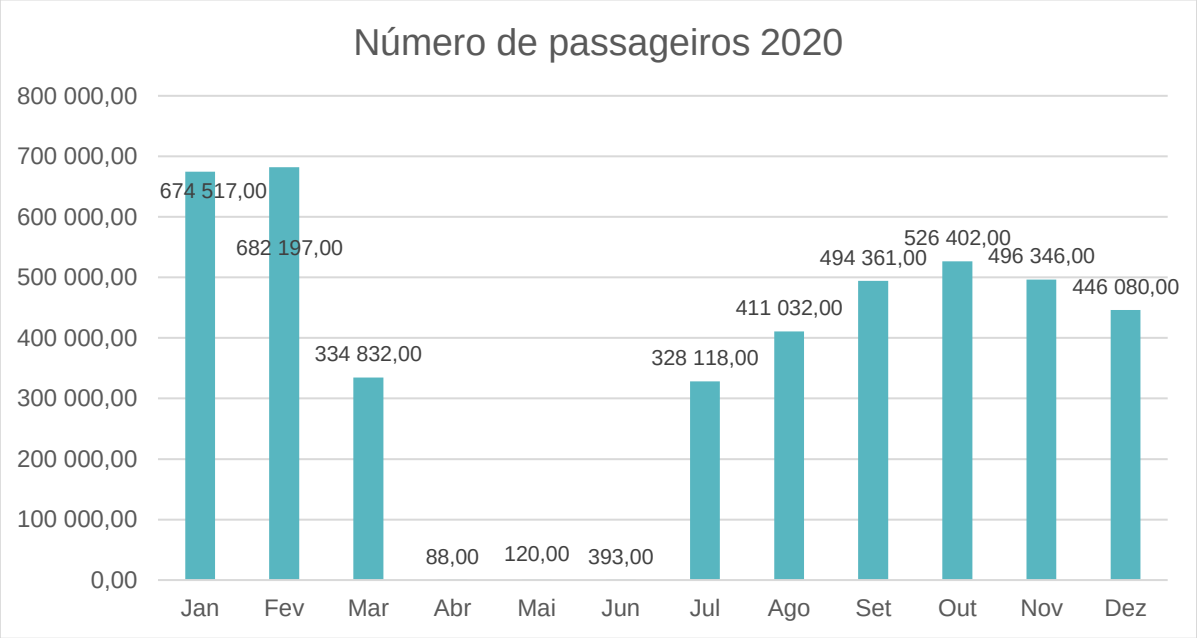


Gráfico 4-6 Representação da baixa quantidade de passageiros nas restrições da descoberta da covid.

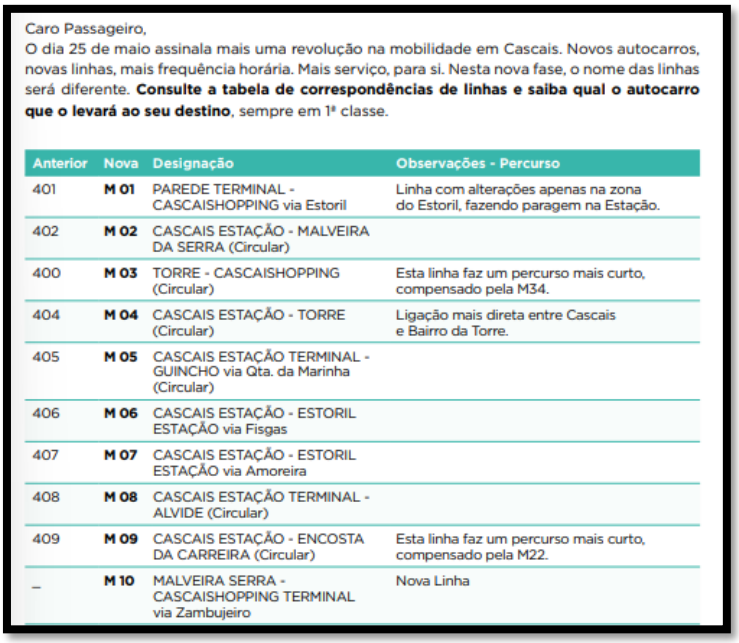
Em síntese, toda a base de dados de informação sobre a procura nas várias linhas de TP no concelho de Cascais foi sujeita a uma rigorosa pesquisa de *outliers* para permitir melhorar a robustez dos dados e das análises subsequentes.

4.1.4 Caracterização da rede de autocarros

Considerando a forma como os dados foram disponibilizados pela Cascais Próxima, que corresponde à desagregação dos números de passageiros por linhas, resolveu-se então caracterizar essas linhas de forma individual, para que pudessem ser posteriormente agregadas em grupos semelhantes para uma maior confiabilidade da pesquisa. Além disso, seguindo a ideia de agregação por semelhança, buscou-se também outros tipos de agregação, como por exemplo região geográfica, função ou proximidade a equipamentos públicos, tais como universidades.

Para essa análise exploratória de dados das linhas, o primeiro ponto a ser levado em consideração foram as **mudanças na oferta (veículos e/ou linhas)** que ocorreram nos últimos anos. No ano de 2021 houve uma troca na operação da frota de autocarros em Cascais, antes gerida pela Scotturb, que no final de maio passou a ser gerida pela Cascais Próxima e pela Martin SA. Essa troca de operação levou a algumas mudanças estruturais na frota de autocarros, nomeadamente a introdução de novas linhas, mudança de nomenclatura das linhas, mudança de percurso de algumas linhas existentes e mudanças de horários.

A Figura 4-1 mostra parte do arquivo informativo sobre a mudança de operação.



Caro Passageiro,
O dia 25 de maio assinala mais uma revolução na mobilidade em Cascais. Novos autocarros, novas linhas, mais frequência horária. Mais serviço, para si. Nesta nova fase, o nome das linhas será diferente. **Consulte a tabela de correspondências de linhas e saiba qual o autocarro que o levará ao seu destino**, sempre em 1ª classe.

Anterior	Nova	Designação	Observações - Percurso
401	**M 01**	PAREDE TERMINAL - CASCAISHOPPING via Estoril	Linha com alterações apenas na zona do Estoril, fazendo paragem na Estação.
402	**M 02**	CASCAIS ESTAÇÃO - MALVEIRA DA SERRA (Circular)	
400	**M 03**	TORRE - CASCAISHOPPING (Circular)	Esta linha faz um percurso mais curto, compensado pela M34.
404	**M 04**	CASCAIS ESTAÇÃO - TORRE (Circular)	Ligação mais direta entre Cascais e Bairro da Torre.
405	**M 05**	CASCAIS ESTAÇÃO TERMINAL - GUINCHO via Qta. da Marinha (Circular)	
406	**M 06**	CASCAIS ESTAÇÃO - ESTORIL ESTAÇÃO via Fisgas	
407	**M 07**	CASCAIS ESTAÇÃO - ESTORIL ESTAÇÃO via Amoreira	
408	**M 08**	CASCAIS ESTAÇÃO TERMINAL - ALVIDE (Circular)	
409	**M 09**	CASCAIS ESTAÇÃO - ENCOSTA DA CARREIRA (Circular)	Esta linha faz um percurso mais curto, compensado pela M22.
–	**M 10**	MALVEIRA SERRA - CASCAISHOPPING TERMINAL via Zambujeiro	Nova Linha

Figura 4-1 Tabela de correspondência de linhas (Câmara Municipal de Cascais, 2021)

Com base nesse documento, as linhas foram divididas em **linhas novas e linhas antigas**, sendo as antigas, apenas aquelas em que a rota fosse similar ou com poucas alterações, para que não houvesse grandes diferenças no trajeto que pudessem alterar ou diminuir a significância estatística do estudo.

Na sequência, outro ponto da segregação foi a **esfera geográfica**, tendo em vista que Cascais está dividido entre 4 regiões. Sendo Cascais e Estoril as regiões mais populosas e Carcavelos uma região que tem apresentado um maior crescimento nos últimos anos, e as regiões mais afastadas do litoral, que são Alcabideche e São Domingos de Rana como pode ser observado na Figura 4-2.

Figura 4-2 As regiões de Cascais, (Câmara Municipal de Cascais, 2016).

Dessa maneira, a primeira caracterização das linhas foi por zona leste e oeste, sendo definida por Zona Cascais, o que inclui U.F Cascais e Estoril e Alcabideche e Zona Carcavelos, que inclui U.F Carcavelos e Parede e São Domingos de Rana.

Uma outra caracterização das linhas, foi a sua **classificação quanto ao traçado e função**. Com a análise do traçado de todas as linhas, observou-se a presença de três tipos de linhas como mostra a Figura 4-3. Quanto ao traçado, existem linhas locais e circulares, e locais que são circulares, como é o caso da M25 na zona de Carcavelos. Quanto a sua função elas enquadram-se na classificação de alimentadoras ou não alimentadoras, como é o caso da M18 e M28 respectivamente. **Sendo segregadas então em linhas locais/circulares e alimentadoras**.

O traçado de todas as linhas encontra-se na totalidade no Anexo B – Traçado rota dos autocarros.

Figura 4-3 - Diferentes tipologias de linhas na rede da Cascais Próxima.

Para o aprofundamento do estudo individualizado partiu-se para a **análise dos parâmetros** de cada linha, tendo sido calculada a sua procura média geral ao longo dos anos, o desvio padrão e a variância. A análise dos parâmetros tem a intenção de documentar quais as linhas que possuem uma maior ou menor variabilidade em relação as outras. Todos os gráficos se encontram em no Anexo A – . No Gráfico 4-7 apresenta-se o exemplo da linha M22 que é uma linha com variabilidade média.

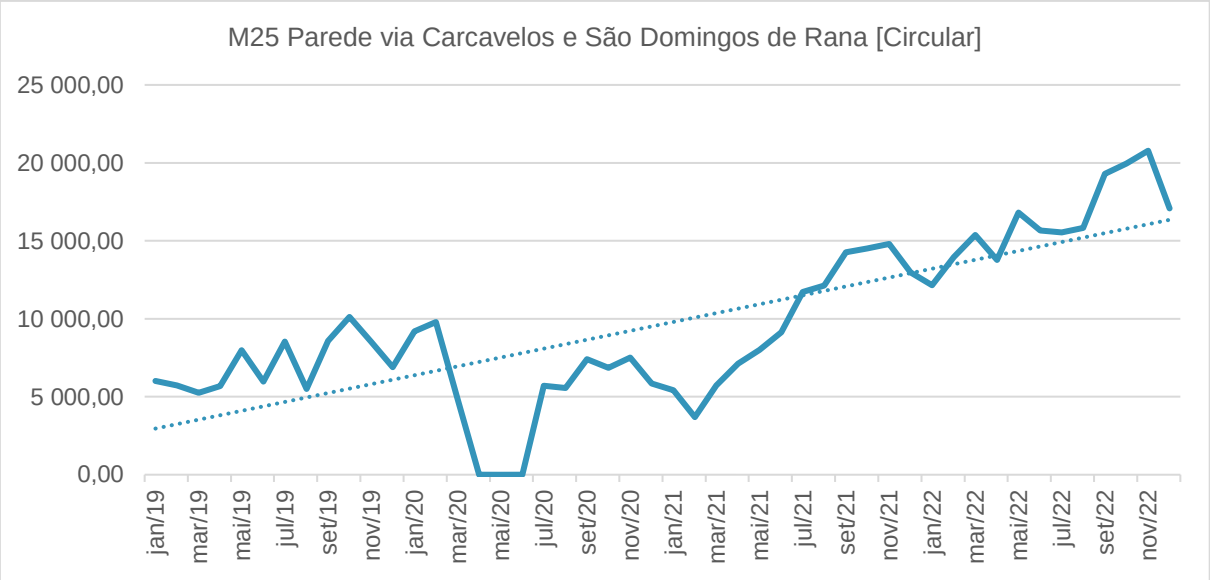


Gráfico 4-7 Gráfico do período para a linha M22

No Quadro 4-2 apresentam-se a título de exemplo os valores calculados para obtenção do grau de variabilidade da linha M22.

Quadro 4-2 Quadro cálculo dos parâmetros dos das observações

Carreira	Trajeto	Antiga nomenclatura	Média	Desvio Padrão	Variabilidade (%)
M22	Carcavelos Estação - Cascais Terminal [Via Manique]	**462**	76 531,45	24 906,72	33%

Com base na perspectiva anterior foram adotadas pela autora quatro tipos de classificações, no quesito variabilidade das observações, tendo como base a divisão em quartis. Essa classificação pode ser observada a seguir:

- Baixa variabilidade - 0 a 25%
- Média variabilidade - 26 a 50%
- Alta variabilidade - 51 a 75%
- Muito elevada variabilidade - 76 a 100%

4.2 Definição das hipóteses a investigar

Levando em consideração as análises exploratórias que foram descritas no início deste capítulo e os objetivos do trabalho, chegou-se à conclusão que as hipóteses a serem testadas em relação às variáveis a analisar seriam:

- Na regressão global, incluindo todas as linhas, quais as variáveis têm um maior poder explicativo para a variação da regressão?
- As linhas que servem as universidades são mesmo mais afetadas em período de início de ano letivo?
- Existe uma diferença significativa entre as linhas alimentadoras e circulares em relação a procura?
- A região geográfica, Cascais ou Carcavelos interfere de alguma maneira com a procura?

4.3 Regressão linear

Tendo sido verificadas e descritas as características dos dados e colhidas informações através de gráficos, documentos e mapas, por um lado, e tendo sido identificadas as hipóteses a investigar em relação à procura, pelo outro, inicia-se o processo de preparo para o desenvolvimento da regressão linear, que quantificará e qualificará cada variável da equação de regressão.

4.3.1 Estrutura do modelo de regressão

A proposta de estrutura do modelo a utilizar nesta investigação inicial corresponde a um modelo de regressão linear baseado em variáveis do tipo *dummy* com eventual potencial explicativo do comportamento da procura.

Pretende-se, pois, identificar a capacidade explicativa das variações da procura (Y) do TP em Cascais a partir das seguintes variáveis explicativas potenciais: características do período letivo (X1), períodos covid (X2), introdução do título de transportes Navegantes na AML (X3),

introdução do título de transportes Viver Cascais pela Cascais Próxima (X4) e época do ano (X5).

A equação inicial a investigar é a seguinte:

$$\boldsymbol{Y(procura)} = \boldsymbol{x1}\ (\boldsymbol{período\ letívo}) + \boldsymbol{x2}\ (\boldsymbol{covid}) + \boldsymbol{x3}\ (\boldsymbol{navegantes}) + \boldsymbol{x4}\ (\boldsymbol{viver\ cascais}) + \boldsymbol{x5}\ (\boldsymbol{época\ do\ ano}) + \boldsymbol{b}$$

A identificação destas variáveis esteve por um lado relacionada com o tipo de dados disponibilizados pela Cascais Próxima e por outro com as hipóteses de investigação definidas em 4.2.

Numa segunda fase, posterior à criação deste conjunto pioneiro de modelos explicativos, poderão vir a ser identificadas novas questões de investigação e vir a ser incluídas novas variáveis com potencial explicativo, do tipo *dummy*, ou mesmo de variação cronológica (como a inclusão de um fator cronológico de crescimento da procura). No entanto, essa abordagem saí fora do âmbito dos trabalhos desenvolvidos no presente estágio.

4.3.2 Classificação das variáveis

Quanto à classificação das variáveis, a variável dependente (procura), por ser uma contagem do número de passageiros por mês, é caracterizada como variável quantitativa contínua. As variáveis independentes, por serem qualitativas, representando a ausência ou presença da uma característica, tiveram que passar pelo processo de codificação *dummy* para que pudessem entrar no modelo de regressão, em que o número de variáveis binarias corresponde a (n-1), sendo n o número de categorias, assim como mostra o Quadro 4-3

Quadro 4-3 Codificação das variáveis qualitativas

Variável	Característica	Codificação
Covid	Restrição	1
	Não restrição	0
Período letivo	Aulas	1
	Ferias	0
T. navegantes	Sim	1
	Não	0

Variável	Característica	Codificação
T. Viver Cascais	Sim	1
	Não	0
Época do ano	Verão	1
	Inverno	0

4.3.3 Amostragem

O segundo passo do procedimento de elaboração de uma regressão é a amostragem. Inicialmente, buscou-se fazer uma análise de regressão com todos os dados para uma visão mais abrangente do conjunto de informações e das suas relações. Denominámos essa regressão por regressão A.

Em seguida optou-se por ensaiar uma agregação de linhas que a autora julgou ser pertinente para análise e estudo, em conformidade com item 4.1.2, no qual foram analisados três tipos de amostras:

- Regressão B: região geográfica;
- Regressão C: tipologia;
- Regressão D: proximidade à faculdade NOVA SBE.

O Quadro 4-4 mostra a relação de todas as linhas e sua classificação quando à tipologia, localidade e condição, sendo marcadas em verde as linhas novas, já que com a adição dessas linhas, a configuração da rede ganha outro formato. Este quadro foi utilizado como base para agregação das amostragens desejadas.

Quadro 4-4 Relação das linhas e sua classificação

Linhas	Variabilidade	Tipologia	Região	Condição	Proximidade NOVA SBE
M01	Média	Alimentadora	Cascais	Linha com alterações	
M02	Média	Alimentadora/Circular	Cascais	Linha antiga	
M03	Média	Circular	Cascais	Linha com alterações	
M04	Média	Alimentadora/Circular	Cascais	Linha com alterações	
M05	Média	Alimentadora/Circular	Cascais	Linha antiga	
M06	Média	Alimentadora	Cascais	Linha antiga	
M07	Média	Alimentadora/Circular	Cascais	Linha antiga	
M08	Média	Alimentadora	Cascais	Linha antiga	
M09	Média	Alimentadora/Circular	Cascais	Linha com alterações	
M10	Média	Alimentadora/Circular	Cascais	Linha nova	

Linhas	Variabilidade	Tipologia	Região	Condição	Proximidade NOVA SBE
M11	Média	Alimentadora	Cascais	Linha antiga	
M12	Média	Alimentadora	Cascais	Linha com alterações	
M13	Média	Alimentadora	Cascais	Linha com alterações	
M14	Média	Alimentadora/Circular	Cascais	Linha antiga	
M15	Média	Alimentadora/Circular	Cascais	Linha antiga	
M16	Alta	Local	Carcavelos	Linha antiga	Sim
M17	Pouca	Alimentadora	Cascais	Linha nova	
M18	Média	Alimentadora	Cascais	Linha nova	
M19	Média	Alimentadora/Circular	Cascais	Linha com alterações	
M20	Pouca	Alimentadora	Cascais	Linha nova	
M21	Média	Alimentadora	Carcavelos	Linha nova	Sim
M22	Média	Alimentadora	Mista	Linha nova	
M23	Média	Alimentadora	Mista	Linha com alterações	Sim
M24	Média	Alimentadora	Carcavelos	Linha antiga	Sim
M25	Alta	Alimentadora	Carcavelos	Linha com alterações	
M26	Média	Alimentadora	Carcavelos	Linha com alterações	
M27	Média	Circular	Cascais	Linha antiga	
M28	Muito alta	Local	Cascais	Linha com alterações	
M29	Média	Alimentadora	Mista	Linha nova	
M30	Média	Alimentadora	Carcavelos	Linha nova	
M31	Média	Alimentadora	Mista	Linha nova	
M32	Média	Alimentadora	Carcavelos	Linha nova	
M33	Média	Circular	Carcavelos	Linha nova	
M34	Média	Alimentadora	Cascais	Linha nova	
M35	Média	Alimentadora	Carcavelos	Linha nova	
M36	Média	Alimentadora	Carcavelos	Linha nova	
M37	Alta	Circular/Local	Carcavelos	Linha antiga	
M38	Alta	Alimentadora	Cascais	Linha antiga	
M39	Alta	Alimentadora	Cascais	Linha antiga	
M40	Muito alta	Local	Cascais	Linha antiga	
M41	Alta	Circular/Local	Carcavelos	Linha antiga	
M42	Média	Local	Carcavelos	Linha nova	
M43	Média	Circular	Cascais	Linha nova	
M44	Alta	Alimentadora	Cascais	Linha nova	

4.3.4 Nível de significância

Para o desenvolvimento das análises de regressão, optou-se pela escolha de um nível de significância de 5%, conferindo ao modelo uma maior robustez, com uma confiabilidade de 95%. A razão da escolha deste percentual se deve ao fato de que este nível é considerado um padrão razoável e amplamente aceito nas mais diversas áreas de estudo.

4.3.5 Teste de correlação

Para a execução do teste de correlação, recorreu-se ao Excel para cálculo dos coeficientes de cada variável, podendo ser analisados os resultados no Quadro 4-5:

Quadro 4-5 Teste de correlação

Coeficiente de correlação de Pearson					
	Navegantes	*Viver Cascais*	*Covid*	*Época Ano*	*Período Letivo*
Navegantes	1,0000				
Viver Cascais	**0,4432**	1,0000			
Covid	0,1336	0,3015	1,0000		
Época Ano	0,2500	-0,0403	-0,1336	1,0000	
Período Letivo	-0,1243	-0,0175	-0,0581	**-0,4971**	1,0000

Em conformidade com a classificação para o coeficiente de correlação de Pearson, as variáveis Viver Cascais e Navegantes apresentam uma correlação com algum significado, juntamente com Período Letivo e Época do ano. Esse fator indica a presença de uma colinearidade média entre as variáveis, ou seja, elas estão relativamente relacionadas entre si, entretanto por ser uma relação de intensidade média, poderá não interferir assim tanto na interpretação dos coeficientes. Essa situação será aferida no desenvolvimento das análises. Para as outras variáveis que apresentam um baixo coeficiente, é bom resultado para a regressão.

4.3.6 Testes de regressão

Como foi dito, os testes de regressão foram separados conforme o que a autora julgou ser o motivo para a variação da procura de passageiros. Em primeiro lugar foi efetuada uma análise de regressão global, para se ter uma base da quantificação da participação de cada variável independente na explicação da variação da procura. Em seguida foram ensaiadas as regressões por grupo, para verificação se aquelas segmentações à priori fazem sentido ou não.

4.3.6.1 Análise da população completa

A primeira análise de regressão, foi efetuada utilizando todas as observações para todas as linhas. Ao se realizar a regressão, a primeira constatação foi em relação à constante “b”, a interseção da reta de regressão com o eixo YY, o valor esperado de Y quando as variáveis independentes são zero.

Quando os dados estão estruturados em séries temporais, e se considera a variável “b”, assume-se que existe um valor esperado para a variável dependente antes do início da série temporal, o que não faz muito sentido já que o valor da variável dependente pode ser influenciado por fatores sazonais.

Além disso, considerar a inclinação “m” igual a zero, seria o mesmo que afirmar que o modelo seria uma reta horizontal constante, indicando que a variável y não muda em relação ao tempo, tornando-o em um modelo restrito e inadequado. Dessa maneira, nas análises não se fixa o parâmetro mudo em 0.

O procedimento de análise para todas as regressões será feito da seguinte maneira:

- Regressão linear com todas as variáveis para a amostragem selecionada;
- Análise dos parâmetros de regressão como o coeficiente de correlação múltipla, análise de variância e análise de significância de cada variável;
- Repetição da regressão linear somente com as variáveis que apresentaram significância nessa primeira análise;
- Análise gráfica;
- Resultados.

Esse procedimento será repetido em todas as análises e será progressivamente sendo descrito de forma mais resumida.

Quadro 4-6 Regressão população geral

SUMÁRIO DOS RESULTADOS						
Estatística de regressão						
R múltiplo	0,974486502					
Quadrado de R	0,949623942					
Quadrado de R ajustado	0,919586336					
Erro-padrão	148492,8274					
Observações	45					
ANOVA						
	gl	*SQ*	*MQ*	*F*	*F de significância*	
Regressão	5	1,66264E+13	3,32528E+12	150,805598	2,04285E-24	
Residual	40	8,82005E+11	22050119783			
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	440389,8687	80599,95391	5,463897277	2,67428E-06	277491,2854	603288,4521
ViverCascais	37765,76558	58569,32354	0,644804538	0,522735015	-80607,25286	156138,784
Covid19	-246014,5786	58923,75536	-4,175134071	0,000156326	-365103,9305	-126925,2267
EpocaAno	39529,22757	51524,35797	0,767194957	0,447470553	-64605,38432	143663,8395
PeriodoLetivo	216988,4023	53018,12695	4,092721013	0,000200904	109834,7706	324142,0339

Partindo para as análises, quanto ao sumário de resultados, o **coeficiente de correlação múltipla (R)** apresenta um ótimo valor para a medida de relação linear entre as variáveis independentes e a variável dependente, apresentando uma correlação bastante boa. O quadrado de R também apresenta um bom resultado para o percentual de variabilidade de Y que é explicado pelos X. O valor calculado para o quadrado de R ajustado de 91,95 % também apresenta resultados coesos.

O teste de **análise de variância** informa os valores do F calculado e o F significância. Para que se rejeite a hipótese nula e confirme a existência do modelo os resultados dessas medidas dever ser respectivamente maior que um e menor que o valor do nível de significância escolhido. Dessa maneira um F=150,80 e F de significância = 2,04x10^-24 indicam a existência do modelo.

Quanto a análise da **significância de cada variável**, através do valor P, nota-se que as variáveis Viver Cascais e Época do ano não são significativas para o modelo, apresentando um valor de p maior que 0,05. Isso ocorre porque quando se verifica o período de incidência da variável Viver Cascais no Gráfico 4-8 -Incidência Viver Cascais, nota-se que houve um aumento na procura nos anos seguintes, entretanto esse aumento não é explicado por essa variável, o que impede de afirmar que esse aumento pode ter sido em decorrência do título.

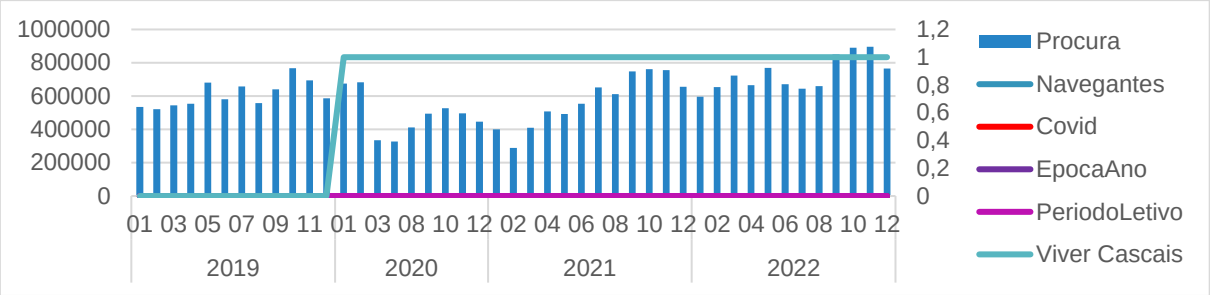


Gráfico 4-8 -Incidência Viver Cascais

Para a variável época do ano, como mostra o Gráfico 4-9 , acontece algo semelhante. No período de incidência do verão, a variável não explica a variação na procura. São períodos que apresentam aumento no número de passageiros de comparado aos meses anteriores, com exceção do período de incidência de covid, entretanto a variável não é significativa, portanto, o verão não explica este aumento.

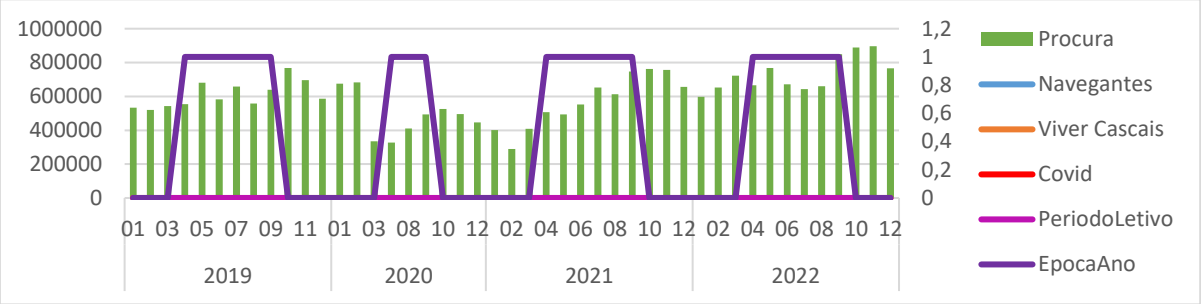


Gráfico 4-9 - Incidência Época do Ano

Quanto as variáveis Navegantes, Covid e Período letivo, estas apresentam resultados significativos que explicam a regressão sendo classificadas como 1º, 2º e 3º lugar respectivamente na explicação da regressão. Nesse âmbito, repete-se o procedimento de regressão linear apenas com essas variáveis significativas, como mostra o Quadro 4-7.

Quadro 4-7 Regressão geral apenas com variaveis significantivas

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

Nota-se que o valor de do coeficiente de determinação (R) aumentou ligeiramente e o valor de F calculado aumentou significativamente. O aumento do coeficiente R pode ser explicado devido ao melhor ajuste do modelo, ao se remover as variáveis que não são explicativas resulta-se em um melhor ajuste dos pontos observados, o que reflete no valor de R.

Desta maneira, conclui-se que a uma taxa de explicação de 92,22%, as variáveis significativas que contribuem para o resultado da procura são:

- **Navegantes**: através do valor do coeficiente da variável Navegantes, pode-se chegar à conclusão de que a partir de sua aparição em abril de 2019 até o final de 2022 contribuiu para um aumento do número de passageiros em 502.105 Através do Gráfico 4-10, pode-se afirmar que parte do aumento que se deu a partir de abril de 2019, pelo menos quase meio milhão de passageiros pode ser atribuído ao título navegantes.

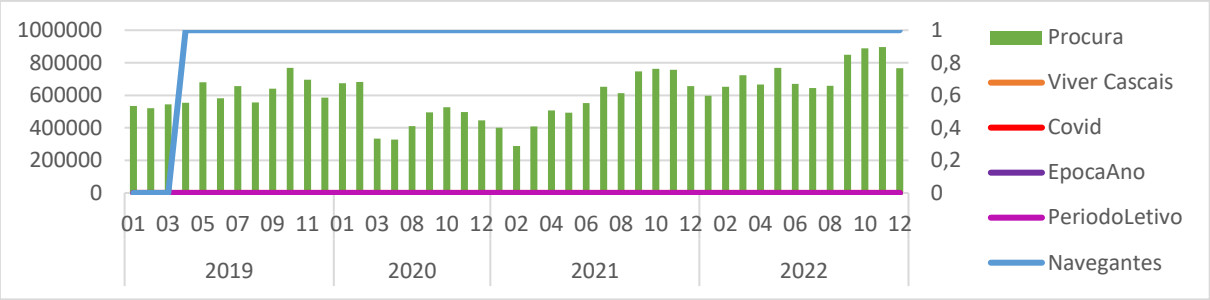


Gráfico 4-10 Período de incidência título navegantes

- **Covid**: o poder explicativo dessa variável pode ser facilmente verificado no Gráfico 4-11. No período de incidência da Covid, a procura demonstrou uma queda bastante significativa se comparada a pequenos declínios em outros períodos. Sendo assim, a incidência dessa variável consegue explicar muito bem o que houve com a procura nesse intervalo de tempo, ela foi responsável pelo decréscimo de 244.726 passageiros.

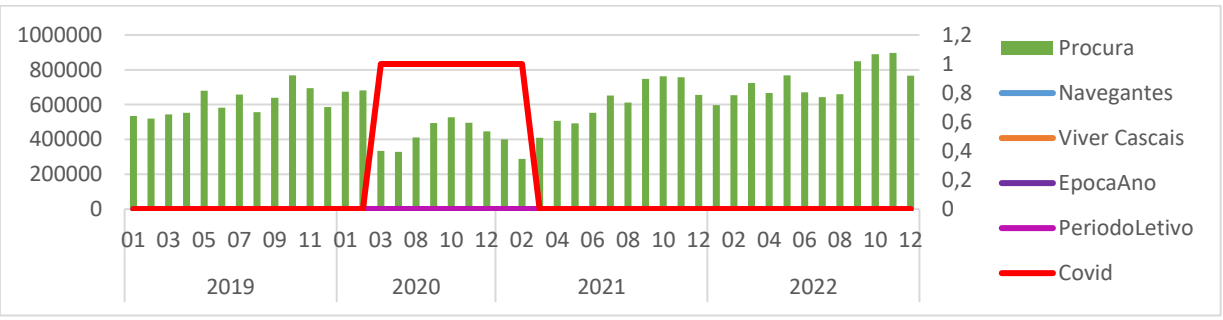


Gráfico 4-11 Período de incidência covid

- **Período Letivo**: O que a regressão mostra é que para os períodos de incidência dessa variável, Gráfico 4-12, há um aumento de aproximadamente 201.480 passageiros.

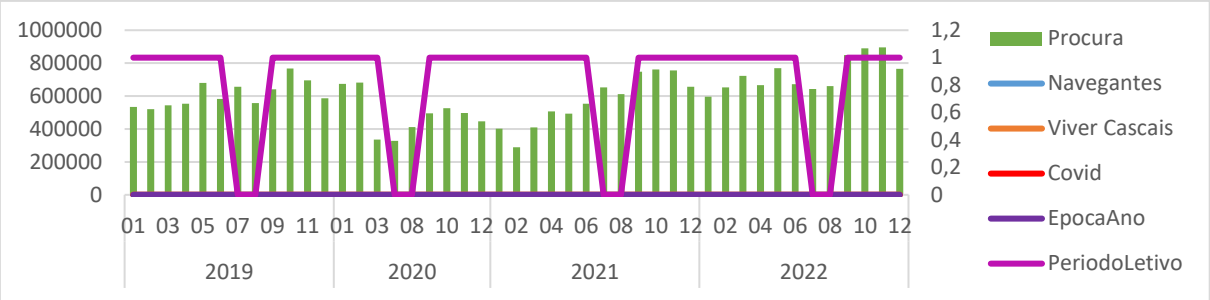


Gráfico 4-12 Período incidência Período Letivo

De modo geral, o modelo **apresentou boa robustez** ao explicar as variáveis em análise. Os resultados obtidos foram coesos e a análise gráfica muito coerente, o que permite concluir que para além de um alto poder explicativo, o modelo está devidamente calibrado e opera eficientemente.

4.3.6.2 Agregação por região

Seguindo o mesmo procedimento da análise anterior, agora de forma um pouco mais direta, iniciam-se os testes para a análise da regressão B, baseada no segregamento de linhas conforme as zonas geográficas: zona de Cascais e zona de Carcavelos. A primeira etapa para as linhas de Cascais, resultou no Quadro 4-8.

Quadro 4-8 Regressão linear amostragem por região (Cascais)

SUMÁRIO DOS RESULTADOS						
Estatística de regressão						
R múltiplo	0,971602262					
Quadrado de R	0,944010956					
Quadrado de R ajustado	0,913412051					
Erro-padrão	92338,84963					
Observações	45					
ANOVA						
	gl	*SQ*	*MQ*	*F*	*F de significância*	
Regressão	5	5,75046E+12	1,15009E+12	134,8850961	1,594E-23	
Residual	40	3,41059E+11	8526463152			
Total	45	6,09152E+12				
	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	0	#N/D	#N/D	#N/D	#N/D	#N/D
Navegantes	296258,5722	50120,31326	5,910948136	6,30587E-07	194961,6406	397555,5039
ViverCascais	-103016,3749	36420,7757	-2,828505788	0,007274011	-176625,5084	-29407,24141
Covid19	-42740,6237	36641,17575	-1,166464307	0,250334439	-116795,2023	31313,95489
EpocaAno	40694,33758	32039,931	1,270113147	0,211383137	-24060,77846	105449,4536
PeriodoLetivo	163758,2606	32968,817	4,967065109	1,31496E-05	97125,79593	230390,7253

E a segunda etapa resultou no Quadro 4-9.

Quadro 4-9 Regressão linear amostragem por região apenas com variáveis significativas

SUMÁRIO DOS RESULTADOS						
Estatística de regressão						
R múltiplo	0,969030754					
Quadrado de R	0,939020601					
Quadrado de R ajustado	0,912307297					
Erro-padrão	94043,73113					
Observações	45					
ANOVA						
	gl	*SQ*	*MQ*	*F*	*F de significância*	
Regressão	3	5,72007E+12	1,90669E+12	215,5857339	3,9636E-25	
Residual	42	3,71457E+11	8844223365			
Total	45	6,09152E+12				
	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	0	#N/D	#N/D	#N/D	#N/D	#N/D
Navegantes	335284,8346	39344,31421	8,521811634	1,05915E-10	255884,794	414684,8752
ViverCascais	-122687,0302	35386,77335	-3,467030719	0,001227463	-194100,4301	-51273,63042
PeriodoLetivo	148462,4984	30568,74809	4,856675777	1,69311E-05	86772,2672	210152,7296

Nessa regressão alguns aspectos chamaram a atenção de forma diferente. O primeiro deles foi a ligeira diminuição do valor de R na regressão com as varáveis explicativas e o segundo refere-se ao coeficiente da variável Viver Cascais estar apresentando um valor negativo.

Quanto a diminuição do coeficiente de determinação, existem alguns motivos que podem contribuir para essa queda. O primeiro deles está relacionado a possível perda de informações, quando se retira uma variável não explicativa pode-se perder algumas informações úteis que parecem insignificantes à primeira vista, entretanto podem conter informações relevantes para a variável resposta. Além disso, o tamanho da amostra também pode diminuir o valor de R, com uma amostra menor, o modelo pode ter um pouco menos de precisão.

Quanto à variável Viver Cascais, era esperado que fosse contribuir para um aumento na procura, já que a gratuitidade do título de transporte, empiricamente leva o cientista de dados a acreditar que mais passageiros fossem utilizar o TP. Entretanto, as análises mostram o contrário, uma queda na procura que deve abrir portas a pesquisa adicional para uma melhor perceção do evento.

Quando se analisa o Gráfico 4-13, nota-se que para a linhas selecionadas realmente há uma queda na procura após a inclusão do passe Viver Cascais.

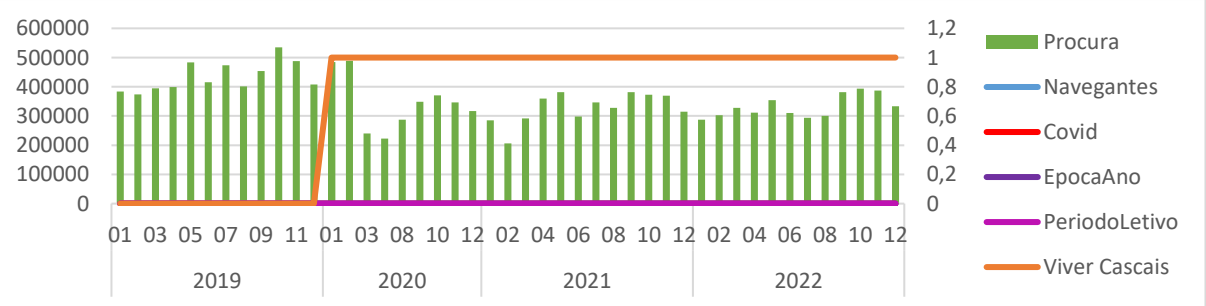


Gráfico 4-13 Incidência Viver Cascais na amostragem B

Entretanto, na análise anterior com todas as linhas, não existe essa queda, pelo contrário, nota-se um aumento na procura no período pós introdução do título. Tendo sito tomado esse conhecimento, optou-se então pela criação de um gráfico de demonstrasse a diferença entre todas as linhas, e as linhas da amostragem por região. No Gráfico 4-14 pode-se observar que para todas as situações a tendencia é de alta, e quando se segrega as linhas a tendência é de queda.

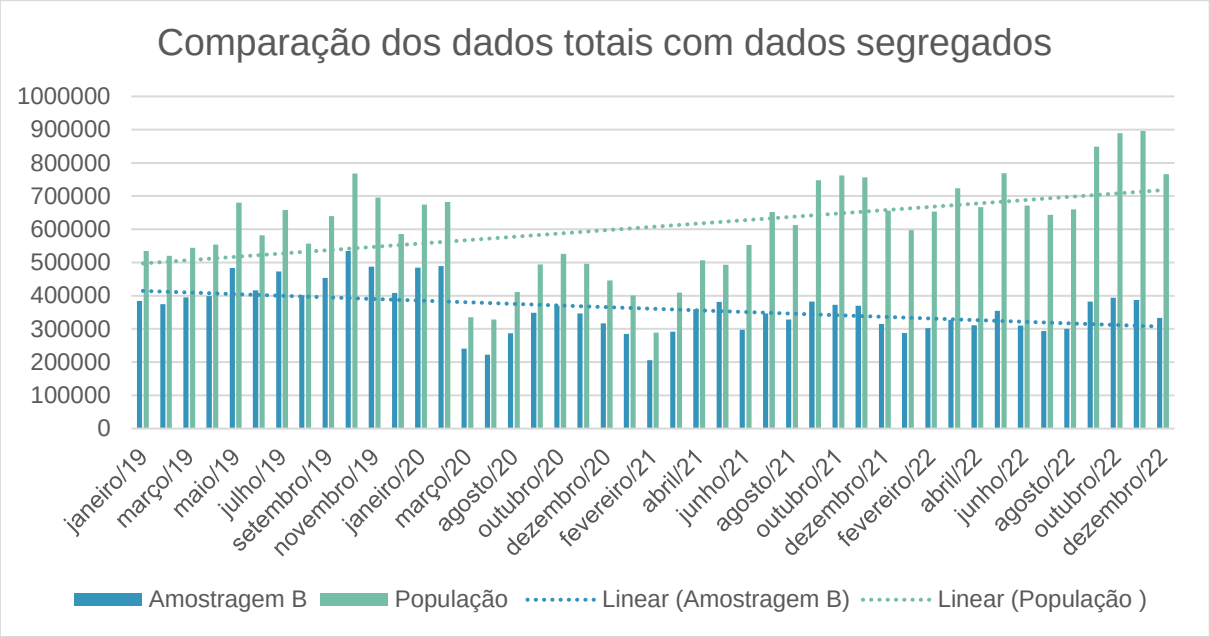


Gráfico 4-14 Dados totais e dados segregados

Essa constatação leva a crer que muito provavelmente **existe uma variável explicativa que não foi considerada no modelo, ou então a segregação das linhas deve ser feita de outra maneira** para que se justifique a queda nessa região. Sendo assim partiu-se para a investigação desses fatos.

A primeira ideia vem da mudança administrativa que ocorreu em 2021 de acordo com a Figura 4-1. Com a mudança na oferta do TP, regiões que eram servidas por apenas uma linha, passaram a ser servidas por várias, o que leva a pensar que houve uma redistribuição de passageiros. O indivíduo que antes tinha apenas uma opção para se deslocar, agora tem outras. Ou seja, a linha antiga pode até apresentar uma queda, mas a verdade é que representa uma redistribuição dos passageiros.

Para **analisar essa teoria**, partiu-se para o estudo de um grupo de linhas específico, as linhas M03, M10 e M17. A linha M03 é uma linha antiga com 41 paragens que começa e termina no Cascais Shopping e está representada na Figura 4-4.

Figura 4-4 Rota linha M03

Após maio de 2021, surgiram duas novas linhas M10 e M17, em que parte do trajeto é correspondente com o a M03. Como mostra a Figura 4-5 a M10 cobre grande parte do trajeto na M03, percorrendo um total de nove paragens, e a linha M17 cobrindo um percurso menor da M03, contabilizando três paragens. A partir de então, calculou-se a regressão desse grupo de linhas.

Figura 4-5 Rota M03 com acréscimo da M17 e M10

Quadro 4-10 Regressão linhas M03, M10 e M17.

SUMÁRIO DOS RESULTADOS						
Estatística de regressão						
R múltiplo	0,943146564					
Quadrado de R	0,889525441					
Quadrado de R ajustado	0,853477985					
Erro-padrão	7077,084768					
Observações	45					
ANOVA						
	gl	*SQ*	*MQ*	*F*	*F de significância*	
Regressão	5	16131133432	3226226686	64,41486251	8,58656E-18	
Residual	40	2003405153	50085128,82			
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	10838,47498	3841,348543	2,821528652	0,007406833	3074,81997	18602,12998
ViverCascais	11950,14566	2791,381071	4,281087158	0,000112978	6308,554069	17591,73724
Covid19	-17600,56445	2808,273093	-6,267397745	1,98886E-07	-23276,29608	-11924,83281
EpocaAno	-135,5897073	2455,621968	-0,055216034	0,956241233	-5098,586834	4827,40742
PeriodoLetivo	2039,609065	2526,814159	0,80718602	0,424333238	-3067,272847	7146,490977

Com esta nova regressão apresentada no Quadro 4-10**Erro! A origem da referência não foi encontrada.**, com o segregamento feito com base na região de abrangência de uma linha antiga, com as linhas novas complementares, chega-se à conclusão de que as variáveis que explicam a procura neste caso são Navegantes, Viver Cascais e Covid19, que juntos conseguem explicar 85,35% da regressão. Esse valor ainda continua apresentando um valor próximo de um, ou seja, o modelo se ajusta aos dados de forma significativa.

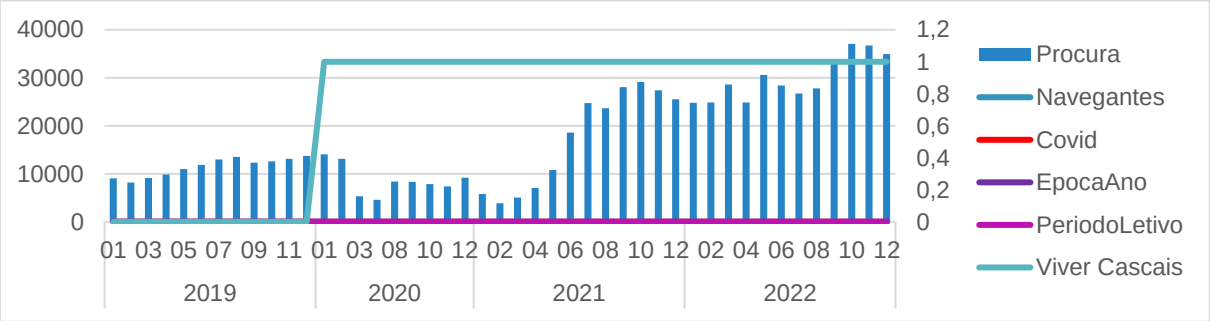


Gráfico 4-15 Período de incidência do Viver Cascais

Nesta regressão, nota-se que a variável Viver Cascais, **apresenta significância** com um coeficiente indicando um aumento na procura na ordem de 11.950 passageiros, representando coerência ao que se era minimamente esperado, e estatisticamente robusto e confiável.

Com essa tipologia de segregação das linhas nota-se que os resultados são diferentes, pois já que houve um aumento na oferta, de forma física, as análises das características devem levar esse fator em consideração. Não se deve analisar somente as linhas antigas, mas sim uma linha antiga e as novas que fazem parte dela, de maneira que tenham uma proporção justa, não discrepante e senso crítico no momento da análise.

Essa nova forma de agrupamento de linhas, mostrou que de certa forma, o aumento da oferta, gerou um aumento no número de passageiros, dessa maneira, optou-se por quantificar esse aumento relacionado a adição de novas linhas. Para isso, foi duplicado o modelo de regressão, onde se substituiu a variável “época do ano”, que não apresentou resultados significativos em nenhuma circunstância, por uma nova variável denominada “novas linhas”.

Essa variável, também é do tipo *dummy*, e sua codificação é 1 “atributo presente” a partir de abril de 2021. A regressão pode ser observada no Quadro 4-11, e a análise gráfica na Figura 4-6 s seguir.

Quadro 4-11 Regressão com variável "novas linhas"

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

A partir da análise de regressão, nota-se que com um poder explicativo de 91,95% as variáveis Navegantes, Novas Linhas e Período letivo explicam a variável dependente. Pela Figura 4-6 percebe-se que o grande aumento na procura se deu com a inclusão de novas linhas, e através da disposição das variáveis, o modelo consegue explicar isso muito bem.

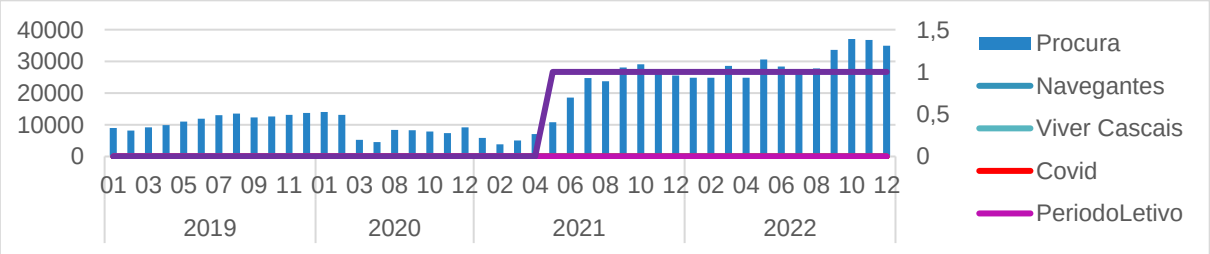


Figura 4-6 Regressão com "Novas Linhas"

Feitas essas análises, pode-se concluir que efetivamente houve uma diminuição da procura na linha M03, entretanto foi devido ao redirecionamento de passageiros. Antes esses passageiros transportavam-se apenas pela M03, já que era a única opção, com as novas linhas

existem mais duas opções de transportes, o que implica que parte da clientela da M03 foi distribuída para a M10 e M17.

Essa conclusão, somada à regressão com a variável "Novas Linhas" informa o analista dos dados que a introdução de novas linhas, e introdução do título Viver Cascais foram medidas que aumentaram de fato a procura, podendo essa situação ser quantificada estatisticamente.

4.3.6.3 Agregação por linhas que servem a universidade

Levando em consideração o modelo de agrupamento de linhas da análise do tópico anterior, optou-se por realizar as próximas análises com um tipo de segmentação semelhante. Dessa maneira decidiu-se analisar a linha M16, que é uma linha antiga com paragem inicial/final na universidade NOVA SBE e as novas linhas com troços correspondentes M21, M23 e M24.

Figura 4-7 Linhas que servem a NOVA SBE.

Cujos resultados são:

Quadro 4-12 Regressão linhas que servem a NOVA SBE

SUMÁRIO DOS RESULTADOS						
Estatística de regressão						
R múltiplo	0,966729144					
Quadrado de R	0,934565238					
Quadrado de R ajustado	0,907639773					
Erro-padrão	14291,23968					
Observações	45					
ANOVA						
	gl	*SQ*	*MQ*	*F*	*F de significância*	
Regressão	3	1,22515E+11	40838420759	199,9535567	1,68227E-24	
Residual	42	8578060326	204239531,6			
Total	45	1,31093E+11				
	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	0	#N/D	#N/D	#N/D	#N/D	#N/D
Navegantes	40722,75774	4542,538848	8,964757177	2,63814E-11	31555,5432	49889,97227
Covid19	-22460,75347	5377,507393	-4,176796391	0,00014602	-33313,00274	-11608,50419
PeriodoLetivo	20226,85165	4645,342123	4,354222169	8,39523E-05	10852,17171	29601,5316

O Quadro 4-12 da ANOVA, já realizado apenas com as variáveis significativas, mostra que para esse caso, o título Navegantes, o Covid e o Período Letivo são relevantes para o modelo com um grau de explicação a 90,76%. A variável do Período Letivo mostra que houve um aumento da procura em aproximadamente 20.227 passageiros. Portanto, parte do aumento da procura que ocorre nesse período, é estatisticamente justificado pela busca de estudantes em época de aulas.

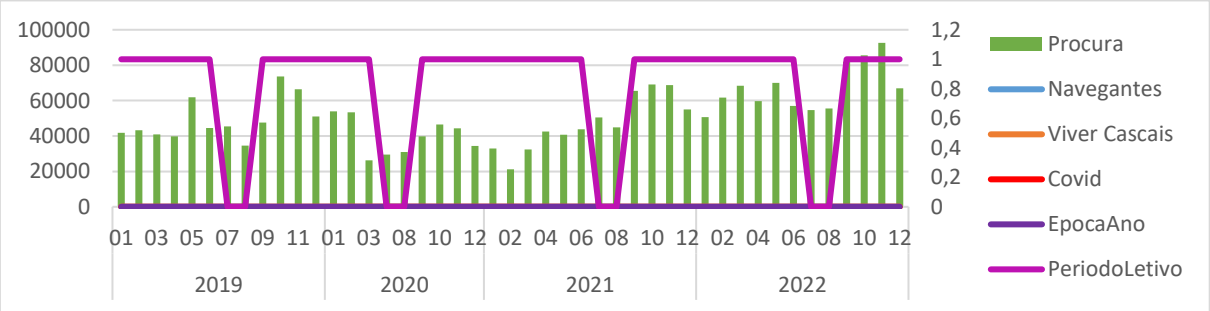


Gráfico 4-16 Gráfico incidência variável período letivo para linhas da NOVA SBE

Quando se analisa esse grupo de linhas no modelo adicional de regressão, com a variável "novas linhas", percebe-se que a variável período letivo continua sendo relevante do ponto de

vista estatístico, confirmando a hipótese de relevância na procura, e dando importância também à reestruturação das linhas como sendo um fator de aumento da procura, dando uma confiabilidade de 91,01% como evidencia o Quadro 4-13.

Quadro 4-13 Regressão com variável "novas linhas"

	m	σ_m	t Stat	P-value	R^2	R^2 Ajust	F	F Signif	G_{Lib}
Navegantes	25891,93954	4781,43975	5,415093	2,74122E-06	0,936923376	0,910110203	207,9522705	7,92744E-25	42
Novas Linhas	19229,05134	4335,88288	4,434864	6,51317E-05					
PeriodoLetivo	21290,5739	4558,92392	4,670088	3,08402E-05					

4.3.6.4 Agregação por tipologia

Na agregação por tipologia de linhas, onde se buscou saber se a incidência de alguma variável poderia ter um efeito diferente em cada tipo de linha, notou-se que entre as linhas circulares/locais e alimentadoras não existem diferenças significativas pelo menos não por algum motivo que esteja expresso nas cinco variáveis analisadas, portanto acredita-se que, com base nestas variáveis, a procura de uma linha não varia em relação a sua tipologia.

As variáveis que se demonstraram significativas são o Navegantes, Covid-19 e Período Letivo representando acréscimos na procura para o título de transporte e período letivo e decréscimo para o Covid-19, como pode ser observado no Quadro 4-14 e Quadro 4-15 com um ajuste da reta de regressão de poder explicativo na volta de 91% e 92%, o que indica robustez para o modelo.

Quadro 4-14 Regressão linhas circulares/locais

	m	σ_m	t Stat	P-value	R^2	R^2 Ajust	F	F Signif	G_{Lib}
Navegantes	56809,19424	6058,79002	9,376327	7,40088E-12	0,939324525	0,912625693	216,7357289	3,57772E-25	42
Covid19	-32798,46638	7172,46219	-4,57283	4,20619E-05					
PeriodoLetivo	28044,20704	6195,90794	4,526247	4,87737E-05					

Quadro 4-15 Regressão linhas alimentadoras

	m	σ_m	t Stat	P-value	R^2	R^2 Ajust	F	F Signif	G_{Lib}
Navegantes	416126,2078	38863,2503	10,70745	1,40182E-13	0,949460499	0,923244332	263,0110448	8,43891E-27	42
Covid19	-183087,323	46006,7426	-3,97958	0,000267746					
PeriodoLetivo	173860,0053	39742,7737	4,374632	7,87385E-05					

Para além dos resultados expressos na tabela ANOVA, os gráficos também mostram que não há diferenças muito significativas, além do volume de passageiros, entre os

agrupamentos de linhas circulares/locais e alimentadores como pode ser observado na página 82.

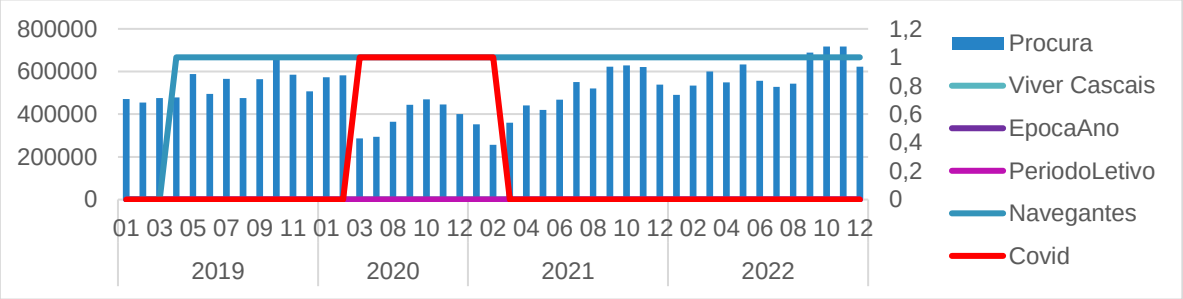


Gráfico 4-17 Incidência covid e navegantes nas linhas circulares

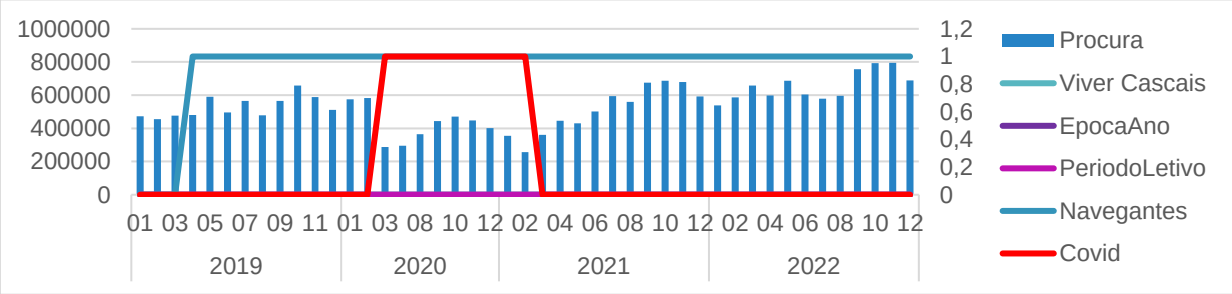


Gráfico 4-18 Incidência covid e navegantes linhas alimentadoras

Mais uma vez, nota-se que a estrutura da procura em ambos os casos segue aparentemente a mesma forma, não diferindo muito, a ponto de ter relevância para o modelo.

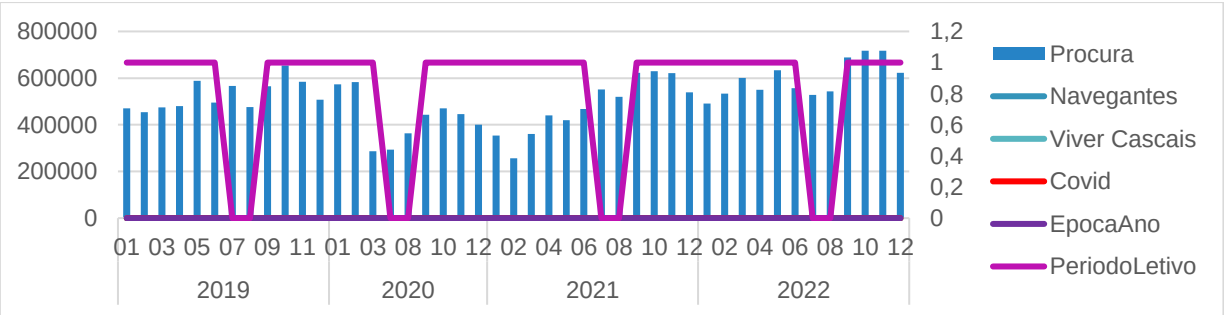


Gráfico 4-19 Incidência Período letivo linhas circulares

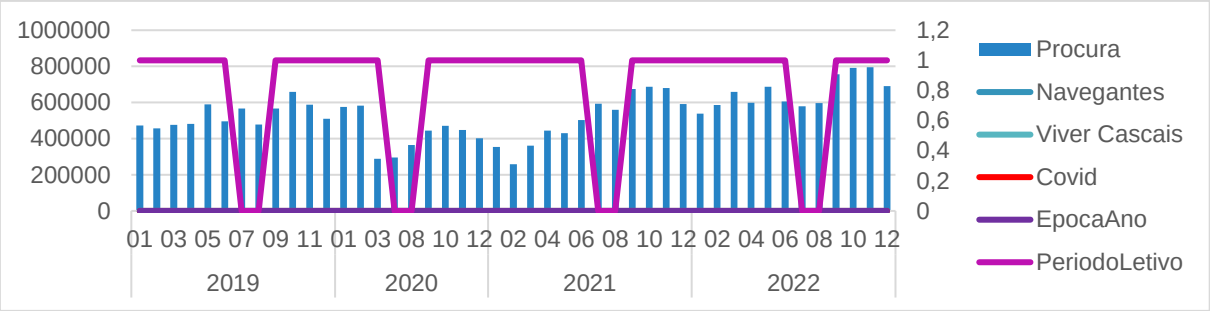


Gráfico 4-20 Incidência Período Letivo linhas alimentadoras

Esta última análise mostra que a forma como os passageiros procuram as linhas depende muito mais das suas necessidades de deslocamento, que da tipologia da linha.

5 Conclusões

5.1 Principais conclusões

Neste capítulo, é apresentada uma análise abrangente dos resultados da pesquisa, resumindo os principais achados e conclusões que emergiram ao longo do estudo. Para facilitar a compreensão e visualização dos resultados, estes foram organizados em um quadro resumido apresentado no Quadro 5-1 e no Quadro 5-2, onde a legenda encontra-se abaixo:

Legenda	
	A partir deste ponto a análise foi feita com diferenciação da nova estrutura de rede
	Variáveis com resultados significativos
Nota:	Análises que contém I, II referem-se respetivamente a análises com agrupamento específico, além disso as suas variações II.I e I.I referem-se respetivamente a análises com as variáveis "iniciais" e com adição da variável "novas linhas".

Quadro 5-1 - Quadro síntese resultados 1

Análises				
	Análise com todas as linhas		Análise dos grupos de linhas com influência das novas linhas	
	Análise população	**Análise por região I**	**Análise por região II**	**Análise por região II.I**
Covid	Apresentação de um bom poder explicativo em 91,95% com todas as variáveis e 92,22% só com variáveis explicativas.	Apresentação de um bom poder explicativo em 91,34% com todas as variáveis e 91,23% só com variáveis explicativas.	Apresentação de um poder explicativo entre bom e razoável com 85,35% com todas as variáveis e 91,23% só com variáveis explicativas.	Apresentação de um poder explicativo entre bom e razoável com 85,35% com todas as variáveis e 91,23% só com variáveis explicativas.
Navegantes	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura
Viver Cascais		x - Diminuição da procura na procura	x - Aumento na procura	
Época do ano				
Período letivo	x - Aumento na procura	x - Aumento na procura		x - Aumento na procura
Covid	x - Diminuição na procura		x - Diminuição na procura	
"Novas Linhas"				x - Aumento na procura

Quadro 5-2 Quadro síntese resultados 2

Análises				
Análise dos grupos de linhas com influência das novas linhas				
	Servidão a universidades I	**Servidão a universidades I.I**	**Tipologia de linhas**	
			Circulares	**Alimentadoras**
Covid	Apresentação de um poder explicativo bom com 90,76% com todas as variáveis e 91,01% só com variáveis explicativas.	Apresentação de um poder explicativo bom com 90,76% com todas as variáveis e 91,01% só com variáveis explicativas.	Apresentação de um poder explicativo bom com 91,26% só com variáveis explicativas.	Apresentação de um poder explicativo bom com 93,32% só com variáveis explicativas.
Navegantes	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura
Viver Cascais				
Época do ano				
Período letivo	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura	x - Aumento na procura
Covid	x - Diminuição na procura		x - Diminuição na procura	x - Diminuição na procura
"Novas Linhas"		x - Aumento na procura		

Para além das informações trazidas pelos quadros, existem outros resultados que valem a pena ser mostrados. Em termos globais estima-que a Covid trouxe uma perda estimada de cerca de 250 mil passageiros por mês no período entre março de 2020 e fevereiro de 2022, e que os períodos escolares trazem um aumento de 200 mil passageiros, provavelmente estudantes

Em termos de volume de viagens e de passageiros a pesquisa em 2022, a frota de 44 viaturas realizou cerca de 560 mil viagens, com uma oferta de cerca de 410 milhões de lugares-km e 106,5 milhões de passageiros-km transportados, correspondendo a uma taxa média de ocupação de 26% (dados da Cascais Próxima)

A pesquisa realizada neste trabalho final de mestrado, teve como objetivo a análise explicativa do comportamento da procura do transporte público em Cascais através da construção de um modelo de regressão linear cujos **resultados apresentaram-se muito coesos, sólidos e satisfatórios, contribuindo com valiosos *insights* para o campo da análise de mobilidade**.

A utilização da ferramenta contribuiu para o fornecimento de informações relevantes através do cálculo dos coeficientes de cada variável, parâmetros de significância do modelo e coeficientes de ajustamento. Tendo sido essa ferramenta testada e aprovada no quesito robustez, o modelo oferece a vantagem de **se adaptar a dados em constante evolução, permitindo atualizações contínuas, o que vir a constituir uma ferramenta muito prática para análises futuras.**

A relevância desse estudo vai muito além dos resultados obtidos. Ela **confirma estatisticamente as suspeitas preexistentes, tornando o modelo uma prova matemática bem fundamentada sobre o comportamento da procura com as variáveis estudadas,** fazendo com que a análise tenha grande potencial para **enriquecer a gestão da mobilidade no Concelho de Cascais, nomeadamente do TP**.

Apesar dos avanços efetuados neste trabalho, verificaram-se algumas **limitações** como a disponibilidade limitada de dados, compreendendo apenas um período de quatro anos, e a reestruturação da rede, que implicou uma nova metodologia de análise. Outro fator, que pode ser considerado limitante é a ausência de dados que contabilizem a saída dos passageiros em cada paragem, que poderiam fazer uma análise mais ampla sobre os troços das linhas de TP mais procurados.

Por fim, a pesquisa proporcionou *insights* sólidos e práticos para o campo da mobilidade urbana, por meio de um modelo de regressão robusto o analista está bem-posicionado para

oferecer contribuições significativas às discussões sobre a gestão da mobilidade e consequentemente realizar intervenções mais eficazes. Reconhecendo as limitações atuais, prevê-se um futuro, onde a disponibilidade de mais dados e análises aprofundadas podem construir o caminho para uma mobilidade urbana mais inteligente e eficiente.

5.2 Perspectivas futuras

No que diz respeito às perspectivas futuras deste estudo, além das considerações conceituais já abordadas, é essencial considerar aspectos práticos que podem enriquecer mais a análise. Primeiramente, uma discussão mais profunda dos resultados com a Cascais Próxima para entender quais objetivos da análise foram alcançados e identificar quais as próximas etapas, o que envolverá uma hierarquização das prioridades e definição de metas específicas para futuras investigações.

Além disso, a expansão do conjunto de variáveis explicativas é uma direção promissora para superar as limitações das cinco variáveis utilizadas nesse estudo. Isso permitirá uma análise mais abrangente e aprofundada dos fatores que influenciam a variação da procura dos transportes públicos, nomeadamente os autocarros de Cascais geridos pela empresa em questão.

Outro fator que pode ser considerado é o aumento da disponibilidade de dados, a medida que a disponibilidade e qualidade dos dados aumenta, abrem-se novas oportunidades para análise mais sofisticadas e significativas. Com a aplicação de princípios de ciência de dados, como técnicas avançadas de tratamentos de dados, pode-se refinar ainda mais os modelos e maximizar a sua capacidade de ajustamento, permitindo resultados cada vez mais robustos.

A evolução constante da infraestrutura de transportes e mobilidade urbana exige que o modelo acompanhe essas mudanças. A reestruturação do modelo para acomodar a nova configuração da rede permitirá análises mais ágeis e simplificadas ao eliminar a necessidade de dividir análise entre linhas novas e linhas antigas, pode-se ter uma visão mais inteira dos padrões de mobilidade, facilitando a tomada de decisões. Ainda sobre a reestruturação do modelo pode-se adicionar a possibilidade de incluir mais variáveis.

Referência Bibliográficas

Afonso, A., & Nunes, C. (2019). *Probabilidades e Estatística Aplicações e Soluções em SPSS*. Editora Escolar

Almeida, J., Sousa, N., Bizerra, R., & Morais, R. (2015). *Indicadores de desempenho em transporte urbano de passageiros: estudo de aderência de indicadores em pesquisa de satisfação de usuários em um terminal de ônibus de São Paulo*. *Novembro*, 16.

André, M. R. (2022). *Arranca a grande modernização da Linha de Cascais. O que vai mudar?* https://lisboaparapessoas.pt/2022/12/07/linha-de-cascais-modernizacao/. Acessado em 21 de abril de 2023.

Bruce, P., & Bruce, A. (2019). *Estatistica Prática para Cientistas de Dados*. Editora Alta Books

Camara Municipal de Cascais. (n.d.). *Cascais Próxima*. https://www.cascais.pt/empresa-municipal/cascais-proxima. Acessado em 02 de julho de 2023

Câmara Municipal de Cascais. (n.d.-a). *MobiCascais*. https://www.cascais.pt/. Acessado em 02 de julho de 2023

Câmara Municipal de Cascais. (n.d.-b). *Primeiro autocarro a hidrogênio do país já circula em Cascais*. https://www.cascais.pt/noticia/primeiro-autocarro-hidrogenio-do-pais-ja-circula-em-cascais. Acessado em 02 de julho de 2023

Câmara Municipal de Cascais. (2016). *Revisão da Carta Educativa do Concelho de Cascais e Elaboração do Plano Estratégico Educativo Municipal: Fase II - Relatório Intercalar*. https://www.cascais.pt/sites/default/files/anexos/gerais/new/revisao_da_carta_educativa_do_concelho_de_cascais_fase_ii.pdf. Acessado em 08 de julho de 2023

Camara Municipal de Cascais. (2020). *O MobiCascais*. https://mobi.cascais.pt/geral/quem-somos. Acessado em 02 de julho de 2023

Câmara Municipal de Cascais. (2021). *Novo serviço de transporte público - Tabela de correspondência de linhas*. Acessado em 02 de julho de 2023

Câmara Municipal de Cascais. (2023). *Já foram criados mais de 100.000 cartões Viver Cascais*. https://www.cascais.pt/noticia/ja-foram-criados-mais-de-100000-cartoes-viver-cascais. Acessado em 02 de julho de 2023

Campos, S. M. (2019). *Métodos Estocásticos da Engenharia II Capítulo 4-Inferência: intervalos de confiança*. *2*, 1–76. Acessado em 29 de agosto de 2023

Cascais - A Riviera Portuguesa. (2017). https://mundodosviajantes.com/cascais-a-riviera-portuguesa/. Acessado em 02 de julho de 2023

CCDRN (2008). Comissão de Coordenação e Desenvolvimento Regional do Norte. *Manual do Planeamento de Acessibilidades e Transportes: Transportes Públicos*.

Cuesta, Y. (2021). *Quais são os suplementos de BI no Excel*. https://biist.pro/complementos-bi-en-excel-power-bi-desktop. Acessado em 13 de agosto de 2023.

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

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

Ferraz, C. P., & Torres, I. G. E. (2004). *Transporte público urbano*. Editora Rima

Han, J., & kamber, M. (2002). *Data mining: concepts and techniques*. *2*. Editora Morgan

Han, J., Pei, J., & Tong, H. (2023). *Data Mining - Concepts and Techniques*. Editora Morgan

Henrique, M., & Molin, D. (2016). *Introdução ao Estudo de Probabilidade e Estatística com auxílio do software R*. *August*. Universidade Tecnológica Federal do Paraná.

Kapel, S. G. B. (2020). *Análise de dados agregados aplicada à mobilidade urbana - Um framework analítico para a geração de conhecimento*. Biblioteca Digital de Teses e Dissertações da USP

Machado, M. (2012). *Modelos de previsão aplicados à optimização da gestão das actividades de um Call Center*. Relatório de Estágio Universidade de Lisboa

Magalhães, M. N., & Lima, A. C. de. (2018). *Noções de Probabilidade e Estatística* (7th ed.). Editora Usp

Makridakis S, Wheelwright SC, H. R. (1997). *Forecasting methods and applications*. Editora Wiley

Marques, P. (2017). *O Sistema Integrado de Mobilidade Mobicascais*. 86. https://estudogeral.sib.uc.pt/bitstream/10316/82845/1/. Dissertação

Martins, M. E. G. (2019). *Regressão linear simples*. *7*(3), 2–4. https://doi.org/10.24927/rce2019.045. Revista de ciência elementar

Microsoft. (n.d.). *O que é o PowerBI*. https://powerbi.microsoft.com/pt-br/what-is-power-bi/. Acessado em 04 de abril de 2023.

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

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis*. Editora Wiley

Morettin, P. A., & Bussab, W. O. (2017). *Estatística Básica*. Editora Saraiva

Oiseth, S., Jones, L., & Maza, E. (2022). *Testes Estatísticos e Representação de Dados*. https://www.lecturio.com/pt/concepts/testes-estatisticos-e-representacao-de-dados/#lecturio-toc__Testes Estatísticos. Acessado em 09 de julho de 2023.

ONU News. (2020). *Organização Mundial da Saúde declara novo coronavírus uma pandemia*. https://news.un.org/pt/story/2020/03/170688 Acessado em 16 de junho de 2023

Piana, C., Machado, A., & Selau, L. (2009). Estatística Básica. *Estatística Básica*, 155–171. minerva.ufpel.edu.br/~markus.stein/Apostila_EB.pdf. Dissertação UFPEL

Portal Viva. (n.d.). *Passes navegante*. https://www.portalviva.pt/pt/homepage/títulos-de-transporte/uso-frequente/passes-navegante.aspx Acessado em 10 de junho de 2023

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. Editora O'Reilly

Triola, M. F. (2017). *Livro Introducao à estatística* (11th ed.). Editora LTC

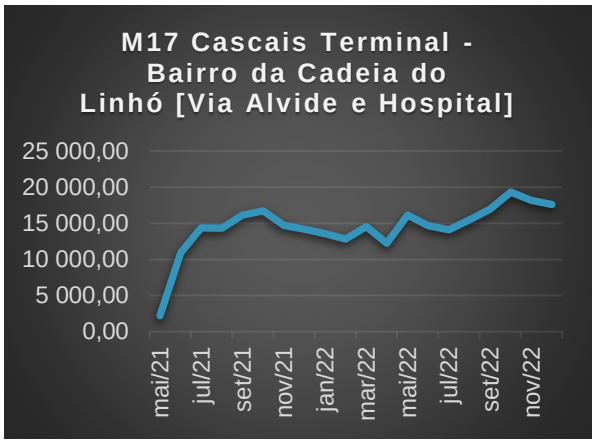
Turkey, J. (1977). *Explanatory Data Analysis*. Editora Pearson

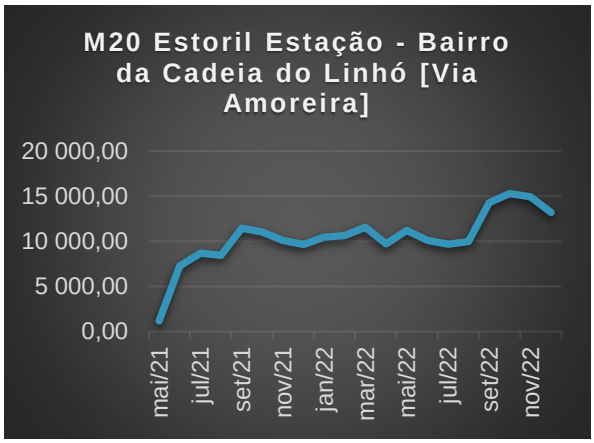
ANEXOS

Anexo A – Variabilidade das linhas

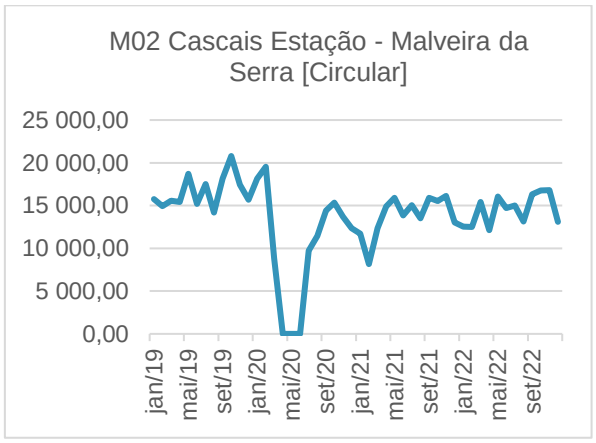
Os gráficos com fundo escuro representam as linhas novas, e os gráficos com o fundo claro representam as linhas antigas.

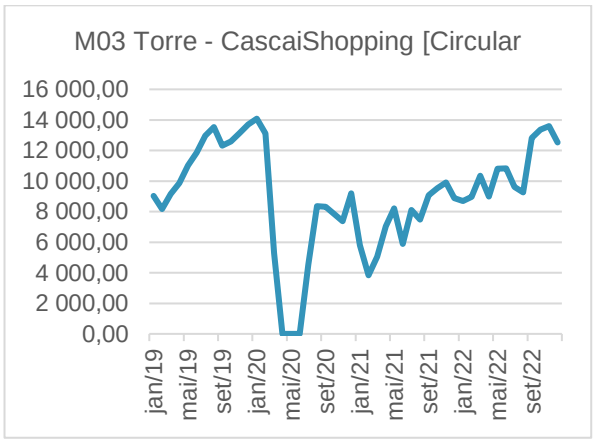
- Linhas com baixa variabilidade

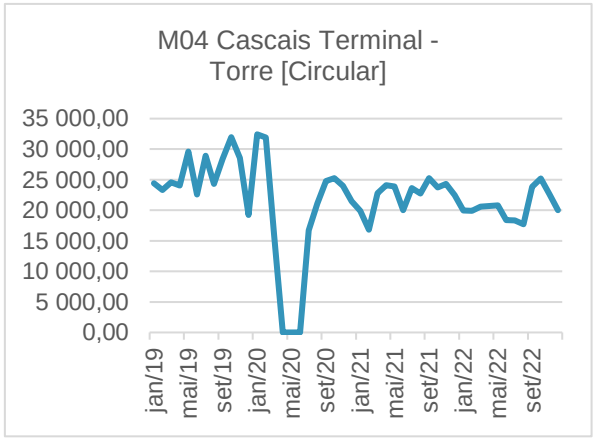


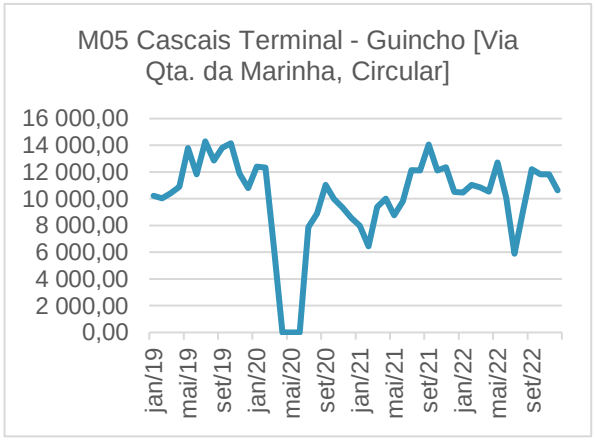


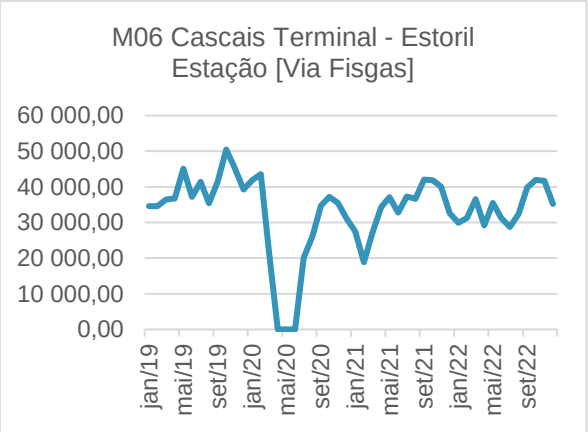
- Linhas com variabilidade média

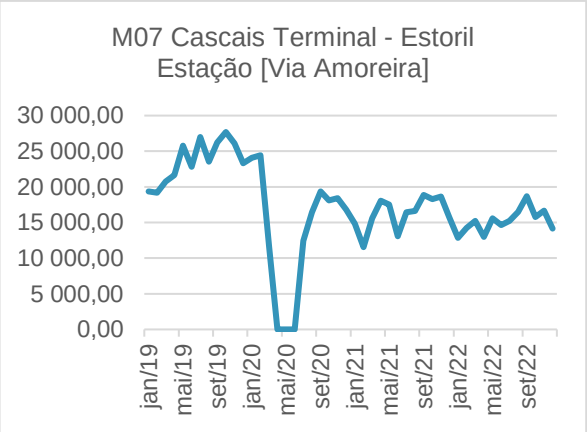


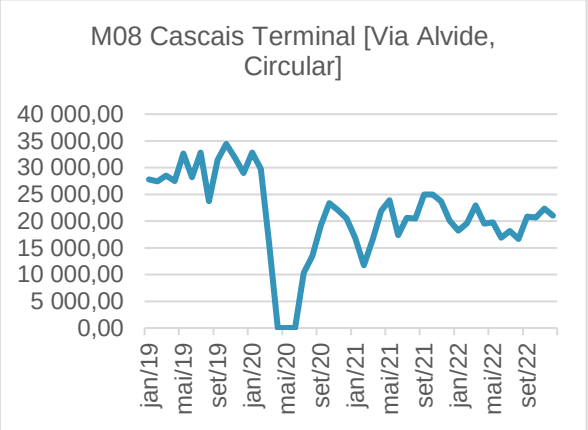


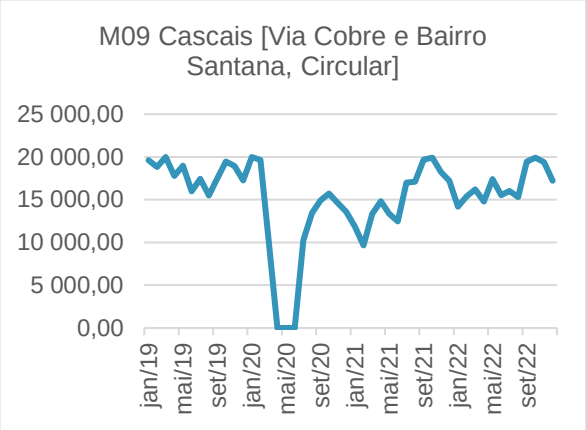


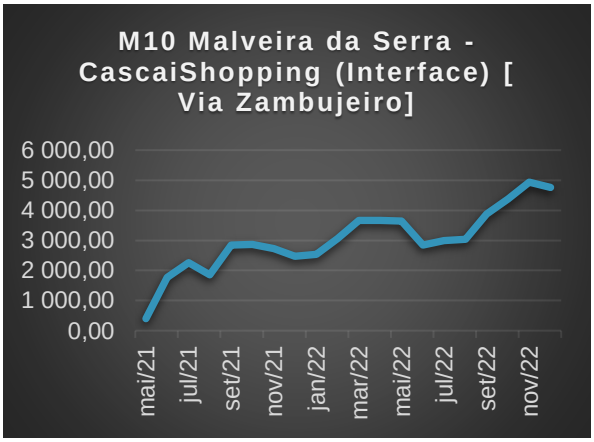


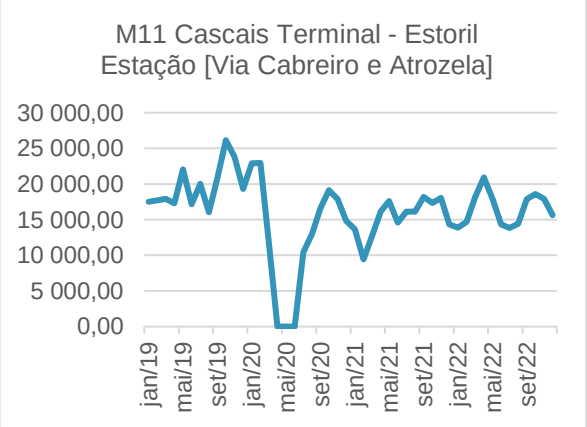

M06 Cascais Terminal - Estoril Estação [Via Fisgas]
60 000,00
50 000,00
40 000,00
30 000,00
20 000,00
10 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

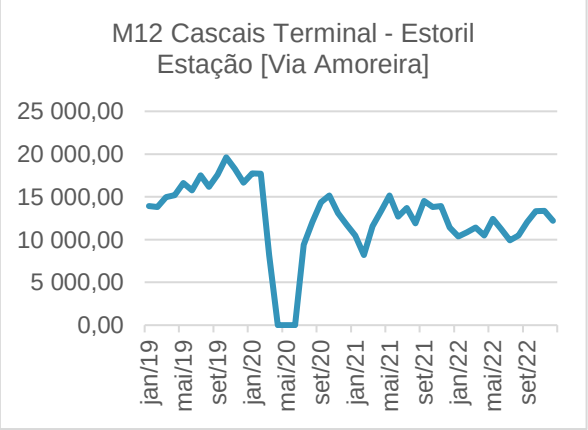

M07 Cascais Terminal - Estoril Estação [Via Amoreira]
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

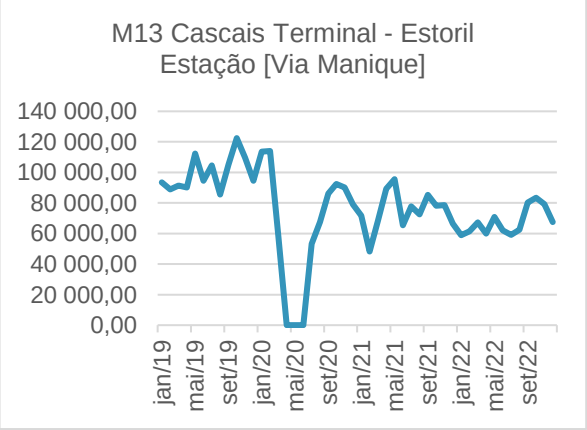

M08 Cascais Terminal [Via Alvide, Circular]
40 000,00
35 000,00
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

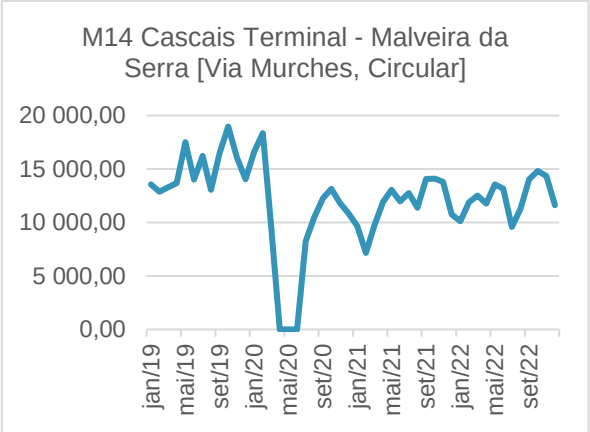

M09 Cascais [Via Cobre e Bairro Santana, Circular]
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

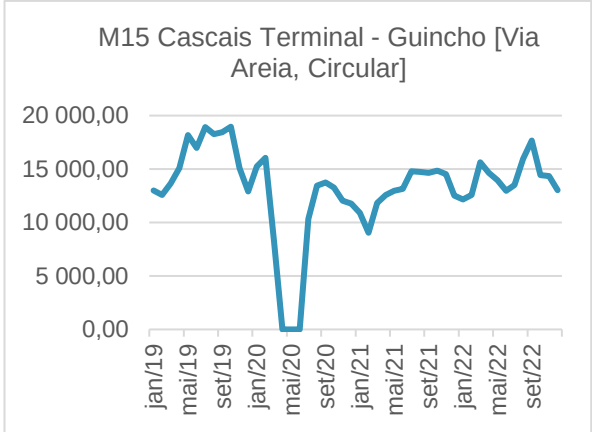

M10 Malveira da Serra - CascaiShopping (Interface) [Via Zambujeiro]
6 000,00
5 000,00
4 000,00
3 000,00
2 000,00
1 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

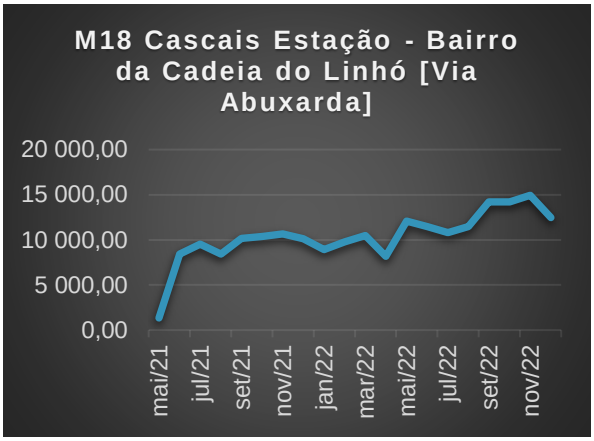

M11 Cascais Terminal - Estoril Estação [Via Cabreiro e Atrozela]
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

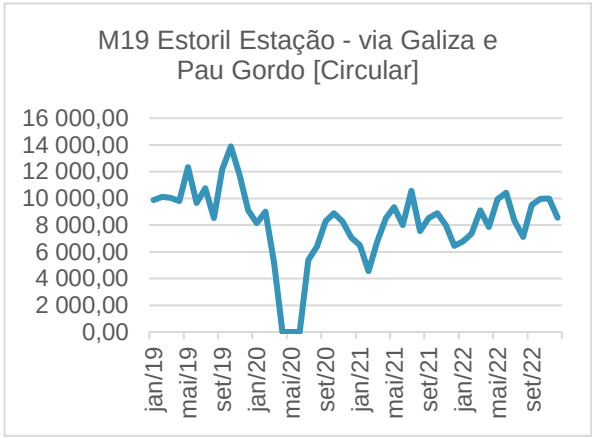

M12 Cascais Terminal - Estoril Estação [Via Amoreira]
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

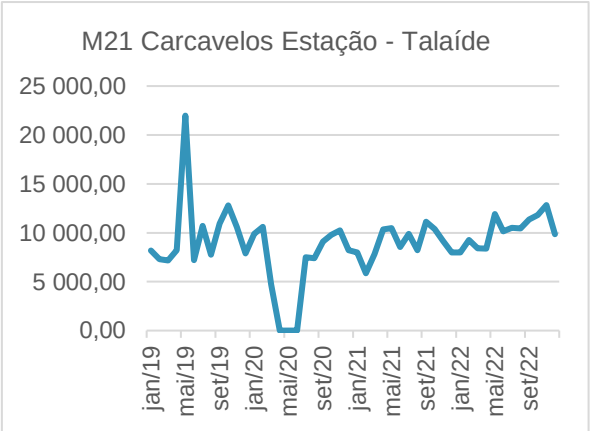

M13 Cascais Terminal - Estoril Estação [Via Manique]
140 000,00
120 000,00
100 000,00
80 000,00
60 000,00
40 000,00
20 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

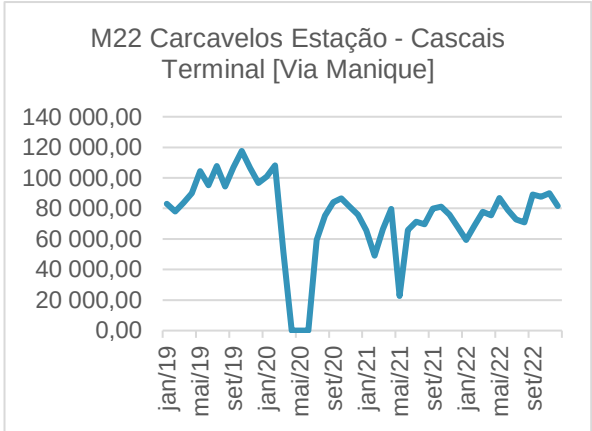

M14 Cascais Terminal - Malveira da Serra [Via Murches, Circular]
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

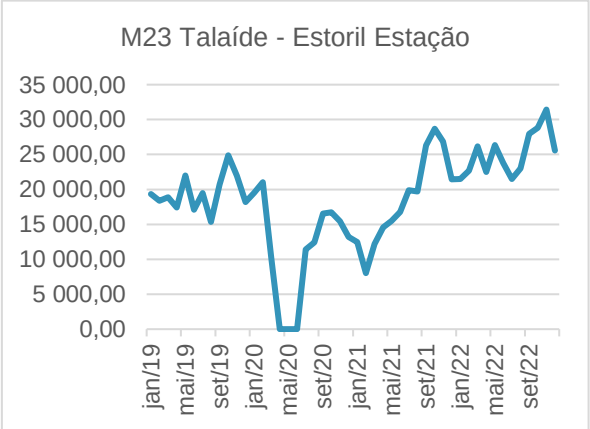

M15 Cascais Terminal - Guincho [Via Areia, Circular]
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

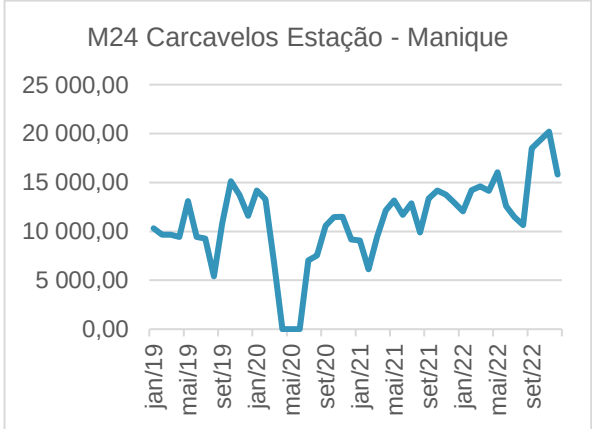

M18 Cascais Estação - Bairro da Cadeia do Linhó [Via Abuxarda]
20 000,00
15 000,00
10 000,00
5 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

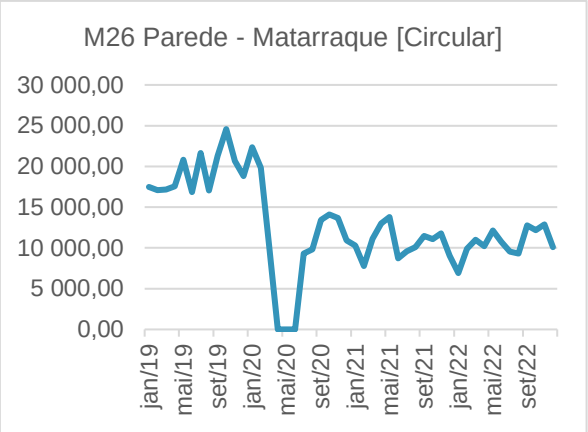

M19 Estoril Estação - via Galiza e Pau Gordo [Circular]
16 000,00
14 000,00
12 000,00
10 000,00
8 000,00
6 000,00
4 000,00
2 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

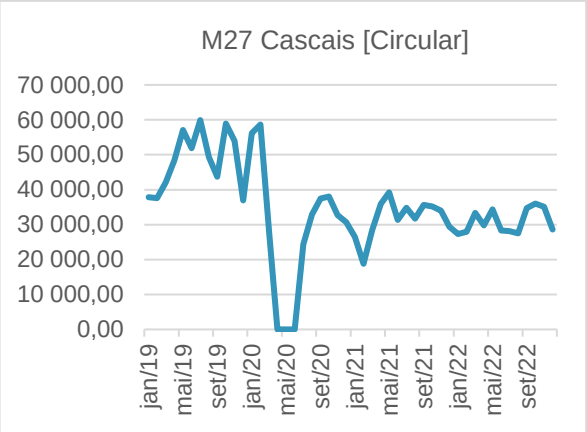

M21 Carcavelos Estação - Talaíde
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

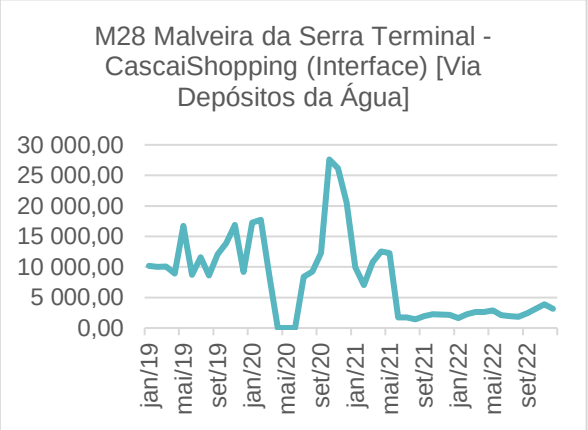

M22 Carcavelos Estação - Cascais Terminal [Via Manique]
140 000,00
120 000,00
100 000,00
80 000,00
60 000,00
40 000,00
20 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22


M23 Talaíde - Estoril Estação
35 000,00
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

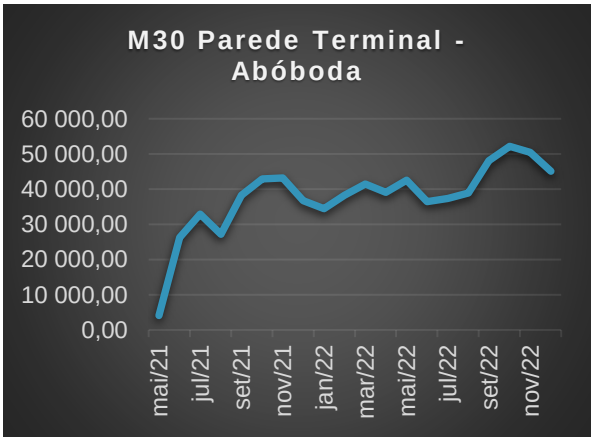

M24 Carcavelos Estação - Manique
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

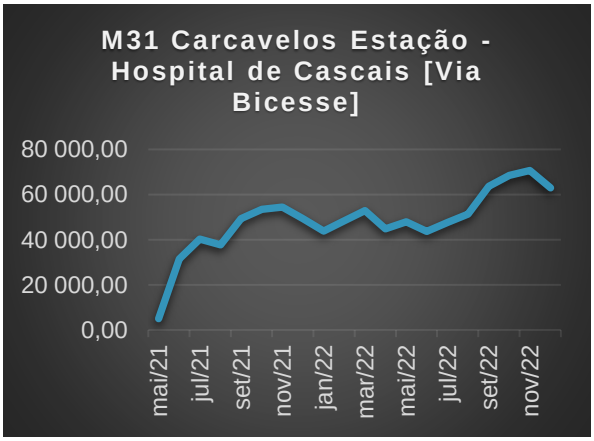

M26 Parede - Matarraque [Circular]
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

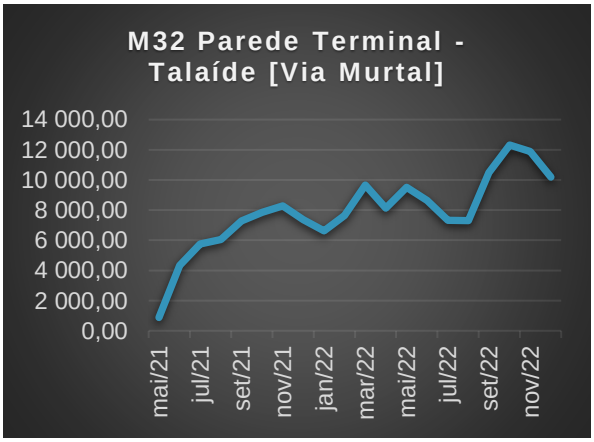

M27 Cascais [Circular]
70 000,00
60 000,00
50 000,00
40 000,00
30 000,00
20 000,00
10 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

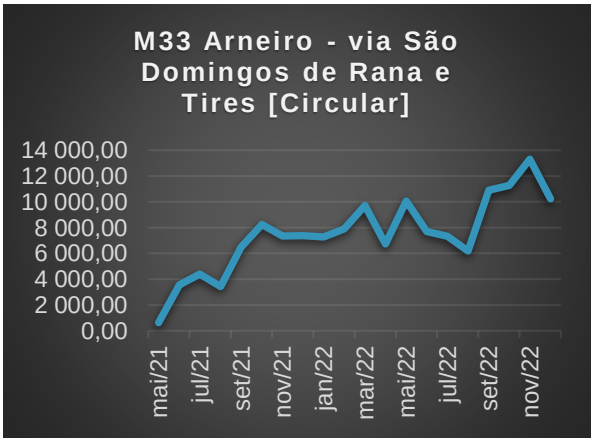

M28 Malveira da Serra Terminal - CascaiShopping (Interface) [Via Depósitos da Água]
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
jan/19
mai/19
set/19
jan/20
mai/20
set/20
jan/21
mai/21
set/21
jan/22
mai/22
set/22

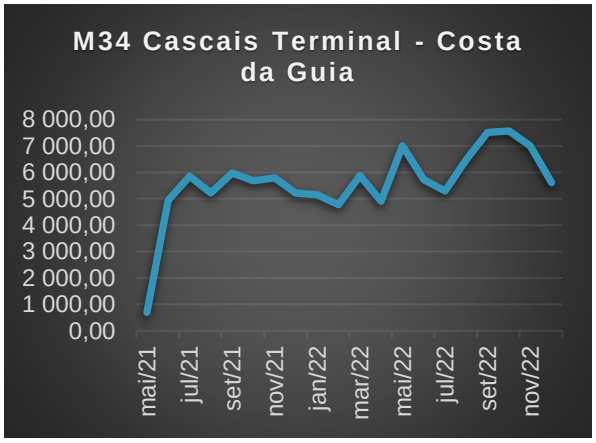

M29 São João do Estoril Estação - CascaiShopping
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

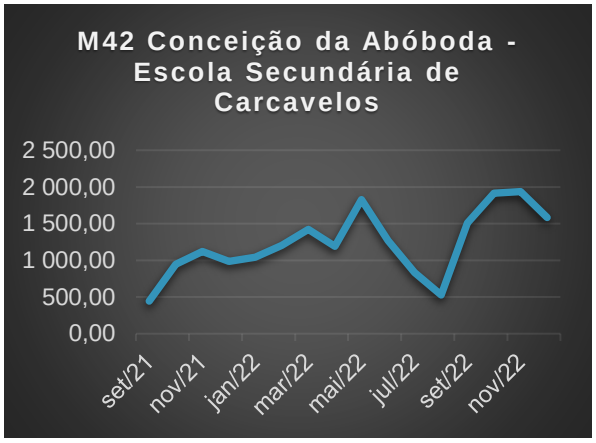

M30 Parede Terminal - Abóboda
60 000,00
50 000,00
40 000,00
30 000,00
20 000,00
10 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

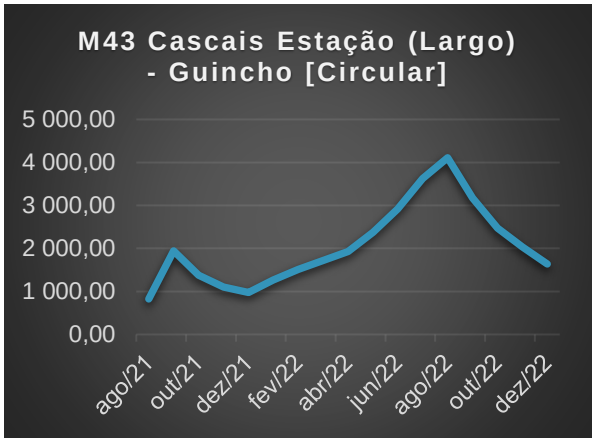

M31 Carcavelos Estação - Hospital de Cascais [Via Bicesse]
80 000,00
60 000,00
40 000,00
20 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

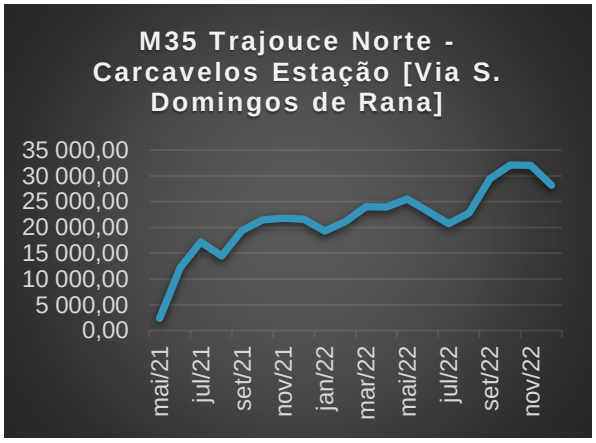

M32 Parede Terminal - Talaíde [Via Murtal]
14 000,00
12 000,00
10 000,00
8 000,00
6 000,00
4 000,00
2 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

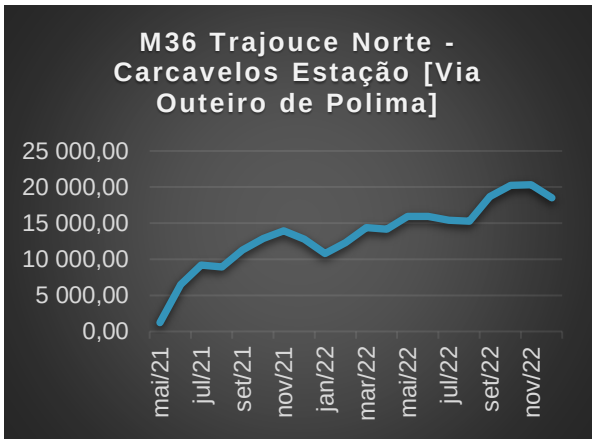

M33 Arneiro - via São Domingos de Rana e Tires [Circular]
14 000,00
12 000,00
10 000,00
8 000,00
6 000,00
4 000,00
2 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22


M34 Cascais Terminal - Costa da Guia
8 000,00
7 000,00
6 000,00
5 000,00
4 000,00
3 000,00
2 000,00
1 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

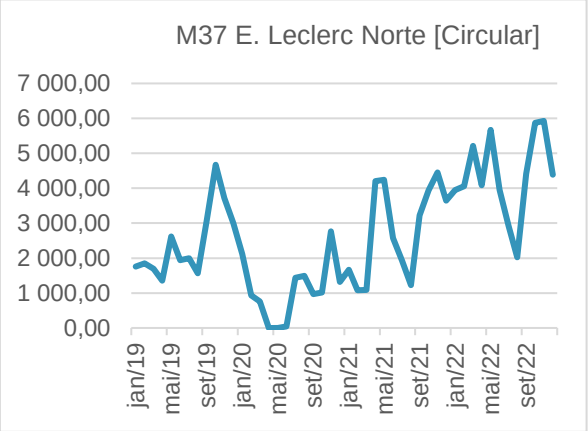

M42 Conceição da Abóboda - Escola Secundária de Carcavelos
2 500,00
2 000,00
1 500,00
1 000,00
500,00
0,00
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

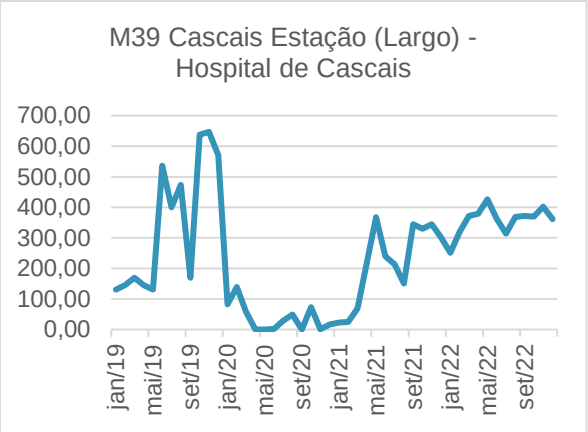

M43 Cascais Estação (Largo) - Guincho [Circular]
5 000,00
4 000,00
3 000,00
2 000,00
1 000,00
0,00
ago/21
out/21
dez/21
fev/22
abr/22
jun/22
ago/22
out/22
dez/22

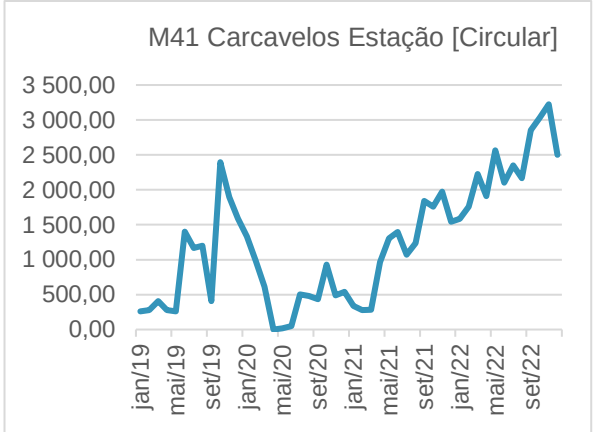

M35 Trajouce Norte - Carcavelos Estação [Via S. Domingos de Rana]
35 000,00
30 000,00
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

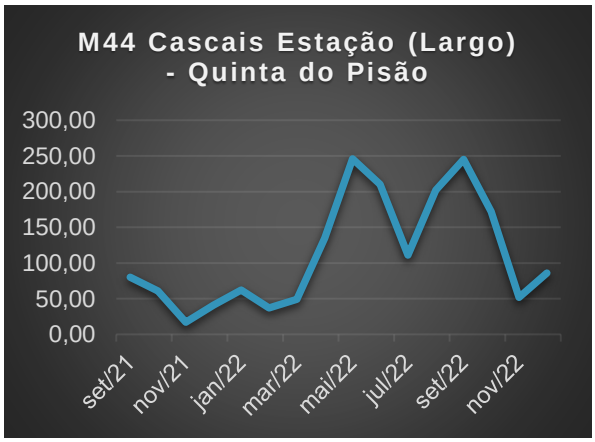

M36 Trajouce Norte - Carcavelos Estação [Via Outeiro de Polima]
25 000,00
20 000,00
15 000,00
10 000,00
5 000,00
0,00
mai/21
jul/21
set/21
nov/21
jan/22
mar/22
mai/22
jul/22
set/22
nov/22

- Linhas com variabilidade elevada

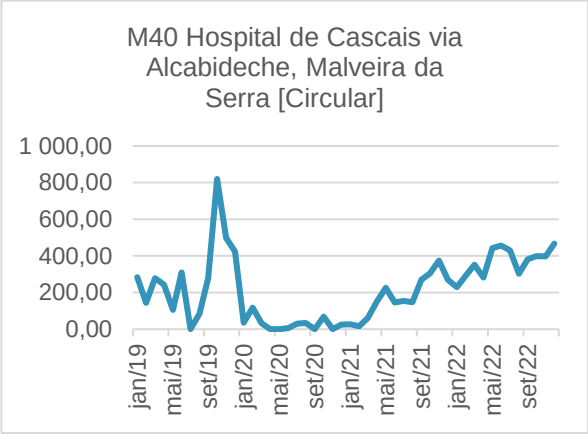




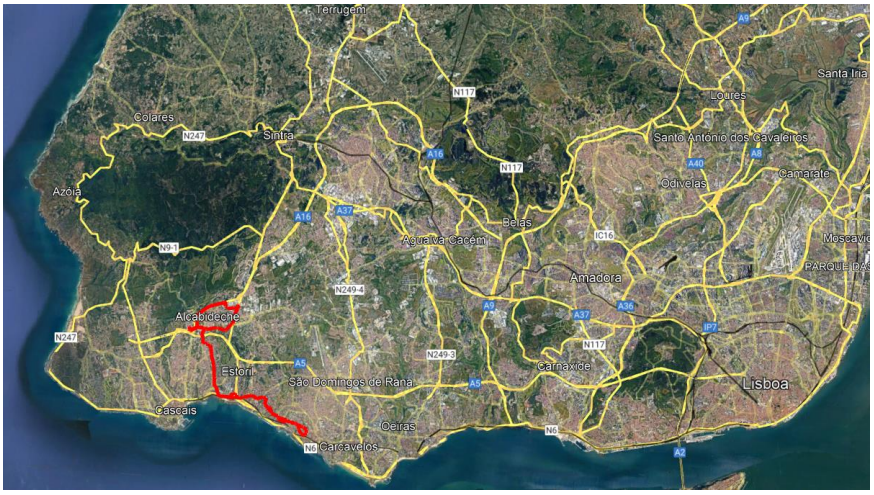
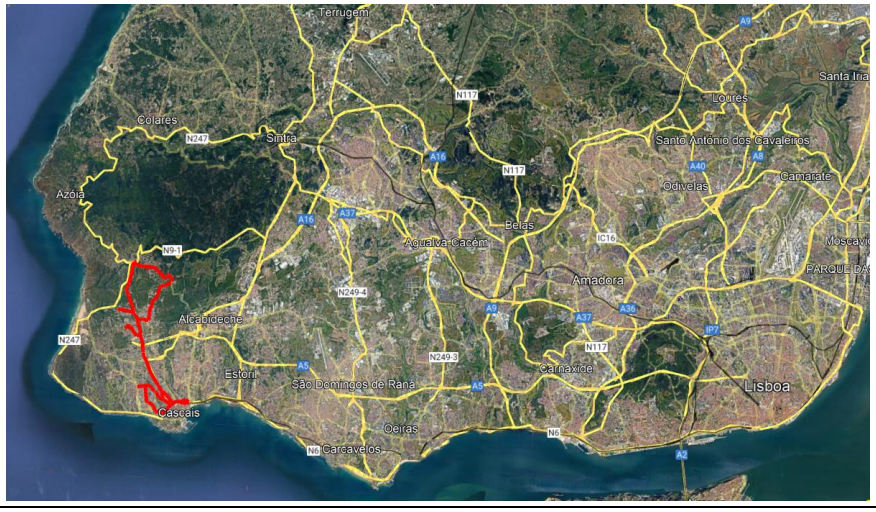


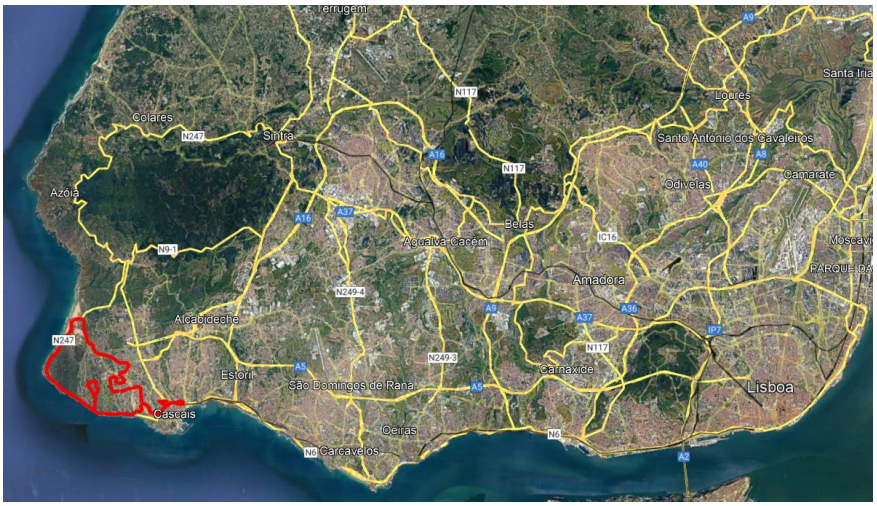


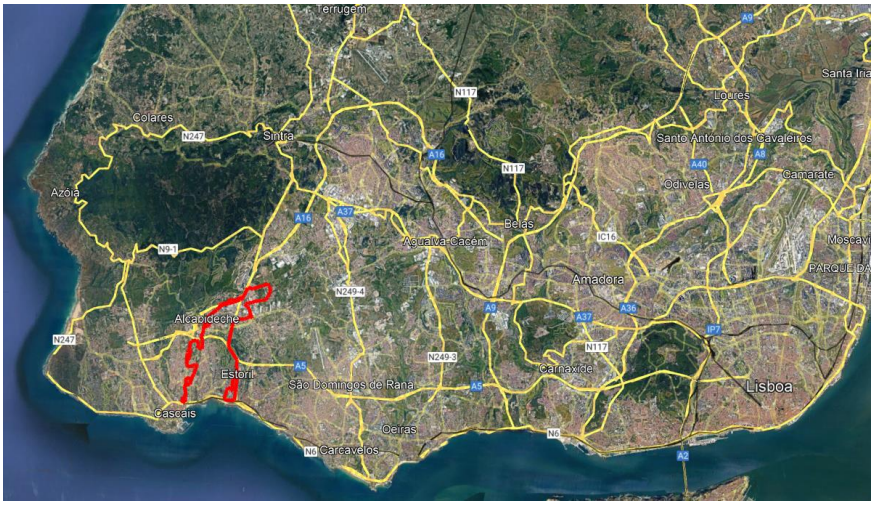
- Linhas com variabilidade extrema

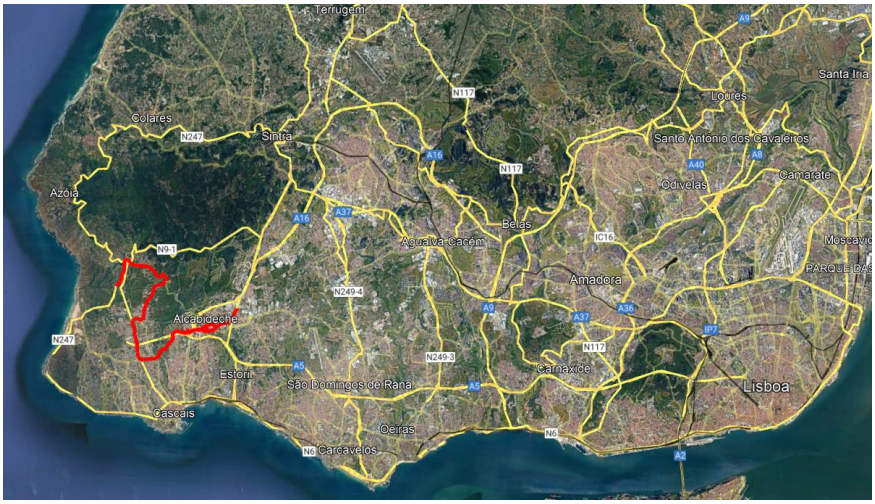
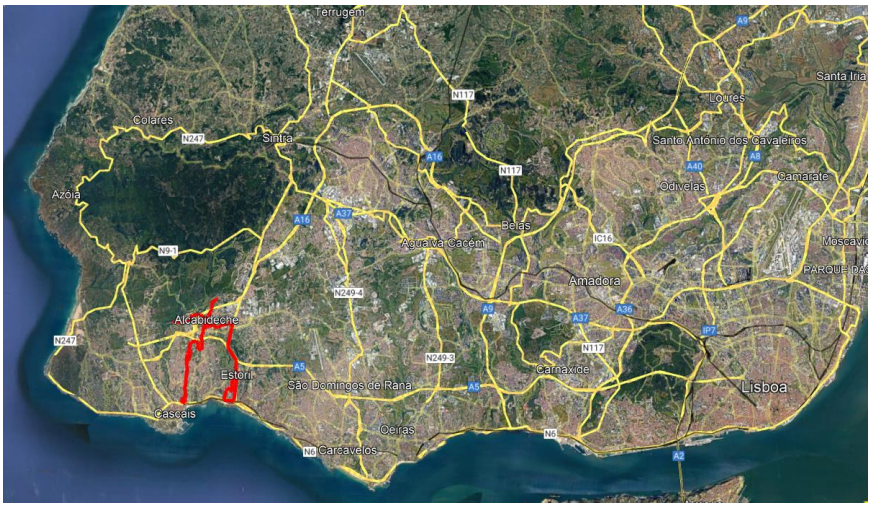


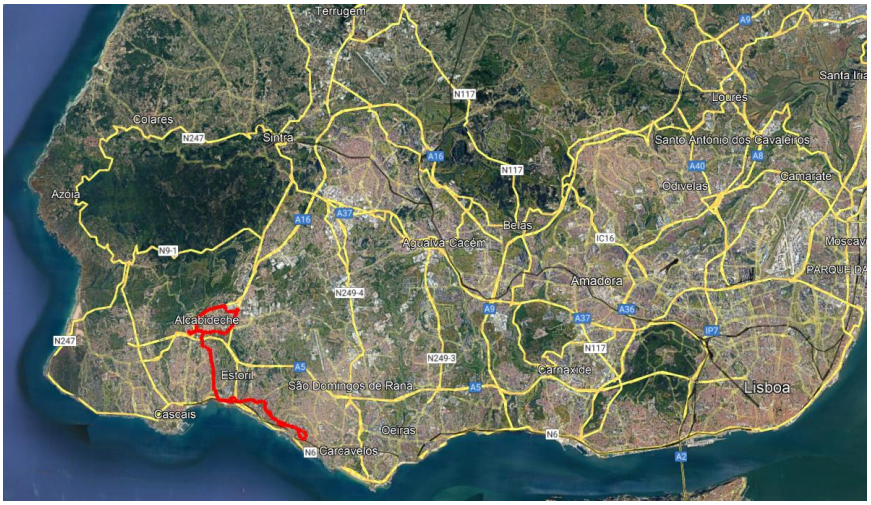
Anexo B – Traçado rota dos autocarros

Carreira	Traçado
M01	
M02	

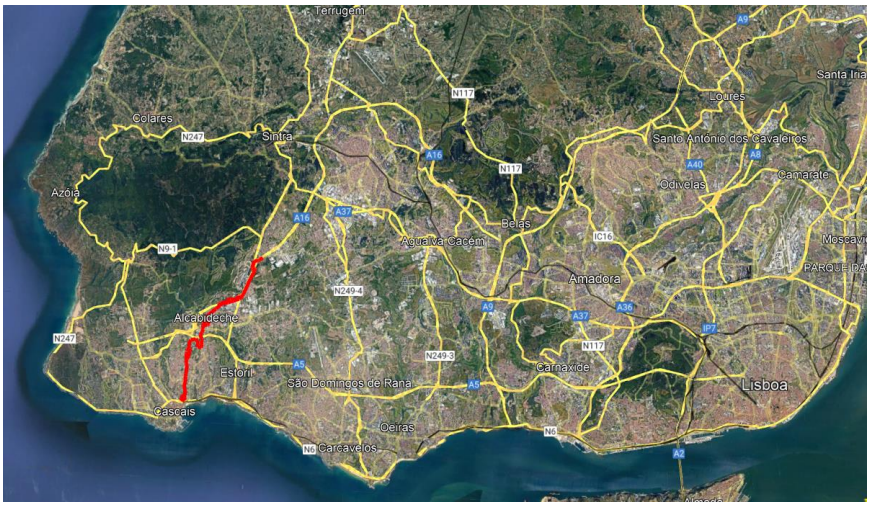
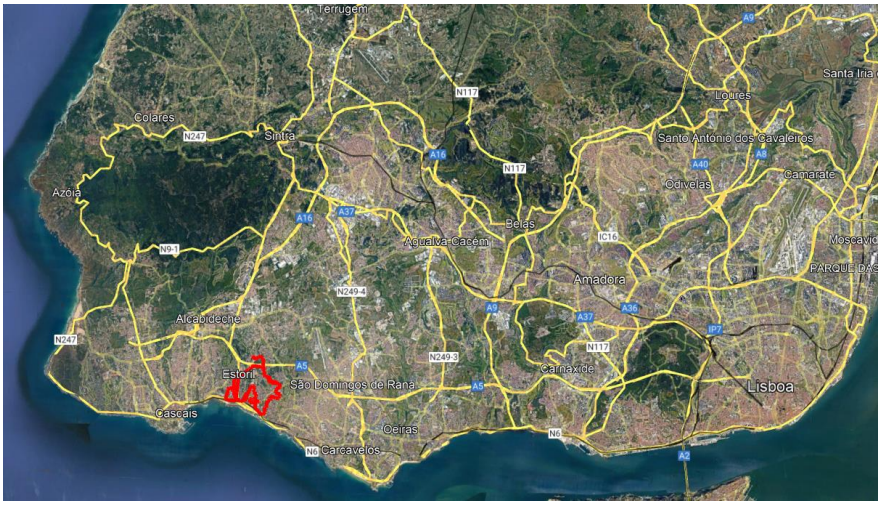
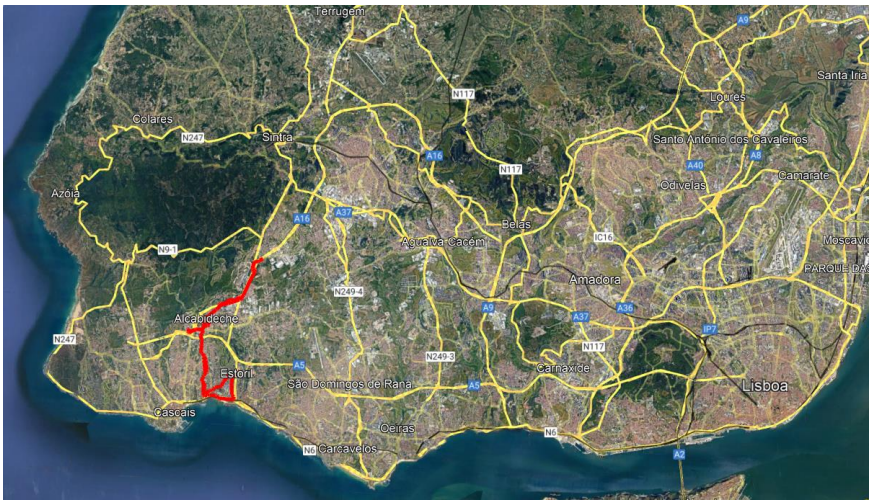
Carreira	Traçado
M03	
M04	
M05	

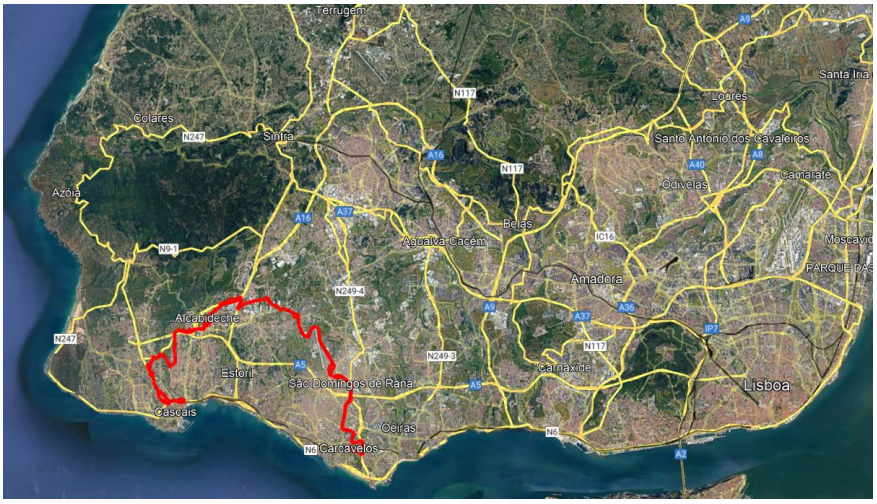
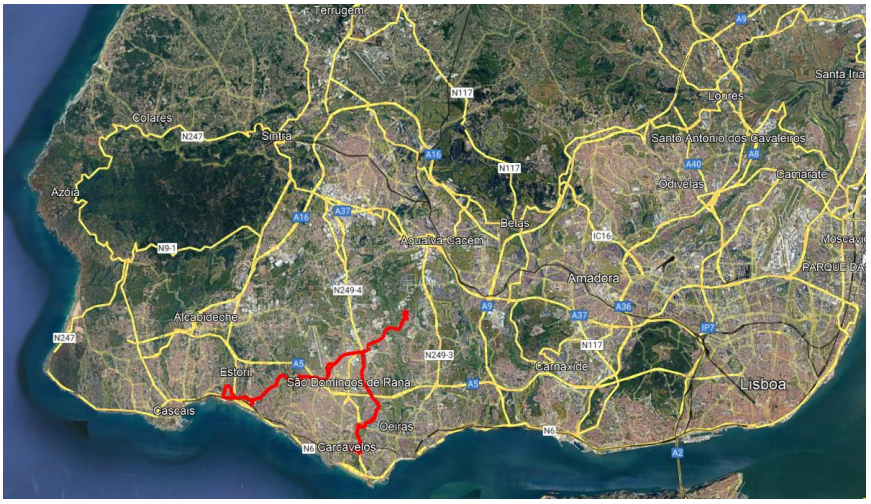
Carreira	Traçado
M06	
M07	
M08	

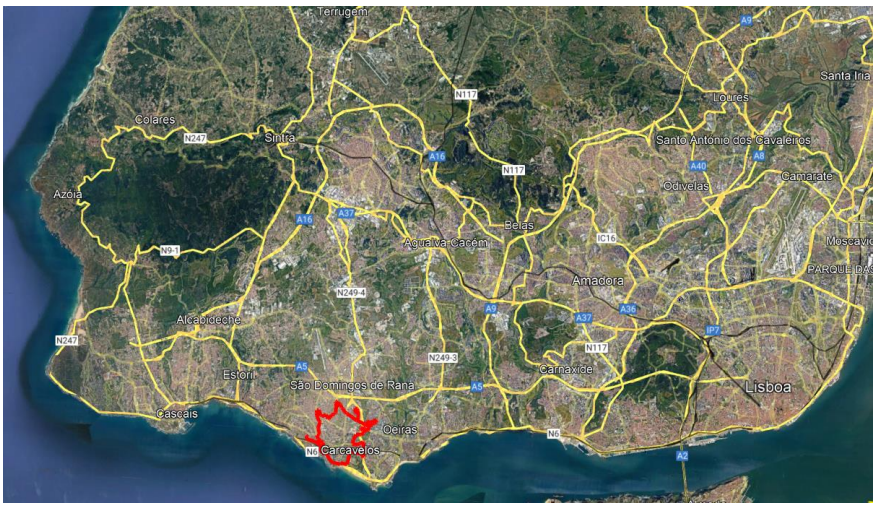
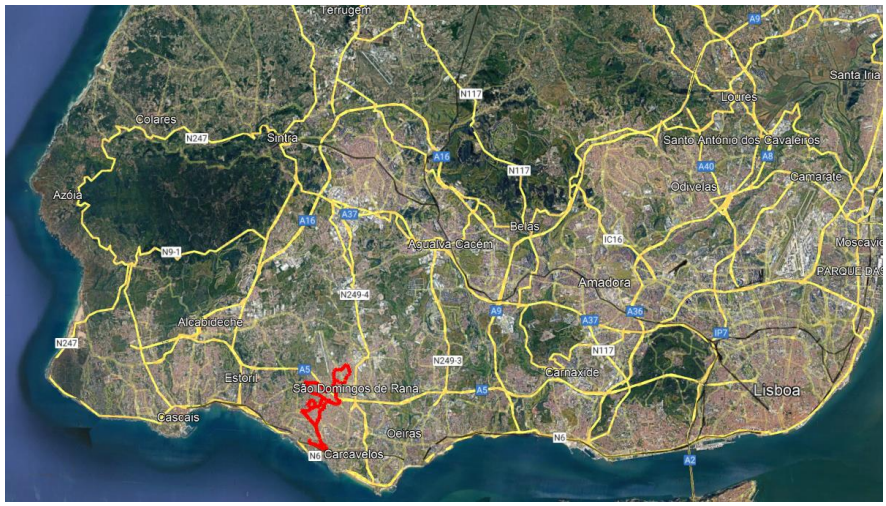
Carreira	Traçado
M09	
M10	
M11	

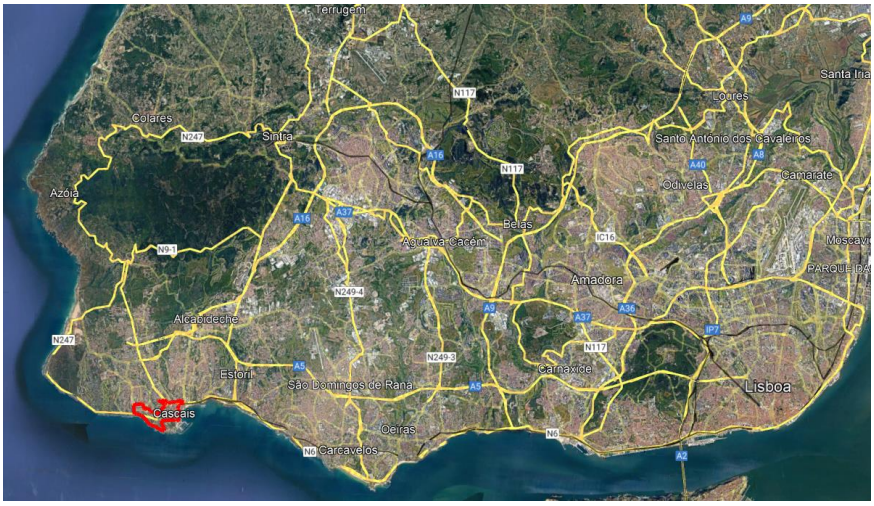
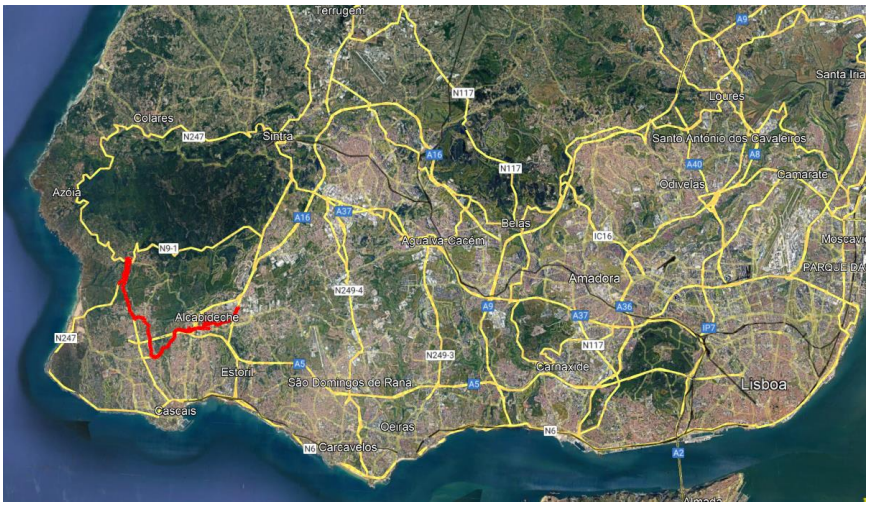
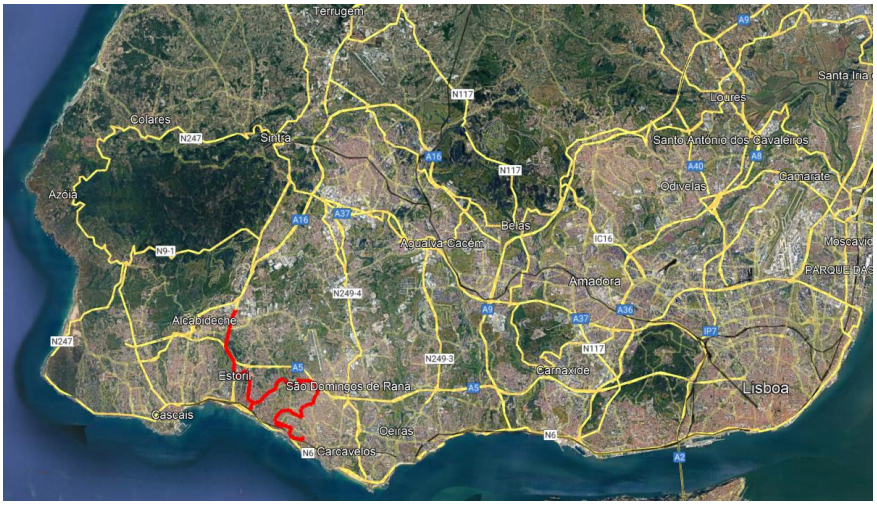
Carreira	Traçado
M12	
M13	
M14	

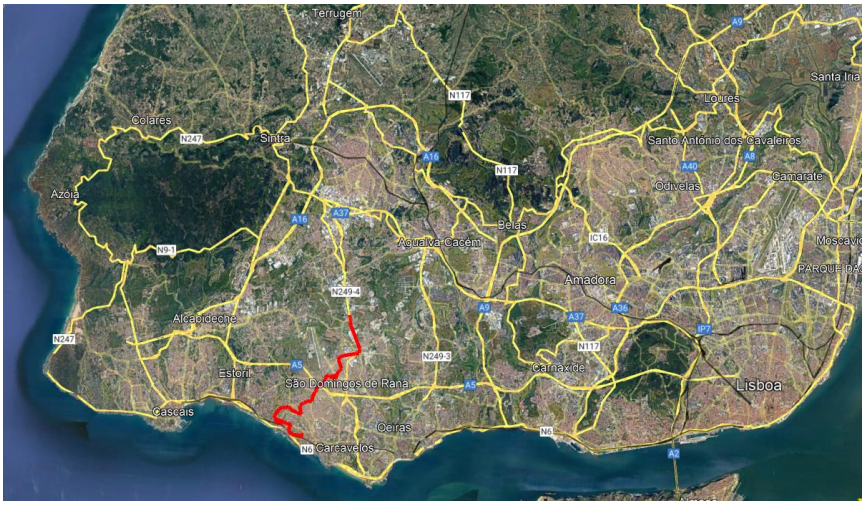
Carreira	Traçado
M15	
M16	
M17	

Carreira	Traçado
M18	
M19	
M20	

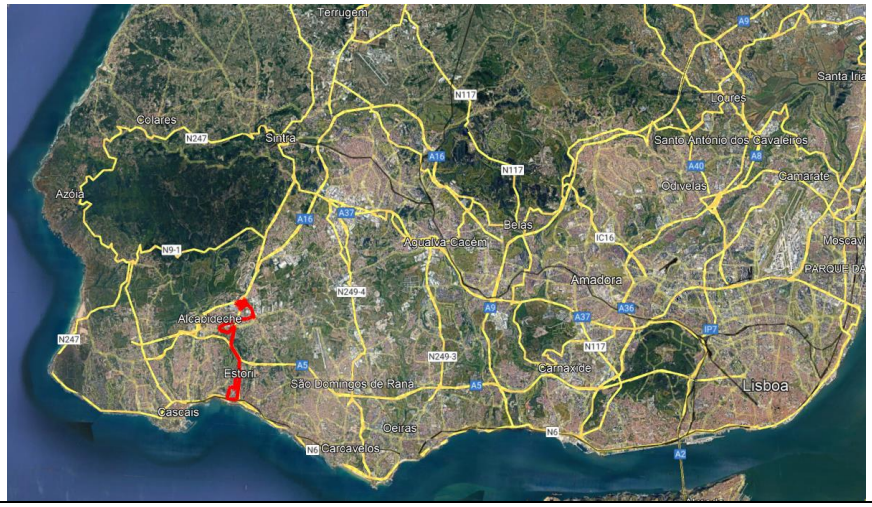
Carreira	Traçado
M21	
M22	
M23	

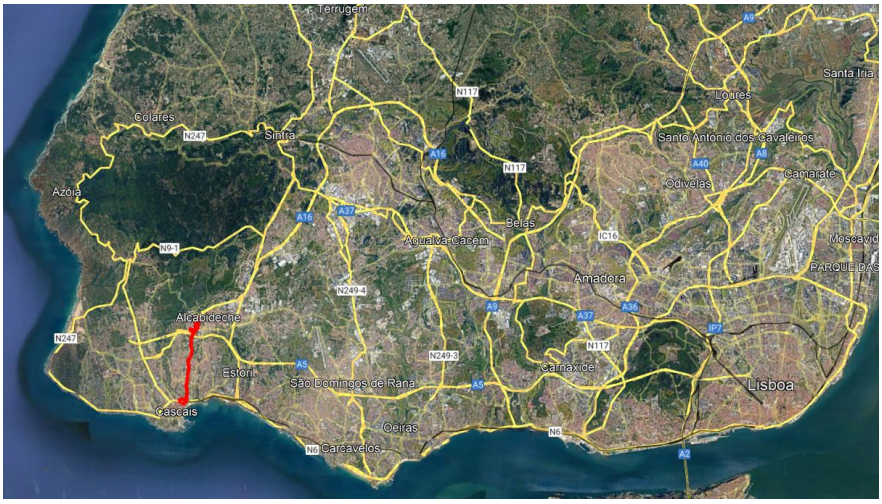
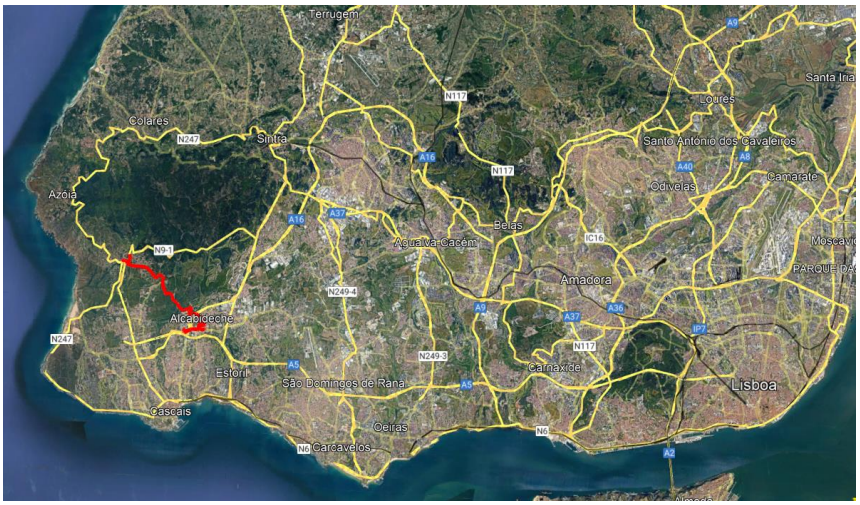
Carreira	Traçado
M24	
M25	
M26	

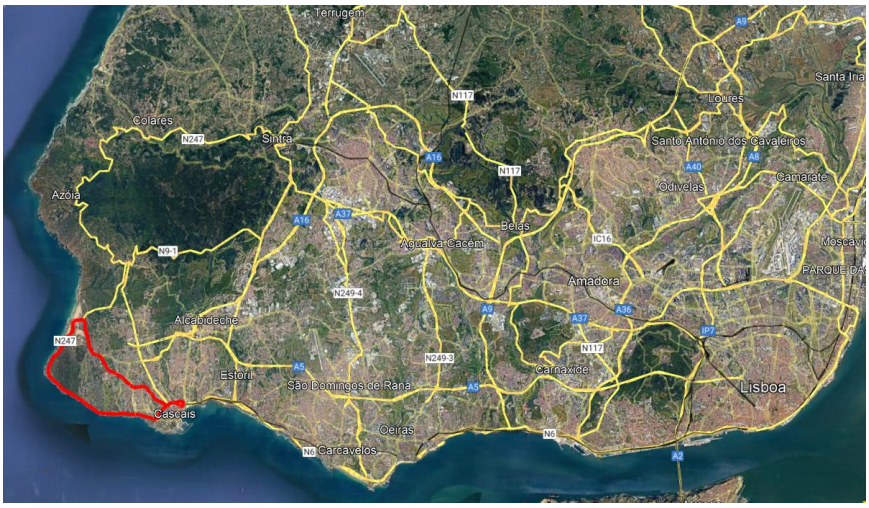
Carreira	Traçado
M27	
M28	
M29	

Carreira	Traçado
M30	
M31	
M32	

Carreira	Traçado
M33	
M34	
M35	

Carreira	Traçado
M36	
M37	
M38	

Carreira	Traçado
M39	
M40	
M41	

Carreira	Traçado
M42	
M43	
M44