U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Telco Voice Traffic: Database Construction and Visualization

Inês Alexandra Cunha Ferreira

Dissertation

Master in Modeling, Data Analysis and Decision Support Systems

Supervised by:

João Manuel Portela da Gama

Bruno Miguel Delindro Veloso

Cláudia Isabel Maia Dias

2022

Acknowledgments

I would first like to acknowledge my supervisors, Professor João Gama and Professor Bruno Veloso, for their feedback and guidance throughout this dissertation.

I am also thankful to NOS for giving me the opportunity of attending this curricular internship at the Planning and Management Control Department, which led me to the development of this project. I would also like to thank my colleagues in this department for their warm and helpful attitude towards me. Particularly, I would like to express my deepest gratitude to my internship supervisor, Cláudia Dias, for all the guidance, insightful suggestions, patience and, most importantly, all the knowledge shared with me.

I want to extend my sincere thanks to my parents. Without you, this dissertation would not be possible. Your endless support and encouragement are invaluable.

Finally, I would like to thank everyone that directly or indirectly contributed to the success of this dissertation.

Abstract

In Portugal, the Telecommunications' Industry is a very competitive sector and, therefore, telecommunications' companies need to analyze their customers' consumption evolution to gain a competitive advantage and mark their position in the industry. Also, it is essential that these companies study the revenues and traffic made by their clients to report to regulatory entities.

This dissertation is within the scope of a curricular internship in the Planning and Management Control Department of NOS. This department has different data sources feeding its models, some of them even having the same data, and needs to build complex models based on the best combination of the different data sources. Given their complexity and lack of visualization, they end up not having the expected results in terms of understandability and usability.

Thus, this work focused on creating a unique relational database, to which were applied data visualization techniques (to both study the voice traffic's metrics present in the database and compare the data of the different data sources). The data used relates to the company's voice services: the traffic and its revenues considered by segment and traffic type, from January of 2019 to May of 2022.

For the construction of this database, several reports had to be created in SAP BusinessObjects BI Launch Pad and imported into Excel, where the database was created. All the tables of the relational model were imported into Power BI, where the relationships between them were established and managed. After all the data present in Power BI, a dashboard was created. This dashboard contains two parts: part I is the study of the voice traffic generating revenues; part II is the integrity check. In Part I of the dashboard, the metrics relating to the traffic generating revenues can be found. These metrics are the minutes, the revenues, the number of calls, among others. All the data found in this part can be studied by traffic type and by customer segment, which allows the department to better study the consumption of the clients, in terms of both traffic and revenues. After this, a comprehensive study of these pages was made and some findings came from it.

In Part II of the dashboard, which is probably the most important part, the three most relevant data sources used by the department (which are the ones present in the database) can be compared using all the different types of voice traffic. After the construction of this part, a study of it was made, which led to relevant findings that even influenced the structure of the first part of the dashboard.

Keywords: Relational Database Construction; Star Schema; Galaxy Schema; Dashboard; Data Integrity; Data Warehouse; Telecommunications; Voice Traffic; Voice Revenues

Resumo

Em Portugal, a Indústria das Telecomunicações é um setor muito competitivo e, por isso, estas empresas precisam de analisar o consumo dos seus clientes para obter uma vantagem competitiva e marcar a sua posição na indústria. Além disso, também é fundamental que estas empresas estudem as receitas e o tráfego dos seus clientes para reportar às entidades reguladoras.

Esta dissertação insere-se no âmbito de um estágio curricular no Departamento de Planeamento e Controlo de Gestão da NOS. O departamento tem diferentes fontes de dados a alimentar os seus modelos, algumas delas até tendo os mesmos dados, e precisa de construir modelos complexos baseados na melhor combinação destas. Dada a sua complexidade e falta de visualização, acabam por não ter os resultados esperados em termos de compreensão e usabilidade.

Por isso, este trabalho focou-se na criação de uma base de dados relacional única, à qual foram aplicadas técnicas de visualização (tanto para estudar as métricas de tráfego de voz presentes na base de dados quanto para comparar os dados das diferentes fontes). Os dados contidos nela referem-se aos serviços de voz da empresa: o tráfego e as suas receitas consideradas por segmento e tipo de tráfego de janeiro de 2019 a maio de 2022.

Para a construção desta base de dados, vários relatórios tiveram que ser criados no SAP BusinessObjects BI Launch Pad e importados para o Excel, onde a base de dados foi criada. Todas as tabelas do modelo relacional foram importadas para o Power BI, onde as relações entre elas foram estabelecidas e geridas. Após todos os dados estarem presentes no Power BI, foi criado um dashboard. Este dashboard contém duas partes: a parte I é o estudo do tráfego gerador de receitas de voz; a parte II é o estudo da integridade.

Na Parte I do dashboard encontram-se as métricas relativas ao tráfego gerador de receita. Essas métricas são os minutos, as receitas, o número de chamadas, entre outras. Todos os dados encontrados nesta parte podem ser estudados por tipo de tráfego e por segmento de cliente, o que permite ao departamento estudar melhor o consumo dos clientes, tanto em termos de tráfego como de receitas. Depois disso, foi feita uma análise dessas páginas e dela surgiram conclusões.

Na Parte II do dashboard, que é provavelmente a parte mais importante, as três fontes de dados mais relevantes utilizadas pelo departamento (que são as presentes na base de dados) podem ser comparadas utilizando todos os diferentes tipos de tráfego de voz. Após a construção desta parte, foi feito um estudo da mesma, que levou a conclusões relevantes que influenciaram a estrutura da primeira parte do dashboard.

Palavras-chave: Construção de Base de Dados Relacional; Star Schema; Galaxy Schema; Dashboard; Integridade de Dados; Data Warehouse; Telecomunicações; Tráfego de Voz; Receitas de Voz

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Problem Description

The uncertainty of customer behaviour, especially in volatile economic and social contexts, is one of the main challenges companies face. Therefore, analyzing their clients' consumption evolution and needs is crucial for companies (to explore market trends, pricing and marketing strategies, etc.).

In Portugal, the Telecommunications Industry is a very competitive sector that comprises much risk, not only due to the competitiveness but also to the regulatory risk. Because of this, telecommunications companies need to analyze their customers' consumption evolution to gain a competitive advantage and mark their position in the industry. Besides this, it is also essential that these companies study the revenues and traffic made by their clients to report to regulatory entities.

This dissertation is within the scope of a curricular internship in the Planning and Management Control Department of NOS, a Portuguese Telecommunications Company.

For the last years, the company's Planning and Management Control Department has faced some difficulties in gathering good and complete data to produce valuable information for the top management's decision-making process and its report to the telecommunications' regulatory entities (such as ANACOM).

NOS's information systems have been undergoing a significant transformation and are still evolving. This transformation has also been taking place in the Data Warehouse, which is not yet fully unified nor cohesive. This led to the present situation, where the department has different data sources, some of them even having the same data, and needs to build complex models based on the best combination of the different data sources.

In the case of telecommunications' companies, in what concerns the voice services, it is crucial to gather different kinds of data about traffic, such as real minutes, billed minutes, interconnection costs and revenues, customer revenues, etc. This is vital to build models that help management understand the customers and their tendencies (to answer their needs by better adjusting their services), know the company services' costs and revenues, report to regulatory entities, forecast future voice traffic, etc. For these models to be trustworthy and meaningful to the company, the data with which they are built needs to be accurate and incorrupt (which can be a difficult task

when there are different data sources).

Currently, there is a lack (and a need) of a single database where every member of the Planning and Management Control team can retrieve the data needed for their everyday work. Even when using the same data (for different models, currently built in Excel), they can retrieve it from different data sources, leading to some incongruencies in the information created with it. Also, the voice traffic models currently in use by the department (which are built in excel) are too complex and lack visualization. Because of this, some information can go unnoticed.

Given the problem faced by the department, this curricular internship has different goals. The first is to build a unique database where the team of the Planning and Management Control Department can retrieve the data of the metrics used to build the models. To support this database, a module of data integrity will also be made to study possible incongruencies and clean the data to get the best possible data and, consequently, the best possible information from it. Lastly, data visualization techniques will be applied to the database to comprehensively visualize the data and its trends, patterns and abnormalities, enabling better decision-making and reporting processes.

1.2 Dissertation Structure

The present dissertation is structured as follows: Chapter 2 includes the main insights drawn from telecommunications, data warehouse, dimensional modelling, data quality, data visualization and dashboards' literature. Chapter 3 contains an elucidation of the data and its relevant hierarchies, a description of the data sources relevant to this study, an explanation of the models in use by the department, the methodology and the software used. Then, Chapter 4 concerns the explanation of the database and integrity check construction, according to the different phases of the methodology. Chapter 5 contains all the steps to build the study of the voice traffic generating revenues part of the dashboard, and there, one can find the images of the dashboard constructed. Also, in this chapter, the findings of the metrics present in this part of the dashboard are explained. Then, Chapter 6 contains the integrity check part of the dashboard, including its construction and findings. Lastly, Chapter 7 includes the conclusions of the work made and some suggestions of future work to the department. Lastly, the references that support this dissertation and the appendix.

Chapter 2

Related Work

2.1 The Telecommunications Sector and Data Mining

Currently, the amount of data created by companies is increasing exponentially, especially in the telecommunications industry, one of the first to adopt Business Intelligence (BI) and Data Mining (DM) technologies. The reason for this was probably that these types of companies have a vast customer base, which leads to a daily generation and storage of enormous amounts of high-quality data, and operate in a rapidly changing and highly competitive environment (Weiss, 2009).

According to Holsheimer and Siebes (1994), data mining is "the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data. These relationships represent valuable knowledge about the database and its objects and if the database is a faithful mirror, of the real world registered by the database". Therefore, industries need to employ data mining to discover the hidden trends and patterns in their large amounts of data (Sumathi & Sivanandam, 2006).

In any industry, the BI and Data Mining applications depend on two crucial factors: the existence of business problems that could be successfully addressed and solved with the help of BI and DM technologies and the availability of data for the application of such technologies (Kabakchieva, 2009). The telecommunications sector fulfils these requirements.

Even though telecom companies have Call Detail Records (CDR) for each call placed on their network, they cannot use them directly for data mining since the goal of DM applications is to extract knowledge at the customer level and not at the individual call level (Weiss, 2005). This is a significant point since telecommunications is a very competitive sector with customers becoming increasingly demanding and the constant change in technology, enhancing the need for mining the data and extracting value from it.

Most of the authors encountered during the research for this report only refer to Data Mining in the telecommunications industry for fraud detection, network fault isolation and prediction, and marketing. Although these are crucial subjects for telecom companies, they are not the only areas where DM technologies can be employed, as will be demonstrated throughout this work.

2.2 Data Warehouse

Increasingly, organizations are analyzing current and historical data to identify useful patterns and support business strategies. The emphasis is on complex, interactive, exploratory analysis of enormous data sets created by integrating data from all parts of a company (Gama & Veloso, 2020).

Two technologies are related to data mining: data warehouse and OLAP (Sumathi & Sivanandam, 2006).

The data warehouse is built from a wide range of sources and is the core of the Business Intelligence system, which is built for data analysis and reporting (Taylor, 2021b).

Data Warehousing is a process for collecting, storing and managing data from various sources to provide relevant business insights (Taylor, 2021b). A Data Warehouse is typically used to connect and analyze business data from heterogeneous sources (the operational systems); that is, data is extracted from the operational systems and is even sometimes combined with additional information from third parties (Hobbs, Hillson, Lawande, & Smith, 2005). The operational systems (or source systems) are the ones that capture the business transactions, are outside the data warehouse (one has little to no control over the content and format of the data in them), and their main priority is to process performance and availability (Kimball & Ross, 2013).

According to Inmon (2002), the data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant data collection in support of management's decisions. They are subject-oriented, which is organized around the company's applications. The integration characteristic is the most important since data is fed from multiple heterogeneous sources into the data warehouse. Therefore, as the data is fed to the data warehouse, it needs to be converted, reformatted, summarized, etc., to be consistent. The DW is nonvolatile because once data is in there, it is stable (does not change). That is, when the data is loaded, it is loaded in a static format and adding new data will not change it (the history of the data is kept). Lastly, it is time-variant since there is some form of time marking to show the moment in time during which the record occurred, allowing to recall of the data of a particular period.

Related to the DW subject, there is the concept of Data Mart, which is a departmental subset of the warehouse data focusing on specific subjects and, therefore, on specific data. A data warehouse could be a union of all the integrant data marts (Chen, 2003). The data in the data mart is denormalized, summarized, and shaped by the operating requirements of a single department (Inmon, 2002).

When talking about DW, it is also important to consider the concept of dimensional modelling, which Ralph Kimball developed. Dimensional modelling is a data structure technique optimized for data storage in a Data warehouse. According to Kimball and Ross (2013), dimensional modelling is the preferred technique for presenting analytic data since it makes databases simple and enables the users to visualize a set of data clearly and tangibly, which is the core of understandability. Dimensional models deliver understandable data to business users and fast query performance.

For Kimball and Ross (2013), dimensional modelling is the most feasible technique for delivering data to Data Warehouse/Business Intelligence users. The dimensional model is simple (which benefits the business users because the data is easier to understand and navigate and allows

for performance benefits since these schemas can be processed more efficiently) and symmetric and is smoothly extensible to accommodate changes in the user behaviour.

In the Data Warehouse/Business Intelligence architecture proposed and developed by Kimball and Ross (2013), first, there are the operational systems that were already mentioned. Because of their characteristics, there is a need for the Extract, Transformation and Load (ETL) System, which comprises a work area, instantiated data structures, and a set of processes. Extraction is the first step of this system and comprises the reading and understanding of the source data and copying the data needed into the ETL system for further manipulation; at this point, the data already belongs to the Data Warehouse. After that, many transformations can add value to the data, such as data cleansing (for example, correcting misspellings, parsing into standard formats, etc.), combining data from multiple sources, and de-duplicating data. The final step of this system is the Loading and physical structuring of data into the presentation area's target dimensional models (it is supposed to deliver the dimension and fact tables of the star schema); this step involves many processes such as surrogate key assignments, splitting or combining columns to represent the appropriate data values, etc.

After the ETL system, the Kimball and Ross (2013) architecture possesses a presentation area to support business intelligence. The data is organized, stored, and made available for direct querying by users, report writers, and other analytical BI applications. In this area, the data is presented, stored and accessed in dimensional schemas (relational star schemas or OLAP cubes). It must contain detailed, atomic data (it cannot have merely aggregated data), it should be structured around business process measurement events, and all the dimensional structures must be built using standard, conformed dimensions.

The Kimball and Ross architecture ends with the business intelligence applications that query the presentation area, which they define as the range of capabilities provided to business users to leverage the presentation area for analytic decision making.

2.3 Dimensional Models: Star Schema and OLAP Cube

As stated in Kimball and Ross (2013), dimensional models can be instantiated in both relational databases, referred to as star schemas, or multidimensional databases, known as online analytical processing (OLAP) cubes. A star schema hosted in a relational database is good physical support for constructing an OLAP cube. The cube is often the final deployment step of a dimensional DW/BI system or may exist as an aggregate structure based on a more atomic relational star schema.

2.3.1 Star Schema

The multidimensional model of a DW in the form of a star schema is the most common modelling paradigm (Han, Kamber, & Pei, 2012). Star schemas consist of fact tables (containing the bulk of the data, with no redundancy) linked to associated dimension tables (one for each existing dimension) via primary/foreign key relationships (Kimball & Ross, 2013). The schema graph resembles a starburst, with the dimension tables displayed in a radial form around the central fact table (Han et al., 2012).

Fact Table

Facts are the measurements that result from a business process event and are almost always numeric (and additive, crucial to BI applications). The term "fact" represents a business measure. Therefore, each row in a fact table will represent a measurement event. The data on each row is at a specific level of detail, referred to as the grain. One of the main rules of dimensional modelling is that all the measurement rows in a fact table must be at the same grain (Kimball & Ross, 2013).

Fact tables have foreign keys (as many as the dimension tables) that connect to the dimension tables' primary keys. The tables satisfy referential integrity when all the keys in the fact table correctly match their respective primary keys in the corresponding dimension tables (Kimball & Ross, 2013). One of the most important characteristics of the fact table is that it has a composite key, that is, a key composed of a subset of the foreign keys (which are primary keys in the dimension tables). The fact table contains two columns: the foreign keys connecting to the dimension tables and the measures (the numeric facts) (Shams Raza & Nayak, 2014).

The granularity of the fact table represents the level of detail of the data stored in it. High granularity means that the data is low in the data hierarchy. It is very detailed (defines data to the most precision) and can be called atomic data. Low granularity refers to data that is high in the data hierarchy. It is aggregated data and not very detailed (Shams Raza & Nayak, 2014).

Dimension Table

The dimension tables contain the textual context associated with a business process measurement event, that is, the facts (Kimball & Ross, 2013). Dimension tables contain the descriptive attributes that BI applications use for filtering and grouping the facts (the attributes in these tables serve as the primary source of query constraints, groupings, and report labels). With the grain of a fact table firmly in mind, all the possible dimensions can be identified. Kimball and Ross (2013) also states that a dimension should be single-valued when associated with a given fact row.

It is also important to remember that dimension tables cannot be joined and tend to have fewer rows than fact tables. Despite this, they can be wide, with many columns (Kimball & Ross, 2013). Also, the primary keys of each of the dimension tables (each dimension table has a single primary key, which serves as the basis for referential integrity) are always part of the composite key of the fact table (Kimball & Ross, 2013); (Shams Raza & Nayak, 2014).

2.3.2 OLAP cube

OLAP stands for Online Analytical Processing, an array-based multidimensional database that enables processing and analyzing numerous data dimensions much more rapidly and efficiently than a traditional relational database (IBM, 2020). This system manages large amounts of historical data, enables summarising and aggregating, and stores and manages information at different levels of granularity (Gama & Veloso, 2020). OLAP gives the answers to multidimensional business questions quickly and easily (Sumathi & Sivanandam, 2006).

The OLAP cube extends the single table by adding extra layers, each layer including additional dimensions, usually the next level in the hierarchy of the dimension (IBM, 2020).

The OLAP is a natural and flexible extension of the data warehouse. The detailed and historical data constitutes the best foundation for the OLAP level of data (Inmon, 1996).

Both Inmon (1996) and Kimball and Ross (2013) believe that the star schema can be used to build the OLAP environment and that OLAP cubes can be equivalent in content to, or more often derived from, a relational star schema.

According to Inmon (1996), the OLAP environment is sometimes called the data mart, the lightly summarized or departmentally structured data warehouse level. The OLAP environment is customized for the department it serves, a characteristic of the data marts.

OLAP cubes allow four types of multidimensional analysis: roll-up, drill-down, pivot, and slicing and dicing (IBM, 2020).

The roll-up operation allows to aggregate data at different levels of a dimension hierarchy (Gama & Veloso, 2020), that is, it converts more detailed data into less detailed data, moving up in the hierarchy (reducing the number of dimensions in the cube) (IBM, 2020). The drill-down operation performs the inverse of the roll-up one, moving down in the hierarchy by converting less detailed data into more detailed data.

The pivot operation allows transposition or rotation on selected dimensions (Gama & Veloso, 2020) to display a new representation of the data, providing dynamic multidimensional views of the data (IBM, 2020).

The slice operation creates a sub-cube by selecting a single dimension from the main OLAP cube, for example, by selecting only the sales data of the year 2019 (time dimension) (IBM, 2020). The dice operation isolates a sub-cube by selecting several dimensions within the main OLAP cube, for example, selecting only the sales data of 2019 (time dimension) for a specific product (product dimension).

2.4 Data Quality

There is a strong relationship between data quality and the quality of the decisions made with it (Howard, Lubbe, & Klopper, 2011).

Kimball and Ross (2013) assert that three factors have converged to put data quality concerns near the top of the list for executives. The first one is that the knowledge workers of current times believe that data is a crucial requirement for them to function in their jobs. The second is that most organizations understand that their data sources are very distributed; therefore, they need to integrate them effectively. The third is that careless data handling will no longer be overlooked or excused.

For companies, better data means fewer mistakes, lower costs, better decisions, and better products. Further, the companies that do not give data quality its due importance will have difficulties in surviving in future business environments (Redman, 2017). The reasoning behind this is ”garbage in, garbage out”. Low-quality input/data will produce low-quality output/decisions, making organizations lose money.

However, how does one measure data quality? Bakhshaliyeva (2021) provides six indicators that companies can use to assess if their data fits their needs. The first one is completeness, and companies should identify the critical data elements required for analysis and ensure that those

elements are present and flawless. The second one is consistency. Companies must ensure that data stored in separate locations are in sync and uniform. The third is conformity; since data is sometimes collected and stored in inconsistent formats and data types, the companies must ensure that each record conforms to the same standards. The fourth is accuracy; inaccurate data makes it difficult for businesses to derive any insights from it since the base data used for the analysis cannot be trusted. The fifth one is the integrity of the data, which is often used interchangeably with data quality. However, data integrity refers to the plenitude of the relationships between data sets. The last is timeliness; high-quality data is complete and available for analysis immediately.

According to Sivathanu, Wright, and Zadok (2005), several factors cause unexpected or unauthorized changes to stored data. Therefore, quick detection of integrity violations is vital for the reliability and safety of the stored data.

Data stored on a storage device or transmitted across a network can get corrupted due to hardware or software malfunctions. A malfunction in hardware could also trigger software problems resulting in serious damage to stored data. Bugs in software could also result in unexpected modifications of data. Unreliable networks can also corrupt data that passes through them. Besides this, highly critical and confidential information is stored electronically and accessed through several different interfaces, leading to security vulnerabilities. Damage to data integrity can cause more critical problems than confidentiality breaches since important information may be modified by malicious programs, malicious users, or faulty system components. Then, there are user errors, which can compromise data integrity at the application level, leading to integrity violations (Sivathanu et al., 2005).

During the ETL process proposed by Kimball and Ross (2013), the data cleaning occurs. Those cleaning subsystems aim to assemble technology to support data quality. Their goals should comprise early diagnosis and triage of data quality issues, requirements for source systems and integration efforts to supply better data, specific descriptions of data errors expected to encounter in the ETL, a framework for capturing all data quality errors and for precisely measuring data quality metrics over time and, lastly, the attachment of quality confidence metrics to final data.

Data from disparate sources must be reconciled when constructing a data warehouse. This is because it can make mistakes in its mapping and transformation that may lead to corrupt data in the data migration process. These kinds of errors may create issues such as missing records, missing values, incorrect values, duplicated records, poorly formatted values, broken relationships across tables or systems, etc. (Taylor, 2021a).

As found in Guo, Liu, and Sun (2018), there are a variety of strategies to improve data quality. This improvement can be made in two ways throughout the whole life cycle of data: one is from the perspective of prevention. That is, in every stage of the life cycle of the data, strict data planning and restriction are required to prevent dirty data from occurring. The other is after the post-diagnosis. That is, because of the evolution or integration of data, dirty data will emerge gradually, and a specific algorithm or technique should be used to detect the dirty data. When talking about data quality issues, Giordano and Onions (2021) states that Data Governance may also be used as the solution for those problems (for prevention).

Al-Ruithe, Benkhelifa, and Hameed (2019) and Stedman (2020) explain Data Governance

as being the "process of managing the availability, usability, integrity and security of the data in enterprise systems", based on agreed-upon internal data standards and policies that also control data usage (who can take what action, upon what data, in what situations, using what methods). Effective Data Governance ensures that data is consistent and trustworthy and is used by the right people in the right way. It is increasingly critical as organizations face new data privacy regulations and rely more and more on data analytics to help optimize operations and drive business decision-making. In Alhassan, Sammon, and Daly (2019), it is said that Data Governance mainly focuses on who holds the decision rights related to an organization's data assets to ensure the quality, consistency, usability, security, privacy, and availability of the data.

According to Kimball and Ross (2013), one of the main goals of Data Governance is to agree on data definitions, labels and domain values so that everyone is "speaking the same language". Otherwise, the same thing may be described by different words, the same words may describe different things, and the same value may have different meanings. Kimball and Ross (2013) believe that defining a foundation of master descriptive conformed dimensions requires effort. However, it is an effort worth undertaking. Another goal of Data Governance is establishing policies and responsibilities for data quality and accuracy and data security and access tools. For Kimball and Ross (2013), strong Data Governance is required for conforming information, regardless of the technical approach.

2.5 Data Visualization and Dashboards

Humans have a very visual mind, and many professionals need to see and understand the data graphically in order to transform it into comprehensible information that can be used to help them make decisions (Aparicio & Costa, 2015). Dashboards are cognitive tools and can help improve the "span of control" over many business data (Brath & Peters, 2004).

A dashboard is used in organizations to visualize useful data for the decision-making process. It aims to inform while not distracting the users from their actual tasks. Because of this requirement, the data displayed in the dashboards need to be summarized using charts, tables, etc. (Janes, Sillitti, & Succi, 2013). Dashboards help users identify patterns, trends and abnormalities, reason about what they see and help them make effective decisions (Brath & Peters, 2004).

As found in Wind (2005), properly created dashboards provide the mechanism to drive effective management and resource allocation decisions since they explicitly link key metrics and the drivers and processes that affect them.

Given that there is a fragmented data overload in the current times (multi-channel management, an increase in product lines and services, etc.), there is a need for greater data organization. Also, human processing capabilities are limited, and managerial biases arise from shortcuts in information processing and decision making (Pauwels et al., 2009). Organizations (such as Telco companies) provide a wide range of services and, therefore, may use disparate systems to manage and maintain their operations. Therefore, data may be generated from multiple operational systems. They are managing that data correctly and applying dashboard techniques to it to bring information all into one place, providing better data visibility.

Dashboards are essential for many purposes. They enforce consistency in measures and

measurement procedures, help monitor metrics for corrective action, may be used for planning, communicate to important stakeholders, etc. A dashboard should provide a framework for recognizing good performance, diagnosing poor performance, and evaluating different options for corrective action, and it should be a source of organizational learning, increased profitability and decision making (Pauwels et al., 2009).

According to the research of Pauwels et al. (2009), for developing a dashboard, first, it is necessary to select the key metrics; then populate the dashboard with data; and, after that, establish the underlying relationships between the dashboard items and metrics (metrics alone do not address cause-and-effect relationships). Later, applying the dashboard's underlying model to scenario planning and budget setting (for example, by making historical pattern studies or using what-if analysis). Dashboards are very useful for creating information with the data and staying on top of issues, but they also enable their users to find new business scenarios and profitable opportunities.

As found in the research of Pauwels et al. (2009), the adoption and success of dashboards depend on five primary factors. First is the demand, which depends on the users, the organizational decision style, interdepartmental relations, and the industry (different industries have different requirements for dashboards). The second factor is the supply, such as the availability of the metrics, the sophistication of the dashboard, the visual display (since information can be displayed in several different ways) and the drill-down capabilities (going from a lower to a higher level of detail). The third factor is the fit between the demand and supply since the information provided should be in line with the decision-making responsibilities of the users, and the metrics in the dashboard should be those that are crucial for the industry or the company. The fourth factor is the implementation process, which involves many people and processes, and many errors along the way; therefore, cooperability and communication are fundamental for this to be accomplished. Lastly, it is required that decision-makers are convinced that they will perform better with the dashboard, that the numbers are reliable, and that they need to have high expectations about the dashboard while still considering the existence of initial problems.

It is very important to consider that it is possible to make ineffective dashboards. This may happen due to the enormous amount of charts and graphs available, together with the lack of training in data visualization. The decision-makers need access to good and precise information, and, because of that, it is crucial to choose the right metrics and visuals (Brath & Peters, 2004).

To end this section, Few (2006) describes the dashboard as "a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance". However, Shaffer (2018) disagrees, especially with the part where Few (2006) says that the visual display must be consolidated on a single screen. Shaffer (2018) states that there are times when a consolidated view is, in fact, necessary, but that is not always the case. He also states that not every question can, or should, be answered at a glance.

Chapter 3

Methodology

3.1 Data Description

To better understand the problem, it is essential to describe some concepts about the data being handled by the department. The data used for this project was real data provided by the company. However, it is important to note that, due to the competitiveness of the telecommunications market, some confidentiality issues arise. This confidentiality does to not allow to show and describe everything in detail and, therefore, in the cases where it is not possible to explain in detail, the description will be kept at a more abstract level.

First, it is essential to note that, even though NOS offers many services, the problem in question will only involve voice services. The company's voice traffic (made by its clients) and its revenues will be considered by segment from January of 2019 to May of 2022.

Regarding NOS's client base, it can be divided into two segments: Consumer (also referred to as B2C) and Business (also referred to as B2B). Both of these segments will be considered in the model. Also, these segments can be divided into sub-segments. In the case of the B2C segment, the sub-segments are the services offered: bundle-wired, bundle-wireless and stand-alone. In the case of the business segment, since its service is mostly stand-alone, the sub-segment is different. However, it cannot be disclosed due to confidentiality issues. Therefore, the sub-segments of the business segment are generically referred to as B2B_1, B2B_2, B2B_3 and B2B_4.

Regardless of the sub-segment, the technology offered can be Fixed Voice (if the call is originated from a fixed telephone) or Mobile Voice (if the call is originated from a mobile cell phone).

It is also important to note that if the service relates to a stand-alone consumer with mobile technology, it can either be a postpaid subscription (subscribers receive a bill after each month of service) or a prepaid subscription (clients pay an upfront cost for one month of service). In the case of the other types of customers, there is no need to distinguish between postpaid or prepaid subscriptions since they only subscribe to postpaid.

One possible and straightforward representation of the two customer segments can be found below.

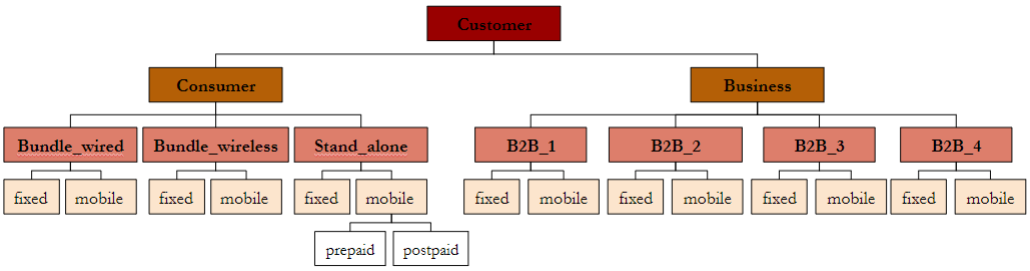

Figure 3.1: Hierarchy of the Customers

As said above, the service contracted by the consumer can be in a bundle (wired or wireless) or stand-alone and, therefore, the minutes made can be on-bundle or off-bundle since the contract of each different client contains a limited amount of voice traffic. The minutes are on-bundle if they are part of those limited minutes included in the bundle, and off-bundle if the amount of minutes made exceeds the limits of the contract. If the minutes are off-bundle, they will represent an extra consumption, resulting in an extra charge to the client. These extra consumptions are the ones that represent extra revenue to the company.

Regarding the traffic, it can either be Basic (if it is from a client to another person) or Special Services (if it is from a client to a service, such as TV contests, customer service lines, etc.).

In the case of the basic traffic, the calls can either be national (if they are made to destinations inside of Portugal) or international (if they are made to clients of other operators outside of Portugal). National calls can either be to a mobile or fixed destination's technology while International calls can either be to destinations inside (DEEA) or outside (FEEA) the European Economic Area.

It should also be mentioned that the International traffic, particularly the DEEA, is regulated and has a pre-defined maximum price tariff (the current tariff is of 0,19€/min, since May of 2019).

In the case of the special services traffic, the calls can be to non-geographical numbers (NNG), such as the TV contests and some customer service lines, or to short numbers (NC or NCurtos), which most commonly have five digits (but can have 3 to 6 digits, shorter than the usual 9 digit numbers of the National Numbering Plan) and are mainly dedicated to customer service support lines and information services.

The NNG numbers are numbers with no geographical significance (geographical significance means that the number refers to a physical/geographical location of the network), and the ranges included in the study are 707x, 76x (includes the lines 760x and 761x), 800x, 808x and 882x. These lines represent different services, according to their first 3 digits: the 707x refer to universal access services, mainly dedicated to customer support lines, and where the maximum retail prices are fixed (€0.09/min (without VAT) for calls originating on fixed networks and €0.13/min (without VAT) for calls originating on mobile networks, with billing per second from the first minute); the 76x lines relate to premium rate numbers used in TV contests and polls and have a predefined maximum retail price per call, which is €0.60 (excluding VAT) per call for 760x, and €1 (excluding VAT) per call for 761x, regardless of the duration, time and origin of the call; the 800x range relates to "green numbers", which are free of charge; the 808x are the "blue numbers" and relate

to shared cost numbers also with maximum retail prices fixed, depending on the peak/off peak time; lastly, the 882x is the virtual calling card service with no special service retail tariff.

It is also important to note that the National Health Service line (SNS 24) is part of the NNG 808x line. This service became free in March of 2020 due to the pandemic. This means that the 808x line increased in minutes (due to the calls to SNS 24) but not in revenues.

NCurtos comprise the lines 16x (dedicated to TELCO Operators' customer support lines), 18x (national informative service regarding telephone service subscribers) and Others (which in this project will mainly be called Outrosx, and relate to remote sales lines, activation lines and other services).

In order to best understand the hierarchy of the traffic, the representation below was designed.

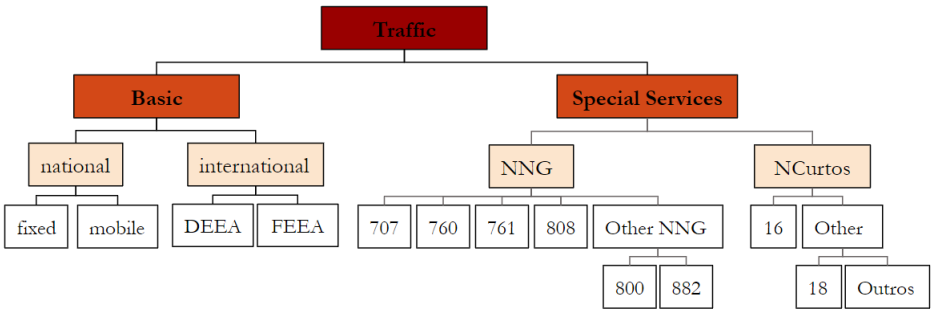

Figure 3.2: Hierarchy of the Traffic

Regardless of the type of traffic, the calls may be on-net (if they are made to NOS's clients) or off-net (if they are made to clients of other operators). In what concerns the Basic National traffic studied, the traditional P&L view is based on the split of the traffic between on-net and off-net (fixed and mobile). Because of this, the decision made first was to divide the Basic National traffic in this way.

However, it is more significant to the department to study between Mobile and Fixed destination's technology instead of OnNet or OffNet. For studying the traffic revenues, the real difference in the price paid by the clients (for the calls they made) is between fixed and mobile destination's technology and not between on-net and off-net calls. For example, let's imagine a client with mobile technology making calls to national basic numbers; the study of this traffic's revenues is more relevant if it is made in terms of the destination's technology (if the client called a mobile or fixed number) than if it was made in terms of the destination's operator (if the client called another NOS client - on-net - or clients from other operators - off-net). Therefore, it was decided to study Basic – Mobile traffic (includes OnNet mobile and OffNet mobile traffic) and Basic – Fixed traffic (includes OnNet fixed and OffNet fixed traffic). This decision was also made due to findings during the integrity check (section 6.2.1).

Additionally, it is worth noting that this distinction between on-net and off-net traffic can only be made to the special services and to the basic national traffic. The international traffic can never be on-net (remember that being on-net means that the calls are made within NOS' network).

Regarding the metrics of the database being constructed, these include: Real Airtime (or Minutes), Real Airtime On-bundle, Real Airtime Off-bundle, Billed Airtime, Billed Airtime On-bundle, Billed Airtime Off-bundle, Number of Calls (only important for Special Services traffic), Revenues (in euros), % of Billed Minutes Off-Bundle, Billing Factor (in %), Average Revenue per Minute (RMM), Average Revenue per Call (RMC) and, lastly, the proxy to NOS's market share in the lines 760x and 761x (NOS Weight).

Regarding the traffic measured in Minutes, it is imperative to distinguish between Real and Billed Minutes (and between on-bundle and off-bundle, which was mentioned previously). Real Minutes are the actual minutes made by NOS's clients, while the Billed Minutes are the number of minutes billed by the company to their clients (for example, a client making a call of 1 minute and 20 seconds could be billed as making 2 minutes, depending on the tariff plan contracted).

The Number of Calls is only essential when studying the 760x and 761x lines of the Special Services traffic since these are value-added service phone lines dedicated to contests and polls and are charged a universal tariff per call. Therefore, these two lines need to be studied using the number of calls instead of the number of minutes, since the calls to these two lines are also always of very few seconds and, therefore, the minutes have no significant meaning.

In this study, the Revenues will be the focus in what concerns amount of euros. Despite this, it is important to understand some concepts about the traffic's revenues and costs. Regarding the Basic Traffic calls (client to client): if they are on-net, they only generate revenue to NOS, but if they are off-net, they generate both a revenue and an interconnection cost (because the call ended in another operator's network and, therefore, a fee must be paid). Regarding the Special Services Traffic calls (client to service): if they are on-net, they only generate a PVP revenue (the retail price that the client has to pay for that service call), but if they are off-net, they generate not only the PVP revenue but also a PVP operator cost (that PVP must be handed off to the operator that owns that service, therefore this revenue and cost cancels out), an Origination revenue (because the call originated in NOS's network) and a Billing and Collection revenue (this fee is due to the risk that NOS incurs when billing the client that made the call, since the client may not pay).

Lastly, the % of Billed Minutes Off-Bundle (or % Off-Bundle) is a division between the OffBundle Billed Airtime and the Billed Airtime and represents the percentage of billed minutes that are off-bundle. The Billing Factor (in %) is a division between the Billed Airtime and the Real Airtime and shows the percentage of minutes that the client is billed in the minutes they actually make. The Average Revenue per Minute (RMM) is the Average Revenue per OffBundle Billed Minute and is a division between the Revenues and the OffBundle Billed Airtime. The Average Revenue per Call (RMC) is a division between the Revenues and the Nbr. Of Calls. The proxy to NOS's market share in the lines 760x and 761x (NOS Weight) allows one to study the weight of these lines belonging to NOS in the totality of the calls made to these lines by the clients of NOS. This is done by dividing the on-net real minutes made to these lines by the total of real minutes (on-net plus off-net) made to these lines (this metric will be further explained in section 5.4.1).

The different granularities across the data in this department imply the need for aggregation and disaggregation of the data to analyze its evolution at both the lower and higher level of granularity. Also, it is essential to note that the Planning and Management Control department needs to handle revenues, costs, margins, and calculated metrics at the different levels of granularity,

since all this data is necessary for planning (budgeting and forecasting), controlling and reporting (both to internal and to external stakeholders as well as regulatory entities).

3.2 Data Sources Description

A study of the data sources and the data warehouse structure used by the department was made, and a visual representation of it can be found below.

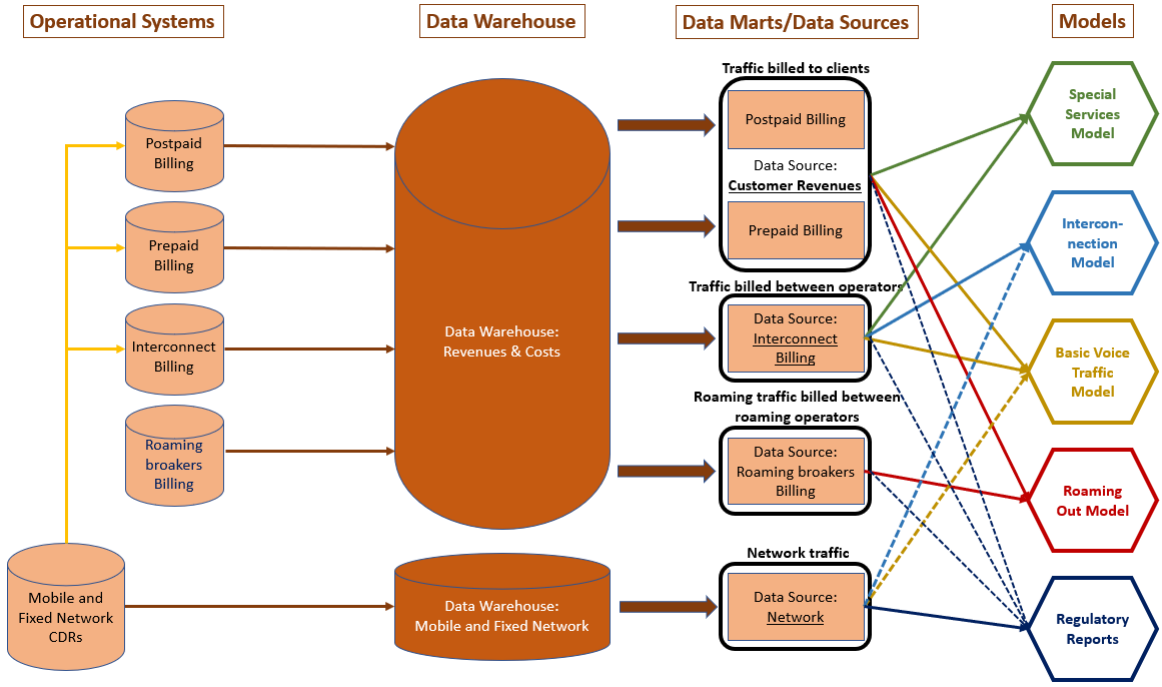

Figure 3.3: Representation of what is used by the department

As already said, the department's models gather data from different data sources. For this study, there are three relevant data sources: the Customer Revenues (will be called CR), the Network and the Interconnection Billing (the INTEC).

In what concerns the data present in each data source, the sources are not 100% compatible. This is because they have different systems behind them, with different necessities, and performing different but complementary roles.

The INTEC data source is fed by the interconnection billing system existing between operators (therefore presents only off-net traffic) and its quality is guaranteed by the business rules. In this data source, the necessary detail is by traffic interconnection route (that is, traffic originating in a national or international operator and ending in a client of NOS and/or traffic originating in a client of NOS and ending in another national or international operator). It is possible to register international traffic's delivery flows via routes with national operators (a client of NOS calls an international number and that call is handed off to another national operator that, in its turn, delivers it to that international operator) that will be typified as a national destination in

INTEC and as an international destination in Network and CR (since the billing to the client and the network CDR are of international destination). This may explain some of the differences existing between the data sources in the basic traffic.

Still considering the INTEC data source, because the necessary detail is by traffic interconnection route, in the case of the international traffic, the minutes include not only the traffic (made in Portugal by NOS's clients) destined to international operators, but also the traffic regarding calls received by NOS's clients when they are in roaming out of Portugal, and therefore are connected to other international operators' networks, that is, when the clients are abroad and receive calls.

The Network data source has all the CDRs (Call Detail Records). This data source is fed on a daily basis, with CDRs coming from the different technologies across NOS's network architecture and platforms, which are constantly evolving. Therefore, this data source is more prone not only to have some information issues (when traffic is not properly identified, or when some CDRs are missing in the data integration process), but also to be subject to some delay in keeping up with the network's architecture and platforms developments (depending on the Data Warehouse proper developments). Nevertheless, the Network data source is very rich, in terms of the detailed data it provides, so it is a very relevant source for the reporting and business analysis.

Figure 3.3 shows that one arrow leading to the Interconnection Model and one arrow leading to the Basic Voice Traffic Model are dashed. This is because these models only use the data from the Network for very few model lines or only just for projection since this is the most updated source (it has daily data, which enable the monthly estimates). However, according to the members of the department, the Network Traffic data source may be the most unreliable one because it has delays in fully integrating the developments of Network CDRs. Despite this, this is the main source for the regulatory reports because it has convenient data according to the details needed. Later on, in the integrity check, it was discovered that, actually, the Network data source is more trustworthy than the CR data source.

In what concerns the CR data source, it is a business intelligence data mart that joins both postpaid billing and prepaid charging systems. The CR data source's role is to gather revenues information (not traffic) and is fed by these operational systems which assure the correct charging to NOS's clients. Therefore, the CR data source contains only the traffic details essential for these billing purposes. This means that this source may not have enough detail to perform a fully detailed integrity check, in what concerns the real voice traffic (for instance, it will not properly distinguish between on-net and off-net calls or even between fixed and mobile destination's technology).

It should also be mentioned that, in order to complete the information coming from the operational systems, the CR data source was recently developed so that additional traffic metrics were added, based on some algorithms applied to both the postpaid and prepaid charging systems, which enabled the calculation of the On-Bundle Real and Billed Minutes (more on this can be found in section 6.2.3).

Also, the prepaid charging systems are fed by the same CDRs as the Network data source and, therefore, the CR source can have the same problems as the Network.

Because of all these specific characteristics of the different data sources and given that all of them serve different purposes, they are not completely comparable.

Lastly, it is also worth mentioning that, from the Data Warehouse perspective, the data marts are directly on the right side of it, having only subsets of the DW. However, when looking from the constructed models perspective (when looking from the department perspective), these data marts are considered data sources of these models. Because of that, the terms data marts and data sources will be used interchangeably.

3.3 Study of the Models Currently in Use by the Department

First, to understand the department's needs, it is necessary to study the models (present on the right side of figure 3.3) already in use by the team, the data included in them, and where it comes from (the data sources). This is because the database to be constructed will have the data that the team uses for its models.

Regarding the models constructed by the department, they exist not only to help the company understand the monthly evolution of the clients' traffic consumption, divided by Basic and Special Services Traffic, but also to evaluate NOS' margin related to these consumptions.

Nevertheless, why are there four models when there could be only one with all the data studied? The different models represent different requirements for the department.

The Special Services Traffic model needs to be separated from the Basic Traffic Model because the Special Services traffic is heavily regulated, especially in the lines related to customer service (lines 16x, 707x and 808x). These lines have been going through regulatory changes, and therefore is very important to study them separately. Also, the TV contests lines (760x and 761x) have a unique seasonality and very low margins, despite the high revenues, increasing the relevance of studying them individually because of the different impacts they have both on Turnover and EBITDA.

The Interconnection Model is essential to study how much NOS is paying its competitors for the consumption of their clients. It is worth mentioning that the amount of traffic included in the bundles is growing bigger and bigger, which means that, even though NOS will not have an extra revenue from those minutes, it still needs to pay the destination operators for those calls that end in their network.

The Basic Traffic Model is especially relevant to study the evolution of the Basic National Traffic margin, particularly in the mobile technology, since this is very valuable to NOS. The value of this traffic is even increasing, which can be confirmed in section 5.2.2, given that it was found in the dashboard that the minutes of clients with fixed technology are decreasing in the basic traffic, while the minutes of clients with mobile technology are increasing.

Lastly, the Regulatory Reports are the most important since, as was already said, the telecommunications industry is very regulated (the Portuguese regulator is ANACOM), and there are periodic reporting obligations regarding voice traffic and related revenues. Additionally, NOS is also required to report to BEREC (which is the Body of European Regulators for Electronic Communications), namely in what concerns international traffic, that also has strict rules that establish a limit for the tariffs the companies charge their clients for international calls inside the European Economic Area (EEA).

3.4 Methodology

The study of the data, the data sources, the models in use by the department and the data warehouse structure used by the department was already made (and can be found above), and a visual representation of it can be found in figure 3.3. Taking that schema into consideration, the visual representation of one of the project's goals (the database creation one) can be the following.

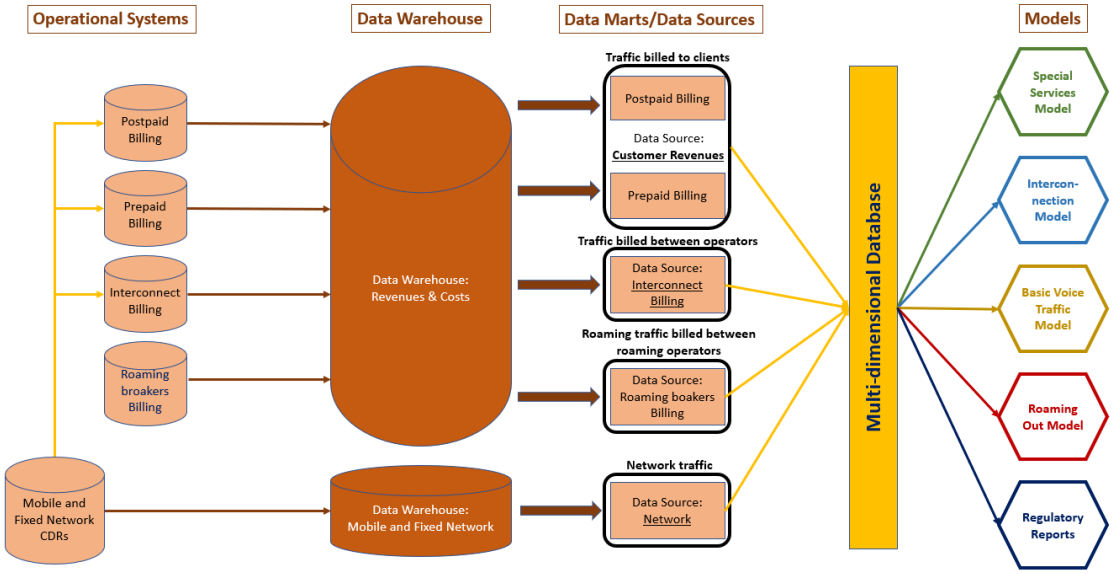

Figure 3.4: Representation of what is expected to be implemented

The multidimensional database will be constructed by reconciling the data from the different data marts. According to Han et al. (2012), Inmon (1996) and Kimball and Ross (2013), the multidimensional model can exist in the form of a star schema since the OLAP database can be derived from a star schema. Because of this, the choice was to organize the OLAP environment as this type of structure.

According to Kimball and Ross (2013), the design of the star schema can be divided into four steps: selection of the business process, declaration of the grain, identification of the dimensions, and identification of the facts.

The business processes are the operational activities performed by the organization. These business process events generate metrics that translate into facts (in a fact table). Most fact tables focus on the results of a single business process. Choosing the business process is very important because it defines a specific design target and allows the grain, dimensions and facts to be declared (Kimball & Ross, 2013). Kimball and Ross (2013) also state that dimensional models should not be designed solely to answer specific questions or to build specific reports since the business users' requirements may change over time.

For the problem in hand, the business process was already selected from the beginning and refers solely to the voice traffic made by NOS's clients, as was mentioned above in section 3.1.

The declaration of the grain is the crucial step in a dimensional design. This grain establishes exactly what a fact table row represents, and it must be defined before the dimensions or facts since these must be consistent with the grain. This consistency enforces uniformity on all dimensional designs, which is critical to BI applications and the ease of use. Kimball and Ross (2013) encourage to start by focusing on atomic data and alert to the fact that different grains must not be mixed in the same fact table. In the case of the problem under study, the declaration of the grain was chosen to be: one row per different type of call/voice traffic. In this case, a different type of call refers to the different possible combinations of customers, traffic and period.

Concerning the dimension tables, Kimball and Ross (2013) raised awareness to the fact that the textual data should be put into dimensions where they can be correlated more effectively with the other textual attributes in that dimension and consume much less space. Also, these authors state that the designer should not store redundant text in fact tables since text belongs in the dimension tables (the text may only be present in the fact table if it is unique for every row). Lastly, attributes should consist of real words rather than cryptic abbreviations.

Having this, and the declaration of the grain in mind, the dimension tables (containing the descriptive attributes of the business process, as explained in section 2.3.1) were constructed. It is important to note that Kimball and Ross (2013) also mentioned that data governance is crucial in the development of the dimension tables. A glossary of the dimensions and the facts will be constructed to cover this point, which will help the members of the Planning and Management Control Department navigate the model better. This glossary will be accompanied with a document regarding the maintenance of the database and of the dashboard.

The facts, as said in section 2.3.1, are the measurements that result from the business process event and are almost always numeric. These facts were already mentioned in section 3.1 as being the metrics required. Another relevant thing referred to by Kimball and Ross (2013) is that the designer must not try to fill the fact table with zeros representing no activity because these zeros would overwhelm most fact tables. Instead of zeros, these cells will not be filled and will be blank.

Following all these steps, and according to the description of the data presented above, a proposal of a star schema was designed and is the following.

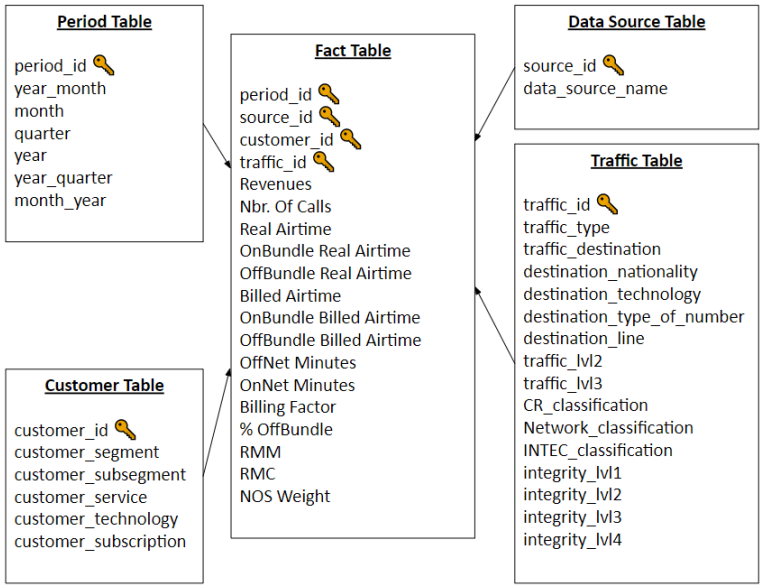

Figure 3.5: Star Schema designed

In figure 3.5, the dimensions can be seen as Period, Customer, Data Source and Traffic. Each of these dimensions comprises the relevant attributes that describe them.

The dimensions' attributes and the textual values in them can be found in section 4.1.

Later, because, as said before, the data can be extracted from different data sources of the data warehouse and because even the same metric can be extracted from different sources (this happens for the Real Minutes, for example), this data will need to have its integrity checked and, if needed and possible, corrected.

In the case of the problem in hand, the members of the department, through the use of this data, already understand that some of the data sources are not as reliable as others. This information will be used to do this integrity check and correction.

Lastly, the database constructed will be imported into Power BI and used to create a dashboard where the data needed by the department will be presented, enabling the retrieval of important information. This database, when in Power BI, will also be used to perform the integrity check.

3.5 Software

In order to retrieve the data from the different existing data sources of the data warehouse, SAP BusinessObjects BI Launch Pad (which will be referred to as BO) and Microsoft's Excel will be used.

SAP BusinessObjects BI Launch Pad has a simple and intuitive user interface, which makes it easier to access, view, organize and manage the BI objects. This software allows the user to export data and information to other business intelligence applications (such as Microsoft Excel), and save it to a determined location (SAP, 2017).

The multidimensional database will be built in Microsoft's Excel, where the reports retrieved from SAP BusinessObjects will be imported.

Lastly, the database will be imported into the Power BI Desktop application and, there, the relationships of the database will be created. The dashboard regarding the study of the traffic's revenues and the integrity check will be created in this software.

Power BI is a platform that opens BI to everyone (Lachev, 2015). It offers a set of tools that connect knowledge workers with the piles of data they would all like to find insights in to share (Becker & Gould, 2019). This application allows analysts to design data models and reports. Power BI dashboards summarize important metrics so that one can get a quick high-level view of how your business is doing at a glance (Lachev, 2015).

According to Becker and Gould (2019), because of disparate data sources and widely differing reporting needs, simplicity, flexibility, and wide distribution are key features. These features are provided by Power BI and the related Power Query tools within Excel, along with advanced analysis capabilities. Power BI Desktop combines Power Query with a visualization and analysis layer, where custom calculations beyond standard descriptive statistics can be made, and reports can be created.

Power Query, also known as Get and Transform, is presented as a graphical user interface window that opens when summoned from Power BI Desktop. Here, users compile a series of steps (query) interpreted to construct a table of data to be loaded to Power BI's visualization tool. The resulting tables can then be filtered and sectioned to focus on the data needed for the specific project and can be loaded to the Power BI's visualization layer (Becker & Gould, 2019).

In Power BI Desktop, multiple tables can be used in a set of visualizations and linked together in a data model similar to those of traditional relational database systems. Putting all the data into a single table before beginning exploratory analysis and visualization is not necessary. A Power BI report contains one or more pages on which one or more visuals can be grouped. Assembling a visualization is a drag-and-drop process that comprises different types of techniques. Bar, line, and area charts are provided in many forms and scatter plots, treemaps, and map-based displays are also available. Dashboard-type presentations are enabled by tables, pivot matrices, and various indicator cards to spotlight a single figure. Fonts, colours, and other presentation settings can be easily customized. Power BI provides easy construction and interactivity. The visualizations on any report page can interact with each other, enabling filtering based on side-by-side visualizations (Becker & Gould, 2019).

Power BI allows the reporting process automation, connecting different sources of data in one data model, interactive and multidimensional reports, possibility to track correlation among the data, quick and easy analysis, standardization and reliability (lower risk of human error), and collaboration and safe report sharing with any member of the organization (Becker & Gould, 2019).

Lachev (2015) asserts that the most common way of bringing data into Power BI Desktop is by importing it from relational databases. When importing tables from a relational database that supports referential integrity and has table relationships defined, Power BI Desktop detects these relationships and applies them to the model. However, when no table joins are defined in the data source, or when the data imported is from different sources, Power BI Desktop might be unable to detect these relationships upon import. Because of this, the model must be revisited

and the appropriate relationships must be created before analyzing the data.

The relationship between tables is an essential subject. The dimensional model mentioned in this work will be constructed in Excel, which will require several sheets (as many as the number of dimension tables plus the fact table). Therefore, when importing this model to Power BI, the relationships between the tables/sheets will need to be established and managed.

Chapter 4

Database Construction and Dashboard Preparation

The first thing to do when constructing the database was to build the necessary dimension tables.

4.1 Dimension Tables

4.1.1 Data Source Table

The data source table is where the data sources' names (Customer Revenues, referred to as CR throughout the work; Interconnect Billing, referred to as INTEC; and Network) and their id (unique to each data source) is.

Later, it was decided that for the dashboard's Part I (the one with the study of the traffic's revenues), only the fact table of the CR data source would be used. This decision was made after performing the integrity check, where it was concluded that there are differences between the different data sources (making it preferable to have the metrics coming from only one data source) and the need to study some metrics that are only available in this data source, such as the revenues and the billed minutes (on and off-bundle). Also, the fact tables (needed to perform the integrity check) were kept separately in Power BI; therefore, the data source table was no longer necessary. Despite this, the table was kept in case of future needs of the department.

data_source_id	data_source_name
101	Customer Revenues
102	INTEC
103	Network

Figure 4.1: Data Source Dimension Table

4.1.2 Period Table

The period table concerns the period when the traffic occurred, being this a monthly period. The period id (unique to each year month), the year-month, the month, the year, the quarter, the year quarter and the month year (in date form) can be found in different columns of this table.

Most of these fields ended up not being needed since the last one allows to travel up and down the period hierarchy (year, quarter, month) in Power BI since it is in date form.

period_id	year_month	month	year	quarter	year_quarter	month_year
201901	201901	jan	2019	Q1	2019Q1	jan/19
201902	201902	fev	2019	Q1	2019Q1	fev/19
201903	201903	mar	2019	Q1	2019Q1	mar/19
201904	201904	abr	2019	Q2	2019Q2	abr/19
201905	201905	mai	2019	Q2	2019Q2	mai/19
201906	201906	jun	2019	Q2	2019Q2	jun/19
201907	201907	jul	2019	Q3	2019Q3	jul/19
201908	201908	ago	2019	Q3	2019Q3	ago/19
201909	201909	set	2019	Q3	2019Q3	set/19
201910	201910	out	2019	Q4	2019Q4	out/19
201911	201911	nov	2019	Q4	2019Q4	nov/19
201912	201912	dez	2019	Q4	2019Q4	dez/19
202001	202001	jan	2020	Q1	2020Q1	jan/20
202002	202002	fev	2020	Q1	2020Q1	fev/20
202003	202003	mar	2020	Q1	2020Q1	mar/20
202004	202004	abr	2020	Q2	2020Q2	abr/20
202005	202005	mai	2020	Q2	2020Q2	mai/20
202006	202006	jun	2020	Q2	2020Q2	jun/20
202007	202007	jul	2020	Q3	2020Q3	jul/20
202008	202008	ago	2020	Q3	2020Q3	ago/20
202009	202009	set	2020	Q3	2020Q3	set/20
202010	202010	out	2020	Q4	2020Q4	out/20
202011	202011	nov	2020	Q4	2020Q4	nov/20
202012	202012	dez	2020	Q4	2020Q4	dez/20

Figure 4.2: Part of the Period Dimension Table

4.1.3 Customer Table

The customer table contains the segment's attributes to which the customers (those who made the voice traffic) belong. In this table, one can find the segmentation used by the department (which is called PCG Detailed Segment and is very detailed), along with the customer id (unique to each detailed segment) and the aggregations made: the customer subscription (prepaid or postpaid), the customer technology (mobile or fixed origination), the customer service (standalone, bundle, bundle wired or bundle wireless), the customer segment (consumer or business) and the customer subsegment (consumer, B2B_1, B2B_2, etc.).

Later on, other aggregations were made. These aggregations on the consumer and business segments (which will not be shown due to confidentiality issues) are according to some more detailed rules of the department. The goal is to offer some flexibility to the dashboard user, so that they can study different customers aggregated by specific characteristics.

In figure 4.3 only part of the used customer table can be found. This is because the detailed segment used by the department and the other aggregations made cannot be shown due to confidentiality issues. Also, not all the entries are shown since they do not all fit into one image.

customer_id	customer_segment	customer_subsegment	customer_service	customer_technology	customer_subscription
101103	Consumer	Consumer	Bundle_wireless	Mobile	Postpaid
103202	Consumer	Consumer	Stand_alone	Fixed	Postpaid
103201	Consumer	Consumer	Stand_alone	Fixed	Postpaid
101102	Consumer	Consumer	Bundle_wireless	Mobile	Postpaid
101201	Consumer	Consumer	Bundle_wireless	Fixed	Postpaid
101200	Consumer	Consumer	Bundle_wireless	Fixed	Postpaid
101101	Consumer	Consumer	Bundle_wireless	Mobile	Postpaid
101100	Consumer	Consumer	Bundle_wireless	Mobile	Postpaid
103200	Consumer	Consumer	Stand_alone	Fixed	Postpaid
102104	Consumer	Consumer	Bundle_wired	Mobile	Postpaid
102103	Consumer	Consumer	Bundle_wired	Mobile	Postpaid
102201	Consumer	Consumer	Bundle_wired	Fixed	Postpaid
102200	Consumer	Consumer	Bundle_wired	Fixed	Postpaid
102102	Consumer	Consumer	Bundle_wired	Mobile	Postpaid
102101	Consumer	Consumer	Bundle_wired	Mobile	Postpaid
102100	Consumer	Consumer	Bundle_wired	Mobile	Postpaid
243100	Business	B2B_1	Stand_alone	Mobile	Postpaid
243101	Business	B2B_1	Stand_alone	Mobile	Postpaid
243103	Business	B2B_1	Stand_alone	Mobile	Postpaid
243102	Business	B2B_1	Stand_alone	Mobile	Postpaid
243200	Business	B2B_1	Stand_alone	Fixed	Postpaid
233103	Business	B2B_2	Stand_alone	Mobile	Postpaid
233102	Business	B2B_2	Stand_alone	Mobile	Postpaid
233101	Business	B2B_2	Stand_alone	Mobile	Postpaid

Figure 4.3: Part of the Customer Dimension Table

4.1.4 Traffic Table

The last dimension table, the traffic table, is a way of understanding and making compatible the different detailed traffic types across the different data sources in order to be able to compare them. Therefore, one can find in this table the classifications existing in the different data sources (which have different names for the same traffic).

In this table, one can find the different aggregations of the traffic along with the traffic id (unique to each detailed type of traffic). The traffic type (basic traffic or special services traffic), the traffic destination (international or national), the destination nationality (International - DEEA, International - FEEA or National), the destination technology (International - DEEA, International - FEEA, OnNet – Fixed, OnNet – Mobile, OffNet – Fixed, OffNet – Mobile, OffNet or OnNet), the destination type of number (International, OnNet, OffNet - Fixed, OffNet - Mobile, NCurtos or NNG), the destination line (International, OnNet – Fixed, OnNet – Mobile, OffNet – Fixed, OffNet – Mobile, NCurtos, NNG 707x, NNG 760x, NNG 761x, NNG 808x or Other NNG) and other different types of aggregations can also be found in this table.

As was already explained in section 3.1, in what concerns the Basic National traffic studied, the traditional P&L view is based on the split of the traffic between on-net and off-net (fixed and mobile). Because of this, the first decision was to divide the Basic National traffic this way. However, it is more significant to the department to study between Mobile and Fixed destination technology instead of OnNet and OffNet. Therefore, it was decided to study Basic – Mobile traffic (includes OnNet mobile and OffNet mobile traffic) and Basic – Fixed traffic (includes OnNet fixed and OffNet fixed traffic).

Additionally, concerning the special services traffic, it is important to mention that the CR data source does not have a perfect distinction between on-net and off-net traffic, so this data source aggregates these two destinations into one. Therefore, and because it is relevant for the integrity analysis of this type of traffic to distinguish between on-net and off-net, it was decided

(together with the internship supervisor at NOS) to equalize the Off-Net traffic of the CR data source to the one from the INTEC data source and then, calculate the On-net traffic of the CR data source by subtracting this Off-Net to the total traffic. This choice is due to the assumption that the INTEC data source is very reliable in what regards off-net traffic.

This assumption of the INTEC always having reliable data is because this source is fed by the billing system existing between operators, which has its quality guaranteed by the business rules. Therefore, the department has solid confidence in this data source.

The special services traffic of the CR data source (which comprises both on-net and off-net) was classified in the traffic table as being off-net (only because all types of traffic need to be classified). Also, to be able to attribute the INTEC off-net to the CR off-net, the totality of the CR special services traffic needs to be classified as off-net.

Because of this need to calculate the on-net and off-net minutes, another aggregation was made in the traffic table to make it easier for the end-user. The types of traffic that are purely off-net (in all data sources) are classified as 1 in the aggregation-id column, the ones that are purely on-net (in all data sources) are classified as 2, and the ones that are purely off-net in both the INTEC and the Network data sources but are a sum of on-net and off-net in the CR data source are classified as 3.

It is also important to note that some exceptions exist in the CR special services traffic. The traffic classified as NCurtos 18x and NNG 882 is totally off-net (this was discovered during the integrity check and was later confirmed by one of the members of the department).

traffic_id	aggregation_id	traffic_type	traffic_destination	destination_nationality	destination_technology	destination_type_of_number	destination_line	traffic_lvl2	traffic_lvl3	integrity_lvl1	integrity_lvl2	integrity_lvl3	integrity_lvl4	CR_classification	Network_classification	INTEC_classification
51000	1	Basic Traffic	International	DEEA	DEEA	International	International	International	DEEA	Basic Traffic	International	International	International	Voz - Internacional - DEEA		
52000	1	Basic Traffic	International	FEEA	FEEA	International	International	International	FEEA	Basic Traffic	International	International	International	Voz - Internacional - FEEA		
50000	1	Basic Traffic	International	International	International	International	International	International		Basic Traffic	International	International	International		Voz - Internacional	Voz - Internacional
53100	2	Basic Traffic	National	National	OnNet - Fixed	OnNet	OnNet - Fixed	National	Basic - Fixed	Basic Traffic	National	Basic - Fixed	Basic - Fixed	Voz - OnNet - Fixo	OnNet - Fixo	
53200	2	Basic Traffic	National	National	OnNet - Mobile	OnNet	OnNet - Mobile	National	Basic - Mobile	Basic Traffic	National	Basic - Mobile	Basic - Mobile	Voz - OnNet - Móvel	OnNet - Móvel	
54100	1	Basic Traffic	National	National	OffNet - Fixed	OffNet - Fixed	OffNet - Fixed	National	Basic - Fixed	Basic Traffic	National	Basic - Fixed	Basic - Fixed	Voz - Nacional Fixo - Tráfego Básico	Voz - Nacional Fixo - Tráfego Básico	Voz - Nacional Fixo - Tráfego Básico
54200	1	Basic Traffic	National	National	OffNet - Mobile	OffNet - Mobile	OffNet - Mobile	National	Basic - Mobile	Basic Traffic	National	Basic - Mobile	Basic - Mobile	Voz - Nacional Móvel - Tráfego Básico	Voz - Nacional Móvel - Tráfego Básico	Voz - Nacional Móvel - Tráfego Básico
63014	2	Special Services	National	National	OnNet	Ncurtos	NCurtos	Ncurtos	Ncurtos 16x	Special Services	Ncurtos	Ncurtos 16x	Ncurtos 16x		ONNET_NCURT_16	
63015	2	Special Services	National	National	OnNet	Ncurtos	NCurtos	Ncurtos	Ncurtos Outros	Special Services	Ncurtos	Ncurtos Outros (18x + outros)	Ncurtos 18x		ONNET_NCURT_18	
63016	2	Special Services	National	National	OnNet	Ncurtos	NCurtos	Ncurtos	Ncurtos Outros	Special Services	Ncurtos	Ncurtos Outros (18x + outros)	Ncurtos Outros		ONNET_NCURT_OUTROS	
63021	2	Special Services	National	National	OnNet	NNG	NNG 707	NNG	NNG 707	Special Services	NNG	NNG 707	NNG 707		ONNET_NNG_707	
63022	2	Special Services	National	National	OnNet	NNG	NNG 760	NNG	NNG 760	Special Services	NNG	NNG 760	NNG 760		ONNET_NNG_760	
63023	2	Special Services	National	National	OnNet	NNG	NNG 761	NNG	NNG 761	Special Services	NNG	NNG 761	NNG 761		ONNET_NNG_761	
63024	2	Special Services	National	National	OnNet	NNG	Other NNG	NNG	Other NNG	Special Services	NNG	NNG 800	NNG 800		ONNET_NNG_800	
63025	2	Special Services	National	National	OnNet	NNG	NNG 808	NNG	NNG 808	Special Services	NNG	NNG 808	NNG 808		ONNET_NNG_808	
63026	2	Special Services	National	National	OnNet	NNG	Other NNG	NNG	Other NNG	Special Services	NNG	NNG 882	NNG 882		ONNET_NNG_882	
64014	3	Special Services	National	National	OffNet	Ncurtos	NCurtos	Ncurtos	Ncurtos 16x	Special Services	Ncurtos	Ncurtos 16x	Ncurtos 16x	Voz - Serv NCurtos 16Out	OFFNET_NCURT_16	Voz - Serv NCurtos 16
64015	1	Special Services	National	National	OffNet	Ncurtos	NCurtos	Ncurtos	Ncurtos Outros	Special Services	Ncurtos	Ncurtos Outros (18x + outros)	Ncurtos 18x	Voz - Serv NCurtos 18	OFFNET_NCURT_18	Voz - Serv NCurtos 18
64016	3	Special Services	National	National	OffNet	Ncurtos	NCurtos	Ncurtos	Ncurtos Outros	Special Services	Ncurtos	Ncurtos Outros (18x + outros)	Ncurtos Outros	Voz - Serv NCurtos Outros	OFFNET_NCURT_OUTROS	Voz - Serv NCurtos Outros
64021	3	Special Services	National	National	OffNet	NNG	NNG 707	NNG	NNG 707	Special Services	NNG	NNG 707	NNG 707	Voz - Serv NNG 707	OFFNET_NNG_707	Voz - Serv NNG 707
64022	3	Special Services	National	National	OffNet	NNG	NNG 760	NNG	NNG 760	Special Services	NNG	NNG 760	NNG 760	Voz - Serv NNG 760	OFFNET_NNG_760	Voz - Serv NNG 760
64023	3	Special Services	National	National	OffNet	NNG	NNG 761	NNG	NNG 761	Special Services	NNG	NNG 761	NNG 761	Voz - Serv NNG 761	OFFNET_NNG_761	Voz - Serv NNG 761
64024	3	Special Services	National	National	OffNet	NNG	Other NNG	NNG	Other NNG	Special Services	NNG	NNG 800	NNG 800	Voz - Serv NNG 800	OFFNET_NNG_800	Voz - Serv NNG 800
64025	3	Special Services	National	National	OffNet	NNG	NNG 808	NNG	NNG 808	Special Services	NNG	NNG 808	NNG 808	Voz - Serv NNG 808	OFFNET_NNG_808	Voz - Serv NNG 808
64026	1	Special Services	National	National	OffNet	NNG	Other NNG	NNG	Other NNG	Special Services	NNG	NNG 882	NNG 882	Voz - Serv NNG 882	OFFNET_NNG_882	Voz - Serv NNG 882

Figure 4.4: Traffic Dimension Table

Then, the process of constructing the fact tables could begin.

4.2 Fact Tables

In the first approach to the database, one unique fact table was built by appending the fact tables of the different data sources in Excel. This approach was discarded because having three fact tables in Excel (one for each data source, respecting each source's information structure) can be more beneficial to the department since these require less manual work and are easier to maintain and update than a unique fact table. Also, it was discovered that it is easier to append the three tables in Power BI (using the "Append Queries as New" functionality to form the total

fact table) to achieve the desired unique fact table. Despite this, it was later decided that to study the traffic generating revenues, only the CR fact table was necessary; therefore, there is no need to append the fact tables. However, the three fact tables are still needed in the Excel and Power BI databases to conduct the integrity check, and, to do this, it is more practical to have the tables separated.

In order to build the 3 fact tables in Excel, 4 files had to be constructed, one with the metrics from the CR data source, one with the metrics from the INTEC data source and three with the metrics from the Network data source (one from the Fixed Network, which concerns the traffic made from fixed technology of clients with fixed wired voice, the other from the Mobile Network, which concerns the traffic made from mobile technology of clients with mobile voice or fixed wireless voice, and the last one bringing these two together to form the total Network file).

The database was later more developed by adding the integrity check (which was previously made in a separate excel file). This Integrity Check Module will be extensively explained in section 4.3.

4.2.1 File from the CR data source

This type of file was already in use by the department to build their models, but they had to separate the basic off-net traffic (mobile and fixed), the basic on-net traffic (mobile and fixed), the international traffic and the special services traffic into different reports given that all of them came from different levels of the information hierarchy of BO. During the work to build this database, it was discovered that it was possible to use one of the levels of information in the BO system (which is not being used for any other aggregation) to do the appropriate aggregation of each type of traffic. That aggregation was made and uploaded to BO; therefore, all types of traffic could then be extracted from one single level and don't need to be separated into different reports.

Additionally, all the metrics regarding the number of minutes (real and billed: total, off-bundle and on-bundle) and the call metrics need to be in a different report from the revenues metric. This is because different filters are applied to both the revenues report and the minutes/calls report to ensure that the information is accurate and in accordance with business rules.

Both the Minutes Report and the Revenues Report have three identifier columns, which identify the year and month when the traffic occurred (Year Month), the detailed segment of the customer that made the traffic (PCG Detailed Segment) and the type of traffic made (Traffic Type Aggregated). These BO reports contain some filters that cannot be disclosed due to confidentiality issues. Nevertheless, in regards to the filters that can be disclosed, a filter was applied to the PCG Detailed Segment so that the null values are not brought to the reports, and a filter was applied to the Traffic Type Aggregated so that the null values and the traffic types classified in this field as Unknown are also not brought to the reports. Later on, after the metrics findings (more specifically, the finding in section 5.3.2), it was decided to filter out all the Detailed Descriptions containing "unknown" (these existed even though the Unknown traffic types were filtered out of the Traffic Type Aggregated field). The Detailed Description (present in all data sources) is the most informative field since it has the traffic type at the atomic level.

The report with the minutes' metrics has seven more columns: the amount of total real minutes made (Real Airtime), the amount of on-bundle real minutes (OnBundle Real Airtime), the amount of off-bundle real minutes (OffBundle Real Airtime), the amount of total billed minutes (Billed airtime), the amount of on-bundle billed minutes (OnBundle Billed Airtime), the amount of off-bundle billed minutes (OffBundle Billed Airtime), and the amount of calls made (Nbr. Of Calls).

The revenues' metric report has only one more column besides the three identifier ones. This column is the one with the amount of voice traffic's revenues, that is, the amount of money NOS charges for the traffic made by its clients (Revenues).

After exporting the BO file with the two reports into Excel, some manipulation must be made. To have all the metrics in the same table, the XLOOKUP function of Excel is used to bring the Revenues column of the Revenues Report to the Minutes Report, using a concatenation of the three identifier columns to match the correct period, segment and type of traffic between the two reports. The total table presents 11 columns.

Then, to build the final table to bring to the database file, the XLOOKUP function is once again used to build the period_id, customer_id and traffic_id columns. The period_id column is built by searching the Year Month column's values in the column year_month of the period table of the database and returning the period_id. The customer_id column is built by searching the PCG Detailed Segment values in the database's respective column PCG Detailed Segment of the customer table and returning the customer_id. The traffic_id column was built by searching the Traffic Type Aggregated values in the column CR_classification of the traffic table of the database and returning the traffic_id.

The final table to bring to the database (to the CR fact table) presents 12 columns: data_source_id, which in the case of this table is 101 (this id can be found in the data source table of the database); period_id; customer_id; traffic_id; Real Airtime; OnBundle Real Airtime; OffBundle Real Airtime; Billed Airtime; OnBundle Billed Airtime; OffBundle Billed Airtime; Nbr. Of Calls; and, lastly, Revenues.

Lastly, when already in the database file, four other columns are added to the fact table:

1. aggregation_id (built using the XLOOKUP function to look for the traffic_id in the traffic table and return the aggregation_id)

2. period_customer_traffic (which is a concatenation of the period_id, customer_id and traffic_id to help fill the column OffNet Minutes)

3. OffNet Minutes (built using the IFS function to check whether the aggregation_id is 1, 2 or 3. If it is 1 (off-net in all data sources), then the cell takes the value of the Real Airtime of that entry; if it is 2 (on-net in all data sources), then the cell takes the value of 0; if it is 3 (off-net in INTEC and Network and a mix of both in the CR data source), then the cell value is given by the XLOOKUP function, which will search for the period_customer_traffic of that entry in the corresponding column of the INTEC fact table and will return the corresponding Real Airtime (of the INTEC fact table)).

4. OnNet Minutes (built using the IFS function to check whether the aggregation_id is 1, 2 or 3. If it is one, then the cell takes the value of 0; if it is two, then the cell takes the value

of the Real Airtime of that entry; if it is 3, then the cell takes the value of the difference between the Real Airtime and the OffNet Minutes of that entry).

Four other metrics should be part of this fact table: Billing Factor (which is a division between the Billed Airtime and the Real Airtime), % OffBundle (which is a division between the OffBundle Billed Airtime and the Billed Airtime), the RMM Billed OffBundle (which is the Average Revenue per OffBundle Billed Minute and is a division between the Revenues and the OffBundle Billed Airtime), and the RMC (which is the Average Revenue per Call and is a division between the Revenues and the Nbr. Of Calls).

At first, these four metrics were calculated in the excel database, but it did not work. This happened because the database is at the finest level of detail and the plots in Power BI are at a more aggregated level, making it mandatory for the software to make sums of the metrics, which of course, does not work well on unitary/percentual metrics. These metrics need to be added to the Power BI software as "New Measures" to make them work.

The expressions used to calculate the metrics were the following:

```
% Off-Bundle = CALCULATE((SUM(CR_fact_table[OffBundle Billed Airtime])/SUM(CR_fact_table[Billed Airtime]))*100)
```

Figure 4.5: % Minutes Off-Bundle measure

```
Billing Factor (%) = CALCULATE((SUM(CR_fact_table[Billed Airtime])/SUM(CR_fact_table[Real Airtime]))*100)
```

Figure 4.6: Billing Factor measure

```
RMC = CALCULATE(SUM(CR_fact_table[Revenues])/SUM(CR_fact_table[Nbr. Of Calls]))
```

Figure 4.7: RMC (Average Revenue per Call) measure

```
RMM = CALCULATE(SUM(CR_fact_table[Revenues])/SUM(CR_fact_table[OffBundle Billed Airtime]))
```

Figure 4.8: RMM (Average Revenue per Minute) measure

4.2.2 File from the Network data source

As said before, this data source can be divided into two "data sources", the Fixed Network and the Mobile Network and, because of this, two files need to be created and then joined to create the total Network file to take to the database.

In a first approach to the database, these two "data sources" (which are not exactly data sources but different and complementary extractions of the Network data source) had different data_source_id but, after further consideration, it was decided to have only one data_source_id for the totality of the Network data source, since the values present on the Fixed Network and the

Mobile Network are not studied separately, they are complementary and need to be considered as belonging to only one data source.

First, it is very important to note that the Mobile Network data cannot be all extracted from one single file since the complexity of the DW architecture does not allow to retrieve all the months simultaneously. Therefore, for this data source, one file must be created for each month. Regarding the Fixed Network data source, the DW structure allows for the extraction of all periods in one report.

The different files were created in BO reports, one file for the Fixed Network and 41 files (from January of 2019 to May of 2022) for the Mobile Network (each of these files contains only one report). All these files have three identifier columns, which identify the year and month when the traffic occurred (Year Month), the detailed segment of the customer that made the traffic (PCG Detailed Segment) and the type of traffic made (Traffic Type Aggregated). These BO reports contain some filters that cannot be disclosed due to confidentiality issues. More in regards to the filters that can be disclosed, a filter was applied to the PCG Detailed Segment so that the null values are not brought to the reports, and a filter was applied to the Traffic Type Aggregated so that the null values and all the traffic types concerning incoming calls are also not brought to the reports.

All the files have 2 more columns, the number of total real minutes (Real Airtime) and the number of calls made (Nbr. Of Calls).

Since all the files have the same structure, after exporting them into excel, the tables in them are all manually appended to form one unique table in one unique file, which still needs some manipulation.

Since the field identifying the type of traffic (Traffic Type Aggregated) does not contain the appropriate aggregations, an aggregation must be made using an excel file called "Master Traffic Type Network" to standardize the types of traffic across data sources. To proceed to such aggregations, the XLOOKUP function is used to search for the value of the column Traffic Type Aggregated in the corresponding column of the file "Master Traffic Type Network" and returns the corresponding value of the column Agregação_Uniformizada.

It is important to note that the "Master Traffic Type Network" excel file was built for the exact purpose of standardizing the aggregations made by the Fixed and Mobile Network "data sources", in order for them to have a real reading in the required context. This file contains two columns: one with the Traffic Type Aggregated and one that uniforms the aggregation (Agregação_Uniformizada).

To finish this aggregation, a pivot table must be made so that there are no duplicated identifier columns. To do this, the XLOOKUP function is once again used to build the period_id, customer_id and traffic_id columns. The period_id column is built by searching the Year Month values in the column year_month of the period table of the database and returning the period_id. The customer_id column is built by searching the PCG Detailed Segment values in the database column PCG Detailed Segment of the customer table and returning the customer_id. The traffic_id column is built by searching the values of the aggregation made to the Traffic Type Aggregated in the column Network_classification of the traffic table of the database and returning the traffic_id. After this, the pivot table is finally made, having as rows the data_source_id, the period_id, the customer_id, and the traffic_id, and as columns the values of the sum of the Real

Airtime and the sum of the Nbr. Of Calls. The pivot table is designed to have a tabular layout, repeating all item labels and disabling the subtotals and grand totals for both rows and columns so that it can be brought to the database.

The final table to bring to the database (to the Network fact table) presents six columns: data_source_id, which in the case of this table is 103 (this id can be found in the data source dimension table of the database); period_id; customer_id; traffic_id; Real Airtime; and, lastly, Nbr. Of Calls.

Lastly, when already in the database file, four other columns are added to the fact table:

1. agregation_id (built using the XLOOKUP function to look for the traffic_id in the traffic table and return the aggregation_id)

2. period_customer_traffic (which is a concatenation of the period_id, customer_id and traffic_id)

3. OffNet Minutes (built using the IFS function to check whether the aggregation_id is 1, 2 or 3. If it is 1 (off-net in all data sources), then the cell takes the value of the Real Airtime of that entry; if it is 2 (on-net in all data sources), then the cell takes the value of 0; if it is 3 (off-net in INTEC and Network and a mix of both in the CR data source), then the cell takes the value of the Real Airtime of that entry).

4. OnNet Minutes (built using the IFS function to check whether the aggregation_id is 1, 2 or 3. If it is one, then the cell takes the value of 0; if it is two, then the cell takes the value of the Real Airtime of that entry; if it is 3, then the cell takes the value of 0).

4.2.3 File from the INTEC data source

In the case of the INTEC data source, the course of action must be different. The data from this source needs to be subject to some adjustments (made to fine-tune the segments' allocation of interconnection traffic); therefore, the data used must account for these adjustments; it cannot simply be retrieved from the BO platform. Because of this, the data from this source must come from the INTEC model built in excel. Therefore, the metrics of this data source for the models built by the department cannot come from the database being constructed. It must happen on the contrary direction (the data goes from the model to the database).

Also, in this data source, the interconnection cost concept appears. As was explained in section 3.1, when a client of NOS makes a call to clients of other operators, NOS has to pay a value to that other operator. Therefore, if a call is off-net (the only type present in this data source), that same call creates both a revenue and a cost. Only the revenue appears in the CR data source, while in the INTEC data source, only the cost appears.

So, to build the INTEC fact table to bring to the database, a file is constructed in excel where all the different combinations of Year Month, PCG Detailed Segment and Traffic Type Aggregated can be found in the entries (in the identifier columns). Then, the values of the Real Airtime, the Cost and the Nbr. Of Calls are filled by using the XLOOKUP function to search for the customer segments in the correct period and in the correct sheet of the INTEC model files (each detailed type of traffic has a different sheet). There are different files for each metric

(airtime, value and calls) and each traffic type (basic traffic and special services). The special services traffic is in the special services files, while the international, national off-net – fixed, and national off-net – mobile traffic can be found in the different basic traffic files (remember that the INTEC data source only has off-net traffic).

Then, the XLOOKUP function is once again used to build the period_id, customer_id and traffic_id columns. The period_id column is built by searching the Year Month values in the column year_month of the period table of the database and returning the period_id. The customer_id column is built by searching the PCG Detailed Segment values in the database column PCG Detailed Segment of the customer table and returning the customer_id. The traffic_id column is built by searching the values of the Traffic Type Aggregated in the column INTEC_classification of the traffic table of the database and returning the traffic_id.

The final table to bring to the database (to the INTEC fact table) presents six columns: data_source_id, which in the case of this table is 102 (this id can be found in the data source table of the database); period_id; customer_id; traffic_id; Real Airtime; Cost; and, lastly, Nbr. Of Calls.

Lastly, when already in the database file, three other columns are added to the fact table:

1. agregation_id (built using the XLOOKUP function to look for the traffic_id in the traffic table and return the aggregation_id)

2. period_customer_traffic (which is a concatenation of the period_id, customer_id and traffic_id to help fill the columns of OffNet Minutes and OnNet Minutes in the CR data source)

3. OffNet Minutes (which, in the case of this data source, presents the values of the Real Airtime column).

4. OnNet Minutes (which in the case of this data source always presents the value of zero since it only has off-net traffic).

4.3 Integrity Check

Several approaches to the voice traffic's real minutes integrity check were taken, and, in the first ones, this process was made on an excel file apart from the database.

These first approaches to the process were made so that in the columns, there were the possible year month values, and in the lines, there were the types of traffic under study, separated by data source (in each cell, there was the number of real minutes made in the correspondent year month for the correspondent type of traffic in that source), and the differences between those values (both in absolute and percentage forms).

In what concerns the data sources, as was stated before in section 4.1, in all the approaches to the integrity check, it was assumed that the INTEC data source always has reliable values in what regards off-net traffic. Again, this has a relevant impact in what concerns the special services traffic of the CR data source because this data source aggregates both on-net and off-net destinations and, therefore, the special services' off-net traffic of the CR data source is equalized

to the one from the INTEC data source and then, the special services' on-net traffic of the CR source is calculated by subtracting this off-net to the total traffic.

In the first integrity check, the types of traffic were mainly studied in the way they were extracted from the BO platform, according to the information structure of each data source. The types of traffic compared between the three data sources were: International, OnNet (aggregating fixed and mobile destination's technology given the knowledge that the CR data source does not distinguish well between fixed and mobile destination's technology), OffNet - Fixed, OffNet - Mobile, NCurtos 16x, NCurtos 18x, NCurtos Outros, NNG 707x, NNG 760x, NNG 761x, NNG 800x, NNG 808x, and NNG 882x. In the case of the special services, it is important to mention that for the Network data source, the off-net and on-net were aggregated; for example, the OffNet - NCurtos 16x was aggregated with the OnNet - NCurtos 16x to study the traffic of the NCurtos 16x.

Given the difficulty of the CR data source in distinguishing between on-net and off-net traffic, a new approach to the integrity check was taken. It was decided to ignore this distinction between on-net and off-net traffic (that is, aggregating these two types of traffic) and study the minutes based on their destination's technology (mobile or fixed). Since the special services are not divided into mobile or fixed technology, the traffic is studied based on the lines (NNG 707x, NNG 800x, etc.). In both the CR and Network data sources: the OffNet - Fixed traffic was aggregated with the OnNet - Fixed traffic to study the Basic - Fixed traffic, and the OffNet - Mobile traffic was aggregated with the OnNet - Mobile traffic to study the Basic - Mobile traffic. The same was done to the special services in the Network data source, just like in the previous approach. In the CR data source, concerning the special services, they were already retrieved from the BO containing both the OnNet and OffNet traffic aggregated. The types of traffic compared between the three data sources were: International, Basic - Fixed, Basic - Mobile, NCurtos 16x, NCurtos 18x, NCurtos Outros, NNG 707x, NNG 760x, NNG 761x, NNG 800x, NNG 808x, and NNG 882x.

However, this approach for the integrity check also did not work since the minutes of the three data sources were not comparable. While the CR and Network data sources had both off-net and on-net traffic, the INTEC data source had only off-net traffic and, therefore, the integrity check could not be made since no conclusions could come from it.

Again, given the difficulty of the CR data source in distinguishing between on-net and off-net traffic and adding the need to be able to compare the different sources, a new approach to the integrity check was taken. The types of traffic compared between the three data sources were now: International, Basic - Fixed, Basic - Mobile, NCurtos 16x, NCurtos 18x, NCurtos Outros, NNG 707x, NNG 760x, NNG 761x, NNG 800x, NNG 808x, and NNG 882x. The difference was that, in this approach, the distinction between on-net and off-net traffic was not ignored. For all these types of traffic (except for the International, because it does not have OnNet), the following was done: The Off-Net traffic was compared for the INTEC and Network data sources. The Off-Net traffic of the CR data source was equalled to the one from the INTEC data source, and then the On-net traffic of the CR data source was calculated by subtracting this Off-Net from the total traffic. Lastly, the CR and Network data sources were compared using the total, off-net and on-net traffic. This equalization of the off-net of the CR data source to the one from the INTEC data source was made, not only to the special services but also to the

basic traffic (except the international traffic). Since the off-net traffic of the CR data source is the one from the INTEC, these two sources were not compared (they were only compared in the international traffic). Also, because of this, the off-net traffic from the Network is only compared to the traffic from INTEC.

In this case, as it is believed that the values from the INTEC data source are trustworthy, if the Network (off-net) is aligned with the INTEC, then the Network is also trustworthy (this was the comparison made first). Therefore, when comparing the CR and Network, if they are not aligned, the problem resides in the CR data source.

As the OffNet Minutes of the CR data source are equaled to the OffNet Minutes of the INTEC data source, in the special services traffic, the values of the column CR-INTEC should be all zero. However, in reality, this did not happen. When exploring the visuals of this table in Power BI, it was discovered that there were some significant differences between the OffNet Minutes of the CR data source and the OffNet Minutes of the INTEC data source in the Special Services.

This happens because there may be traffic missing from the CR data source (which should not happen since, if the traffic actually occurs, it needs to be read by all the data sources). This probably happens due to what was explained in section 3.2. When the traffic presents no charge to the clients, it may not be fully accounted for in the CR data source, and it might not even exist in this source. When building the OffNet Minutes column in the CR fact table, for the special services traffic, a search of the key period_customer_traffic is done on the INTEC fact table, and the corresponding minutes are returned. Since there is a search for the traffic present in the CR data source, the non-present traffic (which is present in the INTEC data source) does not appear in this built column, leading to significant differences (especially in the line NCurtos 18x).

Because of this relevant issue, a new table was built in the database. This table contains only the Special Services Traffic and has a column (OffNet Minutes) that makes, for each traffic type and year month, the subtraction between the corresponding OffNet Minutes of the INTEC data source and the OffNet Minutes of the CR data source (the ones that appear when building the column).

This new table CR_missing_according_to_INTEC contains 7 columns: data_source_id (which will be 101 –> CR data source); period_id (of each year month); customer_id (will be always zero because it does not matter for the integrity check); traffic_id; aggregation_id; period_customer_traffic; OffNet Minutes (built by making subtractions of SUMIFS functions).

At this point of the integrity check, the Network data source was believed to be more reliable than the CR data source, despite the department's initial beliefs. This is because the Network data source seems to be more aligned with the INTEC data source than the CR source (this can be found in section 6.2).

Despite this integrity check already allowing us to make some conclusions about the data sources, it was arduous to do it, given that we could only see the differences' values and not their evolution, trends or averages. These integrity checks made (on a different excel file) did not allow us to graphically visualize the differences between the sources, which is not very intuitive. Therefore, it became imperative to execute the integrity check in Power BI in order to be able to compare the data of the different fact tables while allowing to visualize these differences, making it easier to draw conclusions.

For the integrity check, the OffNet Minutes and OnNet Minutes columns of all the fact tables were used to make the necessary distinction between on-net and off-net traffic. In it, it is possible to study (monthly, from January 2019 to May 2022) the following aggregated traffic types: Total Basic Traffic, International, Basic - Fixed, Basic - Mobile, Total Special Services Traffic, Total NCurtos Traffic, NCurtos 16x, NCurtos Outros (18x + Outros), NCurtos 18x, NCurtos Outros, Total NNG Traffic, NNG 707x, NNG 760x, NNG 761x, NNG 800x, NNG 808x, and NNG 882x.

In this final integrity check, the CR data source's off-net minutes of the basic traffic (Total Basic Traffic, International, Basic - Fixed and Basic - Mobile) were not equalled to the minutes present in the INTEC data source for these types of traffic. This was only done to the special services traffic.

A summary table of the different approaches can be found in appendix A.1

To conduct the integrity study of the data sources, the appropriate traffic aggregations needed to be made (they can be seen in figure 4.4). Also, measures were constructed in Power BI. These measures calculate the differences (in absolute and in percentage) between the different data sources, in what concerns off-net minutes, on-net minutes and total minutes. These measures can be found below.

1. CR-INTEC (Abs) calculates the absolute difference between the OffNet Minutes of the CR data source (plus the OffNet Minutes from the table that contains the minutes missing from the CR data source, according to the INTEC source) and the Minutes of the INTEC data source

```
CR-INTEC (Abs) = CALCULATE(SUM(CR_fact_table[OffNet Minutes])+SUM(CR_missing_according_to_INTEC[OffNet Minutes])
-SUM(INTEC_fact_table[OffNet Minutes]))
```

Figure 4.9: CR-INTEC (Abs) measure

2. CR-INTEC (%) is calculated by dividing the absolute difference by the Minutes of the INTEC data source

```
CR-INTEC (%) = CALCULATE(((SUM(CR_fact_table[OffNet Minutes])+SUM(CR_missing_according_to_INTEC[OffNet Minutes])
-SUM(INTEC_fact_table[OffNet Minutes]))/SUM(INTEC_fact_table[OffNet Minutes]))*100)
```

Figure 4.10: CR-INTEC (%) measure

3. Network-CR OffNet (Abs) calculates the absolute difference between the OffNet Minutes of the Network data source and the OffNet Minutes of the CR data source (plus the OffNet Minutes from the table that contains the minutes missing from the CR data source, according to the INTEC source)

```
Network-CR OffNet (Abs) = CALCULATE((SUM(Network_fact_table[OffNet Minutes]))
-(SUM(CR_fact_table[OffNet Minutes])+SUM(CR_missing_according_to_INTEC[OffNet Minutes])))
```

Figure 4.11: Network-CR OffNet (Abs) measure

4. Network-CR OffNet (%) calculated by dividing the absolute difference by the OffNet Minutes of the CR data source (plus the OffNet Minutes from the table that contains the minutes missing from the CR data source, according to the INTEC source)

```
Network-CR OffNet (%) = CALCULATE((((SUM(Network_fact_table[OffNet Minutes])-(SUM(CR_fact_table[OffNet Minutes])
+SUM(CR_missing_according_to_INTEC[OffNet Minutes]))))/(SUM(CR_fact_table[OffNet Minutes])
+SUM(CR_missing_according_to_INTEC[OffNet Minutes]))*100)))
```

Figure 4.12: Network-CR OffNet (%) measure

5. Network-CR OnNet (Abs) calculates the absolute difference between the OnNet Minutes of the Network data source and the OnNet Minutes of the CR data source

```
Network-CR OnNet (Abs) = CALCULATE((SUM(Network_fact_table[OnNet Minutes]))-(SUM(CR_fact_table[OnNet Minutes])))
```

Figure 4.13: Network-CR OnNet (Abs) measure

6. Network-CR OnNet (%) calculated by dividing the absolute difference by the OnNet Minutes of the CR data source

```
Network-CR OnNet (%) = CALCULATE((((SUM(Network_fact_table[OnNet Minutes])
-SUM(CR_fact_table[OnNet Minutes]))/SUM(CR_fact_table[OnNet Minutes]))*100))
```

Figure 4.14: Network-CR OnNet (%) measure

7. Network-INTEC (Abs) calculates the absolute difference between the OffNet Minutes of the Network data source and the Minutes of the INTEC data source

```
Network-INTEC (Abs) = CALCULATE((SUM(Network_fact_table[OffNet Minutes]))-(SUM(INTEC_fact_table[OffNet Minutes])))
```

Figure 4.15: Network-INTEC (Abs) measure

8. Network-INTEC (%) calculated by dividing the absolute difference by the Minutes of the INTEC data source

```
Network-INTEC (%) = CALCULATE((((SUM(Network_fact_table[OffNet Minutes])
-SUM(INTEC_fact_table[OffNet Minutes]))/SUM(INTEC_fact_table[OffNet Minutes]))*100))
```

Figure 4.16: Network-INTEC (%) measure

9. Network-CR Total (Abs) calculates the absolute difference between the Total Minutes of the Network data source and the Total Minutes of the CR data source

```
Network-CR Total (Abs) = CALCULATE((SUM(Network_fact_table[Real Airtime]))-(SUM(CR_fact_table[Real Airtime])))
```

Figure 4.17: Network-CR Total (Abs) measure

10. Network-CR Total (%) calculated by dividing the absolute difference by the Total Minutes of the CR data source

```
Network-CR Total (%) = CALCULATE((((SUM(Network_fact_table[Real Airtime])
-SUM(CR_fact_table[Real Airtime]))/SUM(CR_fact_table[Real Airtime]))*100))
```

Figure 4.18: Network-CR Total (%) measure

At this point, it is very important to understand that the columns named ”OnNet Minutes” and ”OffNet Minutes” in the different fact tables built are not needed in the dashboard’s Part I, which is the one with the study of the traffic generating revenues (since, again, the final choice was to not study between on-net and off-net), but are needed to build the integrity check.

4.4 Power BI Preparation

After loading the data (the CR fact table, the Network fact table, the INTEC fact table, the CR missing according to INTEC fact table, and the three dimension tables – traffic, period and customer) into Power BI, it needs to be transformed. It is important to eliminate all the blank rows and columns that may be in the tables (coming from excel) so that there is no space being unnecessarily occupied.

After this, it is necessary to check the relationships between the different tables. Each fact table must have three relationships (one with each dimension table, through the ids). If the Power BI software does not recognize these relationships automatically, they need to be inserted manually.

By choosing the Model view, the relationships can be studied.

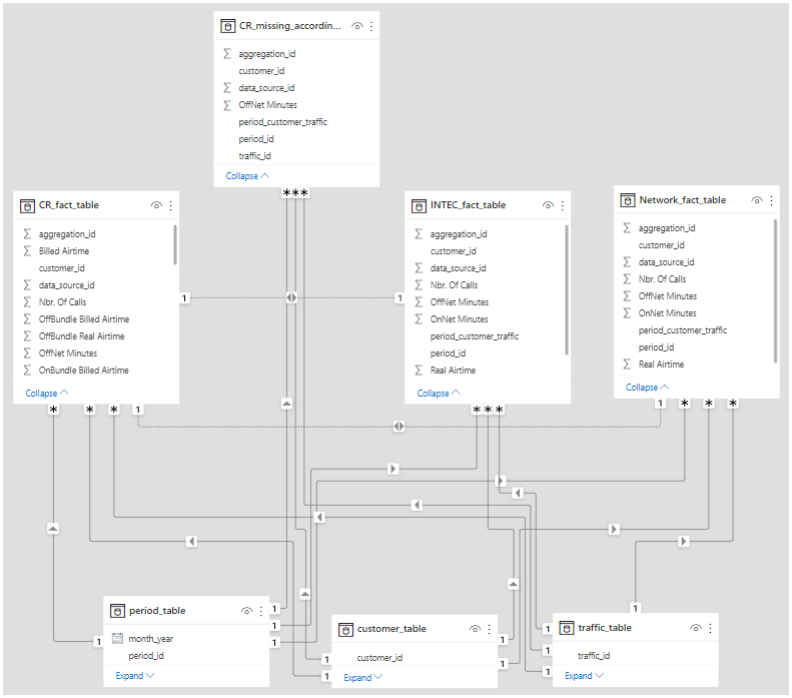

Figure 4.19: Model View of Power BI

If any relationship is missing from the model, it must be inserted manually since the software did not recognize them. This can also be confirmed in the "Manage Relationships" option in the modelling panel. It is also in this functionality that the missing relationships will be added. For example, if it is necessary to create a relationship between the CR fact table and the period table, the following must be made:

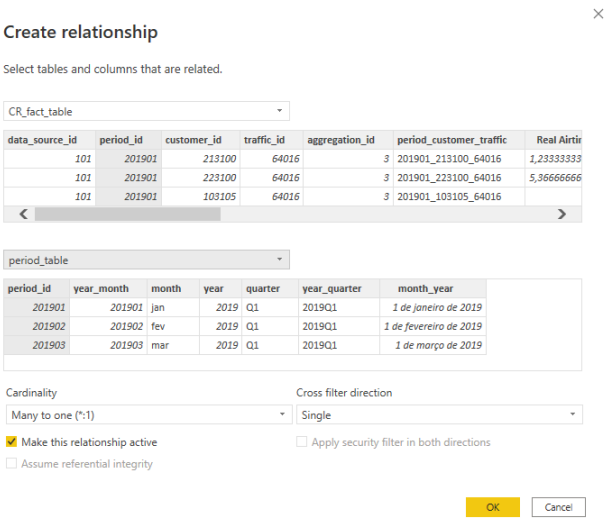

Figure 4.20: Creating a relationship between the CR fact table and the period dimension table

The table to put first is the fact table (and the correct column must be chosen) and the table

to put second is the dimension table (and the correct column must be chosen). The cardinality of the relationships should always be ”Many to one (*:1)”, implying that the id appears many times in the fact table but only once in the dimension table, to avoid errors in the model. The cross filter direction should always be ”Single”.

After creating the necessary new relationships, there should be twelve active relationships (between the four fact tables and the three dimension tables). After this, the dashboard implementation can begin.

4.5 Dashboard's Structure

The first and most important notion when building the dashboard was to make it the most flexible possible while ensuring its simplicity. This is because the end-user must be able to choose what they want to see and study in the most direct and intuitive way possible.

As said before, it was decided to bring only the CR data source to the Power BI dashboard to study the voice traffic generating revenues. Also, it was later decided to visualize the integrity check in the dashboard. Therefore, the dashboard needed to be separated into two parts: one to study the voice traffic generating revenues and the other to visualize the integrity check by comparing the different data sources.

4.6 Dashboard's Introduction

First, an introduction page was created, where the structure and contents of the dashboard can be found.

On the left side of the introduction page, there are the contents of the first part of the dashboard (study of the voice traffic generating revenues) and some other information, such as the metrics used, along with their meaning, unit of measure and form of calculation (when applicable). Also, the revenues' metrics can be studied by choosing the traffic type and customer segment one desires to see. If none is chosen, these are all aggregated together. Therefore, the traffic and customer hierarchies through which one can navigate are shown so that the user can see exactly what they can choose to see.

On the right side of the introduction page are the contents of the second part of the dashboard (integrity check) and other information that makes the different comparisons easier to understand. Also, the traffic comparisons to be visualized can be chosen by the user; therefore, the traffic hierarchy is shown. Remember that to perform the Integrity Check, there is no possibility to drill down into the customer segment hierarchy, as it is studied as a whole and, thus, the customer hierarchy is not present on this side of the page.

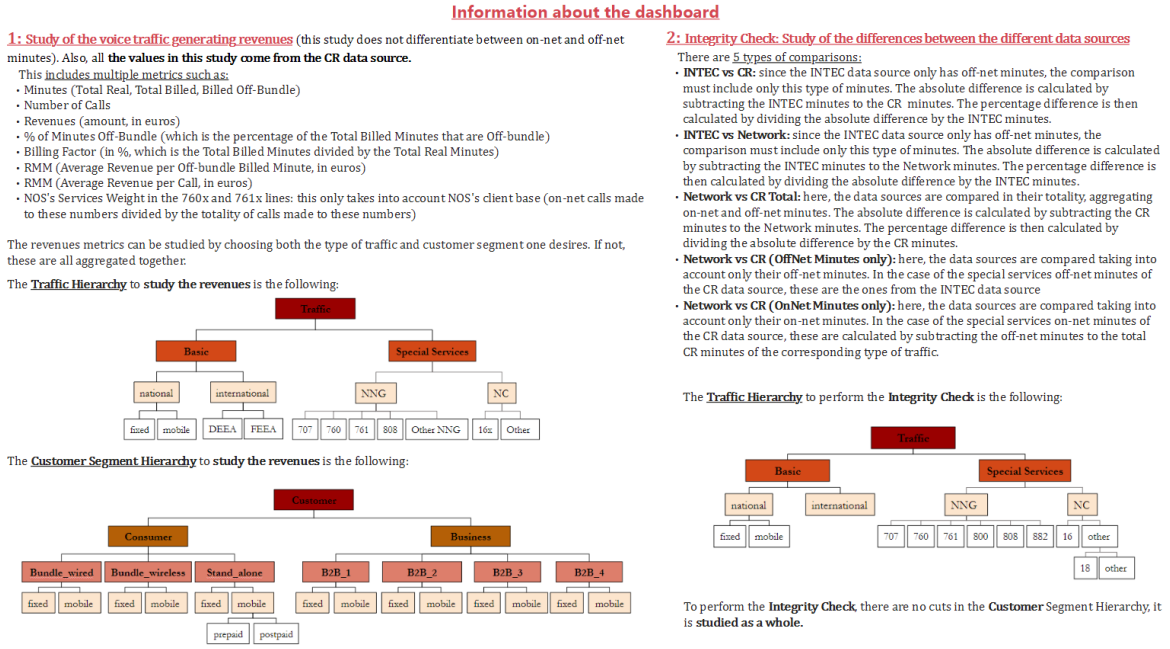

Figure 4.21: Introduction page of the dashboard

One important thing to note is that the traffic hierarchy to study the revenues is different from the traffic hierarchy to do the integrity check, which is more disaggregated. This is because the integrity check was made so that the sources are compared in their most detailed way (to see where the differences actually lie), and, by doing this, some conclusions were drawn that influenced the decision to study the revenues in a more aggregated way. In the special case of the international traffic, it is not divided into DEEA and FEEA in the integrity check only because both the INTEC and Network data sources do not have this ”built-in” split and therefore, to compare the sources, the international traffic needs to be compared as a whole.

One last relevant note is that there are some filters applied to the totality of the dashboard; that is, there are filters applied to all the pages simultaneously.

The first filter is applied to the customer_segment field of the customer table so that only the Business and the Consumer segments appear.

The second filter is applied to the month_year field of the period table so that only the values of the current month (the most recent one, which in this case is May of 2022) and of the months previous to that one appear. This filter will need to be updated every month, when the values of all the data sources are updated in the database.

The last filter is applied to the traffic_type field of the traffic table so that only the Basic and Special Services traffic appears.

Figure 4.22: Filters applied to all the pages of the dashboard simultaneously

Chapter 5

Part I: Study of the voice traffic generating revenues

This part of the dashboard contains seven pages: Y2D, Real Airtime's Evolution, Revenues' Evolution, Other Important Metrics, 76x's Evolution, Zoom in the Consumer Segment, and Airtime vs Revenues.

The literature research stated that dashboards should have everything on one screen (one page), which was said by Few (2006). However, this is not always the case (as stated by Shaffer (2018)).

In the particular case of the dashboard being designed, it was not desirable to have all the visualizations on one screen. It is preferable to have multiple pages to study specific metrics, types of traffic or customer segments (which have a specific behaviour and require special attention). One of the goals of the dashboard was for it to be the most informative possible, which required some detail in the visualizations and, therefore, increased the number of plots to build, making it not advisable to have everything on one screen (it would be too confusing and not viable).

Additionally, the findings were very important to make decisions about what to bring to the database and, consequently, to the dashboard.

Lastly, it is important to note that due to confidentiality issues, the plots shown and the findings explained cannot be completely disclosed. Therefore, the plots' values will not be shown and the findings will be kept at a more aggregated level.

5.1 Year to Date - Y2D

5.1.1 Construction

This page has a clustered column chart that offers a wide view of the evolution of the real minutes over the four years under study (2019 to 2022) by type of traffic.

To build this clustered column chart, the field of the traffic table chosen for the x-axis values was the traffic_lvl3, which can be seen in figure 4.4. The legend chosen for the chart was the year from the month_year date hierarchy of the period table. The field chosen for the y-axis was the Real Airtime from the CR fact table.

Here, the values of the minutes of each year comprise the months from January to the month under study, all aggregated together. This allows us to check how the year is going so far as a whole when comparing with the same period of the last three years.

The page also has two slicers. The first one presents the hierarchy of the quarters and months of the month_year hierarchy of the period table. It allows the users to choose the quarters and months they want to study. This slicer offers flexibility because, by having this slicer, the user can individually choose the months from January to the month under study (to actually study the year to date), or they can choose other months/quarters that they might want to study.

The second slicer allows the user to choose only the type(s) of traffic they want to see. This slicer is a traffic hierarchy, which is built by using the fields traffic_type, traffic_lvl2 and traffic_lvl3 (in this order) from the traffic table (figure 4.4). When no traffic type is chosen, they all appear in separate blocks of clustered columns. When all the types of traffic appear, some of them lose visibility because of the difference in the scales of the different types of traffic. For example, the Basic - Fixed traffic has a very high amount of minutes compared to the NCurtos Outros; therefore, this last traffic seems always to have null values, which is not true. One possible solution for this problem (besides giving the ability to choose the traffic) was to change the scale of the y-axis (which shows the amount of minutes) to a logarithmic type. This solution was tested, but it led to the loss of visibility of the real scale in terms of minutes of the different traffic types, which are relevant to preserve in the visualization.

Additionally, it was decided that there was a need to be able to see the percentage differences between the different years to see how much the minutes increased/decreased from one year to the other (when considering the same period). To do this, four new quick measures were built. In these measures, the quick calculation chosen was "Percentage difference from filtered value", the base value chosen was "Sum of Real Airtime", and the filter chosen was the year from the month_year hierarchy (in one measure, 2019 was chosen, in the other 2020 was chosen, in the other 2021 was chosen, and in the last 2022 was chosen).

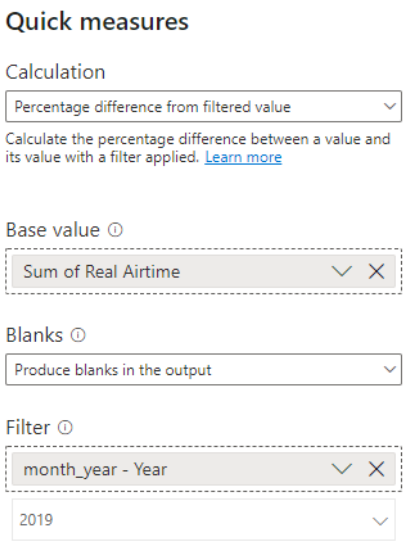

Figure 5.1: Quick measure to calculate the percentage differences between the different years and 2019

These quick measures were put as tooltips. When the user hovers over the column of one specific year, the percentage differences between that year's minutes and the other three years' minutes appear.

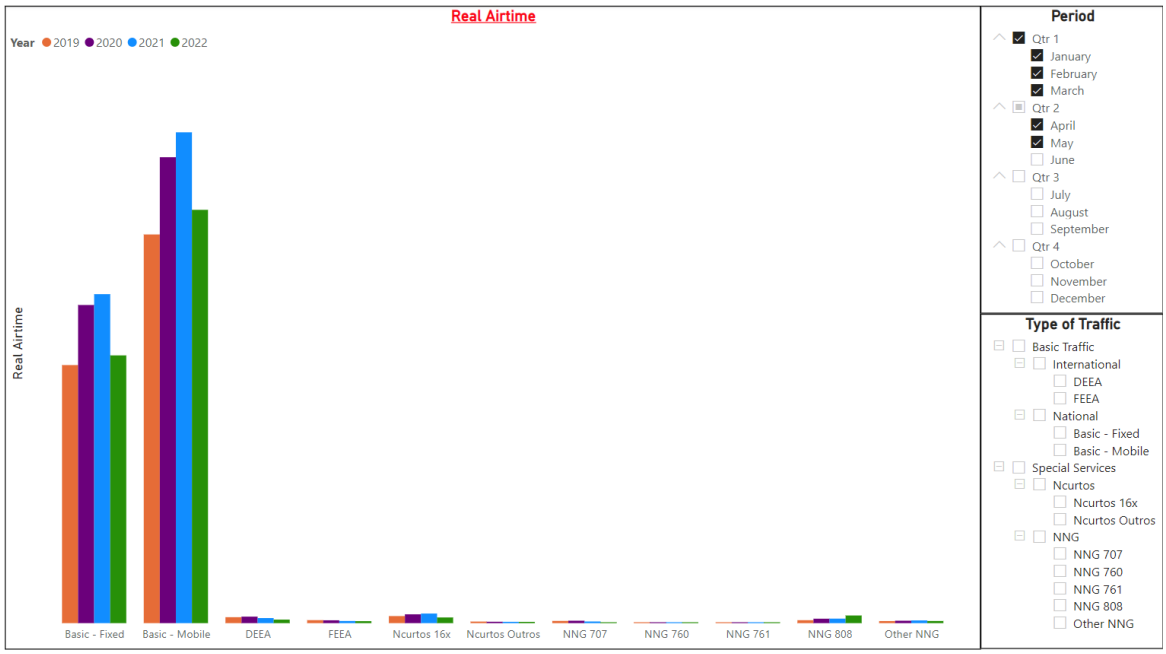

Figure 5.2: Year to Date - Y2D page

5.1.2 Findings

Clearly, the Basic - Fixed and the Basic - Mobile traffic are the dominant ones in what concerns the real minutes made by NOS's clients. Also, for the same months, the minutes of this traffic increased from 2019 to 2020 and from 2020 to 2021 (due to the pandemic). Then, from 2021 to 2022, the minutes decreased but are still above the 2019 values.

Considering now the International traffic, the FEEA traffic has been decreasing (in minutes) since 2019, while the DEEA traffic increased from 2019 to 2020 and then also started to decrease, being now below the 2019 values. Also, the DEEA traffic is always above the FEEA traffic; that is, NOS's clients make more minutes to destinations inside the European Economic Area than to destinations outside of it.

Now, starting the special services study with the NCurtos traffic. For the same months, the NCurtos 16x traffic increased in minutes from 2019 to 2020 and from 2020 to 2021 (due to the pandemic), decreasing again in 2022, being now below the 2019 values. For the NCurtos Outros traffic, this one decreased from 2019 to 2020 and from 2020 to 2021, increasing in 2022 (but still being below the 2019 values). Also, it is important to note that NOS's clients make more minutes to 16x lines (lines dedicated to customer service support) than to other lines of the NCurtos traffic.

Regarding the NNG lines, the NNG 808x is the line with the highest amount of minutes, which has been increasing since 2019. This was expected since the National Health Service is part of this line. Now, for the NNG 707x, this line increased in minutes from 2019 to 2020 and then started to decrease due to additional regulations specific to the lines dedicated to B2C customer service support, being now much lower than the 2019 values. The Other NNG lines increased from 2019 to 2020 and then again in 2021, decreasing in 2022 (while still being above the 2019 values).

Concerning the NNG 760x and NNG 761x, it was mentioned earlier that for the study of these lines, it is more relevant to use the number of calls made since the charging of these lines is made by call and not by minutes. Despite this, for a first and more general study of the overall traffic made by NOS's clients, these lines were still added to this page studying the minutes. Later on, in their separate page, the more appropriated metrics can be studied more deeply. The chart shows that the line NNG 760x has been decreasing since 2019, following a downward trend, in the sequence of its substitution by the NNG 761x for the past few years. For the NNG 761x, the conclusion is very different. This line's minutes increased in 2020 and 2021, decreasing in 2022 but still exceeding the 2019 and 2020 values.

5.2 Real Airtime and Revenues' Evolution

5.2.1 Construction

On the dashboard's Real Airtime's Evolution page, five line charts can be found along with two slicers.

All the line charts on this page are of Real Minutes; to build them, the values chosen for the y-axis are the ones of the Real Airtime field of the CR fact table. Also, they all have as the x-axis

the month_year field of the period table. It is important to note that the axis' values need to be the ones of the month_year field and not the date hierarchy.

In the top center of the page, a more general chart can be found, which shows the evolution of the total real minutes of the Basic Traffic and Special Services Traffic. For this chart, the legend chosen was the traffic_type field of the traffic table (which can be seen in figure 4.4).

Then, there are four charts at the bottom of the page. The first two regard the basic traffic in a more detailed way, while the last two regard the special services traffic.

The first bottom chart shows the evolution of the National traffic (including Basic - Fixed and Basic - Mobile) and the total International traffic (including both DEEA and FEEA destinations). Therefore, this chart uses the traffic_destination field of the traffic table as legend. Also, it has a filter on the traffic_type field to show only the Basic Traffic.

The second bottom chart shows the evolution of the International traffic in a more detailed way, by having two lines: one for the DEEA traffic and one for the FEEA traffic. Therefore, this chart uses the destination_nationality field of the traffic table as legend. Also, it has a filter on the traffic_destination field to show only the International.

The third bottom chart shows the evolution of the Special Services traffic separated into NCurtos (containing the lines 16x and Outros) and NNG (containing the lines 707x, 808x, 760x, 761x, and Other NNG - that contains the NNG 800x and NNG 882x). Therefore, this chart uses the destination_type_of_number field of the traffic table as legend. Also, it has a filter on the traffic_type field to show only the Special Services.

The fourth (and last) bottom chart shows the evolution of the NNG traffic in a more detailed way by having three lines: one for the NNG 707x traffic, another for the NNG 808x traffic and the last for the Other NNG. Therefore, this chart uses the destination_line field of the traffic table as legend. Also, it has a filter on the destination_line field to show only the NNG 707x, the NNG 808x and the Other NNG.

Here, it is important to understand that the NNG 760x and NNG 761x lines present a different behaviour than the other special services lines because these are value-added service phone lines dedicated to contests and polls and charged a universal tariff by call. Hence, these two lines need to be studied using the number of calls (instead of the number of minutes) and the average revenue per call (instead of the average revenue per minute). Therefore, the NNG 760x and the NNG 761x is not included in the fourth bottom chart and will have their own page, where the convenient metrics can be studied.

Lastly, there are two slicers.

The first has the traffic hierarchy, which is built by using the fields traffic_type, traffic_lvl2 and traffic_lvl3 (in this order) from the traffic table (figure 4.4). When no traffic is chosen, all the traffic types appear in the correspondent charts (aggregated or not, depending on what the chart shows). By choosing a specific type(s) of traffic in this slicer, the user can focus on what they want to study.

The second slicer presents the customer's segment hierarchy, which is built by using the fields customer_segment, customer_subsegment, customer_technology, customer_subscription and PCG Detailed Segment (in this order) from the customer table (all but the last can be found in figure 4.3). When no customer segment is chosen, they all appear aggregated. By choosing specific segments, subsegments, etc., the user can visualize what they want to study.

By having both these slicers, the user can study the minutes of the specific types of traffic made by specific customer segments.

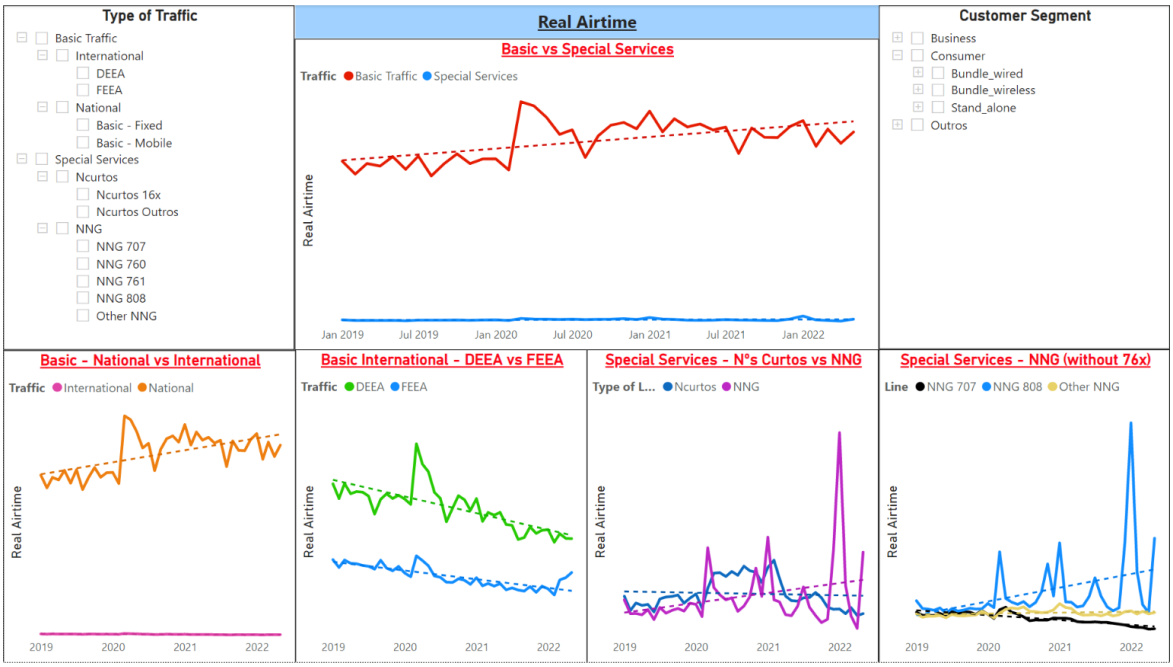

Figure 5.3: Real Airtime's Evolution page

The Revenues' Evolution page of the dashboard is built the same way as the Real Airtime's Evolution page. The difference lies in the values chosen for all the line charts. On the Real Airtime's Evolution page, the values chosen are the ones of the Real Airtime field of the CR fact table while in this Revenues' Evolution page the values chosen are the ones of the Revenues field of the CR fact table.

Therefore, this page presents five line charts (Basic vs Special Services; National vs International; DEEA vs FEEA; NCurtos vs NNG; and lastly, the detailed NNG) and two slicers (type of traffic and customer segment), just like the last page.

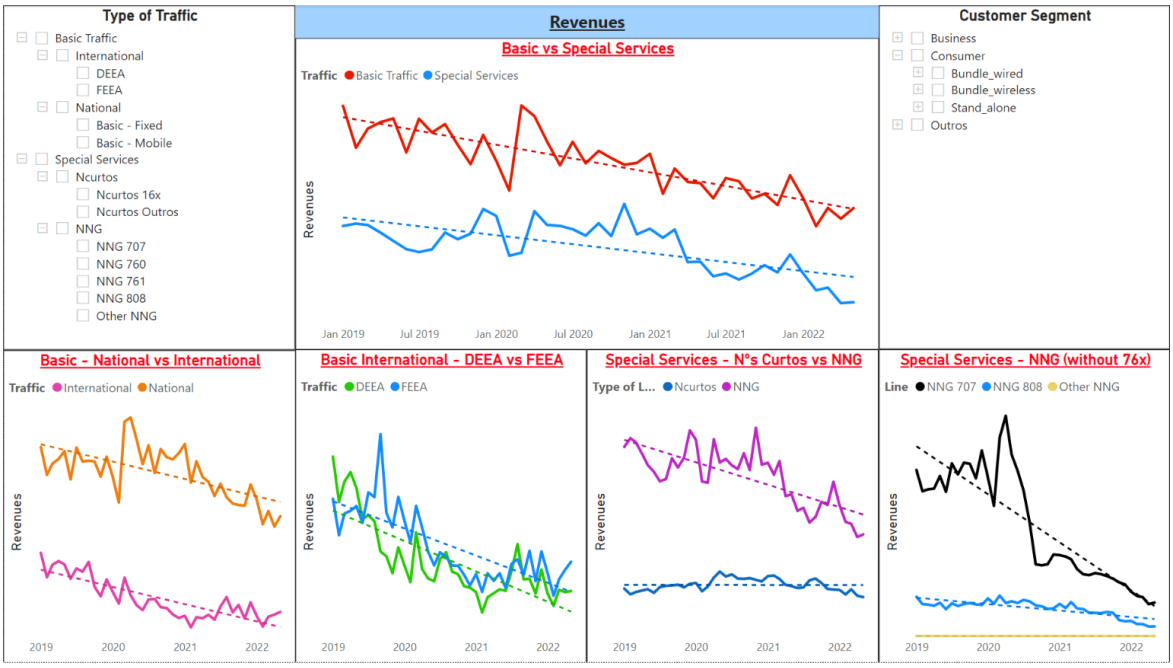

Figure 5.4: Revenues' Evolution page

5.2.2 Findings

Moving to the analysis of the Real Airtime's Evolution page and the Revenues' Evolution page, more conclusions could be drawn, considering the customer segments. In this analysis, the customers will be divided into customers with fixed technology and customers with mobile technology.

As mentioned before, the bundles of the clients with fixed technology include traffic to basic national (only to fixed destinations) and to international (only in off-peak hours and to a certain list of countries, mostly DEEA); because of this, despite the large amounts of real minutes made, these will have low revenues. Therefore, the revenues of the special services will weigh more than those from the basic traffic, especially the lines 76x, 707x and 808x.

When considering the traffic evolution of the customers with fixed technology, the basic traffic (national and international) seems to be decreasing, especially the international traffic, which has a more accentuated decline. Despite this trend, both had a sudden increase in March of 2020, an effect of the start of the pandemic in Portugal. For the special services traffic, the NCurtos traffic is higher than the NNG traffic when measured in minutes (due to more and longer calls, mainly to the customer support lines 16x). However, they are both declining. Studying further the NNG lines (without the NNG 76x), the NNG 808x is the dominant one, followed by the NNG 707x and, lastly, the Other NNG. Another important finding was that there are very visible peaks in the minutes of the total NNG minutes, which coincide with peaks in the NNG 808x. These peaks occurred in March and November of 2020, January and July of 2021, and January of 2022, which coincide with pandemic peaks. Remember that the National Health Service is part of this line.

Studying the Revenues' Evolution page for the same customers (fixed technology), the revenues of the basic national traffic are declining (and this type of traffic is the one with the most revenues, mainly due to the mobile destination, which is not included in the bundles), just as happens with the minutes. Concerning the International traffic, while the DEEA minutes are significantly above the FEEA minutes, their revenues are very close to each other because the DEEA minutes are mostly included in the bundles. Moreover, the tariff charged for the extra consumption is regulated (the last decrease occurred in May 2019) and lower than that of the FEEA traffic. Also, despite the NCurtos traffic having more minutes than the NNG traffic, in the revenues, the contrary happens. This is because the 76x lines' revenues greatly influence the NNG lines' revenues. In the NNG lines (without the 76x), the NNG 707x has the biggest revenues, followed by the NNG 808x and, lastly, the Other NNG.

For the clients with mobile technology, the basic traffic (national and international) seems to be increasing (contrary to the traffic made by clients with fixed technology). The national traffic seems to be increasing (the basic national fixed traffic has a smaller weight than the basic national mobile traffic, despite both being included in the bundles), while the international traffic (especially the FEEA) is declining (due to the increase in the use of voice, messaging and video services based on internet access, such as Whatsapp, Facebook, Facetime, etc.). For the special services traffic, the NCurtos traffic has more minutes than the NNG traffic (except in the pandemic peaks); however, they are very close. Studying the NNG lines (without the 76x) further, the NNG 808x is the dominant one, followed by the Other NNG and, lastly, the NNG 707x. The pandemic peaks mentioned in the fixed technology clients are even more visible in the mobile technology clients.

In what concerns the Revenues' Evolution page for the same customers (mobile technology), on the contrary to what happens to the minutes, in which the basic traffic is much higher than the special services traffic, the revenues of the special services traffic are very close to the ones of the basic traffic, sometimes even being above it (this is because the customers with mobile technology have basic national traffic included in their bundles, and not special services traffic). Again, despite the minutes of the basic national traffic being much higher than the international traffic, their revenues are closer together (given that international traffic is not included in the bundles). Regarding the international traffic in a more detailed way it is the same, the DEEA has more minutes than the FEEA but the revenues of the FEEA traffic are very close (and mostly above) the ones of the DEEA traffic. This is because the tariff charged to the extra consumption of DEEA traffic is regulated (last decrease occurred on may 2019) and therefore lower to that of the FEEA traffic. Now regarding the special services traffic, the total NNG traffic has more revenues than the NCurtos traffic. The NNG 707x is the one with the highest revenues, followed by the NNG 808x and, lastly, the Other NNG.

5.3 Other Important Metrics

5.3.1 Construction

On this page of the dashboard, other metrics that are considered important to the study of voice traffic revenues can be found, such as the billed off-bundle minutes, the billed minutes, the

real minutes, the percentage of minutes off-bundle, the billing factor, and the average revenue per minute (RMM).

The way to calculate the metrics (the percentage of minutes off-bundle, the billing factor, and the average revenue per minute (RMM)) was already mentioned at the end of section 4.2.1.

This page contains four line charts and two slicers.

The two slicers are the ones already present on the last two pages of the dashboard: one regarding the traffic hierarchy, which allows for the user to choose the type of traffic to study; and the other regarding the customer hierarchy, which allows for the user to choose the customers' segments to study. With this, the user can study the mentioned metrics for the specific types of traffic made by specific customer segments.

On this page, the slicer of the type of traffic is especially important. This is because in the three bottom line charts, there is a line for each existing type of traffic. This means that, when no traffic is chosen, all the lines appear in the charts and, therefore, it becomes hard to study the metrics for the different types of traffic. By choosing the type of traffic, it is easier to see and study the evolution of the desired metric for that specific traffic.

Before explaining the different line charts, it is important to note that all of them have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy.

The top center line chart shows three lines: one with the evolution of the Real minutes, the other with the evolution of the Billed minutes, and the last one with the evolution of the Off-Bundle Billed minutes. To build this chart, the y-axis was built with the fields Billed Airtime, Real Airtime, and OffBundle Billed Airtime of the CR fact table. In each line of this chart, all the traffic types and customer segments are aggregated together. To study a specific traffic and/or customer segment, they need to be chosen in the provided slicers.

The three bottom charts all have as legend the traffic_lvl3 field from the traffic table, which can be found in figure 4.4. For the y-axis, the first bottom chart has the % Off-Bundle metric, the second has the Billing Factor (%) metric, and the last has the RMM metric. Remember that all the metrics were built in the Power BI software (and are now part of the CR fact table). Also, the traffic_lvl3 field was filtered in all the charts so that it does not show blank values.

As said before, the NNG 760x and NNG 761x lines present a different behaviour, and because of that, these two lines need to be studied using different metrics and will have their own page where the convenient metrics can be studied. Because of this, there is a filter on this page. This filter filters these two lines out using the destination_line field of the traffic table.

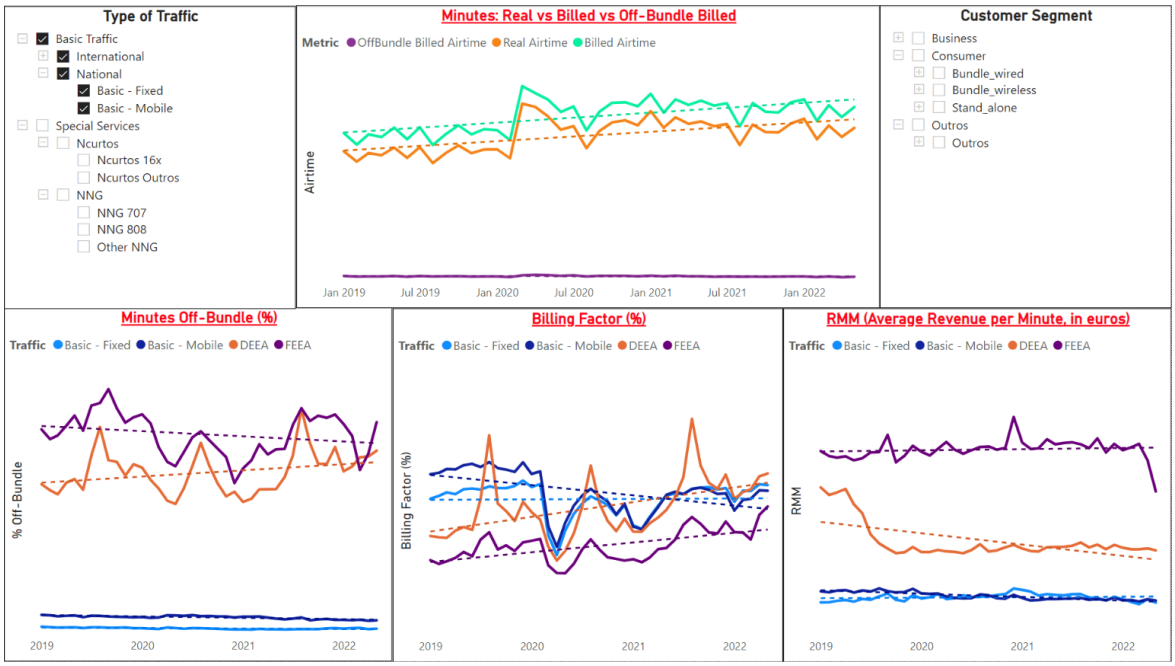

Figure 5.5: Other Important Metrics page (showing only the Basic Traffic)

5.3.2 Findings

Starting with the minutes chart, it can be seen that when no type of traffic and no customer segment are chosen, the billed airtime is always above the real airtime (the main contribution to this comes from the customers of mobile technology, where the charging is traditionally based on a higher billing factor). The off-bundle billed minutes are even lower than the real minutes (due to the increase of the minutes included in the bundles, the off-bundle minutes represent a small part of the total real minutes). This means that, when all the traffic and customers are aggregated together, the billing factor is always above 100%, while the percentage of minutes off-bundle is very low (closer to zero).

Studying the types of traffic in more detail and starting with the international traffic, the FEEA traffic has, as expected, a higher percentage of off-bundle minutes than the DEEA traffic and this is because it is more likely that the bundles contain minutes to destinations inside the European Economic Area than to outside of it. The contrary happens in what regards the billing factor, which is always higher for the DEEA traffic than for the FEEA. This is because in the DEEA traffic, the clients with fixed technology, who have a lower billing factor, have a higher weight than the clients with mobile technology. Despite this, the average revenue per minute of the FEEA traffic is higher (because the DEEA maximum tariff is regulated at 0.19€/min). Also, in the RMM chart, it is very visible when the tariffs for calls made to destinations inside the EEA dropped, which happened in the second and third quarters of 2019.

In this analysis, it was discovered that, for a specific detailed segment in the FEEA traffic, the percentage of minutes off-bundle had a sudden drop in May of 2021 and continued to decrease over the following months. This was not expected since it was thought that this percentage

should be around 100%. A deeper study was made by checking the total billed minutes and the off-bundle billed minutes of the detailed descriptions of the FEEA traffic type. It was concluded that there was a detailed description in which the total billed minutes were increasing too much and the off-bundle billed minutes were not increasing in the same proportion. This detailed description belonged to unknown traffic and should not be accounted for in any traffic type. Because of this, a decision was made so that all the detailed descriptions containing "unknown" were filtered out of the BO reports (this filter was applied in the BO platform) and, therefore, were no longer brought to the database and, consequently, to the dashboard. This data issue was no possible to identify in the heavy models built in excel by the department (explained in section 3.4), given the lack of visualization.

Now, for the basic national traffic, the Basic - Mobile traffic has a higher percentage of off-bundle minutes than the Basic - Fixed traffic. Despite this, these percentages are very low (and seem to be decreasing to zero), and this is because there has been an increase in the number of clients with bundles that include Basic National minutes, therefore resulting in a progressive decrease of the off-bundle minutes. Nevertheless, it is important to remember that the bundles are different regarding the customers segment. The bundles of customers with mobile technology include a large amount of national minutes (both to fixed and mobile destinations), resulting in a very low percentage of off-bundle minutes in the basic national traffic. On the contrary, the bundles of customers with fixed technology do not include minutes to national mobile destinations (only national fixed and some international destinations), presenting a higher percentage of off-bundle mobile minutes. In what concerns the billing factor, the two traffic types are very close to each other. The average revenue per minute seems to have a downward trend for both traffic types.

Regarding the special services traffic, these calls should always have 100% of minutes off-bundle given that no bundle includes minutes to these lines. Also, the billing factor for these types of traffic has no reading due to their characteristics. For the special services (excluding the 76x lines), the only important metric (of the ones on this page of the dashboard) is the RMM, which will have different values according to the retail price of each line (section 3.1).

5.4 76x's Evolution

5.4.1 Construction

This page was built so that the lines NNG 760x and NNG 761x could be studied separately, using appropriate metrics. The calls to these lines are charged a universal tariff by call and, therefore, studying the minutes would not lead to meaningful conclusions about the evolution of this traffic. Because of this, the number of calls must be studied instead of the number of minutes. Consequently, instead of studying the average revenue per minute (RMM), the average revenue per call (RMC) must be studied.

Additionally, in what concerns these lines, it was decided that it would be useful to have a chart depicting the evolution of NOS's market share regarding these value-added service phone lines, that is, the percentage of NNG 760x and NNG 761x that NOS represents, in terms of traffic. However, to do this, it was necessary to have data from the totality of the Portuguese

market concerning these lines, which is not publicly available on a structured and regular basis. Given this difficulty, and in order to find a proxy of this market share, it was possible to build a metric that allows one to study the weight of these lines belonging to NOS in the totality of the calls made to these lines by NOS's clients. This is done by dividing the on-net real minutes made to these lines by the total of real minutes (on-net plus off-net) made to these lines.

To do this, a new measure was built in Power BI.

Here, it is important to note that in the CR data source, there is no distinction between the on-net and off-net in these lines. Therefore, as was explained throughout this work, the off-net minutes of the CR source are equaled to the ones of the INTEC and the on-net minutes are calculated by difference. This is not applied to the number of calls and, therefore, to calculate this metric, the number of minutes was used. Also, since this metric is a division between the on-net and the total, doing this with the minutes or the number of calls should provide the same outcome.

The expression made to calculate it is the following:

```
NOS Weight = CALCULATE((SUM(CR_fact_table[OnNet Minutes])/SUM(CR_fact_table[Real Airtime]))*100)
```

Figure 5.6: NOS Weight measure

Regarding the layout, this page contains four line charts and one customer segment slicer.

The customer segment slicer is the same as in the Real Airtime's Evolution, Revenues' Evolution and Other Important Metrics pages. It is built by using the fields customer_segment, customer_subsegment, customer_technology, customer_subscription and PCG Detailed Segment (in this order) from the customer table (all but the last can be found in figure 4.3). Again, when no customer segment is chosen, they all appear aggregated as a whole.

Since the traffic type is fixed on this page, there is no slicer regarding the traffic.

All the line charts have the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy. Also, they all have as the legend the destination_line field of the traffic table.

In the first top line chart, the NOS Weight metric built before can be found on the y-axis (the lines of this chart in figure 5.7 will not be shown due to confidentiality issues). In the second top chart, this axis presents the Nbr. of Calls metric of the CR fact table. In the first bottom chart, the y-axis has the Revenues metric from the CR fact table, while in the second bottom chart, it has the RMC metric from the CR fact table.

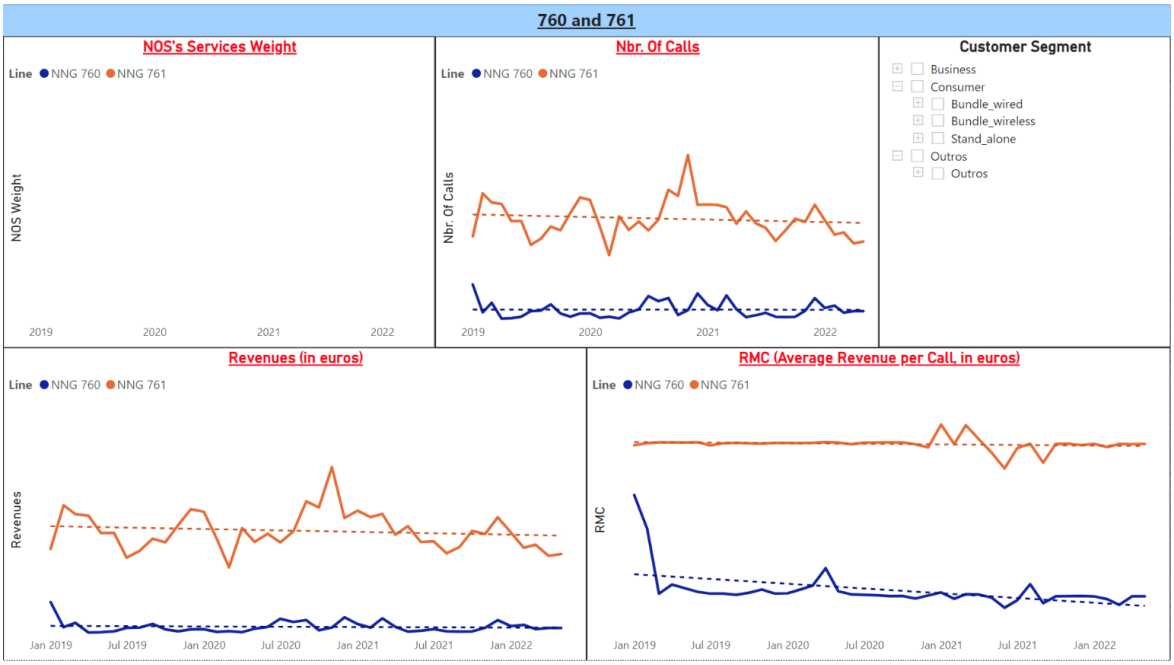

Figure 5.7: 76x's Evolution page

5.4.2 Findings

Considering the number of calls, NOS's clients call more to the 761x lines than to the 760x lines, given that the 761x lines have been replacing the 760x lines. Also, the number of calls to these lines seems to be at the same level over the period under study, with some peaks due to the pandemic (these lines were used for charitable purposes). The same happens for the revenues.

The average revenue per call (RMC) is, as expected, very steady, being around €1 for the 761x lines (which have a maximum retail price per call of €1) and around €0,60 for the 760x lines (which have a maximum retail price per call of €0,60).

5.5 Zoom in the Consumer Segment

As was said in section 4.1.3, some other aggregations were made in the customer table so that the user can study different customers aggregated by specific characteristics. This page contains two of those other aggregations, which cannot be shown due to confidentiality issues.

Because these aggregations overlap, it was impossible to have them together in one chart; therefore, six line charts are present on this page.

In the top three charts, a filter was applied to the first aggregation field of the customer table so that only the desired value appears. In the bottom three charts, a filter was applied to the second aggregation field of the customer table so that only the other desired value appears.

All the line charts have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy.

In both the first top chart and the first bottom chart, the y-axis contains both the Real Airtime and the OffBundle Billed Airtime metrics of the CR fact table.

In both the second top chart and the second bottom chart, the y-axis contains the Revenues metric of the CR fact table.

In the third top chart and the third bottom chart, the y-axis contains the RMM metric built earlier.

This page also contains a slicer with the types of traffic. This slicer has the traffic hierarchy, which is built by using the fields traffic_type, traffic_lvl2 and traffic_lvl3 (in this order) from the traffic table (figure 4.4), just like before. When no traffic is chosen, they all appear aggregated in the lines of the charts. To study specific types of traffic, the user must choose them in this slicer.

5.6 Airtime vs Revenues

This last page of the dashboard enables a better study of the relationship between the revenues (in euros) and the off-bundle billed minutes, which are the ones that present an extra charge to the clients and, therefore, extra revenues for the company. It also allows studying the % of these off-bundle minutes and the billing factor.

This page contains one line and stacked column chart, two line charts and one slicer.

Again, all the charts have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy.

The line and stacked column chart have in the column y-axis the OffBundle Billed Airtime field from the CR fact table and in the line y-axis the Revenues field of that same fact table.

The first line chart has as the y-axis the % Off-Bundle metric built earlier, while the second line chart has the Billing Factor metric, also built before. The two weren't put together in one line chart given the significant differences in values (the % of minutes off-bundle is always under 5% while the billing factor is always above 100%).

Lastly, the slicer has the traffic hierarchy, which is built by using the fields traffic_type, traffic_lvl2 and traffic_lvl3 (in this order) from the traffic table (figure 4.4), just like before. When no traffic is chosen, they all appear aggregated in the columns and lines of the charts. To study specific types of traffic, the user must choose them in this slicer.

Figure 5.8: Airtime vs Revenues page

Chapter 6

Part II: Integrity Check

This part of the dashboard contains three pages.

The choice was to have a more general first page where one can view the comparisons between the CR and INTEC data sources, between the Network and INTEC data sources and between the CR and Network data sources (in their totality, that is, off-net plus on-net real minutes). In this first page, one can only compare the data sources in what concerns the total basic traffic and the total special services traffic (it is also possible to study the differences regarding the international, the total national, the total NNG traffic and the total NCurtos traffic). In this page, only percentage differences can be studied.

It is very important to remember that all the comparisons regarding the INTEC data source contain only off-net minutes, since this data source only has that purpose.

The second and third pages allow to check for the differences in a more detailed way, presenting both absolute and percentage differences.

Lastly, it is important to note that due to confidentiality issues, in section 6.1, the lines in the charts will not be shown. The figures in this section are only for demonstration purposes.

6.1 Construction

6.1.1 Aggregated Comparison

As mentioned before, this first page contains the comparisons between the CR and INTEC data sources, between the Network and INTEC data sources and between the CR and Network data sources (in their totality, that is, off-net minutes plus on-net real minutes) for more aggregated types of traffic.

This page contains three line charts along with one slicer.

All the line charts have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy. They also have as the legend the integrity_lvl1 field of the traffic table.

The first top chart presents the differences, in percentage, between the CR and INTEC data sources; therefore, the y-axis is the CR-INTEC (%) measure.

The second top chart contains the differences, in percentage, between the Network and INTEC data sources and, therefore, the y-axis is the Network-INTEC (%) measure built earlier.

The bottom chart presents the differences, in percentage, between the Network and CR data sources and, therefore, the y-axis is the Network-CR Total (%) measure built earlier.

The slicer contains the integrity_lvl1, and integrity_lvl2 fields (in this order) of the traffic table. This slicer was added to this page so that the user can choose if they only want to see the most aggregated differences (basic and special services traffic) or if they also want to see the differences in what concerns the total International traffic, the total National traffic, the total NNG traffic or the total NCurtos traffic.

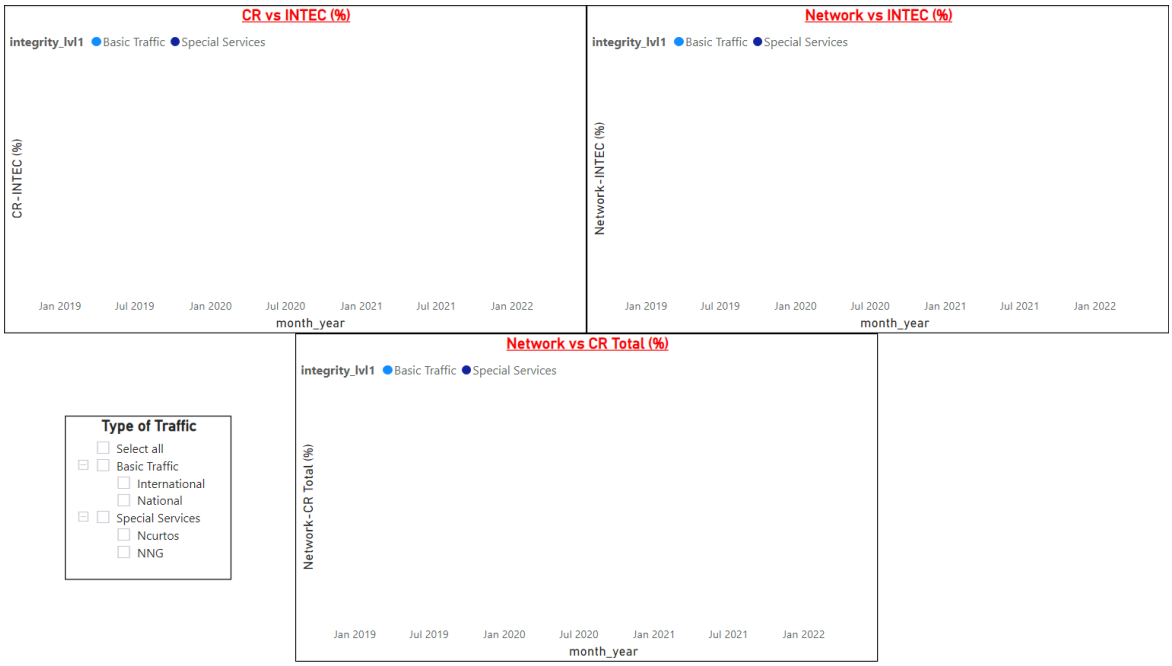

Figure 6.1: Aggregated Comparison page

6.1.2 INTEC Comparisons

On this page, one can find the differences between the INTEC data source and the CR and Network data sources, in both percentage and absolute values, in a more detailed way.

As said before, it is very important to always have in mind that the comparisons regarding the INTEC data source regard only off-net minutes.

This page contains four line charts along with one slicer.

All the line charts have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy. They also have as the legend the integrity_lvl3 field of the traffic table.

The first top chart presents the differences, in percentage, between the CR and INTEC data sources and, therefore, the y-axis is the CR-INTEC (%) measure.

The second top chart presents the differences, in percentage, between the Network and INTEC data sources and, therefore, the y-axis is the Network-INTEC (%) measure.

The first bottom chart presents the differences, in absolute values, between the CR and INTEC data sources and, therefore, the y-axis is the CR-INTEC (Abs) measure.

The second bottom chart presents the differences, in absolute values, between the Network and INTEC data sources and, therefore, the y-axis is the Network-INTEC (Abs) measure.

The slicer contains the integrity_lvl1, integrity_lvl2, integrity_lvl3, and integrity_lvl4 fields (in this order) of the traffic table. This slicer was added to this page because when all the traffic types are present in the charts, some of them loose visibility due to the significant differences in other traffic types. By having this slicer, the user can choose the traffic types they want to study in a detailed way.

Figure 6.2: INTEC Comparison page

6.1.3 CR vs Network

In this page of the dashboard, the user can find the differences between the CR and Network data sources, in both percentage and absolute values, in a more detailed way.

This page contains four line charts and one slicer.

All the line charts have as the x-axis the month_year field of the period table. Also, the values of this axis need to be the ones of the month_year field and not the date hierarchy. They also have as the legend the integrity_lvl3 field of the traffic table.

The first top chart presents the differences, in percentage, between the CR and Network data sources in what concerns their off-net minutes and, therefore, the y-axis is the Network-CR OffNet (%) measure.

The second top chart contains the differences, in percentage, between the CR and Network data sources in what concerns their on-net minutes and, therefore, the y-axis is the Network-CR OnNet (%) measure.

The first bottom chart presents the differences, in absolute values, between the CR and Network data sources in what concerns their off-net minutes and, therefore, the y-axis is the Network-CR OffNet (Abs) measure.

The second bottom chart presents the differences, in absolute values, between the CR and Network data sources in what concerns their on-net minutes and, therefore, the y-axis is the Network-CR OnNet (Abs) measure.

Just like before, the slicer contains the integrity_lvl1, integrity_lvl2, integrity_lvl3, and integrity_lvl4 fields (in this order) of the traffic table. By having this slicer, the user can choose the traffic types they want to study in a detailed way.

Figure 6.3: CR vs Network page

6.2 Findings

First, it is very important to mention that the findings of the integrity check process were incredibly important since they allowed us to make improvement decisions in what concerns the integrity check and the actual database.

In the very initial approach to the integrity check process, which was already mentioned in section 4.3, some findings were made.

Despite this, a comprehensive study was made with the Detailed Descriptions of the traffic of the CR and Network data sources. In this study, it was discovered that these two data sources

are not 100% compatible in terms of the Detailed D escription. In the case of the CR data source, some of the descriptions are less detailed than one may expect, being very broad and unspecific, making it impossible to draw meaningful conclusions on why these differences occur.

For example, while the Network data source's Detailed Descriptions always allow checking the operator to which the calls are destined, the CR data source sometimes only states the destination's city, making it impossible to know the operator and, therefore, compare.

Lastly, the first assumptions taken, based on the experience of the members of the department, were: the INTEC data source always has good and trustworthy values in what regards off-net traffic, and the CR data source does not distinguish well between traffic made to fixed and to mobile destination's technology nor between traffic made to off-net and on-net.

6.2.1 Basic Traffic

International Traffic

In what concerns the International traffic, the Network data source is aligned with INTEC, except in some months, due to the delay in implementing in the Network data source the recent developments of the Network CDRs, leading to incomplete integration of the total traffic made in the period (this can be found in section 3.2).

The CR data source seems to also be aligned with the INTEC, with most differences under 6% (in absolute value), except in some months.

It is worth noting that in one month of 2020, there is a very noticeable peak in the INTEC differences with both the CR and Network data sources. This peak is not visible in the difference between the CR and Network. This could mean that, in that period, there was some issue with esthe data of the INTEC source.

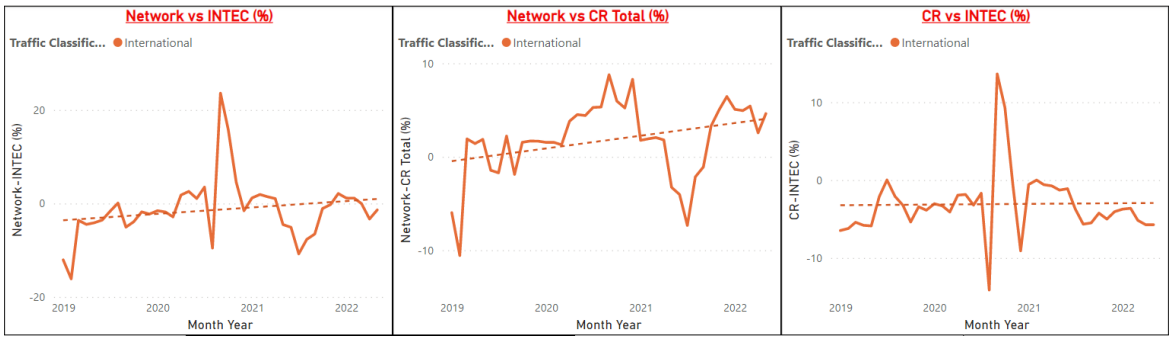

Figure 6.4: Differences, in percentage, between the data sources for the International traffic

National Traffic

In the first approaches to the integrity check, the differences in the data sources in the basic national traffic were studied by dividing it into OffNet - Fixed, OffNet - Mobile, and OnNet. Since this division was previous to the visualization of the integrity check in Power BI, the differences will not be shown.

First, note that the OnNet traffic does not exist in the INTEC data source so we can only compare the CR and Network data sources with each other.

For the Fixed and Mobile OffNet traffic, the Network and INTEC data sources are mostly aligned, except for the months of May to September of 2021, because of some delay of the Network data source in incorporating the totality of recent developments of Network CDRs.

In what concerns the comparison of the CR data source both with the Network and INTEC data sources, for the Fixed and Mobile, Off-Net and On-Net traffic, significant differences were found (over 30%), even though with opposite signs, indicating that there may be some misclassification of traffic details between the different traffic groups. It appears to exist mobile destination's technology communications being classified as fixed in the CR source.

As it was mentioned before, the CR data source only has the detail available in the operating billing systems, therefore, it can have some significant differences when comparing with the other two sources.

Because of this, it was decided to study the basic national traffic in a different way. This type of traffic was then divided into Basic - Fixed (contains both OffNet - Fixed and OnNet - Fixed traffic) and Basic - Mobile traffic (contains both OffNet - Mobile and OnNet - Mobile traffic). However, even though this distinction has shortened the gap, there were still some relevant differences, given that the CR data source does not distinguish well between fixed and mobile destination's technology. This is particularly relevant in the customers with mobile technology, because these clients have an increasing amount of minutes included in their bundles, which include fixed and mobile destination's technology as well as on-net and off-net destinations. Moreover, the tariffs charged for the extra consumption are the same for both national mobile and fixed destinations, as well as for both on-net and off-net destinations.

In contrary to what as been said in the previous paragraphs, the CR and INTEC are very aligned in what concerns the Basic - Mobile traffic.

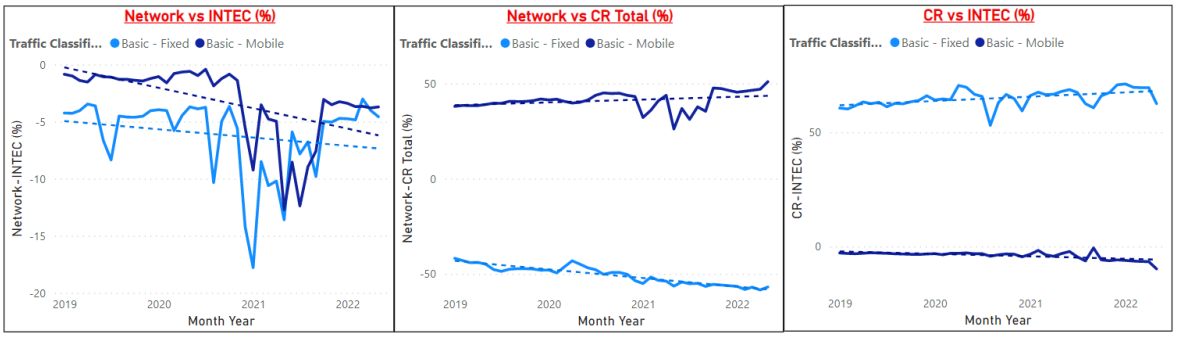

Figure 6.5: Differences, in percentage, between the data sources for the National traffic, divided into Basic - Mobile and Basic - Fixed

When looking at the basic national traffic as a whole, the data sources are all actually very aligned. This confirms the thought that there is misclassification of traffic details between the different traffic groups.

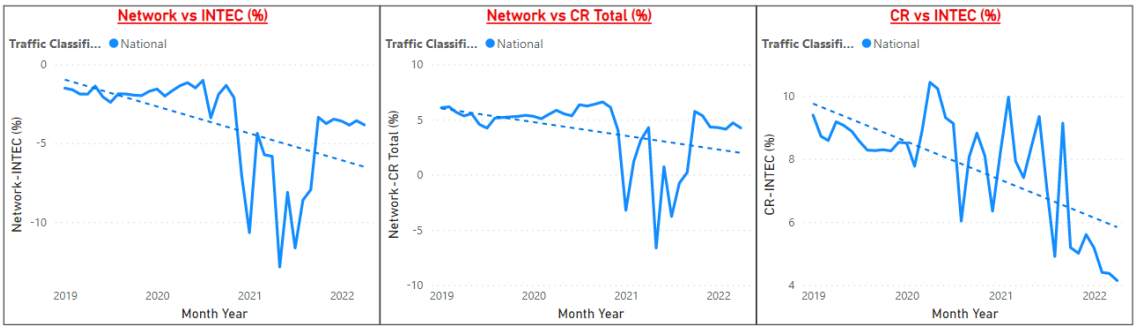

Figure 6.6: Differences, in percentage, between the data sources for the National traffic as a whole

6.2.2 Special Services Traffic

Before beginning the study of the Special Services traffic's differences, it is important to remember that the CR data source contains both the on-net and off-net traffic aggregated together and, therefore, the on-net traffic is calculated by removing the INTEC's special services minutes, which are the off-net. Because of this, the differences between the CR and INTEC data sources should always be null for the special services. Since the differences between these two data sources are null, they will not be mentioned nor shown graphically, except in some specific cases where they are relevant.

Note that, in the figures found in this section, the comparisons between the Network and INTEC data sources comprise only the off-net minutes.

NCurtos

The study of the differences of the NCurtos lines will begin with the line 16x. This is because this line is the one that weighs the most (about 80%) in the minutes of the entire NCurtos lines.

When comparing the INTEC and Network data sources, they are quite aligned (differences are under 10% (in absolute) of the INTEC minutes).

Comparing the CR and Network data sources, they are quite different (even more in the months of 2021 mentioned before), being the CR always higher than the Network.

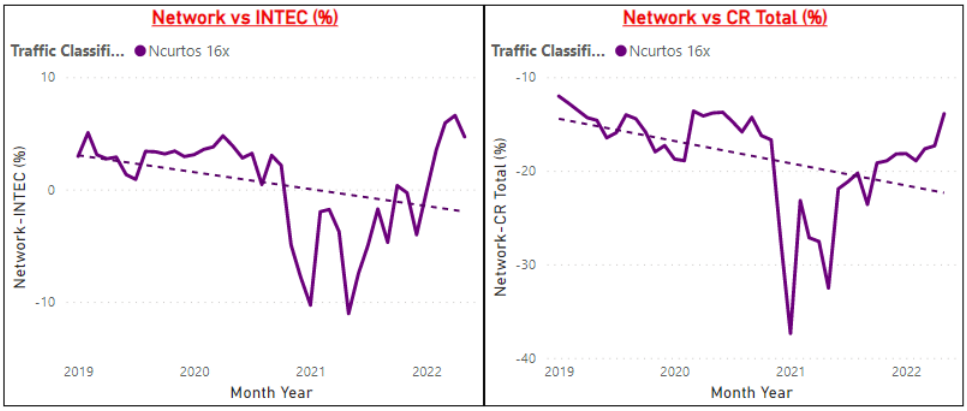

Figure 6.7: Differences, in percentage, between the data sources for the NCurtos 16x traffic

Now, for the NCurtos 18x. This is the line that weighs the least (under 1%) in the minutes of the entire NCurtos lines.

The Network and INTEC data sources are very aligned in what concerns this line.

Now for the CR source, it has very low amounts of minutes, which can be seen when comparing the Network and CR (Network is always very above the CR). There are a lot of minutes missing from the CR data source for this type of traffic. This may be due to wrong classifications of calls' descriptions (that is, this type of calls is being wrongly classified as other types of calls), given that there is no sufficient detail in this data source to know for sure which category the calls fall into.

Just like was said before, there was an approach where the traffic was studied regardless of the destination's operator and, therefore, the INTEC should be much smaller than the other two data sources. In this approach, for this type of traffic, it was discovered that the number of minutes of the CR data source (which contains on-net and off-net traffic) is way smaller than the INTEC data source ones (which has only the off-net traffic), which should not happen.

Also in that approach, it was discovered that the Network data source (in its totality) is very aligned with the INTEC data source, which shouldn't happen since, again, the Network data source has both on-net and off-net traffic, and the INTEC has only off-net. That is, the Network data source should have much more minutes than the INTEC. One possible explanation is the nonexistence of on-net traffic in this line, that is, NOS possesses little to no lines of this type. In fact, when checking the on-net minutes of the Network data source for this type of traffic, it was found that they are null, which confirms the suspicion. This was also confirmed by a member of the department.

The issue with this type of traffic resides in the fact that the CR data source has low amounts of minutes when compared with the other two data sources and, therefore when equalizing the Off-Net minutes to the ones from the INTEC and doing the On-Net minutes by difference (of the total minutes of CR and the Off-Net minutes), these On-Net minutes are always negative, which does not make sense. Thus, in this case, it was decided that the off-net minutes of the CR data source must be the ones extracted from the BO platform and are not equaled to the ones from the INTEC data source. The on-net minutes of this type of traffic in the CR data source are, therefore, null. The graphs in figure 6.8 were obtained after this decision.

Then, when comparing the minutes of the CR source with the ones of the INTEC source, it was discovered that the CR (just like when comparing it to the Network) has much less minutes than the INTEC data source.

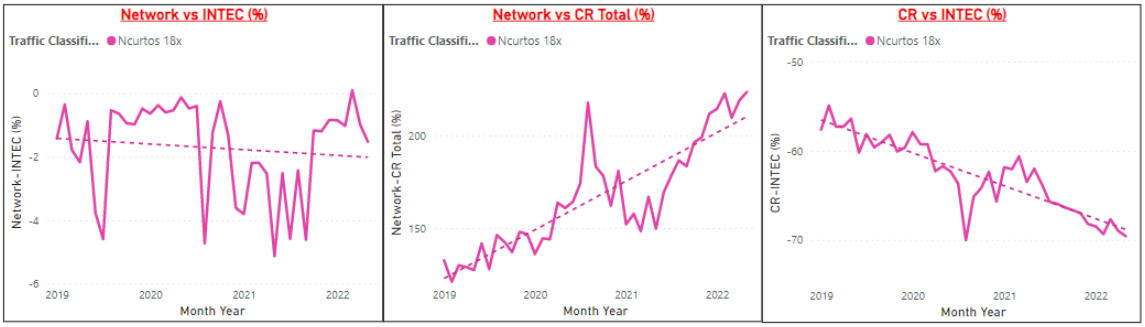

Figure 6.8: Differences, in percentage, between the data sources for the NCurtos 18x traffic

In the case of the NCurtos Outros, this line represents under 20% of the minutes of the entire NCurtos lines.

When comparing the INTEC and Network data sources, the Network presents much more minutes of this type of traffic. When studying this problem in a more detailed manner, it was discovered that it is mainly due to a specific type of CDRs, which has a relevant amount of minutes in the Network data source but, since it is a traffic that generates no revenue (the clients are not charged for it), it is not fully accounted in the CR data source. A deeper study was made, in order to assess if those calls were actually appearing in the invoice received by NOS's clients. It was discovered that those calls were not in the invoice of the clients that made them and, because of that, the minutes of those specific calls are not being completely accounted in the CR data source. Later on, the specific type of CDRs concerning those calls were removed from the database (in both the CR and the Network sources) and, consequently, the dashboard. This significantly decreased the differences, in the on-net minutes (since these calls are only on-net) but it didn't solve the differences in the off-net minutes. Despite this, the decrease in the difference in the on-net minutes was not as relevant as it was expected, therefore, that specific type of CDRs is not the only one creating the differences.

Now, comparing the CR and Network data sources, they are very different (CR has much less minutes). The differences are over 2 times the CR minutes, that is, the CR is always smaller than the Network. Again, this may be due to wrong classifications of calls' descriptions in the CR data source (that is, this type of calls is being wrongly classified as other types of calls) or to wrong classifications of calls' descriptions in the Network data source (that is, other types of calls are being wrongly classified as this type of calls), given the lack of sufficient detail in these sources.

In this line, the plots of the differences cannot be shown due to confidentiality issues.

Later, given the significant differences in the NCurtos 18x and the NCurtos Outros, it was decided to aggregate them together in the first part of the dashboard, which concerns the study of the traffic's revenues. This aggregation was also made in the integrity check, which resulted in smaller differences between the data sources.

In what concerns the NCurtos lines, the main thoughts are that there seems to be calls belonging to the 16x line being classified as NCurtos Outros in the Network data source (since this source is below the CR in the line 16x but is very above the CR and the INTEC in the Outros line). Therefore, it is necessary to perform a deeper study with the help of the Data Warehouse team, in order to find out if it is possible to have more detail both in the Network and CR data sources, which will allow for the correct classification of the different lines, making these sources more comparable.

NNG

Starting with the 707x line, it weighs less than 10% in the totality of the NNG minutes.

The Network and INTEC data sources are very aligned, again having some higher differences in some months of 2021, due to the delay in implementing the developments of the Network CDRs in the Network data source.

When comparing the CR and Network data sources, they are very different. The CR is always smaller than the Network (CR has much less minutes). This may be due to wrong classifications of calls' descriptions in the CR data source (that is, this type of calls is being wrongly classified as other types of calls), given the lack of detail in this source. It may also be due to the billing systems feeding the CR source, since if the calls present no charge to the clients, the minutes may not be fully accounted for.

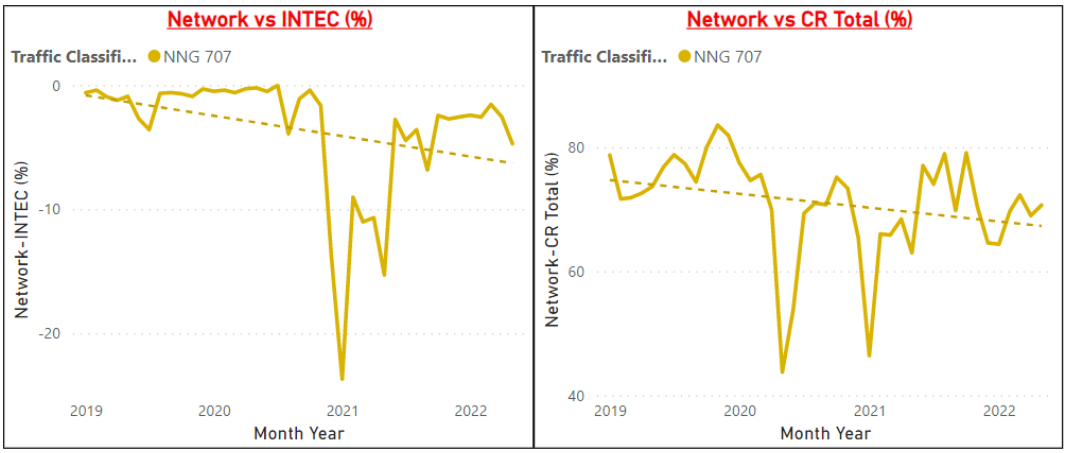

Figure 6.9: Differences, in percentage, between the data sources for the NNG 707x traffic

Now, for the 76x lines (that is, the lines 760x and 761x), they account for less than 1% of the total NNG minutes. This is because, once again, the calls to these lines are always of very few seconds and are, therefore, should be measured in terms of number of calls.

The Network and INTEC data sources are quite aligned, having some higher differences in some months of 2021, which was already mentioned.

For the 760x line, when comparing the CR and Network data sources, they are quite different (CR has much less minutes) up until the beginning of 2021 (at this point there were some developments in the CR data sources that allowed to have more detail of the destination numbers). However, these differences seem to be decreasing, becoming closer to zero.

For the 761x line, the CR and Network data sources are very aligned.

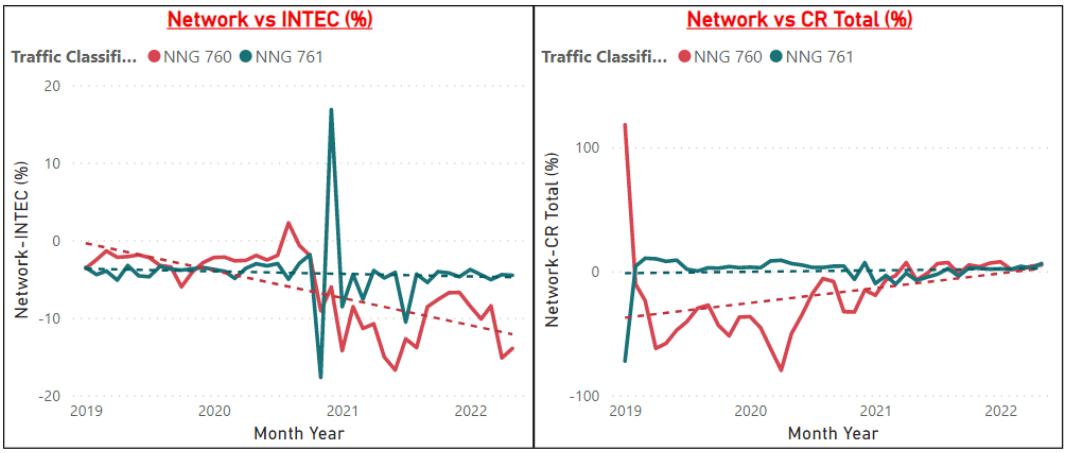

Figure 6.10: Differences, in percentage, between the data sources for the NNG 76x traffic

Now, for the 800x line, it accounts for less than 20% of the total NNG minutes. It is very important to understand that this is a free line and, therefore, there is no charge to the clients when they call this line. Because of this, it is possible that the CR data source does not fully account for the minutes of this line (the explanation for this can be found in section 3.2).

The Network data source always has more minutes than the INTEC data source. This difference seems to be increasing over time, and may indicate some classification issues regarding the off-net traffic within the Network data source.

The CR and Network data sources in their totality are very different (CR is always smaller, probably because of what was mentioned before - free line). Despite of this, this difference seems to be decreasing over time.

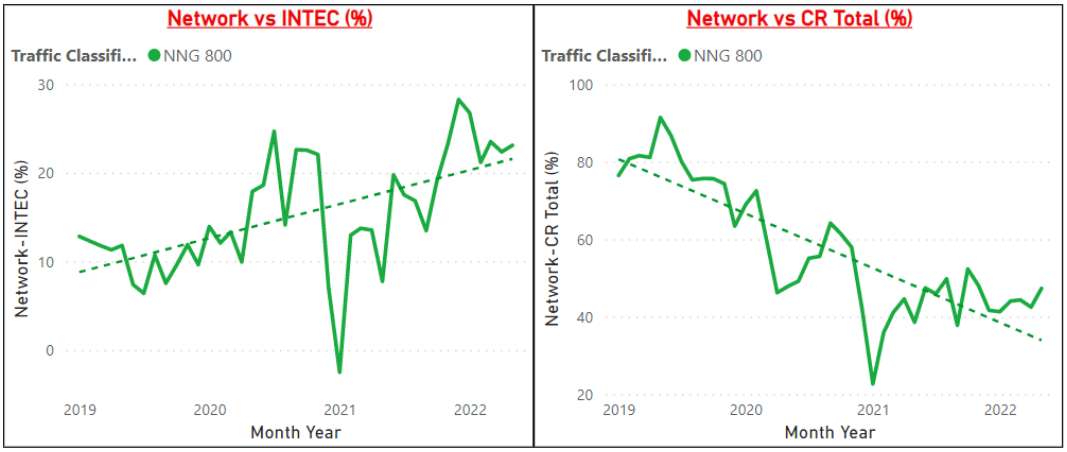

Figure 6.11: Differences, in percentage, between the data sources for the NNG 800x traffic

Moving to the 808x line, this is the line that weighs the most in the total of the NNG lines, account for around 70% of it. This line contains the number of the National Health Service,

which has no charge for the clients (since March of 2020). Because of this, just like in the previous line, it is possible that the CR data source does not fully account for its minutes.

The Network and INTEC data sources are very aligned, again having some higher differences in some months of 2021, due to the delay in implementing in the Network source the developments of the Network CDRs.

In what concerns the CR and Network data sources in their totality, they are not very aligned. The differences between these two data sources are relevant, although being below 20% of the CR minutes. The Network seems to always be above the CR data source, which can be due to the gratuity of this line.

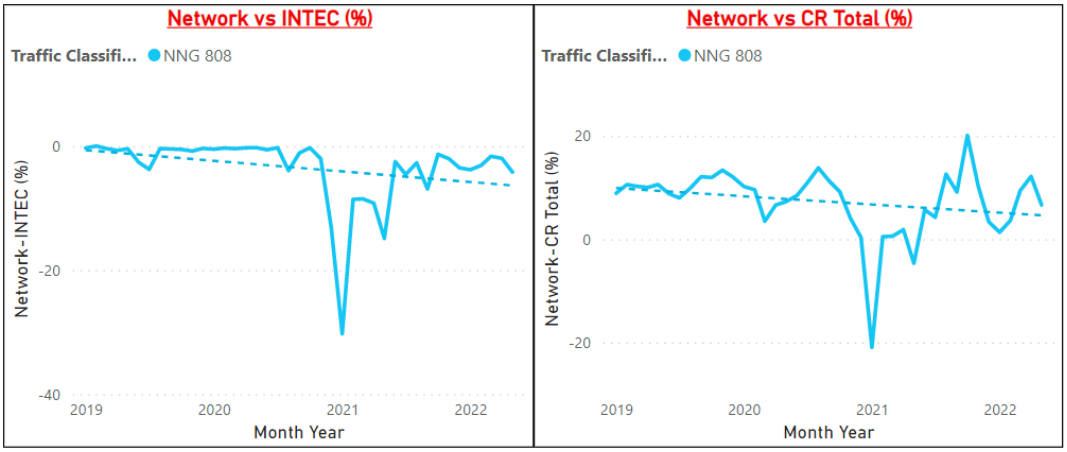

Figure 6.12: Differences, in percentage, between the data sources for the NNG 808x traffic

In the case of the NNG 882x, it was discovered that this line does not have On-Net, it is all Off-Net. The 882x line refers to Virtual Calling Cards, which can be bought by anyone (being a client of NOS or not) so these calls can never be considered On-Net. Therefore, in this specific case, it was decided that the off-net minutes of the CR data source must be the ones extracted from the BO platform and are not equaled to the ones from the INTEC data source. The on-net minutes of this type of traffic in the CR data source are, therefore, null. The graphs in figure 6.13 were obtained after this decision.

The Network and INTEC data sources are very aligned, having some peaks (which should be further studied). The same happens for the differences between the CR and INTEC and for the differences between the CR and Network data sources.

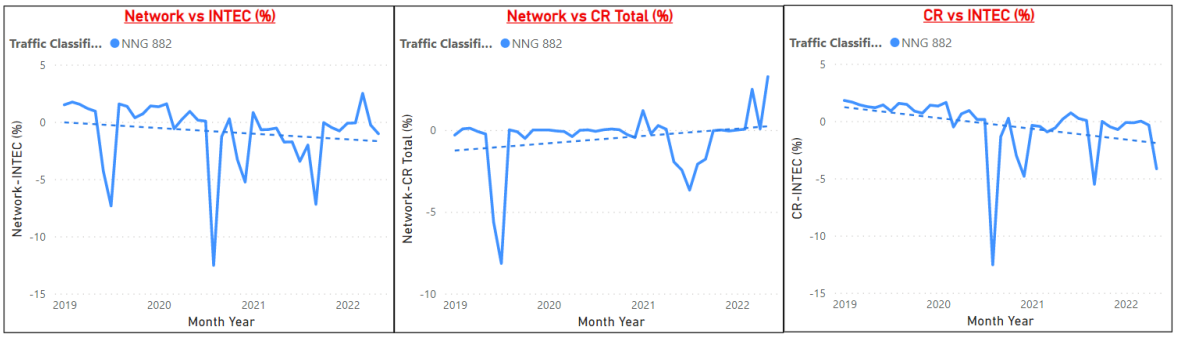

Figure 6.13: Differences, in percentage, between the data sources for the NNG 882x traffic

6.2.3 Other Integrity Findings

Negative values in the minutes on-bundle

During the integrity check, it was discovered that some values in the metrics OnBundle Real Airtime and OnBundle Billed Airtime were negative, which should not happen since the number of minutes cannot be negative.

This was brought to the attention of the members of the department, which explained that the Real Airtime and OffBundle Real Airtime come from different systems which are not 100% compatible (there are no direct keys between the tables). The OnBundle Real Airtime is calculated by subtracting the OffBundle Real Airtime to the Real Airtime. Since these last two are not 100% compatible, this can lead to some negative values in the OnBundle Real Airtime.

The member of the team explained that, when these systems were built, they had a choice of keeping the negative values in the minutes on-bundle or changing them to zero. Since it is desirable that the total minutes are equal to the sum of the minutes on-bundle and the minutes off-bundle, the negative values were kept.

A study was made on these negative values and it was discovered that the majority of them belong to customers with prepaid subscriptions. One of these was chosen and two integrity checks were made for it (following the last approach): the first keeping the negative values in the OnBundle Real Airtime and the other changing them to zero (only when they are negative) and changing the Real Airtime to the OffBundle Real Airtime. The conclusion was that changing the negative values to zero and the total minutes to the off-bundle minutes actually increased the differences between the data sources. Because of this, it was decided to keep the negative values in the OnBundle Real Airtime.

Chapter 7

Conclusion

In this dissertation, all the project steps developed within the scope of a curricular internship at the Planning and Management Control department of NOS were presented.

Currently, the team has a very heavy workload when building the voice services' models needed (mentioned in section 3.3), mainly because of the existence of three different data sources (section 3.2) feeding them (the same metric can come from different data sources, leading to incongruencies in the information created). Also, because of the models' complexity, no data visualization techniques are applied to them; therefore, the department does not fully extract the best possible information from them. As said before, the different models are necessary because they represent different requirements, but that does not mean that the data used to build them cannot come from a unique database.

Therefore, this project's goals consisted in: building a unique database where the Planning and Management Control Department team can retrieve the data of the metrics used to build the models they need; apply data visualization techniques to the built database; and perform a data integrity check, where possible incongruencies can be studied and the data cleaned as much as possible.

This project had a relevant impact on the work of the department, especially concerning the quality of the information created with the data retrieved, the efficiency in creating it and its usefulness (in terms of understandability through the application of data visualization techniques).

Concerning the data used for the project development, described in section 3.1, it is real data provided by the company and comprises different granularities. Despite NOS having many services, this work focused on voice traffic only (comprising basic and special services traffic). Also, the department works with different data sources, but in this project, only three were considered (Customer Revenues, Interconnect Billing and Network), which are the ones containing the relevant metrics for the study in question.

Concerning the construction of the database and following the methodology proposed by Kimball and Ross, a dimensional model in the form of a star schema was designed. This dimensional model was applied in Excel, however, with some modifications. Four dimension tables were constructed (in four sheets of Excel, where different aggregations were made so that different levels of the hierarchies can be studied), along with four fact tables (in other four sheets of Excel). Three of those fact tables (CR fact table, INTEC fact table and Network fact table)

contain the metrics retrieved from each of the three data sources, while the last one (CR missing according to INTEC) was constructed during the integrity check when it was discovered that not all combinations of period, customer and traffic are compatible in all the data sources. The four fact tables are linked to the four dimension tables through relationship keys. This forms a galaxy schema instead of a star schema.

Several metrics were constructed both in Excel and in Power BI.

In Excel, two metrics were made in each of the fact tables so that on-net and off-net minutes can be distinguished in the integrity check (this distinction is not relevant for the study of the voice traffic generating revenues). These metrics were especially important in the CR's special services since the off-net minutes of these types of traffic in this data source were equaled to the ones of the INTEC source.

In Power BI, other metrics were calculated, both for Part I and Part II of the dashboard. For the first part of the dashboard (Part I: Study of the voice traffic generating revenues), there was the need to calculate the metrics regarding the % of Minutes OffBundle, the Billing Factor, the Average Revenue per OffBundle Billed Minute (RMM), the Average Revenue per Call (RMC) and the proxy to NOS's Market Share in the 76x lines (NOS Weight). Also, four quick measures were built in order to show the percentage difference between the years in the Y2D page.

For the second part of the dashboard (Part II: Integrity Check), it was necessary to calculate the differences (both absolute and in percentage) between the different data sources (off-net differences between the INTEC and CR, the INTEC and Network, the Network and CR; on-net differences between the Network and CR; and the total differences between the Network and CR).

After all the metrics constructed, the dashboard implementation was performed. Then, the dashboard was analyzed which led to the discovery of important findings.

In Part I, it was discovered that the clients with fixed technology have basic national and international traffics included in their bundles and, therefore, despite the large amounts of minutes, these traffics will have low revenues. Therefore, the revenues of the special services will weigh more than the ones from the basic traffic, especially the lines 76x, 707x and 808x. The clients with mobile technology have been having a decline in the international traffic and an increase in the national traffic. For these clients, the basic national fixed traffic has smaller weight than the basic national mobile traffic, despite both being included in the bundles. Once again, the revenues of the special services weigh more than the ones from the basic traffic, especially the lines 76x, 707x and 16x.

In Part II, it was discovered that, given the way of functioning of the data sources and the lack of detail in them, there are several differences between them, some of them even being very significant (when studied in the detailed way). It is thought that there is misclassification of traffic details between the different traffic groups.

Given all that was found during the internship, one suggestion made is that NOS's team responsible for the Data Warehouse should assure the correct integration of the data coming from the different systems behind the Data Warehouse. Despite the different requirements for the data sources (that will always lead to differences between them), the differences should be smaller than they actually are. Also, the Planning and Management Control department should work with the Data Warehouse team so that the CR and Network data sources present more

detailed data than the data they present right now, so that the sources are more comparable and the traffic better classified.

Also, it was discovered that there is a significant amount of minutes belonging to unknown types of traffic in the Network data source. This was notified to the Data Warehouse team, which is currently working on solving that issue.

In the future, it would be interesting that the database and dashboard built evolve to include not only the voice traffic but also mobile data or SMS traffic.

Bibliography

Alhassan, I., Sammon, D., & Daly, M. (2019, 06). Critical success factors for data governance: a telecommunications case study. *Journal of Decision Systems*, *28*, 1-21. doi: 10.1080/12460125.2019.1633226

Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2019, 11). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, *23*. doi: 10.1007/s00779-017-1104-3

Aparicio, M., & Costa, C. J. (2015, jan). Data visualization. *Commun. Des. Q. Rev*, *3*(1), 7–11. Retrieved from https://doi.org/10.1145/2721882.2721883 doi: 10.1145/2721882.2721883

Bakhshaliyeva, M. (2021). *The ultimate guide to managing data quality*. Retrieved 2022-01-04, from https://www.greenbird.com/news/utilities-data-quality

Becker, L. T., & Gould, E. M. (2019). Microsoft power bi: extending excel to manipulate, analyze, and visualize diverse data. *Serials Review*, *45*(3), 184–188.

Brath, R., & Peters, M. (2004). Dashboard design: Why design is important. *DM Direct*, *85*, 1011285–1.

Chen, Z. (2003). Data warehousing and data marts. In H. Bidgoli (Ed.), *Encyclopedia of information systems* (p. 521-533). New York: Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B0122272404000368 doi: https://doi.org/10.1016/B0-12-227240-4/00036-8

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Incorporated. Retrieved from https://books.google.pt/books?id=qWER8Im-WYIC

Gama, J., & Veloso, B. (2020). *Olap and power bi*. University Lecture.

Giordano, L., & Onions, P. (2021, 09). Data governance is not data management!

Guo, A., Liu, X., & Sun, T. (2018). Research on key problems of data quality in large industrial data environment. In *Proceedings of the 3rd international conference on robotics, control and automation* (p. 245–248). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3265639.3265680 doi: 10.1145/3265639.3265680

Han, J., Kamber, M., & Pei, J. (2012). 4 - data warehousing and online analytical processing. In J. Han, M. Kamber, & J. Pei (Eds.), *Data mining (third edition)* (Third Edition ed., p. 125-185). Boston: Morgan Kaufmann. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780123814791000046 doi: https://doi.org/10.1016/B978-0-12-381479-1.00004-6

Hobbs, L., Hillson, S., Lawande, S., & Smith, P. (2005). 1 - data warehousing. In L. Hobbs, S. Hillson, S. Lawande, & P. Smith (Eds.), *Oracle 10g data warehousing* (p. 1-22). Burling-

ton: Digital Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9781555583224500035 doi: https://doi.org/10.1016/B978-155558322-4/50003-5

Holsheimer, M., & Siebes, A. P. (1994). *Data mining: The search for knowledge in databases.* (Tech. Rep.). Amsterdam, the Netherlands.

Howard, G., Lubbe, S., & Klopper, R. (2011, 01). The impact of information quality on information research. *Alternation Special Edition: Management, Informatics and Research Design II*, *18*, 288-305.

IBM. (2020). *Olap.* Retrieved 2022-01-04, from https://www.ibm.com/cloud/learn/olap

Inmon, W. H. (1996). *Building the data warehouse (2nd ed.).* USA: John Wiley & Sons, Inc.

Inmon, W. H. (2002). *Building the data warehouse* (3rd ed.). USA: John Wiley & Sons, Inc.

Janes, A., Sillitti, A., & Succi, G. (2013, 01). Effective dashboard design. *Cutter IT Journal*, *26*, 17-24.

Kabakchieva, D. (2009). Business intelligence applications and data mining methods in telecommunications: A literature review.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley Publishing.

Lachev, T. (2015). *Applied microsoft power bi: Bring your data to life!* Prologika Press.

Pauwels, K., Ambler, T., Clark, B., LaPointe, P., Reibstein, D., Skiera, B., … Wiesel, T. (2009, 10). Dashboards as a service : Why, what, how, and what research is needed? *Journal of Service Research*, *12*, 175-189. doi: 10.1177/1094670509344213

Redman, T. C. (2017). *What is data reconciliation? definition, process, tools.* Retrieved 2022-01-04, from https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

SAP. (2017). *Sap help portal: Getting started with bi launch pad.* Retrieved 2021-12-23, from https://help.sap.com/viewer/baaf5c869e824d07ab0109e7b093348e/4.2.4/en-US

Shaffer, J. A. (2018). *The definition of a dashboard.* Retrieved 2022-01-03, from https://www.dataplusscience.com/DashboardDefinition.html

Shams Raza, M., & Nayak, A. (2014). A study on designing a layered star schema for data mining optimization. In *2014 conference on it in business, industry and government (csibig)* (p. 1-5). doi: 10.1109/CSIBIG.2014.7056948

Sivathanu, G., Wright, C. P., & Zadok, E. (2005). Ensuring data integrity in storage: Techniques and applications. In *Proceedings of the 2005 acm workshop on storage security and survivability* (p. 26–36). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1103780.1103784 doi: 10.1145/1103780.1103784

Stedman, C. (2020). *What is data governance and why does it matter?* Retrieved 2022-01-05, from https://searchdatamanagement.techtarget.com/definition/data-governance

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its applications.* doi: 10.1007/978-3-540-34351-6

Taylor, D. (2021a). *What is data reconciliation? definition, process, tools.* Retrieved 2022-01-04, from https://www.guru99.com/what-is-data-reconciliation.html#2

Taylor, D. (2021b). *What is data warehouse? types, definition & example.* Retrieved 2021-12-29, from https://www.guru99.com/data-warehousing.html

Weiss, G. M. (2005). Data mining in telecommunications. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 1189–1201). Boston, MA: Springer US.

Retrieved from https://doi.org/10.1007/0-387-25465-X_56 doi: 10.1007/0-387-25465-X_56

Weiss, G. M. (2009). Data mining in the telecommunications industry. In J. Wang (Ed.), *Encyclopedia of data warehousing and mining, second edition* (pp. 486–491). IGI Global. Retrieved from https://doi.org/10.4018/978-1-60566-010-3.ch076 doi: 10.4018/978-1-60566-010-3.ch076

Wind, Y. J. (2005). Marketing as an engine of business growth: a cross-functional perspective. *Journal of Business Research*, *58*(7), 863-873. Retrieved from https://www.sciencedirect.com/science/article/pii/S0148296304000141 (Special Section: Cross-functional cases in management education) doi: https://doi.org/10.1016/j.jbusres.2004.01.002

Appendix A

Appendix

A.1 Approaches to the Integrity Check

Traffic Types by Data Source			Integrity Check Approaches					
CR	**Network**	**INTEC**	**1st Approach**	**2nd Approach**	**3rd Approach**	**4th Approach**		
International - DEEA International - FEEA	International	International	International	International	International	International		Total Basic
OffNet - Fixed	OffNet - Fixed	OffNet - Fixed	OffNet - Fixed	Basic - Fixed	Basic - Fixed*	Basic - Fixed		
OnNet - Fixed	OnNet - Fixed		OnNet					
OnNet - Mobile	OnNet - Mobile			Basic - Mobile	Basic - Mobile*	Basic - Mobile		
OffNet - Mobile	OffNet - Mobile	OffNet - Mobile	OffNet - Mobile					
Serv NCurtos 16Out	OffNet - Serv NCurtos 16	OffNet - Serv NCurtos 16	NCurtos 16	NCurtos 16	NCurtos 16	NCurtos 16*		Total Special Services
Serv NCurtos 18	OnNet - Serv NCurtos 18		NCurtos 18	NCurtos 18	NCurtos 18*	NCurtos 18	NCurtos Outros (18 + Outros)	
	OffNet - Serv NCurtos 18	OffNet - Serv NCurtos 18						
Serv NCurtos Outros	OnNet - Serv NCurtos Outros		NCurtos Outros	NCurtos Outros	NCurtos Outros*	NCurtos Outros*		
	OffNet - Serv NCurtos Outros	OffNet - Serv NCurtos Outros						
Serv NNG 707	OnNet - Serv NNG 707		NNG 707	NNG 707	NNG 707*	NNG 707*		
	OffNet - Serv NNG 707	OffNet - Serv NNG 707						
Serv NNG 760	OnNet - Serv NNG 760		NNG 760	NNG 760	NNG 760*	NNG 760*		
	OffNet - Serv NNG 760	OffNet - Serv NNG 760						
Serv NNG 761	OnNet - Serv NNG 761		NNG 761	NNG 761	NNG 761*	NNG 761*		
	OffNet - Serv NNG 761	OffNet - Serv NNG 761						
Serv NNG 800	OnNet - Serv NNG 800		NNG 800	NNG 800	NNG 800*	NNG 800*		
	OffNet - Serv NNG 800							
Serv NNG 808	OnNet - Serv NNG 808		NNG 808	NNG 808	NNG 808*	NNG 808*		
	OffNet - Serv NNG 808	OffNet - Serv NNG 808						
Serv NNG 882	OnNet - Serv NNG 882		NNG 882	NNG 882	NNG 882*	NNG 882		
	OffNet - Serv NNG 882							
				Basic traffic: Aggregating on-net and off-net minutes. Special Services Traffic: Aggregating on-net and off-net minutes.	Basic traffic: Off-net minutes of the CR equaled to the ones from INTEC and on-net minutes calculated by difference. Special Services Traffic: Off-net minutes of the CR equaled to the ones from INTEC and on-net minutes calculated by difference.	Basic traffic and NCurtos 18x and NNG 882x: Off-net and On-net minutes of the CR data source not equaled to the ones of the INTEC source. Special Services Traffic (except NCurtos 18x and NNG 882x): Off-net minutes of the CR equaled to the ones from INTEC and on-net minutes calculated by difference. Adding visualization.		
			Problems: - Difficulty of the CR source in distinguishing between on-net and off-net traffic - Lack of comparability between the data sources (CR and Network: off-net + on-net; INTEC: off-net only)	**Problems:** - Difficulty of the CR source in distinguishing between on-net and off-net traffic - Lack of comparability between the data sources (CR and Network: off-net + on-net; INTEC: off-net only)	**Problems:** - Lack of visualization of the differences between the data sources			
			*Off-net minutes of the CR equaled to the ones from INTEC and on-net minutes calculated by difference					

Figure A.1: Approaches to the integrity check

Direção de Planeamento e Controlo de Gestão
NOS Corporate Center

Porto, 30th June 2022

José Varejão
Director of Faculdade de Economia do Porto

Dear Professor,

At this final moment of the project, we would like to leave here our feedback from the experience with the curricular internship of Inês Ferreira, within the scope of her master's thesis in Modelling, Data Analysis and Decision Support Systems (MADSAD) at FEP.

For past few years, it has been one of our objectives to implement an integrated analysis of the different models and data sources concerning voice of traffic (volumes, revenues and margins), but the project has been postponed due to continuous evolution resulting from NOS's operational platforms and information systems transformation plan.

This project carried out by Inês finally made this objective possible, and we want to highlight the following messages:

- The project had a direct impact on our work, right from the first few weeks, and culminated in the delivery of two very important outputs for our activity:
    1. Treatment and systematization of information from different data sources, and the respective integrity analysis within a business sense
    2. Construction in Power BI of a very relevant Business Intelligence tool, composed of an integrated database and a visualization dashboard, flexible and simple enough to respond to our needs, and moreover with an easy updating and maintenance process;
- Given the complexity of the information involved, the project was a long journey of "trial and error" discovery, and was only possible due to the work and dedication of Inês, who was tireless in analysing in depth the different angles of the business problem
- It is also important to state that it was very easy to guide and work with Inês, not only because of her natural curiosity and enthusiasm to embrace the project, but also because of her assertiveness in the way she breaks down the problems and explains them in a very simple way. I would also like to highlight her autonomy and rigor, as well as her "right attitude" that made very natural her on-boarding and integration within our team
- For all the experience and competence shown, Inês was recommended internally, and, after the application stage, she was already hired to be a trainee of our NOS Alfa program from Sep'22, being a pleasure to be able to continue to count on Inês' contribution in NOS team.

On our side, we would like to thank the Professors João Gama and Bruno Veloso for their collaboration in this partnership, which in the last two years has allowed us to keep in touch with young talent and state-of-the-art methodologies and academic knowledge, essential to our transformation and innovation journey.

Kind regards,

Cláudia Dias
Head of NOS' TELCO Management Control & Corporate BI