UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Presbycusis patterns in the Portuguese population

Identification and association with epidemiological and genetic factors

Mariana Martins Epifânio Frazão Aparício

Mestrado em Bioestatística

Trabalho de Projeto orientado por:

Orientador:	Prof.ª Marília Antunes
Co-Orientador:	Prof.ª Maria Helena Caria

2017

The important thing is to not stop questioning. Curiosity has its own reason for existing.

Albert Einstein

Acknowledgments

First of all, I would like to thank Professor Marília Antunes for the perfect guidance on this project. I could not have better person to be my supervisor being an honour to have her as my teacher. Since the beginning, she always helped me to have a critical thinking and showed confidence in my work. Secondly, to Professor Helena Caria for all support in biological aspects of the work and for letting me to partipate in the Presbycusis project.
I would like also to thank Professor Lisete who was always very helpful, as well to my colleagues for all good moments during the course.
Finally, to my family, due to values instilled in me that allowed me to be the person who I am. They always give me the strength and the support to follow the way that I want.

Abstract

Presbycusis or Age-Related Hearing Loss (ARHL) is the most prevalent sensorial impairment in the elderly, affecting more than 30% of people older than 65 years old. This condition has a negative impact on quality of life, which may lead to social isolation and the development of some psychiatric disorders. Although there are several studies based on prevalence of Hearing Loss (HL), only a few studies based on audiogram configurations or HL pattern were made.

The main aim of this study was to identify dominant audiogram patterns. Furthermore, based on that a classification procedure was build relying in a sample of 321 individuals aged between 62 and 115 years old. The obtained classification was validated through Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and then compared with audiogram pattern identification procedures existing in the literature. Finally, some statistical models were adjusted to the data in order to investigate the influence of demographic, environmental, medical and genetic factors in both, audiogram pattern and mean quantity of HL.

In this study, the overall prevalence of presbycusis was 79.1%, being significantly different among age groups, increasing gradually with aging. The most common audiogram configuration was High Frequency Steeply Sloping (HFSS) (51.2%), followed by High Frequency Gently Sloping (HFGS) (29.6%) and FLAT (14.5%). Through cluster analysis techniques it was possible to identify three distinct groups of audiogram patterns. These patterns were significantly associated with gender and noise exposure. Besides the audiogram pattern, the mean quantity of HL, increases with the age of the individuals.

The results suggest the existence of three main audiogram patterns, significantly associated with gender and noise exposure and confirm the positive association between age and HL prevalence or mean amount of HL.

Keywords

Audiogram Configuration; Cluster Analysis; Hearing Loss; Hearing Loss Pattern; Presbycusis

Resumo

A Presbiacusia ou Perda Auditiva Associada ao Envelhecimento é a limitação sensorial mais comum, afetando mais de 30% das pessoas com idade superior a 65 anos. Esta condição tem um impacto negativo na qualidade de vida dos indivíduos, podendo levar ao isolamento social e ao desenvolvimento de doenças neurodegenerativas. Embora existam alguns estudos cujo objetivo tenha sido determinar a prevalência da perda auditiva na população, poucos foram efetuados com o intuito de investigar o padrão de perda ou a configuração do audiograma.

O objetivo deste trabalho consistiu na identificação de padrões dominantes de perda auditiva recorrendo à análise de clusters e construção de um procedimento de classificação com base numa amostra de 321 indivíduos com idade compreendida entre os 62 e 115 anos. A classificação obtida foi validada com recurso à análise de componentes principais e à análise discriminante, e posteriormente, comparada com procedimentos de identificação de padrões descritos na literatura. Por fim, foram ajustados alguns modelos estatísticos com o intuito de investigar a influência de fatores demográficos, ambientais, clínicos e genéticos quer nos padrões determinados, quer na perda auditiva média.

Neste estudo, a prevalência de presbiacusia foi de 79.1% sendo significativamente diferente entre faixas etárias, verificando-se um aumento gradual com o avançar da idade. A configuração do audiograma mais comum foi a HFSS (51.2%), seguida da HFGS (29.6%) e da FLAT (14.5%). Através de técnicas de análise de clusters foi possível identificar a existência de três grupos distintos de padrões de audiograma. A distribuição dos indivíduos em cada um desses grupos foi associada significativamente ao género e à exposição ao ruído. Independentemente do padrão, verificou-se que a perda auditiva média dependia da idade.

Os resultados sugerem a existência de três padrões de presbiacusia, significativamente associados ao género e à exposição ao ruído, e confirmam, a associação positiva existente entre a idade e a ocorrência de perda auditiva ou perda média auditiva.

Palavras Chave

Análise de Clusters; Configuração do Audiograma; Padrão de Perda Auditiva; Perda Auditiva; Pres-

biacusia

Contents

List of Figures

List of Tables

Abbreviations

ARHL Age-Related Hearing Loss

BE Better Ear

CM Cluster Membership

COL Cholesterol

CT Computed Tomography

DF Discriminant Function

DNA Deoxyribonucleic Acid

ENT Ears, Nose and Throat

GLM Generalized Linear Model

GRM7 Glutamate Metabotropic Receptor 7

HFGS High Frequency Gently Sloping

HFSS High Frequency Steeply Sloping

HI Hearing Impairment

HL Hearing Loss

HTA Arterial Hypertension

LDA Linear Discriminant Analysis

LE Left Ear

LFA Low Frequency Ascending

MFU Mid Frequency U-shape

MFRU Mid Frequency Reverse U-shape

MRI Magnetic Resonance Imaging

MSE Mean Square of Error

MSR Mean Square due to Regression

mtDNA mitochondrial DNA

NAT2 N-Acetyltransferase 2

PC Principal Component

PCA Principal Component Analysis

PTA Pure-Tone Average

RE Right Ear

SSE Sum of Squared Errors

SSR Residual Sum of Squares

SST Total Sum of Squares

WE Worst Ear

WHO World Health Organization

WSS Within-Cluster Sum of Squares

1

Introduction

Contents

1.1 Motivation

HL is one of the most common chronic health conditions, being the most prevalent sensory disorder worldwide. Due mainly to the medicine and technology evolution, populations are becoming progressively older and consequently more vulnerable to degenerative diseases. The loss of hearing that gradually occurs in most of individuals as they grow older is named presbycusis or ARHL and it is the main cause of HL.

According to World Health Organization (WHO) statistics, by 2025, there will be approximately 1200 million people in the world over 60 years old, from which 500 million will have presbycusis. It is a multifactorial disorder, characterized by a progressive, symmetrical HL that usually starts in high frequencies of hearing. Essentially due to reduction in the ability to communicate, presbycusis has a negative impact on the quality of life of the individuals being associated, for example, with social isolation.

Although several epidemiological studies have been made to investigate the prevalence of presbycusis aimed to have a measure of the amount of HL, studies on the prevalence of audiogram configurations are few. Thus, the main objective of this study consisted in the identification of individual audiogram patterns, regarding their number and type, based on frequency-specific hearing thresholds in a sample of 321 Portuguese elderly with a clinical indication of presbycusis screened during the period between 2007 and 2016. This approach could be very useful for investigating associations between possible factors that contribute to the etiology and progression of presbycusis, as well could enhance the understanding of pathophysiological inner ear deficits associated with the disorder.

1.2 Thesis Outline

This work is organized in six chapters. In Chapter 1 are presented i) an introduction to the problem and the main objectives of this project and ii) a description of the several chapters. In Chapter 2, it drives the literature review of biologic aspects of auditory system (Section 2.1), as well the epidemiological and clinical aspects of presbycusis (Section 2.2). Section 2.3 presents the main and specific objectives of this work. Chapter 3 starts with the presentation of the study population included in the original project (Section 3.1), whereas in Section 3.2 are introduced the procedures employed in the definition of the study sample and in data collection. Finally, Section 3.3 covered the theorical aspects of statistical methods employed during this work. Chapter 4 presents the results and it is divided in seven sections: 4.1) Sample Description, 4.2) Prevalence of HL, 4.3) Prevalence of Audiogram Shape, 4.4) Principal Component Analysis (PCA), 4.5) Linear Discriminant Analysis (LDA), 4.6) Multinomial Logistic Regression Model and 4.7) Linear Regression Model. The first two sections are related to the description and determination of prevalence of presbycusis in our study sample, respectively. Section 4.3, contains all results for prevalence of audiogram pattern applying either a classification method existing in the literature [3], or hierarchical or K-means clustering techniques. The validation of the K-means classification is presented in Sections 4.4 and 4.5, using PCA and LDA, respectively. The models to investigate the influence of several characteristics of the individuals in both, audiogram pattern and mean quantity of HL are described in Section 4.6 and Section 4.7, respectively. In Chapter 5, a discussion of the results is presented, comparing with the ones existing in literature. The project ends with Chapter 6 where a general conclusion is done.

2

Literature Review

Contents

In human body, the auditory system is responsible for the sense of hearing. It includes both the peripheral auditory system and the central auditory system. This Chapter describes the rough basics of the anatomy and physiology of the auditory system — Section 2.1 — and presents the main definitions and aspects of presbycusis or ARHL — Section 2.2. In Section 2.3 the objectives of this work are described.

2.1 Auditory System and Hearing Loss

2.1.1 Auditory System

The auditory periphery is responsible for the first stage of the transduction of sound in a hearing organism. It is composed by three parts, each one with specific function: the outer ear, the middle ear and the inner ear. The outer ear comprises the pinna and the auditory canal. It captures the sound waves coming from the outside and transmits them to the tympanic membrane which marks the beginning of middle ear. Middle ear is an air-filled structure connected to the nasopharyngeal by the Eustachian tube and it is composed of three small bones or ossicles that act as a lever during the sound transmission process: the malleus, in contact with the tympanic membrane, the incus and the stapes which contacts the cochlea at the oval window. It also contains tiny ligaments and muscles that support and adjust tension of the bony chain. One of the major functions of the middle ear is to ensure the efficient transfer of sound from the air to the fluids in the inner ear: it acts as an impedance-matching device that improves sound transmission, reduces the amount of reflect sound and protects the inner ear from excessive sound pressure levels. The inner ear consists of two main parts: a bony labyrinth and, inside this a membranous labyrinth. The membranous labyrinth is surrounded by the perilymph, a fluid similar to cerebrospinal fluid – rich in sodium and poor in potassium and calcium – and filled by endolymph, a potassium-rich fluid. Is this difference in the electrolyte composition of these two fluids that creates an electrochemical environment that allows the sensorineural transduction. The bony labyrinth consists of three structures: the vestibule, cochlea and semi-circular canals. Only the cochlea is involved in hearing process, while the other two structures belongs to the vestibular system that is involved in sensations of balance and motion. Cochlea is a spiral shaped cavity and it has three sections: the *scala tympani* and the *scala vestibuli* both containing perilymph and the *scala media* or cochlear duct which contains endolymph. All these sections are separated by membranes: Reissner's membrane separates the *scala vestibuli* from the cochlear duct while the basilar membrane separates the cochlear duct from the *scala tympani*. The basilar membrane behaves as a finely tuned band-pass filter, with each location along its length responding to a specific or characteristic frequency: high-frequency sounds localize near the base of the cochlea (near the round and oval windows), while low-frequency sounds localize near the apex. Located on the basilar membrane throughout the entire length of cochlear duct is the sensory organ of hearing, the Organ of Corti. It consists of two types of tiny hair cells that possess large *stereocilia* on their superficial surfaces: the outer hair cells that perform an amplifying role and the inner hair cells that make the transduction of basilar membrane motion into electrical signals and transmit them to the brain via the auditory nerve. It is in the central auditory system that occurs all the signal processing and perception.

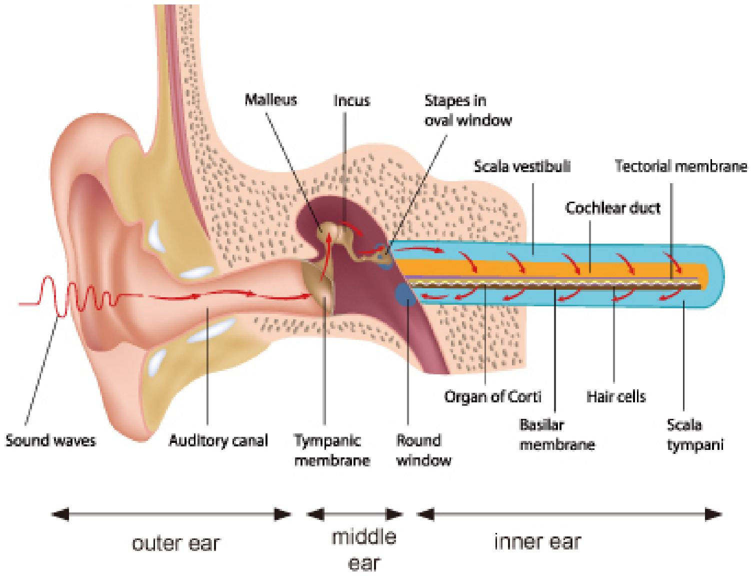

Figure 2.1: Anatomy of the human ear [1].

2.1.2 Hearing Loss

It is clear that hearing plays an essential role in the development of speech and spoken language, being a vital sense for human beings. HL is a major public health problem being one of the most common chronic health condition and the most prevalent sensory disorder worldwide: around 360 million people worldwide is estimated to suffer from disabling HL [4]. Although there is no universal definition accepted, according with WHO [4], a person who is not able to hear as well as someone with normal hearing — hearing thresholds of 25 dB or better in both ears — is said to have some degree of HL. When describing HL there are considered three aspects: type, severity and configuration.

2.1.2.A Type

Depending on the anatomic location of the problem in auditory system it is possible to categorize HL in three basic types: conductive, sensorineural and mixed [5]. Conductive HL indicates an obstruction to air conduction that prevents the efficiently transduction of the sound energy through the auditory canal to the eardrum and the ossicles of the middle ear. Thus, conductive HL may result from problems that affect the external ear or middle ear structures such cerumen obstruction, otitis externa/media, tympanic membrane perforation or damage of the middle ear ossicles. This type of HL can often be treated medically or surgically. Sensorineural HL is used to describe a reduction of auditory threshold sensitivity because sensory receptors of the inner ear are dysfunctional. The pathology may be located in the cochlea and/or the auditory division of the vestibulocochlear nerve. Possible causes of this type of HL are congenital infections, aging and intense noise, and mostly the patients can be habilitated with the use of hearing aids. Finally, mixed HL is a result of the combination of both conductive and sensorineural HL, meaning that there may be a problem in the outer or middle ear, as well in the iner ear or auditory nerve. Other descriptions associated with the type of HL are related to laterality and symmetry. Regarding laterality, if HL occurs in both ears it is named bilateral whereas if only one ear is affected it is named unilateral. Finally, symmetrical HL means that the degree of HL are the same in both ears. When it is different for each ear, HL is said to be asymmetrical.

2.1.2.B Severity

To measure hearing capacity, two parameters must be considered: the frequency or pitch (measured in Hz) and the intensity or loudness of a sound (measured in dB). The degree of HL is quantified for each ear as an indication of severity of the HL. There is no generally accepted definition for the degree of HL and, consequently there are several classifications. In Table 2.1 the definitions according to different authors are presented. In all of them, HL degree is defined on the basis of the average of hearing thresholds (the smallest detectable sound level in a controlled environment) calculated over a certain frequency range – Pure-Tone Average (PTA).

Table 2.1: Commonly accepted hearing threshold ranges for the degree of HL.

Level of HL	**Range (dB)**							
	Goodman [6]	*Clark* [7]	*Loyd et al.* [8]	*Mazzoli et al.* [5, 9]	WHO,*Pascolini et al.* [4, 10]	*Northern et al.* [11]	*Stevens et al.* [12]	*Silverman et al.* [13]
Normal	10-26	-10-15	≤ 25	≤ 20	≤ 25	≤ 15	≤ 19	≤ 26
Slight	27-40	16-25				16-25		27-40
Mild		26-40	26-40	21-40	26-40	26-40	20-34	41-54
Moderate	41-55	41-55	41-55	41-70	41-60	41-65	35-49	55-69
Moderately Severe	56-70	56-70	56-70				50-64	
Severe	71-90	71-90	71-90	71-95	61-80	66-95	65-79	70-89
Profound	≥ 91	≥ 91	≥ 91	≥ 96	≥ 81	≥ 96	80-94	≥ 90

Accordingly with [14], PTA is a good indicator being the gold standard measure for Hearing Impairment (HI).

2.1.2.C Configuration

The configuration of a audiological curve is related to the pattern of HL across frequencies. It can be illustrated in a graph called an audiogram. The general configurations are: flat, sloping, rising and U-shaped. In a flat configuration, hearing thresholds are the same across the speech frequencies. A sloping configuration is characterized by lower hearing thresholds at low frequencies and higher hearing thresholds in high-frequency regions, otherwise in rising configuration there is a better hearing at the higher frequencies. In a U-shaped or trough-shaped configuration the poorest hearing thresholds are in the middle-frequencies. To characterize audiogram configuration, there are several classifications proposed [3, 5, 9, 15, 16].

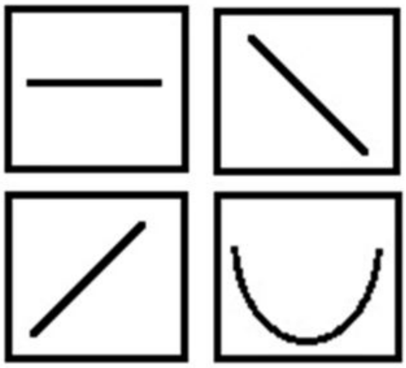

Figure 2.2: General audiogram configurations: flat (top left), sloping (top right), rising (bottom left) and U-shaped (bottom right). Consider represented on the vertical axis the threshold hearing level and on the horizontal axis the frequency.

2.2 Presbycusis or Age-Related Hearing Loss

2.2.1 Epidemiology

The term presbycusis or Age-Related Hearing Loss (ARHL) refers to the loss of hearing that gradually occurs in individuals as they grow old as a result of the physiologic and degenerative processes associated to aging. It is the most frequent cause of HL in the elderly, affecting between 30% to 40% of people older than 65 years of age [17–20]. According with WHO estimates, by 2025 there will be approximately 1200 million people in the world over 60 years, with more than 500 million of individuals who will suffer of presbycusis [21]. The prevalence of presbycusis has been evaluated in large cohorts where audiogram testing was performed (Table 2.2). However, as the different classification of HL, the range of frequencies considered and the criteria to select the study population were not homogeneous, the overall prevalence of HL was different across all studies.

Table 2.2: Overall prevalence of ARHL in different studies.

Reference	n	Age (mean±sd) [range]	HL definition	HL Prevalence (%)
[22]	2448	>50	$PTA_{0.5-4kHz} > 25$ Better Ear (BE)	32.1
[23]	3510	[43-84]	$PTA_{0.5-4kHz} > 25$ either ear	42
[24]	4300	40.8 ±11 [>20]	$PTA_{0.5-8kHz} > 25$ Worst Ear (WE)	18.1
[25]	2956	>50	$PTA_{0.5-4kHz} > 25$ BE	33
[26]	536	[20-69]	$PTA_{0.5-2kHz} > 25$ either ear	26
[17]	3753	65.8 [43-84]	$PTA_{0.5-4kHz} > 25$ WE	45.9
[18]	717	>70	$PTA_{0.5-4kHz} > 25$ BE	63.1
[27]	2765	67.4 [>50]	$PTA_{0.5-4kHz} > 25$ BE	33
[28]	69	[45-93]	$PTA_{0.25-8kHz} > 25$ BE	88.4
[29]	2688	69 [53-97]	$PTA_{0.5-4kHz} > 25$ either ear	51
[19]	3285	49.2±9.9	$PTA_{0.5-4kHz} > 25$ WE	14.1
[30]	548	[72-96]	$PTA_{0.5-4kHz} > 25$ WE	55
[31]	639	[36-90]	$PTA_{0.5-4kHz} > 25$ BE	28.8
[32]	5742	[20-69]	$PTA_{0.5-4kHz} > 25$ either ear	16.1
[33]	110	74.4 ±12.1 [50-96]	$PTA_{0.5,1,2kHz} > 25$ either ear	61.5

PTA: average of hearing thresholds at a certain frequency range.

Example: $PTA_{0.5-4kHz}$ is the average of hearing thresholds at 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz.

2.2.2 Physiopathology

Age-related changes occur in a variety of anatomic, physiologic and auditory functional abilities, either in peripheral or central auditory systems. Schuknecht [2], based on results of audiometric tests and temporal bone histology has classified presbycusis phenotype in six subtypes: sensory, neural, strial or metabolic, cochlear conductive, mixed and indeterminate [34, 35]. According to these results, it was possible to correlate the pathologic findings with audiological clinical manifestations, namely audiogram configuration.

Sensory ARHL refers to degeneration of the organ of Corti that is primarily caused by damaged outer hair cells. It is characterized by an audiogram that is abnormal only at the highest frequencies. Neural ARHL refers to the loss of radial afferent neurons in the cochlea showing a gradual loss of hearing with a slightly greater loss at higher frequencies and a severe decrease in speech discrimination. Strial or metabolic presbycusis shows HL across all frequency range in audiogram being caused by the atrophy of the *stria vascularis* usually in the mid-cochlear to apical regions. The pattern of HL is “flat” with all frequencies affected similarly, including the lower frequencies. Regarding cochlear conductive presbycusis, it has not been verified yet but, it is described as a degenerative change resulting from the stiffness of the basilar membrane of the cochlea and it is manifested by a linearly descending audiogram (greater than 50 dB decline overall) that appeared unexplained by obvious degeneration of any cochlear cells or structures. Finally there were described two more subtypes of presbycusis: mixed which refers to a combination of the above types of presbycusis and indeterminate which corresponds to all cases of presbycusis that show none of the above characteristics.

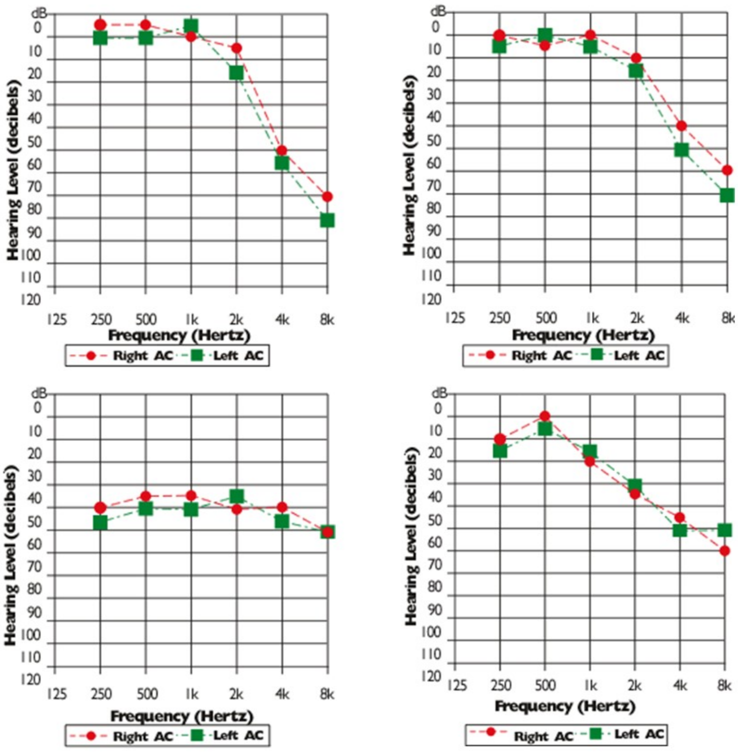

Figure 2.3: Typical audiogram curves of presbycusis according to Schuknecht [2]: sensory (top left), neural (top right), strial (bottom left) and cochlear conductive (bottom right).

2.2.3 Clinical Presentation, Assessment and Treatment

By definition, presbycusis is a bilateral symmetrical progressive sensorineural HL that happens over many years. Although its clinical presentation and progress can be variable, usually it starts affecting the high frequencies of hearing (above 2 kHz) [17, 24, 36–38]. The missing high frequencies are essential to communication, particularly in situations with background noise. Over the time, the loss progresses to the mid and low frequencies (0.5-2 kHz) that are associated with human speech. This frequency range includes most of the voiceless consonants (t, p, k, f, s and ch). As a result of their HL pattern, patients will report being able to hear someone is speaking, but not being able to understand what is being said [17, 29, 39]. Other symptoms that can be involved are tinnitus (sensation of hearing a sound when no external cause is present) [28, 40–42] and dizziness [28, 42]. The impact of HL may lead to adverse effects on elderly quality of life [29, 43] with consequences for the social, functional and psychological well-being. As an example, social isolation [29, 43, 44], depression and anxiety

[45] or neurodegenerative disorders, such as dementia [31, 36] are described in elderly who suffered presbycusis.

The diagnosis of presbycusis should be made based on medical history, physical examination and testing procedures [46]. It is done by audiologists who have expertise in hearing testing, using assistive listening devices and also by otolaryngologists or Ears, Nose and Throat (ENT) clinicians who have specialty training in a range of disorders in the head and neck. Prior to making the diagnosis of presbycusis, through hearing history it is possible to understand potential risk factors involved in HL etiology, such as family history, ototoxic medications and noise exposure. Regarding physical procedures, they usually include otoscopy examination of outer ear to inspect obstructions, infections, congenital malformations and other lesions. The Weber and Rinne tests require a tuning fork and can help to differentiate conductive from sensorineural HL, but accordingly with [47] should not be used for general screening. To obtain a valid and formal measure of HL and to confirm the diagnosis the last phase of HL assessment includes audiologic and advanced tests. The audiologic tests include pure-tone audiometry, speech audiometry, word recognition score tests, tympanometry and acoustic reflexes [46, 48]. Pure-tone audiometry is performed using an audiometer that delivers sounds of specific frequencies at different intensities to determine the patient's hearing threshold (lowest intensity in dB at which this tone is perceived 50% of the time) at each tested frequency. Hearing in each ear is tested from low (125 Hz or 250 Hz) to high frequencies (generaly, 8000 Hz) and the hearing thresholds are recorded on a graph called an audiogram (Figure 3.2). This type of test enables the determination of the degree, type and configuration of HL being the gold standard for its assessment [14]. Sometimes advanced testing, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) is needed to evaluate other unknown possible causes for HL or to confirm the diagnosis.

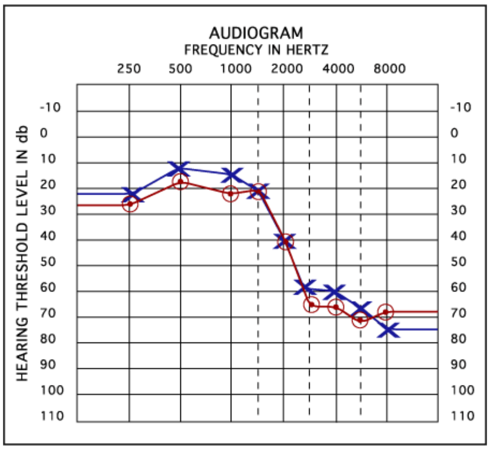

Figure 2.4: Audiogram for left (blue) and right (red) ears: in vertical axis is represented the hearing thresholds (dB) and in horizontal axis is represented the frequencies (Hz).

At present, presbycusis is irreversible and a directed treatment to prevent or reverse its effects is not available [48]. However, there are multiple options that can attenuate or compensate HL in order to improve daily function and well-being [21]. Depending on the severity and type of HL there are different treatment options. Hearing amplification with the use of hearing aids is an effective treatment for people with mild-to-severe HL being associated with improvements in the social, emotional, psychological, and physical well-being [35, 45]. Hearing aids may not allow the patient to understand speech any better, but rather allows only the patient to hear noise at a louder level. Despite hearing aids offer potential help, only a small percentage with presbycusis actually receive effective treatment with amplification [18, 49, 50]. When the conventional amplification strategies are no longer effective due the increased severity of HL, the placement of a cochlear implant replaces the function of the damaged inner ear to provide sound signals to the brain. Is expected that the majority of patients undergoing cochlear implantation achieve significant functional improvement [51, 52]. In order to reduce HL induced deficits of function and activity, such as reduction of communication abilities, which can lead to a negative impact on their quality of life, patients may resort to auditory rehabilitation. It includes interventions such as active listening training, speech reading and communication enhancement [53, 54].

2.2.4 Etiology and risk factors

Presbycusis is a multifactorial disorder being associated with age-related degeneration of auditory structures or nerves, environmental, genetic predisposition and medical conditions [17, 55]. Among non-modifiable risk factors, increasing age [17, 18, 24, 25, 27, 56, 57] and male gender [17–19, 25, 27, 57, 58] were associated with an increasing risk of HL, while African race was a protective factor [18, 30, 32, 59]. Genetic predisposition as shown by heritability studies [23, 58] indicate that genetic phenotype accounts for a substantial portion of HL risk. Some studies have been done to identify some candidate genes related to HL. An association between N-Acetyltransferase 2 (NAT2), a detoxification enzyme, and ARHL has been found in [60, 61]. However, more recently [62] concluded that there is no evidence to support that NAT2 polymorphism are associated to ARHL. Moreover, associations with presbycusis have been reported for Glutamate Metabotropic Receptor 7 (GRM7) [63]. Also, mutations in mitochondrial DNA (mtDNA) are associated with various forms of HL [27, 64]. Beyond medical conditions factors, hypertension and cardiovascular disease [17, 26, 32, 56, 65], diabetes [17, 18, 22, 26, 32, 66–68] and atherosclerosis [69], head injury [56] and peripheral neuropathy [26] were related with an increased risk of HL. Among environmental factors, tobacco smoking [20, 22, 69, 70], noise exposure [17, 19, 22, 37, 70, 71], chemicals [72–74], and ototoxic medications [28, 75–77] were associated with increasing HL. Regarding alcohol consumption and level of education the results are inconsistent. [56, 77, 78] showed no effect of alcohol in HL, whereas [20] reported alcohol consumption as a protective factor for HL. Low education level was associated with an increase of HL [17, 19]. In fact, individuals with lower levels of education may tend to work in occupations that have higher levels of noise exposure, such as labor or manufacturing. On the other hand, accordingly with [56] level of education was no effect on HL.

2.3 Objectives

The main objective of this work aimed to identify, regarding the number and type, presbycusis patterns based on audiometric data collected from a sample of 321 Portuguese elderly with a clinical indication of presbycusis during the period between 2007 and 2016. For that unsupervised classification techniques, namely hierarchical and K-Means clustering, as well an audiometric configuration classification proposed by Wuyts [3] were employed.

The specific aims of the study included:

- Characterization of the obtained groups according to audiological, demographic, environmental, medical and genetic conditions;
- Build models to study the association between both, audiogram patterns or mean quantity of HL with demographic, environmental, medical and genetic factors.

Moreover, these aspects were investigated:

- Differences between RE and LE;
- Prevalence of HL;

3
Material & Methods

Contents

The following chapter presents the methods involved in the finding of the presbycusis patterns. The aim of this chapter was to indicate the study population — Section 3.1 — and the final sample chosen, as well the instruments employed during the data collection — Section 3.2. In addition, Section 3.3 covered the statistical methods used across all steps of the project.

3.1 Study Population

This study is a part of the "Age-related hearing loss: Genetic risk factors and social impact" project. The aim of this project was to investigate the epidemiological and etiological factors associated with ARHL. For the project, subjects from Coimbra, Egas Moniz and Santa Maria Hospital's, as well from some social centers with a clinical indication of presbycusis were recruited. To select the participants to be included in the study, all volunteers underwent a questionnaire related to exclusion criteria. Based on that, subjects with ear diseases or any other pathologies reported to potentially influence hearing were excluded. All participants gave written informed consent and the study was approved by the Ethics Committees of all hospitals and social centers (when applicable).

3.2 Sample and Data Collection

3.2.1 Sample definition

The volunteers who passed the medical exclusion criteria underwent a clinical examination including otoscopy and completed an extended questionnaire on medical history and environmental exposure. The audiological examination was performed by a trained audiological assistant in a sound insulated booth using a portable audiometer for patients from senior residences, the ones from hospital were evaluated at audiology rooms. Pure-tone air conduction thresholds were measured at 250, 500, 1000, 2000, 4000 and 8000 Hz. Audiological exclusion criteria was: unknown pure tone thresholds at least one frequency. Additionally, 9 individuals with unknown age and 4 individuals under 60 years old were excluded. To satisfied audiological exclusion criteria, 123 individuals were excluded remaining 321 (70.2%) (Figure 3.1). Audiological data collection occurred in the period between 2007 and 2016.

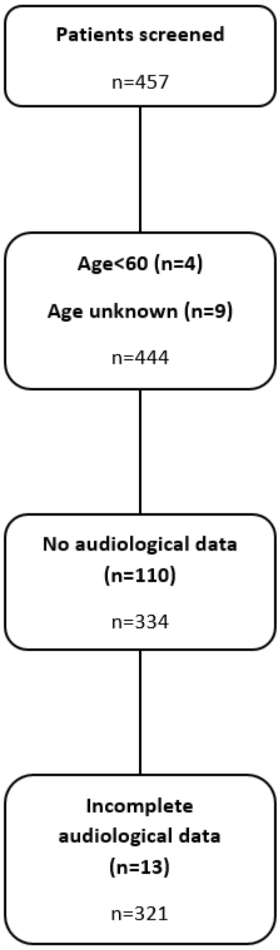

Figure 3.1: Definition of study sample.

3.2.2 Database

A database in excel format was provided by the research team. For the participants, information regarding biodemographic, social, clinical and genetic characteristics was collected. Prior to statistical analysis the entire database was cleaned and organized. Table 3.1 contains the description of all variables included in the database:

Table 3.1: Variables included in the study.

Variable	Description	Type	Categories
id	Patient identification	PREXXX XXX: number with 3 digits	
sexo	Patient gender	Categorical (Nominal)	M: Female H: Male
idade	Patient age when audiological assessment was performed	Continuous	
idade_cat	Patient age when audiological assessment was performed categorized by decades	Categorical (Ordinal)	60: 60-69 years-old 70: 70-79 years-old 80: 80-89 years-old 90: 90-99 years-old 100: 100-109 years-old 110: 110-119 years-old

Continued on next page

Table 3.1 – continued from previous page

Variable	Description	Type	Categories
F250_OD F500_OD F1000_OD F2000_OD F4000_OD F8000_OD	Hearing intensity thresholds measured at 250, 500, 1000, 2000, 4000 and 8000 Hz in the Right Ear (RE) (dB)	Continuous	
F250_OE F500_OE F1000_OE F2000_OE F4000_OE F8000_OE	Hearing intensity thresholds measured at 250, 500, 1000, 2000, 4000 and 8000 Hz in the Left Ear (LE) (dB)	Continuous	
NAT2F	NAT2 phenotype	Categorical (Nominal)	I R S
GRM7	GRM7 genotype	Categorical (Nominal)	A/A A/T T/T
MTDNA	mtDNA haplotype	Categorical (Nominal)	H HV I J K L M N R T U V W X Y
prorui	Noise exposure	Categorical (Nominal)	N: No S: Yes
histfam	Family history of hearing problems	Categorical (Nominal)	N: No S: Yes
medic	Ototoxic medication history	Categorical (Nominal)	N: No S: Yes
zumb	Tinnitus symptoms indicator	Categorical (Nominal)	N: No S: Yes
HTA	Hypertension problems indicator	Categorical (Nominal)	N: No S: Yes
COL	Cholesterol problems indicator	Categorical (Nominal)	N: No S: Yes

3.2.3 Audiological Criteria

As mentioned in the previous chapter there is no definite definition for the HL degree. For the purpose of this study, as recommended by WHO [4, 9], the degree of HL was defined on the basis of average hearing thresholds over the frequencies 0.5, 1, 2, and 4 kHz ($PTA_{0.5,1,2,4kHz}$) on the BE. When justified,

$PTA_{0.5,1,2,4kHz}$ was also determinated for the RE or LE. The severity of HL was based on WHO [4] definition (Table 3.2). A HL was considered asymmetrical if the difference between the left and right ear air conduction thresholds was 20 dB or more for at least two frequencies out of 0.5, 1, and 2 kHz. Unilateral HL occurred if HL was only in one of the ears, what is not expected for presbycusis as physiologic condition.

Table 3.2: HL degree classification according with WHO.

Level	Range (in dB)
Normal	$\leq$25
Mild	25 $<$ db HL $\leq$40
Moderate	40 $<$ db HL $\leq$60
Severe	60 $<$ db HL $\leq$80
Profound	$>$80

3.2.4 Audiogram Shape Identification

One of the approaches used to determine audiogram configurations was based on a classification criteria used in previous studies [41, 79] where both ears were categorized into six accordingly with audiometric configuration classification of Wuyts [3], based on earlier studies [15, 80]. In this classification a Flat (FLAT) audiogram configuration is defined as an audiogram where the difference between the mean of 250/500 Hz hearing thresholds, the mean of 1/2 kHz hearing thresholds and the mean of 4/8 kHz hearing thresholds, is less than 15 dB. A High Frequency Gently Sloping (HFGS) audiogram configuration is defined as an audiogram where the difference between the mean of 500 Hz/1 kHz hearing thresholds and the mean of 4 kHz/8 kHz hearing thresholds is greater than 15 dB and less than 29 dB. A High Frequency Steeply Sloping (HFSS) audiogram configuration is defined as an audiogram where the difference between the mean of 500 Hz/1 kHz hearing thresholds and the mean of 4 kHz/8 kHz hearing thresholds is greater than 30 dB. A Low Frequency Ascending (LFA) audiogram configuration is defined as an audiogram where the difference between the poorer low frequency hearing thresholds and better high frequency ones is greater than 15 dB. A Mid Frequency U-shape (MFU) audiogram configuration is defined as an audiogram where the difference between the poorest hearing thresholds in the mid-frequencies and those at higher and lower frequencies is greater than 15 dB. A Mid Frequency Reverse U-shape (MFRU) audiogram configuration is defined as an audiogram where the difference between the best hearing thresholds in the mid-frequencies and those at higher and lower frequencies is greater than 15. These audiogram configurations are represented in Figure 3.2.

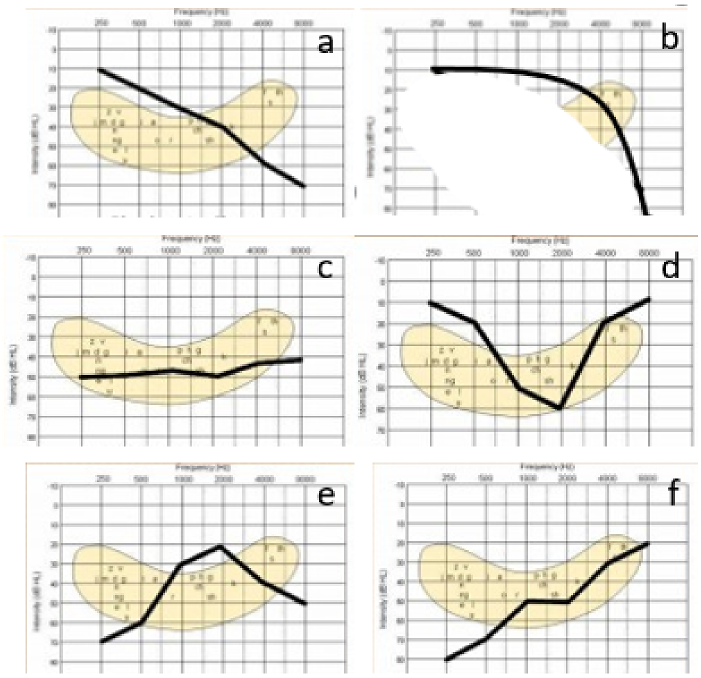

Figure 3.2: Audiogram configurations based on Wuyts classification: a) HFSS, b) HFGS, c) FLAT, d) MFU , e) MFRU and f) LFA.

3.3 Statistical Analysis

Additionally to the classification criteria presented in Subsection 3.2.4, in order to investigate if the individuals could be grouped in a “natural” way a cluster analysis was performed. It was mainly based on the partition of the audiological data set, recurring to hierarchical clustering or K-means techniques. A Principal Component Analysis (PCA) with an exploratory prospective was used to detect possible groups and a Linear Discriminant Analysis (LDA) was implemented with the aim of construct an allocation rule to each one of the groups found. Both procedures were also performed as a validation of the results obtained by cluster analysis. Finally, Generalized Linear Models (GLMs) were adjusted to the audiological data in order to investigate the influence of the variables in both audiogram pattern and amount of HL.

3.3.1 Cluster Analysis

Cluster analysis is an unsupervised classification method that allows the partitioning of a set of data observations into subsets — named clusters — such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. It usually follows a hierarchical strategy or an iterative reallocation procedure.

The data set containing n subjects to be clustered and p variables is usually represented by a data matrix $n{\times}p$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & ... & x_{1p} \\ ... & ... & ... \\ x_{n1} & ... & x_{np} \end{bmatrix} \tag{3.1}$$

where, the x_{ij} element represents the observation of the subject i in variable p. Based on this matrix, a similarity matrix $\mathbf{D}$, $n \times n$, is calculated, where the generic element (i, j) is a similarity measure between the subject i and subject j, d_{ij}. To quantify the similarity between two subjects (i, j) a quantitative distance measure, a similarity or a dissimilarity measure, d_{ij}, can be used. This measure reflects the bigger or smaller differences between the values that these individuals have in the set of p variables. In case of a distance measure it must satisfy the following properties:

- $d_{ij} \geq 0$
- $d_{ii} = 0$
- $d_{ij} = d_{ji}\ (symmetry)$
- $d_{ij} \leq d_{ik} + d_{jk}\ (triangular\ inequality)$

Some of widely used measures for distance between individuals to quantitative data are:

- Euclidean Distance:

$$d_{ij} = \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\| \tag{3.2}$$

  where, $\mathbf{x}_{(i)}$ and $\mathbf{x}_{(j)}$ are the correspondent vectors of the subjects i and j.

- Generalized Euclidean distance:

$$d_{ij} = \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\|_w = \sqrt{\left(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\right)^T \mathbf{W} \left(\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\right)} \tag{3.3}$$

  where, $\mathbf{W}$ is a defined positive matrix. If $\mathbf{W} = \mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma}$ is the covariance matrix of the variables, the distance is known as Mahalanobis distance which is caracterized by being invariant to changes in the scale of the variables.

- Minkowski distance:

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda}\right)^{1/\lambda} \tag{3.4}$$

  When $\lambda = 1$, is designed Manhattan distance. As higher is the value of λ, higher is the weight of subjects very dissimilants.

- Canberra distance:

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})} \tag{3.5}$$

  Generally, this distance is used in non-negative variables.

Another aspect that should be considered in a cluster analysis and it also becomes necessary is how to measure the proximity between cluster groups of individuals in order to aggregate them in the same cluster. Considering K and G clusters, the most used aggregation methods are:

- Nearest Neighbour or Single-Linkage:

$$D_{GH} = \min\{d_{kl}, k \in G, l \in H\} \tag{3.6}$$

The distance between two clusters is the lower distance among an element of a cluster and an element of another cluster.

- Furthest Neighbour or Complete-Linkage:

$$D_{GH} = \max\{d_{kl}, k \in G, l \in H\} \tag{3.7}$$

The distance between two clusters is the higher distance among an element of a cluster and an element of another cluster.

- Average Linkage:

$$D_{GH} = \frac{1}{n_G n_H} \sum_{k=1}^{n_G} \sum_{l=1}^{n_H} d_{kl} \tag{3.8}$$

The distance between two clusters is the average distance of all distances among pairs of elements (one in each cluster group).

- Minimum Variance Method (Ward Criteria):

The objective is to generate clusters in order to minimize the quadratic sum of errors. Considering the inertia of cluster G, i.e., the quadratic sum of differences among each individual and the "mean individual" of this cluster:

$$I_G = \sum_{l=1}^{p} \left[\sum_{k \in G} \left(x_{kl} - \overline{x_l^G} \right)^2 \right] \tag{3.9}$$

where $\overline{x_l^G}$ is the average of the values of variable l with respect to the individuals of cluster G. The distance between G and H will be the increase in the total sum of inertias due to fusion of G and H cluster groups:

$$D_{GH} = \frac{n_G n_H}{n_G + n_H} \|\overline{x_G} - \overline{x_H}\|^2 \tag{3.10}$$

- Centroid Method:

The distance between two clusters is the distance between the gravity centers (centroides) of the clusters:

$$D_{GH} = \|\overline{x_G} - \overline{x_H}\| \tag{3.11}$$

Depending on the chosen aggregation criteria and similarity measure, different clustering maybe obtained. Single Linkage and Complete linkage tend to be overly sensitive to outliers or noisy data. The use of average distance is a compromise between the minimum and maximum distances and overcomes the outlier sensitivity problem [81]. Ward method presents good results for all distance metrics, is sensitive to outliers and tends to combine groups with few elements. Although centroid method is robust to noise data, the results using Euclidian distance are not good in the presence of noise.

One important step in cluster analysis is the determination of the optimal number of clusters. A variety of methods have been suggested, most of them are rather informal and subjective. To overcome this problem, more formal methodologies were proposed by [82] and more recently by [83]. In this work the clustering validity was based on a relative criteria which consists in the evaluation of a clustering structure by comparing it with other clustering schemes, resulting from the same algorithm but with different parameter values, as the number of clusters. For that it was used the R package NbClust that provides 30 indices which determine the number of clusters in a data set and offers also the best clustering scheme from different results to the user [84].

3.3.1.A Hierarchical Clustering

In a hierarchical classification, the data are not partitioned into a particular number of classes or clusters at a single step [85]. Strategies for hierarchical clustering can be further subdivided in agglomerative or divisive methods. Due to high computational complexity of the latter procedure which can lead to inaccurate results, agglomerative approach is probably the most widely used of the hierarchical methods. Agglomerative methods start with a partition of the data, usually with each object forming a separate cluster. At each step, the objects or clusters that are more similar to one another are merged forming a new partition. To represent the process of hierarchical clustering, a tree structure called a dendrogram is commonly used. In this work an agglomerative hierarchical clustering was employed, where the distance measure used to calculate the similarity matrix was the Euclidian distance and the aggregation criteria used to group the most similar clusters of individuals was the Ward method. The steps of the analysis were the following:

1. Data set with n individuals and p variables (frequencies);
2. n clusters each one with one individual;
3. Calculation of similarity matrix $D(n)$;
4. Choose and group of the two most similar individuals, forming a new cluster;
5. Recalculation of the distances between the new cluster and the remaining individuals using the aggregation criteria (Ward criteria). A new similarity matrix $D(n-1)$ is obtained;
6. Repetition of steps (4) and (5), until all individuals are grouped in one cluster.

3.3.1.B Partitioning Methods

Partitioning methods are non-hierarchical classification procedures aimed is to determine a classification for n individuals in k classes in order to optimize an objective partitioning criterion, such as a dissimilarity function based on distance. Partitioning clustering decomposes a data set into a set of disjoint clusters. The most well-known and commonly used partitioning method is K-Means [86]. In K-means algorithm, clusters are represented by centroids which are defined as the average of the objects within the cluster. The main goal of K-means is to find the best partition of the objects in a predefined number of groups, so that the total distance from each object to its corresponding centroid is minimized. Formally, having n objects the goal is to obtain a partition of k sets as compact and as separate as possible in order to minimize the Within-Cluster Sum of Squares (WSS), defined as:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^j - c_j\|^2 \tag{3.12}$$

Where $\|x_i^j - c_j\|$ provides the distance between an individual and the cluster's centroid. In this study the steps of the algorithm were the following [81]:

1. Randomly selection of k observations of the data set using these as initial cluster means;
2. Based on Euclidian distance between the subjects and the cluster center, the remaining subjects are assigned to the most similar cluster;
3. Iteratively improvement of the within-cluster variation and computation of the new mean for each cluster using the subjects assigned to the cluster in the previous iteration;
4. All the subjects are then reassigned using the updated means as the new cluster centers;
5. The process continues until the convergence, that is, the clusters formed in the current round are the same as those formed in the previous round or the maximum number of allowed iterations is reached.

3.3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate data analysis method used primarily as a dimensionality-reduction technique. The basic aim is to describe the variation in a set of correlated variables x_1, x_2,..., x_p by an orthogonal transformation in terms of a new set of uncorrelated variables y_1, y_2,..., y_p each of which is a linear combination of the x variables named Principal Components (PCs). The general hope of PCA is that the first few components will account for a substantial proportion of the variation in the original variables and can, consequently, be used to provide a convenient lower-dimensional summary of these variables that might prove useful for a variety of reasons [87]. PCA essentially rotates the data (via a linear transformation) so that most of the variability in the data is contained in as few dimensions as possible. Assuming that the random vector $\mathbf{X} = [x_1, x_2, \ldots, x_p]^T$ has mean μ and covariance matrix $\Sigma_{p\times p}$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$, the PCs are linear combinations of p variables X_i, $i = 1, \ldots, p$:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p = \mathbf{a}_j^T\mathbf{X},\ j = 1, \ldots, p \tag{3.13}$$

where $\mathbf{a}_1$,$\mathbf{a}_2$,...,$\mathbf{a}_p$ are the p normed eigenvectors associated to the p eigenvalues of Σ, $\lambda_1 > \lambda_2 > ... > \lambda_p$ and $Var\,(Y_j) = \lambda_j$, $j = 1, \ldots, p$.

To determine the PCs, the eigenvalues and the correspondent normed eigenvectors of empiric covariance matrix are calculated. In many situations, the variables are not in the same unit or have diverse variances. In such cases, it is performed a standardization of the variables which is equivalent to use the correlation matrix instead of the covariance matrix to calculate the PCs. The coefficients of the linear combinations, a_{ij}, are estimated in order to have in the first PC the higher variance. The relative importance of a variable X_i to the explanation of the PC, Y_j, is given by a_{ij}^2. A PC is more important as higher is the proportion of total variance it explained by it, i.e., as higher is the value of $\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\lambda_j}{tr(\Sigma)}$. In case of correlation matrix $\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\lambda_j}{p}$. One major aspect in PCA is the number of components that should be retained to provide an adequate summary of a given data set. The most common procedures that have been suggested are the following:

- Retain just enough components to explain some specified, large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested [87].
- Kaiser's Rule: When dealing with the PCA of a sample correlation matrix, Kaiser [88] suggested that only those PCs whose eigenvalues exceed unity (or arithmetic mean in case of sample covariance matrix) should be retained. This rule is popular but controversial, once there is evidence that the cutoff value of 1 is too high [89]. A modified rule proposed by Jolliffe [90], concluded that a more appropriate procedure would be to exclude components extracted from a correlation matrix whose associated eigenvalues are less than 0.7 [87].
- Scree-Plot: The sample eigenvalues from a PCA are ordered from largest to smallest. It is usual to plot the ordered sample eigenvalues against their order number, such a display is called a "scree-plot" [91], after the break between a mountainside and a collection of boulders usually found at its base. If the largest few sample eigenvalues dominate in magnitude, with the remaining sample eigenvalues very small, then the scree-plot will exhibit an "elbow" in the plot corresponding to the division into "large" and "small" values of the sample eigenvalues. It is usually recommended to retain those PCs up to the elbow and also the first PC following the elbow [89].

The next step in a PCA is the interpretation of the PCs. The meaning and the importance of a variable to the PC can be interpreted by the coefficients of linear combinations (a_{ij}), the correlation of the original variables and the PCs (ρ_{ij}) or the loadings (l_{ij}), given by equations 3.14 and 3.15, respectively:

$$\rho_{ij} = \frac{a_{ij}\sqrt{\lambda_j}}{\sigma_i}, \tag{3.14}$$

$$l_{ij} = a_{ij}\sqrt{\lambda_j} \tag{3.15}$$

When using the correlation matrix, $\sigma_i = 1$, which implies $\rho_{ij} = l_{ij}$. Thus if the absolute value of a coefficient of a PC for a given variable is high, the correlation between this PC and this variable will be also high. The variables that should be used in the interpretation of a given PC are the ones that respect the $|\rho_{ij}| \geq \sqrt{\frac{\lambda_j}{p}}$ or, simply, $|\rho_{ij}| \geq 0.5$.

For each subject i it is obtained an observed value to the j PC named score and given by:

$$y_{ij} = a_{1j}x_i1 + a_{2j}x_i2 + \ldots + a_{pj}x_ip,\ i = 1, \ldots, n, j = 1, \ldots, p \tag{3.16}$$

In this study, PCA was used in audiogram data set in an exploratory way in order to help in the detection of groups, and also as a validation method for cluster analysis.

3.3.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised learning multivariate technique concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. The first step consists on discrimination or separation to find characteristics that allows the allocation of objects in different predefined groups. The classification can be defined as a set of rules that will be used to allocate new objects [92]. The main goal is to obtain functions that can be able to classify an observation in one of different populations π_i $(i = 1, \ldots, g)$ based on measures of a number p of features, in order to minimize the possibility of misclassifying cases into their respective groups or categories [93]. In order to obtain a classification rule the following assumptions should be done to the discriminant function model:

- The distribution of the g populations is Normal Multivariate;
- π_i priori probabilities of population occurrence are equal and $\sum_{i=1}^{g} \pi_i = 1$;
- The populations have the same misclassification cost.

Considering that the g populations follow a multivariate normal distribution the linear discriminant function for the $i-th$ population of a random vector $\mathbf{X}$ is given by:

$$\delta_i(\mathbf{x}) = -\frac{1}{2} ln|\Sigma_i| - \frac{1}{2}[\mathbf{x} - \mathbf{u}_i]^T \Sigma_i^{-1} [\mathbf{x} - \mathbf{u}_i] + ln(\pi_i),\ i = 1, \ldots, g \tag{3.17}$$

where Σ_i is the covariance matrix of population i, $\mathbf{x}$ is the vector of features, $\mathbf{u}_i$ is the vector of means of population I and π_i is the vector of probabilities of occurrence of population i. Supposing the equality of covariance matrices Σ=Σ_i, the constant terms can be excluded and the function is thus given by equation 3.18:

$$\delta_i(\mathbf{x}) = \frac{1}{2}[\mathbf{x} - \mathbf{u}_i]^T \Sigma^{-1} [\mathbf{x} - \mathbf{u}_i] + ln(\pi_i),\ i = 1, \ldots, g \tag{3.18}$$

If we explicitly expand the quadratic matrix-vector expression, we can simplify it as:

$$\delta_i(\mathbf{x}) = \mathbf{u}_i^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}\mathbf{u}_i^T \Sigma^{-1} \mathbf{u}_i + ln(\pi_i),\ i = 1, \ldots, g \tag{3.19}$$

Only the first term of the right hand side of the equation depends on $\mathbf{x}$, thus two new variables $\mathbf{w}_i^T = \mathbf{u}_i^T \Sigma^{-1}$ and $w_{i0} = -\frac{1}{2}\mathbf{u}_i^T \Sigma^{-1} \mathbf{u}_i + ln(\pi_i) = -\frac{1}{2}\mathbf{w}_i^T \mathbf{u}_i + ln(\pi_i)$ can be defined. Replacing them in the equation 3.19 we obtain:

$$\delta_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0},\ i = 1, \ldots, g \tag{3.20}$$

The classification rule to allocate an observation $\mathbf{x}$ is to classify $\mathbf{x}$ in π_i if and only if:

$$\hat{\delta}_i(\mathbf{x}) = argmax\left(\hat{\delta}_1(\mathbf{x}), \hat{\delta}_2(\mathbf{x}), \ldots, \hat{\delta}_g(\mathbf{x})\right) \tag{3.21}$$

3.3.4 Multiple Regression Model

Regression analysis is a statistical methodology that describes the relationship between a response or dependent variable and one or more independent variables. Depending on the aim of the model, the independent variables can be named by explanatory or predictor variables. Since one of the objectives of this work is to predict the mean amount of HL, independent variables will be called predictors. Denoting the response variable by Y and the set of predictor variables by X_1, X_2,...,X_p, where p represents the number of predictor variables, this relationship can be expressed by a multiple linear regression model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.22}$$

where $\boldsymbol{\beta}$ is the vector of the regression parameters or unknown coefficients to be determined and $\boldsymbol{\varepsilon}$ is the vector of random errors, such that each individual error has a Normal distribution with mean $E[\varepsilon_i] = 0$ and variance $V[\varepsilon_i] = \sigma^2$. For all $i \neq j$, ε_i and ε_j are uncorrelated so that their covariance is zero ($Cov[\varepsilon_i, \varepsilon_j] = 0$). According to equation 3.22 each observation can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots n \tag{3.23}$$

where y_i represents the $i-th$ observation of the response variable Y, ε_i represents the error term for this observation and x_{i1},x_{i2},...,x_{ip} are the predictor variables.

Once defined the model, the next step is the estimation of the parameters. The most commonly used method is the least squares method that minimize the Sum of Squared Errors (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \tag{3.24}$$

Denoting the vector of the least square estimators as $\hat{\boldsymbol{\beta}}$, they can be obtained solving equation 3.25:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}^T \mathbf{Y} \tag{3.25}$$

Based on the properties of multivariate normal distribution:

- $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $V[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}$;
- Each individual estimator $\hat{\beta}_j$ is normally distributed with $E[\hat{\beta}_j] = \beta_j$ and $V[\hat{\beta}_j] = \sigma^2 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}_{(j+1,j+1)}$;
- Different estimators are not independent, having a covariance matrix with terms $Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}_{(j+1,j+1)}$

An unbiased estimator for σ^2 is then given by:

$$MSE = \hat{\sigma}^2 = \frac{SSE}{n-p-1} \tag{3.26}$$

where MSE is the Mean Square of Error (MSE) and $W = \dfrac{SSE}{\sigma^2} \sim \chi^2_{n-(p+1)}$ and $\hat{\boldsymbol{\beta}}$ and SSE are distributed independently of each other.

After fitting the linear model to a given data set, an assessment is made regarding adequacy of fit. The coefficient of multiple determination R^2 may be interpreted as the proportion of the total variability in the response variable Y associated with the use of the set of predictor variables $X_1, X_2, \ldots, X_p$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \tag{3.27}$$

where SST is the Total Sum of Squares (SST) and SSR is the Residual Sum of Squares (SSR). The coefficient, R^2, can assume values between 0 and 1, being that it will be near to 1 when observed and predicted values are close to each other, what make increase the value of SSE. Once adding more variables to the regression model will increase R^2, an modified measure that adjusts for the number of X variables (and hence parameters) in the model is often suggested:

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST}/(n-1) = 1 - \frac{MSE}{SST/(n-1)} \tag{3.28}$$

In a multiple linear regression a model is inadequate if all predictors variables have a null coefficient. It is equivalent to test:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0 \tag{3.29a}$$

$$H_1 : \exists j = 1, \ldots, p : \beta_p \neq 0 \tag{3.29b}$$

Being the statistic test, under H_0, given by:

$$F = 1 - \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \sim F_{p,n-p-1} \tag{3.30}$$

where MSR is the Mean Square due to Regression (MSR). Accordingly, H_0 is rejected at the significance level α if $F > F_{(p,n-p-1,\alpha)}$.

Based on the properties presented above, it will be possible to make statistical inference regarding the regression coefficients. To investigate the influence of each predictor in the response variable is tested:

$$H_0 : \beta_j = 0 \tag{3.31a}$$

$$H_1 : \beta_j \neq 0 \tag{3.31b}$$

using the following test statistic:

$$T = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\beta j}} \sim t_{n-p-1} \tag{3.32}$$

with $\hat{\sigma}_{\beta j} = \hat{V}[\hat{\beta}_j] = \sqrt{\hat{\sigma}^2 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}_{(j+1,j+1)}} = \sqrt{MSE \left(\mathbf{X}\mathbf{X}^T\right)^{-1}_{(j+1,j+1)}}$. Accordingly, H_0 is rejected at the significance level α if $|t| > t_{(n-p-1,\alpha/2)}$ where t is the observed values of the test statistic. It would mean that the variable X, is a statistically significant predictor of the response variable Y after adjusting for the other predictor variables. An interval with $(1-\alpha)\times 100\%$ of confidence for each β_j can be obtained by:

$$\hat{\beta}_j \pm t_{(n-p-1,\alpha/2)} \times \hat{\sigma}_{\beta j} \tag{3.33}$$

3.3.5 Multinomial Logistic Model

Considering a random variable Y_i, $i = 1,\ldots,n$, that may take one of several discrete values $j = 1,2,\ldots,J$. Let $\pi_{ij} = P(Y_i = j)$ denote the probability that the $i-th$ response falls in the $j-th$ category. Assuming that the response categories are mutually exclusive and exhaustive, we have $\sum_{j=1}^{J}\pi_{ij} = 1$ for each i, i.e., the probabilities add up to one for each individual, and we have only $J-1$ parameters [94]. Assume now that the individuals can be grouped according to their characteristics, forming I disjoint groups. For grouped data it will be convenient to introduce auxiliary random variables representing counts of responses in the various categories. Let n_i denote the number of cases in the $i-th$ group, $\sum_{i=1}^{I} n_i = n$, and let Y_{ij} denote the number of responses from the $i-th$ group that fall in the $j-th$ category, with observed value y_{ij}. The probability distribution of the vector counts $Y_{ij}, j = 1,\ldots,J$ given the total n_i is given by the multinomial distribution:

$$P(Y_{i1} = y_{i1}, ..., Y_{iJ} = y_{iJ}) = \binom{n_i}{y_{i1}, ..., y_{iJ}} \pi_{i1}^{y_{i1}} \ldots \pi_{iJ}^{y_{iJ}} \tag{3.34}$$

The multinomial logistic regression is widely used to model the outcomes of a categorical dependent variable with more than two categories. In this model, it is assumed that the log-odds of each response j follows a linear model, i.e., the linear component to the log of the odds of an observation in the $j-th$ category is compared to the $J-th$ category. That is, it is considered the $J-th$ category to be the baseline category, where logits of the first $J-1$ categories are constructed with the baseline category in the denominator [95]:

$$\eta_{ij} = logit(\pi_{ij}) = log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \alpha_j + \mathbf{x}_i^T\boldsymbol{\beta}_j, i = 1,\ldots,I, j = 1,\ldots,J-1 \tag{3.35}$$

where α_j is a constant and $\boldsymbol{\beta}_j$ is a vector of regression coefficients $(\beta_{1j}, \beta_{2j},\ldots,\beta_{pj})$, for $j = 1,2,\ldots,J-1$, where p is the number of regressor variables. In this model, each $\beta_{lj}, l = 1,...,p$ represents the change in the logit of the probability associated with a unit change in the $l-th$ predictor holding all other predictors constant. In terms of the probabilities π_{ij}, the model can be expressed as:

$$\pi_{ij} = \frac{exp(\eta_{ij})}{\sum_{k=1}^{J} exp(\eta_{ij})},\ j = 1,\ldots,J \tag{3.36}$$

The estimation of the parameters of this model is done by maximum likelihood, where in expression 3.34 the probabilities π_{ij} are viewed as functions of the α_j and $\boldsymbol{\beta_j}$ parameters in equation 3.35. This usually requires numerical procedures as Fisher scoring or Newton-Raphson methods. Thus, adjusting a multinomial model to the data it was possible to investigate the most influential variables in the characterization of an audiogram pattern group.

3.3.6 Additional Analysis

In addition to the statistical methodologies presented in the previous Subsections, to compare hearing thresholds at each frequency the Wilcoxon-Mann-Whitney or the Wilcoxon signed-ranks tests where used when the objective was to compare independent samples or paired samples, respectively. Kruskal-Wallis test followed by a post-hoc analysis using Dunn test for pairwise comparisons were performed in order to detect hearing thresholds differences between more than three independent groups. Chi-square test or Fisher test were employed to investigate the association of HL occurrence with demographic, environmental or medical and genetic conditions and a multiple logistic regression analysis was used to study the influence of having HL associated with demographic, environmental or medical and genetic conditions, adjusting for age.

4

Results

Contents

This chapter provides the results of the project. In Section 4.1 a description of the sample is done and in Section 4.2 the results for HL prevalence are present. Section 4.3 shows the results of the determination of audiogram patterns and Sections 4.4 and 4.5 present the PCA and LDA results, respectively. Finally, in Section 4.6 a multinomial logistic regression model and a regression model are adjusted to the data with the purpose of identifying the variables which have an influence in the cluster membership and in the prediction of the mean amount of HL, respectively.

4.1 Sample Description

A total of 321 individuals were included in the study. They were 62–115 years-old (mean: 77.91, median: 78.00 and sd: 8.21) and 231 (72%) were women. Table 4.1 presents the distribution of participants according with gender and age.

Table 4.1: Distribution of individuals according with demographic characteristics.

Age (in years)	Female (n)	Male (n)	Total (n)
[60-70[	38	22	60
[70-75[	35	16	51
[75-80[	49	21	70
[80-85[	53	15	68
$\geq$85	56	16	72
mean$\pm$sd	78.55$\pm$8.23	76.28$\pm$7.95	77.91$\pm$8.21

Most of individuals did not have history of exposition to noise (62.7%) nor family history of hearing problems (51.0%). Only 23.5% had history of ototoxic medication intake, whereas 69.2%, 56.0% and 61.5% had clinical history of hypertension, cholesterol and tinnitus, respectively. Regarding genetic conditions, S phenotype of NAT2 was the most frequent (56.7%), with R phenotype occurring only in 6.9% of the individuals. The genotype T/T at GRM7 gene is present in 59.8% of individuals, whereas A/A genotype was only represented by 4.8% of individuals. About half of individuals (50.7%) had a H mtDNA haplotype. U mtDNA haplotype was the second most common, being presented in 12.0% of the individuals. Table 4.2 and Table 4.3, respectively, show the detailed description of environmental or medical and genetic characteristics of the individuals in the sample.

Table 4.2: Clinical conditions description of the individuals.

Environmental and Clinical Characteristics		Female (n)	Male (n)	Total, n (%)
Noise Exposure	Yes	68	42	110 (37.3)
	No	149	36	185 (62.7)
	UKN	14	12	26
Family History	Yes	73	21	94 (49.0)
	No	67	31	98 (51.0)
	UKN	91	38	129
Hypertension	Yes	109	37	146 (69.2)
	No	48	17	65 (30.8)
	UKN	74	36	110
Cholesterol	Yes	83	24	107 (56.0)
	No	60	24	84 (44.0)
	UKN	88	42	130
Tinnitus	Yes	88	30	118 (61.5)
	No	52	22	74 (38.5)
	UKN	91	38	129
Ototoxic Medication	Yes	20	11	31 (23.5)
	No	69	32	101 (76.5)
	UKN	142	47	189

UKN: Missing Values

Table 4.3: Genetic conditions description of the individuals.

Genetic Characteristics		Female (n)	Male (n)	Total, n (%)
NAT2 phenotype	I	62	22	84 (36.4)
	R	11	5	16 (6.9)
	S	101	30	131 (56.7)
	UKN	57	33	90
GRM7 genotype	A/A	10	5	15 (4.8)
	A/T	78	32	110 (35.4)
	T/T	136	50	186 (59.8)
	UKN	7	3	10
mtDNA haplotype	H	106	34	140 (50.7)
	HV	13	6	19 (6.9)
	I	5	0	5 (1.8)
	J	5	8	13 (4.7)
	K	6	5	11 (4.0)
	L	9	4	13 (4.7)
	M	3	0	3 (1.1)
	N	1	0	1 (0.4)
	R	4	0	4 (1.4)
	T	12	6	18 (6.5)
	U	22	11	33 (12.0)
	V	5	1	6 (2.2)
	W	2	2	4 (1.4)
	X	5	0	5 (1.8)
	Y	0	1	1 (0.4)
	UKN	33	12	45

UKN: Missing Values

Among all subjects, the mean PTA was 42.75 ± 18.61 dB HL in the RE and 41.65 ± 17.03 dB HL in the LE (p=0.048). Concerning PTA no statistical significant difference was found between men and women either for RE (p=0.886) or LE (p=0.608). Differences between genders were found at frequencies of 250 Hz (RE: p=0.036 and LE: p=0.005), 500 Hz (RE: p=0.049 and LE: p=0.032) and 4000 Hz (RE: p=0.002 and LE: $p<0.001$). Considering overall sample, significant differences in hearing thresholds were found between REs and LEs at frequencies of 500 Hz (p=0.017), 1000 Hz (p=0.015) and 8000 Hz (p=0.016). However, either in male or female sample, no differences in hearing thresholds were found at each frequency between REs and LEs (Table 4.4).

Table 4.4: Hearing intensity thresholds (dB HL) at each frequency for female, male and overall individuals.

Sample	Ear (n)	Frequency (Hz)						PTA
		250	500	1000	2000	4000	8000	
Female	RE	33.20±18.63	35.91±19.88	37.14±19.71	43.46±21.52	54.76±21.67	70.76±21.20	42.82±19.35
	LE	33.00±16.83	34.59±18.21	35.67±19.25	42.19±19.71	54.33±19.82	69.31±20.88	41.69±17.92
p-value		0.580	0.076	0.073	0.356	0.637	0.06	0.103
Male	RE	28.56±15.11	31.44±16.91	34.39±19.80	42.72±19.61	61.78±17.95	73.00±17.09	42.58±16.65
	LE	27.11±13.34	29.56±13.89	32.5±17.26	42.72±18.92	61.44±16.75	71.22±18.80	41.56±14.59
p-value		0.193	0.087	0.078	0.934	0.843	0.136	0.277
Total	RE	31.90±17.82	34.66±19.17	36.37±19.74	43.26±20.97	56.73±20.91	71.39±20.13	42.75±18.61
	LE	31.35±16.13	33.18±17.24	34.78±18.74	42.34±19.47	56.32±19.25	69.84±20.31	41.65±17.03
p-value		0.267	0.017	0.015	0.405	0.595	0.016	0.048

In Figure 4.1 are presented the mean values of hearing intensity thresholds at each frequency tested when considering the BE for each individual, i.e., the ear with a lower $PTA_{0.5,1,2,4kHz}$. As shown, in general considering overall and both genders separately, with increasing age, hearing thresholds at each frequency become higher, which indicate loss of hearing.

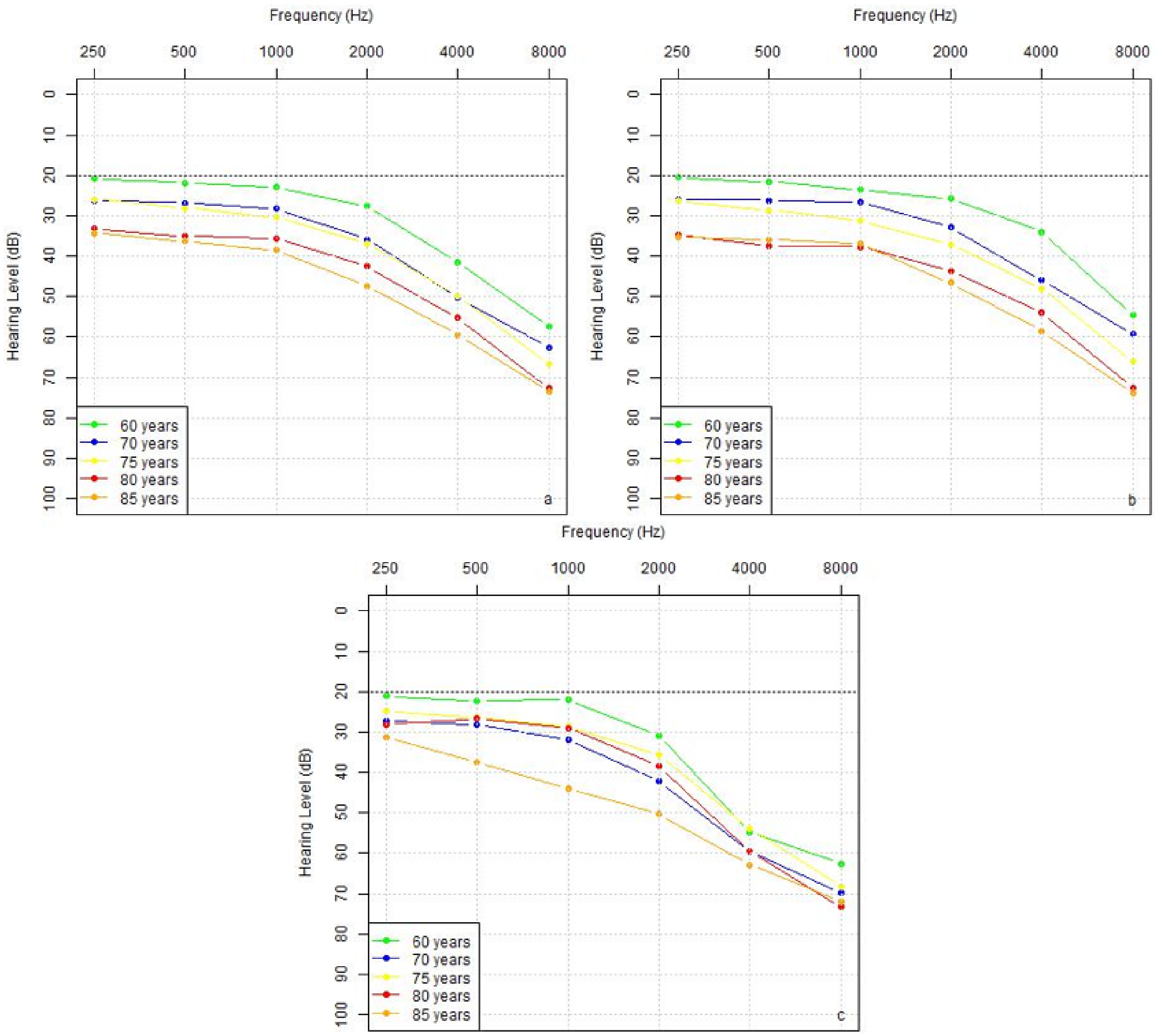

Figure 4.1: Mean hearing intensity thresholds at each test frequency considering a) overall, (b) female and (c) male sample.

There were significant differences between genders in the frequencies of 250 Hz (p=0.049) and 4000 Hz (p<0.001). When controlled by age, differences between males and females were found on age group of 60-70 years-old at 4000 Hz frequency (p<0.001) and also, on age group of 70-75 years-old at high frequencies (4000 Hz: p=0.013 and 8000 Hz: p=0.049) and on age group of 80-85 years-old in the frequency of 500 Hz (p=0.022).

4.2 Prevalence of HL

The estimated prevalence of any type of HL ($PTA_{0.5,1,2,4kHz}$ ≥25dB HL) was 79.1%, being significantly different among age groups ($\chi^2 = 32.4, p < 0.001$), increasing gradually with the increase of age: the prevalence reached 93.1% among those with ≥ 85 years old, while the same rate was 56.7% among those with 60-70 years old (Table 4.5).

Table 4.5: Distribution of age groups in HL levels based on PTA calculated for the BE averaged over 0.5, 1, 2, and 4 kHz.

	60-70 n (%)	70-75 n (%)	75-80 (n (%)	80-85 n (%)	≥85 n (%)	Total n (%)
Normal (≤25db HL)	26 (43.3)	11 (21.6)	18 (25.7)	7 (10.3)	5 (6.9)	67 (20.9%)
Mild (25 < db HL ≤40)	25 (41.7)	24 (47.1)	27 (38.6)	33 (48.5)	20 (27.8)	129 (40.2%)
Moderate (40 < db HL ≤60)	8 (13.3)	14 (27.5)	23 (32.9)	22 (32.4)	39 (54.2)	106 (33.0%)
Severe/Profound (≥60 db HL)	1 (1.7)	2 (3.9)	2 (2.9)	6 (8.8)	8 (11.1)	19 (5.9%)

Although non-significant, the HL prevalence was higher for male (82.2%, n=74) than for female (77.9%, n=180), despite the influence of age $(OR_{(M/F)} = 1.68, CI_{95\%} = [0.88, 3.22], p = 0.113)$. Table 4.6 shows the degree of HL according to the gender.

Table 4.6: Distribution of gender groups and overall sample in HL levels based on PTA calculated for the BE averaged over 0.5, 1, 2, and 4 kHz.

	Female n (%)	Male n (%)	Total n (%)
Normal (≤25db HL)	51 (22.1)	16 (17.8)	67 (20.9)
Mild (25 < db HL ≤40)	89 (38.5)	40 (44.4)	129 (40.2)
Moderate (40 < db HL ≤60)	75 (32.5)	31 (34.4)	106 (33.0)
Severe/Profound (≥60 db HL)	16 (6.9)	3 (3.3)	19 (5.9)

HL occurrence wasn't significantly associated with noise exposure $(\chi^2 = 0.013, p = 0.910)$, familial history $(\chi^2 = 0.23, p = 0.631)$, hypertension $(\chi^2 = 0.75, p = 0.385)$, cholesterol $(\chi^2 = 2.22, p = 0.137)$, tinnitus $(\chi^2 = 0.36, p = 0.55)$, ototoxic medication $(\chi^2 = 0.49, p = 0.486)$, NAT2 phenotype $(p = 0.969)$ and GRM7 genotype $(p = 0.320)$. Same results were obtained using logistic regression analysis when controlling by age: only age still remained a significant association with HL, whereas the influence of the others variables was not significant for HL occurrence. For one-unit increase of age it is expected a 10% increase in the odds of having HL.

4.3 Prevalence of Audiogram Shape

To determine the prevalence of audiogram shape there were performed two methodologies. The first one was based on a classification proposed by Wuyts [3] (Subsection 3.2.4) and the second one was performed using classification techniques, namely hierarchical and K-means clustering, all presented in the previous chapter. The classification proposed by Wuyts [3] was done considering the better and worst ears, and the right and the left ears. In addition, a comparison of hearing intensity thresholds at the respective test frequency for each configuration between RE vs LE and BE vs WE was performed. Morever, clustering techniques were only performed in RE what will be conveniently justified. To minimize the potential differences between the two ears, the subjects with asymmetrical HL were excluded, i.e., individuals who had a difference between the left and right ear hearing thresholds of 20 dB or more for at least 2 frequencies out of 0.5, 1, and 2 kHz. Thus, a total of 297 subjects were included in the determination of HL patterns.

4.3.1 Classification proposed by Wuyts

The most common configuration was HFSS, accounting for more than 50% of the REs, LEs, BEs and WEs. The second most common was HFGS followed by a FLAT configuration. The remaining configurations were rare: MFRU 0.3% for the RE or BE with LFA and MFU configurations not represented. About 174 individuals (58.6%) presented the same audiogram configuration in both ears. These results are presented in Table 4.7.

Table 4.7: Audiogram configuration prevalence (n (%)) for RE, LE, BE and LE.

Configuration	None n (%)	FLAT n (%)	HFGS n (%)	HFSS n (%)	LFA n (%)	MFU n (%)	MFRU n (%)	Mixed n (%)
RE	3 (1.0)	43 (14.5)	88 (29.6)	152 (51.2)	0 (0.0)	0 (0.0)	1 (0.3)	10 (3.4)
LE	8 (2.7)	34 (11.4)	97 (32.7)	150 (50.5)	0 (0.0)	0 (0.0)	0 (0.0)	8 (2.7)
BE	3 (1.0)	49 (16.5)	85 (28.6)	152 (51.2)	0 (0.0)	0 (0.0)	1 (0.3)	7 (2.4)
WE	8 (2.7)	28 (9.4)	100 (33.7)	150 (50.5)	0 (0.0)	0 (0.0)	0 (0.0)	11 (3.7)

The prevalence of audiograms not fitting into any of the above-mentioned configurations was low, being less than 3% considering BE vs WE or RE vs LE. As shown in the above table, there were also some audiograms classified in two classes. Therefore, further analyses were performed taken into account the three most common audiogram configurations: HFSS, HFGS and FLAT. For each one of them it was verified if there were significant differences on hearing intensity thresholds at each tested frequency between RE and LE, and also between BE and WE. The analysis of RE vs LE showed no significant difference in hearing intensity thresholds at all frequencies between both ears in all configurations. However, when comparing BE vs WE differences were found in FLAT configuration on frequency of 2000 Hz ($p=0.040$), in HFSS configuration at frequencies above 250 Hz (500 Hz: $p=0.004$,1000 Hz: $p=0.007$, 2000 Hz: $p=0.002$, 4000 Hz: $p<0.001$ and 8000 Hz: $p=0.009$) and in frequency of 500 Hz for HFGS configuration ($p=0.028$). These results suggest that the amount of HL inside a audiogram configuration is different for BE and WE, whereas for RE and LE this was not verified, that is hearing thresholds were not significantly different from each other (Figure 4.2).

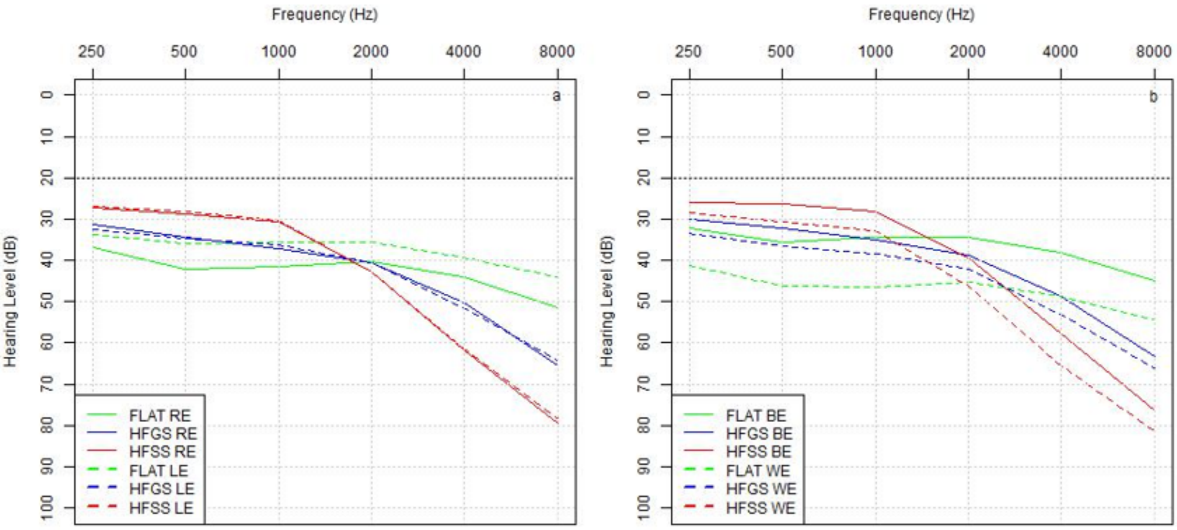

Figure 4.2: Mean hearing intensity thresholds for FLAT, HFGS and HFSS configurations for: a) RE vs LE and b) BE vs WE.

Thus, although audiogram configurations were not always the same for the LEs and the REs (see Table 4.8 and Table 4.9), the different audiogram configurations were equally represented in both ears (Table 4.10). For example, the individuals with LE configuration FLAT are found to have RE configuration FLAT (52.94%), HFGS (29.41%), HFSS (8.82%) and Mixed (8.82%). Similar reading, by row, can be done in Table 4.9 to check the distribution of the LE configurations for each configuration of the RE. Therefore, all further analyses were performed on the RE audiograms of all subjects.

Table 4.8: Percentages of LEs with equal audiogram shape in the RE versus percentages of LEs with different audiogram shape in the RE.

	LE:None	LE:FLAT	LE:HFGS	LE:HFSS	LE:LFA	LE:MFU	LE:MFRU	LE:Mixed
RE:None	12.5	0	1.03	0.67	0	0	0	0
RE:FLAT	62.5	52.94	11.34	2.67	0	0	0	62.5
RE:HFGS	12.5	29.41	44.33	21.33	0	0	0	25
RE:HFSS	12.5	8.82	38.14	74	0	0	0	0
RE:LFA	0	0	0	0	0	0	0	0
RE:MFU	0	0	0	0	0	0	0	0
RE:MFRU	0	0	1.03	0	0	0	0	0
RE:Mixed	0	8.82	4.12	1.33	0	0	0	12.5
Total of LE	100	100	100	100	0	0	0	100

Table 4.9: Percentages of REs with equal audiogram shape in the LE versus percentages of REs with different audiogram shape in the LE.

	LE:None	LE:FLAT	LE:HFGS	LE:HFSS	LE:LFA	LE:MFU	LE:MFRU	LE:Mixed	Total of RE
RE:None	33.33	0	33.33	33.33	0	0	0	0	100
RE:FLAT	11.63	41.86	25.58	36.36	0	0	0	11.63	100
RE:HFGS	1.14	11.36	48.86	36.36	0	0	0	2.27	100
RE:HFSS	0.66	1.97	24.34	73.03	0	0	0	0	100
RE:LFA	0	0	0	0	0	0	0	0	0
RE:MFU	0	0	0	0	0	0	0	0	0
RE:MFRU	0	0	100	0	0	0	0	0	100
RE:Mixed	0	30	40	20	0	0	0	10	100

Table 4.10: Number of subjects by audiogram configurations for RE and LE.

	LE:None	LE:FLAT	LE:HFGS	LE:HFSS	LE:LFA	LE:MFU	LE:MFRU	LE:Mixed	Total of RE
RE:None	1	0	1	1	0	0	0	0	3
RE:FLAT	5	18	11	4	0	0	0	5	43
RE:HFGS	1	10	43	32	0	0	0	2	88
RE:HFSS	1	3	37	111	0	0	0	0	152
RE:LFA	0	0	0	0	0	0	0	0	0
RE:MFU	0	0	0	0	0	0	0	0	0
RE:MFRU	0	0	1	0	0	0	0	0	1
RE:Mixed	0	3	4	2	0	0	0	1	10
Total of LE	8	34	97	150	0	0	0	8	297

Figure 4.3 shows the mean hearing intensity thresholds for FLAT, HFSS and HFGS configurations. There were found significant differences on all hearing thresholds at all frequencies, excepting 2000 Hz (250 Hz: $p=0.007$, 500 Hz: $p<0.001$, 1000 Hz: $p=0.003$, 4000 Hz: $p<0.001$ and 8000 Hz: $p=0.001$), being that in 250 Hz frequency the hearing intensity thresholds were significantly higher in FLAT configuration when compared to HFSS configuration ($p=0.021$) and in 500 Hz frequency HFSS configuration had lower hearing intensity thresholds comparing with FLAT and HFGS configurations ($p<0.001$ and $p=0.009$, respectively). Also, at frequency of 1000 Hz, the hearing thresholds were higher in FLAT and HFGS configurations than in HFSS configuration ($p=0.016$ and $p=0.022$, respectively). However, this tendency is reversed at high frequencies: at 4000 Hz and 8000 Hz, HFSS configuration had significantly higher values than FLAT and HFGS configurations ($p<0.001$ and $p<0.001$, respectively for both frequencies). In addition, FLAT configuration presented lower hearing intensity thresholds when comparing with HFGS at 8000 Hz frequency ($p=0.009$).

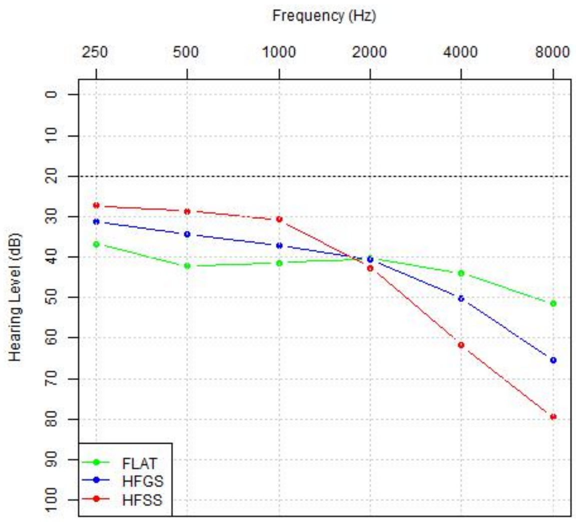

Figure 4.3: Mean hearing intensity thresholds at each frequency tested for FLAT, HFGS and HFSS configurations.

Gender vs Audiogram Configuration

The percentage of males and females per category of audiogram configuration in the RE is shown in Figure 4.4. Among men, the most prevalent configuration was HFSS, comprising the majority of the audiograms (69.1%), followed by HFGS configuration (24.7%) and FLAT configuration (6.2%). In women, also HFSS was the most frequent configuration (47.5%), followed by HFGS (33.7%) and FLAT (18.8%) configurations.

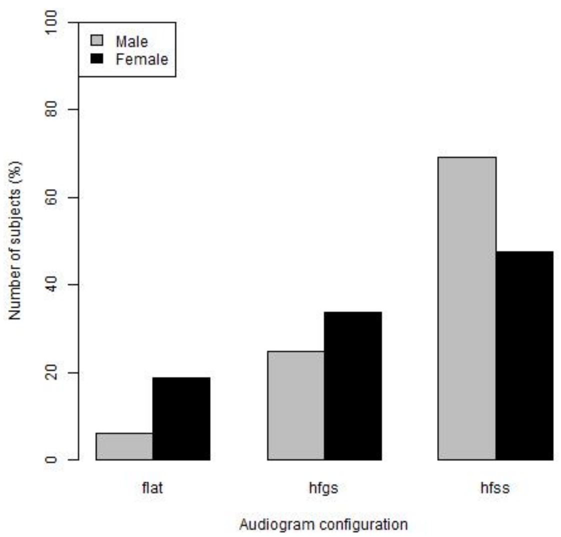

Figure 4.4: The prevalence of the different audiogram configurations for all males and females.

A significant association was found between gender and audiogram configuration ($\chi^2 = 12.603, df = 2, p = 0.002$) (Table 4.11). In fact, the proportion of women with a FLAT configuration was significantly higher than the proportion of men with the same configuration (18.8% vs 6.2%, $\chi^2 = 6.220, df = 1, p =$

0.013), whereas the proportion of men with HFSS configuration was significantly higher when comparing to the proportion of women (69.1% vs 47.5%, $\chi^2 = 10.009, df = 1, p = 0.002$).

Figure 4.5 presents the mean hearing thresholds for females and males at each frequency. Considering FLAT configuration, no significant differences were found in hearing intensity thresholds at each frequency according with gender. For HFGS configuration there were found differences among females and males in all frequencies, excepting 250 Hz (500 Hz: p=0.005, 1000 Hz:p<0.001, 2000 Hz: p=0.005, 4000 Hz: p<0.001 and 8000 Hz: p=0.008), being that males showed higher values comparing with females. In HFSS configuration the differences between genders occurred at all frequencies, excepting 2000 Hz and 4000 Hz (250 Hz:p=0.018, 500 Hz: p=0.009, 1000 Hz:p=0.003 and 8000 Hz: p=0.031), with males presenting lower values until 4000 Hz, where it happens a notch, which is often associated with the presence of environmental exposure.

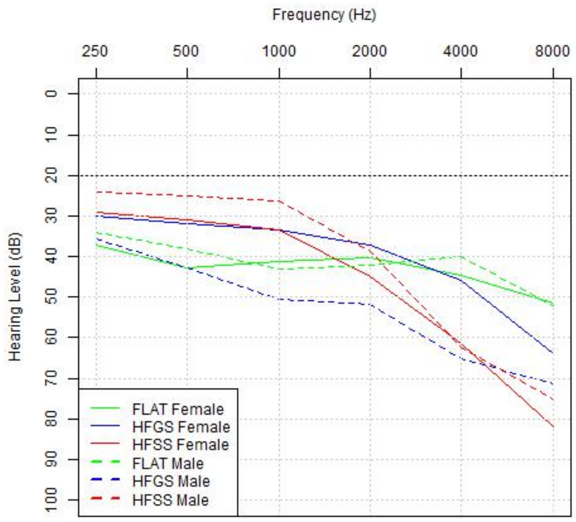

Figure 4.5: Mean hearing intensity thresholds at each frequency tested for FLAT, HFGS and HFSS configurations in females and males.

Noise exposure vs Audiogram Configuration

Noise exposure was associated with audiogram configuration at a significance level of $\alpha = 0.1$. About 62.2% of the individuals exposed to noise, had a HFSS configuration, a percentage significantly higher when compared to the same proportion in non-exposed (48.1%) ($\chi^2 = 4.327, df = 1, p = 0.038$) (Figure 4.6a)). Additionally, an association test performed on the variables gender and noise exposure showed a significant difference in noise exposure between both sexes ($\chi^2 = 11.095, df = 1, p < 0.001$), with the male population being significantly more exposed to noise compared to female population (55.1% vs 31.4%). In non-exposed individuals, the influence of gender in audiogram configuration was not significant. On the other hand, in exposed subjects a significant association between gender and audiogram configuration was found (Fisher exact test, $p = 0.038$), being that 73.7% of males presented a HFSS configuration and only 2.6% presented a FLAT configuration. Among females, 55% had a HFSS configuration, occurring a FLAT configuration in 18.3% (Figure 4.6b)).

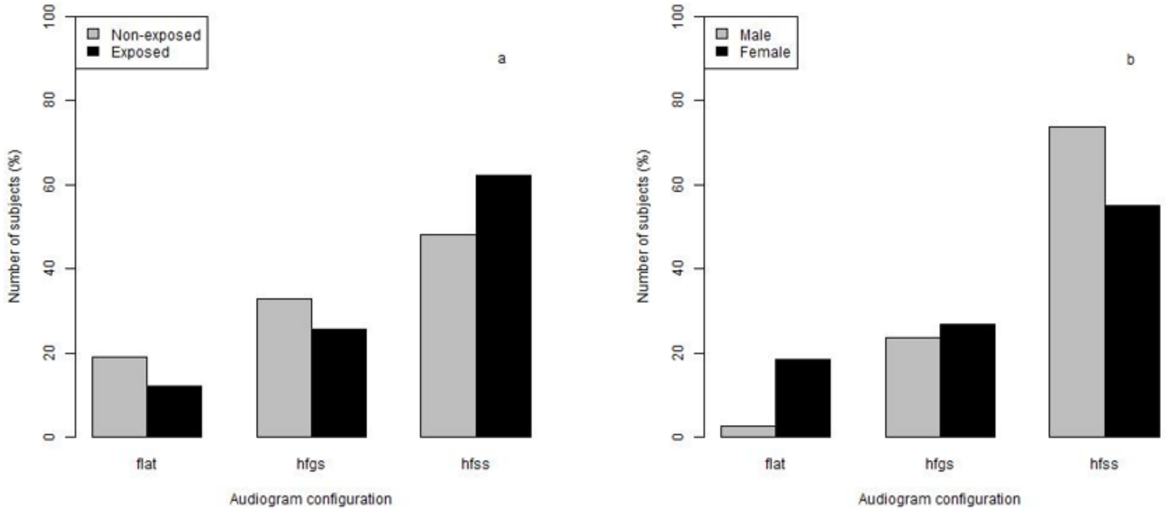

Figure 4.6: The prevalence of the different audiogram configurations for a) exposed and non-exposed and b) exposed females and males.

The distribution of subjects according to demographic, environmental or medical and genetic conditions for the three most prevalent configurations is presented in Table 4.11, Table 4.12 and Table 4.13, respectively. As observed, no significant associations were found between the remaining individuals characteristics and audiogram configuration. Although no significant difference in prevalence of configurations was found by age group ($\chi^2 = 5.457, df = 8, p = 0.708$), within each configuration, when considering age groups, in mean, as age increases, the hearing intensity thresholds at each frequency increased at a same rate. This could indicate that with aging, the audiogram shape remains the same and the amount of HL changes equally for each frequency (Figure 4.7), i.e., the measured hearing intensities thresholds become higher, indicating a HL occurrence (audiograms move down), but the change in HL is the same at each frequency (given by the parallel audiograms).

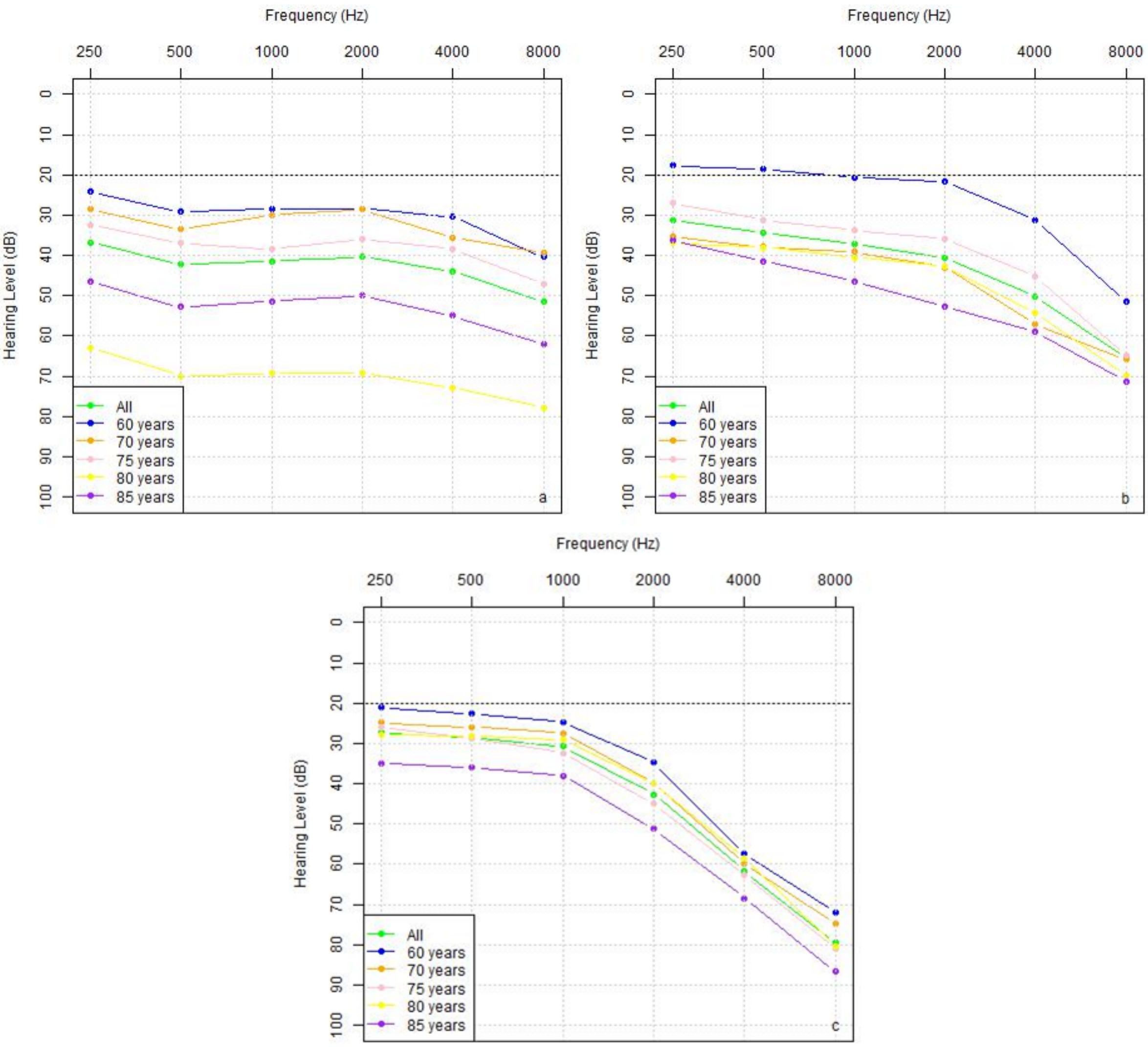

Figure 4.7: Mean hearing intensity thresholds audiogram for: a) FLAT, b) HFGS and c) HFSS configurations.

Table 4.11: Distribution of subjects according to demographic features in FLAT, HFGS and HFSS audiogram configurations.

Characteristic		**Configuration**			**p-value**
		FLAT (n=43)	HFGS (n=88)	HFSS (n=152)	
Gender	Female	38	68	96	0.002
	Male	5	20	56	
Age (in years)	60-70	12	15	26	0.708
	70-75	7	12	26	
	75-80	10	17	36	
	80-85	7	20	31	
	≥85	7	24	33	
	mean±sd	76.40±7.96	78.65±8.13	77.83±7.76	0.505

Table 4.12: Distribution of subjects according to environmental and medical features in FLAT, HFGS and HFSS audiogram configurations.

Characteristic		Configuration			p-value
		FLAT (n=43)	HFGS (n=88)	HFSS (n=152)	
Noise Exposure	Yes	12	25	61	0.079
	No	31	53	78	
	UKN		10	13	
Family History	Yes	16	26	39	0.865
	No	16	27	47	
	UKN	11	35	66	
Hypertension	Yes	23	47	59	0.288
	No	9	15	33	
	UKN	11	26	60	
Cholesterol	Yes	20	34	40	0.134
	No	11	21	44	
	UKN	12	33	68	
Tinnitus	Yes	20	33	49	0.550
	No	9	23	36	
	UKN	14	32	67	
Ototoxic Medication	Yes	3	9	15	0.793
	No	15	26	49	
	UKN	25	53	88	

UKN: Missing Values

Table 4.13: Distribution of subjects according to genetic features in FLAT, HFGS and HFSS audiogram configurations.

Characteristic		Configuration			p-value
		FLAT (n=43)	HFGS (n=88)	HFSS (n=152)	
NAT2 phenotype	I	11	24	37	0.937
	R	3	4	9	
	S	20	33	64	
	UKN	9	27	42	
GRM7 genotype	A/A	1	5	9	0.756
	A/T	13	33	54	
	T/T	29	48	84	
	UKN		2	5	

UKN: Missing Values

The distribution of the individuals according with HL degree for each audiogram configuration is presented in Table 4.14. The majority of the individuals within each configuration had a mild HL. In addition, all individuals with a FLAT configuration and with a normal or severe/profound HL level were women.

Table 4.14: Distribution of audiogram configurations in HL levels based on PTA calculated for the RE averaged over 0.5, 1, 2, and 4 kHz.

	FLAT n (%)	HFGS n (%)	HFSS n (%)
Normal ($\leq$25db HL)	11 (25.6)	18 (20.5)	22 (14.5)
Mild (25 $<$ db HL $\leq$40)	13 (30.2)	30 (34.1)	59 (38.8)
Moderate (40 $<$ db HL $\leq$60)	11 (25.6)	30 (34.1)	58 (38.2)
Severe/Profound ($\geq$60 db HL)	8 (18.6)	10 (11.4)	13 (8.6)
Total	43 (100)	88 (100)	152 (100)

4.3.2 Cluster Analysis to find HL Patterns

Secondly, a cluster analysis was used to find audiogram patterns in audiometry data set. This type of unsupervised learning classification method involved partitioning a set of frequencies of individuals audiograms into groups, such that the audiograms in the same group are similar while those in other groups are dissimilar. In this study two techniques were applied: a hierarchical and a K-means clustering. Hierarchical clustering with the Ward method as aggregation criteria was carried out to explore the pattern of the individual's audiograms. The Ward method ensured homogeneity and relatively equal size of the groups. In a cluster analysis the determination of the number of clusters is a subjective process. However, there are statistical tests which can be used as a guidance to identify the "optimal" number of clusters, in this work, it was decided to use the three-cluster solution, since the objective was to verify if the three most prevalent audiogram configurations determined by the classification proposed by Wuyts [3] (presented in Subsection 4.3.1) were associated to a given cluster. Differences between the variables were calculated using Euclidean distance. Since the main objective was to obtain groups composed by similar audiogram curves having into account only their shape (not the amount or severity of HL), the original data was transformed using the function given by:

```
padrao <-function(x){
 p <-length(x)
 w <-vector()
 w[1] <-0
 for (i in 1:(p-1)){
  for (j in (i+1):p){
   dif <-x[i]-x[j]
   w <-c(w,dif)
  }
 }
 w <-w[-1]
 return(w)
}
```

Figure 4.8: Function employed in each set of six frequencies on the audiological data set before cluster analysis.

In this study, using this function for each set of six frequencies ($f_1 = 250\ Hz$, $f_2 = 500\ Hz$, $f_3 = 1000\ Hz$, $f_4 = 2000\ Hz$, $f_5 = 4000\ Hz$, $f_6 = 8000\ Hz$) associated to an individual, the pairwise differences were calculated, resulting in $(f_1 - f_2, f_1 - f_3, \ldots, f_1 - f_6, f_2 - f_3, \ldots, f_5 - f_6)$. Thus, the output was a vector with dimension ${}^6C_2 = 15$ that contained all of those pairwise differences. This vector contains information regarding the shape and magnitude of the frequencies variation, independently of their position (HL degree) on the audiogram. It allows to consider as similar curves those that have the same shape, wherever they are located (more above or below on the audiogram graph). For example, considering the mean hearing threshold values of individuals in age groups of 75-80 and ≥ 85 years-old in cluster 2 obtained from K-Means clustering (Figure 4.20b) and Table 4.15), the elements of the output vector are then presented in Table 4.16.

Table 4.15: Example of two input vectors for pattern function: .

Frequency (Hz)	**Mean hearing thresholds (dB)**	
	75-80 years-old	≥85 years-old
f_1	27.10	32.93
f_2	30.65	36.90
f_3	36.61	41.21
f_4	50.48	55.52
f_5	68.06	72.07
f_6	80.81	84.66

Table 4.16: Example of two input vectors for pattern function: .

Frequency pairwise difference	**Mean hearing thresholds difference (dB)**	
	75-80 years-old	≥85 years-old
$f_1 - f_2$	-3.55	-3.97
$f_1 - f_3$	-9.52	-8.28
$f_1 - f_4$	-23.39	-22.59
$f_1 - f_5$	-40.97	-39.14
$f_1 - f_6$	-53.71	-51.72
$f_2 - f_3$	-5.97	-4.31
$f_2 - f_4$	-19.84	-18.62
$f_2 - f_5$	-37.42	-35.17
$f_2 - f_6$	-50.16	-47.76
$f_3 - f_4$	-13.87	-14.31
$f_3 - f_5$	-31.45	-30.86
$f_3 - f_6$	-44.19	-43.45
$f_4 - f_5$	-17.58	-16.55
$f_4 - f_6$	-30.32	-29.14
$f_5 - f_6$	-12.74	-12.59

In conclusion, it was expected that the cluster groups would be done considered only the shape of the audiograms since the information now retrieved is on the differences of HL between frequencies and not the amount of HL in the different frequencies.

4.3.2.A Hierarchical Clustering

The dendrogram that represented clusters of similar groups of participants is in Figure 4.9. The cophenetic distance which is a correlation coefficient for the hierarchical cluster tree and signifies how well the dendrogram represents dissimilarities among the data was 0.47.

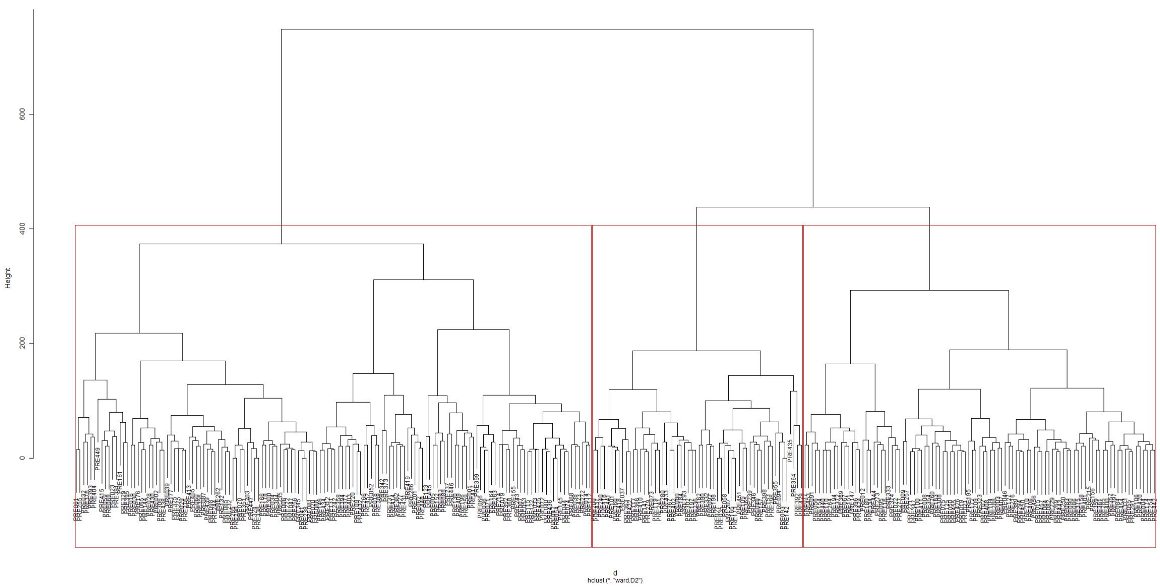

Figure 4.9: Dendogram resulted from hierarchical clustering with Ward method on audiological data transformed.

Thus, considering a partition in three clusters, the distribution of the subjects was the following:

Table 4.17: Distribution of individuals in the clusters obtained from hierarchical clustering.

	Cluster 1	Cluster 2	Cluster 3
n (%)	142 (47.8)	97 (32.7)	58 (19.5)

The mean pattern of audiograms obtained for each cluster is represented in Figure 4.10. The obtained mean patterns for cluster 1, cluster 2 and cluster 3 were much similar to a HFSS, a HFGS and a FLAT configurations, respectively, obtained through Wuyts classification [3] (Figure 4.3).
At all frequencies, significant differences in hearing thresholds among cluster groups were found (250 Hz: $p < 0.001$, 500 Hz: $p < 0.001$, 1000 Hz: $p = 0.001$, 2000 Hz: $p < 0.001$, 4000 Hz: $p < 0.001$ and 8000 Hz: $p = 0.001$), being that in 250 Hz frequency the hearing intensity thresholds were significantly lower in cluster 1 (“HFSS” pattern) than cluster 3 (“FLAT” pattern) ($p < 0.001$) and in 500 Hz and 1000 Hz frequencies, cluster 3 (“FLAT” pattern) had higher hearing intensity thresholds comparing with cluster 1 (“HFSS” pattern) and cluster 2 (“HFGS” pattern) (500 Hz: $p < 0.001$ and $p = 0.001$, respectively and 1000 Hz: $p = 0.002$ and $p = 0.002$, respectively). Cluster 2 (“HFGS” pattern) presented significantly

lower values at frequency of 2000 Hz when compared to cluster 1 (“HFSS” pattern) ($p < 0.001$). At frequency of 4000 Hz, the hearing thresholds were higher in cluster 1 (“HFSS” pattern) than in cluster 2 (“HFGS” pattern) and cluster 3 (“FLAT” pattern) ($p < 0.001$ and $p < 0.001$, respectively). Finally, in contrast with low frequencies, at 8000 Hz, cluster 3 (“FLAT” pattern) had hearing intensity thresholds lower than cluster 1 (“HFSS” pattern) and cluster 2 (“HFGS” pattern)($p < 0.001$ and $p < 0.001$, respectively). Additionally, hearing thresholds of cluster 2 (“HFGS” pattern) were lower than those in cluster 1 (“HFSS” pattern) ($p = 0.004$).

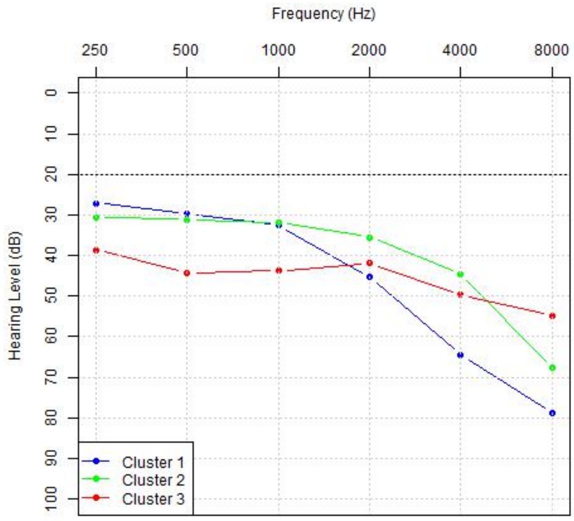

Figure 4.10: Mean hearing intensity thresholds at each frequency tested for the three clusters obtained by hierarchical clustering (Cluster 1: “HFSS” pattern, Cluster 2: “HFGS” pattern and Cluster 3: “FLAT” pattern).

Gender vs Audiogram Configuration

The distribution of audiograms in each cluster according with gender is represented in Figure 4.11. The majority of the male audiograms (67.1%) were allocated to the cluster 1 (“HFSS” pattern), whereas in the other clusters, male audiograms were almost evenly distributed (cluster 2 - “HFGS” pattern: 17.6% and cluster 3 - “FLAT” pattern: 15.3%). In females, the most common audiogram pattern was the one associated to cluster 1 (“HFSS” pattern) (40.1%), followed by those in cluster 2 (“HFGS” pattern) (38.7%) and cluster 3 (“FLAT” pattern) (21.2%).

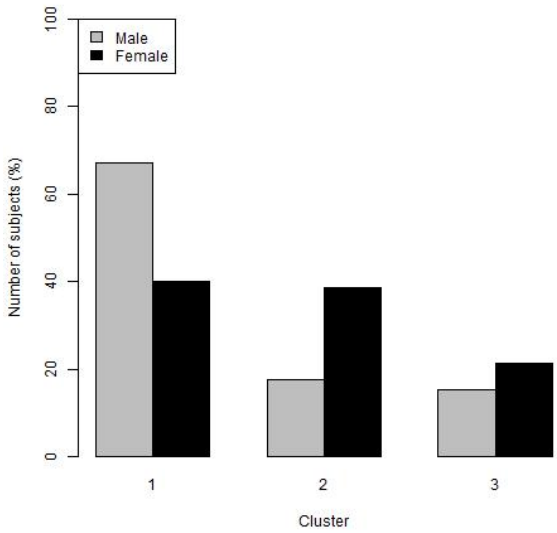

Figure 4.11: The prevalence of the different audiogram cluster patterns obtained by hierarchical clustering for all males and females (Cluster 1: "HFSS" pattern, Cluster 2: "HFGS" pattern and Cluster 3: "FLAT" pattern).

The differences in audiogram pattern distribution between genders was statistically significant ($\chi^2 = 18.538, df = 2, p < 0.001$). The proportion of males with a pattern characteristic of cluster 1 ("HFSS" pattern) was significantly higher than same proportion in females ($\chi^2 = 16.616, df = 1, p < 0.001$), as well the proportion of females with a pattern of cluster 2 ("HFGS" pattern) was significantly higher than those in males ($\chi^2 = 11.266, df = 1, p < 0.001$). These results are consistent with the results obtained for HFSS configuration, in the previous Subsection, however not for HFGS and FLAT configuration.

Figure 4.12 presents the mean hearing thresholds for females and males at each frequency. In relation to cluster 1 ("HFSS" pattern), it was found a significant difference in the hearing intensity thresholds according with gender in frequency of 1000 Hz (p=0.024), whereas for the other clusters no differences were found in hearing intensity thresholds at each frequency. Similar to the HFSS configuration, the pattern represented by cluster 1 ("HFSS" pattern), presented a notch at frequency of 4000 Hz for males. Also, the pattern of cluster 2 ("HFGS" pattern), as in HFGS configuration, had lower hearing intensity thresholds for females comparing with males.

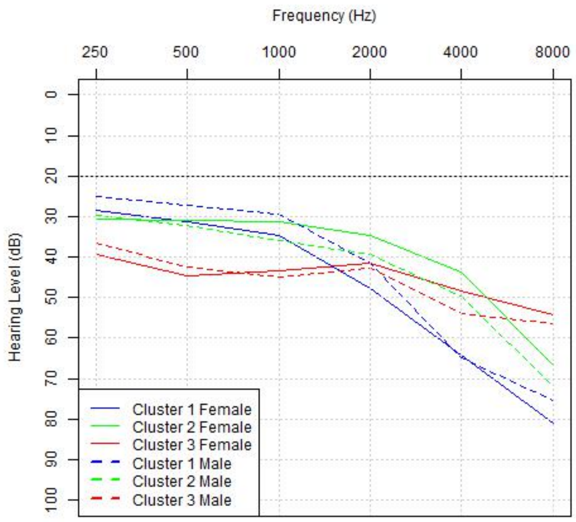

Figure 4.12: Mean hearing intensity thresholds at each frequency tested for cluster 1 ("HFSS" pattern), cluster 2 ("HFGS" pattern) and cluster 3 ("FLAT" pattern) in females and males (hierarchical clustering).

Noise exposure vs Cluster Membership

Additionally to the gender, noise exposure was significantly associated with cluster membership ($\chi^2 = 6.228, df = 2, p = 0.044$), being that of those exposed to noise, more than half were in cluster 1 ("HFSS" pattern) (56.4%), a higher proportion compared to non-exposed individuals (41.5%) ($\chi^2 = 5.087, df = 1, p = 0.024$) (Figure 4.13a)). An association test performed on the variables gender and noise exposure showed a significant difference in noise exposure between both genders ($\chi^2 = 10.412, df = 1, p = 0.001$), with the male population being significantly more exposed to noise compared to the female population (53.4% vs 31.2%). The influence of gender in cluster membership was not significant in non-exposed subjects. Otherwise, in exposed subjects a significant association was found ($\chi^2 = 8.373, df = 1, p = 0.015$). As shown in Figure 4.13b), a higher proportion of males (74.4%) had a pattern of cluster 1 ("HFSS" pattern) and only 7.7% had a pattern of cluster 3 ("FLAT" pattern). Among females, the distribution in cluster 1 ("HFSS" pattern) and cluster 2 ("HFGS" pattern) were approximately similar, with 45.2% of females having a pattern of cluster 1 ("HFSS" pattern) and 35.5% having a pattern of cluster 2 ("HFGS" pattern). Only 19.4% females presented a pattern of cluster 3 ("FLAT" pattern).

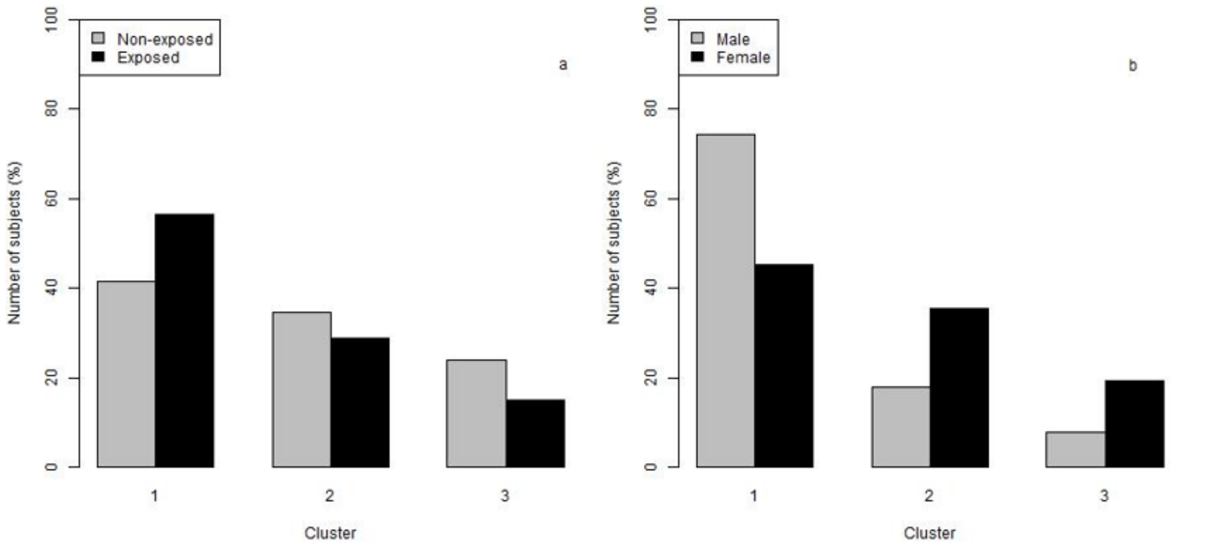

Figure 4.13: The prevalence of the different cluster membership (hierarchical clustering) for a) exposed and non-exposed and b) exposed females and males. Cluster 1: “HFSS” pattern, Cluster 2: “HFGS” pattern and Cluster 3: “FLAT” pattern

The characterization of the obtained clusters according to demographic, environmental or medical and genetic conditions of subjects are represented in Table 4.18, Table 4.19 and 4.20, respectively.

Table 4.18: Distribution of subjects according to demographic features in cluster groups obtained by hierarchical clustering.

Characteristic		**Configuration**			**p-value**
		Cluster 1: “HFSS” (n=142)	Cluster 2: “HFGS” (n=97)	Cluster 3: “FLAT” (n=58)	
Gender	Female	85	82	45	<0.001
	Male	57	15	13	
Age (in years)	60-70	22	20	14	0.408
	70-75	24	11	12	
	75-80	36	20	10	
	80-85	26	26	10	
	≥85	34	20	12	
	mean±sd	78.31±8.49	77.95±7.85	76.90±8.11	0.542

Table 4.19: Distribution of subjects according to environmental and medical features in cluster groups obtained by hierarchical clustering.

Characteristic		Configuration			p-value
		Cluster 1: “HFSS” (n=142)	Cluster 2: “HFGS” (n=97)	Cluster 3: “FLAT” (n=58)	
Noise Exposure	Yes	57	29	15	0.044
	No	71	59	41	
	UKN	14	9	2	
Family History	Yes	37	27	21	0.761
	No	41	33	19	
	UKN	64	37	18	
Hypertension	Yes	58	47	29	0.982
	No	27	21	14	
	UKN	57	29	15	
Cholesterol	Yes	39	34	25	0.481
	No	40	24	18	
	UKN	63	39	15	
Tinnitus	Yes	47	35	29	0.335
	No	31	27	12	
	UKN	64	35	17	
Ototoxic Medication	Yes	13	11	7	0.870
	No	44	29	21	
	UKN	85	57	30	

UKN: Missing Values

Table 4.20: Distribution of subjects according to genetic features in cluster groups obtained by hierarchical clustering.

Characteristic		Configuration			p-value
		Cluster 1: “HFSS” (n=142)	Cluster 2: “HFGS” (n=97)	Cluster 3: “FLAT” (n=58)	
NAT2 phenotype	I	38	27	14	0.863
	R	6	7	3	
	S	58	37	24	
	UKN	40	26	17	
GRM7 genotype	A/A	7	5	3	0.998
	A/T	50	34	21	
	T/T	79	57	33	
	UKN	6	1	1	

UKN: Missing Values

No significant association was found between cluster membership and the remaining characteristics. Although no significant difference in cluster distribution was found according to age of the subjects ($\chi^2 = 8.271, df = 8, p = 0.408$), considering each of the cluster patterns there were found significant differences between age groups in hearing intensity thresholds at each frequency (excepting for cluster 1 - “HFSS” pattern) in frequencies of 4000 Hz, $p = 0.152$ and 8000 Hz, $p = 0.203$), meaning that for the same audiogram shape the measured intensities were different depending on subject age (Figure

4.14a)). As in the results of Subsection 4.3.1, in mean, as age increases, the pattern remains the same, but the hearing intensity thresholds become higher, indicating a HL occurrence.

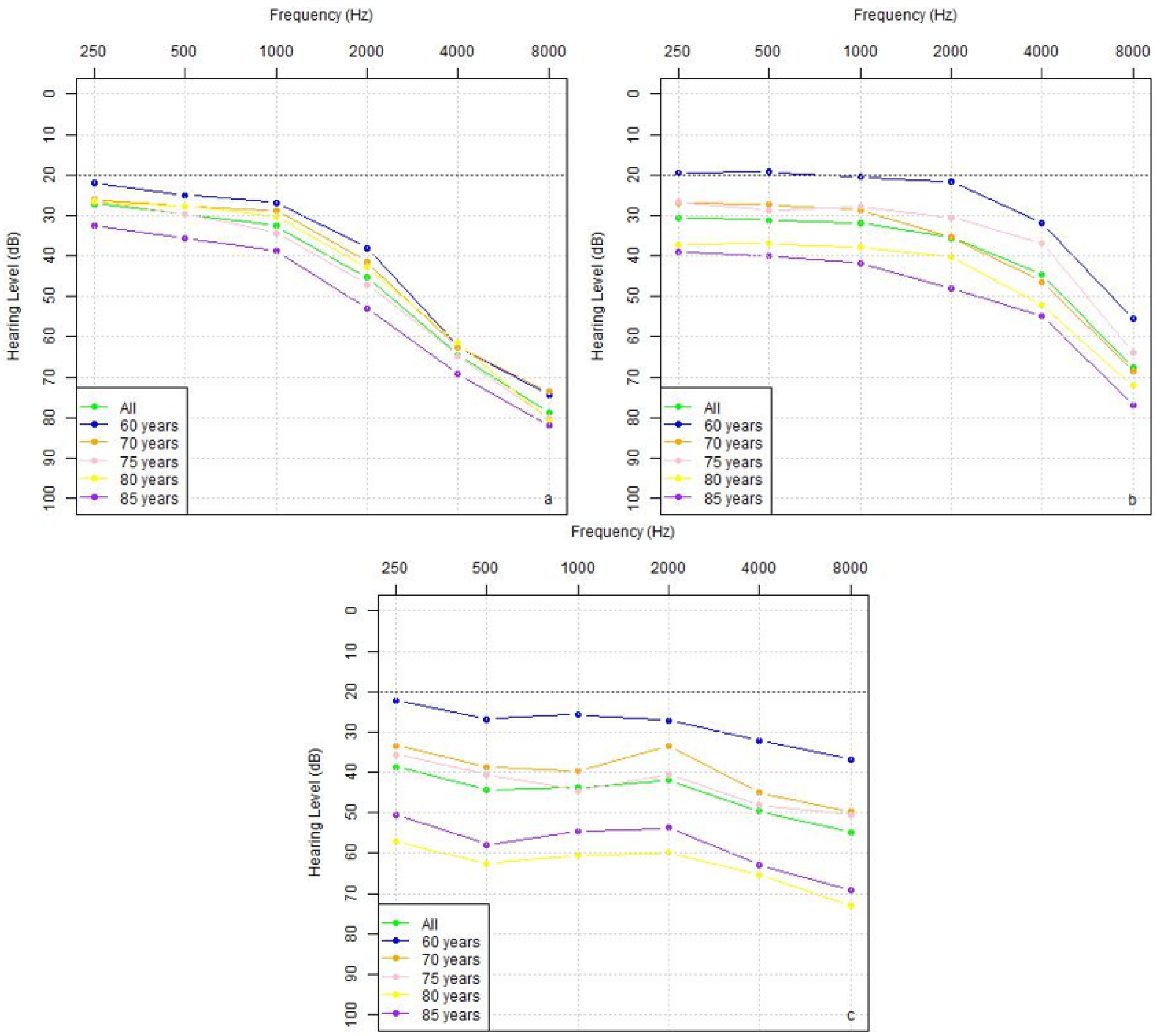

Figure 4.14: Mean hearing intensity thresholds audiogram for: a) Cluster 1 (“HFSS” pattern), b) Cluster 2 (“HFGS” pattern) and c) Cluster 3 (“FLAT” pattern) accordingly with age (hierarchical clustering).

Considering the degree of HL, as presented in Table 4.21, 40.8% of individuals in the cluster 1 (“HFSS” pattern) had a moderate HL and 37.1% of individuals in the cluster 2 (“HFGS” pattern) had a mild HL. Regarding cluster 3 (“FLAT” pattern), mild and moderate levels were the most frequent (both with 29.3%).

Table 4.21: Distribution of hierarchical cluster groups in HL levels based on PTA calculated for the RE averaged over 0.5, 1, 2, and 4 kHz.

	Cluster 1: "HFSS", n (%)	Cluster 2: "HFGS", n (%)	Cluster 3: "FLAT", n (%)
Normal ($\leq$25db HL)	13 (9.2)	28 (28.9)	12 (20.7)
Mild (25 $<$ db HL $\leq$40)	57 (40.1)	36 (37.1)	17 (29.3)
Moderate (40 $<$ db HL $\leq$60)	58 (40.8)	26 (26.8)	17 (29.3)
Severe/Profound ($\geq$60 db HL)	14 (9.9)	7 (7.2)	12 (20.7)
Total	142 (100)	97 (100)	58 (100)

4.3.2.B K-Means Clustering

Based on the dendogram and on the three most common configurations obtained through Wuyts classification [3], the number of clusters considered *a priori* to this analysis was three. The objective of this method was to minimize the WSS which corresponds to the sum of distance functions of each point in the cluster to the k centre (Figure 4.15).

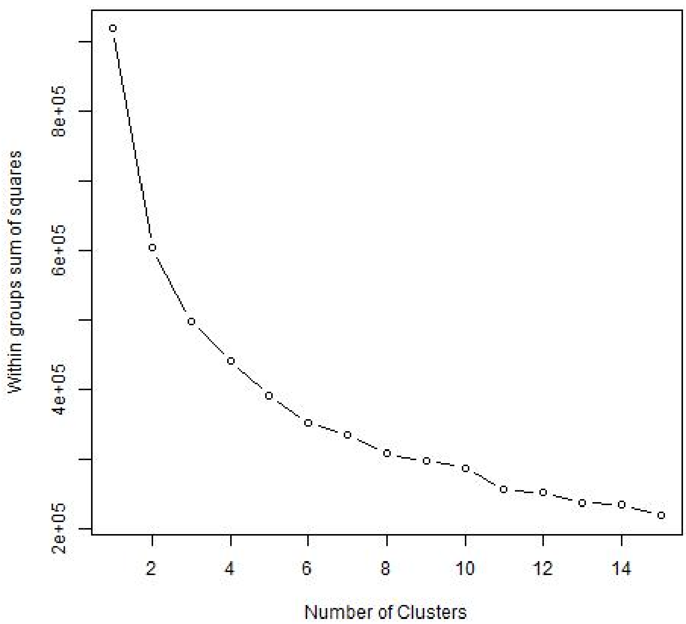

Figure 4.15: WSS plot.

In Table 4.22 is presented the distribution of the individuals in the three clusters derived from K-means algorithm.

Table 4.22: Distribution of individuals in the clusters obtained from K-Means algorithm.

	Cluster 1	Cluster 2	Cluster 3
n (%)	87 (29.3)	120 (40.4)	90 (30.3)

Figure 4.16 presents the mean hearing thresholds for each one of the cluster groups. As shown,

cluster 3 had a pattern similar to a FLAT configuration, whereas cluster 1 and cluster 2 had some characteristics of HFGS and HFSS configurations, respectively.

At all frequencies, there were found significant differences in hearing thresholds among cluster groups (250 Hz: $p < 0.001$, 500 Hz: $p < 0.001$, 1000 Hz: $p = 0.001$, 2000 Hz: $p < 0.001$, 4000 Hz: $p < 0.001$ and 8000 Hz: $p = 0.001$), being that in 250 Hz and in 500 Hz frequencies the hearing intensity thresholds were significantly higher in cluster 3 ("FLAT" pattern) than cluster 1 ("HFGS" pattern) and cluster 2 ("HFSS" pattern) (250 Hz: $p = 0.004$ and $p < 0.001$, respectively and 500 Hz: $p < 0.001$ and $p < 0.001$, respectively). At 1000 Hz and 2000 Hz, cluster 1 ("HFGS" pattern) presented lower hearing intensity thresholds when comparing with cluster 2 ("HFSS" pattern) and cluster 3 ("FLAT" pattern) (1000 Hz: $p = 0.003$ and $p < 0.001$, respectively and 2000 Hz: $p < 0.001$ and $p < 0.001$, respectively). In addition, cluster 2 ("HFSS" pattern) had higher hearing intensity thresholds than cluster 3 ("FLAT" pattern) ($p = 0.022$) at 2000 Hz. Regarding high frequencies, at frequency of 4000 Hz, the hearing thresholds were higher in cluster 2 ("HFSS" pattern) than in cluster 1 ("HFGS" pattern) and cluster 3 ("FLAT" pattern) ($p < 0.001$ and $p < 0.001$, respectively), whereas at 8000 Hz, cluster 3 ("FLAT" pattern) presented lower values of hearing thresholds in comparison with cluster 1 ("HFGS" pattern) and cluster 2 ("HFSS" pattern) ($p < 0.001$ and $p < 0.001$, respectively).

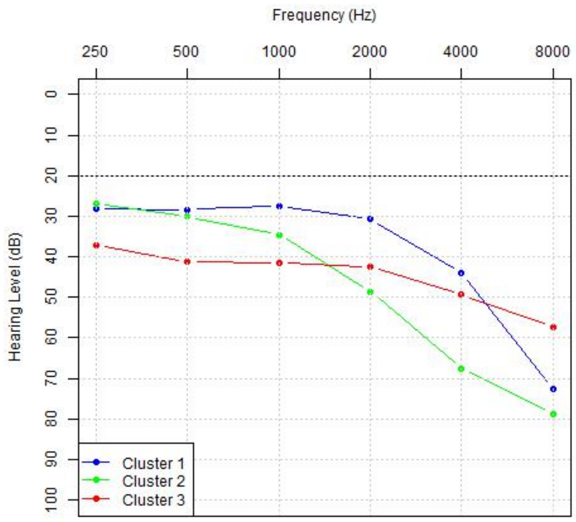

Figure 4.16: Mean hearing intensity thresholds at each frequency tested for the three clusters obtained by K-Means (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

Gender vs Audiogram Configuration

The percentage of males and females per category of cluster membership is presented in Figure 4.17. Among men, 58.8% of audiograms were included in cluster 2 ("HFSS" pattern), 23.5% in cluster 1 ("HFGS" pattern) and 17.6% in cluster 3 ("FLAT" pattern). In females, the audiograms were almost evenly distributed between groups (cluster 1 - "HFGS" pattern: 31.6%, cluster 2 - "HFSS" pattern: 33.0% and cluster 3 - "FLAT" pattern: 35.4%).

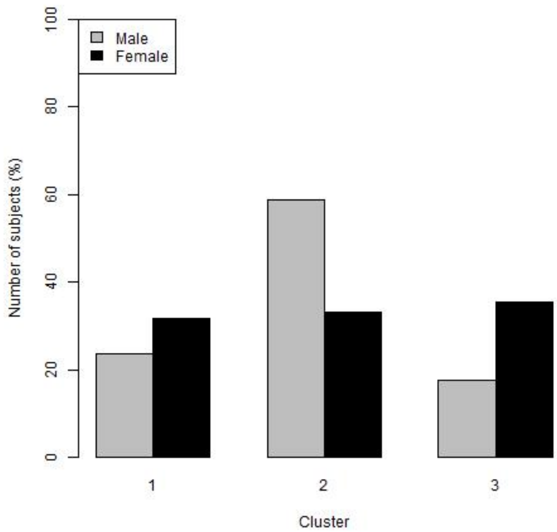

Figure 4.17: The prevalence of the different audiogram cluster patterns obtained by K-means for all males and females (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

A significant association was found between cluster membership and gender ($\chi^2 = 17.644, df = 2, p < 0.001$), being that the proportion of males in cluster 2 ("HFSS" pattern) was higher than the same proportion of females (58.8% vs 33.0%, $\chi^2 = 15.724, df = 1, p < 0.001$), as well the proportion of males in cluster 3 ("FLAT" pattern) was lower than the proportion of females (17.6% vs 35.4%, $\chi^2 = 8.211, df = 1, p = 0.004$).

Figure 4.18 shows the mean hearing thresholds at each frequency for males and females. Regarding cluster 1 ("HFGS" pattern) and cluster 2 ("HFSS" pattern), there were found significant differences in the hearing intensity thresholds according to gender in frequencies of 4000 Hz ($p = 0.035$) and 1000 Hz ($p = 0.045$), respectively. For cluster 3 ("FLAT" pattern) no differences were found in hearing intensity thresholds at each frequency among male and female. As in HFSS configuration and pattern of cluster 1 ("HFSS" pattern) in hierarchical clustering, for male gender, a notch at 4000 Hz in the pattern of cluster 2 ("HFSS" pattern) was observed.

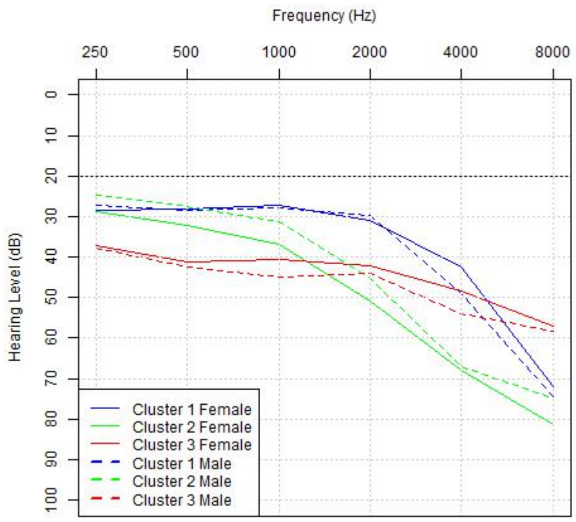

Figure 4.18: Mean hearing intensity thresholds at each frequency tested for cluster 1 (“HFGS” pattern), cluster 2 (“HFSS” pattern) and cluster 3 (“FLAT” pattern) in females and males (K-means).

Noise exposure vs Cluster Membership

As observed in Table 4.23, Table 4.24 and Table 4.25 the results obtained were similar to those obtained in hierarchical clustering. Only noise exposure (and gender) were significantly associated to the individual's cluster membership ($\chi^2 = 8.526, df = 2, p = 0.014$), being that the proportion of exposed to noise in cluster 2 (“HFSS” pattern) was significantly higher than proportion of non-exposed (48.5% vs 34.5%, $\chi^2 = 4.639, df = 1, p = 0.031$), whereas the same proportion in cluster 3 (“FLAT” pattern) was higher in non-exposed than exposed individuals (36.8% vs 20.8%, $\chi^2 = 6.930, df = 1, p = 0.008$) (Figure4.19a)). According with the association tests, the influence of gender in cluster membership in exposed subjects was significant ($\chi^2 = 8.907, df = 1, p = 0.012$), whereas in non-exposed subjects it was not. In fact, 66.7% of exposed males had a pattern of cluster 2 (“HFSS” pattern) and only 10.3% had a pattern of cluster 3 (“FLAT” pattern). In exposed females, the patterns of cluster 1 (“HFGS” pattern) and cluster 2 (“HFSS” pattern) were the most common, corresponding to 35.5% and 37.1%, respectively (Figure 4.19b)).

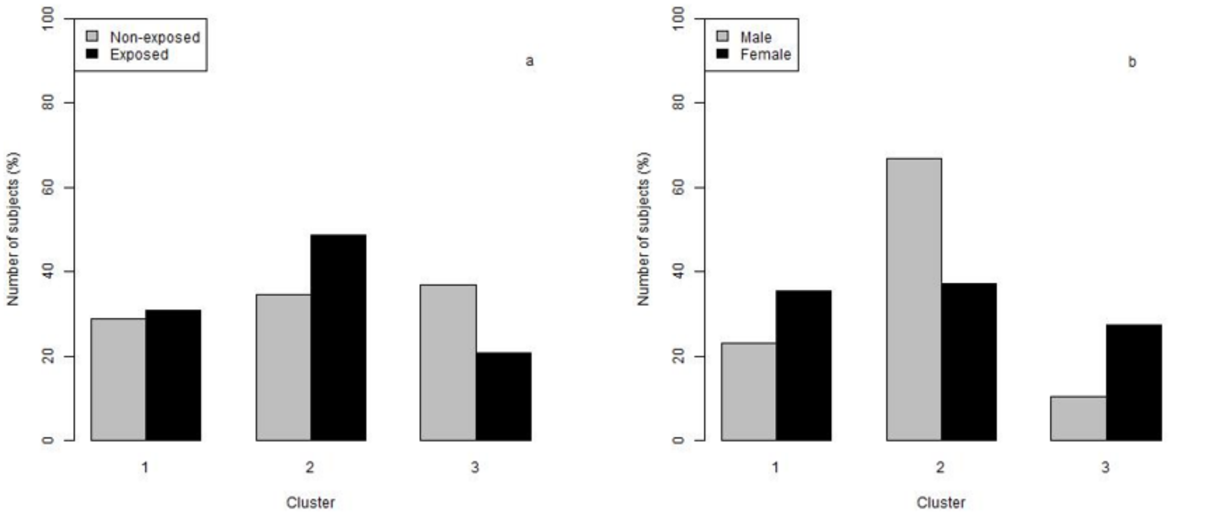

Figure 4.19: The prevalence of the different K-Means cluster membership for a) exposed and non-exposed and b) exposed females and males. Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern

Table 4.23: Distribution of subjects according to demographic features in cluster groups obtained by K-Means.

Characteristic		**Configuration**			**p-value**
		Cluster 1: "HFGS" (n=87)	Cluster 2: "HFSS" (n=120)	Cluster 3: "FLAT" (n=90)	
Gender	Female	67	70	75	<0.001
	Male	20	50	15	
Age (in years)	60-70	20	17	19	0.664
	70-75	12	21	14	
	75-80	19	31	16	
	80-85	20	22	20	
	≥85	16	29	21	
	mean±sd	77.09±7.75	78.22±7.98	78.30±8.91	0.700

Table 4.24: Distribution of subjects according to environmental and medical features in cluster groups obtained by K-Means.

Characteristic		Configuration			p-value
		Cluster 1: "HFGS" (n=87)	Cluster 2: "HFSS" (n=120)	Cluster 3: "FLAT" (n=90)	
Noise Exposure	Yes	31	49	21	0.014
	No	49	59	63	
	UKN	7	12	6	
Family History	Yes	20	36	29	0.342
	No	31	33	29	
	UKN	36	51	32	
Hypertension	Yes	39	51	44	0.970
	No	17	24	21	
	UKN	31	45	25	
Cholesterol	Yes	23	36	39	0.325
	No	24	34	24	
	UKN	40	50	27	
Tinnitus	Yes	26	43	42	0.260
	No	24	25	21	
	UKN	37	52	27	
Ototoxic Medication	Yes	10	12	9	0.602
	No	22	39	33	
	UKN	55	69	48	

UKN: Missing Values

Table 4.25: Distribution of subjects according to genetic features in cluster groups obtained by K-Means.

Characteristic		Configuration			p-value
		Cluster 1: "HFGS" (n=87)	Cluster 2: "HFSS" (n=120)	Cluster 3: "FLAT" (n=90)	
NAT2 phenotype	I	23	31	25	0.635
	R	7	4	5	
	S	32	51	36	
	UKN	25	34	24	
GRM7 genotype	A/A	5	7	3	0.900
	A/T	29	43	33	
	T/T	51	65	53	
	UKN	2	5	1	

UKN: Missing Values

Although no significant difference in prevalence of configurations was found by age group ($\chi^2 = 5.849, df = 8, p = 0.664$), considering each one of the cluster patterns there were found significant differences between age groups in hearing intensity thresholds at each frequency (excepting for cluster 2 ("HFSS" pattern) in frequencies of 4000 Hz, $p = 0.117$ and 8000 Hz, $p = 0.065$), meaning that for the same pattern, the HL quantity, i.e., the measured intensities were different depending on subject age. In mean, as age increases, the pattern remains the same, but the hearing intensity thresholds become higher, indicating a HL occurrence (Figure 4.20).

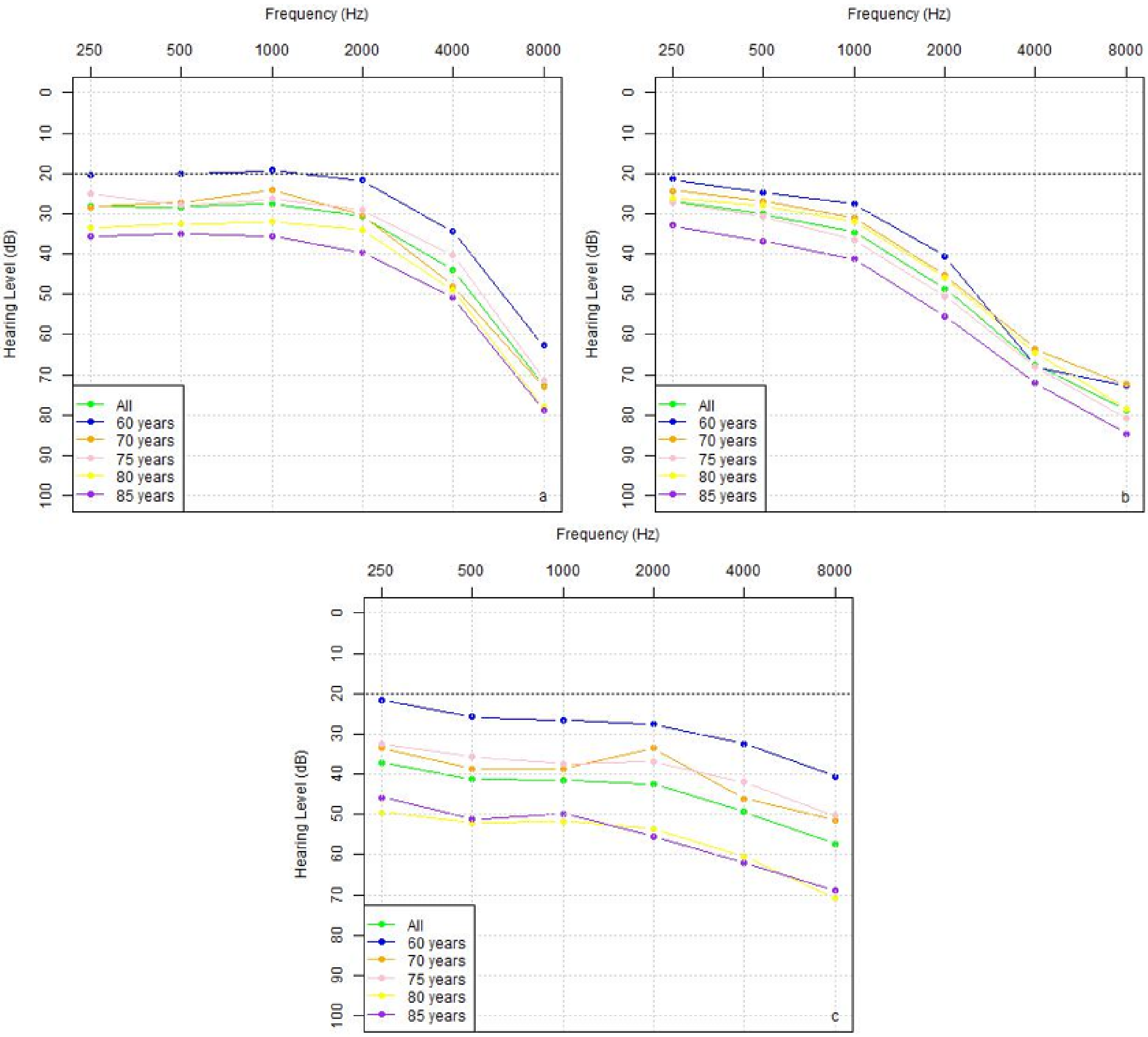

Figure 4.20: Mean hearing intensity thresholds audiogram for: a) Cluster 1 ("HFGS" pattern), b) Cluster 2 ("HFSS" pattern) and c) Cluster 3 ("FLAT" pattern) accordingly with age (K-Means).

In Table 4.26 is presented the individual's distribution in a cluster depending on the HL degree. Most of individuals in cluster 1 ("HFGS" pattern) and 3 ("FLAT" pattern) had a mild HL, whereas the most frequent HL level in cluster 2 ("HFSS" pattern) was the moderate.

Table 4.26: Distribution of K-Means cluster groups in HL levels based on PTA calculated for the RE averaged over 0.5, 1, 2, and 4 kHz.

	Cluster 1: "HFGS", n (%)	**Cluster 2: "HFSS", n (%)**	**Cluster 3: "FLAT", n (%)**
Normal ($\leq$25db HL)	29 (33.3)	8 (6.7)	16 (17.8)
Mild (25 $<$ db HL $\leq$40)	38 (43.7)	41 (34.2)	31 (34.4)
Moderate (40 $<$ db HL $\leq$60)	17 (19.5)	57 (47.5)	27 (30.0)
Severe/Profound ($\geq$60 db HL)	3 (3.4)	14 (11.7)	16 (17.8)
Total	87 (100)	120 (100)	90 (100)

4.3.3 Comparison of the results obtained by the different methodologies

The distribution of the individuals in clusters in both methods were clearly associated with Wuyts classification methodology [3] presented in Subsection 4.3.1 (Table 4.27). All of those with a FLAT configuration determined by Wuyts classification [3] were allocated to the cluster 3 ("FLAT" pattern) in K-Means algorithm. This value was 90.7% for hierarchical clustering. Regarding HFSS configuration, 69.7% were placed in cluster 2 ("HFSS" pattern) in K-means clustering and 83.3% in cluster 1 ("HFSS" pattern) of hierarchical clustering. Finally, the individuals with HFGS configuration were distributed in all clusters, being that in K-Means the majority (45.5%) was allocated to the cluster 1 ("HFGS" pattern), whereas in the hierarchical method, 72.7% of individuals were placed in the cluster 2 ("HFGS" pattern).

Table 4.27: Comparison of the three classification methods employed to find audiogram patterns in audiological data set.

		Configuration								
		None	FLAT	HFGS	HFSS	LFA	MFU	MFRU	Mixed	
		3	43	88	152	0	0	1	10	
Clustering Method										**#**
K-means	Cluster 1: "HFGS"	0	0	40	46	0	0	0	1	87
	Cluster 2: "HFSS"	0	0	14	106	0	0	0	0	120
	Cluster 3: "FLAT"	3	43	34	0	0	0	1	9	90
Hierarchical	Cluster 1: "HFSS"	0	0	13	127	0	0	0	2	142
	Cluster 2: "HFGS"	0	4	64	25	0	0	0	4	97
	Cluster 3: "FLAT"	3	39	11	0	0	0	1	4	58

These results sugest that both approachs, K-Means and hierarchical clustering, recurring to the audiological data transformed, in a very reasonable way, could separate the individuals taken into account only the audiogram shape.

4.4 Principal Component Analysis (PCA)

The purpose of PCA was to find the best low-dimensional representation of the variation in the audiological data set. In this study, from the set of six frequencies (variables) correlated that described the audiogram of an individual the main objective was to find a new set of variables (PCs) with no correlation that were a linear combination of the first ones that allow to explain the data variability without loss of information. PCA involves the calculation of the correlation matrix eigenvalues, each associated to one PC. They represent the contribution of each component to the total variability of the sample, indicating the proportion of the total variance which is explained by each of the PCs. The estimated eigenvalues of the correlation matrix are presented in Table 4.28. As can be seen, the first PC (PC1) explained a high proportion of variance (74.88%).

Table 4.28: Eigenvalues and variance explained by each PC.

Principal Components (j)	Eigenvalues (λ_j)	% Variance Explained by j PC	% Cumulative Proportion Explained
PC1	4.492	74.875	74.875
PC2	0.793	13.211	88.086
PC3	0.362	6.040	94.126
PC4	0.180	2.992	97.118
PC5	0.111	1.857	98.975
PC6	0.062	1.025	100

Since the main objective of this type of analysis is the dimensionality reduction, the next step consisted on the determination of the number of PCs to be retained. For that, the three criteria's presented in the Section 3.3.2 were used. Accordingly with modified Keiser's rule the components to be retained would be those with an eigenvalue higher than 0.7, PC1 and PC2. Another criteria was based on the selection of as many components as necessary to explain 70% to 80% of total variance. As seen in Table 4.28, the first two PCs accounted for more than 80% of the total variance. Finally, when analysing the scree-plot (Figure 4.21) the number of components to be retained must be two, since the eigenvalues corresponding to the following PCs were approximately equal and small which indicates a small contribution to explain the variability in the data.

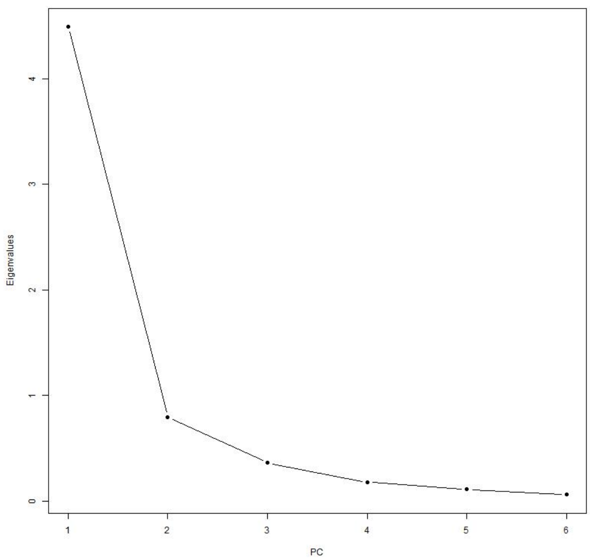

Figure 4.21: Scree-plot of eigenvalues vs number of PC.

The estimated eigenvectors ($\mathbf{a}_j$) for the $j-th$ component are presented in Table 4.29.

Table 4.29: Estimated eigenvectors for each PC.

	Components	
Frequency (Variable)	PC1	PC2
250 Hz	0.413	0.399
500 Hz	0.424	0.422
1000 Hz	0.437	0.257
2000 Hz	0.427	-0.153
4000 Hz	0.395	-0.463
8000 Hz	0.348	-0.599

Thus, each individual score could be determinded by expressions 4.1a) and 4.1b) on component 1 and component 2, respectively.

$$PC1 = 0.413F250 + 0.424F500 + 0.437F1000 + 0.427F2000 + 0.395F4000 + 0.348F8000 \quad (4.1a)$$

$$PC2 = 0.399F250 + 0.422F500 + 0.257F1000 - 0.153F2000 - 0.463F4000 - 0.599F8000 \quad (4.1b)$$

The eigenvectors or weights of PC1 were all positive and approximately equal suggesting that the main source of variability among patients was the overall degree of HL. Thus, it implies that individuals suffering HL at certain frequencies will more than likely suffer loss at the other frequencies as well. Regarding PC2, the eigenvectors were positive for frequencies at or below 1000 Hz, but negatives for higher frequencies. Thus, it differentiated subjects according to whether they have a predominantly high frequency or low frequency HL. As we known, in presbycusis, HL is more evident at high frequencies. A negative score on PC2 would correspond to a sloping audiogram pattern, which will be more pronounced as the difference between low and high frequencies increases (similar to a pattern of HFSS configuration). On the other hand, if the difference between low and high frequencies is lower, the sloping pattern will resemble to a FLAT configuration. The correlations or loadings between each PC and the variables are presented in Table 4.30.

Table 4.30: Correlations (loadings)($\rho_{ij} = l_{ij}$) between each PC and original variables.

	Components	
Frequency (Variable)	PC1	PC2
250 Hz	**0.875**	0.355
500 Hz	**0.898**	**0.376**
1000 Hz	**0.927**	0.229
2000 Hz	**0.904**	-0.136
4000 Hz	0.836	**-0.412**
8000 Hz	0.737	**-0.534**

The higher correlations (loadings), in absolute terms, are related to the variables which have a higher contribute and thus are the most important for a given component. As shown, the lower and medium frequencies were the ones that had a higher contribute to the explanation of the first PC ($|\rho_{ij}| \geq \sqrt{\frac{4.492}{6}} = 0.865$), whereas the higher frequencies negatively correlated, and the frequency of 500 Hz, positively correlated were the ones which were more important for the second PC ($|\rho_{ij}| \geq \sqrt{\frac{0.793}{6}} = 0.364$).

In order to understand the distribution of the subjects in the new space, a distance biplot was drawn (Figure 4.22). It represents the scores of a given subject associated to each principal component, and it also contains the representation of the loading vectors with the indication of the configuration obtained in the previous section by K-Means algorithm. As observed, the individuals were well distributed in the space, as well that individuals with positive scores for PC2 were associated to the cluster 3 which have a typical pattern of a FLAT configuration, while individuals of cluster 2 which is related to a HFSS configuration presented negative scores. Individuals who belongs to cluster 1, a pattern close to a HFGS configuration were found between the other two cluster, having both positive and negative scores.

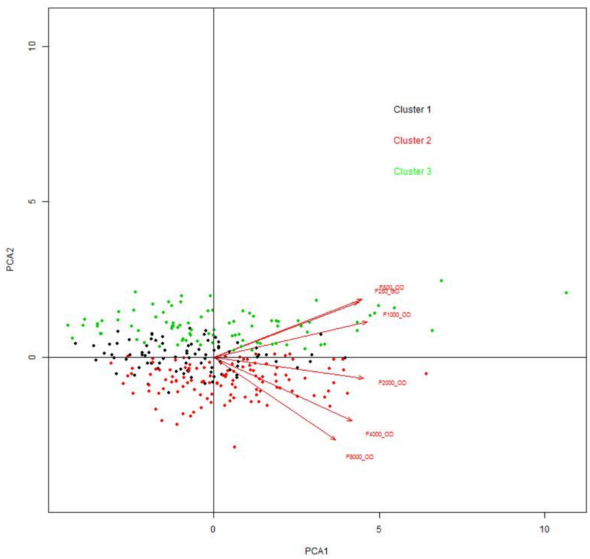

Figure 4.22: Scores and loadings representation (scale factor of 5) - Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern

4.5 Linear Discriminant Analysis (LDA)

The two objectives of LDA were to find the linear combinations of the original frequencies, as well of the two principal components that gives the best possible separation between the clusters resulted from K-Means method in the data set in order to have a rule which could be used to allocate new individuals, based on their characteristics, in each cluster. Thus, it was expected the evaluation of the reliability of the cluster groups obtained through K-Means and consequently the prediction of the audiogram pattern of newly set of six frequencies.

Once the number of groups was three, it was necessary to define two Discriminant Functions (DFs): the first one to distinguish the first group from second and third groups and the second to distinguish the first and second groups from the third group. Defining π_i as the prior probability of belonging to a given cluster $i = 1, 2, 3$, their estimates were $\hat{\pi}_1 = 0.293$, $\hat{\pi}_2 = 0.404$ and $\hat{\pi}_3 = 0.303$.

4.5.1 Original Data

Firstly, it was investigated if the predictors, in this case, the frequencies varied sufficiently over the different clusters. Based on the results of Wilks test ($\lambda = 0.531$, $p - value < 0.001$), the discriminant model was not rejected. Table 4.31 contains the estimated coefficients of the two DFs.

Table 4.31: Estimated coefficients for each frequency in each linear DF.

	DF	
Frequency	LD1	LD2
250 Hz	1.013	0.578
500 Hz	0.581	-0.401
1000 Hz	0.544	-0.103
2000 Hz	-0.285	-0.770
4000 Hz	-1.347	-1.115
8000 Hz	-0.900	1.506

Regarding these results, each DF could be written by equation 4.2a) and equation 4.2b):

$$\delta_1 = 1.013F250 + 0.581F500 + 0.544F1000 - 0.285F2000 - 1.347F4000 - 0.900F8000 \quad (4.2a)$$

$$\delta_2 = 0.578F250 - 0.401F500 - 0.103F1000 - 0.770F2000 - 1.115F4000 + 1.506F8000 \quad (4.2b)$$

The percentages of between-group variance that the first DF (LD1) and the second DF (LD2) were able to explain from the total amount of between-group variance were 75.85% and 24.15%, respectively. As shown in Figure 4.23, although there was some overlap, the two DFs separate reasonable the cluster groups, once the most of the K-Means clusters fall within the boundaries of the matching predicted CM. Moreover, it seemed that clusters 2 (“HFSS” pattern) and 3 (“FLAT” pattern) were rather different, while cluster 1 (“HFGS” pattern) was less separable from the other two. The overall percentage of individuals correctly classified through LDA was 97.31% (Table 4.32), being that the procedure was more likely to misclassify patterns of cluster 1 (“HFGS” pattern) given the lower classification rate (93.1%).

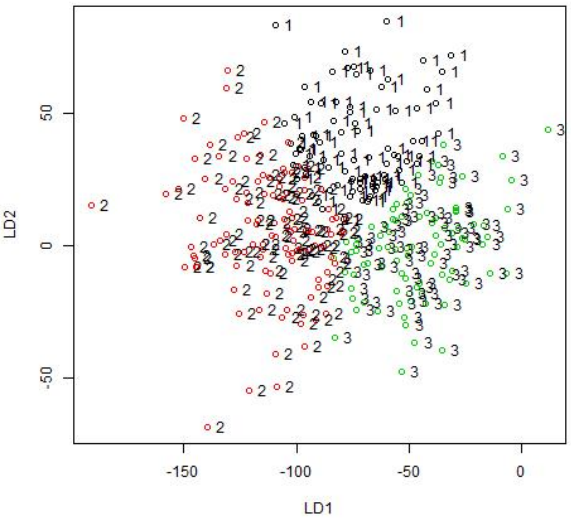

Figure 4.23: Discriminant function values of subjects for predicted groups for audiological data set (Cluster 1: “HFGS” pattern, Cluster 2: “HFSS” pattern and Cluster 3: “FLAT” pattern)

Table 4.32: Confusion matrix for LDA in audiological data: Predicted Groups vs K-Means Clusters.

	Clusters		
Predicted Groups	“HFGS” pattern”	“HFSS” pattern	“FLAT” pattern
“HFGS” pattern”	81	0	2
“HFSS” pattern	6	120	0
“FLAT” pattern	0	0	88

4.5.2 PCA Data

The second part of LDA consisted on the determination of a rule that can separate the clusters taken into account the scores of each PC.

Based on the results of Wilks test ($\lambda = 0.772$, $p - value < 0.001$), the discriminant model was not rejected. The estimated coefficients of the two linear DFs are presented in Table 4.33.

Table 4.33: Estimated coefficients for each PC in each linear DF.

	DF	
Frequency	LD1	LD2
PC1	-0.037	-0.484
PC2	1.995	-0.050

Thus each one of the DF could be written by equation 4.3a) and equation 4.3b):

$$\delta_1 = -0.037PC1 + 1.995PC2 \quad (4.3a)$$

$$\delta_2 = -0.484PC1 - 0.050PC2 \quad (4.3b)$$

The percentages of between-group variance that the first DF (LD1) and the second DF (LD2) were able to explain from the total amount of between-group variance were 97.31% and 2.69%, respectively. As shown in Figure 4.24 the DF values of subjects were not well separated according to predicted groups in comparison with the previous LDA. In fact, the overall percentage of individuals correctly classified was 77.1% (Table 4.34) a lower value when comparing with the previous one, being the classification rate of cluster 1 ("HFGS" pattern) only 60.9%.

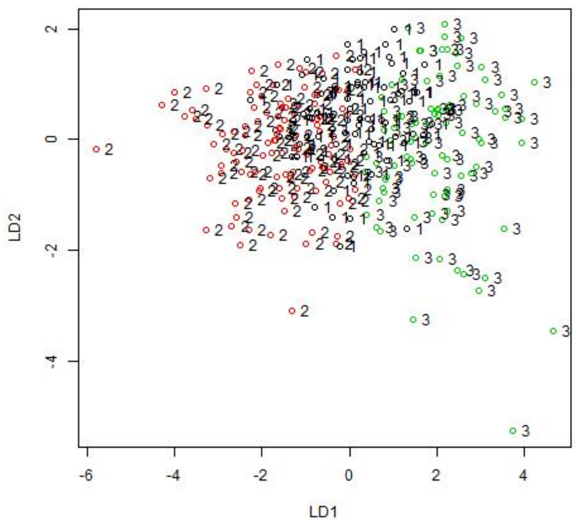

Figure 4.24: Discriminant function values of subjects for predicted groups for scores data (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

Table 4.34: Confusion matrix for LDA in scores data: Predicted Groups vs K-Means Clusters.

	Clusters		
Predicted Groups	"HFGS" pattern	"HFSS" pattern	"FLAT" pattern
"HFGS" pattern	53	20	14
"HFSS" pattern	23	100	0
"FLAT" pattern	11	0	76

4.6 Multinomial Logistic Regression Model

After the formation of the three groups, it was applied a multinomial logistic regression to examine the association between demographic, environmental, medical and genetic characteristics (independent variables) and Cluster Membership (CM) (response variable). For the response variable, the clusters considered were those obtained by K-Means clustering. The independent variables were then analysed for statistical significance in the adjusted models. Once the gender was associated with audiogram pattern, all models were adjusted controlled by it. The referent group for all analyses was the cluster representing the least number of individuals: cluster 1 ("HFGS" pattern). The results are presented in Table 4.35 and Table 4.36.

Table 4.35: Adjusted multinomial model for individual characteristics controlled by gender comparing cluster 2 ("HFSS" pattern) with cluster 1 ("HFGS" pattern).

Model	Variable	$\hat{\beta}$	s.e.	$exp(\hat{\beta})$	C.I.$_{95\%}$	z$_{obs}$	p-value	Regression characteristics
$CM \sim Gender + Age$	Gender: Female Age	-0.935 0.024	0.320 0.018	0.393 1.025	[0.210,0.735] [0.990,1.062]	-2.923 1.384	0.003 0.166	N=297 Pseudo R-squared=0.03 AIC=638.37 $\chi_2 = 19.68$ p-value$<$0.001
$CM \sim Gender + NoiseExposure$	Gender: Female Noise Exposure:Yes	-0.832 0.131	0.343 0.309	0.435 1.140	[-1.503,-0.160] [-0.475,0.736]	-2.427 0.424	0.015 0.672	N=272 Pseudo R-squared=0.11 AIC=585.30 $\chi_2 = 72.75$ p-value$<$0.001
$CM \sim Gender + FamilialHistory$	Gender: Female Familial History:Yes	-0.916 0.657	0.413 0.387	0.400 1.93	[-1.725,-0.107] [-0.101,1.416]	-2.218 1.699	0.027 0.089	N=178 Pseudo R-squared=0.43 AIC=381.11 $\chi_2 = 276.93$ p-value$<$0.001
$CM \sim Gender + Tinnitus$	Gender: Female Tinnitus:Yes	-0.782 0.508	0.409 0.385	0.457 1.662	[-1.583,0.019] [-0.246,1.262]	-1.914 1.320	0.056 0.187	N=181 Pseudo R-squared=0.41 AIC=390.66 $\chi_2 = 267.38$ p-value$<$0.001
$CM \sim Gender + OtotoxicMedication$	Gender: Female Ototoxic Medication:Yes	-0.679 -0.407	0.475 0.510	0.507 0.666	[-1.610,0.252] [-1.406,0.592]	-1.429 -0.798	0.153 0.425	N=125 Pseudo R-squared=0.60 AIC=272.87 $\chi_2 = 385.17$ p-value$<$0.001
$CM \sim Gender + Hypertension$	Gender: Female Hypertension:Yes	-1.002 -0.062	0.401 0.390	0.367 1.946	[-1.788,-0.217] [-0.825,0.702]	-2.502 -0.158	0.012 0.874	N=196 Pseudo R-squared=0.37 AIC=422.08 $\chi_2 = 235.97$ p-value$<$0.001
$CM \sim Gender + Cholesterol$	Gender: Female Cholesterol:Yes	-0.914 0.154	0.433 1.166	0.401 1.946	[-1.763,-0.065] [-0.599,0.907]	-2.110 0.401	0.035 0.689	N=180 Pseudo R-squared=0.42 AIC=387.13 $\chi_2 = 270.91$ p-value$<$0.001
$CM \sim Gender + NAT2\ Phenotype$	Gender: Female NAT2 Phenotype:R NAT2 Phenotype:S	-0.905 -0.938 0.208	0.388 0.698 0.362	0.405 0.391 1.231	[-1.666,-0.144] [-2.307,0.430] [-0.502,0.917]	-2.329 -1.344 0.574	0.020 0.179 0.566	N=214 Pseudo R-squared=0.31 AIC=463.74 $\chi_2 = 198.30$ p-value$<$0.001
$CM \sim Gender + GRM7\ Genotype$	Gender: Female GRM7 Genotype:A/T GRM7 Genotype:T/T	-0.909 0.077 -0.056	0.322 0.645 0.626	0.403 1.080 0.945	[-1.540,-0.278] [-1.187,1.1340] [-1.284,1.171]	-2.823 0.119 -0.090	0.005 0.905 0.929	N=214 Pseudo R-squared=0.05 AIC=627.20 $\chi_2 = 34.84$ p-value$<$0.001

Table 4.36: Adjusted multinomial model for individual characteristics controlled by gender comparing cluster 3 ("FLAT" pattern) with cluster 1 ("HFGS" pattern).

Model	Variable	$\hat{\beta}$	s.e.	$exp(\hat{\beta})$	C.I.$_{95\%}$	z$_{obs}$	p-value	Regression characteristics
$CM \sim Gender + Age$	Gender: Female	0.360	0.384	1.433	[0.676,3.041]	0.938	0.348	N=297 Pseudo R-squared=0.03 AIC=638.37 $\chi_2 = 19.68$ p-value$<$0.001
	Age	0.016	0.019	1.016	[0.980,1.054]	0.863	0.388	
$CM \sim Gender + NoiseExposure$	Gender: Female	0.184	0.408	1.202	[-0.615,0.983]	0.451	0.652	N=272 Pseudo R-squared=0.11 AIC=585.30 $\chi_2 = 72.75$ p-value$<$0.001
	Noise Exposure:Yes	-0.616	0.345	0.540	[-1.292,0.060]	-1.785	0.074	
$CM \sim Gender + FamilialHistory$	Gender: Female	0.857	0.520	2.356	[-0.161,1.875]	1.649	0.099	N=178 Pseudo R-squared=0.43 AIC=381.11 $\chi_2 = 276.93$ p-value$<$0.001
	Familial History:Yes	0.358	0.394	1.43	[-0.414,1.130]	0.909	0.363	
$CM \sim Gender + Tinnitus$	Gender: Female	0.709	0.486	2.032	[-0.244,1.662]	1.458	0.145	N=181 Pseudo R-squared=0.41 AIC=390.66 $\chi_2 = 267.38$ p-value$<$0.001
	Tinnitus:Yes	0.584	0.392	1.793	[-0.184,1.351]	1.490	0.136	
$CM \sim Gender + OtotoxicMedication$	Gender: Female	0.648	0.549	1.912	[-0.429,1.725]	1.180	0.238	N=125 Pseudo R-squared=0.60 AIC=272.87 $\chi_2 = 385.17$ p-value$<$0.001
	Ototoxic Medication:Yes	-0.498	0.538	0.608	[-1.553,0.557]	-0.925	0.355	
$CM \sim Gender + Hypertension$	Gender: Female	0.666	0.499	1.946	[-0.312,-1.643]	1.335	0.182	N=196 Pseudo R-squared=0.37 AIC=422.08 $\chi_2 = 235.97$ p-value$<$0.001
	Hypertension:Yes	-0.097	0.395	0.907	[-0.871,0.677]	-0.246	0.806	
$CM \sim Gender + Cholesterol$	Gender: Female	0.584	0.522	1.793	[-0.440,1.607]	1.118	0.264	N=180 Pseudo R-squared=0.42 AIC=387.13 $\chi_2 = 270.91$ p-value$<$0.001
	Cholesterol:Yes	0.505	0.392	1.657	[-0.264,1.273]	1.288	0.198	
$CM \sim Gender + NAT2\ Phenotype$	Gender: Female	0.634	0.491	1.885	[-0.329,1.597]	1.290	0.197	N=214 Pseudo R-squared=0.31 AIC=463.74 $\chi_2 = 198.30$ p-value$<$0.001
	NAT2 Phenotype:R	-0.380	0.657	0.684	[-1.667,0.908]	-0.578	0.563	
	NAT2 Phenotype:S	0.016	0.379	1.016	[-0.727,0.759]	0.042	0.966	
$CM \sim Gender + GRM7\ Genotype$	Gender: Female	0.349	0.385	1.417	[-0.406,1.103]	0.905	0.365	N=214 Pseudo R-squared=0.05 AIC=627.20 $\chi_2 = 34.84$ p-value$<$0.001
	GRM7 Genotype:A/T	0.634	0.774	1.884	[-0.884,2.151]	0.818	0.413	
	GRM7 Genotype:T/T	0.538	0.757	1.712	[-0.946,2.021]	0.710	0.478	

After the adjustment of these models, the variables were excluded and a likelihood ratio test was performed to investigate if the exclusion of the variable was significant. Based on the results, it was concluded that the model with the independent variables gender and noise exposure was the most parsimonious.

$$\eta_{i2} = log\left(\frac{\pi_i^{(2)}}{\pi_i^{(1)}}\right) = 0.831 - 0.832(Gender = F) + 0.131(NoiseExposure = Yes) \tag{4.4a}$$

$$\eta_{i3} = log\left(\frac{\pi_i^{(3)}}{\pi_i^{(1)}}\right) = 0.094 + 0.184(Gender = F) - 0.616(NoiseExposure = Yes) \tag{4.4b}$$

Based on this model it was possible to conclude:

- Cluster 2 (“HFSS” pattern) relative to cluster 1 (“HFGS” pattern): If a subject is female, the risk of being in the cluster 2 (“HFSS” pattern) would be 56.5% lower than the risk for males, when the noise exposure variable is held constant. For subjects exposed to the noise relative to non-exposed, the relative risk for being in cluster 2 (“HFSS” pattern) would be expected to increase 14% given the gender variable in the model constant.
- Cluster 3 (“FLAT” pattern) relative to cluster 1 (“HFGS” pattern): The expected risk of females being in cluster 3 (“FLAT” pattern) is 20.2% higher than the risk of males, keeping the noise exposure variable constant. Exposed subjects have a risk 46% lower than the risk for non-exposed subjects of being in cluster 3 (“FLAT” pattern), given the gender variable in the model is held constant.

4.7 Linear Regression Model

Regarding the results obtained in the previous sections, it was considered of interest to study if the prediction of the amount of HL was influenced by the audiogram pattern, knowing yet that HL ocurrence was significantly associated with age and the audiogram pattern was significantly associated with gender. The main objective was to investigate if the amount of HL was the same for each pattern or if it varies depending on age according to the pattern. Thus, linear regression models were estimated considering as the response variable the mean of HL, given by $PTA_{0.5,1,2,4kHz}$ in RE, and as predictors variables: age, gender and CM.

$$\begin{aligned}PTA_{0.5,1,2,4kHz_i} &= \beta_0 + \beta_1(Gender_i = Female) + \beta_2 Age_i + \beta_3(Cluster_i = 1) + \\ &\beta_4(Cluster_i = 2) + \beta_5 Age_i \times (Cluster_i = 1) + \beta_6 Age_i \times (Cluster_i = 2) + \varepsilon_i, i = 1, \ldots, n\end{aligned} \tag{4.5}$$

Table 4.37: Adjusted multiple linear regression model for $PTA_{0.5,1,2,4kHz}$ response considering age$*$CM interaction and gender.

Model	Variable	$\hat{\beta}$	s.e.	C.I.$_{95\%}$	t$_{obs}$	p-value	Regression characteristics
$PTA_{0.5,1,2,4kHz} \sim Gender + Age + CM + Age * CM$	Intercept	-39.01	13.33	[-65.25,-12.77]	-2.93	0.004	N=297
	Gender= Female	-0.59	1.91	[-4.34,3.17]	-0.31	0.76	Adjusted R-squared=0.25
	Age	1.06	0.17	[0.73,1.39]	6.28	$<$0.001	AIC=2427.64
	Cluster=1	12.00	20.28	[-27.92,51.91]	0.59	0.55	F$_{(6,290)} = 17.77$
	Cluster=2	48.80	18.57	[12.25,85.35]	2.63	0.009	p-value$<$0.001
	Age:(Cluster=1)	-0.28	0.26	[-0.79,0.23]	-1.09	0.28	
	Age:(Cluster=2)	-0.60	0.24	[-1.07,-0.15]	-2.57	0.011	

Replacing the coefficients estimates in equation 4.5:

$$PTA_{0.5,1,2,4kHz_i} = -39.01 - 0.59(Gender_i = Female) + 1.06Age_i + 12.00(Cluster_i = 1)+ \\ 48.80(Cluster_i = 2) - 0.28Age_i \times (Cluster_i = 1) - 0.60Age_i \times (Cluster_i = 2) + \varepsilon_i, i = 1, \ldots, n \quad (4.6)$$

The influence of gender predictor on the mean quantity of HL was not significant, meaning that $PTA_{0.5,1,2,4kHz}$ was not affected by individual's gender. Adittionally to the little effect of gender on response variable, according with Figure 4.25 it seemed clearly the lack of an interaction among gender and CM, once the lines associated with each cluster pattern were vertically equidistant and parallel. In fact according with the model, the estimated average difference in $PTA_{0.5,1,2,4kHz}$ for men and women with the same audiogram pattern was 0.59.

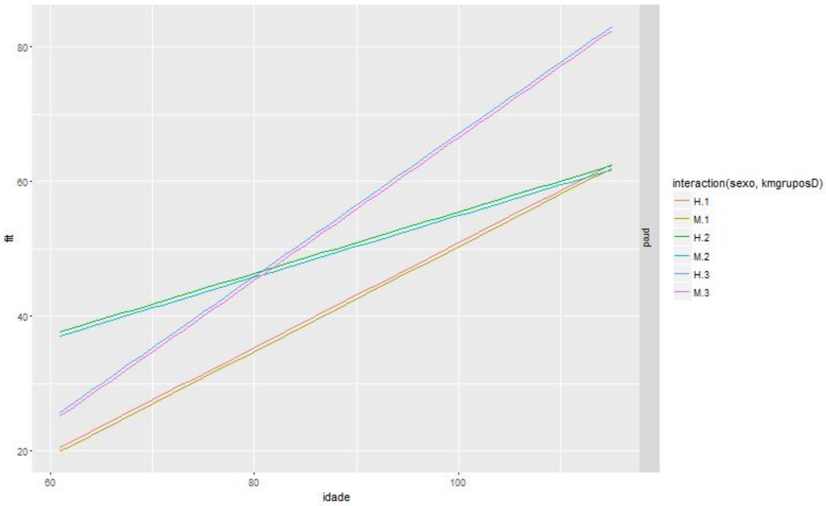

Figure 4.25: Interaction plot of variables included in the model with age*CM interaction and gender (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

Thus, the gender variable was removed and it was observed that its exclusion was not important to the model ($p = 0.76$) having a little effect on the prediction of the response variable. The results for the adjusted model after gender exclusion and the interaction plot are presented in Table 4.38 and Figure 4.26, respectively.

Table 4.38: Adjusted multiple linear regression model for $PTA_{0.5,1,2,4kHz}$ response considering age$*$CM interaction.

Model	Variable	$\hat{\beta}$	s.e.	C.I.$_{95\%}$	t$_{obs}$	p-value	Regression characteristics
$PTA_{0.5,1,2,4kHz} \sim Age + CM + Age * CM$	Intercept	-39.27	13.28	[-65.42,-13.13]	-2.96	0.003	N=297
	Age	1.06	0.17	[0.73,1.39]	6.28	$<$0.001	Adjusted R-squared=0.26
	Cluster=1	11.86	20.24	[-27.98,51.70]	0.59	0.56	AIC=2425.74
	Cluster=2	49.36	18.45	[13.04,85.68]	2.68	0.008	F$_{(5,291)} = 21.37$
	Age:(Cluster=1)	-0.28	0.26	[-0.79,0.23]	-1.08	0.28	p-value$<$0.001
	Age:(Cluster=2)	-0.61	0.23	[-1.07,-0.15]	-2.60	0.01	

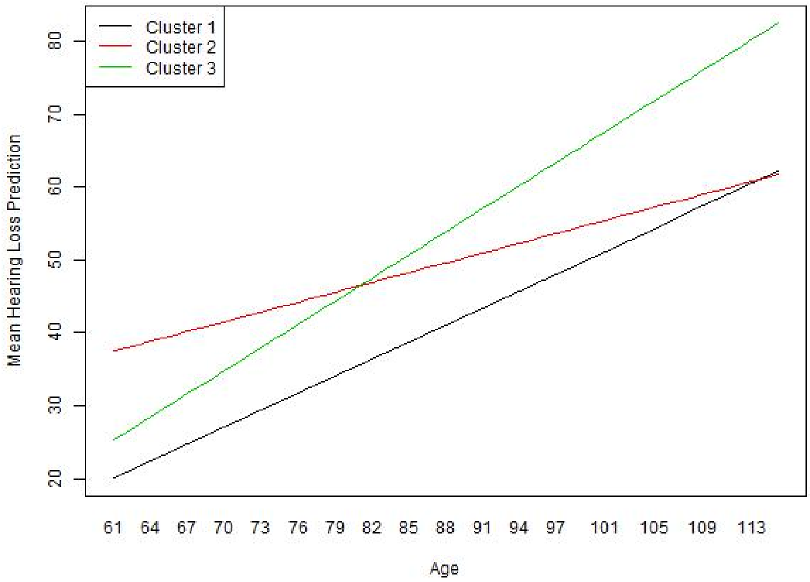

Figure 4.26: Interaction plot of variables included in the model with age*CM interaction (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

Based on this model, age seemed to be a strong predictor of the amount of HL. The influence of cluster 2 ("HFSS" pattern) and also the interaction of cluster 2 ("HFSS" pattern) and age were also importants on the prediction of the $PTA_{0.5,1,2,4kHz}$. To a better interpretation of the model, consider the following equation:

$$\begin{aligned} PTA_{0.5,1,2,4kHz_i} &= -39.27 + 1.06Age_i + 11.86(Cluster_i = 1) + 49.36(Cluster_i = 2) \\ &- 0.28Age_i \times (Cluster_i = 1) - 0.61Age_i \times (Cluster_i = 2) + \varepsilon_i, i = 1, \ldots, n \end{aligned} \tag{4.7}$$

- For individuals with audiogram pattern of cluster 1 ("HFGS") (Cluster1=1, Cluster2=0 and Cluster3=0):

$$PTA_{0.5,1,2,4kHz_i} = -27.41 + 0.78Age_i + \varepsilon_i, i = 1, \ldots, n \tag{4.8}$$

  The effect of age on $PTA_{0.5,1,2,4kHz}$ is 0.78. So, for two individuals with the pattern of cluster 1 ("HFGS"), the older individual would be expected to have a $PTA_{0.5,1,2,4kHz}$ 0.78 dB higher than the individual one year younger.

- For individuals with audiogram pattern of cluster 2 ("HFSS") (Cluster1=0, Cluster2=1 and Cluster3=0):

$$PTA_{0.5,1,2,4kHz_i} = 10.09 + 0.45Age_i + \varepsilon_i, i = 1, \ldots, n \tag{4.9}$$

  The increase effect of one-unit age on $PTA_{0.5,1,2,4kHz}$ in cluster 2 ("HFSS" pattern) is slower than cluster 1 ("HFGS" pattern) and cluster 3 ("FLAT" pattern), being 0.45. That is, $PTA_{0.5,1,2,4kHz}$ increases 0.45 dB for one-unit increase in individual's age.

- For individuals with audiogram pattern of cluster 3 ("FLAT") (Cluster1=0, Cluster2=0 and Cluster3=1):

$$PTA_{0.5,1,2,4kHz_i} = -39.27 + 1.06Age_i + \varepsilon_i, i = 1, \ldots, n \quad (4.10)$$

  Cluster 3 ("FLAT" pattern) is the group that presents a faster increase on $PTA_{0.5,1,2,4kHz}$ with aging. That is, for one-unit increase in individual's age $PTA_{0.5,1,2,4kHz}$ increases 1.06 dB.

According to equations 4.8, 4.9 and 4.10 the amount of HL increases with aging for all cluster patterns. However, the rate of this increase is different for each one of them. It is expected that individuals with a pattern characteristic of cluster 1 ("HFGS") will be present lower values of $PTA_{0.5,1,2,4kHz}$ than the ones with patterns of clusters 2 ("HFSS") and 3 ("FLAT") at each age. Until 81 years old, it will be expected that individuals with pattern of cluster 2 ("HFSS") present higher values for $PTA_{0.5,1,2,4kHz}$ than individuals with pattern of cluster 3 ("FLAT"). After that, an inversion occurs, being the individuals in cluster 3 ("FLAT" pattern) the ones with higher $PTA_{0.5,1,2,4kHz}$ at each age.

To investigate the adequacy of the previous model (consider its name as model 1), a residual analysis was performed. This type of analysis is very important to evaluate if the underlying assumptions of the model regarding the random component (error) are fulfilled. The assumption concerning the normality of the error distribution was checked with a histogram and a normal quantile plot (QQ-plot) of the residuals (Figure 4.27).

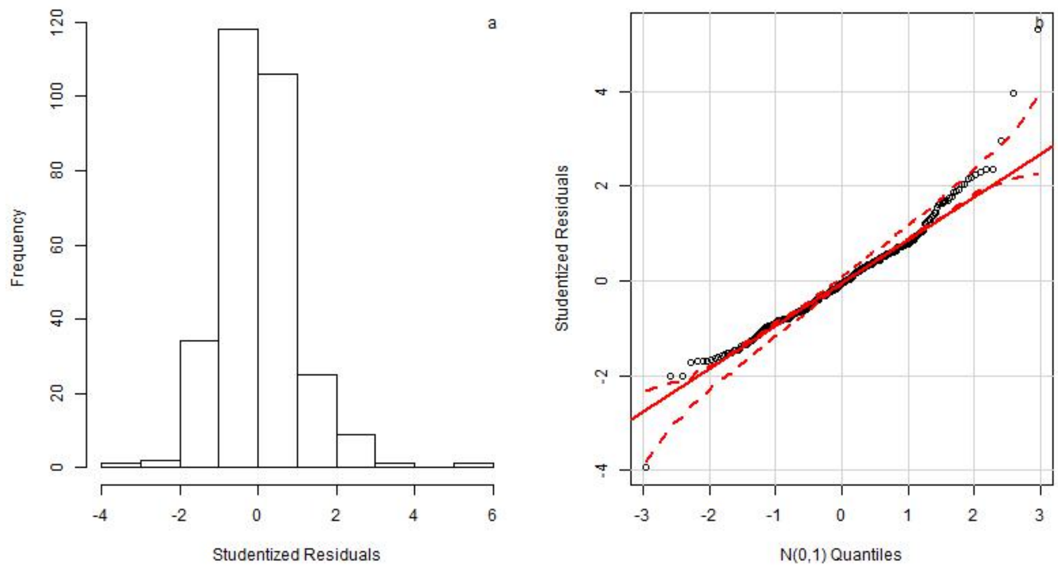

Figure 4.27: Distribution of the residuals of model 1: a) histogram and b) QQ-plot.

In general, the points were symmetrically distributed, however it can be observed that some points were displaced from the line in the QQ-plot, as well, in the left and right sides of the histogram a slight extension was presented. The linearity and homoscedasticity assumption was checked with a plot of the studentized residuals against fitted values (Figure 4.28a)). Moreover, to check if the independence assumption was met, a plot of the studentized residuals against observation number was drawn (Figure 4.28b))

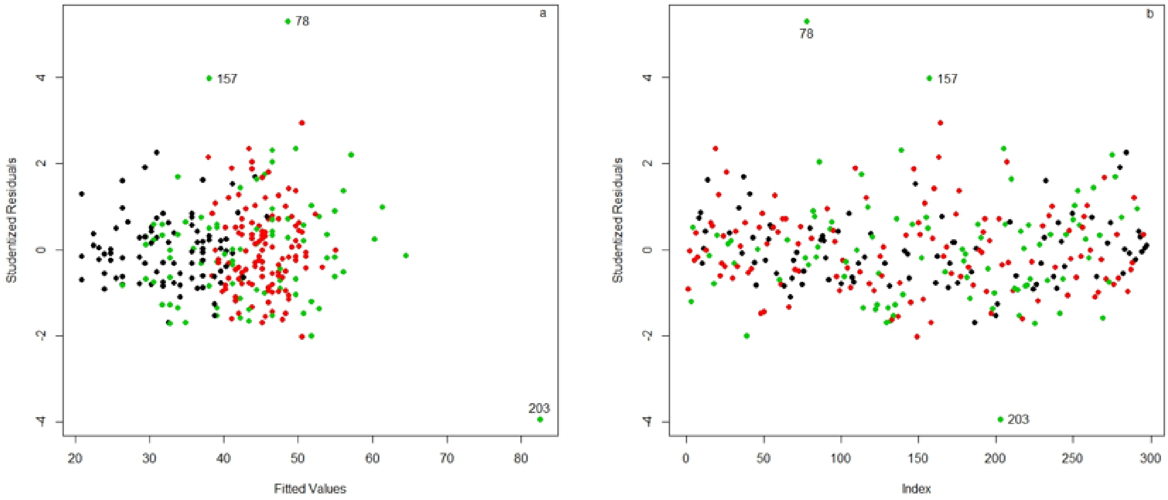

Figure 4.28: a) Studentized residuals vs fitted values and b) Studentized residuals vs observation number of the individuals for model 1.

According to Figure 4.28a), the majority of the residuals were scattered randomly around zero and no pattern in the residuals was presented. This suggest that homoscedasticity assumption was not violated, that is, the errors had a constant variance. Only the residuals of the observations number 78, 157 and 203 were not close to the others, having a higher value. These observations were potential outliers and depending on their influence, the exclusion from the model was considered. Regarding Figure 4.28b), it can be observed the random distribution of the residuals indicating that there was no correlation between the errors of the observations. To detect the existence of influential observations in the model, a plot of Cook's distance against the observation number was drawn (Figure 4.29).

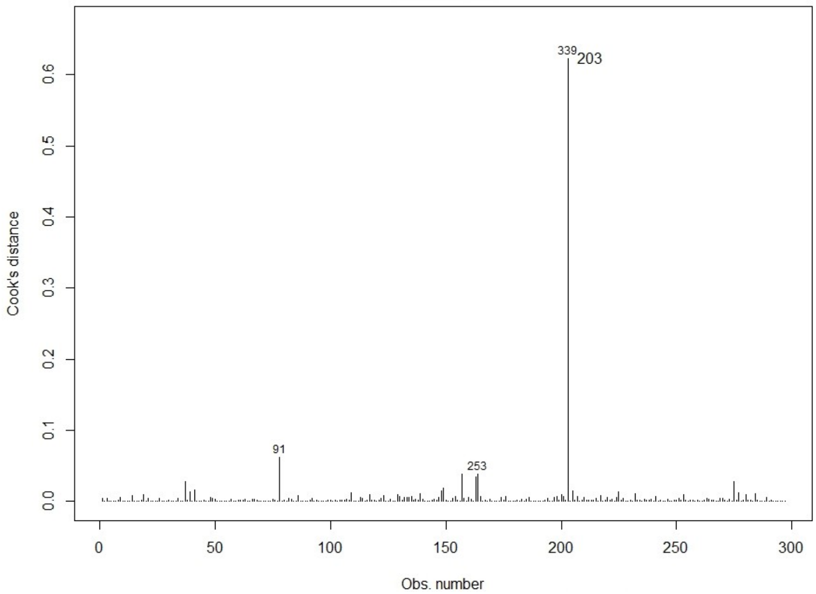

Figure 4.29: Cook's distance vs observation number of the individuals for model 1.

Cook's distance is a measure of the effect in the regression model obtained by the exclusion of an observation and it is given by:

$$D_i = \frac{\epsilon_i^2}{pMSE}\frac{h_{ii}}{(1-h_{ii})^2}, i = 1, \ldots, n \tag{4.11}$$

where h_{ii} is the $i - th$ leverage value of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T}$ and p is the number of coefficients in the regression model. Through the plot it can be seen that observation number 203 (which corresponds to the individual with id: PRE341) had an high influence on the regression model having a higher Cook distance value ($D = 0.622$). In addition, this observation had also a leverage value much higher ($h_{ii} = 0.202$) than leverage values of the other individuals (Figure 4.30). In fact, the fitted value for this individual, who was 115 years-old and from cluster 3 ("FLAT" pattern) was 82.46 dB, being higher than the observed value (33.75 dB).

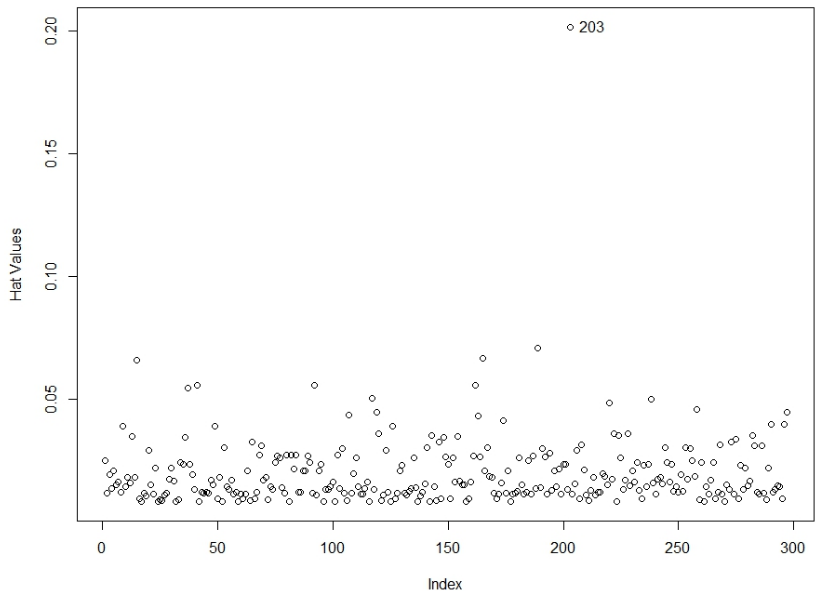

Figure 4.30: Leverage values vs observation number of the individuals for model 1.

Regarding these results it was decided to remove the individual PRE342 from the data set and refit the model in order to have an ideia of the impact of this observation in the prediction of the amount of HL (Table 4.39). Let name this new model as model 2.

Table 4.39: Adjusted multiple linear regression model for $PTA_{0.5,1,2,4kHz}$ response considering age$*$CM interaction excluding individual PRE342 (model 2).

Model	Variable	$\hat{\beta}$	s.e.	C.I.$_{95\%}$	t$_{obs}$	p-value	Regression characteristics
$PTA_{0.5,1,2,4kHz} \sim Age + CM + Age * CM$	Intercept	-63.39	14.34	[-91.60,-35.18]	4.42	$<$0.001	N=296
	Age	1.38	0.18	[1.01,1.74]	7.51	$<$0.001	Adjusted R-squared=0.29
	Cluster=1	35.98	20.68	[-4.73,76.68]	1.74	0.083	AIC=2403.17
	Cluster=2	73.48	19.02	[36.04,110.91]	3.86	$<$0.001	F$_{(5,290)} = 25.49$
	Age:(Cluster=1)	-0.60	0.27	[-1.12,-0.07]	-2.25	0.025	p-value$<$0.001
	Age:(Cluster=2)	-0.93	0.25	[-1.40,-0.45]	-3.82	$<$0.001	

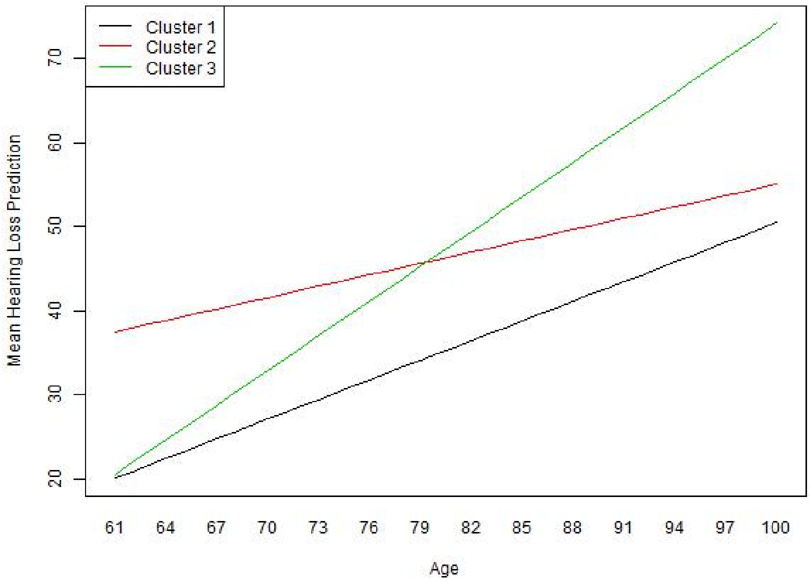

Figure 4.31: Interaction plot of variables included in the model with age*CM interaction excluding the individual PRE342 (Cluster 1: "HFGS" pattern, Cluster 2: "HFSS" pattern and Cluster 3: "FLAT" pattern).

Based on model 2, all the predictors were important in the determination of the amount of HL. As verified in the model 1, the $PTA_{0.5,1,2,4kHz}$ is expected to increase with aging for all clusters but at a different rate. For individuals in cluster 1 ("HFGS" pattern) and cluster 2 ("HFSS" pattern), the effect of age on $PTA_{0.5,1,2,4kHz}$ remained the same. However, for individuals in cluster 3 ("FLAT" pattern) the effect of age changed:

$$PTA_{0.5,1,2,4kHz_i} = -63.39 + 1.38Age_i + \varepsilon_i, i = 1, \ldots, n \tag{4.12}$$

In order to obtain a measure of the impact of the individual PRE342, it was calculated the absolute variation between the estimated $PTA_{0.5,1,2,4kHz}$ values of model 1 and model 2 (Figure 4.32b)). As observed, the maximum variation that can be obtained for $PTA_{0.5,1,2,4kHz}$ using one model or other is 19%, which corresponds to an individual with 60 years-old. On the other hand, the variation will be 0.1% for an individual with 76 years-old.

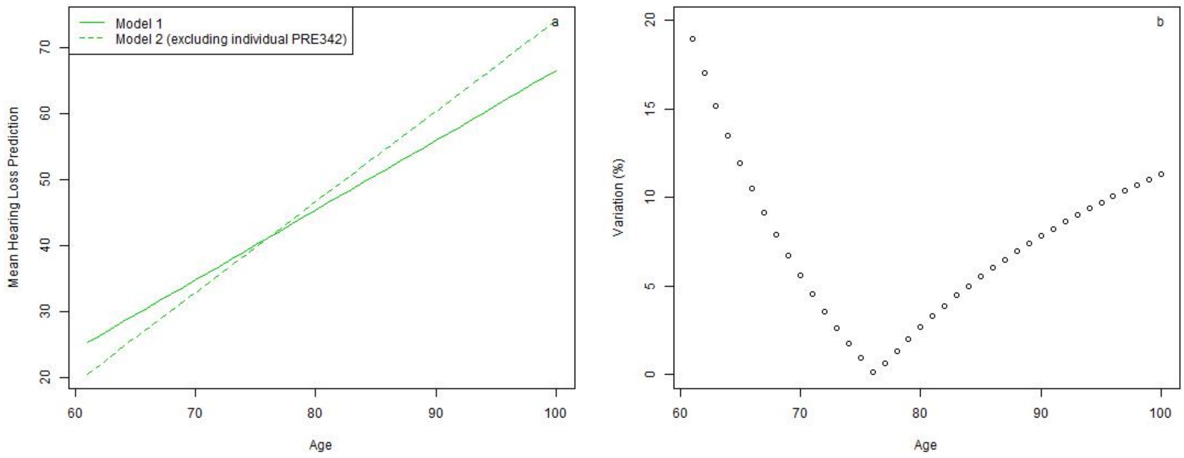

Figure 4.32: Comparison of adjusted models: a) Predicted amount of hearing loss adjusting model 1 and model 2 and b) Absolute variation (%) from model 1 to model 2.

After the exclusion of individual PRE342, the residual analysis was repeated for model 2.

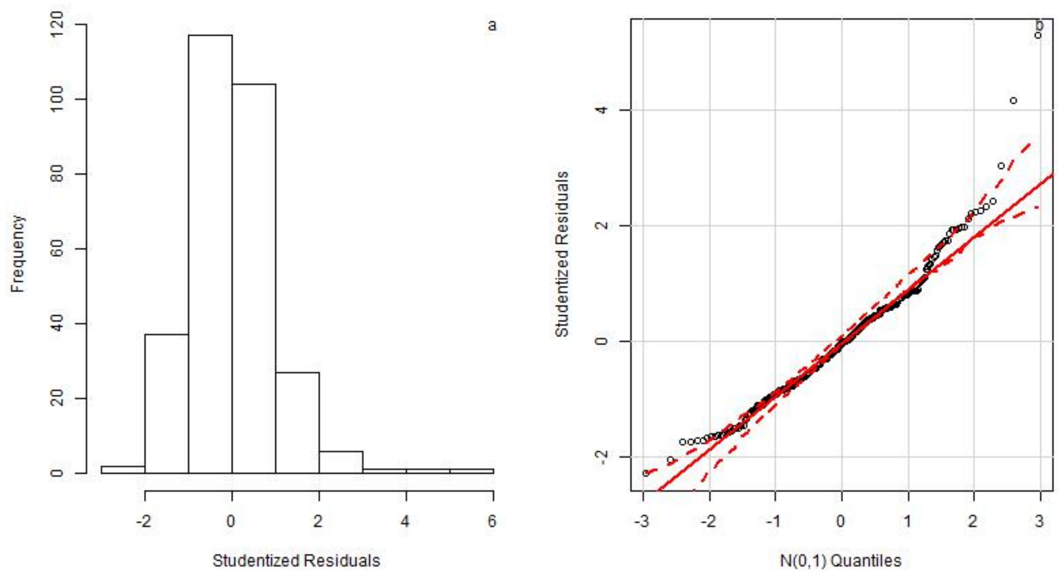

Figure 4.33: Distribution of the residuals of model 2: a) histogram and b) QQ-plot.

The line of the QQ-plot presented in Figure 4.33b) was slightly straighter than the previous one, indicating an apparent normality. Regarding the histogram of the residuals (Figure 4.33a)), however the right extension still existed, with the exclusion of the observation, the extension in the left side disappeared.

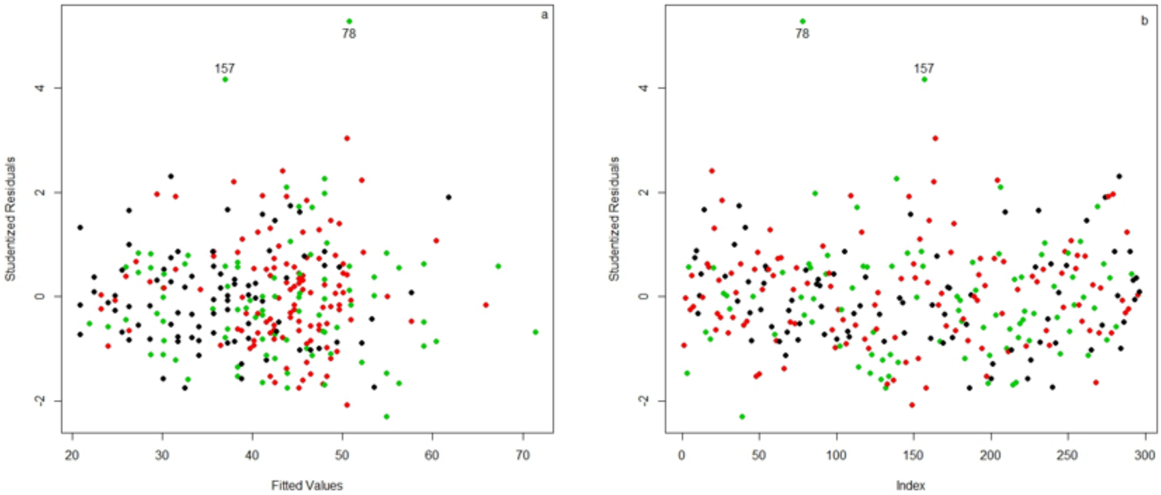

Figure 4.34: a) Studentized residuals vs fitted values and b) Studentized residuals vs observation number of the individuals for model 2.

As shown in Figure 4.34a), the homocedasticity and linearity assumption was not violated, since most of the residuals were randomly distributed. The same conclusion could be done to independence assumption (Figure 4.34b)). Furthermore, the observations 78 and 157 continued to be potential outliers. Concerning the measures of influence, the plots of Cook's distance and leverage values on Figure 4.35 showed that there were no extreme values for both, D_i and h_{ii}, differently from what happened in model 1, where the values from observation 203 stood out.

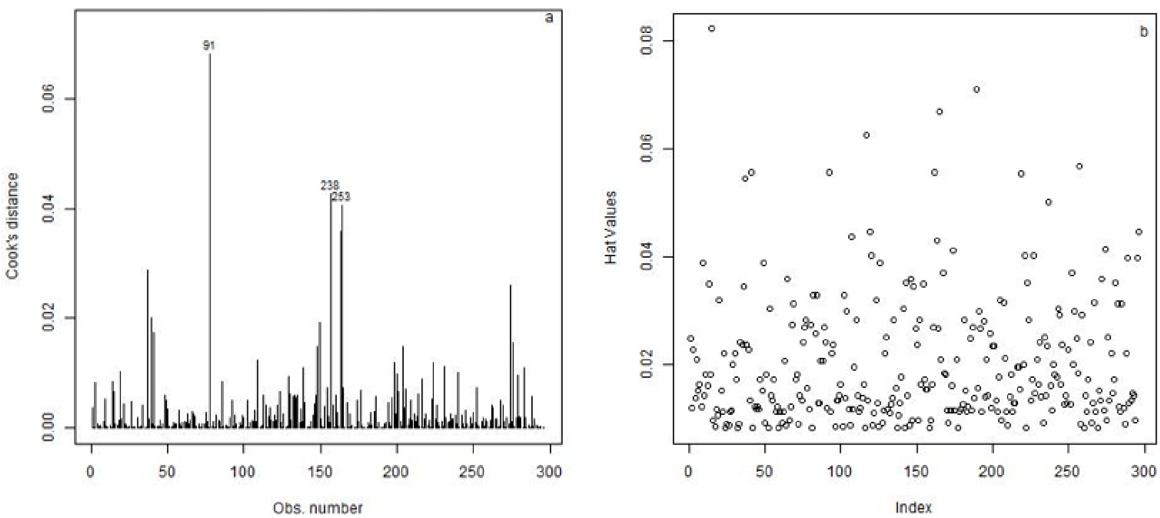

Figure 4.35: a) Cook's distance vs observation number of the individuals for model 2 b) Leverage values vs observation number of the individuals for model 2.

5

Discussion

HL is one of the most common chronic conditions being the most prevalent sensory deficit worldwide. Presbycusis or ARHL is the main cause of HL in the elderly as a result of the degenerative processes of aging, affecting between 30% to 40% of people older than 65 years old. The present project aimed to identify patterns of presbycusis, regarding the number and type, based on audiometric data collected from a sample of 321 Portuguese elderly during the period between 2007 and 2016.

In this study, it was observed a mean and median age when audiological assessment was performed between 75 and 80 years old (77.91 and 78.00 years old, respectively), being most of the individuals women (72%). In contrast to previous studies [17, 24], hearing thresholds were slightly worse for REs than LEs. This pattern held for men and women. For both ears, at high frequencies, men presented higher hearing thresholds than women, whereas at low frequencies the hearing thresholds are higher for women, corroborating findings from [17, 26, 36, 96, 97]. As concluded in [24, 71], at each frequency tested, hearing thresholds were significantly higher in older subjects.

The prevalence of HL ($PTA_{0.5,1,2,4kHz}$ $\geq$25dB HL) was 79.1%, a higher value in comparison with the ones presented in Table 2.2. However, the comparison with other epidemiological studies must be done carefully due to the differences in study population, as well as in HL definition used. In accordance with [17, 19, 24, 25, 44, 56, 57, 98, 99], in the present study, HL prevalence increased significantly with age, being 56.7%, 74.3% and 93.1% for individuals in 60-70, 75-80 and $\geq$85 years old groups, respectively. Although HL was more prevalent in males (82.2%) than females (77.9%), gender was not found to be significantly associated with the occurrence of HL corroborating [20, 50, 56], but in contrast with most studies, where male gender was strongly associated to HL [17–19, 25, 44, 57, 99]. Regarding noise exposure, familial history, hypertension, cholesterol, tinnitus, ototoxic medication, NAT2 phenotype and GRM7 genotype no significant associations were found either in independence tests or in bivariate logistic regression, when controlled by age. Although some of these findings are in accordance with [18, 26] (noise exposure), [18, 19, 25, 65] (hypertension) and [65] (family history), in several studies noise exposure [17, 19, 25, 32] and hypertension [56] were significantly associated with HL.

Prevalence of Audiogram Shape

- Classification proposed by Wuyts [3]

  In this sample the most prevalent audiogram configuration was the HFSS, comprising more than 50% of ears (RE: 51.2%, LE: 50.5%, BE: 51.2% and WE: 50.5%), followed by HFGS (RE: 29.6%, LE: 32.7%, BE: 28.6% and WE: 33.7%) and FLAT configurations (RE: 14.5%, LE: 11.4%, BE: 16.5% and WE: 9.4%). Otherwise, MFU, MFRU and LFA configurations were very rare, accounting for less than 0.3%. The prevalence of each configuration reported by [79] was different, being 37%, 35% and 27% for FLAT, HFGS and HFSS configurations for LE, respectively. Although the audiogram configuration classification used in our study was the same, this difference might be explained by the screened sample of the subjects in [79], which included younger individuals with age between 55 and 65 years old and with any type of HL. Nevertheless, the result is in accordance with the results of [100] study, where 46.2% of LEs and 35.2% of REs were HFSS, 29.2% of LEs and 30.9% of REs were HFGS, while a FLAT configuration was found in 12.0% of LEs and 18.8% of REs. Also, [101] reported that the most common configuration was the HFSS (48.5%), followed by the HFGS (26.9%) and FLAT (24.5%) configurations. Among males and females, HFSS configuration was the most prevalent (Male: 69.1% and Female: 47.5%), followed by HFGS configuration (Male: 24.7% and Female: 33.7%), whereas FLAT configuration was the less common (Male: 6.2% and Female: 18.8%). Although with slightly different percentages, similar findings were found in [79] study for male gender, where HFSS configuration was represented in 41% of ears, followed by the HFGS (35%) and FLAT configuration 24%, as well in [100] study, where 51.2%, 30.3% and 7.3% of male ears had a HFSS, HFGS and a FLAT configuration, respectively. Regarding females, the findings are inconsistent. [79] reported that FLAT configuration is the most common (50%), followed by the HFGS configuration (36%) and the HFSS configuration (14%). Otherwise, [100], reported that 31.5%, 28.3% and 22.1% of female ears had a HFGS, FLAT and a HFSS configuration, respectively, whereas accordingly with [101] results, FLAT and HFSS configurations were almost equally distributed (36.3% and 34.5%, respectively), being the less common configuration the HFGS (29.2%). The higher number of females than males in our sample could explain the differences found in the prevalence of audiogram configurations among our study and the other ones. A significant association between audiogram configuration and gender and noise exposure was found, meaning that the distribution of individuals with a given configuration is different according to gender and noise exposure. In fact, the prevalence of FLAT configuration in females was significantly higher comparing to the one in males, whereas the prevalence of HFSS configuration in males was significantly higher than those in females. In addition, females with any type of HL and a FLAT audiogram configuration tended to have a higher amount of HL (all FLAT configurations in Severe/Profound HL level corresponded to females). [102] who associate FLAT configuration to a strial phenotype (Figure 5.1), suggested that heritability was the most important cause of strial presbycusis. Findings of [36] supported a genetic effect on the inheritance of strial presbycusis in woman, occurring stronger aggregation of hearing levels in woman than in men. In accordance with [103], [104] and [105], hormonal differences between males and females could be associated to the tendency of females in having a larger overall amount of hearing loss in case of a FLAT audiogram. [103] demonstrated that homozygous megalin mutant mice exhibit profound hearing loss at 3 months of age associated with features of presbycusis and a reduced number of microvilli in marginal cells of the *stria vascularis*. Megalin is an endocytic receptor for estrogen and is strongly expressed within the marginal cells of the *stria vascularis* of the cochlea. Additionally, [104] concluded that hormone replacement therapy may have a protective effect on hearing impairment in postmenopausal women, whereas [105] found a higher decline rate in postmenopausal period for women on the same age group.

Considering HFSS configuration, [102] associated to sensory presbycusis which the main cause is outer hair-cell loss. [58] demonstrated that aging itself does not cause sensory presbycusis (Figure 5.1), being strongly related with accumulated environmental exposure. In addition, [106] demonstrated that solvents may induce auditory damage, especially to the outer hair cells, whereas [107, 108] indicated that noise exposure damages Deoxyribonucleic Acid (DNA) in the outer hair-cell. In our sample, the proportion of noise exposed subjects with a HFSS configuration was higher than the proportion of non-exposed subjects (62.2% vs 48.1%). This corroborate [79] results where the prevalence of HFSS configuration increases with increasing noise or solvent exposure. Moreover, the fact that the proportion of a HFSS configuration was higher in males than females, was consisted with the fact that in our study the male population was significantly more exposed to the noise. These findings corroborated [79, 101]. In addition, the mean HFSS configuration audiogram of males presented a notch at 4000 Hz frequency which is typical in subjects exposed to occupational noise [109–112].

In our study, HFGS configuration was present in males and females at a same proportion, as well in exposed and non-exposed subjects. [102] related cochlear presbycusis (Figure 5.1) to a HFGS audiogram configuration, being the main cause the abnormal motion mechanics of the basilar membrane.

No significant associations were found between age, medical and genetic conditions and audiogram configuration. However, [41] concluded that tinnitus is more common in subjects with a HFSS audiogram than in subjects with a FLAT audiogram.

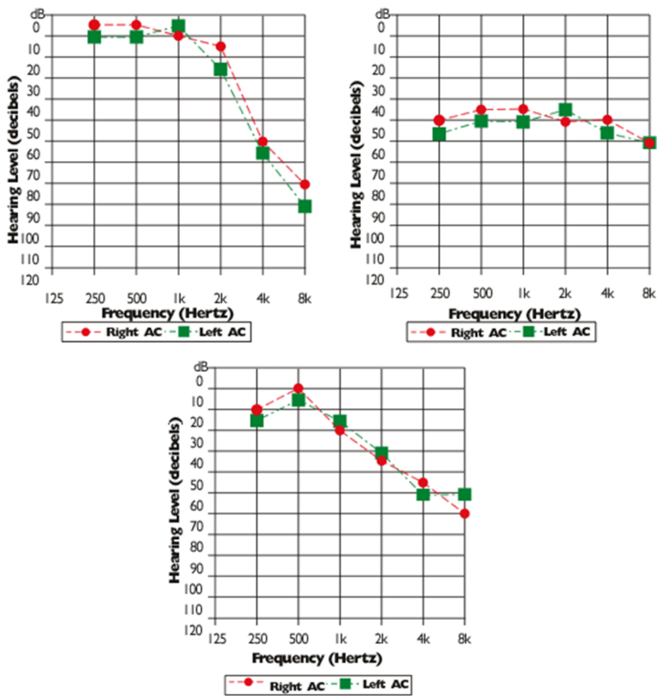

Figure 5.1: Audiogram patterns: sensory presbycusis (top left), metabolic or strial presbycusis (top right) and cochlear conductive presbycusis (bottom).

- Cluster Analysis

Employing hierarchical and K-means clustering, three main audiogram shapes were identified, which resemble to a FLAT, HFSS and HFGS audiogram configurations. Moreover, the results for both clustering techniques were much similar. In hierarchical clustering, 47.8% of the individuals were allocated to cluster 1, followed by cluster 2 (32.7%) and cluster 3 (19.5%). According to our results, the mean pattern of cluster 1, cluster 2 and cluster 3 groups resembled to a HFSS, HFGS and FLAT configurations, respectively. Only gender and noise exposure showed significant association with cluster membership. Most of males (67.1%) and females (40.1%) were presented in cluster 1 (“HFSS” pattern), followed by cluster 2 (“HFGS” pattern) (Male: 17.6% and Female: 38.7%) and cluster 3 (“FLAT” pattern) (Male: 15.3% and Female: 21.2%), being that the proportion of males in cluster 1 (“HFSS” pattern) was significantly higher than the proportion of females, as well the proportion of females in cluster 2 (“HFGS” pattern) was higher than the proportion of males. Regarding noise exposure, in our sample, male gender was more exposed to noise compared to female gender. About 56.4% of exposed individuals belonged to the cluster 1 (“HFSS” pattern), a higher proportion in comparison to non-exposed. In exposed subjects, gender had a significant influence: the proportion of males in cluster 1 (“HFSS” pattern) was high (74.4%), whereas the distribution in cluster 1 (“HFSS” pattern) and 2 (“HFGS” pattern) was much similar among females. In addition, no significant associations were found between cluster membership and age, medical and genetic characteristics. As in classification method proposed by Wuyts [3], a notch at 4000 Hz frequency was presented in the mean audiogram pattern of cluster 1 (“HFSS” pattern) of males.

In relation to K-means results, it was observed a more balanced distribution of subjects in each one of the clusters: 29.3% (cluster 1), 40.4% (cluster 2) and 30.3% (cluster 3), being that the mean pattern of cluster 3 was similar to a FLAT configuration, whereas cluster 1 and cluster 2 had some characteristics of HFGS and HFSS configurations, respectively. As in hierarchical clustering only gender and noise exposure were significantly associated with cluster membership. A higher proportion of males was found in cluster 2 (“HFSS” pattern), when comparing with the same proportion in female gender (58.8% vs 33.0%). Otherwise, the proportion of males was lower than proportion of females in cluster 3 (“FLAT” pattern) (17.6% vs 35.4%). 23.5% of males and 31.6% of females were allocated to cluster 1 (“HFGS” pattern). Regarding noise exposure, it was verified that the prevalence of exposed individuals in cluster 2 (“HFSS” pattern) was higher than non-exposed (48.5% vs 34.5%). Inversely, in cluster 3 (“FLAT” pattern) the prevalence of non-exposed was higher than exposed. In exposed subjects, most of males had a pattern of cluster 2 (“HFSS” pattern) (66.7%). Among exposed females, cluster 1 (“HFGS” pattern) and cluster 2 (“HFSS” pattern) had approximately equal proportions. In addition, the mean audiogram pattern of cluster 2 (“HFSS” pattern) of males contained also, a notch at 4000 Hz frequency which is typical in subjects exposed to noise. No significant associations were found between remaining characteristics and cluster membership.

Few studies were made using classification techniques to group similar audiograms shapes using audiological data. In [113] study, hierarchical clustering failed to reveal natural clusters in the data. However, results of [114] suggested a sloping sensory HL in a participant with self–reported noise exposure, tinnitus, and vertigo. [115, 116] determined presbycusis patterns using K-means clustering. Results of [115] are in accordance with the fact that men and women tended toward different audiometric configurations. Additionally, men displayed notched and sharply sloping configurations that were not dominant in data from women. [116] identified four audiogram shape subtypes of HL: flat shape, sloping shape, 2–4 kHz abrupt loss shape and 8 kHz dip shape, suggesting that GRM7 genotype TT occurs more frequently in patients with sloping shape and 2–4 kHz abrupt loss patterns.

Methodology Validation

- PCA

  In this study, the first two PCs explained a high proportion of variability, with PC1 accounting for 74.88% of variance and PC2 for 13.21%. The PC1 eigenvectors were all positive and very close to each other, indicating that the main source of variability among subjects was the overall degree of HL. On the other hand, the eigenvectors of PC2 were positive for frequencies at or below to 1000 Hz, and negative for higher frequencies. Thus, a negative score on PC2 would correspond to a sloping HL pattern, while a positive score would represent a flatter HL pattern. These findings are consistent with [117] results where it was performed a PCA on a set of 11462 RE audiograms with intensity hearing thresholds collected at the same frequencies used in our study. Although the number of components retained had been four, the first two PCs accounted for 72.9% of overall variability, being that the coefficients (eigenvectors) of the first one were all negative and approximately equal and the coefficients of the second one were negative to frequencies at or below to 1000 Hz, but positive for higher frequencies.

- LDA

  The results indicated that the classifier used to obtain automated classifications of audiogram pattern of HL exhibited a high degree of classification accuracy with the K-Means cluster groups when performed in the original data set. When using the scores of the PCs, the reliability of the classifier is lower.

Mulinomial Logistic Model

Additionally to gender, only noise exposure seemed to influence the CM. These results were in aggreement with what was obtained with the association tests. Thus, based on the model males and exposed to noise individuals were more likely to belong to cluster 2 (“HFSS” pattern) instead of cluster 1 (“HFGS” pattern) than females and non-exposed individuals, respectively. On the other hand, the risk of being in cluster 3 (“FLAT” pattern) was higher for females and non-exposed to noise individuals when compared to the risk of males and exposed subjects, respectively.

Linear Regression Model

The influence of age and CM on the prediction of the amount of HL ($PTA_{0.5,1,2,4kHz}$) was significant, contrary to the effect of gender. The amount of HL is expected to increase with aging for each audiogram pattern at a different rate: 0.78 for cluster 1 (“HFGS” pattern), 0.45 for cluster 2 (“HFSS” pattern) and 1.06 for cluster 3 (“FLAT” pattern). Moreover, there was a interaction effect among age and CM, that is individuals until 81 years old and a pattern characteristic of cluster 2 (“HFSS”) present a higher amount of HL than those in cluster 3 (“FLAT” pattern). After that, this tendency is reversed. Otherwise, individuals with an audiogram pattern of cluster 1 (“HFGS”) at all ages present lower amount of HL than the other two groups. Thus, independently of the gender, for the same audiogram pattern, it will be expected that the amount of HL increases with the aging. The audiogram pattern define the rate of this increase.

6

Final Remarks

In this study, audiogram patterns were identified in subjects with a clinical indication of presbycusis trough unsupervised classification methods, namely cluster analysis. It was also investigated the influence of age, gender and audiogram pattern (CM) on the prediction of the amount of HL ($PTA_{0.5,1,2,4kHz}$). The results suggested that: (1) there are three main presbycusis audiogram patterns that resemble to a FLAT, a HFSS and a HFGS configurations; (2) the variables gender and noise exposure are associated with the audiogram shape, whereas age is strongly associated to the prevalence and mean amount of HL; (3) the amount of HL increases with aging at a different rate depending on audiogram pattern.

Most of the results are consistent with the previously documented, however there are some discrepancies namely in the association tests for prevalence of HL and in the prevalence of audiogram configuration using the classification proposed by Wuyts [3]. As mentioned, only age was strongly associated with HL occurrence, whereas the prevalence of audiogram configurations was different from that determined in some studies, particularly for female gender. These results can be related with some limitations of the study possible derived from study sample: a) lack of information on environmental, medical and genetic variables at data collection time; b) the asymmetry of the distribution of cases for each one of the categories and c) the methods of information collection, which in some of the variables, such as environmental and medical variables, are particularly subjective.

Finally, it should be realized that presbycusis or ARHL is a public health concern that requires a special attention from the scientific community, health care providers and the general population. Therefore, in all research strategies, it is extremely important to have a deeper understanding of the disorder, in order to help in the development of preventive actions or in the determination of the best and most effective type of treatment.

Bibliography

[1] Pak Mitro. (2016) Anatomy of thorax and brain, ear canal diagram. Acessed: 14/06/2016. [Online]. Available: http://www.anatomytubuh.com/10427/ear-canal-diagram/

[2] H. Schuknecht, "Further observations on the pathology of presbycusis." *Arch Otolaryngol.*, vol. 80, no. 4, pp. 369–382, 1964.

[3] F. Wuyts, P. Van de Heyning, and F. Declau, "Audiometric criteria for linkage analysis in genetic hearing impairment." in *Developments in genetic hearing impairment.*, 1998, vol. 1, pp. 54–9.

[4] WHO. (2015) Deafness and hearing loss. Acessed: 21/03/2016. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[5] M. Mazzoli, G. Van Camp, V. Newton, N. Giarbini, F. Declau, and A. Parving. (2003) Recommendations for the description of genetic and audiological data for families with nonsyndromic hereditary hearing impairment. Acessed: 10/04/2016. [Online]. Available: http://hereditaryhearingloss.org

[6] A. Goodman, "Reference zero levels for pure-tone audiometers." *American Speech-Language-Hearing Association*, vol. 7, pp. 262 – 263, 1965.

[7] J. G. Clark, "Uses and abuses of hearing loss classification." *American Speech-Language-Hearing Association*, vol. 23, pp. 493 – 500, 1981.

[8] L. Loyd and H. Kaplan, *Audiometric interpretation: a manual of basic audiometry.* University Park Press: Baltimore, 1978.

[9] A. Martini and M. Mazzoli, "European Working Group on Genetics of Hearing Impairment." *International Journal of Pediatric Otorhinolaryngology*, vol. 49, no. 1, pp. 155 – 158, 1999.

[10] D. Pascolini and D. Smith, "Hearing impairment in 2008: A compilation of available epidemiological studies." *International Journal of Audiology*, vol. 48, no. 7, pp. 473 – 85, 2009.

[11] J. Northern and M. Downs, *Hearing in Children.*, 2nd ed. Williams Wilkins Co Baltimore, 1978.

[12] G. Stevens, S. Flaxman, E. Brunskill, M. Mascarenhas, C. Mathers, and M. Finucane, "Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries." *European Journal of Public Health*, pp. 1–7, 2011.

[13] H. Davis and S. Silverman, *Hearing and Deafness.*, 4th ed. New York: Holt, Rinehart & Winston, 1978.

[14] D. Sindhusake, P. Mitchell, W. Smith, M. Golding, P. Newall, D. Hartley, and G. Rubin, "Validation of self-reported hearing loss. The Blue Mountains Hearing Study." *International Journal of Epidemiology*, vol. 30, no. 6, pp. 1371–8, 2001.

[15] X. Liu and L. Xu, "Nonsyndromic hearing loss: an analysis of audiograms." *Annals of Otology, Rhinology & Laryngology*, vol. 103, no. 6, pp. 428–433, 1994.

[16] M. Paparella, D. Hanson, K. Rao, and R. Ulvestad, "Genetic sensorineural deafness in adults." *Annals of Otology, Rhinology & Laryngology*, vol. 84, no. 4, pp. 459–72, 1975.

[17] K. Cruickshanks, T. Wiley, T. Tweed, B. Klein, J. Mares-Perlman, and D. Nondahl, "Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin." *American Journal of Epidemiology*, vol. 148, no. 9, pp. 879–86, 1998.

[18] F. Lin, R. Thorpe, S. Gordon-Salant, and L. Ferrucci, "Hearing loss prevalence and risk factors among older adults in the United States." *J Gerontol A Biol Sci Med Sci.*, vol. 66A, no. 5, pp. 582–590, 2011.

[19] S. Nash, K. Cruickshanks, B. Klein, F. Nieto, G. Huang, J. Pankow, and T. Tweed, "The prevalence of hearing impairment and associated risk factors The Beaver Dam Offspring Study." *Arch. Otolaryngol. Head Neck Surg.*, vol. 137, no. 5, pp. 432–439, 2011.

[20] P. Dawes, K. Cruickshanks, D. Moore, M. Edmondson-Jones, A. McCormack, H. Fortnum, and K. Munro, "Cigarette smoking, passive smoking, alcohol consumption and hearing loss." *Journal of the Association for Research in Otolaryngology*, vol. 15, no. 4, pp. 663–674, 2014.

[21] G. Sprinzl and H. Riechelmann, "Current trends in treating hearing loss in elderly people: A review of the technology and treatment options – A mini-review." *Gerontology*, vol. 56, no. 3, pp. 351–8, 2010.

[22] B. Gopinath, V. Flood, C. McMahon, G. Burlutsky, J. Brand-Miller, and M. P., "Dietary glycemic load is a predictor of Age-Related Hearing Loss in older adults." *The Journal of Nutrition*, vol. 140, no. 12, pp. 2207–12, 2010.

[23] L. Raynor, J. Pankow, M. Miller, G.-H. Huang, D. Dalton, K. R., B. Klein, and K. Cruickshanks, "Familial aggregation of Age-Related Hearing Loss in an epidemiological study of older adults," *American Journal of Audiology*, vol. 18, no. 2, pp. 114–118, 2009.

[24] M. Janghorbani, A. Sheikhi, and S. Pourabdian, "The prevalence and correlates of hearing loss in drivers in Isfahan, Iran," *Archives of Iranian Medicine*, vol. 12, no. 2, pp. 128–134, 2009.

[25] B. Gopinath, E. Rochtchina, J. Wang, J. Schneider, S. Leeder, and P. Mitchell, "Prevalence of Age-Related Hearing Loss in older adults: Blue Mountains Study." *Archives of Internal Medicine*, vol. 169, no. 4, pp. 415–6, 2009.

[26] K. Bainbridge, H. Hoffman, and C. Cowie, “Risk factors for hearing impairment among U.S. adults with diabetes.” *Archives of Internal Medicine*, vol. 34, no. 7, pp. 1540–1545, 2009.

[27] N. Manwaring, M. Jones, J. Wang, E. Rochtchina, C. Howard, P. Newall, P. Mitchell, and C. Sue, “Mitochondrial DNA haplogroups and Age-Related Hearing Loss.” *Arch. Otolaryngol. Head Neck Surg.*, vol. 133, no. 9, pp. 929–33, 2007.

[28] O. Sogebi, O. Olusoga-Peters, and O. Oluwapelumi, “Clinical and audiometric features of presbycusis in Nigerians.” *African Health Science*, vol. 13, no. 4, pp. 886–92, 2013.

[29] D. Dalton, K. Cruickshanks, B. Klein, R. Klein, T. Wiley, and D. Nondahl, “The impact of hearing loss on quality of life in older adults.” *The Gerontologist*, vol. 43, no. 5, pp. 661–668, 2003.

[30] S. Pratt, L. Kuller, E. Talbott, K. McHugh-Pemu, A. Buhari, and X. Xu, “Prevalence of hearing loss in black and white elders: Results of the cardiovascular health study.” *J Speech Lang Hear Res.*, vol. 52, no. 4, pp. 973–989, 2009.

[31] F. Lin, J. Metter, R. O’Brien, S. Resnick, A. Zonderman, and L. Ferrucci, “Hearing loss and incident dementia.” *Arch Neurol.*, vol. 68, no. 2, pp. 214–220, 2011.

[32] Y. Agrawal, E. Platz, and J. Niparko, “Prevalence of hearing loss and differences by demographic characteristics among US adults.” *Archives of Internal Medicine*, vol. 168, no. 14, pp. 1522–1530, 2008.

[33] J. Fei, L. Lei, Z. Su-Ping, L. Ke-fang, Z. Qi-you, and Y. Shi-ming, “An investigation into hearing loss among patients of 50 years or older.” *Journal of Otology*, vol. 6, no. 1, pp. 44–49, 2011.

[34] K. Ohlemiller and R. Frisina, *Auditory Trauma, Protection and Repair*. Springer, 2008, ch. 6, pp. 145–194.

[35] K.-Y. Lee, “Pathophysiology of Age-Related Hearing Loss (peripheral and central).” *Korean J Audiol*, vol. 17, pp. 45–49, 2013.

[36] G. Gates, L. Gibbons, S. McCurry, P. Crane, M. Feeney, and E. Larson, “Executive dysfunction and presbycusis in older persons with and without memory loss and dementia.” *Cogn Behav Neurol.*, vol. 23, no. 4, pp. 218–223, 2010.

[37] K. Kiely, B. Gopinath, P. Mitchell, M. Luszcz, and K. Anstey, “Cognitive, health, and sociodemographic predictors of longitudinal decline in hearing acuity among older adults.” *J Gerontol A Biol Sci Med Sci.*, vol. 67, no. 9, pp. 997–1003, 2012.

[38] I. Maccà, M. L. Scapellato, M. Carrieri, S. Maso, A. Trevisan, and G. B. Bartolucci, “High-frequency hearing thresholds: effects of age, occupational ultrasound and noise exposure,” *International Archives of Occupational and Environmental Health*, vol. 88, no. 2, pp. 197–211, 2015.

[39] S. Arlinger, “Audiometric profile in presbycusis.” *Acta Otolaryngol (Stockh)*, vol. 476, pp. 85–90, 1991.

[40] B. Seimetz, A. Teixeira, L. Rosito, L. Flores, C. Pappen, and C. Dall'igna, "Pitch and loudness tinnitus in individuals with presbycusis." *International Archives of Otorhinolaryngology*, 2015.

[41] K. Demeester, A. van Wieringen, J.-J. Hendrickx, V. Topsakal, E. Fransen, L. Van Laer, D. De Ridder, G. Van Camp, and P. Van de Heyning, "Prevalence of tinnitus and audiometric shape." *Acta Otolaryngol (Stockh)*, vol. 7, pp. 37–49, 1991.

[42] N.-C. Chang, K.-Y. Ho, and W.-R. Kuo, "Audiometric patterns and prognosis in sudden sensorineural hearing loss in Southern Taiwan." *Arch. Otolaryngol. Head Neck Surg.*, vol. 133, no. 6, pp. 916–922, 2005.

[43] A. Simpson, K. Simpson, and J. Dubno, "Health-related quality of life in older adults: Effects of hearing loss and common chronic conditions." *Healthy Aging Research*, vol. 4, pp. 1–5, 2015.

[44] J. Schneider, B. Gopinath, M. Karpa, C. Mcmahon, E. Rochtchina, S. Leeder, and P. Mitchell, "Hearing loss impacts on the use of community and informal supports." *Age ageing*, vol. 39, no. 4, pp. 458–464, 2010.

[45] S. Kochkin and C. Rogin, "Quantifying the obvious: The impact of hearing aids on quality of life." *The Hearing Review*, 2000.

[46] T. Kim and J. Chung, "Evaluation of Age-Related Hearing Loss." *Age ageing*, vol. 17, no. 2, pp. 50–53, 2013.

[47] A. Bagai, P. Thavendiranathan, and A. Detsky, "Does this patient have hearing impairment?" *Journal of American Medical Association*, vol. 295, no. 4, pp. 416–428, 2006.

[48] B. Yueh, N. Shapiro, C. MacLean, and P. Shekelle, "Screening and management of adult hearing loss in primary care." *Journal of American Medical Association*, vol. 289, no. 15, pp. 1976–1985, 2003.

[49] W. Chien and F. Lin, "Prevalence of hearing aid use among older adults in the United States." *Archives of Internal Medicine*, vol. 172, no. 3, pp. 292–293, 2012.

[50] O. Abdel-Hamid, O. Khatib, A. Aly, M. Morad, and S. Kamel, "Prevalence and patterns of hearing impairment in Egytp: a national household study," *La Revue de Santé de la Méditerranée orientale*, vol. 13, no. 5, pp. 1170–1180, 2007.

[51] H. Francis, N. Chee, J. Yeagle, A. Cheng, and J. Niparko, "Impact of cochlear implants on the functional health status of older adults." *Laryngoscope*, vol. 112, no. 8 Part 1, pp. 1482–1488, 2002.

[52] Y.-J. Shin, B. Fraysse, O. Deguine, O. Valè, M.-L. Laborde, D. Bouccara, O. Sterkers, and A. Uziel, "Benefits of cochlear implantation in elderly patients." *Arch. Otolaryngol. Head Neck Surg.*, vol. 122, no. 4, pp. 602–606, 2000.

[53] A. Boothroyd, "Adult aural rehabilitation: What is it and does it work?" *Trends in Amplification*, vol. 11, no. 2, pp. 63–71, 2007.

[54] R. Sweetow and C. Palmer, "Efficacy of individual auditory training in adults: A systematic review of the evidence." *J Am Acad Audiol*, vol. 16, no. 7, pp. 494–504, 2005.

[55] T. Yamasobaa, F. Linb, S. Someyae, A. Kashioa, T. Sakamotoa, and K. Kondo, "Current concepts in Age-Related Hearing Loss: Epidemiology and mechanistic pathways." *Hear Res.*, vol. 303, pp. 30–38, 2013.

[56] A. Lasisi, T. Abionab, and O. Gurejeb, "The prevalence and correlates of self-reported hearing impairment in the Ibadan study of ageing." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 104, no. 8, pp. 518–523, 2010.

[57] W. You-qin, Y. Chong-ling, X. Shi-wen, X. Xiao-hong, L. Fei, L. Yu-qing, and Y. Qing, "A report of WHO ear and hearing disorders survey in Guizhou Province." *Journal of Otology*, vol. 5, no. 2, pp. 61–67, 2010.

[58] G. Gates, N. Couropmitree, and R. Myers, "Genetic associations in age-related hearing thresholds." *Arch. Otolaryngol. Head Neck Surg.*, vol. 125, pp. 654–659, 1999.

[59] F. Lin, P. Maas, W. Chien, J. Carey, L. Ferrucci, and R. Thorpe, "Association of skin color, race/ethnicity, and hearing loss among adults in the USA." *Journal of the Association for Research in Otolaryngology*, vol. 13, no. 1, pp. 109–117, 2012.

[60] E. Eyken, G. Van Camp, E. Fransen, V. Topsakal, J. Hendrickx, K. Demeester, P. Van de Heyning, E. Mäki-Torkko, and S. Hannula, "Contribution of the N-acetyltransferase 2 polymorphism NAT2*6A to age-related hearing impairment." *J Med Genet*, vol. 44, pp. 570–578, 2007.

[61] N. Ates, M. Unal, L. Tamer, E. Derici, S. Karakas, B. Ercan, Y. Pata, Y. Akbas, Y. Vayisoglu, and H. Camdeviren, "Glutathione S-transferase gene polymorphisms in presbyacusis." *Otol Neurotol*, vol. 26, no. 3, pp. 392–397, 2005.

[62] P. Dawes, H. Platt, M. Horan, W. Ollier, K. Munro, N. Pendleton, and A. Payton, "No association between apolipoprotein e or N-Acetyltransferase 2 gene polymorphisms and Age-Related Hearing Loss." *Laryngoscope*, vol. 125, pp. E33–E38, 2015.

[63] R. Friedman, L. Van Laer, M. Huentelman, S. Sheth, E. Van Eyken, J. Corneveaux, W. Tembe, and R. Halperin, "GRM7 variants confer susceptibility to age-related hearing impairment." *Human Molecular Genetics*, vol. 18, no. 4, pp. 785–796, 2009.

[64] H. Mostafa, M. Saad, A. El-Attar, G. Ahmed, S. Berrettini, F. Forli, G. Siciliano, and M. Mancuso, "Mitochondrial DNA (mtDNA) haplotypes and dysfunctions in presbyacusis." *Acta Otorhinolaryngol Ital*, vol. 34, pp. 54–61, 2014.

[65] S. Park, K. Johnson, K. Shea-Miller, and A. Chicchi, "Hearing loss and cardiovascular disease risk factors in older adults." *The Journal of Nutrition, Health & Aging*, vol. 11, no. 6, pp. 515–518, 2007.

[66] C. Horikawa, S. Kodama, S. Tanaka, K. Fujihara, R. Hirasawa, Y. Yachi, H. Shimano, N. Yamada, K. Saito, and H. Sone, “Diabetes and risk of hearing impairment in adults: A meta-analysis.” *J Clin Endocrinol Metab*, vol. 98, no. 1, pp. 51–58, 2013.

[67] M. Çayönü, M. Çapraz, A. Acar, A. Altundağ, and M. Salihoğlu, “Hearing loss related with type 2 diabetes in an elderly population.” *Int Adv Otol*, vol. 10, no. 1, pp. 72–5, 2014.

[68] P. Mitchell, B. Gopinath, C. McMahon, E. Rochtchina, J. Wang, S. Boyages, and S. Leeder, “Relationship of type 2 diabetes to the prevalence, incidence and progression of Age-Related Hearing Loss.” *Diabetic Medicine*, vol. 26, pp. 483–488, 2009.

[69] K. Nomura, M. Nakao, and E. Yano, “Hearing loss associated with smoking and occupational noise exposure in a Japanese metal working company.” *Int Arch Occup Environ Health*, vol. 78, pp. 178–184, 2005.

[70] S. Ferrite and V. Santana, “Joint effects of smoking, noise exposure and age on hearing loss.” *Occupational Medicine*, vol. 55, pp. 48–53, 2005.

[71] R. Albera, M. Lacilla, L. Piumetto, and A. Canale, “Noise-induced hearing loss evolution: influence of age and exposure to noise.” *Eur Arch Otorhinolaryngol*, vol. 267, pp. 665–671, 2010.

[72] T. Morata, “Chemical exposure as a risk factor for hearing loss.” *Journal of Occupational and Environmental Medicine*, vol. 45, no. 7, pp. 676–681, 2003.

[73] T. Morata, D. Dunn, L. Kretschmer, G. Lemasters, and R. Keith, “Effects of occupational exposure to organic solvents and noise on hearing.” *Scand J Work Environ Health*, vol. 19, pp. 245–254, 1993.

[74] S.-J. Chang, C.-J. Chen, C.-H. Lien, and F.-F. Sung, “Hearing loss in workers exposed to toluene and noise.” *Environmental Health Perspectives*, vol. 114, no. 8, pp. 1283–1286, 2006.

[75] Y. Chen, W.-G. Huang, D.-J. Zha, J.-H. Qiu, J. Wang, S.-H. Sha, and J. Schacht, “Aspirin attenuates gentamicin ototoxicity: From the laboratory to the clinic.” *Hearing Research*, vol. 226, pp. 178–182, 2007.

[76] E. Selimoglu, “Aminoglycoside-induced ototoxicity.” *Current Pharmaceutical Design*, vol. 13, pp. 119–126, 2007.

[77] S. Curhan, R. Eavey, J. Shargorodsky, and G. Curhan, “Prospective study of alcohol use and hearing loss in men.” *Ear Hear.*, vol. 32, no. 1, pp. 46–52, 2011.

[78] L. Brant, S. Gordon-Salantt, J. Pearson, L. Klein, C. Morrell, E. Metter, and J. Fozard, “Risk factors related to age-associated hearing loss in the speech frequencies.” *J Am Acad Audiol*, vol. 7, pp. 152–160, 1996.

[79] K. Demeester, A. van Wieringen, J. Hendrickx, V. Topsakal, L. Fransen, E. van Laer, G. Camp, and G. Heyning, “Audiometric shape and presbycusis.” *International Journal of Audiology*, vol. 48, pp. 222–232, 2009.

[80] A. Parving and V. Newton, “Guidelines for description of inherited hearing loss.” *J. Audiological Med*, vol. 4, pp. ii–iii, 1995.

[81] J. Han, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[82] G. W. Milligan and M. C. Cooper, “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[83] A. Weingessel, E. Dimitriadou, and S. Dolnicar, “An examination of indexes for determining the number of clusters in binary data sets,” SFB “Adaptive Information Systems and Modeling in Economics and Management Science”, Tech. Rep. Working Paper 29, 1999.

[84] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An R package for determining the relevant number of clusters in a data set,” *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014.

[85] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, 4th ed. Wiley Publishing, 2009.

[86] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, 1967, pp. 281–297.

[87] B. S. Everitt, *An R and S-PLUS companion to multivariate analysis*. Springer, 2005.

[88] H. F. Kaiser, “The application of electronic computers to factor analysis,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 141–151, 1960.

[89] A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[90] I. T. Jolliffe, “Discarding variables in a Principal Component Analysis,” *Applied Statistics*, vol. 21, pp. 160–173, 1972.

[91] R. B. Cattell, “The scree test for the number of factors,” *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[92] R. A. Johnson and D. W. Wichern, Eds., *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[93] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[94] G. Rodríguez. (2007) Lecture notes on generalized linear models. [Online]. Available: http://data.princeton.edu/wws509/notes/

[95] S. A. Czepiel, "Maximum likelihood estimation of logistic regression models: Theory and implementation."

[96] J. Jerger, R. Chmiel, B. Stach, and M. Spretnjak, "Gender affects audiometric shape in presbyacusis," *J Am Acad Audiol*, vol. 4, pp. 42–49, 1993.

[97] E. Bedin, A. Franz, C. Zadro, M. G. Persico, M. Ciullo, U. Hladnik, D. Dolcetta, D. Grasso, P. Riccardi, and T. Nutile, "Age-Related Hearing Loss in four Italian genetic isolates: An epidemiological study," *International Journal of Audiology*, vol. 48, pp. 531–538, 2009.

[98] N. Al-Ruwali and A. Hagr, "Prevalence of presbycusis in the elderly Saudi Arabian population," *Journal of Taibah University Medical Sciences*, vol. 5, no. 1, pp. 21–26, 2010.

[99] D. Wilson, P. Walsh, L. Sanchez, A. Davis, A. Taylor, G. Tuckere, and I. Meaghera, "The epidemiology of hearing impairment in an Australian adult population," *International Journal of Epidemiology*, vol. 28, pp. 247–252, 1999.

[100] S. Hannula, "Hearing among older adults-an epidemiological study," Master's thesis, Faculty of Medicine of the University of Oulu, 2011.

[101] K. Kaya, A. Koç, I. Sayin, S. Gneş, S. Canpolat, B. Şimşek, and F. Kayhan, "Etiological classification of presbycusis in Turkish population according to audiogram configuration," *Kulak Burun Bogaz Ihtis Derg*, vol. 25, no. 1, pp. 1–8, 2015.

[102] H. Schuknecht and M. Gacek, "Cochlear pathology in presbycusis," *The Annals of otology, rhinology, and laryngology*, vol. 102, no. 1 Pt 2, pp. 1–16, 1993.

[103] M. Froehlicher, A. Liedtke, K. Groh, H. López-Schier, S. Neuhauss, H. Segner, and R. Eggen, "Estrogen receptor subtype $\beta 2$ is involved in neuromast development in zebrafish (danio rerio) larvae," *Developmental Biology*, vol. 330, no. 1, pp. 32–43, 2009.

[104] C. Hederstierna, M. Hultcrantz, A. Collins, and U. Rosenhall, "Hearing in women at menopause. prevalence of hearing loss, audiometric configuration and relation to hormone replacement therapy," *Acta Oto-Laryngologica*, vol. 127, no. 2, pp. 149–155, 2007.

[105] J. Svedbrant, R. Bark, M. Hultcrantz, and C. Hederstierna, "Hearing decline in menopausal women â a 10-year follow-up," *Acta Oto-Laryngologica*, vol. 135, no. 8, pp. 807–813, 2015.

[106] A. Fuente and B. McPherson, "Organic solvents and hearing loss: The challenge for audiology," *International Journal of Audiology*, vol. 45, no. 7, pp. 367–381, 2006.

[107] S. A. Moussavi-Najarkola, A. Khavanin, R. Mirzaei, M. Salehnia, A. Muhammadnejad, and M. Akbari, "Noise-induced outer hair cell's dysfunction and cochlear damage in rabbits," *Iranian Red Crescent Medical Journal*, vol. 14, no. 10, pp. 647–656, 2012.

[108] H. Wei-ju, S. Xiao-rui, and A. Nuttall, “Noise exposure induced cochlear hair cell death pathways in guinea pig,” *Journal of Otology*, vol. 5, no. 1, pp. 51–56, 2010.

[109] D. McBride and S. Williams, “Audiometric notch as a sign of noise induced hearing loss,” *Occup Environ Med*, vol. 58, pp. 46–51, 2001.

[110] D. Nelson, R. Nelson, M. Concha-Barrientos, and M. Fingerhut, “The global burden of occupational noiseinduced hearing loss,” *American Journal of Industrial Medicine*, vol. 48, no. 6, pp. 446–458, 2005.

[111] A. Job, Y. Pons, L. Lamalle, A. Jaillard, K. Buck, C. Segebarth, and C. Delon-Martin, “Abnormal cortical sensorimotor activity during “target” sound detection in subjects with acute acoustic trauma sequelae: an fMRI study,” *Brain and Behavior*, vol. 2, no. 2, pp. 187–199, 2012.

[112] L. Ristovska, Z. Jachova, and N. Atanasova, “Frequency of the audiometric notch following excessive noise exposure,” *Archives of Acoustics*, vol. 40, no. 2, pp. 213–221, 2015.

[113] P. Allena and D. Eddins, “Presbycusis phenotypes form a heterogeneous continuum when ordered by degree and configuration of hearing loss,” *Hear Res.*, vol. 264, no. 1-2, pp. 10–20, 2010.

[114] R. Baiduc, J. Lee, and S. Dhar, “Toward a test battery for differential categorization of Age-Related Hearing Loss,” *Journal of Indian Speech Language Hearing Association*, vol. 28, no. 2, pp. 25–28, 2014.

[115] L. Ciletti and G. Flamme, “Prevalence of hearing impairment by gender and audiometric configuration: Results from the national health and nutrition examination survey (1999–2004) and the keokuk county rural health study (1994–1998),” *J Am Acad Audiol*, vol. 19, pp. 672–685, 2008.

[116] H. Luo, T. Yang, X. Jin, X. Pang, J. Li, Y. Chai, L. Li, Y. Zhang, L. Zhang, and Z. Zhang, “Association of GRM7 variants with different phenotype patterns of Age-Related Hearing Impairment in an elderly male Han Chinese population,” *PLOS ONE*, vol. 8, no. 10, pp. e77 153–e77 153, 2013.

[117] M. Anwar and M. Oakes, “Data mining of audiology patient records: factors influencing the choice of hearing aid type,” *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. S6, 2011.