BASES DE DATOS EN R
ANÁLISIS GRÁFICO Y ESTADÍSTICO DE VALORES ATÍPICOS Y AUSENTES.

Jorge Cordero Sánchez

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA. FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

Trabajo Fin Máster en Ingeniería de Computadores

Septiembre 2013

Directora:

Victoria López
Colaboradora de dirección:

Beatriz González

Autorización de difusión

Jorge Cordero Sánchez

Septiembre 2013

El abajo firmante, matriculado en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "Bases de datos en R", realizado durante el curso académico 2012-2013 bajo la dirección de Victoria López del Departamento de Arquitectura de Computadores y con la colaboración de dirección de Beatriz González del Departamento de Estadística e Investigación Operativa, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Resumen

Este trabajo se centra en el tratamiento de valores atípicos y valores ausentes. Para ello se incluye una mejora a las caras de Chernoff, para poder localizar en un entorno multivariante valores atípicos de manera unívoca. Dicha mejora se desarrolla en una base de datos propia de 531 ejemplares de cabras de Guadarrama con un total de 21 variables, con resultados satisfactorios. En la misma base de datos se verifica la correlación existente entre el perímetro torácico y el peso en caprinos, consiguiendo una ecuación con resultados considerablemente buenos. También se consiguen hallar más correlaciones en las medidas morfológicas de los caprinos gracias a esta base de datos. Estas correlaciones son la anchura de la caña con el perimetro de la caña, la altura de la cruz con la altura de medio dorso y esta última con la altura de la grupa.

También se dispone de dos bases de datos de repostajes de carburante, una de diesel con 231 registros y otra de gasolina 95 con 109, sobre las que se descartan posibles correlaciones entre sus variables, así como se desarrolla un sistema experto el cual se testea haciendo uno de un comparador de estimadores, el cual ha sido también desarrollado en este trabajo y permite comparar métodos propios de estimación desarrollado por usuarios y compararlos entre sí.

Palabras clave

Lista de palabras clave: caras de Chernoff resaltando valores atípicos, localizar outliers entorno multivariante, correlación morfología cabras, sistema experto repostajes, comparador de estimadores.

Abstract

This work is centred in the treatment of outliers and missing values. For this proposal here is developed an improvement of Chernoff face's, which consists of finding outliers accurately and univocally in a multivariate environment. This improvement is developed in our on-house developed database with 531 specimens of coats of Guadarrama, with 21 variables for each and satisfactory results. With this database is verified the correlation between the bust measurement and the weight in coats, getting an equation that responds well to tests. Also other correlations with the morphological measures of the coats are discovered: the leg's width with the leg's perimeter, the cross's height and the back half height, and finally, the correlation between latter and rump height.

This work also shows the work on other two databases of refueling, the first with 109 records of 95 octano's petrol and the last one with 231 records of diesel. Any variable correlations in those databases are discarded, and then an expert system is developed, which is tested with an estimators comparator. This developed comparator allows to compare user designed estimation methods.

Keywords

List of keywords: localize outliers in Chernoff's faces, outliers in multivariate environment, correlations in coat's morphology, refueling's expert system, estimators comparator.

Índice general

Agradecimientos

Quisiera agradecer el apoyo prestado por Victoria López, y Beatriz González así como agradecer el compromiso y la gran capacidad de trabajo en equipo mostrada por mis compañeros Jorge Martínez Ladrón de Guevara y Óscar Sánchez, sin ellos habría sido más arduo llegar hasta aquí. Además sería egoísta ignorar el apoyo prestado día a día por el resto de mis compañeros, especialmente Ainhoa por perder infinidad de tardes y no perder la paciencia. Finalmente pero no menos importante, a mi madre.

A todos ellos y a los que dejo en el tintero, gracias.

Dedicatoria

Para ti, que has creído que merece la pena detenerse en este trabajo.

Capítulo 1

Introducción

La ausencia de datos, así como la presencia de valores atípicos puede suponer en la práctica un problema cuando se trabaja con bases de datos, en especial en el campo de la bioinformática, donde incluso la pérdida de un dato puede suponer la no detección de enfermedades presentes. En este proyecto se va a trabajar en esta línea con el objetivo de ayudar a localizar y visualizar valores atípicos, así como con la estimación de datos perdidos.

Actualmente las Caras de Chernoff representan una de las pocas alternativas disponibles para poder mostrar gráfcamente un perfil multivariante. Este gráfico asigna a las variables de una base de datos (hasta un máximo de dieciocho) rasgos físicos de la cara y dibuja una cara por cada individuo de la tabla. Así por ejemplo, en psicología se pueden detectar individuos con rasgos depresivos. Los rasgos faciales son fácilmente interpretables por el ojo humano, pero cuando la base de datos tiene muchos individuos, la labor de diferenciación es tediosa para el investigador y se complica cuando el número de variables es también grande.

En este trabajo se elabora un gráfico basado en las caras de Chernoff que utiliza una paleta de colores para resaltar sobre el gráfico original aquellos rasgos que representan un valor atípico en una o varias variables de uno o varios individuos de una base de datos.

Para detectar estos rasgos atípicos u outliers se utiliza en primer lugar un diagrama de caja o Box Plot. Este gráfico está basado en cuartiles y permite visualizar de una forma sencilla los individuos de una única variable. Está compuesto por un rectángulo central, la çaja", y dos brazos, los "bigotes". Es un gráfico que suministra información sobre los valores

mínimo y máximo, los cuartiles del 25 %, 50 % y 75 %, la existencia de outliers y la simetría de la distribución.

Posteriormente se elabora un código que permite dibujar e interpretar sobre las caras de Chernoff dichos outliers diferenciándolos por colores. El procedimiento diseñado se aplica a una base de datos de cabras de la sierra de Guadarrama suministrada por personal de la Facultad de Veterinaria. También se hace un estudio de correlación entre las variables de dicha base de datos y se compara con estudios de regresión existentes para otro tipo de ganado, llegando a conclusiones similares de una buena estimación del peso de un ejemplar en función de su perímetro torácico. La base de datos de las cabras de la sierra de Guadarrama es una tabla típica del Análisis Multivariante. Otro aspecto importante del trabajo es el estudio de una base de datos de Series Temporales, concretamente de carburantes, proporcionada por personal de la Facultad de Matemáticas. Dichos datos se comparan con los extraídos de la página web del Ministerio de Industria, Energía y Turismo [INE]. Se estudia la relación euros/litros y kilómetros/litros y se desarrolla un sistema experto para estimar valores perdidos y un comparador de estimadores. Se estima también el consumo de litros cada cien kilómetros y se comprueba que coincide con las especificaciones del fabricante.

En este trabajo toda la programación está realizada con el lenguaje de programación R. Es necesario hacer una breve introducción a los dos aspectos principales sobre los que gira el trabajo, que son el lenguaje de programación R y el tratamiento que hace de las bases de datos. Cabe destacar, que en esta memoria se presta especial atención a la lectura de datos con R.

La razón por la que se ha desarrollado este trabajo bajo el lenguaje R es que es el lenguaje más utilizado en el ámbito de la bioinformática y la bioestadística. Esto es debido a una razón muy sencilla, en un mundo en el que la investigación pública tiene dificultades para financiarse, R, además de ser un lenguaje de programación, es un software para el análisis gráfico y estadístico de datos bien documentado `www.r-progect.org`, con una amplia comunidad de usuarios y un gran abanico de librerías, 4644 paquetes el 24 de Junio

de 2013, bajo licencia GNU[GNULicense,]. Dicha licencia es de código libre lo que implica una reducción del coste en un proyecto de desarrollo por no tener que invertir parte de los fondos en la adquisición de software.

A continuación, antes de explicar con más detalle qué es R y cuáles son sus cualidades, se va a explicar qué es la bioinformática y algunas de las bases de datos públicas más relevantes de este sector.

1.1. Bioinformática

La bioinformática es la aplicación de técnicas y tecnologías informáticas a la biología.

En 1977 se realiza la primera secuenciación de ADN (ácido desoxiribonucleico). Fue la del Phi-X174 [Sanger et al., 1977], un bacteriófago con 5386 nucleótidos que codifican 11 proteínas. Actualmente ya se ha secuenciado el ADN de cientos de seres vivos, son cientos de cadenas de miles de datos, la manipulación de esta información sin ayuda computacional sería una ardua tarea y en muchos casos frustrante, lo que un ordenador potente puede procesar en poco tiempo, del orden de segundos o minutos, un equipo de trabajo podría necesitar del órden de semanas para desarrollar ese mismo trabajo. Por lo que la bioinformática se destapa como un campo fundamental en el ámbito de la biología. El término bioinformática lo acuña Paulien Hogeweg en 1970 para referirse a esta relevancia que comienza a tomar la informática en el ámbito de la biología siendo una parte fundamental en el tratamiento, almacenamiento y transmisión de datos. La importancia de la informática en la biología ha ido en aumento desde los años 70, por dos razones principales. La primera es que a día de hoy cada vez se analizan mayor número de secuencias, y y la segunda y más importante es que no solo se analizan esas secuencias, sino que se almacenan las secuencias y los resultados, compartiéndolos en la red lo que agiliza las investigaciones, ya que evita tener que realizar el mismo trabajo repetidas veces suponiendo un ahorro en tiempo y dinero considerablemente alto.

El beneficio que extrae la biología de la informática se puede apreciar observando el

rendimiento de la secuenciación con tecnologías de alto rendimiento, uno de los analizadores de ADN más utilizados durante la corta historia de la bioinformática es el ABI 3730XL cuya capacidad diaria es de 1Mb por día. Menos de 10 años después nacía el Applied Biosystems SOLiD que es capaz de analizar 3000Mb por cada ejecución, lo que supone una multiplicación por 1000 en la capacidad de secuenciación. Como se puede apreciar es similar el espíritu que tienen dichas investigaciones con el lenguaje de programación R, ya que ambas basándose en la filosofía de compartir datos pretenden conseguir la máxima mejora en el menor tiempo posible. Dicha información se almacena en bases de datos públicas.

1.2. Aportaciones de este trabajo

Como se explica con más detalle en el Punto 2.1, el sector de las bases de datos públicas en el ámbito de la bioinformática es muy completo y en él hay implicados un gran número de laboratorios con un personal cualificado, por lo que no tiene sentido querer realizar otra base de datos bioinformática. Por ello este trabajo se centra en el análisis exploratorio de datos. El análisis exploratorio de datos es el análisis necesario que hay que realizar sobre los datos recopilados.

Para ello se han desarrollado varias ideas, la primera de ellas es la localización de los valores atípicos en entornos multivariante, permitiendo que sea muy sencilla y que se realice de una manera totalmente visual mediante el uso de un código de colores aplicado a las caras de Chernoff (las cuales permiten representar gráficamente entornos multivariante), permitiendo así que no sea necesaria una persona especialista en ningún sector, ni en el de los datos que se estén tratando ni en el sector estadístico; la segunda de ellas, se centra en la búsqueda de relaciones entre variables en una base de datos y se han hallado los coeficientes de correlación en el caso de las mejores correlaciones. Entre dichas correlaciones se ha verificado una hipótesis de que el peso y el perímetro torácico de un caprino están estrechamente relacionados; también se ha realizado una estimación de datos mediante el uso de un sistema experto, aplicado al sector de los carburantes en base al gasto de dos

vehículo tanto de motor diesel como de motor gasolina. En dicho apartado se han comparado conocimientos expertos y determinado cuales de ellos obtienen los mejores resultados; por los resultados obtenidos en el apartado anterior se ha creado un comparador de estimaciones, ya que se considera que los sistemas expertos mejoran en tiempo y resultados muchos otros métodos de estimación. Por ello se ha permitido la comparación de un sistema del propio usuario con diversos métodos de estimación conocidos para facilitar al usuario el trabajo a la hora de desarrollar un sistema experto, también permite al usuario comparar dos sistemas expertos y diversas funcionalidades más, explicadas en el Punto 3.4

Recopilación de datos

Algunas veces a la hora de recopilar datos pueden tener cierta pérdida, en función de su origen la causa puede ser de una naturaleza u otra.

Actualmente la mayoría de los datos se informatizan, pero aún así pueden diferir en el momento en que esto sucede. El momento se entiende como el tiempo que pasa desde que el dato se crea hasta que se informatiza, algunos ya se crean informatizados directamente, y es importante, ya que cuanto menos se trabajen los datos antes de informatizarlos mejor, porque se ayudará a evitar posibles modificaciones erróneas por manipulación humana del dato.

A continuación se explican las diversas naturalezas posibles de los datos y las vulnerabilidades de cada una.

Datos procedentes de laboratorio

En el caso de que los datos procedan de un laboratorio se tiene que la pérdida de datos puede deberse a que algún utensilio del laboratorio, más comúnmente la lente, contenga suciedad, como puede ser una mota de polvo. Otra fuente plausible para la pérdida de datos es que la muestra no sea suficiente para la extracción de los datos. También puede suceder simplemente que al realizar el análisis no se consigan extraer los datos.

Datos procedentes de códigos de barras

Otro posible origen de los datos es mediante el uso de las pistolas para leer códigos de barras, una vez que se ha realizado el análisis y se ha clasificado se utiliza la pistola para leer la etiqueta del tubo, el código de barras, y ver así los datos de dicho análisis.

Los problemas en este aspecto suelen estar más relacionados con la configuración de la recogida de datos de la pistola que con la lectura del código de barras.

Datos procedentes de encuestas

También pueden provenir de encuestas como registros, cuestionarios o entrevistas. Dicho sistema de extracción de datos se suele utilizar en la búsqueda de relaciones entre los hábitos de las personas y posibles derivaciones en ciertas enfermedades.

En ese caso la fuente puede ser o bien una encuesta informatizada o una a papel, en el caso de ser informatizada el error disminuye, ya que el único error plausible es en el que el usuario aporta un dato erróneo.

Pero en el caso de que la encuesta sea a papel la probabilidad de error aumenta de manera considerable ya que los datos aportados por el usuario pueden ser ilegibles o simplemente que en el proceso de almacenamiento a la base de datos se cometa algún error de transcripción.

Volver a recopilar los datos en algunos casos puede ser muy costoso y en otros incluso puede ser inviable, ya que quizás las condiciones hayan cambiado y por lo tanto los resultados no serían congruentes. En estos casos se ofrece como una alternativa real la estimación de datos de la cual se habla en el Punto 2.2.

1.3. Estructura de la memoria

Esta memoria está estructurada de la siguiente forma, además de esta breve introducción al trabajo, en la que están incluidas las aportaciones en el ámbito de la investigación. En el Capítulo 2 se ve el contexto en el que se ha desarrollado el trabajo, con algunas de las bases de datos bioinformáticas más importantes, todas ellas públicas. A continuación

se mencionan los métodos de estimación estadística y estudios de análisis de datos más destacados y para finalizar el capítulo algunos métodos de eliminación de datos controlados. En el Capítulo 3 se verá el análisis exploratorio de datos realizado, que contiene la carga de archivos, observación de outliers en las caras de Chernoff, completar datos ausentes usando correlaciones y un sistema experto, métodos implementados para completar datos, así como la comparativa entre ellos mediante el uso de un comparador de estimadores. En el Capítulo 4 se hace una breve introducción a las bases de datos utilizadas en el proyecto y se realiza la carga de archivos sobre la misma. En el Capítulo 5 se muestran los resultados obtenidos en la localización de valores atípicos utilizando las caras de Chernoff, así como las correlaciones existentes entre las variables, mostrando finalmente los resultados obtenidos al aplicar dichas correlaciones. En el Capítulo 6 se muestra por qué no se pueden aplicar correlaciones en el caso de las bases de datos de carburantes, así como la aplicación de conocimiento experto y la consecuente creación de un sistema experto así como su evaluación, realizada con un comparador de estimaciones propio. Finalmente en el Capítulo 7 se muestran las conclusiones a las que se han llegado al realizar el trabajo, enfatizando los puntos más importantes y se habla de los trabajos futuros que se pueden realizar, continuando esta línea de investigación. Dichos trabajos futuros se dividen por secciones y dentro de cada sección por orden de relevancia.

Capítulo 2

Bases de datos en bioinformática

Lo primero que hay que destacar en este capítulo es que se pretende que el trabajo se pueda aplicar a diversas bases de datos, bioinformáticas claro, esto es que si cualquiera de las funciones desarrolladas fuera útil para alguna base de datos que se pueda aplicar sin necesidad de modificaciones.
Pese a que se ha trabajado con bases de datos propias, mostradas en el Punto 4.1, en bioinformática hay una gran variedad de bases de datos públicas, algunas de las cuales se describen a continuación.

Una vez vistas las bases de datos públicas que existen en la actualidad en el mundo de la bioinformática, no es difícil apreciar que, en el caso de este trabajo, no se puede aportar nada a dichas bases de datos. Por un lado no se dispone de datos suficientes, ya que son millones de datos almacenados frente a los cientos que existen en las bases de datos de las que se disponen en este trabajo, explicadas en el Punto 4.1, además tampoco se puede conseguir equiparar los recursos humanos ni económicos.

Por ello este trabajo pretende ayudar a las bases de datos bioinformáticas, incluyendo las anteriormente citadas, centrándose en la localización visual de outliers y la estimación de datos perdidos.

2.1. Bases de datos bioinformáticas públicas

En la biología se puede afirmar que un descubrimiento tiene como precio el dinero que de un modo u otro se ha invertido en conseguirlo. De ahí la importancia de que las bases de datos biológicas sean públicas ya que hacen que no sea necesario que distintos laboratorios realicen los mismos experimentos y las mismas pruebas para conseguir datos base y no solo eso, si no que los datos disponibles son mucho mayores de los que cualquier laboratorio podría tener por sí mismo, al estar compartidos entre muchos de ellos. Dichas bases de datos también permiten que un usuario, con escasez de medios, pueda emprender una investigación, al menos teórica, reduciendo los gastos iniciales e incluso permite un primer acercamiento a los datos sin necesidad de tener un equipo especializado a su disposición. En función de los datos almacenados y de la finalidad las bases de datos se pueden agrupar de diferentes formas, a continuación se muestran los principales tipos de bases de datos con sus principales bases de datos.

Bases de datos de secuencias nucleótidos

A continuación se muestran las principales bases de datos de secuencias de nucleótidos las cuales colaboran entre sí para mejorar los datos ofrecidos a los usuarios.

UKBiobank

UKBiobank[UKBiobank,] es una base de datos biológica que almacena datos de más de 500.000 británicos de entre 40 y 69 años, dichos datos fueron recopilados entre 2006 y 2010, los datos incluyen pruebas tanto de saliva, como de orina y sangre para analizarlas. La información de la base de datos incluye información de las personas sobre las que se han realizado las pruebas y consentimiento para que se realice un seguimiento de su estado de salud. La finalidad de UKBiobank es poder prevenir, diagnosticar y tratar enfermedades utilizando como casos base los datos almacenados.

European Bioinformatics Institute (EBI)

El EBI[EBI,] es una organización sin ánimo de lucro y parte del European Molecular

Biology Laboratory (EMBL), tiene bases de datos de acceso gratuito a disposición de los científicos, las cuales abarcan todos los campos de la biología molecular y permiten realizar consultas sobre dichos datos. Pero su tarea no se queda aquí sino que además se aseguran de coordinar los datos biológicos en toda Europa. También pretenden facilitar el acceso a las tecnologías puntas en el ámbito bioinformático así como ayuda a científicos para progresar en sus estudios.

National Center for Biotecnology Information (NCBI)

El NCBI[NCBI,] es parte del United States National Library of Medicine (NLM) que a su vez es una división del National Institutes of Health. Hospeda, entre otras, la base de datos GenBank desde 1992. Esta base de datos contiene información de secuencias de ADN y permite recolectar una gran cantidad de datos en el ámbito de la bioinformática, además permite a los usuarios compartir sus datos y seleccionar una fecha de compartición de cara a que no aparezcan en la base de datos antes de que se haya realizado la publicación de la investigación. A parte de GenBank en el NCBI está almacenada PubMed[PubMed,]. Si bien es cierto que no contiene bases de datos de bioinformática eso no implica que no sea una base de datos sin interés ya que contiene más de 22 millones de referencias relacionadas con la literatura bioinformática de MEDLINE y en muchos casos permite el acceso a artículos completos.

DNA Data Bank of Japan (DDBJ)

El DDBJ[DDBJ,] contiene una de las más importantes bases de datos de secuencias nucleótidos y la principal de Asia. Está coordinado con las otras dos organizaciones de bases de datos de nucleótidos más importantes que son los anteriormente citados EMBL y GenBank. La unión de estas tres bases de datos se denomina International Nucleotide Sequence Database Collaboration(INSDC). El principal objetivo del DDBJ es mejorar la calidad del INSD.

Metabases de datos

Las metabases de datos, como puede intuirse por su nombre, son aquellas que almacenan información sobre bases de datos, las cuales no se utilizan para buscar datos como tal, si no para buscar una tipología de datos y poder encontrar aquellos datos que mejor se ajusten a los requerimientos del usuario.

MetaBase

MetaBase[MetaBase,] es una base de datos realimentada por los usuarios, en la que se pueden encontrar todas las bases de datos biológicas que existen actualmente. A día de hoy tiene más de 1801 entradas, lo que convierte a MetaBase en un claro referente ya que se pueden realizar búsquedas para que el usuario encuentre más fácil las bases de datos que esté buscando. Funciona de la misma manera que Wikipedia[1], donde los usuarios aportan el contenido, de hecho la apariencia de ambas es prácticamente idéntica.

Genómicas

Como su propio nombre indica las bases de datos genómicas almacenan secuencias de genomas, las bases de datos genómicas han sido muy importantes para convertir a la genómica en una de los campos más vanguardistas en el mundo biológico. Una de las más destacadas en este sector es Ensembl[ENSEMBL,], que es una base de datos que almacena los genomas de diversas especies desde 1999, conteniendo en la actualidad el genoma de 75 especies.

Proteínas

En 1838 Jöns Jacob Berzelius acuña el nombre de proteína para referirse a la sustancia rica en nitrógeno presente en las células tanto de animales como de vegetales. Son parte fundamental para la vida y considerablemente complejas, por ello, los datos almacenados de las proteínas difieren según la función de la base de datos que lo almacena. Algunas de ellas las enunciamos a continuación.

[1] www.wikipedia.org

Hay bases de datos que almacenan información de las secuencias de aminoácidos, almacenando el orden en el que se concatenan los aminoácidos para formar las proteínas, que representa la estructura primaria de una proteína.

Uniprot[Uniprot,] es una base de datos que almacena datos sobre las secuencias de proteínas y sus funciones. Colabora con INSDC, siendo esta colaboración la fuente de más de un 99 % de los datos contenidos en UniProt.

También hay bases de datos proteómicas, son aquellas que almacenan información sobre la estructura de las proteínas y su función. En esta especialización destacan Proteomics Identifications Database (PRIDE)[PRIDE,], pertenece al EBI y es el sector encargado del almacenamiento y manejo de la base de datos proteómicas.

Para finalizar, cabe destacar las bases de datos de interacciones proteína-proteína, que almacenan información relativa a las relaciones entre proteínas. Estas relaciones se dan entre dos o más proteínas y son importantes para la mayoría de funciones biológicas. Para un mayor detalle de las interacciones entre proteínas, se puede consultar el trabajo de Golemis y Adams[Golemis and Adams, 2002].

En este ámbito destaca BIND (Biomolecular interaction network database)[Gilbert, 2005], es muy útil ya que proporciona apoyo visual para las funciones de las proteínas y la localización ayudando a facilitar la investigación.

Bases de Datos de Ácido ribonucleico (ARN)

Otro tipo de bases de datos públicas que hay en bioinformática son las bases de datos que almacenan información del ARN. Desde que se descubrieron los ácidos nucleicos en el año 1868 gracias a los estudios realizados por Firedrich Miescher se ha descubierto que tienen un papel muy importante para los seres vivos.

Mientras el ADN contiene la información genética el ARN es mucho más versátil, regulando la expresión génica, ayudando en la duplicación del ADN, incluso teniendo actividad catalítica. También es el encargado de transmitir la información para sintetizar las proteínas, para lo cual se utilizan diversos ARN. El ARNm es el ARN mensajero, contiene información

genética idéntica de genes que vayan a expresarse y es utilizada por los ribosomas para unir los aminoácidos de manera correcta y formar una proteína.

Posteriormente el ARNt se elimina para frenar la síntesis de las proteínas y evitar excesos proteicos. El ARN de transferencia, el ARNt, también es el encargado del transporte de los aminoácidos que están en el citoplasma celular hasta los ribosomas para construir las proteínas. Para ello cada molécula de ARNt es la encargada del transporte de un aminoácido concreto.

Otro ARN es el ARN ribosómico, ARNr, el cual es parte de los ribosomas y participa activamente en la unión de los aminoácidos en el proceso de síntesis de proteínas.

Existen más tipos de ARN y para muchos de ellos hay una base de datos pública, también hay bases de datos que incluyen varios tipos como es la RFAM que es la abreviatura de Rna FAMilies del inglés, familias de ARN, [Gardner et al., 2011]. Los tipos que incluye son ncRNA (ARN no codificante) su nombre es debido a que no se traduce en una proteína; elementos cis-reguladores, que son las regiones, tanto de ADN como de ARN, encargadas de regular la expresión génica de las moléculas de ADN donde esté; y por último el *self-splicing* ARN, este es el que realiza el mismo su propio empalme, *splicing*. El empalme del ARN es el proceso por el cual se eliminan ciertos fragmentos secuenciales denominados intrones. En circunstancias normales, en el caso de no darse el *self-splicing*, son necesarias proteínas externas para realizar el empalme, pero en ciertos casos es viable que el propio ARN realice él mismo el empalme y es el proceso denominado *auto-splicing*.

Estructura de carbohidratos

Los cabrohidratos son biomoléculas cuya composición tiene moléculas de carbono, hidrógeno y oxígeno, son los encargados de almacenar y consumir la energía en los seres vivos. La estructura de los carbohidratos es tanto los átomos de Carbono (C), Hidrógeno (H) y Oxígeno (O) que los forman como la disposición e interrelación que tienen entre sí. Además es parte fundamental de los mismos ya que aún teniendo el mismo número de átomos con igual proporcionalidad la disposición de los mismos hace que sea un tipo de carbohidrato u

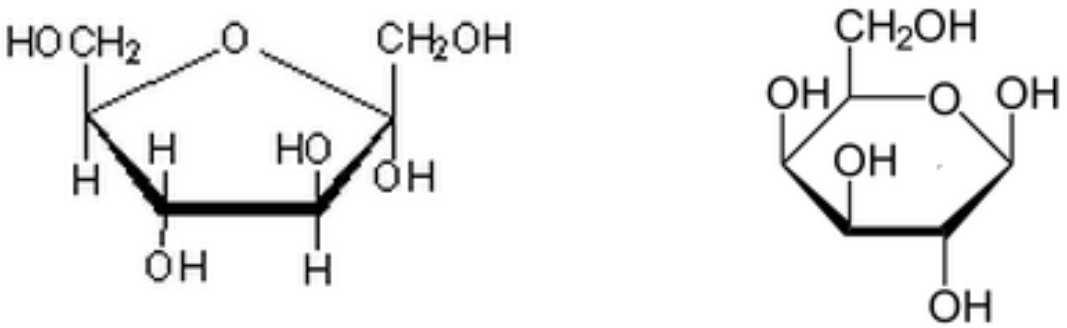


Figura 2.1: *A la izquierda la estructura de la fructosa, a la derecha la de la galactosa.*

otro, como puede pasar en el caso de la Figura 2.1 que es lo que sucede con la galactosa y la fructosa, teniendo así distintas propiedades como la densidad de cada una.

Por lo citado anteriormente la estructura de los carbohidratos es relevante en la biología y por ello se creó el EuroCarbDB que es un proyecto europeo que contiene una base de datos con herramientas e información de la estructura, biosíntesis y biología de los sacáridos así como información glucómica. El EuroCarbDB es un ejemplo de la dificultad intrínseca que existe para mantener un proyecto de libre acceso a la información contenida y es que desde 2009 el proyecto se encuentra detenido por falta de inversores.

Microarrays

Las bases de datos de microarrays [MacBeath and Schreiber, 2000] son unas de las más importantes de R, ya que R tiene desarrolladas herramientas libres para hallar estadísticas, realizar análisis y visualizar los microarrays. Los Microarrays nacen hacia mediados de la década de los 90, como una herramienta para monitorizar simultáneamente los niveles de expresión de varios genes en un conjunto de células. Es una herramienta potente ya que produce una gran cantidad de datos, por lo que el científico debe saber utilizar toda esa información obtenida.

Funciona introduciendo un gen en cada diana del microarray, de dichos genes se mide la concentración en algunas determinadas células. De las células se extrae el ARN para

obtener el ADNc y se le añade un marcador, cada uno se coloca con una diana y si es factible hibridarán. Finalmente se evaluará el material genético que se ha fijado en cada diana.

Se utiliza por tres principales razones, para la genotipificación, detectar genes concretos para localizar mutaciones y poder controlar el riesgo de enfermedades, para medir cuantas veces se expresa cierto gen en un tejido, y para detectar las copias de ADN, pudiendo detectar posibles tumores por un crecimiento descontrolado de determinados genes en un cromosoma.

Uno de los problemas que presentan es que a la hora de verificar un número elevado de hipótesis, como puede ser a la hora de buscar enfermedades, se pueden descartar ciertas hipótesis que en el caso de haberse propuesto de manera aislada no se habría descartado. Si bien es cierto que se están desarrollando trabajos como el trabajo desarrollado por María Isabel Salazar [Salazar Mendoza, 2011] en el que haciendo uso de las aproximaciones bayesianas se reduce.

Una de las principales bases de datos de microarrays es el ArrayExpress[Rustici et al., 2013], a cargo del EBI, es muy importante tanto en rendimiento como en datos, los datos recopilados son acordes a los estándares Minimum Information About A Microarray Experiment (MIAME) y Minimum Information About a Sequencing Experiment (MINSEQE)[2].

Formato FASTA

De las bases de datos citadas anteriormente, muchas de ellas tienen los datos almacenados en formato FASTA, dicho formato consiste en representar mediante una única letra ácidos nucleicos y aminoácidos, su uso está muy extendido en la bioinformática. Cada una de las cadenas de aminoácidos y ácidos nucleicos está en una única línea, o lo que es lo mismo, las cadenas están separadas por saltos de línea (\n). La codificación varía en función de la naturaleza de los datos, así en el caso de tener representadas cadenas de aminoácidos y

[2]No se consideran especialmente relevantes los estándares de cara al proyecto, para más información de MIAME puede consultar [Brazma et al., 2001] y de MINSEQE[MINSEQE,].

Código	Significado aminoácidos	Significado ácidos nucleicos
A	Alanina	Adenosina
B	Asparagina	G o T o C
C	Cisteína	Citosina
D	Ácido aspártico	G o A o T
E	Ácido glutámico	-
F	Fenilalanina	-
G	Glicina	Guanina
H	Histidina	A o C o T
I	Isoleucina	-
K	Lisina	Cetona [G o T]
L	Leucina	-
M	Metionina	Amino [A o C]
N	Asparagina	Cualquiera
O	Pirrolisina	-
P	Prolina	-
Q	Glutamina	-
R	Arginina	Purina [G o A]
S	Serina	Interacción fuerte [G o C]
T	Treonina	Timidina
U	Selenocisteína	Uracilo
V	Valina	G o C o A
W	Triptófano	Interacción débil [A o T]
Y	Tirosina	Pirimidina [T o C]
Z	Glutamina	-
X	cualquiera	Máscara
*	parada de traducción	-
-	gap	gap

Cuadro 2.1: *Código FASTA de aminoácidos y ácidos nucleicos.*

ácidos nucleicos la representación es la mostrada en la Tabla 2.1.

Se puede apreciar que en el caso de que no haya una sección de la cadena de aminoácidos nos encontraremos con lo que se denomina como *gap*, dicha sección con pérdida de datos se puede recuperar mediante diversos sistemas, entre ellos está una variante del algoritmo de *Needleman Wunsch*, desarrollado en lenguaje R por Óscar Sánchez, para la Universidad Complutense de Madrid.

En el caso de los ácidos nucleicos, es importante destacar que solamente hay cinco tipos

Código	Significado
A	Adenosina
C	Citosina
G	Guanina
T	Timidina
U	Uracilo
R	Purina [G o A]
Y	Pirimidina [T o C]
K	Cetona [G o T]
M	Amino [A o C]
S	Interacción fuerte [G o C]
W	Interacción débil [Ao T]
B	G o T o C
D	G o A o T
H	A o C o T
V	G o C o A
N	Cualquiera
X	máscara
-	gap

Cuadro 2.2: *Código FASTA de ácidos nucleicos.*

base que son Adenosina, Citosina, Guanina, Timidirina y Uracilo, el resto sirven para acotar un valor no exacto, esto se da cuando no se sabe exactamente cuál es el ácido nucleico pero sí cuáles no pueden estar en esa posición.

2.2. Métodos estadísticos para analizar los datos

Se denomina estimación estadística a las técnicas que permiten aportar un valor aproximado para un dato concreto. Hay varias formas de aplicarla, a continuación se van a ver las técnicas más utilizadas en la actualidad divididas en dos grupos, el primero de ellos que incluye algunas de las técnicas más sencillas y otro grupo de técnicas más complejas utilizadas en la actualidad.

Como se pretenden desarrollar estimadores cabe destacar que para poder calificar a un estimador como un buen estimador debe cumplir ciertas cualidades. Estas cualidades son insesgabilidad, eficiencia, consistencia y suficiencia.

Insesgabilidad es la propiedad de un estimador que determina que en cierto modo tiene imparcialidad a la hora de obtener los datos. Esto es que tienen las mismas posibilidades, aproximadas, de obtener una estimación que esté por encima del dato real como de obtener una que esté por debajo y que las magnitudes de dichas diferencias son similares. Viene determinado por la diferencia entre la esperanza y el valor esperado teniendo mayor insesgabilidad cuanto menor es dicha diferencia.

La eficiencia viene determinada por la varianza de la resta entre los valores del estimador y los valores reales. Lo idóneo es que tenga una varianza de 0, en cuyo caso no tendrá diferencias con respecto a los datos que se desean calcular. Una varianza cercana a 0, en función de la magnitud de los datos, nos indica, con una probabilidad alta, que dicho estimador obtendrá estimaciones cercanos a los deseados.

La exactitud con que un estimador estima un dato, es una propiedad muy importante ya que en el caso de comparar dos estimadores, por mucho que uno sea insesgado y otro tenga sesgo, si el segundo es considerablemente mejor que el primero podría no tener sentido decantarse por este, ya que el segundo aportaría mejores valores estimados.

La consistencia es una propiedad muy deseable ya que determina la escalabilidad del estimador. Esto es que si hay un estimador que para un tamaño poblacional medio aporta buenos resultados, que dichos buenos resultados se mantengan en el caso de que el tamaño de la población aumente o disminuya. Lo idóneo en un estimador consistente es que según aumenta el tamaño de la población mejore su estimación.

La suficiencia de un estimador se entiende como el uso de la máxima información que contenga la muestra, de modo que ningún otro estimador pueda obtener mayor información sobre el parámetro a hallar, como puede ser la media que utiliza todos los datos existentes de una variable estadística a estimar, al contrario que la moda que solamente cuenta el más repetido.

2.2.1. Técnicas de estimación estadística

Hay un conjunto de técnicas de estimación estadística que aportan datos con una relación *aproximación/coste* muy buena, ya que no tienen un alto coste computacional y ofrece una aproximación suficientemente buena.

Estas técnicas se utilizan habitualmente para tener un caso base invirtiendo un tiempo razonablemente pequeño. Por ello son consideradas bastante útiles, ya que sirven para compararlas con otras técnicas más elaboradas y evaluar, en cierto modo, su calidad. A continuación se explican algunas de ellas.

Media

Una forma muy sencilla de estimar un dato es mediante la media del conjunto, $\bar{x}$, donde x es na variable estadistica con n datos, tomando i valores entre 1 y n . La Fórmula 2.1 muestra la fórmula de la media no tiene en cuenta la varianza por lo que si los datos sufren muchas variaciones no se reflejarán. Tampoco tiene en cuenta la tendencia y por lo tanto si una serie tiene un valor eminentemente creciente la estimación de un dato ofrecerá una estimación que salvo que el dato a estimar esté hacia el centro de la serie no ofrecerá una estimación buena, por ofrecer un dato demasiado elevado en los primeros casos o muy pequeño en los últimos.

$$\bar{x_j} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

Una derivación de esto son las medias móviles, la idea es la misma que la de la media pero en vez de realizar la estimación mediante una media global se realiza sobre una media local como se muestra en la Fórmula 2.2. El hecho de que sea más o menos local varía de la naturaleza de los datos, en el caso del ejemplo se toman 7 valores para realizar la media móvil por hacer referencia a un coste semanal.

$$\bar{coste_j} = \frac{\sum_{i=j-3}^{j+3} coste_i}{7} \tag{2.2}$$

Centrándose en el ejemplo, el coste de un producto a lo largo de la semana, no es difícil tender a pensar que si se quiere generalizar la fórmula, en función del día de la semana

pueden cambiar muchos elementos, como puede ser el consumo de una persona que suele aumentar de manera notable los fines de semana. En ese caso se pueden aplicar pesos a la estimación, dichos pesos asignados en función del día que se quiera hallar darán mayor o menor relevancia a cada día. En la Fórmula 2.3 se aprecia que ahora hay una matriz de pesos y_{ji} donde la variable i corresponde al peso que tiene el día de la semana a estimar (j), de lunes a domingo.

$$\bar{coste}_j = \frac{\sum_{i=j-3}^{j+3} coste_i * y_{ji}}{7} \tag{2.3}$$

Por otro lado existen técnicas que tienen en cuenta un mayor número de factores y por lo tanto son más completas que las anteriores y por lo general más precisas. Algunas de las técnicas más utilizadas son las que se muestran a continuación.

Regresión lineal

La primera regresión lineal de la historia fue el método de los mínimos cuadrados de Legendre. Se puede utilizar como una forma de estimar la relación entre dos variables. Su nombre de regresión lineal no implica que sea una recta, sino que se llama así para distinguirlo de las técnicas de regresión que utilizan modelos basados en funciones matemáticas. La Función 2.4 muestra la función de la regresión lineal.

$$Y = \sum \beta_k X_k + \varepsilon \tag{2.4}$$

Correlación

Para estimar los datos también se pueden utilizar otras variables que sí tengan datos, para ello habrá que utilizar la correlación. Es una relación lineal entre dos variables estadísticas. Se divide en tres componentes, la fuerza que es el grado de relación, el sentido que mide si ambas variables crecen a la vez o si según crece una decrece la otra, relación positiva y negativa respectivamente y por último se tiene la forma que indica la línea que define el ajuste, linea recta, curva monotónica o curva no monotónica. Para medir dicha correlación se utilizan coeficientes de correlación, el más conocido es el coeficiente de correlación de

Pearson [Pearson, 1896].

Para saber si hay una correlación fuerte entre las variables hay que calcular los coeficientes de correlación entre todas las variables y tener en cuenta solamente aquellos que se consideren suficientemente fuertes, en este trabajo se considera una correlación suficientemente fuerte aquella con coeficientes superiores al 80 %. En el caso de que la correlación sea alta hay que buscar la forma de dicha correlación, y encontrar la función a la que se ajusta. Pudiendo utilizar una regresión lineal en el caso de que la correlación sea lineal.

El caso ideal a la hora de hallar una relación entre dos variables es el mostrado en la Fórmula 2.5.

$$y_i - \hat{y}_i \approx 0 \tag{2.5}$$

Siendo y, la variable que se quiere hallar en función de otra, la variable x, e $\hat{y}$ la estimación en función de dicha variable. Dándose esta condición ideal para cada muestra i perteneciente al conjunto de datos. Para facilitar la búsqueda de dicha condición se puede crear una variable error como la mostrada en la Fórmula 2.6.

$$error_i = y_i - \hat{y}_i \tag{2.6}$$

Así se tiene que lo ideal es que la variable *error* tenga una media lo más cercana a 0 posible y una varianza cuyo valor sea pequeño en comparación con la magnitud de y. Cumpliéndose lo anterior se tiene que una buena estimación tendrá la cualidad mostrada en la Fórmula 2.7.

$$R^2 = \frac{S_y^2 - S_{error}^2}{S_y^2} \tag{2.7}$$

R^2 representa la fuerza de la correlación, que indica en qué medida se ajusta una aproximación a la variable que se pretende aproximar. Además S_y determina la varianza de y mientras que S_{error} representa la del error.

En este trabajo se va a realizar una aproximación lineal, por lo que las funciones serán de la forma $\hat{y} = a + bx$ y por lo tanto hay que conocer el valor de las variables a y b para conocer la aproximación[3].

[3]Se puede apreciar que b determina la dirección de la pendiente de la relación, así si b es positiva ambas

Para poder hallar a y b lo que se tiene que hacer es minimizar el error total, para ello generalizando la Fórmula 2.6 y evitando compensación de errores, haciendo uso de cuadrados para que siempre tenga un valor positivo, obtenemos la Fórmula 2.8, en la cual n representa el número total de muestras.

$$error(a,b) = \sum_{i=0}^{n}(y_i - a - bx_i)^2 \tag{2.8}$$

Se puede apreciar que $\hat{y}_i$ ya está desglosada como $a + bx_i$, tal y como se ha comentado anteriormente.

A continuación usando el procedimiento de los mínimos cuadrados y derivando tanto por a como por b se obtienen las ecuaciones para poder hallar los valores de a y b. Estas ecuaciones son mostradas en las Fórmulas 2.9 y 2.10 respectivamente

$$a = \bar{y} - b\bar{x} \tag{2.9}$$

$$b = \frac{S_{xy}}{\sqrt{S_x} * \sqrt{S_y}} \tag{2.10}$$

Siendo b el coeficiente de correlación entre x e y mientras que S_{xy} la covarianza entre x e y. Una vez se sabe como hallar los valores de a y b solamente hay que aplicarlo a las dos variables que se quieran comprobar para ver si realmente están correlacionadas.

Sistemas expertos

En muchos casos la mejor alternativa para estimar un dato puede ser realizar un sistema experto.

Un sistema experto consiste en la aplicación de ciertos conocimientos sobre un problema específico que puedan ayudar a solucionarlo, emulando así el proceder que tendría un experto en la materia. Pero para aplicar dicho conocimiento de una manera correcta y por lo tanto el sistema experto sea efectivo se tienen que cumplir dos capacidades.

La primera de ellas es que el conocimiento se debe aplicar de una manera sencilla facilitando la creación de una serie de reglas que se basan en hechos. La segunda y última

variables crecerán al unísono mientras que si es negativa el crecimiento de una indicará el decrecimiento de la otra

-	A	T	C	G
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Cuadro 2.3: *Ejemplo de matriz de pesos del algoritmo de Needleman-Wunsch.*

es la capacidad para modificar dichas reglas con relativa facilidad ya que la perspectiva de los propios expertos puede verse modificada de algún modo, lo que conllevaría a tener que modificar el sistema experto.

Un ejemplo de aplicaciones de sistemas expertos en bioinformática se da con las matrices de pesos utilizadas en la variante del algoritmo de Needleman-Wunsch desarrollado por Óscar Sánchez [Sánchez, 2013] (aún sin publicar). Dicho algoritmo es usado generalmente en el ámbito de la bioinformática para alineamiento de secuencias de proteínas o ácidos nucleicos. Para conseguirlo hace uso de una matriz de pesos que se puede observar en la Tabla 2.3, dichos pesos por supuesto han sido asignados previamente por un experto para optimizar los resultados.

Redes bayesianas

Dentro del ámbito de los sistemas expertos destacan las redes bayesianas. Las redes bayesianas son gafos dirigidos acíclicos compuestos por un conjunto de nodos donde cada uno de ellos representa a una variable. Los arcos que unen los nodos, dos a dos, determinan la dependencia entre variables, así si se tiene $x \rightarrow y$ significa que la variable y tiene cierta dependencia con la variable x y se denomina que x es padre de y. Además cada nodo contiene una distribución condicional de probabilidad, donde el experto es el que decide la topología de la red así como las distribuciones de probabilidad. A continuación se explican más detalladamente los Modelos de Markov, que se consideran las redes bayesianas más simples.

Modelos ocultos de Markov

En bioinformática en muchos casos se aplican sistemas que funcionan en otras ramas de

la informática, este es el caso de los Modelos ocultos de Markov. En éstos se parte de la suposición de que el sistema es un proceso de Markov, por lo que es necesario explicarlo para poder comprender los Modelos ocultos.

Éstos son sucesos aleatorios que dependen del tiempo y que cumplen la propiedad de Markov. Dicha propiedad obliga a que las transiciones de un estado con otro solamente pueda suceder con el inmediatamente anterior o el inmediatamente posterior, lo que significa que si el estado actual es el estado i las únicas transiciones válidas son a $i-1$ y a $i+1$. Además de lo anteriormente citado, los modelos ocultos de Markov deben su nombre a que no se conocen los estados a ciencia cierta, sino que se conocen las variables afectadas por dicho estado. Dichas variables toman los valores en función de las distribuciones de probabilidad existentes en cada estado.

Un ejemplo sencillo de Markov, puede ser que sabiendo el tiempo atmosférico en este momento, *soleado*, *lluvioso* o *nublado* que en este caso serán las variables, se puede intuir en qué estación estamos haciendo uso de las distribuciones de probabilidad, ya que en cada estación hay unas condiciones u otras. En invierno puede haber un 0,4 de probabilidades de que esté *lluvioso*, un 0,4 de probabilidades de que esté *nublado* y un 0,2 de que esté *soleado*. Mientras que en verano las probabilidades de soleado pueden ser de 0,8.

En la bioinformática se utilizan para la detección de genes. Uno de los problemas que tienen es que necesitan un número suficiente de casos de entrenamiento, para que la precisión sea suficientemente buena

2.3. Análisis de bases de datos con bioestadística

Es un lenguaje de programación estadístico ya que es la implementación libre, distribuído bajo licencia GNU, del lenguaje S y con una única condición de uso que se respeten los términos de licencia de GNU General Public License version 2.

Ross Ihaka y Robert Gentleman comenzaron su desarrollo en 1993 y si bien es cierto que sigue en continuo desarrollo desde el 29 de Febrero del 2000 dispone de una versión

suficientemente estable para usarla en producción.

R es utilizado principalmente para realizar análisis estadísticos y por ello tiene tanta relevancia en el mundo de la bioinformática, además ofrece una amplia gama de paquetes con todo tipo de funcionalidades en los distintos campos de la informática esto lo convierte en un lenguaje con un gran potencial. Está en continua expansión, ya no solo su grupo de desarrollo se encarga de desarrollar nuevas versiones de R sino también por parte de los usuarios dado que los paquetes de R son de código libre y por lo tanto abiertos a modificaciones.

R proporciona un amplio abanico de herramientas estadísticas como pueden ser modelos lineales y no lineales, así como test estadísticos y análisis de series temporales. También permite de una manera muy sencilla crear gráficas, lo que nos permite realizar gráficos de alta resolución para poder localizar algunos datos concretos con mayor claridad como pueden ser outliers utilizando box plots o una serie de datos mediante un plot, más adelante se explican ambas funciones de manera más detallada.

Además R permite el desarrollo de bibliotecas en C, C++ y Fortran y hacer una carga dinámica de ellas. De hecho muchas de las funciones nativas de R están desarrolladas en lenguaje C ya que a la hora de realizar código computacionalmente exigente tiene mayor rendimiento en tiempo de computación si se realiza en C que en el propio R.

Es importante destacar que R facilita la integración con bases de datos así como utilizar Perl o Python gracias a ciertas bibliotecas, también permite interactuar con Weka haciendo uso de RWeka5, que permite leer y escribir ficheros .arff y disponer de minería de datos en R. Actualmente hay bibliotecas para cargar y guardar archivos en formato .xls y .xlsx que son los utilizados por el paquete de programas Office para almacenar las tablas y manipularlas desde el programa Excel, que en muchos casos serán bases de datos, de hecho las tres bases de datos del proyecto están en dichos formatos.

También se usan los datos en formato CSV (Comma-Separated Value). Este formato sirve para almacenar datos de tablas, donde cada fila está separada por un salto de línea,

mientras que los datos de una misma fila están separados mediante dos tipos de separadores, coma (,) o punto y coma (;). Para evitar conflictos a la hora de interpretar los datos, como puede darse en el caso de los decimales, se pueden poner entre comillas (").

Es un formato muy útil y ampliamente extendido en la carga y lectura de datos masivos, usando la librería ff [Adler et al., 2010], dicha librería permite cargar de manera rápida datos de bases de datos con gran carga de información. Por ejemplo en el caso de querer almacenar el número de veces que aparecen ciertos números, para almacenar que el número π ha aparecido 14 veces se tiene, 3,1415,14, donde 3,1415 representa el valor del número π y 14 el número de veces que ha aparecido repetido. En el propio número π hay una coma lo cual crearía conflicto dando tres variables en esta fila una con valor 3, la siguiente con valor 1415 y por último una con valor 14. En cambio si se introducen los valores entre comillas quedaría como se muestra a continuación "3,1415","14"dando únicamente los dos valores esperados, otra posible solución era utilizar como separador el punto y coma quedando 3,1415;14.

R además tiene *Lexical scoping* que es el uso de variables solamente en el ámbito privado, por lo que solo se pueden utilizar en la función en que son definidas, esto es muy importante ya que al ser un lenguaje interpretado si no fuera así habría acceso a todas las variables auxiliares de las funciones y se tendría que o bien obligar al usuario a eliminar las variables cuando implementa una función, como sucede en lenguajes como C, o tener control de basura, como es el caso de Java, de hecho es una especie de control de basura, ya que al final de las funciones se eliminan las variables locales.

R permite la inclusión de paquetes de una manera muy sencilla, solamente hay que descargarse el paquete objetivo usando la función $install.package(nombrePaquete)$ y posteriormente utilizarlo mediante la función $library(nombrePaquete)$, además los paquetes que están en la página oficial de R están ordenados según su naturaleza y función para que la accesibilidad a los mismos sea más sencilla.

Es importante destacar en primera instancia dos proyectos relacionados con R, el primero de ellos está relacionado con lo citado anteriormente de poder compartir con la comunidad un

proyecto desarrollado por el usuario y es *The Comprehensive R Archive Network* (CRAN).

CRAN

CRAN[4] nace el 23 de Abril de 1997 con tres espejos y 12 paquetes descargables para el lenguaje R. Su objetivo es el de facilitar el desarrollo global y conjunto de R, aportando un sitio único donde poder descargar los paquetes que necesite el usuario, así consigue que sea más sencillo localizar los paquetes existentes e incluso permite que los propios usuarios suban sus paquetes desarrollados, esto es muy importante ya que de hecho muchos de los paquetes han sido desarrollados por los usuarios. Si bien es cierto que la rigurosidad para poder agregar un paquete propio es muy elevada.

Además es la red oficial de R, en dicho proyecto se encuentra el propio R, para Linux, Windows y Mac. Consiste tanto en una red ftp como en servicios web dispuestos alrededor del mundo, todos están sincronizados para tener así todos la misma versión de código y la misma documentación. Actualmente hay más de 4680 paquetes en un total de 90 servidores, incluyendo uno en Madrid.

Bioconductor

El segundo proyecto que es importante destacar es *Bioconductor*[Gentleman et al., 2004]. Ya que es uno de los proyectos más influyente para este trabajo. Nace en el 2001 con el objetivo de desarrollar software para ayudar a analizar datos en trabajos de laboratorio en biología molecular. El software está desarrollado en R y se centran principalmente en los análisis de microarrays. El proyecto proporciona un amplio conjunto de métodos para analizar datos genéticos también facilita el rápido desarrollo de software escalable. Actualmente tienen 671 paquetes de software y una amplia comunidad de usuarios.

En este proyecto así como todos los que componen la librería BioSeq, se pretende imitar el modelo de crecimiento de Bioconductor. Este modelo consiste en comenzar desarrollando librerías en R útiles para la comunidad, para lo cual es necesario conocer como crear librerías

[4] http://cran.r-project.org/

en R, proyecto desarrollado por Jorge Martínez, creando así una base para mejorar en tiempo y calidad el desarrollo de las librerías.

Gestión de bases de datos

También es necesario gestionar las bases de datos en R, cargarlas en una variable en R y realizar las modificaciones pertinentes para conseguir los objetivos. Es importante destacar que en ninguna de las tres bases de datos se tienen la totalidad de los datos, el 100 % de los datos, esto es que tiene pérdida de datos, la cual se representa como NA, del inglés not available. En R hay paquetes que permiten cargar los datos desde archivos como .xls, .xlsx, usado frecuentemente para almacenar datos y verlos desde el programa Excel de Microsoft Office, o el editor de tablas de LibreOffice[5] que es un programa de características similares al anterior pero de código libre.

Boxplots

Para la búsqueda de datos anómalos es muy extendido el uso de diagramas de caja (*boxplots*). Los diagramas de caja son gráficos mediante los cuales se muestran diversos datos referentes a la distribución de la muestra. Es importante destacar que cada *boxplot* hace referencia a cada variable, ya que los datos atípicos es una característica de la propia variable y su distribución, por lo tanto para poder hallar los valores representados en los *boxplots* se ordenan los datos de manera ascendente. Como se puede ver en la Figura 2.2 el boxplot permite localizar fácilmente los valores orientativos, ya que al ser un gráfico en muchos casos no mostrará el valor exacto, del primer cuartil ($q1$), el tercer cuartil ($q3$), la mediana ($Mediana$), el Rango Inter Cuartílico (RIC), así como los dos límites, el inferior (L_i)y el superior (L_s), por último está el máximo valor menor que L_s denominado ($max(x)$), el mínimo valor mayor que $L_i(min(x))$ y los valores atípicos ($Atipico$), a continuación se hace una breve descripción de dichos datos para facilitar su comprensión.

Los cuartiles dividen la distribución y como su propio nombre indica la dividen en 4

[5]http://www.libreoffice.org/

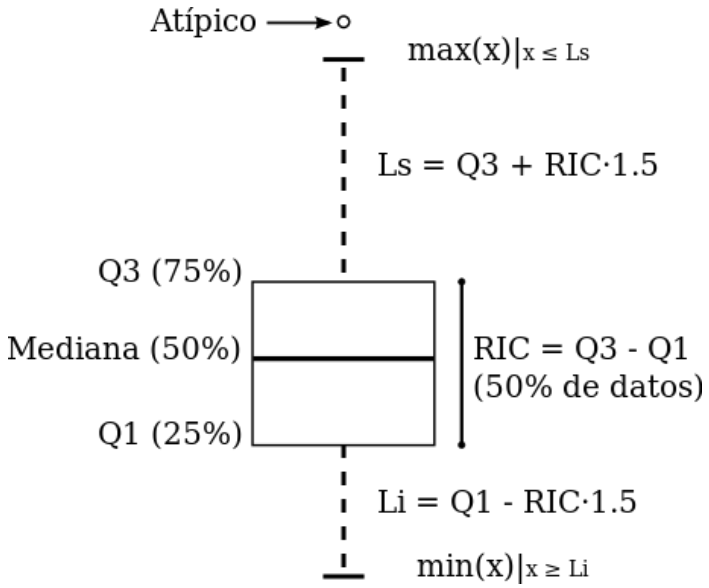


Figura 2.2: *Ejemplo de un* boxplot *destacando la información revelada en el mismo.*

partes, así el $q1$ pertenece al elemento que está inmediatamente por encima de un 25 % del total, el $q2$ o *Mediana* es el elemento que deja por debajo el 50 % de la distribución y el $q3$ es aquel que deja por debajo el 75 % de los elementos, como se puede apreciar para hallar los cuartiles no tiene importancia los valores, si no solamente la distribución de los datos.

Por último el $RIC = q3 - q1$ es la distancia que hay entre $q3$ y $q1$ y comprende todos aquellos elementos contenidos entre el $q1$ y el $q3$, quedando así un 50 % de la distribución en su interior y un 50 % fuera.

Para poder calcular L_s y L_i hay que realizar dos operaciones, mostradas en las Fórmulas 2.11 y 2.12.

$$L_s = q3 + RIC * 1{,}5 \tag{2.11}$$

$$L_i = q1 - RIC * 1{,}5 \tag{2.12}$$

El $min(x)$ y $max(x)$ como se ha explicado anteriormente son el mínimo valor superior a L_i y el máximo valor inferior a L_s respectivamente, los elementos desde q_1 hasta $min(x)$ y desde q_3 hasta $max(x)$ están representados por unas líneas discontinuas que junto a las líneas perpendiculares que las cortan se denominan bigotes.

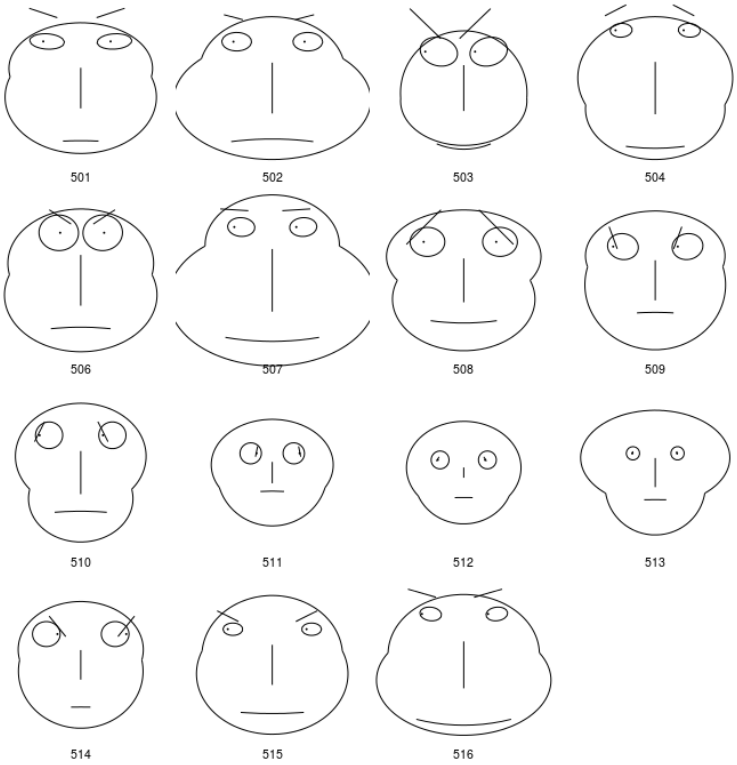


Figura 2.3: *Ejemplo de las caras de Chernoff aplicadas a 15 registros de 18 variables.*

Así los denominados *bigotes* van desde $q3$ hasta $max(x)$, el superior, y desde $q1$ hasta $min(x)$, el inferior, todos aquellos valores que no están en el rango $[L_i, L_s]$ son los llamados *outliers* o valores atípicos y son los datos que hay que tratar.

Caras de Chernoff

Actualmente el único medio plausible para poder mostrar gráficamente un entorno multivariante son las Caras de Chernoff[Chernoff, 1973].

Las caras de Chernoff convierten las variables de una tabla en rasgos físicos de una cara, así cada entrada en la tabla, cada registro, es representado por una cara teniendo así tantas caras como registros existan, en la Figura 2.3 se aprecia un ejemplo de las caras de Chernoff, usando el paquete implementada en R denominado TeachingDemos[6]. Las características físicas sobre las que se reflejan las variables para formar las caras son las mostradas en la

[6]Las caras de Chernoff se representan usando la función faces2 del paquete.

Número de Variable	¿Qué representa?
1	Anchura del centro
2	Superior Vs inferior, altura de la separación
3	Altura de la cara
4	Ancho de la mitad superior de la cara
5	Ancho de la mitad inferior de la cara
6	Largo de la nariz
7	Altura de la boca
8	Curvatura de la boca (abs <9)
9	Ancho de la boca
10	Altura de los ojos
11	Distancia entre los ojos (.5-.9)
12	Ángulo de ojos y cejas
13	Elipse de los ojos
14	Tamaño de los ojos
15	Posición izquierda/derecha de los ojos
16	Altura de las cejas
17	Ángulo de las cejas
18	Ancho de las cejas

Cuadro 2.4: *Representación de cada variable en las caras de Chernoff.*

Tabla 2.4.

Es importante destacar dos puntos, el primero que el orden de las variables es importante, ya que en función de su posición se utiliza para representar una parte u otra de la cara. Y el segundo que el número máximo de variables a representar es 18, si bien es cierto que es un número considerablemente elevado, es una limitación a tener en cuenta.

2.4. Métodos de depurar y/o eliminar datos

Para poder realizar comparaciones entre diversos métodos de estimación no se puede realizar sobre datos que realmente estén perdidos, ya que no se conocerá cuál de las estimaciones ha sido mejor, ni si quiera se podrá saber si alguna de las estimaciones ha aportado un dato satisfactorio para poder considerarla buena. Por ello es necesario que la pérdida de datos sea controlada, para poder compararlas. Dicha pérdida de datos es ficticia, no se

pierden los datos sino que se crea una base de datos auxiliar con los datos perdidos, pero obviamente la original se mantiene. La pérdida de datos se puede realizar de manera totalmente aleatoria en función de un parámetro que es la pérdida de información, en porcentaje de datos perdidos. Este método es el llamado MCAR (Missing completely at random). La pérdida de datos también puede hacerse mediante el método MAR (Missing at Random). El método MAR en cambio da mayor probabilidad de pérdida de datos si las variables anteriores de dicha fila no tienen una pérdida de datos y asigna el valor normal de pérdida para filas con valores perdidos.

Esto implica que con menor porcentaje de pérdida de datos se pueda conseguir, prácticamente una pérdida de datos por fila, sobre todo cuanto mayor es el número de variables, lo que implica mayor número de columnas.

En muchos casos puede ser interesante que haya pocos datos perdidos por fila, pero que la cantidad de datos perdidos sea suficientemente alta como puede ser en el caso de que haya correlaciones entre variables.

2.5. Métodos para comparar datos estimados

Lo primero que hay que hacer en este caso es verificar si en efecto es posible comparar dos estimadores. Teniendo en cuenta que los estimadores son variables aleatorias con cierta distribución de probabilidad, se puede afirmar que tienen ciertas propiedades como la varianza o el sesgo que permiten conocer la calidad de los estimadores y por lo tanto permite comparar unos con otros. Si bien por lógica, se puede afirmar que en el caso de que no fuera posible conocer la calidad de un estimador no tendrían sentido, ya que no se podría decidir entre uno u otro. Es importante destacar los métodos que existen para comparar cuál de entre dos o más estimaciones es la que más se ajusta a los datos reales. Para ello se tienen que tener en cuenta las cualidades que debe tener un buen estimador. Dichas cualidades son insesgabilidad, eficiencia, consistencia y suficiencia y a continuación se explican formas de comparar dichas cualidades.

Insesgabilidad

La insesgabilidad es la no presencia de sesgo y como se ha explicado anteriormente se calcula hallando la diferencia entre la esperanza y el dato obtenido y buscando que dicho dato sea mínimo. Es importante destacar que la ecuación de la esperanza (E) de una variable estadística (x), mostrada en la Función 2.13, depende de la probabilidad de cada suceso, por lo que se tiene que conocer la distribución de los datos (p) para poder hallarla, en la ecuación hay un total de n valores sin repetición en x, ya que las repeticiones están representadas por p_i, para cada $i \in (0, n)$.

$$E(x) = \frac{1}{n} * \sum_{i=0}^{n} x_i * p_i \tag{2.13}$$

Se puede apreciar que la función de la esperanza se ajusta a la media de la variable aleatoria, esto es debido a que no tiene sesgo ($s(x)$) en cuyo caso la función de la esperanza es la mostrada en la Fórmula 2.14, por lo tanto la fórmula de la esperanza entre un estimador insesgado y uno sesgado es esta última y solamente difieren en el valor del sesgo.

$$E(x) = \hat{x} + s(x) \tag{2.14}$$

La distribución de un conjunto de datos representa la frecuencia con la cual dichos datos toman ciertos valores. En circunstancias normales, teniendo un conjunto de datos suficientemente grande se tiene que el conjunto de datos se acerca, suficientemente a una distribución de datos normal, representada como una campana de Gauss, que se puede apreciar en la Figura 2.4. Obviamente hay más tipos de distribuciones como son la binomial o la de Poisson, pero lo más común es que los datos se ajusten a la normal, de ahí su nombre.

Eficiencia

A la hora de medir la eficiencia se tiene que lo óptimo para cumplir esta propiedad es que un estimador tenga un mínimo valor en la varianza de la diferencia entre el valor real y el que se ha estimado. Por lo tanto para conocer la eficiencia la única posibilidad es hallar dicho valor. En este caso es importante evitar que los errores se anulen entre ellos como

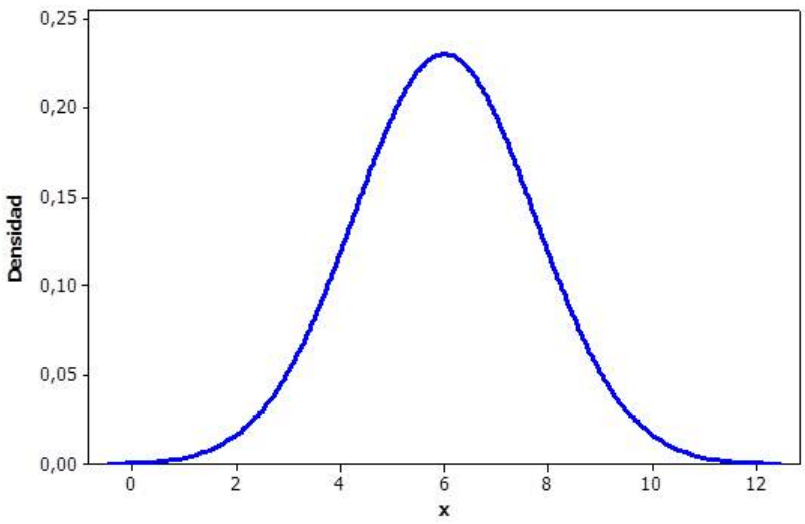


Figura 2.4: *Muestra de una distribución de datos normal.*

sería el caso de tener dos errores de altas magnitudes pero distinto signo, si se sumaran los errores obtenidos se tendría un error cercano a 0 y se determinaría que la estimación ha sido muy buena, en términos de eficiencia, pero sería una conclusión errónea.

$$ECM(x) = \frac{\sum_{i=0}^{n} x_i^2}{n} \tag{2.15}$$

Por ello es necesario no solo hallar el error, si no hallarlo de manera correcta para lo cual se puede hallar el error cuadrático medio (ECM), mostrado en la Fórmula 2.15, que consiste en elevar a la potencia de 2 todos los errores (x) con la finalidad de evitar lo mencionado anteriormente y a continuación dividirlo entre el número de elementos (n).

Se puede por lo tanto definir el ECM puntual como el cuadrado de la esperanza entre el valor esperado y el obtenido siendo la situación ideal que sea 0, ya que es el caso en que ambos valores coinciden, dicha ecuación está reflejada en la Fórmula 2.16.

$$ECM(\hat{x}, x) = E[(\hat{x} - x)^2] \tag{2.16}$$

Y siguiendo con esto y según indica la Fórmula 2.17, siendo por lo tanto en ECM la suma

entre la varianza de una variable aleatoria ($var(x)$) y el sesgo ($(E(\hat{x}) - x)^2$)) al cuadrado.

$$E[(\hat{x}-x)^2] = E[(\hat{x}-E[\hat{x}]+E[\hat{x}]-x)^2] = E[(\hat{x}-E[\hat{x}])^2]+(E[\hat{x}]-x)^2 = Var(\hat{x})+(E(\hat{x})-x)^2 \tag{2.17}$$

Por lo tanto se puede aceptar que el ECM es un compromiso entre la eficiencia de un estimador y la insesgabilidad, ya que tiene en cuenta ambas propiedades.

Consistencia

La consistencia de un estimador es la capacidad de mantener sus prestaciones, o mejorarlas, según aumenta el tamaño de la población, por lo que para medir la consistencia de un estimador se debe realizar un proceso de pruebas. Este proceso de pruebas consiste en realizar estimaciones de diversa magnitud sobre una misma base de datos, para así poder comparar los resultados obtenidos de todas.

La única diferencia entre las pruebas es el tamaño de datos, debiendo cumplir que en el caso de ordenar las pruebas de menor a mayor magnitud de datos, se obtenga que a mayor magnitud de datos mejores resultados se obtienen, tal como muestra la Fórmula 2.18. En esta fórmula la calidad se mide haciendo uso del ECM y por lo tanto un valor más grande hace referencia a un peor resultado.

Por lo tanto los resultados obtenidos con una muestra pequeña de la población deben de ser peores o como mucho mejores que los obtenidos con una muestra de tamaño medio y esta a su vez debe ser peor o a lo sumo igual que para el total de la población.

$$ecm(p_1) >= ecm(p_2) >= ... >= ecm(p_n) \tag{2.18}$$

Suficiencia

A la hora de valorar que un estimador tenga una buena suficiencia no es posible realizarlo de manera automática, al contrario que en los casos anteriores, ya que solamente un experto puede tener conciencia de si es o no posible obtener mayor información de los datos. No es posible utilizar una fórmula o desarrollar un código para comprobar si exprime al máximo

los datos, ya que por mucho que se detectara si un código recorre todos los datos disponibles de la tabla, no se puede asegurar de manera genérica para cualquier variable que se vaya a estimar que para que la estimación sea lo más precisa posibel se utilicen las relaciones existentes entre todas las variables de la base de datos.

En el siguiente capítulo se va a ver el trabajo desarrollado en este proyecto, el trabajo se muestra según el orden en el que hay que tratar los datos, primero se cargan, a continuación se realiza la localización visual de outliers, seguido de la búsqueda de correlaciones, posteriormente se aplica un sistema experto para los casos en los que las correlaciones no son suficientes. Por último se guardan los datos en un formato determinado por el usuario.

Capítulo 3

Análisis exploratorio de datos

A continuación se va a realizar el análisis exploratorio de datos en R. Para ello en cada sección primero se enumeran las librerías utilizadas de CRAN necesarias para esa sección. Seguido de la explicación de cómo se carga y almacena una base de datos en este trabajo, seguido de la búsqueda de outliers en la base de datos de las cabras haciendo uso de las caras de Chernoff y de la búsqueda de correlaciones en la misma base de datos y las ecuaciones pertinentes para las correlaciones aceptables. Finalmente se aplica un sistema experto a las bases de datos de los carburantes y se analizan sus resultados aplicando un comparador de estimadores.

3.1. Carga de datos

Para poder realizar cualquier tratamiento de datos primero es necesario cargarlos desde R, para poder manipularlos en dicho lenguaje. Para ello en este trabajo se han utilizado librerías del proyecto CRAN. Estas librerías son *XLConnect*[XLConnect,] usada para cargar datos con extensión *.xls* y *.xlsx*, para conocer estas extensiones se ha utilizado el paquete *tools*. Como no solamente se cargan archivos *.xls* y *.xlsx* se ha utilizado también la librería *foreign* [1], la cual permite la carga de formatos CSV y FASTA. Una vez se conocen las liberías usadas se puede explicar el trabajo realizado.

Para comprender a la perfección el funcionamiento de algunas de las funciones desarro-

[1] http://www.ats.ucla.edu/stat/r/faq/inputdata_R.htm

lladas es importante destacar una propiedad muy interesante de R y es que a diferencia de otros lenguajes no necesita que el usuario asigne valores a todos los parámetros de entrada, ya que se les puede asignar valores por defecto.

Esta particularidad de R, también existente en Prolog entre otros, sirve para dotar a las funciones de una mayor versatilidad además de hacer que se ajusten al conocimiento y nivel del programador o usuario, ya que en muchos casos los parámetros tienen valores por defecto que el usuario podrá introducir o no en función de sus necesidades u objetivos.

Las bases de datos en R se almacenan como variables de tipo *data.frame*, que internamente está compuesto de tantos tipos como columnas haya sin la necesidad de que sean del mismo tipo, ya que entonces sería una lista.

Para la carga de archivos hay 3 funciones específicas y una global. Las funciones específicas permiten cargar archivos con extensiones .xls y .xlsx, formato FASTA y formato CSV, mientras que la global aúna estos tres casos.

3.1.1. Carga de archivos .xls y .xlsx

La primera carga que se va a ver es la carga de archivos con extensiones de hoja de cálculo de *Excel*. En R existen ya funciones para cargar dichos archivos, en este trabajo se ha utilizado una de estas funciones para cargar archivos y para ello se ha tenido que hacer uso del paquete *XLConnect*, que permite hacer la conexión con el archivo y cargar los datos. Si bien es cierto que permite cargar otro tipo de formatos este trabajo se ha centrado en la carga de las extensiones de archivo anteriormente citadas.
La función de carga tiene tres parámetros, *archivo*, *hoja="No definida"* y *create=FALSE*. El parámetro *archivo* hace referencia al archivo que hay que cargar, en el caso de que la extensión del archivo sea incorrecta dará un aviso y se frenará la carga; *create* indica si se ha de crear la hoja en caso de que no exista a la hora de cargarla, lo cual puede resultar útil si se quiere comenzar desde cero una base de datos o si es necesario que dicho archivo exista para el correcto funcionamiento de otras funciones; *hoja* en el caso de que el archivo

a cargar tenga varias hojas de cálculo esta variable indica cuál de todas es la que hay que cargar.

Se ha agregado una funcionalidad a la carga estándar del paquete, si bien no se ha incluido en el paquete en sí, sino que se ha agregado a la función de carga realizada en este trabajo.

Dicha funcionalidad es que, dado que R es un lenguaje interactivo, al introducir el usuario la carga de un archivo con extensión de hoja de cálculo, se comprueba las hojas que tiene, en el caso de tener más de una y no haber sido especificada ninguna por el usuario, cuando *hoja="No definida"* se pregunta qué hoja desea cargar el usuario, mostrando los nombres de las hojas de cálculo disponibles. Así se busca que sea más sencillo para el usuario dado que no necesita tener un conocimiento previo de las hojas de cálculo. Si bien es cierto que el nombre de las hojas de cálculo ha de ser suficientemente esclarecedor, para que el usuario reconozca la que quiere cargar.

3.1.2. Carga de archivos FASTA

Como se ha dicho anteriormente en el Punto 2.1 en la bioinformática es muy importante el formato FASTA, por ello pese a que no es un formato útil para las bases de datos utilizadas en este trabajo, ya que habría una pérdida de datos considerable en cualquiera de ellas, existe una función para poder cargar datos en dicho formato. Esto es debido a que se quiere tener la mayor versatilidad posible y para ello es indispensable que la librería pueda cargar datos en este formato.

Al igual que en la carga de hojas de cálculo se utiliza un paquete para la carga de archivos en formato FASTA, dicho paquete es **PAQUETE FASTA** y la función que se utiliza de dicho paquete es *read.table*, en este caso no se realiza ninguna modificación sobre dicha función sino un simple encapsulado.

La función que realiza la carga tiene los siguientes parámetros. *archivo* al igual que en el caso anterior es el archivo del que se desea extraer los datos; *headerFun=TRUE* se pone

a *TRUE* en el caso de que se quiera cargar con los datos la cabecera de la base de datos, el nombre de las columnas. En este caso por defecto está a *TRUE* por querer cargarla completa si bien se permite al usuario la posibilidad de que modifique dicho valor; *sepFun=”* es el separador utilizado entre cada uno de los valores de cada fila, el valor por defecto ” (carácter vacío) hará que al llamar a la función read.table tome como separadores espacios en blanco, uno o más, tabulados, saltos de línea y retornos de carro.

3.1.3. Carga de archivos CSV

Otro de los formatos que es importante cargar en R, es el formato CSV, explicado en el Punto 2.3, en este caso sí es útil cargarlo, ya que dicho formato no implica ningún tipo de pérdida para las bases de datos.
Para cargar los datos en formato CSV se utiliza la función *read.csv* del paquete **PAQUETE CSV**, con los mismos parámetros y función de los mismos que en el punto anterior.

3.1.4. Carga genérica

Además se ha incluido una función genérica para la carga de archivos, que lo que hace es aunar las anteriormente citadas, para ello se le introducirá por parámetro el *archivo="BasesDatos.xlsx"*, con idéntica función que en los casos anteriores a diferencia de que en este caso se introduce como valor por defecto las bases de datos que existen de ejemplo, explicadas en el PuntoBases de datos propias ; también incluye los parámetros citados en los puntos anteriores, *hoja="No definida"* y *create=FALSE* pertenecientes a la carga de hojas de cálculo así como *headerFun=TRUE* y *sepFun=* para cargar tanto CSV y FASTA.

3.2. Observación de datos anómalos

Para poder realizar la observación de datos anómalos es necesario utilizar varias librerías de R, para ello se utiliza las funciones *boxplot* y *plot*, las cuales están incluidas en la pa-

quetería estándar de R- También se utiliza la librería *TeachingDemos*, la cual en la función faces2 incluye las caras de Chernoff, en R al descargar los paquetes se descarga el código de los mismos por lo que se puede manipular y agregar funcionalidades a las funciones, como se ve más adelante.

En uno de los puntos que se centra este trabajo es en la representación gráfica de los datos anómalos, ya es realmente útil la localización de outliers desde una visión global de los datos.

En este caso se quiere analizar si los valores anómalos pueden ser fácilmente identificables para una persona que no tenga por qué tener conocimientos de matemáticas. Esto es permitir al usuario acotar de manera considerable los outliers, además de eso daremos un aspecto visual a esta localización de los outliers.

En algunos casos, los datos que estén en esos valores serán perfectamente aceptados, por ejemplo dada una tabla de repostajes como es el caso de la tabla de diesel, se aceptarán los outliers de los litros repostados, ya que en ciertas ocasiones se puede repostar una cantidad pequeña de diesel por un lado y en otras circunstancias una cantidad elevada para, por ejemplo, llenar el depósito. En cambio hay situaciones en las que los outliers representan valores claramente erróneos, por ejemplo, en el caso de la relación de la Fórmula 3.1,donde se muestra la fórmula del consumo (c), litros (l) consumidos cada 100 Kilómetros (km), es lógico pensar que la presencia de outliers puede indicar un valor erróneo, sobre todo aquellos más extremos. Por ello es necesario investigar las posibles relaciones entre las distintas variables.

$$Consumo = fraclitrosKM/100 \tag{3.1}$$

Para localizar los outliers se ha decidido hacer uso de los boxplots, ya que permite visualizar de manera sencilla la presencia o no de los mismos.

Al usar los outliers, como se aprecia en la Figura 3.1, se puede apreciar que implican un problema y es que la visualización de los mismos es demasiado genérica, ya que no aporta datos sobre las muestras exactas que contiene los outliers, por lo tanto si hubiera una muestra completamente errónea, que esté en una magnitud distinta del resto por ejemplo,

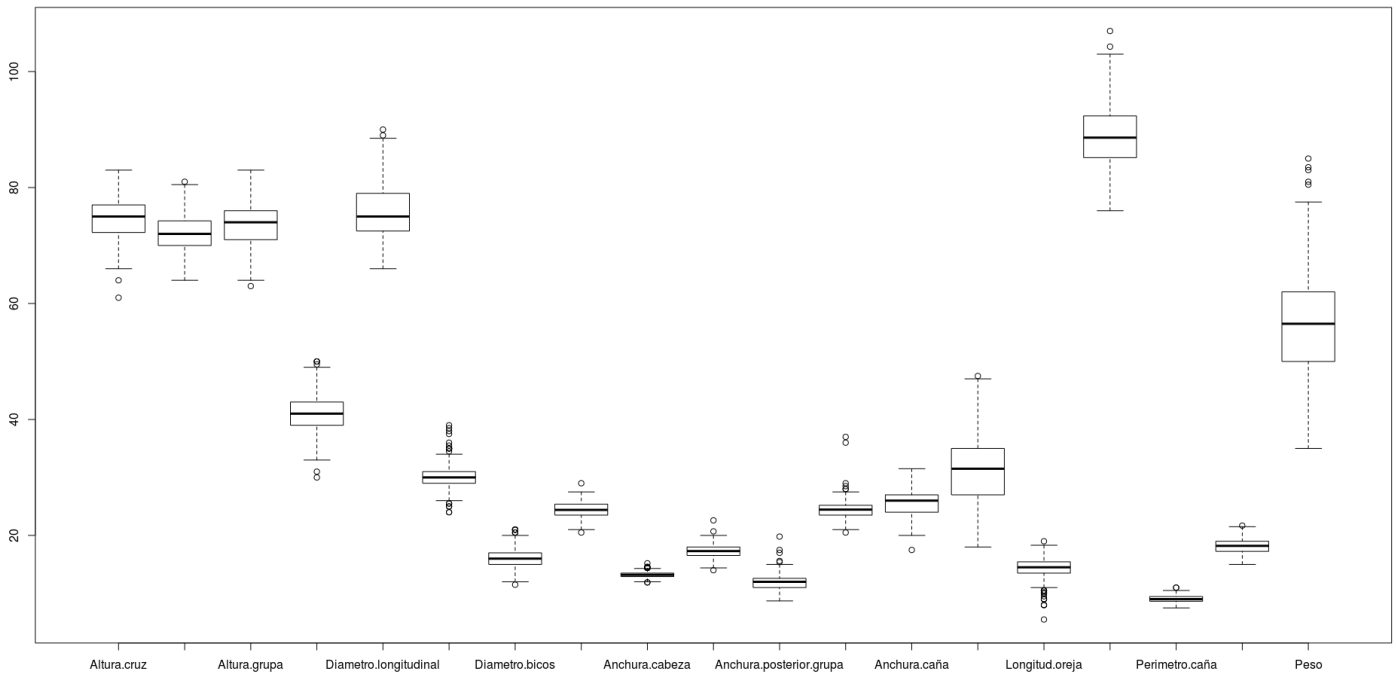


Figura 3.1: *Ejemplo de un boxplot*

no se podría localizar de una manera sencilla utilizando los boxplots.

Por lo que en este trabajo se les considera una herramienta interesante, pero que para los objetivos del mismo no es suficientemente específica. Hay que tener en cuenta que se habla siempre desde el punto de vista de la visualización, ya que realmente se pueden recoger los outliers y mostrarlos tal cual, pero quizás enseñar los datos así a una persona no especializada, como puede ser un ganadero, para el caso de las cabras podría no ser suficientemente esclarecedor.

Por ello se utilizan las caras de Chernoff, que ya están desarrolladas en R, en el paquete *faces*. El objetivo es aprovechar la visión global de los datos aportada por las caras de Chernoff para poder mostrar un aspecto igual de global para los outliers.

Se va a probar si las caras de Chernoff permiten localizar outliers de una manera rotundamente clara, ya que si deja lugar a dudas no tendría fiabilidad y por lo tanto no resultaría útil. Se puede apreciar un ejemplo del uso de las caras de Chernoff en la Figura 3.2 dicho

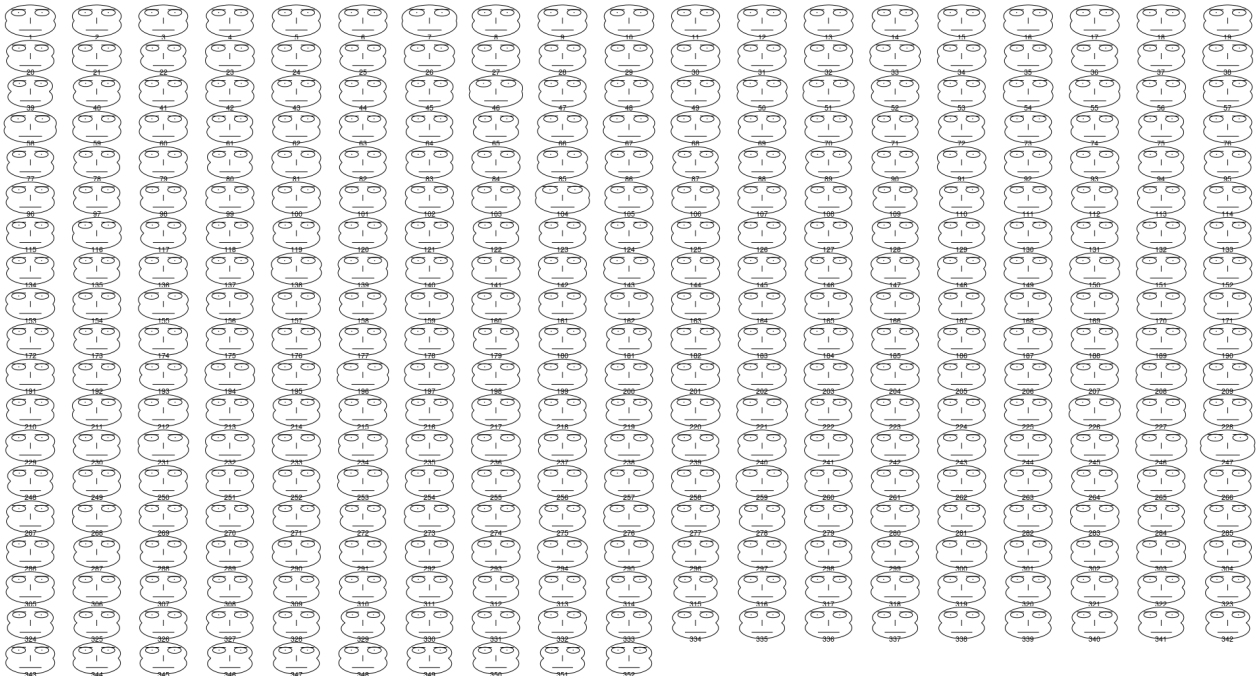


Figura 3.2: *Ejemplo de las caras de Chernoff*

ejemplo es un conjunto de 333 muestras con una única variable y un total de 17 outilers. Al intentar localizar los outliers se ve que la búsqueda de outliers directamente sobre las caras es poco fiable y por lo tanto que hay que buscar otra alternativa.

Además, en el caso de tener un menor número de datos y un mayor número de variables tampoco es del todo fiable. Ya que como se puede ver en la Tabla 2.4, en algunos casos las variables se *difuminan* entre sí, como sucede con las variables 9 y 10, esto es que no es fácilmente apreciable saber dónde comienza una u otra, ya que tienen relación entre ellas, a la hora de realizar los trazos de la cara al menos, siendo trazas simples si varias variables hacen referencia a los ojos no es fácil determinar dónde termina una y comienza otra. Por eso el sistema propuesto para realizarlo es utilizar un código de colores para recalcar aquellos trazos de las caras donde se localice un dato anómalo.

Así en circunstancias normales se tiene que los trazos de las caras serán de color negro, mientras que si uno fuera un *outlier* se recalcaría de un color en función del número de variable que sea, con esto se pretende que sea perfectamente claro saber la variable que se

está remarcando, aunque la zona de trazado coincida con otras.

3.3. Observación de datos ausentes

En este caso lo que se pretende es estimar datos ausentes, para lo cual se buscarán correlaciones entre las variables, para ello se utiliza la librería *corrgram*, la cual permite correlaciones entre varias variables, mostrando siempre correlaciones de una a una, esto es correlaciones directas entre dos variables.

Para poder estimar los datos que no están presentes o los datos anómalos es necesario localizarlos primero. Para ello se tendrán que realizar dos acciones, la primera de ellas es localizar valores perdidos. Para ello se usará la función is.na(*var*), a la cual se puede introducir cualquier tipo de variable y devolverá si es un NA o no, devolviendo TRUE o FALSE respectivamente. Es importante destacar que en R no es viable la comparación directa *var*==**NA** que se podría hacer en otros lenguajes como es el caso de Java *var*==**null**, de ahí la importancia de la función. Además en el caso de introducir una variable de tamaño superior a 1 devolverá tantos valores *booleanos* como valores haya, lo cual es muy útil a la hora de buscar datos que no estén definidos en bases de datos.

3.3.1. Investigar relaciones

Para poder investigar las relaciones entre las variables es muy útil el uso de las correlaciones, explicadas en el Punto 2.2.1, ya que muestra la regresión lineal de dos variables así como el coeficiente de su correlación.

3.4. Sistema Experto

En el caso de los carburantes hay que centrarse en las estimaciones con ausencia parcial de datos, para ello lo primero que hay que hacer es tener unos datos base para saber si los datos estimados pueden ser considerados buenos, que estiman valores razonablemente cercanos a los valores reales.

En este caso se va a utilizar el conocimiento de un experto, para conseguir hallar los datos faltantes. Para ello hay que cumplir las dos capacidades de un sistema experto,la primera de ellas, que se creen reglas de una manera sencilla, de modo que sea más sencillo la realización de reglas basadas en hechos.
Y la segunda que dichas reglas estén creadas de tal modo que sea fácil modificarlas, ya que el razonamiento de un experto puede variar con el tiempo y se debe asegurar que es viable un cambio en ese sentido.

Se va a partir de una premisa y es que siempre que se introduce un dato se conoce la fecha. Por lo tanto para estimar un dato se puede utilizar el campo fecha, que siempre tiene valor, esto es importante, ya que si hubiera que estimar fechas se podrían arrastrar fallos y que los valores obtenidos no sean buenos. La suposición anterior está refutada observando las bases de datos ya que estas contienen todas las fechas, por lo que se puede afirmar que la suposición es correcta.

En este caso se desea poder hallar un dato ausente en cualquiera de las variables menos en fecha (euros, litros y km.totales) haciendo uso del resto de datos de los que se disponen. Para ello es posible que haya que ayudarse de nuevas variables, como es el caso de los kilómetros recorridos con el repostaje actual (km), que se calculan restando a los kilómetros totales (KM.totales) de la siguiente fila los de la actual. Es importante destacar que en el caso de que alguno de los dos datos se haya perdido el resultado de la resta será NA, por lo que no aporta datos erróneos.

Continuando con las formas de estimar un dato en estas bases de datos, hay varias formas de poder realizar la estimación de los mismos, ya que los litros repostados se pueden hallar tanto en función de los *euros*, como de los km como de la $fecha$ o mediante combinaciones de las tres, si bien es cierto que una de las formas tiene que ser la que aporte mejores resultados. Pero no solo eso, sino que los resultados pueden variar en función del orden en el que se estiman, ya calculándolo por columnas o por filas, en este caso el cálculo por columnas se denomina *por variables* y el cálculo por filas se denomina *cronológico* ya que

los registros tienen están introducidos siguiendo el orden cronológico. Antes de determinar el orden de estimación de los datos es importante que se traten ciertas relaciones entre las variables, ya que algunas de ellas están estrechamente relacionadas.

Relación Euros/Litro

Se sabe que hay relación entre los euros y los litros, por lo que hay que estudiar dicha relación y conseguir sacar el mayor partido a la relación a la hora de realizar el sistema experto, para ello es necesario conseguir un registro externo con el precio del litro y ver si realmente los valores de las bases de datos se ajustan a los del registro.

Por supuesto debe usarse un registro tanto de diesel como de gasolina 95 y lo más cerca de la zona donde se ha repostado que sea posible.

Relación kilómetros/Litros

La nueva variable, km, se crea pensando que pueda tener alguna relación con las otras variables o al menos facilitar los cálculos a la hora de conseguir estimar alguna variable. Éstas variables son los litros y los euros, en un principio se comprobará la de los litros porque puede ser la que tenga una mayor correlación, ya que se conoce que existe una relación real entre los litros del depósito y los kilómetros que se pueden realizar con dichos litros, incluso dicha relación se muestra continuamente en los vehículos que tienen indicadores electrónicos a modo de referencia para el usuario, donde muestran los kilómetros restantes del depósito en función de los litros y el consumo.

Estimación de las variables

A continuación se explica como se va a aplicar el conocimiento del experto para estimar las variables, se explica primero las formas de estimar cada variable, y a continuación se explican varios métodos que calculan las variables en distinto orden, de entre los cuales hay que determinar cuales es el mejor.

Litros

La forma más sencilla y fiable de estimar los litros es mediante los euros (e) de dicha fila, calculando previamente la relación euros/litro del mes correspondiente (d_i). Una vez se

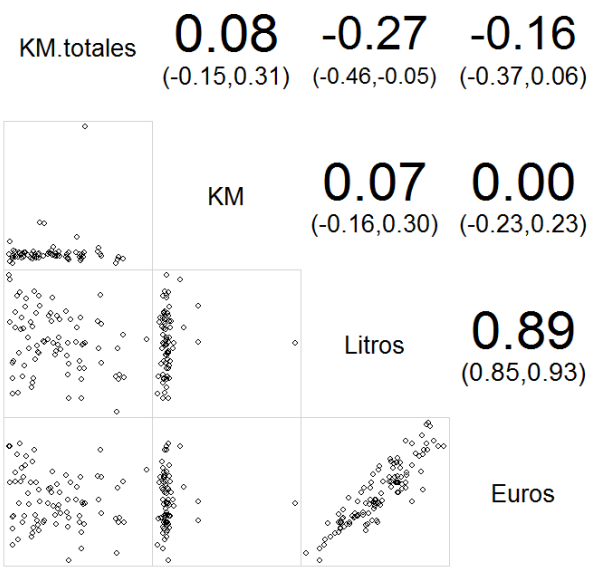


Figura 3.3: *Correlación entre las variables de gasolina incluyendo los km*

tenga la relación se estima cuanto cuesta en ese período y se calculan los litros (l), tal como muestra la Fórmula 3.2.

$$l = \frac{e}{d_i} \tag{3.2}$$

$$l = \frac{km}{100} * \bar{c} \tag{3.3}$$

En el caso de no tener tampoco los euros se ha de utilizar el consumo medio tal como muestra la Fórmula 3.3 haciendo uso de la variable km así como del consumo medio (c).

Finalmente si solamente se tuviera la fecha, en el caso de no tener ni Litros ni Euros, ni KM.totales, se calcularán los kilómetros totales para después calcular los Litros.

Euros

En el caso de tener los litros se calcularán mediante el precio del litro, al igual que en el caso anterior pero despejando los euros (e) en este caso, en el caso de no tener los litros se calcularán previamente, ya que se considera que las relaciones entre los euros y los kilómetros totales, parciales o la fecha no son suficientemente buenas.

$$e = l * d_i \tag{3.4}$$

Km.totales

La primera alternativa para calcular los KM.totales de una fila concreta (kt_i), es usando los kilómetros recorridos hasta el momento (kt_{i-1}), los litros (l) y el consumo medio (c), hallando los kilómetros recorridos con dichos litros repostados, tal como muestra la Fórmula 3.5. En el caso de que no estén disponibles se utilizará la Fórmula 3.6 que usa la fecha de dicha fila (f_i), así como la fecha de la fila anterior (f_{i-1}), haciendo un cálculo aproximado de los kilómetros recorridos diariamente, de media ($\frac{\sum_{j=1}^{i} k_j}{f_i - f_1}$) y multiplicándolo por los días transcurridos entre el registro anterior (f_{i-1}) y este (f_i). Esto se basa en que en muchas circunstancias un conductor va a realizar un número similar de kilómetros con el coche, al menos aproximado, ya que en circunstancias normales, un individuo que tenga coche y lo use con relativa frecuencia es porque va a trabajar en el y por lo tanto es fácil pensar que los kilómetros realizados dos días distintos pueden ser similares.

Si bien es cierto que dicho razonamiento puede ser perfectamente erróneo, ya que si el usuario es taxista o comercial no tiene por qué haber ninguna relación entre los kilómetros realizados y los días transcurridos, pero es un razonamiento del experto que como se ha explicado anteriormente puede variar con el tiempo y perfeccionarse.

$$kt_i = kt_{i-1} + \frac{l * 100}{\bar{c}} \tag{3.5}$$

$$kt_i = \frac{\sum_{j=1}^{i} k_j}{f_i - f_1} * (f_i - f_{i-1}) \tag{3.6}$$

Orden en que se estiman los datos

En algunos casos el orden en que se estiman los datos puede ser determinante en la calidad de un estimador, por ello en este trabajo se aboga por utilizar varios métodos y determinar cuál de ellos es el que ofrece mejores resultados.

Cronológicamente

Esta forma se basa en que todos los datos anteriores están calculados, por lo que se pueden utilizar esos datos anteriores para calcular el siguiente NA.

A la hora de evaluar un registro puede ir desde no tener ningún dato ausente, fila completa, hasta un máximo de tres datos ausentes, fila solo con fecha, a continuación se desglosará el proceder en función de dicha información, del número de datos ausentes.
Como es lógico, en el caso de no haber datos ausentes, no hay que realizar ninguna estimación.

En el caso de haber un único dato ausente, se puede aplicar cualquiera de las fórmulas citadas anteriormente, teniendo en cuenta que algunas fórmulas van a ofrecer mejores prestaciones, lo cual hay que demostrar. Las Fórmulas usadas son 3.2 y 3.3 para estimar los litros, la Fórmula 3.4 para estimar los euros y las Fórmulas 3.5 y 3.6 para estimar los kilómetros;

Cuando el número de datos ausentes sea dos, que es el mismo caso en que solamente se contenga un dato, más allá de la fecha. El objetivo en este caso es conseguir hallar los litros,en el caso de que no se tuvieran, ya que mediante los litros se pueden hallar las otras dos variables, mediante la Fórmula 3.4 se hallan los euros y mediante la Fórmula 3.5 se estiman los kilómetros totales. Para hallar los litros hay que usar las Fórmulas 3.2 y 3.3, en función de los datos disponibles.

En el caso más extremo, en el que solamente se contenga la fecha de un registro, se deben hallar primero los kilómetros totales mediante la Fórmula 3.6 y posteriormente operar como en el caso anterior en que solamente faltaban dos datos.

Con este método se tiene un problema y se da en el caso de que en el primer registro exista pérdida de datos, *NAs*, obviamente no se puede hacer uso de los datos precedentes para estimar esta fila, ya que no existen, por lo cual hay que modificar la estrategia para este primer caso.

En el caso de que haya ausencia de un único dato, se puede hallar haciendo uso de las funciones anteriormente citadas, pero ninguna de las dos funciones usadas para hallar los kilómetros totales es válida, por lo que hay que realizar una ampliación en este sentido. Para ello se va a partir de la siguiente suposición, al menos existen dos datos en la columna de

kilómetros totales, dicha suposición es una condición nimia en una base de datos de más de 100 elementos ya que se solicita que contenga algo menos de un 2 % de los datos para esa variable.

Partiendo de eso la Fórmula para estimar los kilómetros es la 3.7[2], donde los kilómetros totales del primer registro (kt_1) se hallan en función del coeficiente de kilómetros por día ($\frac{k_q - k_p}{f_q - f_p}$) usando el primer registro que contenga datos válidos en los kilómetros (p), y el segundo (q) y multiplicando por la resta entre la fecha de la posición p (f_p) y la primera fecha (f_1).

$$kt_1 = k_p - \frac{k_q - k_p}{f_q - f_p} * (f_p - f_1) \tag{3.7}$$

Una vez hallado esto se pueden estimar los litros y los euros haciendo uso de estos kilómetros estimados.

Por variables

Otro posible sistema es localizar los datos por variables, esto significa que no importa ya el orden cronológico de los datos, sino que los datos se completan de variable en variable, esto es que al ir a completar la variable i y registro j sabemos que todos los datos de las variables ya completadas, están disponibles para continuar estimando valores, así como los valores anteriores a j de la variable i.

El problema aquí se da al determinar el orden en el que se estimarán las variables. Tomando como referencia el caso base del punto anterior, la primera variable que se hallará serán los kilómetros, ya que haya los datos que haya es posible conseguir hallar esta variable, ya que como mínimo contaremos con los datos necesarios para poder usar la Fórmula 3.7.

Hay que observar qué propuesta de las anteriores ofrece mejores resultados y a posteriori decidir cual de entre las fórmulas para obtener los kilómetros es la mejor.

Una vez se tienen los kilómetros se hallan los litros, ya que existe una relación directa entre ellos, la mostrada en la Fórmula 3.3. Por último se estiman los valores faltantes de la

[2]Cabe destacar que en R los datos comienzan desde el número 1, al contrario que muchos otros lenguajes de programación, esto es porque el registro 0 está reservado para aportar información del contenido como el tipo de datos que contiene o las dimensiones del mismo

variable euros.

Este método es importante en el caso del diesel, que no es posible calcular el consumo medio usando datos reales del coche, como se ha citado anteriormente, por lo tanto si se calculan los kilómetros en primera instancia es posible calcular una estimación del consumo medio, para así poder estimar a continuación los litros y euros.

Uno de los apartados interesantes del sistema experto es conocer qué camino de los citados anteriormente es el que más se ajusta a la base de datos. Ya que el hecho de estimar antes o después un dato puede ser determinante de cara a que el resto de datos estimados se ajuste más a los datos reales. Además como ya se ha dicho en este trabajo, para comparar dos estimadores no es suficiente con hacer una comparación puntual sino realizar varias pruebas para que sea más fiable. Por ello se ha desarrollado el comparador de estimadores.

3.5. Completar o modificar datos

Una vez se han localizado los datos que se desean completar o modificar, ya sean datos perdidos o aquellos que se consideren incorrectos respectivamente, el siguiente paso es encontrar con qué dato sustituirlo. Para introducir este nuevo dato se puede utilizar cualquiera de las técnicas explicadas en el Punto 2.2.

En este trabajo se van a completar los datos de dos maneras principales, la primera de ellas es buscando relaciones entre las variables, correlaciones, en el caso de encontrarlas se tendrán que buscar los parámetros correctos y ver la calidad de las correlaciones. Además basándose en los estudios realizados por [García Lara et al., 2009], donde demuestran la relación directa entre el perímetro torácico y el peso en vacunos, con un coeficiente de correlación bastante elevado, del 95,35 %, se pretende comprobar si dicha correlación tiene también buenos resultados en caprinos.

3.6. Comprobar completados

Para comprobar que los datos completados son *aceptables* se puede realizar de varias maneras. Una de ellas es comparar el algoritmo con un algoritmo base, como puede ser hallar la media de los valores conocidos. Esta opción es muy interesante cuando se busca tener una referencia rápida y como punto de partida a la hora de empezar a sacar ciertas conclusiones sobre el algoritmo desarrollado. Si bien es cierto que puede ser una mala alternativa en el caso de no tener muchos valores de la variable que se desea estimar. Esto es debido a que en el caso de que el número de valores conocidos no sea suficientemente representativo la media o prácticamente cualquier valor que se pueda hallar de dichos datos pueden no ser útiles, por lo tanto tomarlas como referencia puede llevar a conclusiones erróneas.

Por lo tanto, es interesante comparar dos algoritmos distintos para completar los datos. Este método se aplica muy comúnmente ya que es la forma más directa de comprobar dos algoritmos entre sí. El problema que se tiene en este caso es que se necesitan conocer todos los datos de la variable que deseamos estimar, que son las circunstancias normales, ya que en caso de no conocer todos los datos no se va a poder conocer la bondad de las estimaciones. En este trabajo se ha preferido centrarse en el uso del ECM, para comparar las diversas relaciones, así como una relación entre la desviación máxima y el ECM.

Como se verá en el Punto 5.2, las correlaciones no son suficientes en el caso de las bases de datos de los carburantes, gasolina y diesel, por lo que hay que buscar otras alternativas.

3.7. Comparador de estimadores

Dado que se pretende que el trabajo sea lo más genérico posible, para que así pueda ser utilizado en un mayor número de sectores de la bioinformática, se ha creído oportuno centrarse en la realización de un *evaluador de estimación.*

Se ha denominado *evaluador de estimación* a un programa que permite evaluar una estimación, como su propio nombre indica, para ello se puede seleccionar un conjunto de técnicas, de las anteriormente citadas, además de permitir al usuario la introducción de más

de una técnica para así compararlas y que en función de los resultados obtenidos llegue a sus conclusiones pertinentes.

Motivación para realizar el *evaluador de estimación*

Una de las razones por las que se ha creado dicho *evaluador de estimación* es considerar que es una buena manera de brindar al usuario la opción de introducir cierto conocimiento experto y así poder evaluarlo con otras estimaciones. Por lo tanto, y como es obvio, el usuario introducirá su propia tabla sobre la que quiere realizar la estimación, ya que en caso contrario no tendría sentido el *comparador de estimación*.

En dicho *evaluador de estimación* no se incluyen técnicas avanzadas, ya que lo que se pone a servicio del usuario es la opción de situar a su función de estimación en diversas situaciones para que así conozca su rendimiento en función de diversos factores o comparar dos estimaciones en función de dichos factores.

Para que el usuario pueda introducir sus funciones de estimación, estas deben cumplir ciertas condiciones, para así asegurar el correcto funcionamiento del estimador. Estas condiciones hacen referencia a la cabecera de las funciones así como al valor devuelto de las mismas. El estimador introduce en las funciones dadas por el usuario, o las por defecto, unos parámetros concretos para permitir el funcionamiento genérico del mismo, dichos parámetros son *tablaMod* que hace referencia a la tabla sobre la que hay que realizar las estimaciones y columnas, que indica las columnas que hay que analizar en busca de *NAs* para sustituirlos por valores reales. Así el valor devuelto por la función será la totalidad de la tabla habiéndole aplicado las estimaciones pertinentes.

Rendimiento de las estimaciones

Los factores de los que se trata son la insesgabilidad, la eficiencia, la consistencia y la suficiencia explicadas con detalle en el Punto 2.5, así como el tiempo de cómputo, que en ciertos casos puede ser esencial, hay situaciones en las que se busca una rápida reacción del sistema, si bien es cierto que en la mayoría de casos a la hora de estimar un dato el tiempo

de cómputo va a quedar relegado a un segundo plano, ya que se busca la mejor estimación posible.

Para comparar los estimadores en base a estos parámetros, y de cara a que el usuario pueda introducir las funciones de verificación que crea convenientes, es necesario que al igual que con las funciones de estimación, las de verificación cumplan las condiciones de ajustarse a unos parámetros y un valor de salida de la función. Los parámetros en este caso serán, *tablaOriginal* que contiene los datos originales para que puedan ser verificados los datos estimados contenidos en la *tablaEstimada*. Para poder conocer los datos sobre los que hay que realizar las verificaciones se deben introducir los elementos que hay que evaluar en el parámetro *comprobacionNA*, que contiene todos los valores que han sido convertidos en *NA* y por último la columna sobre la que se realiza la estimación en la variable *columnas*.

Se recomienda al usuario que la devolución de estas funciones sea un dato único ya que en el caso de ser un vector o una matriz puede perder legibilidad y no ser suficientemente exclarecedores, sobre todo en el caso de que se apliquen varias funciones, ya que se devuelven los resultados como un único vector. Por esto se decide utilizar el ECM como medida de evaluación entre estimadores, ya que representa un compromiso entre la insesgabilidad y la eficiencia, como se explica en el Punto 2.5.

Por otro lado en el caso de la consistencia se considera que, por el momento, debe depender del usuario ya que en muchos casos el usuario puede querer hacer unas pruebas de consistencia muy específicas, como puede ser realizar pruebas con múltiples cantidad de datos, para ver como se van escalando los resultados y si se ajusta a lo que realmente esperan del estimador.

En el caso de la suficiencia tampoco se evalúa, ya que resulta imposible saber si ha conseguido extraer toda la información posible de los datos o no.

Para asegurar el correcto funcionamiento del *comparador de estimación* se realiza un tratamiento previo de la tabla introducida por el usuario. El tratamiento consiste en la eliminación de los valores *NA* de dicha tabla, con ello se pretende que sea posible tener unos

valores precisos de los resultados. Pero este proceso puede hacer caer, como se ha dicho anteriormente, en conclusiones erróneas, por asegurar que la estimación pueda ser correcta basándose en un conjunto no representativo de los datos.

Por ello se ha incluido una alternativa para aportar precisión a la estimación. Permitir al usuario introducir un conjunto de datos completo donde se seguirá realizando la eliminación de *NAs*, pero que en el caso de ser completo no debería de eliminar ninguno de los registros de la base de datos.

Pérdida de datos controlada en el *comparador de estimación*

Una vez se tiene la base de datos sin ningún valor NA, se permite al usuario realizar una pérdida controlada de datos, esto es, que se elimine una cierta cantidad de datos, en %, para que la base de datos de entrada en el estimador tenga ausencia de datos *real*, aunque haya sido forzada.

La diferencia con respecto a la ausencia de datos real, es que en este caso se va a tener una manera de comparar los resultados obtenidos mediante el uso del estimador, ya que se mantiene la base de datos original, para que sirva de comparación con la base de datos que contiene estimaciones.

Para la pérdida de datos se van a utilizar las dos alternativas comentadas en el Punto 2.4, pudiendo elegir el propio usuario entre ambas. En ambos casos se introduce la pérdida de datos porcentual citada anteriormente, en el caso de la pérdida de datos MCAR es totalmente aleatoria, por lo que si se quiere que los datos contengan al menos un *NA* por fila, para que así todos los registros puedan tener pérdidas, habrá que introducir un valor superior a $1/numVariables$.

Aunque si se desea asegurar la presencia de *NA* en cada uno de los registros lo idóneo es utilizar la pérdida de datos MCAR, ya que dicha pérdida tiene como propiedad, que si no se ha realizado una pérdida de datos previa en dicha fila las posibilidades de pérdida aumentan por cada variable que no sufre la conversión a *NA*, mientras que si ya ha realizado

alguna conversión vuelve al valor inicial. En este caso introduciendo un menor porcentaje de pérdida se puede conseguir tener una pérdida en la mayoría de las filas.
Además de este modo se reducen las posibilidades de que en caso de tener una pérdida en una fila pueda haber un mayor número de ellas, lo cual no interesa en absoluto, ya que cuantos más datos haya perdidos en una fila puede ser que menos alternativas haya para completar dicho dato en el caso de existir correlaciones entre ellos.

Selección de las variables a estimar

Pero un usuario perfectamente puede no querer estimar todos los datos, si no estimar un número determinado de variables o una única, de hecho lo más común es que se desarrolle un buen estimador para cada variable y luego se prueben en común ya que en el caso de querer estimarlos todos a la vez puede darse la situación de que las estimaciones de uno entorpezcan a las del otro.
Por ello el *comparador de estimación* permite al usuario seleccionar las variables, columnas, a las que desea aplicar la pérdida de datos, usando el parámetro *col* que tiene como valor por defecto 0, que es la aplicación a todas las variables de la base de datos.

Además, combinando lo citado hasta ahora, en el caso de que el usuario introduzca el mismo número de estimaciones y de columnas a estimar se interpreta que el usuario quiere que se aplique un método a a cada una de ellas. Por lo tanto el método de estimación i está asociado a la variable a estimar i, donde i es un número menor que el número total de estimaciones introducidas por el usuario y también menor que el número de variables contenidas en la base de datos.
En caso contrario si desea aplicar cada técnica de estimación a cada variable, debe poner a true la variable *compara*. Esto hace que si hubiera i variables y j métodos, se realizarán un número de $i * j$ combinaciones devolviendo un número $i * j$ resultados uno por cada una de las combinaciones, mientras que si no se activara y $i \neq j$, devolvería el error pertinente ya que no se pueden aplicar j estimaciones a i variables .

Aumentando la precisión de los resultados

Pero el hecho de aplicar una técnica de estimación a un conjunto de datos y devuelva un valor con buen resultado, no implica que dicha técnica de estimación sea buena. Puede darse el caso de que teniendo dos técnicas de estimación similares A y B, donde la técnica A sea ligeramente mejor que la técnica B, la técnica B obtenga un resultado puntual mejor que la técnica A, lo cual podría llevar a conclusiones erróneas.

Por ello, y para finalizar, se ha incluido una variable más, que es la variable *iteraciones*, la cual determina el número de iteraciones que se realiza. Con ello se pretende eliminar los resultados puntualmente mejores de algunas estimaciones, ya que realizando un número alto de iteraciones permite generalizar suficiente como para considerar que un método es mejor que otro.

También es muy útil para conseguir un tiempo de computación más fiable de los métodos de estimación. Ya que teniendo en cuenta una única iteración los tiempos de computación son más vulnerables a errores por redondeo a la hora de que el sistema devuelva los tiempos, o incluso que en algunas circunstancias un método puede tardar más o menos en función de los datos ausentes, tanto del número como de la naturaleza, en cambio aumentando el número de iteraciones se consigue que los resultados se acerquen más a la realidad. También se le muestra al usuario el tiempo máximo y el mínimo porque en función de las circunstancias puede ser imprescindible para el usuario.

3.8. Guardar datos

Para poder almacenar los datos en archivos externos a R y así que las modificaciones perduren más allá de la sesión actual, es necesario utilizar ciertas librerías, la primera de ellas *tools*, la cual al igual que en el caso de la carga permite conocer la extensión del archivo de destino. Para almacenar los datos en formato *.xls* y *.xlsx* se utiliza la librería *dataframes2xls* y para almacenarlos en CSV o FASTA al igual que en el caso de la carga

de datos, la librería *foreign*. Guardar datos funciona de forma similar a cargar datos, pero en este caso el usuario debe seleccionar los datos que se desean guardar, el archivo al que se quieren exportar esos datos y el formato o extensión en que se quieren guardar. En el caso de querer guardarlo como archivo .xls o .xlsx no es necesario especificar el parámetro *formato* ya que dicho parámetro es solamente para conocer si se quieren guardar los datos en formato FASTA o CSV. En el caso del formato FASTA o CSV solamente hay que elegir el formato en el que se desea guardar.

En el siguiente capítulo se va a ver la aplicación de lo citado en el Capítulo 3, mostrando los procedimientos para conseguirlo y los resultados obtenidos de los mismos.

Capítulo 4

Bases de datos propias y carga de archivos en R

Para poder aplicar lo citado en el Capítulo 3, es necesario hacer uso de algún conjunto de datos sobre el que probarlo. Además en función de la naturaleza de los datos es necesario aplicar un análisis exploratorio u otro, no tiene mucho sentido utilizar las caras de Chernoff para buscar outliers en una tabla con dos variables, ya que está orientado a un mayor número de variables y quizás los boxplots sean más indicados en ese caso.

Para aplicar dichos análisis se han utilizado unas bases de datos propias, las cuales se explican a continuación.

4.1. Bases de datos propias

A continuación se realiza una descripción de las bases de datos contenidas en el proyecto. Pese a tener tres bases de datos dos de ellas tienen una naturaleza prácticamente idéntica por lo que se describen como si fueran una, para no redundar en la información.

4.1.1. Base de datos de las cabras de guadarrama

La base de datos de cabras de guadarrama es una base de datos recopilada y cedida por la Facultad de Veterinaria de la Universidad Complutense de Madrid. En esta base de datos hay un total de 531 ejemplares de cabras de Guadarrama con un total de 21 variables

cada registro. Las cabras de Guadarrama, también llamadas guadarrameña, es una raza de cabra natural de la Sierra de Guadarrama y es una de las Razas Autóctonas de Protección Espacial que está en peligro de extinción según el Catálogo Oficial de Razas del ganado en España[1].

Estas cabras tienen ciertas cualidades físicas que varían con el sexo, en el caso de los machos su peso está entre los 70 y 80 Kilogramos y su alzada de cruz está entre los 80 a 84 centímeros. Mientras que en el caso de las hembras tienen un peso entre 50 y 55 Kilogramos y una alzada de cruz de 72 a 75 centímetros.

Si bien es cierto que no todas se ajustan a esos parámetros de manera exacta, de hecho en el caso de la base de datos de las cabras, donde todas son cabras de Guadarrama, muy pocas cumplen con ambas características, un 13.73 % en el caso de los machos y un 10.42 % en el caso de las hembras. Esto puede ser debido a que no todas las cabras son cabras maduras, si no que muchas de ellas son cabras de menos de 4 años, que es considerada la edad madura de la cabra.

Es importante destacar una cualidad de esta base de datos, la proporción entre hembras y machos está totalmente desequilibrada, teniendo 480 hembras de las 531 muestras, un 90,395 % ,mientras que tenemos 51 machos de las 531 muestras, un 9,605 %. Pero este suceso es algo perfectamente normal en un rebaño mixto.

Los rebaños mixtos son aquellos que tienen dos propósitos, que dedican sus cabras tanto al sector cárnico, aportando carne al mercado, como al sector lácteo, en función del sexo varían las aportaciones de cada miembro. Las hembras tienen dos funciones principales, la primera de ellas es parir para asegurar la perpetuidad del rebaño y segunda, derivada de la primera, es aportar leche no solo para el cabrito, si no para venderla.

Pero los machos no pueden aportar esas funciones por razones evidentes, dedicándose al apareo así como a la producción de carne para vender en el mercado. Al venderse en el mercado se tiene que ajustar a los requisitos de éste y por lo tanto se venden como cabrito

[1] http://www.feagas.com/index.php/es/razas/especie-caprina/del-guadarrama

lechal que son las cabras con menos de 4 meses, por ello es por lo que no hay una gran cantidad de machos a ninguna edad. En la base de datos las edades de las cabras están divididas en tres grupos que son, andoscos que van desde los 2 años hasta los 3 años una vez superados entran en la edad trasandosca hasta que cumplen los 4 años que pasan a ser cerrados. Por ello la desproporcionalidad de machos y hembras, además un macho puede ser el mismo encargado de preñar a varias cabras en una misma temporada.

La base de datos de las cabras tiene un total de 21 variables, con un nombre suficientemente claro para que el usuario comprenda perfectamente de que variable se trata. En el caso de haber más de una palabra, estas se separarán mediante el uso de puntos (.). De esas 21 variables dos de ellas vienen representadas por un único carácter, estas variables son el sexo, que puede ser hembra (H) o macho (M) y la edad que como hemos dicho antes está dividida en tres bloques, andosco (A) entre 2 y 3 años; trasandosco (T) entre 3 y 4 años; cerrado(C) más de 4 años. Mientras que las otras 18 variables son medidas morfológicas expresadas mediante números *double*.

A continuación se van a explicar cada una de las características restantes, expresadas todas en centímetros. Además en la Figura 4.1 se aprecian algunas de las zonas de las cabras comprendidas en la base de datos.

Las alturas hacen referencia a la distancia, en centímetros, desde el suelo hasta un punto concreto, determinado por el resto del nombre de la variable. Las alturas que hay son *altura.cruz* donde la cruz es la primera vértebra torácica; *altura.dorso* que es la altura hasta el punto medio de la región del dorso; *altura.grupa* que es la altura hasta el punto medio de la distancia o la línea imaginaria entre las puntas del íleon; por úlitmo la *altura.hueco* es la altura hasta el esternón.

Los diámetros existentes son el *diametro.longitudinal*, distancia desde la base de la cola hasta el pecho del animal; el *diametro.dorso* medida circundante del dorso, desde la cruz al esternón; el *diametro.bicostal* la anchura del torax, medida desde una axila hasta la otra; y

Figura 4.1: *Señalización de las partes de una cabra.*

finalmente el *diametro.cabeza* la distancia desde los ollares hasta el atlas, primera vértebra de las cervicales.

Hay cuatro medidas longitudinales, dichas medidas son *longitud.cabeza* que es la longitud desde la apófisis rostral del hueso nasal hasta el temporal, desde donde acaba el cuello hasta donde termina la mandíbula inferior.; la *longitud.cuerno* es la medida desde el nacimiento del cuerno hasta el final del pitón; la *longitud.oreja* medida desde el punto en que la oreja sale de la cabeza hasta el final de la misma; y por último la *longitud.grupa* que abarca desde la cruz hasta la grupa.

Otras medidas son las anchuras, hay un total de cuatro medidas de anchura, el *ancho.cabeza* es la distancia entre ambas aberturas orbitales del canal supraorbitario, que son las zonas exteriores de los ojos, las dos zonas más alejadas de la cabeza; el *ancho.anterior.grupa* es la distancia interilíaca; el *ancho.posterior.grupa* es la distancia interisquiática; el *ancho.cana*

es el ancho de la caña, la longitud desde la rodilla hasta el menudillo.

Los perímetros almacenados en la base de datos son el *perimetro.toracico* que es la medida de la circunferencia del tórax; el *perimetro.cana*, medida de la circunferencia del hueso de la caña; *perimetro.corvejon* medida de la circunferencia de la zona del corvejón.

De los datos anteriormente citados, solamente 346 de las cabras no tienen ningún tipo de ausencia de datos, tienen los datos completos, mientras que 185 tienen ausencia de datos en alguna de las variables.

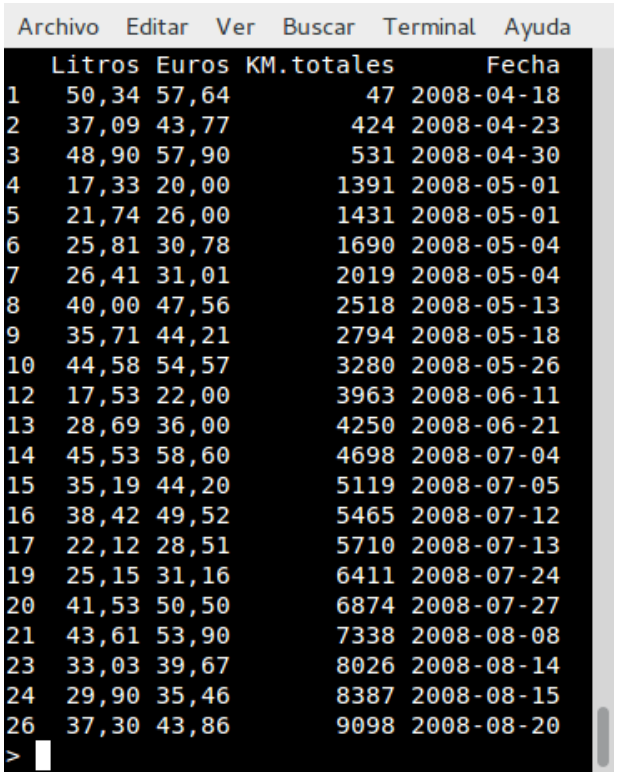
4.1.2. Bases de datos de repostajes

En este trabajo se dispone de otras dos bases de datos, dichas bases de datos son de naturaleza muy similar ya que en ambas se almacena información de repostajes, teniendo tres diferencias principales, la primera y más importante es que en la base de datos de *Diesel* se almacenan repostajes con carburante diesel, mientras que en la base de datos de *Gasolina*95 se almacenan repostajes de gasolina 95, como su propio nombre indica.

La segunda diferencia es que en el caso de la base de datos de repostajes diesel tiene un mayor número de variables ya que las bases de datos tienen en común 4 variables principales que son *Litros* que son los litros repostados en ese repostaje, *Euros* que indica los euros gastados en el repostaje, *Km.totales* muestra los kilómetros totales que lleva el vehículo hasta ese momento, mostados en el cuenta kilómetros[2] por último se tiene la *Fecha* que muestra la fecha a la que se ha realizado el repostaje.

Además de las varibles anteriormente citadas, que son todas las que tiene *Gasolina*95, se tiene que diesel contiene más variables, concretamente 3 más. La primera de ellas es *Tipo* que muestra el tipo de diesel repostado, de tal modo que en el caso de ser diesel e+ se marcará la columna con la cadena 'e+', mientras que en el caso de no serlo se quedará en blanco. La segunda de ellas es una columna de *Notas* que contiene información variada del

[2]Esta cifra corresponde a los kilómetros totales, no a los kilómetros reseteables que también se muestran



Archivo Editar Ver Buscar Terminal Ayuda

	Litros	Euros	KM.totales	Fecha
1	50,34	57,64	47	2008-04-18
2	37,09	43,77	424	2008-04-23
3	48,90	57,90	531	2008-04-30
4	17,33	20,00	1391	2008-05-01
5	21,74	26,00	1431	2008-05-01
6	25,81	30,78	1690	2008-05-04
7	26,41	31,01	2019	2008-05-04
8	40,00	47,56	2518	2008-05-13
9	35,71	44,21	2794	2008-05-18
10	44,58	54,57	3280	2008-05-26
12	17,53	22,00	3963	2008-06-11
13	28,69	36,00	4250	2008-06-21
14	45,53	58,60	4698	2008-07-04
15	35,19	44,20	5119	2008-07-05
16	38,42	49,52	5465	2008-07-12
17	22,12	28,51	5710	2008-07-13
19	25,15	31,16	6411	2008-07-24
20	41,53	50,50	6874	2008-07-27
21	43,61	53,90	7338	2008-08-08
23	33,03	39,67	8026	2008-08-14
24	29,90	35,46	8387	2008-08-15
26	37,30	43,86	9098	2008-08-20

>

Figura 4.2: *Muestra de la tabla de gasolina 95.*

gasto en ese momento, esta columna es importante, ya que en esta base de datos también se almacenan cambios de ruedas, líquido limpiaparabrisas, revisiones y en general cualquier gasto relativo al vehículo, además en el caso de que el pago no se realice de una manera corriente, como puede ser el uso de puntos de una tarjeta, se deberá indicar en esta columna y desglosar el pago en la columna *Desglose*.

La tercera y última diferencia es que en el caso del diesel se tienen 231 registros mientras que en el caso de la gasolina se tienen 109, esto es debido a que en la base de datos del diesel se tienen datos desde Enero de 2003 hasta Septiembre de 2012, mientras que en el caso de la tabla de gasolinas los datos almacenados son desde Abril de 2008 hasta Agosto de 2012.

En este trabajo no se van a utilizar las tres columnas extra de la base de datos de diesel, solamente para que el precio de los repostajes sea el idóneo, por ejemplo no se tendrá en cuenta los repostajes con e+, ya que adulterarían los resultados dando un precio más alto del carburante.

Así las bases de datos de gasolina y diesel tendrán la forma mostrada en la Figura 4.2, donde se muestra una sección de la tabla de gasolina 95, ambas tienen el mismo formato.

4.2. Carga de archivos

En este trabajo se ha creado un *script* (guión) de carga que permite cargar todos los archivos necesarios, así como las bibliotecas, para probar los resultados hallados en el trabajo, dicho guión se arranca escribiendo por consola de R el comando *source*("*carga.R*").

Una vez explicada la base de datos y sus variables se pueden explicar los análisis exploratorios realizados sobre esta base de datos. Para ello es necesario cargar los datos, por lo que hay que seguir los pasos explicados en el Capítulo 3.1. Como se puede ver en la Figura 4.3, es el caso en el que se carga la base de datos desde un archivo *.xls* y por lo tanto al comprobar que el archivo contiene más de una *hoja* pregunta al usuario qué hoja desea cargar. Finalmente almacena los datos en la variable asignada.

```
> tablaGasolina<-cargaArchivo()
Las hojas del archivo excel son las siguientes:
'Gasolina95'    'Diesel'        'Cabras'
Introduzca El nombre de la hoja correspondiente
Gasolina95
>
```

Figura 4.3: *Ejemplo de una carga de archivos de la base de datos 'Gasolina95'.*

Lo que se pretende con la base de datos de las cabras son dos cosas, la primera de ellas ver si es viable hallar outliers en las variables en función de irregularidades en las caras de Chernoff, de cara a que los ganaderos viéndolas puedan encontrar ciertos valores poco comunes, y en segundo lugar ver si se puede hallar o estimar el peso del animal en función de otras variables. Para ello también se carga la base de datos de las cabras de Guadarrama, se puede cargar como se ha mostrado anteriormente en el caso de la gasolina o también desde un archivo con formato csv, como sería el caso en el comando
$tablaCabras <- cargaArchivo("cabras.csv")$.

Capítulo 5

Aplicaciones a la base de datos de las cabras de Guadarrama

En el caso de las cabras de Guadarrama se va a realizar un tratamiento tanto de los datos con valores atípicos, como de los datos ausentes. En el caso de los valores atípicos se va a realizar un método de localización de estos valores en entornos multivariante, mediante el uso de las caras de Chernoff y aplicándole modificaciones a las mismas. Por otro lado en el caso del tratamiento de los posibles valores ausentes se van a buscar relaciones entre las variables mediante el uso de las correlaciones.

5.1. Localización de outliers

En el caso de la base de datos de las Cabras de Guadarrama el trabajo se centra en la localización de outliers. Esto es debido a que la base de datos es un entorno con un número elevado de variables y un número de muestras suficientemente grande, lo que hace que al realizar las caras de Chernoff se tenga un número interesante de caras con, a su vez, un número elevado de variables. De hecho no se pueden tener en cuenta todas las variables de la base de datos para realizar las caras, ya que hay un total de 21 variables mientras que en las caras de Chernoff el máximo de variables son 18, por lo que hay que organizar el rebaño previamente.

A la hora de organizar el rebaño, hay que tener en cuenta que las variables presentes en

la base de datos son medidas fisiológicas. Éste es un dato importante, y gracias a esto la organización del rebaño se convierte en algo trivial. La razón es que teniendo un total de 21 variables, hay que agruparlos en función de dos de ellas, como mínimo. Además las caras de Chernoff obligan a que los datos introducidos sean únicamente numéricos y en este caso las únicas variables no numéricas son *edad* y *sexo* representadas mediante un único carácter.

Hay que tener cuidado a la hora de buscar outliers en un conjunto de datos tan dependiente de la edad de los individuos, ya que como resulta obvio una cabra de 2 años no tendrá su cuerpo totalmente desarrollado y por lo tanto no tiene sentido buscar outliers a un conjunto de todas las hembras, ya que perfectamente las adultas pequeñas no se apreciarían como outliers por *camuflarse* entre las cabras de menor edad. Juntando todo esto se tiene que al agrupar las cabras en función de su edad y sexo hay seis conjuntos de 19 variables, todas ellas numéricas y se puede discernir entre una cabra adulta pequeña y una cabra de menor edad.

Antes de comprobar el número de elementos que contendrá cada conjunto hay que eliminar los *NAs* ya que los datos no definidos no pueden ser tenidos en cuenta en las caras de Chernoff.

Una vez eliminados los *NAs* el total de individuos son 346 mientras que los seis conjuntos quedan así: hay un total de 15 Machos Andoscos(MA), 9 Machos Trasandoscos(MT) y 18 Machos Cerrados(MC); también tenemos 24 Hembras Andoscas(HA), 54 Hembras Trasandoscas(HT) y finalmente 226 Hembras Cerradas(HC).

5.1.1. Búsqueda de trazos anómalos en las caras.

Se va a analizar si los outliers se pueden localizar de algún modo en las caras de Chernoff, buscando formas atípicas o patrones de ciertos comportamientos en el dibujo de las caras que puedan indicar la presencia de los mismos.

Para saber si los outliers son visibles de algún modo en las caras de Chernoff es necesario antes localizar los outliers. Para ello se va a utilizar la función *boxplot*, ya implementada en

R, que permite ver en un diagrama todos los boxplot que contenga una variable dada, esa variable puede ser una base de datos completa o una columna, variable, de dicha base de datos.

Para poder realizar las pruebas hay que seleccionar como conjunto de entrenamiento a un conjunto reducido, de los seis citados anteriormente, por lo que los conjuntos que interesan pueden ser MA, MT, MC, HA y HT ya que HC es demasiado elevado en número de individuos como para poder considerarse conjunto de entrenamiento. Estos 5 son los válidos, pero se puede apreciar que en el caso de los conjuntos de los machos no son elementos suficientes como para tomarlos como elementos de entrenamiento, por lo que por esa misma razón se usarán las HT por encima de las HA ya que son mayores en número. En el caso de que las HT no contengan ningún outlier se tomarán las HA como conjunto de entrenamiento.

En la Figura 5.1 se puede ver cómo las hembras trasandoscas tienen una serie de outliers. Estos outliers se reparten como se enuncia a continuación. El *diametro.longitudinal* contiene 1 outlier inferior; el *diametro.dorso* contiene 2 superiores y 1 inferior; La *anchura.cabeza* contiene 3 superiores; La *anchura.posterior.grupa* contiene 2 superiores y 1 inferior; la *longitud.grupa* contiene 3 superiores y 1 inferior; la *longitud.oreja* 1 inferior; el *perimetro.caña* 1 superior y el *peso* 1 superior. En total son 12 superiores y 5 inferiores.

El objetivo entonces es poder localizar dichos *outliers* en las caras de Chernoff, pero no solamente eso, si no que ha de ser de manera unívoca pudiendo reconocerlos y que además un valor normal no se interprete como un *outlier* ni un *outlier* como un valor normal.

En la Figura 5.2 se aprecia como en este caso al usar un conjunto suficientemente representativo existe un problema. El problema es que la información recibida es demasiado alta, desembocando en un exceso de información para el ojo humano humano. Esto dificulta en exceso la tarea de discernir los *outliers* haciendo que en muchos casos sea prácticamente imposible.

Por lo que hay que modificar el razonamiento, ahora se busca reducir el número de individuos representados, para ello no es suficiente con dividir los conjuntos mostrados en

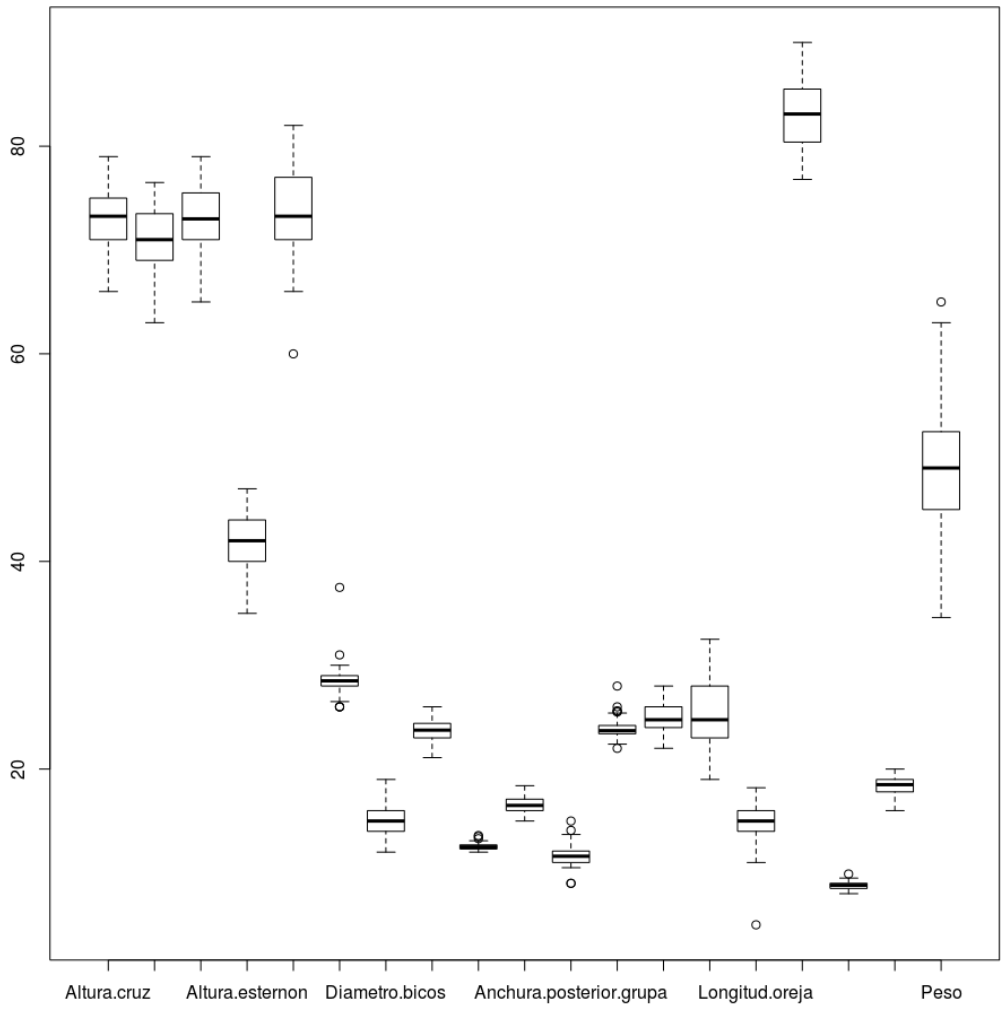


Figura 5.1: *Boxplot de las hembras trasandoscas*

dos subconjuntos, ya que dividir en dos subconjuntos puede implicar que ciertos valores normales se consideren atípicos porque no haya presentes más valores atípicos en esa muestra de individuos, por no ser representativa y por tanto las conclusiones sean erróneas. Por lo que es necesario seleccionar un conjunto de datos menor y por lo tanto ya no se busca que sea un conjunto suficientemente representativo sino uno con tamaño tal para poder llegar a conclusiones y comprobar si son escalables aplicándolo a otros mayores.

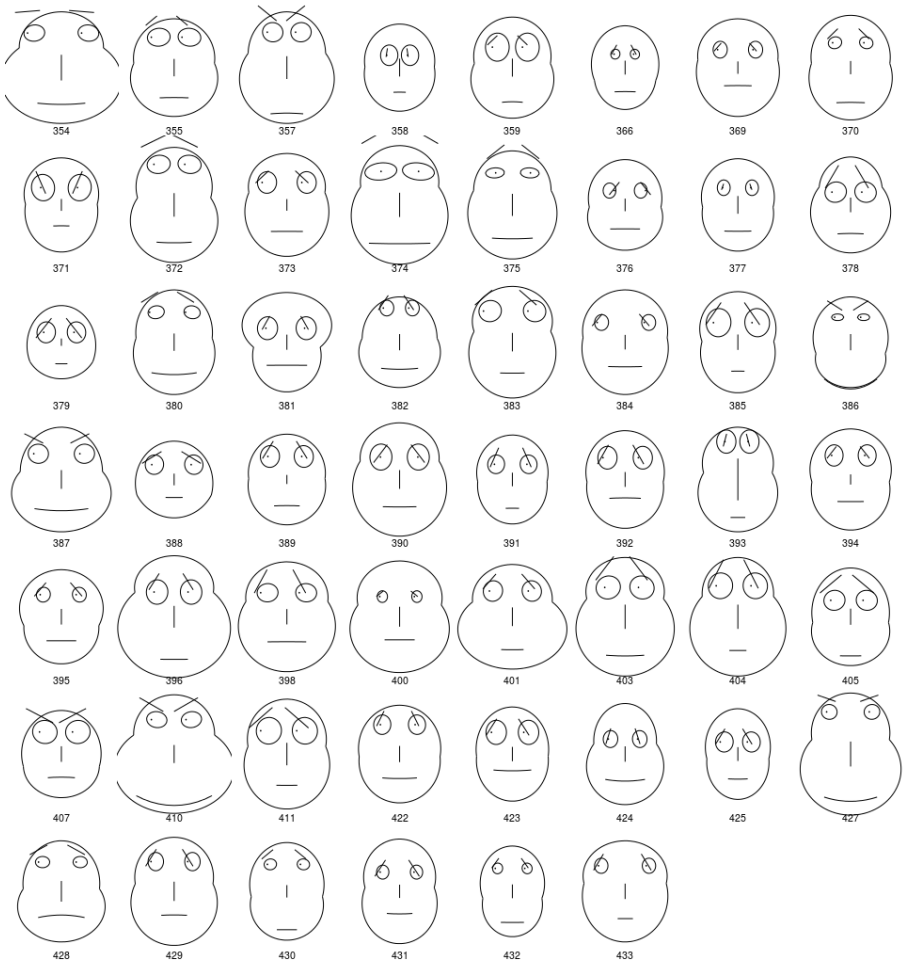


Figura 5.2: *Caras de Chernoff de las hembras trasandoscas*

El conjunto de datos que a seleccionar en este caso es el de MT, además para continuar simplificándolo y poder llegar antes a las conclusiones pertinentes se va a tomar como caso base un conjunto pequeño, con una única variable ya que si se incluyen todas las variables se corre el riesgo de que los *outliers* de unas y otras se diluyan, haciendo que identificarlos a simple vista sea más complicado. Esto es debido a que como se muestra en la Tabla 5.1, varias variables hacen referencia a una misma zona de la cara.

Por supuesto la variable que se seleccione debe contener *outliers*, en este caso se utilizará el *perimetro.toracico*, por dos razones. La primera de ellas y la principal es que contiene *outliers*, la segunda es que es deseable que la variable a usar sea representativa, y de entre

Zona de la cara	Variables					
Contorno superior cabeza	1	2	3	4	-	-
Contorno inferior cabeza	1	2	3	5	-	-
Nariz	6	-	-	-	-	-
Boca	7	8	9	-	-	-
Ojos	10	11	12	13	14	15
Cejas	12	16	17	18	-	-

Cuadro 5.1: *Variables relacionadas con cada zona de la cara en las caras de Chernoff.*

ellas una de las más destacadas es el *perimetro.toracico*. Si bien es cierto que quizás el peso sea la variable más representativa hay un problema y es que no todos los ganaderos se pueden costear una báscula para poder pesar a los animales lo que hace que termine cobrando mayor relevancia el *perimetro.toracico* ya que es considerablemente más barata una cinta métrica.

En este caso concreto no es posible utilizar como conjunto de validación a los otros dos conjuntos de machos ya que ninguno de ellos tienen outliers, y por lo tanto no es posible localizar o no, los *outliers*. En la misma situación se encuentran las HT, ya que tampoco contienen ningún *outlier*. Por lo que hay dos conjuntos para la validación y la prueba, y hay que validar con las HA y probarlo con las HC ya que estas últimas son más de la mitad y por lo tanto en caso de hacerlo al revés habría un conjunto de prueba demasiado pequeño.

En la Figura 5.3 se pueden observar los outliers que los MT presentan en la variable *perimetro.toracico*. En el boxplot se aprecia que hay un *outlier*, concretamente el del individuo 523 con valor de 85,6.

Para poder analizar correctamente las caras de Chernoff antes hay que ver qué representa cada variable en función de su orden en la base de datos. Al volver sobre la Tabla 2.4 se puede observar que en este caso concreto, en el que solamente hay una única variable, la única variación posible es la *Anchura del centro*. Esto significa que la única diferencia entre las caras es el ancho de la cabeza. Ahora al observar las caras 5.4 podemos apreciar como la muestra 523 tiene el contorno de la cabeza ligeramente más estrecho que el resto, lo que hace que destaque porque a ojos humanos parezca más pequeña que las demás. Partiendo

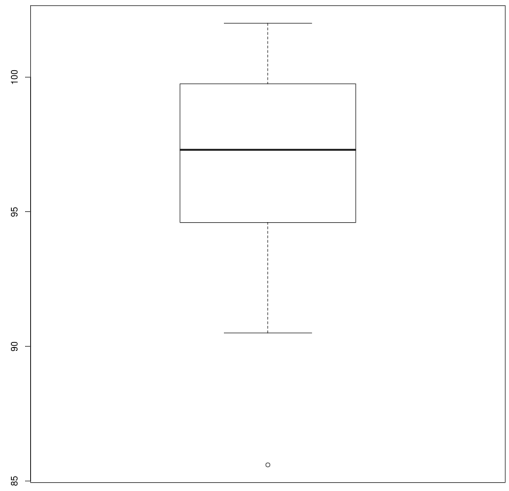


Figura 5.3: *Boxplot del perimetro.toracico de los machos trasandoscos*

de este hecho se pueden hacer suposiciones.

Parece, como es lógico, que a menor valor de la variable, con respecto al resto de elementos, menor es la representación de esta. Esto es lógico, ya que las caras de Chernoff se representan normalizando todo a 1 y realizando los cálculos pertinentes en función de los resultados obtenidos. El objetivo y reto ahora es ver si realmente se pueden apreciar las diferencias en esta característica sencilla, que unas caras sean tengan mayor o menor anchura, y mediante ello localizar los outliers.

Para ello se ha de observar el boxplot de las HA, para validar la propuesta. Se observa apreciar en la Figura 5.5 que las HA contienen un *outlier*, concretamente uno superior, dicho valor atípico corresponde al elemento 438 con valor de 85,5.

Este es un ejemplo de por qué no se pueden incluir todos los conjuntos en uno único en la búsqueda de *outliers*. Hay un MT con un *perimetro.toracico* pequeño para su edad y sexo con una medida de 85,6 mientras que por el otro lado hay una HA que es grande para su sexo y edad y en cambio mide 85,5.

Continuando con la explicación, ahora se observa la Figura 5.6. El ejemplar 438 efectivamente parece que es más grande que el resto, pero además parece que el 446 también es

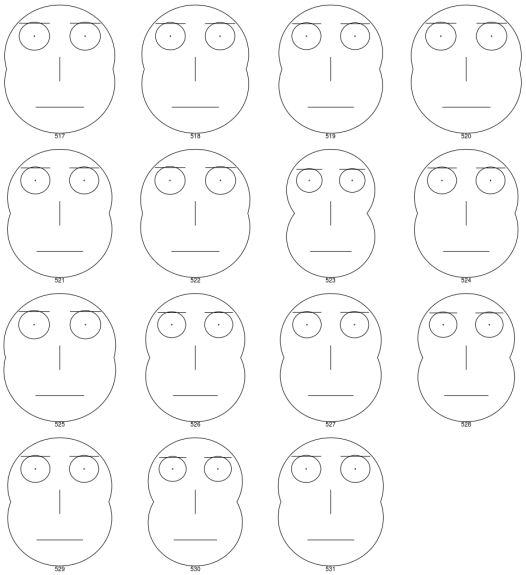


Figura 5.4: *Caras de Chernoff del perimetro.toracico de los machos trasandoscos*

un *outlier*, en este caso inferior, y gracias a haber realizado el boxplot anteriormente se sabe que no es cierto y que solamente hay un único *outlier*.

Por lo tanto no es posible llegar a la conclusión de que de manera unívoca se puedan encontrar los *outliers* observando los datos, sean cuales sean, ya que se ha encontrado un caso que tiene una respuesta negativa a esta hipótesis.

Pero aún no se sabe si los resultados obtenidos con un mayor número de variables permiten llegar a una conclusión más correcta, por lo tanto se procede a realizar una hipótesis por esta vía.

La hipótesis consiste en que sí es posible localizar los outliers con un mayor número de variables. Si el razonamiento inicial es que los outliers se podían difuminar entre ellos, ahora el razonamiento se basa en la posibilidad de que al haber un mayor número de variables y no depender de una única los registros con *outliers* es posible que tengan una cara *más rara* que el resto y se localicen más fácil y unívocamente .

El número de variables a tener en cuenta son 10, en este caso para trazar la hipótesis se seleccionan las 10 primeras variables, sin importar que algunas de ellas no contengan ningún tipo de valores atípicos, ya que forma parte de una situación perfectamente normal.

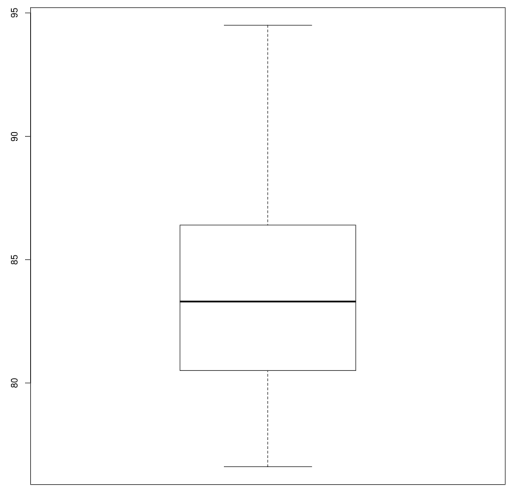


Figura 5.5: *Boxplot del perimetro.toracico de las hembras andoscas*

El proceso de localización en este caso consiste primero en localizar aquellas caras con *outliers* y posteriormente precisar las que contienen los *outliers*, en el caso de que alguna de las dos partes fuera inviable, la hipótesis quedaría descartada, ya que no se busca solamente saber que algún dato es incorrecto sino concretar cuál es.

El conjunto de datos sobre el que se va a trabajar tiene que ser suficientemente grande, ya que hay un número considerables de variables. En el caso anterior habiendo una única variable tener 18 registros aportaba un conjunto suficientemente variado y rico, pero ahora, teniendo 10 variables, aportaría un conjunto escaso. Por ello se selecciona como conjunto de prueba el de las HT ya que hay 54 registros que parece un número suficientemente grande y seleccionar las HC dejaría un bajo porcentaje del total de individuos para verificar la hipótesis, lo cual no es viable.

Antes de pasar a ver las caras de Chernoff y observar los resultados obtenidos hay que conocer cuales son las zonas de la cara que hay que observar y sobre los que hay que buscar los *outliers*. Como se explica en el Punto 2.4 el orden en el que se introducen las variables en las caras de Chernoff importa, ya que en función de uno u otro, se modifcarán ciertas zonas de dichas caras. En este caso interesan las 10 primeras variables, la relación entre ellas

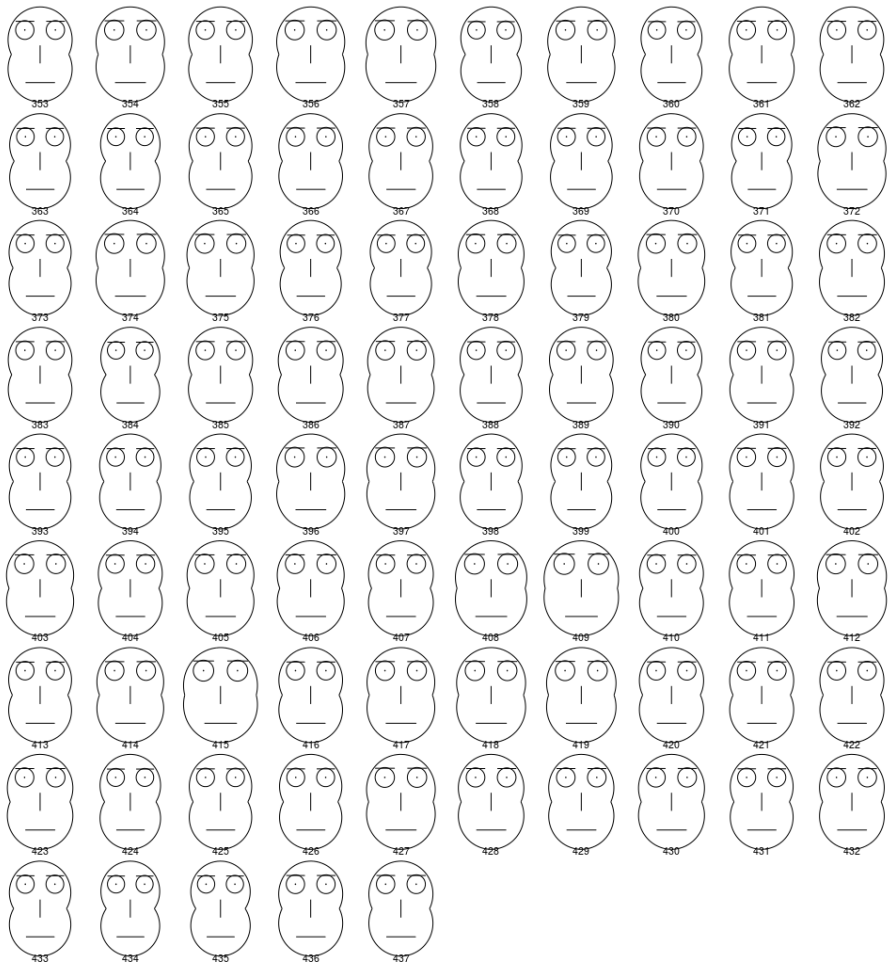


Figura 5.6: *Caras de Chernoff del perimetro.toracico de las hembras andoscas*

y su representación en las caras. Además el orden en el que se han introducido las variables en las caras de Chernoff es el siguiente: *altura.cruz*, *altura.medio.dorso*, *altura.grupa*, *altura.esternon*, *diametro.longitudinal*, *diametro.dorso*, *diametro.bicos*, *longitud.cabeza*, *anchura.cabeza* y *anchura.anterior.grupa*.

Por lo que en teoría se deben de observar las variables que se conoce que deben resaltar sobre el resto. Si bien es cierto que se deben observar todas ya que tanto en el caso de que alguna de las mencionadas anteriormente no destaque o que destaque alguna no mencionada anteriormente la hipótesis no será correcta.

Aunque se van a analizar los datos de manera conjunta se ha decidido decidido, de cara

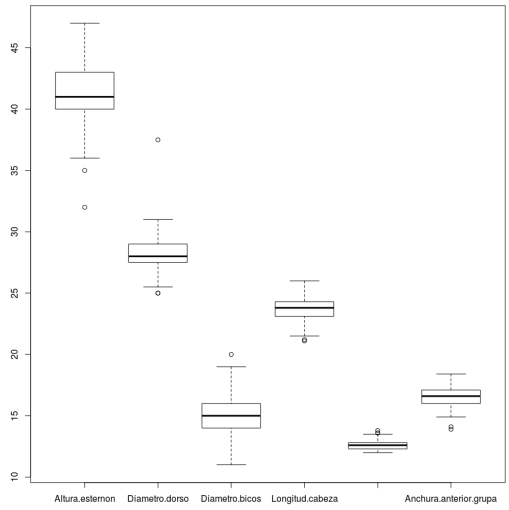


Figura 5.7: *Boxplot de 6 variables de las hembras trasandoscas*

al usuario, que para mejorar la vista de los datos se divida la muestra de los boxplot en dos gupos. El primero, en la Figura 5.8, con datos superiores a 50 cm, que consta de las siguientes 4 variables: *altura.cruz*, *altura.medio.dorso*, *altura.grupa* y *diametro.longitudinal*; y el segundo, mostrado en la Figura 5.7 con datos inferiores a 50 cm, que consta de estas 6 variables: *altura.esternon*, *diametro.dorso*, *diametro.bicos*, *longitud.cabeza*, *anchura.cabeza* y *anchura.anterior.grupa*.

Se puede apreciar que para estar trabajando con 10 variables y 85 individuos hay pocos valores atípicos, son un total de 15 distribuidos de la siguiente forma: en la *altura.medio.dorso*, hay 1 *outlier* concretamente en el individuo 368, dicha variable es la número 2 por lo que corresponde a la *altura de la separación* que exactamente consiste en el punto vertical en el que se juntan el trazo superior y el inferior del contorno de la cabeza. Al ser inferior esta medida debe ser menor que en el resto de individuos, por lo que debe de estar más abajo; otros 2 valores atípicos hay en en la *altura.esternon*, los 2 son inferiores con valor 32 y 35 en los individuos 409 y 424 respectivamente. La *altura.esternon* corresponde a la variable número 4 representada por el *ancho de la mitad superior de la cara*, así por ser inferiores tendrán una anchura menor que el resto de individuos; en el *diametro.longitudinal* hay 1

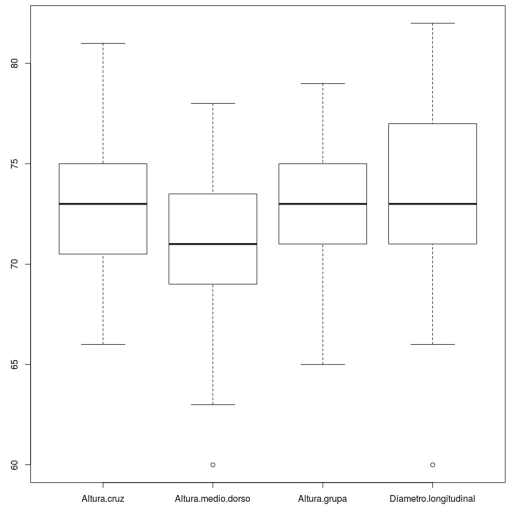


Figura 5.8: *Boxplot de 4 variables de las hembras trasandoscas*

outlier inferior situado en el individuo 425, esta variable está en la posición 5 por lo que corresponde al *ancho de la mitad inferior de la cara* al ser un outlier inferior tenemos que debe tener, al igual que en el caso anterior, una anchura menor, la única diferencia es que en el caso anterior corrspondía a la parte superior y en este caso es la parte inferior; en el *diametro.dorso* hay 2 outliers inferiores y 1 outlier superior, están en los registros 434 y 436 los inferiores ambos con valor 25 y 393 el superior con valor 37.5. *diametro.dorso* corresponde a la variable 6, por lo que hace referencia al *largo de la nariz*, al tener tanto outliers superiores como inferiores deberán distinguirse del resto, en el caso del valor atípico superior por ser más largo y en el caso de los dos valores atípicos inferiores por tener un menor tamaño que el resto de los individuos; en el *diametro.bicos* hay 1 outlier superior posicionado en el 409, al ser la variable 7 le corresponde el trazo de la *altura de la boca* y por ser un outlier superior tendrá un valor superior al del resto, teniendo así la boca más alta que el resto de individuos; en la *longitud.cabeza* hay 2 outliers inferiores con valores 21.1 y 21.2 y en los individuos 359 y 435 respectivamente. *longitud.cabeza* corresponde a la *curvatura de la boca* ya que es la variable que está en la posición 8, estos valores atípicos tendrán una curvatura menor que el resto de elementos; en la *anchura.cabeza* hay 3 outliers

superiores con valor 13.6, en las posiciones 386 y 412, y con valor 13.8, en la posición 408. Al ser la variable 9 le corresponde el *ancho de la boca* y al ser superiores tendrán un mayor ancho que el resto de elementos; en la *anchura.anterior.grupa* hay 2 outliers inferiores que corresponden a los individuos 364 y 434 con valores 13.9 y 14.1 respectivamente. Ocupa la posición 10 entre las variables y por lo tanto es la que determina la *altura de los ojos*, por esto y por ser outliers inferiores estos dos valores atípicos tendrán una menor altura de los ojos que el resto de individuos;

Así hay un total de 15 *outliers* distribuidos en 8 variables y en los siguientes individuos: 359, 364, 368, 386, 393, 408, 409 (con dos valores atípicos), 412, 424, 425, 434 (con dos valores atípicos), 435, 436. Al observar los outliers del conjunto de datos puede verse que se dan dos circunstancias que son propicias para nuestra propuesta. La primera de ellas es que hay dos individuos con más de un valor atípico, 409 y 434. Esto es importante ya que así puede evaluarse si en el caso de haber más de un único outlier el ojo humano es capaz de localizar ambos sin problemas. Además se da otra circunstancia y es que existen tres elementos consecutivos con outliers, 434, 435 y 436, al igual que en el caso anterior esto es importante, ya que en algunas circunstancias podemos obviar algunos valores que nos interesan por estar cerca de otros que destaquen más.

Por lo tanto para poder seguir en esta línea de desarrollo debería haber un rasgo anómalo en cada una de estas caras, y que no se diera en ninguna otra que no fuera un outlier. De hecho para poder realizar pruebas reales, sobre personas no implicadas en la propuesta, deberíamos de poder determinar, las personas implicadas en el proyecto, dichas variables de manera unívoca de modo que sea posible localizar todos aquellos outliers y no tomar como outlier aquellos valores que no lo sean. En este caso se parte con una ventaja y es que se conoce qué representa cada variable y se sabe lo que se está buscando, que en principio se cuenta con que no sea necesario por parte de un usuario normal que lo conozca.

Pueden verse las caras de Chernoff en la Figura 5.9. Al observarlas, en busca de posibles *outliers*, se aprecia que destacan ciertos individuos, a continuación se han enumerado orde-

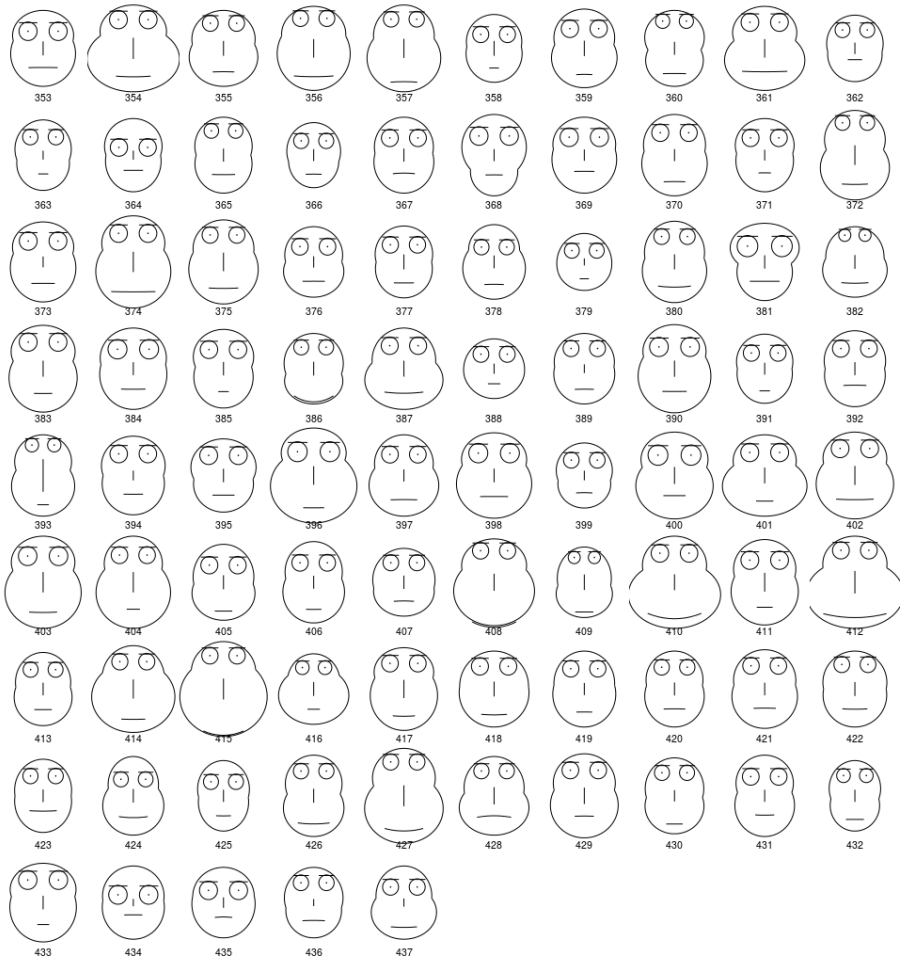


Figura 5.9: *Chernoff de 10 variables de las hembras trasandoscas*

nados por la variable por la que destacan, es importante tener en cuenta que solamente se han buscado individuos que destaquen por los atributos con outliers.

Al analizar la *altura de la separación* puede verse que los individuos 368, 381 y 432 destacan con valores pequeños y el 399 con un valor aparentemente más grande mientras que con otros 3 individuos, 379, 388 y 435, se produce el problema de que no se distingue bien donde se unen las partes superior e inferior del contorno de la cara; En el *ancho de la mitad superior de la cara* se observa que el registro 381 tiene el tamaño considerablemente más grande que el resto de registros, en este caso al igual que en el homólogo de la mitad inferior de la cara no se da el problema anterior ya que no es tan importante donde se separan

ambos sino la anchura máxima que toman; El *ancho de la mitad inferior de la cara* tiene como valores visualmente destacados 4 elementos que destacan por parecer grandes, son los siguientes: 354, 410, 412, 415. En especial el individuo 412 que parte del contorno inferior sobrepasa el espacio destinado para dibujar al individuo; A la hora de analizar el *largo de la nariz* se puede ver que el 379 y el 436 parecen tener una nariz corta mientras que el 393 aparentemente tiene una nariz más larga de lo normal; Al observar la *altura de la boca* los elementos 386, 408, 410, 412 y 415 tienen una boca baja, en especial el 408 que incluso no está contenida en el resto del rostro. Por otro lado el 434 parece tener la boca más alta de lo normal; Siguiendo con la boca, al observar la *curvatura de la boca*, donde algunos de los elementos anteriormente citados se repiten, los elementos 386, 408, 410, 412 y 415 tienen una curvatura positiva, representando una cara feliz, lo que implica un valor superior por lo tanto serían valores más grandes de lo habitual. También destaca la curvatura del 428 que muestra una curvatura negativa, representando una cara triste, por lo que sería un valor menor que el resto; Para finalizar el apartado de la boca hay que analizar el *ancho de la boca*, al observarlo se ve con que el 361, el 410 y el 412 tienen bocas grandes y el 385, el 388 y el 379 muestran bocas pequeñas; Por último la *altura de los ojos* solamente tiene un individuo que destaca que es el 382 con unos ojos considerablemente elevados.

Ahora hay que comparar los outliers reales con los localizados a la hora de ver las caras de Chernoff. Se realizará la comparativa en función de las variables, ya que quizás ciertas características sean más fáciles de ver que otras.

Para elaborar la tabla con los resultados se ha decidido discernir entre un valor atípico inferior y superior, para ello el individuo viene acompañado a continuación con una i o una o respectivamente. La Tabla 5.2 contiene 5 columnas, la primera de ellas muestra la zona de la cara que se va a evaluar; la segunda los individuos con los outliers reales y la letra en función de que sea superior o inferior; la tercera muestra los outliers localizados sobre las caras de Chernoff; la cuarta columna muestra el % de los outliers reales localizados en las caras de Chernoff; finalmente la quinta columna, % de fallo, muestra en % cuantos supuestos

Variable	**Outliers reales**	**Outliers localizados**	**% de reales acertados**	**% de fallo**
Altura separación	368i	368i 381i 432i 399s	100 %	75 %
Ancho mitad superior	409i 424i	381s	0 %	100 %
Ancho mitad inferior	425i	354s 410s 412s 415s	0 %	100 %
Largo nariz	436i 434i 393s	379i 436i 393s	66 %	33 %
Altura boca	409s	386i 408i 410i 412i 415i 434s	0 %	100 %
Curvatura boca	359i 435i	386s 408s 410s 412s 415s 428i	0 %	100 %
Ancho boca	386s 412s 408s	361s 410s 412s 385i 388i 379i	33 %	83 %
Altura ojos	364i 434i	382s	0 %	100 %
Totales	14	31	28,57 % (4/14)	87,09 % (27/31)

Cuadro 5.2: *Resultados obtenidos al observar las caras de Chernoff.*

outliers localizados en las caras de Chernoff realmente no lo son.

Se puede observar en la 5.2 que los resultados no son nada satisfactorios, ya que solamente se han conseguido conseguido localizar un 28,57 % de los outliers. Pero esa no es la peor noticia de todas, ya que si se consiguiera localizar pocos de los outliers cometiendo muy pocos errores se podría seguir avanzando en esta línea para buscar mejorar los resultados, pero no es el caso. El mayor de los problemas es que en la búsqueda de outliers se han encontrado 31 posibles outliers, de los cuales solamente 4 son correctos y obteniendo por lo tanto un 87,09 % de falsos outliers localizados.

Esto puede ser debido a que en algunos casos la diferencia entre un valor atípico y uno considerado normal puede ser mínima, por lo que a la hora de normalizar los datos e introducirlos en las caras de Chernoff, la diferencia entre uno y otro sea prácticamente imperceptible al ojo humano.

La evaluación de la hipótesis es negativa y por tanto hay que buscar otra alternativa para localizar los outliers. Pese a que los resultados obtenidos no son satisfactorios, se puede

llegar a pensar que las caras de Chernoff son la mejor alternativa actual para la localización de outliers en un entorno multivariante, ya que gracias a ellas pueden verse a la vez todos los datos de un gran número de individuos.

5.1.2. Completando las caras de Chernoff

El objetivo es conseguir que sean perfectamente localizables los outliers, para ello se pretende encontrar una alternativa unívoca, de modo que no tenga una dependencia tan alta ni del criterio ni de la percepción humana. Así la alternativa planteada es marcar las zonas de un modo diferenciador, para así que resalten de mayor manera y no se confundan con valores cercanos a los atípicos.

En este trabajo se propone como rasgo diferenciador la inclusión de colores, de modo que cada característica de las caras que sea un *outlier* se trace en un color diferente a los datos normales, que se trazan en negro. Para llevar a cabo esto hay que tener en cuenta varios aspectos; el primero de ellos es que el código de colores debe ser lo suficientemente claro, ya que en el caso de dar lugar a confusiones se podrían confundir valores atípicos de unas variables con otras, en el caso de que hubiera varios tonos de verdes y un usuario no diferenciara entre algunos de ellos; el segundo aspecto a tener en cuenta es que una misma zona puede estar representada por varias variables, en el caso del contorno superior de la cabeza representa a las variables 1, 2, 3, 4 las cuales hacen referencia la anchura del centro, la altura de la separación, altura de la cara y ancho de la mitad superior de la cara respectivamente, como se representa en la Tabla 5.1. Por lo tanto no es suficiente marcar con un trazo más grueso una zona, ya que en ese caso no se podría concretar suficiente aún sobre qué variable puede ser, sino que hay que utilizar otro método.

Para poder incluir esta modificación en las caras se va a modificar la librería $TeachingDemos$, concretamente en las funciones $faces2$ y $faces2.plot$, la primera es función usada para representar las caras de Chernoff, mientras que la segunda es la función que utiliza $faces2$ para pintar cada cara. Como cada cara se pinta mediante una función que no tiene conocimiento

Número de Variable	¿Qué representa en R?	Categorización
1	Anchura del centro	Contorno cabeza
2	Superior Vs inferior, altura de la separación	Contorno cabeza
3	Altura de la cara	Contorno cabeza
4	Ancho de la mitad superior de la cara	Contorno superior cabeza
5	Ancho de la mitad inferior de la cara	Contorno inferior cabeza
6	Largo de la nariz	Nariz
7	Altura de la boca	Boca
8	Curvatura de la boca (abs <9)	Boca
9	Ancho de la boca	Boca
10	Altura de los ojos	Ojos
11	Distancia entre los ojos (.5-.9)	Ojos
12	Ángulo de ojos y cejas	Cejas
13	Elipse de los ojos	Ojos
14	Tamaño de los ojos	Ojos
15	Posición izquierda/derecha de los ojos	Ojos
16	Altura de las cejas	Cejas
17	Ángulo de las cejas	Cejas
18	Ancho de las cejas	Cejas

Cuadro 5.3: *Categorización de cada variable de las caras de Chernoff.*

de la situación de esta cara con respecto al resto de caras, y por lo tanto no se pueden localizar outliers dentro de esta función, es necesario pasar por parámetro el hecho de que tenga o no tenga outliers. En este caso se va a pasar por parámetro un vector con 6 valores, los cuales determinan el color en hexadecimal que debe tener cada categoría (contorno superior, contorno inferior, nariz, boca, ojos y cejas), de modo que antes de dibujar cada sección se consulta el color y se aplica. Por lo tanto en la función $faces2$ hay que hacer un tratamiento previo de los datos, localizando outliers de cada variable y determinando los colores que se aplicará a cada rasgo de la cara.

Para poder hacerlo se han realizado categorizaciones para saber en qué zona hay que pintar cada variable, asignando cada variable a una única zona de la cara, las categorizaciones están mostradas en la Tabla 5.3.

Se puede apreciar que en el caso de la variable 12 hace referencia tanto a ojos como cejas, pero en este trabajo se ha decidido asignar dicha variable a las cejas ya que, sin

Categoría	Variables de Chernoff				
	1	**2**	**3**	**4**	**5**
Contorno superior cabeza	1	2	3	4	-
Contorno inferior cabeza	1	2	3	5	-
Nariz	6	-	-	-	-
Boca	7	8	9	-	-
Ojos	10	11	13	14	15
Cejas	12	16	17	18	-

Cuadro 5.4: *Variables relacionadas con cada categorización en la modificación de las caras de Chernoff.*

tener en cuenta esta, tiene un menor número de variables (3) que los ojos (5). Una vez se ha visto las categorizaciones donde se representan los outliers, se va a ver qué colores utilizar para marcarlos, se pueden apreciar a qué categorización corresponde cada variable y el número que ocupa en dicha categorización en la Tabla 5.4. Los colores deben ser lo más exclarecedores posibles para evitar confusiones del usuario con respecto a los mismos. Así el código de colores utilizado debe abarcar al menos un abanico de 18 colores, tantos como variables haya, ya que puede darse el caso de que cada variable contenga al menos un outlier. Para conocer la relación entre los colores y los outliers, el objetivo es que los colores sean sencillamente reconocibles, por lo que se va a realizar dividiendo en 18 partes el espectro de colores y asignando al siguiente color el más alejado posible de los colores que ya han sido asignados. Como caso base se tomarán los colores primarios, amarillo, cyan y magenta y a continuación se realizará lo citado anteriormente, por lo que la segunda terna de colores serán los colores secundarios azul, verde y rojo y a continuación los colores terciarios. En la Figura 5.10 se pueden apreciar los colores para facilitar la comprensión de los mismos y en la Tabla 5.5 está la relación entre las variables y los colores, tomando aquellos que interesan en este trabajo. Una vez se tienen los colores que se van a utilizar se puede ver su aplicación sobre un caso concreto. Al aplicarlo sobre el caso anterior, el de las 10 variables de las hembras trasandoscas, se obtienen buenos resultados. En este caso funciona perfectamente, pero existe un pequeño problema y es que como se ha dicho antes algunos rasgos comparten una misma zona de la cara, juntando eso a que un cierto individuo puede tener más de un

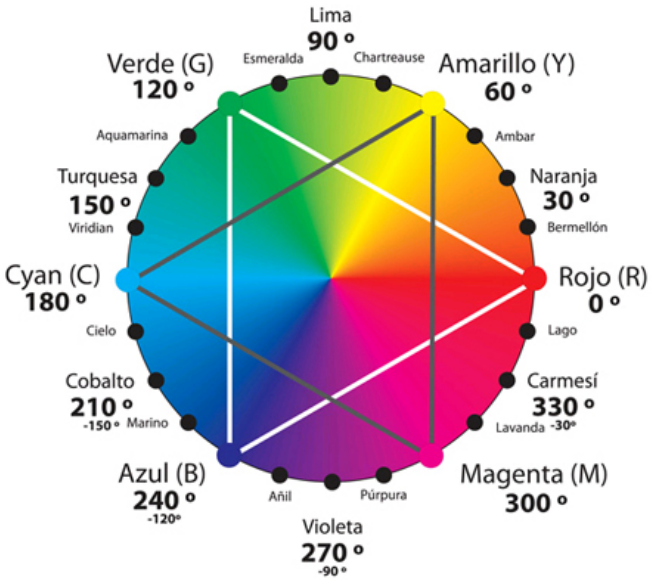


Figura 5.10: *Espectro con los principales colores.*

valor atípico, como ya se ha visto, se tienen que un individuo puede tener un valor atípico en dos variables que pertenezcan a la misma zona, teniendo el valor atípico en la variable 1 y 2 por ejemplo lo cual supone un problema para este sistema, ya que se vería únicamente uno de ellos.

Por lo tanto este sistema solamente ayuda a diferenciar qué rasgo contiene el outlier, pero en el caso de que en una misma zona haya más de un outlier no es capaz de mostrar todos. Por ello se propone una mejora de este sistema, que es mejorar el código de colores.

5.1.3. Mejorando el código de colores

Se pretende localizar todos los outliers y combinaciones de los mismos ampliando el abanico de colores. Así se amplía el abanico de 18 colores, con tantos colores como sean necesarios para cubrir las posibles combinaciones de outliers en todos los rasgos. El número de combinaciones posibles se conoce en base al número de variables que se representan por rasgo y hallando las posibles combinaciones sin repeticiones. Las combinaciones sin repeticiones vienen representadas mediante la Fórmula 5.1, donde p representa el número de posibilidades que hay, el número de variables que comparten zona; m el número de

Número de Variable	Color del outlier
1	Cyan
2	Magenta
3	Amarillo
4	Azul
5	Verde
6	Rojo
7	Turquesa
8	Violeta
9	Naranja
10	Cobalto
11	Carmesí
12	Lima
13	Añil
14	Bermellón
15	Aquamarina
16	Lavanda
17	Chartreause
18	Cielo

Cuadro 5.5: *Correspondencia de cada color y cada variable.*

elementos que se pueden combinar, en este caso son aquellos que pueden contener outliers; y n el número de elementos que tienen el outlier, tomando valores desde 1 hasta el máximo posible, m.

$$p = \sum_{n=1}^{m} \binom{m}{n} \tag{5.1}$$

Así se tiene que en el caso del contorno de la cabeza se tiene que el contorno superior tiene un máximo de 15 posibles combinaciones igual que el contorno inferior que también puede tener 15 combinaciones; mientras que la nariz únicamente tiene una posible combinación, ya que está representada por una única variable; en el caso de la boca tiene 7 posibles distribuciones de los valor atípicos; así como los ojos tienen 31 y las cejas 15. Por lo tanto se tienen un total de 84 colores.

84 colores son un número desmesurado de colores, ya que uno de los problemas que tienen los colores es que pueden llegar a difuminarse unos con otros, en cuyo caso dejarían de ser esclarecedores y podrían dar lugar a confusiones.

Es por esto que se realiza otra modificación, el código de colores no se aplica de manera global sino que se aplica por categorizaciones. Esto es, que para cada categorización de la cara se pueden repetir los colores, así se pueden abarcar las combinaciones posibles con menor número de colores siendo los resultados más claros. Por lo tanto, por ejemplo, el color naranja ya no hace referencia a la novena variable, sino que en función de si está localizado en una categorización u otra hace referencia a variables distintas, si bien es cierto que en todos los casos hará referencia a la novena combinación posible de la categorización. El significado de los posibles colores se puede apreciar en la Tabla 5.6, como se puede observar, los colores (*Color de la combinación*) hacen referencia a diversas combinaciones (*Tipo de combinacion*), donde se mostrarán únicamente las variables que contengan outliers. Es importante destacar que dichas variables no están representadas por su posición global, sino por su posición dentro de la categorización, para facilitar la compresión se muestra el siguiente ejemplo.

En el caso de los ojos, las variables que participan en la creación de los mismo son la 10, la 11, la 13, la 14 y la 15; por lo tanto a la hora de asignar los colores se toman de la siguiente forma, la variable 10 será representada como la variable 1 local, así como la 11 será la 2, la 13 la 3, la 14 la 4 y la 15 la 5.

Para que sea más claro, se ha decidido utilizar una distribución de colores global a todas las variables, de modo que en el caso de la nariz el outlier siempre será de color cyan, mientras que en el caso de, por ejemplo, las cejas abarcarán los primeros 15 colores. Esto es para evitar en la medida de lo posible los colores similares, de modo que solamente en el caso de los ojos tendremos más de 15 posibles colores.

Como se puede observar en la tabla se han tenido que introducir nuevos colores, ya que se han terminado de introducir los ternarios de la Figura 5.10, en vez de seguir introduciendo colores derivados de los anteriores, lo cual podría dificultar la localización de los mismos, se ha optado por incluir colores más esclarecedores como son: plata, oro, gris claro, marrón oscuro, gris oscuro, marrón claro y rosa claro.

Tipo de combinación	Color de la combinación	Representación hexadecimal
1	Cyan	#00FFFF
2	Magenta	#FF0099
1 y 2	Amarillo	#FFFF00
3	Azul	#0000FF
1 y 3	Verde	#008000
2 y 3	Rojo	#FF0000
1, 2 y 3	Turquesa	#48D1CC
4	Violeta	#EE82EE
1 y 4	Naranja	#FFA500
2 y 4	Azul Cobalto	#104E8B
3 y 4	Carmesí	#DC143C
1, 2 y 4	Lima	#00FF00
1, 3, y 4	Añil	#4B0082
2, 3 y 4	Berenjena	#8B3A62
1, 2, 3 y 4	Verde Pálido	#9AFF9A
5	Lavanda	#E6E6FA
1 y 5	Verde Césped	#7CFC00
2 y 5	Azul Cielo	#87CEEB
3 y 5	Púrpura	#9370D8
4 y 5	Ámbar	#CD6600
1, 2 y 5	Viridian	#218868
1, 3 y 5	Azul Marino	#000080
1, 4 y 5	Ladrillo	#EE2C2C
2, 3 y 5	Azul Claro	#ADD8E6
2, 4 y 5	Plata	#C0C0C0
3, 4 y 5	Oro	#EEAD0E
1, 2, 3 y 5	Gris Claro	#D3D3D3
1, 2, 4 y 5	Marrón Oscuro	#993300
1, 3, 4 y 5	Gris Oscuro	#2F4F4F
2, 3, 4 y 5	Marrón Claro	#DEB887
1, 2, 3, 4 y 5	Rosa Claro	#FFB6C1

Cuadro 5.6: *Correspondencia de cada color y cada variable.*

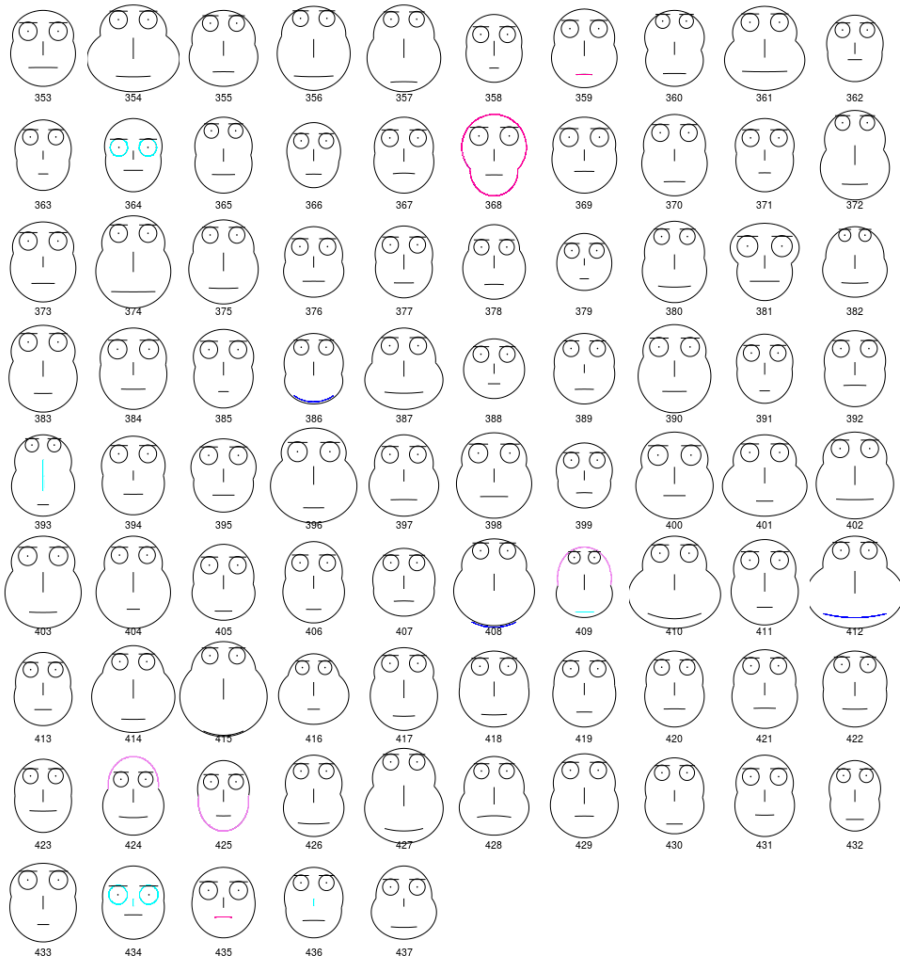


Figura 5.11: *Modificación del código de colores en las caras de Chernoff (sobre las hembras trasandoscas).*

En la Figura 5.11 puede apreciarse un ejemplo, el mismo que en el caso anterior, con la nueva codificación de colores. Pese a que en una primera instancia pueda parecer más complejo es totalmente necesaria la asignación de colores por categorizaciones, ya que como se ha dicho antes, en caso contrario se tendría un abanico de colores demasiado amplio y confuso.

Esta modificación ha sido probada por un total de 30 personas de diversas edades, comprendidos entre los 24 y los 60 años, y sin que necesariamente tuvieran conocimientos informáticos o estadísticos. Al localizar el usuario los colores, haciendo uso de las tablas necesarias para saber a qué colores corresponde cada valor, las cuales son la Tabla 5.6 y la

Individuo (Var)	Categorización	Color	Localizado	Correcta Interpretación
359 (8)	Boca	Magenta	30	26
364 (10)	Ojos	Cyan	30	12
368 (2)	Contorno Cabeza	Magenta	30	30
386 (9)	Boca	Azul	26	18
393 (6)	Nariz	Cyan	30	29
408 (9)	Boca	Azul	24	13
409 (4)	Contorno Superior	Violeta	24	17
409 (7)	Boca	Cyan	30	21
412 (9)	Boca	Azul	30	27
424 (4)	Contorno Superior	Violeta	30	30
425 (5)	Contorno Inferior	Violeta	30	30
434 (6)	Nariz	Cyan	13	10
434 (10)	Ojos	Cyan	30	13
435 (8)	Boca	Magenta	30	30
436 (6)	Nariz	Cyan	24	21
Errores	-	-	6	-

Cuadro 5.7: *Estadísticas de las caras de Chernoff modificadas.*

Tabla 5.4, se consiguen los resultados mostrados en la Tabla 5.7.

Es importante apreciar en la tabla de resultados que se reduce drásticamente el número de outliers que localiza el usuario y finalmente no lo son (6), ya que en este caso se deja de buscar una forma que destaque por encima del resto y se buscan colores, pero todavía se interpretan algunos casos en los que las personas que han realizado las pruebas creen ver ciertos colores cuando realmente en esa categorización no hay ningún tipo de valor atípico, Esto es debido en parte a que al explicar las instrucciones a los usuarios que han realizado las pruebas, se toma como ejemplo la primera cara, lo que hace que algunos resultados reflejen la primera cara con zonas coloreadas cuando realmente no está.

Más allá de eso el único caso posible para que el usuario localice valores atípicos que no son es que se confunda a la hora de determinar de qué color es el valor atípico, que sucede en el caso de localizar el azul que algunas de las personas lo han confundido con el azul marino, o el cyan confundido también con el turquesa o viridian.

Uno de los resultados que más destaca es la baja localización de la nariz cyan en el

individuo 434 (43 % de localizaciones). Teniendo en cuenta que otros outliers con similares características tienen un 100 % de localización y una correcta interpretación en el 96 % de los casos como sucede con el individuo 393, la baja localización en el 434 puede deberse a la presencia de un valor atípico de idéntico color en los ojos, lo que hace que se difuminen entre ellos.

Además la mayoría de los outliers se localizan, sucede en un 91.33 % de los casos, si bien es cierto que de estos valores bien localizados se interpretan bien un 79.56 % de los casos, lo cual sería deseable que fuera más elevado.

La mejoría de los resultados es evidente, ya que lo normal es que se detecten los outliers y que no se detecten valores no atípicos como atípicos, si bien es cierto que la situación idónea, que el error fuera 0 y tanto la localización como la correcta interpretación fuera del 100 %.

Pero el hecho de conseguir localizar, en mayor o menor medida, los valores anómalos puede no ser suficiente si no se consigue representar en las caras de Chernoff un mayor número de variables. En este trabajo al tener variables morfológicas se van a buscar correlaciones entre ellas para poder conseguirlo, no de manera directa pero sí al menos de cierta forma. Con ello se pretende que todas las variables estén representadas de alguna manera en las caras de Chernoff, en el caso de la base de datos de las cabras solamente se necesita una única relación, ya que se tienen 21 variables y los rebaños están agrupados en función de dos de ellas, *sexo* y *edad*, por lo que hays que representar de algún modo las 19 variables restantes.

En el punto siguiente se muestra cómo hacer uso de las correlaciones en R para poder localizar las relaciones y posteriormente cómo se ha definido la recta de regresión.

5.2. Correlación y regresión.

Se van a buscar correlaciones entre las variables de la base de datos de las cabras, para lo cual se utiliza la función de R *corrgram* y una vez se han localizado las variables que tienen

cierta correlación entre ellas se aplican las funciones descritas en el apartado 3.5.

Se va a comprobar si tiene alguna relación la *edad* y el *sexo* de las cabras a la hora de realizar correlaciones entre sus variables. Se tomará la mejor correlación entre las variables de uno de los rebaños para compararla a nivel global. Al igual que en el caso anterior se tienen que dividir las cabras en rebaños, ya que en caso contrario podría no encontrarse relación alguna entre las variables. Al realizarlo, y teniendo en cuenta las características de un buen estimador, mostadas en el Punto 2.5, se tiene que es deseable que los datos obtenidos con un conjunto de datos aumente según aumenta el tamaño de los datos. Por ello puede ser lógico pensar que seleccionar como conjunto de entrenamiento las hembras trasandoscas (85 individuos), o las andoscas (43), es un buen comienzo, ya que existe un conjunto con mayor número de individuos, como son las hembras cerradas (350), y otros conjuntos con menor número de individuos como son cualquiera de los conjuntos de machos, ya sean andoscos (16), trasandoscos (15) o cerrados (20) [1]. Pero hay que tener en cuenta que en un conjunto de solamente 85 individuos los casos poco corrientes que se puedan dar en el rebaño destacan más, ya que cada individuo representa casi el 1.2 % del rebaño entero. Por ello se realizarán las pruebas sobre el rebaño de hembras cerradas, así se pretende trabajar con un conjunto suficientemente amplio como para que los individuos con morfología poco corriente no inviten a tomar decisiones erróneas. Así se selecciona el mayor conjunto de datos posible y se intenta llegar a conclusiones a partir de dicho conjunto.

Aplicando la función *corrgram* a las hembras cerrradas se consiguen los resultados mostrados en la Figura 5.12, se puede apreciar como la correlación más fuerte se da entre la variables *peso* y *perimetro.toracico*, con un 84 % de coeficiente de correlación, siendo una correlación buena y positiva por lo que si una variable aumenta la otra tenderá a aumentar y si decrece tenderá a decrecer.

Es muy interesante que la variable *peso* tenga una correlación con la variable *perimetro.toracico*

[1] Es importante destacar que en este caso el número de individuos válidos difiere con los que hay en las pruebas de las caras de Chernoff, esto es debido a que en este caso el número de variables es menor y por lo tanto hay menor número de NA lo que hace que haya un mayor número de individuos válidos.

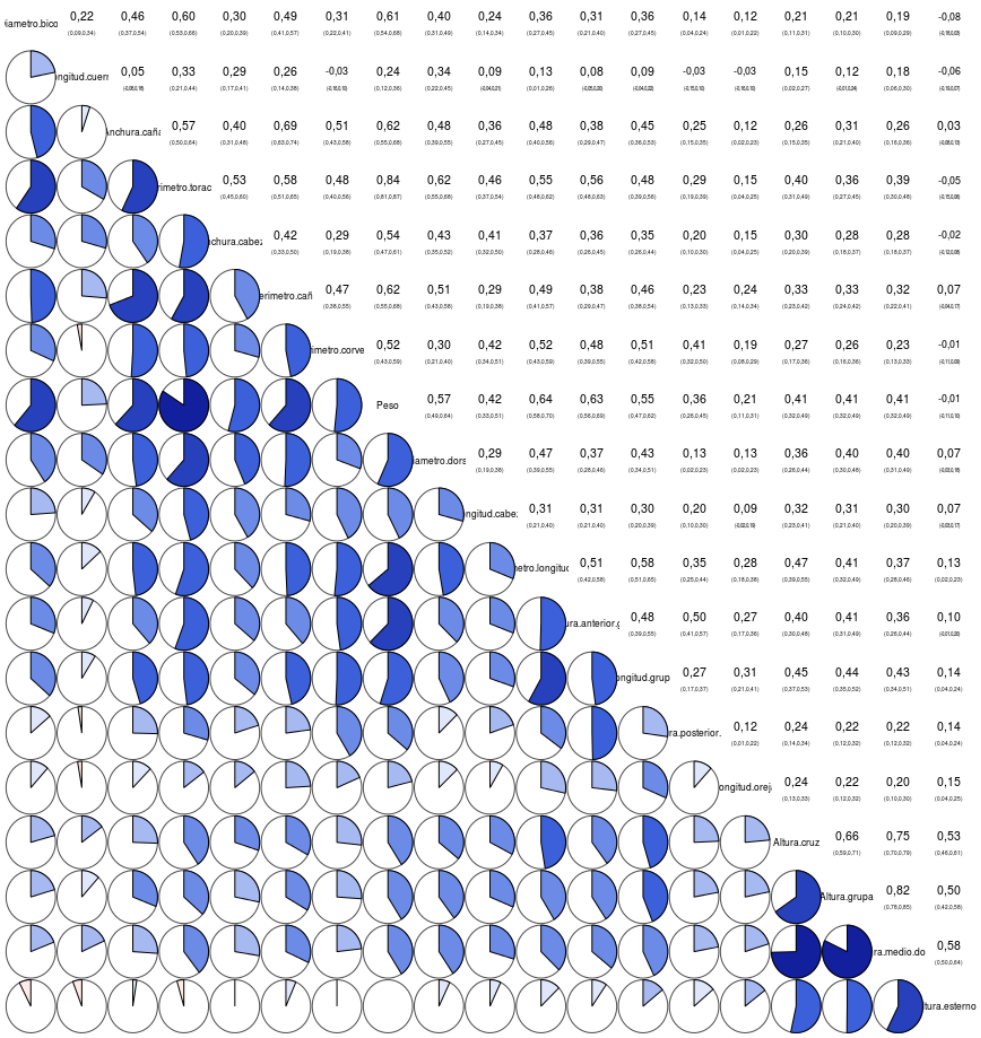


Figura 5.12: *Correlograma de las 352 hembras cerradas*

ya que como se ha dicho en el punto anterior, no todos los ganaderos pueden costearse una báscula por lo que si se verifica una correlación entre una y otra se permite que los ganaderos conozcan el peso aproximado de un individuo sin necesidad de invertir ese dinero, con el simple hecho de medir el perímetro torácico del animal.

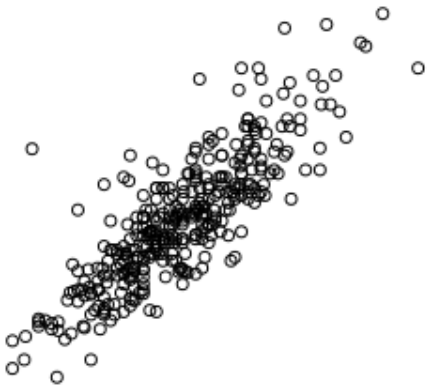


Figura 5.13: *Nube de puntos del peso y perímetro torácico en las hembras cerradas*

Correlación entre peso y perímetro torácico

Esto corrobora lo citado en los estudios de [García Lara et al., 2009] en los cuales se calcula el peso en función del perímetro torácico, por lo que en este trabajo se va a estudiar la correlación y la precisión de la misma. Se puede observar, en la Figura 5.13, la nube de puntos de ambas variables. Se puede apreciar en ella la correlación entre las variables y que dicha correlación es claramente lineal, por lo que hay que proceder tal como se explica en el Apartado 2.2.1.

Por lo tanto la correlación que se desea calcular es de la forma $\hat{y} = a + bx$, donde x representa el *perimetro.toracico* e y representa el *peso*, y lo que hay que hacer es hallar los valores de a y b para que se ajuste lo mejor posible al conjunto de entrenamiento. Las fórmulas para hallar a y b son $a = \bar{y} - b\bar{x}$ y $b = \frac{S_{xy}}{\sqrt{S_x} * \sqrt{S_Y}}$ respectivamente. Por ello se necesita hallar previamente las varianzas tanto de xy como de x, representadas mediante la fórmula *varianza*.

Así los resultados obtenidos de las varianzas son $S_{xy} = 39,88489$, $S_x = 28,16673$ y $S_y = 80{,}06601$ y por lo tanto el valor de b es $b = 0{,}84$ y usando este último resultado obtenido y las medias de ambas variables, $\bar{x} = 88,80086$ y $\bar{y} = 56,71229$ se halla el de a siendo $a = -17{,}8043$. Quedando finalmente la función de correlación entre ambas variables queda como se muestra en la Figura 5.2.

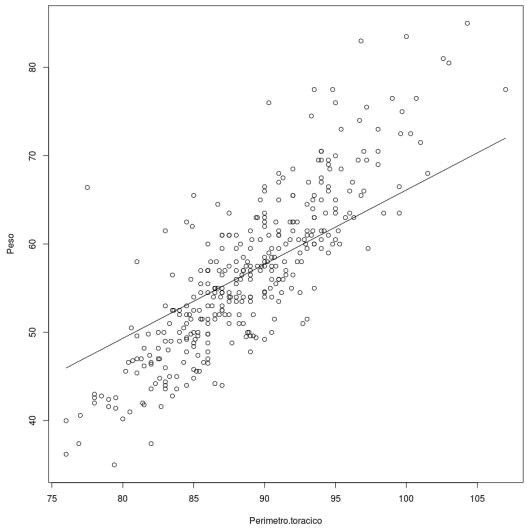


Figura 5.14: *Línea de regresión de la correlación entre peso y perímetro torácico en las hembras cerradas.*

$$\hat{y_i} = -17{,}8043 + 0{,}84 * x_i \tag{5.2}$$

Verificando dicha función de correlación, con los datos de los que se dispone, los individuos reales, se obtiene la línea de regresión mostrada en la Figura 5.14. Dicha correlación tiene un R^2 sobre esa sección de la base de datos de 0.8308, el cual se ha de usar posteriormente para compararlo con el global.

Si ahora se hallan las correlaciones globales, se obtienen buenas correlaciones, mostradas en la Figura 5.15.

De estas correlaciones destaca que la relación entre el *peso* y el *perimetro.toracico* se incremente, esto denota una cuestión muy sencilla y es que al aumentar la correlación entre ambas variables puede afirmarse que estimar la variable *peso* usando el *perimetro.toracico* puede ser una buena estimación, ya que según las características citadas en el Apartado 2.5 para ser un buen estimador, se tiene que es deseable que según aumenta el conjunto de datos se consigue una mejor estimación para poder afirmar que sea *consistente*.

Además de eso también cumple que los datos ofrecidos son considerablemente buenos en

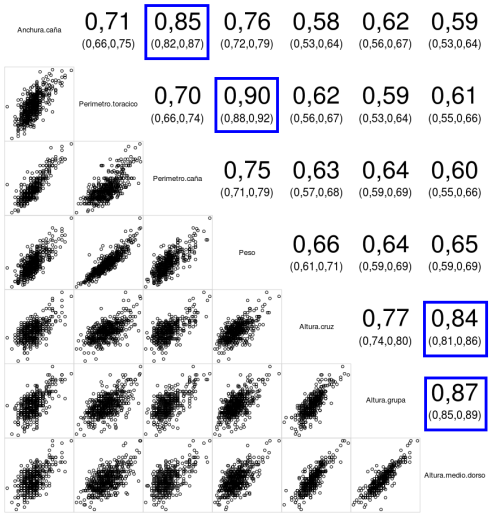


Figura 5.15: *Mejores correlaciones de la tabla total de cabras*

términos de *eficiencia*, ya que tiene un ECM de 46.90, lo que hace que los valores ofrecidos puedan ser considerados buenos.

Además de la mejoría entre la correlación *peso* y *perimetro.toracico*, que sube del 84 % en el caso de las hembras cerradas hasta un 90 % en el caso de observar todo el rebaño sin distinción de edad o sexo, también destacan las otras correlaciones, ya que en el caso de la *anchura.caña* y *perimetro.caña*, parece obvio que tuvieran alguna correlación (85 %), ya que pertenecen ambas a la misma parte del animal.

En el caso de la *altura.medio.dorso* y la *altura.grupa* se puede observar que salvo algunos valores dispersos es bastante uniforme, con un coeficiente de 87 % que es bastante bueno. La *altura.cruz* también tiene una buena correlación con la *altura.medio.dorso*, del 84 %, si bien es cierto que no es tan buena como la mostrada en el caso anterior.

En la Tabla 5.8 se pueden observar las ecuaciones que tienen todas las correlaciones así como el R^2 de cada una de ellas. Para hallar las ecuaciones se ha utilizado el coeficiente de correlación (b), que multiplica a la x, en este caso la x viene dada por la variable que se encuentra a la derecha en la relación. Para hallar la a, que es el valor constante en la ecuación, se han utilizado las medias del *peso* (55.71), la *altura.caña* (25.75), el *perimetro.toracico*

Correlación	Ecuación	R^2
Peso-Perimetro.toracico	$\hat{y} = -23{,}81 + 0{,}90x$	0.8852
Anchura.caña-Perimetro.caña	$\hat{y} = 17{,}963 + 0{,}85x$	0.6025
Altura.cruz-Altura.medio.dorso	$\hat{y} = 13{,}907 + 0{,}84x$	0.9093
Altura.grupa-Altura.medio.dorso	$\hat{y} = 11{,}239 + 0{,}87x$	0.9434

Cuadro 5.8: *Tabla de las mejores correlaciones globales en la base de datos de las cabras.*

Ecuación	R^2 **Total**	R^2 **Hembras cerradas**
$\hat{y} = -23{,}81 + 0{,}90x$	0.8852	0.8492
$\hat{y} = -17{,}8043 + 0{,}84x$	0.8656	0.8308

Cuadro 5.9: *Comparativa correlación perímetro torácico - peso.*

(87.80), el *perimetro.caña* (9.163), la *altura.cruz* (74.635), la *altura.grupa* (74.136) y la *altura.medio.dorso* (72.295).

Se puede observar como los resultados de la estimación del peso mediante el perímetro torácico es considerablemente buena, ya que R^2 es de 0.8852 cifra que es ligeramente peor que la obtenida en la correlación en las hembras cerradas, se va a aplicar la fórmula de las hembras cerradas en el total de la base de datos, para determinar si se obtienen mejores estimaciones, también se aplicará la función extraída en este caso a la sección de las hembras cerradas. Los resultados se pueden apreciar en la Tabla 5.9.

Cabe destacar que la función que utiliza toda la base de datos obtiene mejores valores en el R^2 tanto globalmente como localmente que la que utiliza solamente la sección de las hembras cerradas y que dicha diferencia se amplía según se amplía el conjunto de datos. Por lo que se puede admitir que es un buen estimador ya que a mejor cantidad de datos mejores resultados obtiene, obteniendo mejores resultados cuanto mayor es la base de datos, tanto en la función de las hembras cerradas como en la función global.

También en el caso de otras 2 correlaciones se obtienen buenos resultados, ya que en el caso de en el la altura cruz con la altura de medio dorso un R^2 de 0.9093 y entre la altura de la grupa y la de medio dorso uno de 0.9434. En cambio la relación entre la altura de la caña y el perímetro de la caña no obtiene buenos resultados teniendo un R^2 de 0.6025, valor que es considerablemente bajo y por lo tanto no se le puede considerar bueno.

Capítulo 6

Aplicaciones a la base de datos de la gasolina

En el capítulo siguiente se van a estudiar las relaciones entre las variables de la tabla de repostajes, concretamente la de gasolina95. Observando la Figura 6.1, que muestra las correlaciones entre las variables de las bases de datos de gasolina y diesel, se puede apreciar que en el caso de la base de datos de *Gasolina* se puede observar una buena correlación entre euros y litros, de un 89 %, lo cual puede hacer pensar que tienen una estrecha correlación. Pero al observar la base de datos de *Diesel* se puede apreciar que la correlación no es tan buena, ya que es de un 80 %. Siendo ambas bases de datos de repostajes y teniendo la base de datos de diesel mayor número de registros que la de gasolina95, se puede admitir que la correlación no es tan buena como en una primera instancia pudiera parecer, ya que con el doble de datos se ha reducido casi un 10 % la correlación, cuando lo que hubiera sido realmente deseable es que dicha cifra hubiera ido en aumento, o al menos que se hubiera mantenido.

Pensando detenidamente en ello y conociendo la situación, se sabe que existe una relación entre los euros pagados y el combustible repostado, el problema es que dicha relación es lineal, pero no estacionaria, ya que en función del tiempo varía de manera impredecible. Esto hace que según se contenga un mayor número de registros se reduzca la correlación y por lo tanto que dicha estimación del dato no tenga una de las propiedades más interesantes

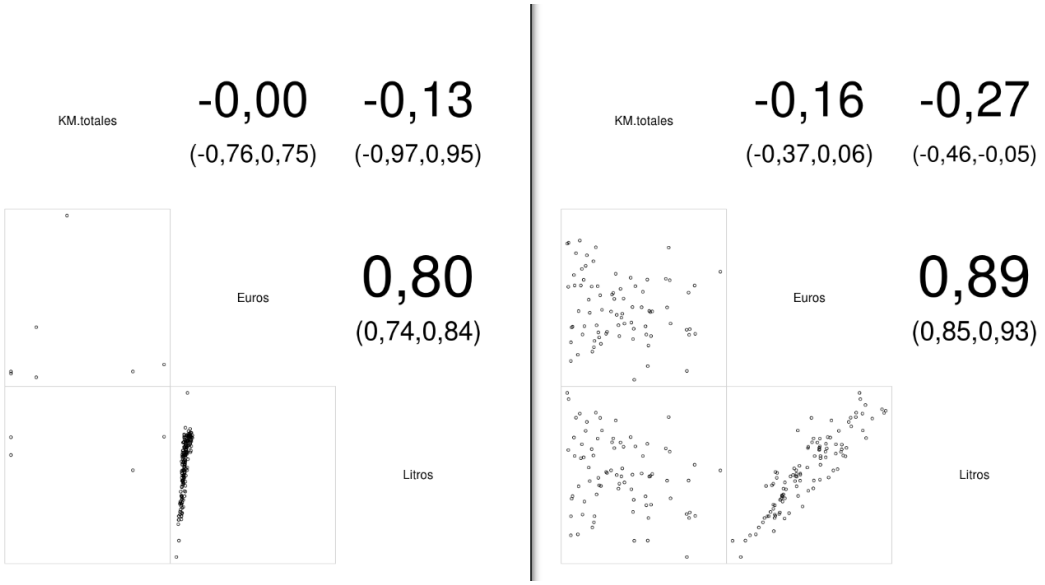


Figura 6.1: *A la izquierda las correlaciones del diesel y a la derecha de la gasolina*

que es la de la consistencia.

Por ello en el caso de estas dos variables no se van a poder utilizar las correlaciones, pero esto se ha sabido aplicando, de manera mínima, un conocimiento experto, ya que se sabe que sí existe una relación, por lo que a continuación se va a seguir esa línea en la búsqueda de una buena estimación.

6.1. Sistema experto

Para seguir con la línea del sistema experto y poder aplicarlo a las bases de datos de los carburantes es necesario buscar relaciones posibles entre las variables. En este caso existen relaciones demostradas, como que cuanto mayor sea la cantidad de litros repostados más kilómetros se pueden hacer o como que si en un período de tiempo pequeño se pagan dos cantidades diferentes es muy probable que se hayan repostado más litros en el pago mayor.

Pero eso depende de otros factores, por ello hay que estudiar dichas relaciones, que es lo que se realiza a continuación.

6.1.1. Relación Euros/Litro

Es lógico pensar que existe una relación entre los euros gastados y los litros repostados, si bien es cierto que no es una correlación porque dicha relación es aleatoria con el tiempo, pero se puede hallar de manera local, de modo que el precio del carburante en un registro tiene un valor cercano al del registro anterior y al del registro siguiente.

También se puede tomar como referencia el precio del carburante, que será la relación Euros/L. Es importante destacar que no se busca el precio del crudo, que no tiene por qué ajustarse en absoluto al precio del litro, sino que se necesita el precio de venta al público, que es el precio que tiene en las gasolineras.

Dichos datos se extraen de la página web del Ministerio de Industria, Energía y Turismo [INE,]. Los datos están agrupados por meses, por lo que hay que agrupar los datos propios también en meses de cara a realizar la comparación en la misma escala. Los datos del Ministerio son las medias de precio de combustible por comunidades autónomas, por lo que en el caso de tener varios datos en el mismo mes se ha de calcular la media. La comunidad autónoma que interesa en este caso es la de Madrid que es donde se han realizado los repostajes.

Existe un problema y es que los distintos repostajes del coche bien se han podido realizar en distintas gasolineras que a su vez pueden tener mayor o menor relación Euros/L y que muy probablemente será distinta de la media por comunidad autónoma o sencillamente que se han realizado en otra comunidad, por lo que los datos a estimar no solo pueden basarse en los datos extraídos del ministerio. Para confirmar esto se van a comparar los datos del ministerio con los de las bases de datos, solamente los conocidos ya que se quiere ver la semejanza tanto para la gasolina como el diesel.

Como lo que interesa es el precio del carburante en ambos casos se ha decidido omitir la columna de *KM.totales* ya que en en el caso de haber NAs en dicha columna, pérdida de datos, al contemplar solamente las muestras completas, sin NAs, la eliminaría y realmente no afecta para hallar el precio del combustible ya que solo se necesita el coste, litros repostados

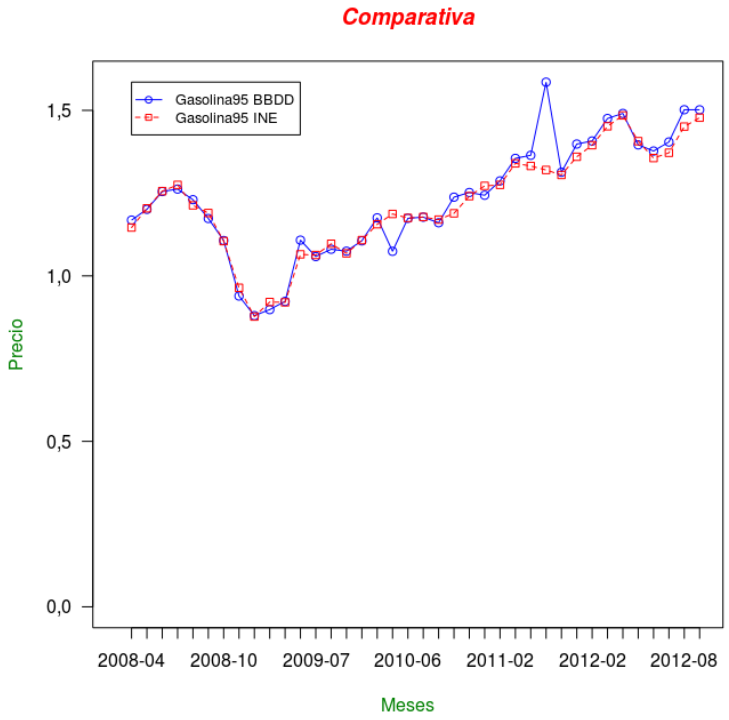


Figura 6.2: *Comparativa precios gasolina95 2008-2012*

y la fecha. Además en el caso de la tabla de diesel el precio del carburante varía en función de si es diesel normal o *e+*, para lo cual hay que realizar una función que elimine los repostajes con *e+*, ya que interesan aquellos repostajes realizados con diesel normal.

Para ello dicha función localizará aquellas filas que tengan en la variable *tipo* el valor *e+*. A continuación se eliminan las columnas no deseadas de la tabla, dejando solamente las tres mencionadas anteriormente, *euros*, *litros* y $fecha$ y eliminará los NAs. Los resultados obtenidos se muestran en las figuras 6.2 y 6.3.

Como puede verse la base de datos de la gasolina es considerablemente buena, ya que se obtiene un ECM de 0,1545 Euros con un fallo mínimo de 0,1127 Euros y un fallo máximo de 0,2654 Euros siendo números muy pequeños para considerar fiables los datos y más teniendo en cuenta que la gráfica de los datos es prácticamente idéntica. En cambio en el caso del diesel, no se da la misma situación y es que los resultados obtenidos no se ajustan a los resultados dados por el Ministerio. Esto tiene una sencilla explicación, y es que los datos

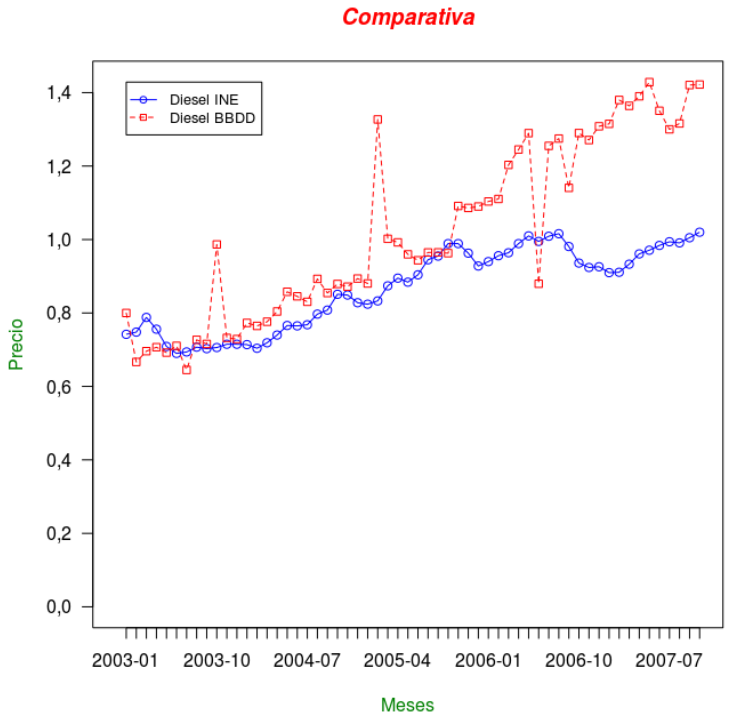


Figura 6.3: *Comparativa precios diesel 2003-2012*

del INE son las medias del precio en las gasolineras, al haber gasolineras que difieren en los precios es por lo que es fácil que no cuadren a la perfección. De hecho si se ha ajustado tanto la base de datos de gasolina se debe o bien a que el precio de la gasolina varía menos entre gasolineras que el del diesel, o que se trate de algo casual, pero lo más corriente es que los datos no se ajusten tanto.

Por lo tanto se puede utilizar sin problemas el precio real de la gasolina para hallar el coste que tendrá en cierto momento en una gasolinera y así conseguir una mejor estimación de los datos, si bien es cierto que en cálculos futuros, se debe revisar periódicamente que los precios del INE sean similares a los obtenidos por los vehículos y asegurar así que sigan siendo buenos. En cambio, en el caso del diesel, no se consideran suficientemente fiables por lo que hay que utilizar otras técnicas.

Cabe destacar que la relación Euros/L es válida en ambos sentidos, esto es que es válida tanto para estimar litros a partir de euros, como para estimar euros a partir de litros.

Para obtener las estadísticas de este punto, una vez se han cargado todos los datos, solamente hay que escribir el comando

$comparativaCosteMensual(tablaGasolina = NA, tablaDiesel = NA)$

Este comando muestra la comparativa de los datos de la siguiente forma, si no se introducen datos se compararán las dos bases de datos del INE, tanto la de gasolina 95 como la de diesel; en el caso de introducir datos en el parámetro *tablaGasolina*, comparará la base de datos introducida en ese parámetro con los datos de gasolina del INE funcionando de manera análoga en el caso de *tablaDiesel*. En el caso de introducir una base de datos en cada parámetro comparará ambas.

6.1.2. Relación Kilómetros/Litros

A continuación se evalúa si realmente existe una relación entre los litros repostados y los kilómetros recorridos, en el Punto 3.4 se explica como se hallan. En caso de existir dicha relación posibilita el uso de la Fórmula 3.3, lo que es muy interesante de cara a poder mejorar la calidad del estimador, aunque no fuera la mejor opción, lo cual se ve más adelante, permite tener mayor versatilidad lo cual resulta muy interesante.

En la Figura 3.3, como era de esperar la correlación entre los *euros* y los *km* es baja, incluso nula, con un coeficiente de correlación del 0 %. Además se puede apreciar que la correlación entre los *litros* y los kilómetros es muchísimo peor de lo que se esperaba, es decir que fuera una buena correlación y se ha encontrado una que está en torno al 7 %, lo que implica que no hay ningún tipo de correlación entre ellos.

Esto puede deberse a que no siempre se realiza un repostaje en el preciso momento en que el depósito tiene un nivel de carburante muy bajo, sino que muchas veces se reposta aunque no sea así, haciendo que hallar los kilómetros recorridos mediante el uso de una correlación en esta base de datos sea complejo.

En este caso la presencia de outliers puede hacer que el dato estimado sea desmesuradamente erróneo, por ello, se ha decidido utilizar el consumo medio, en esta situación es

04-2008	05-2008	06-2008	07-2008	08-2008	09-2008	10-2008
10.1436012	8.2262830	6.2884354	7.8765152	7.9583565	6.9097938	15.1679389
11-2008	12-2008	02-2009	03-2009	06-2009	07-2009	08-2009
11.0158730	3.6348140	9.9657321	2.1122924	8.4986877	8.5025604	8.2312860
09-2009	03-2010	04-2010	06-2010	07-2010	08-2010	11-2010
3.8200873	11.0606287	1.9526221	4.6576763	7.6727273	2.8292968	7.8842105
12-2010	01-2011	07-2011	08-2011	11-2011	08-2012	
7.6012208	0.9960248	4.9007092	3.3128917	0.5806485	NA	

Figura 6.4: *Consumo medio por mes de la base de datos de Gasolina95*

necesario calcular los kilómetros totales, ya que los kilómetros no forman parte de la base de datos original y realizar dicho cálculo en función de las otras dos variables posibles, que son euros y fecha puede ser muy precario y propiciar un error alto, en cambio en cierto modo regirse por el consumo medio, en circunstancias normales puede ofrecer un error relativamente bajo y por lo tanto se considera la mejor alternativa.

El consumo medio se pretende hallar por meses y solamente en el caso de que no se obtengan valores satisfactorios se hallaría el global. Ya que en circunstancias normales un vehículo incrementará su consumo con el paso del tiempo y por lo tanto puede ser interesante tenerlo en cuenta, siempre y cuando los datos ofrecidos sean satisfactorios. Los resultados se muestran en la Figura 6.4, como puede observarse no son satisfactorios, ya que todavía contiene valores muy dispersos, como es la diferencia entre Diciembre de 2010 (7.6012208) y Enero de 2011 (0.9960248), donde claramente este último dato es erróneo, ya que una variación tan alta en un solo mes no es común y menos que un vehículo, no híbrido tenga el consumo de Enero de 2011.

Por lo tanto se va a hallar el consumo medio total, usando estos consumos medios de los meses, al hacerlo se obtiene un consumo medio de 6.607727, lo cual es un consumo medio posible, de hecho se ajusta al patrón del manual del coche, por lo tanto se toma como consumo medio base para la base de datos de Gasolina95.

En la tabla de diesel en cambio no es posible hacer esto ya que los *km.totales* de los que se dispone son un número de datos muy escaso (11) lo que hace que no sea conveniente

tomarlo como referencia, ya que puede llevar a conclusiones erróneas. [1].

6.2. Resultados del comparador de estimadores

En el Punto 3.7 se introdujo un método para comparar estimadores basado en la pérdida aleatoria de datos (en %). Para comenzar a comparar los métodos citados anteriormente es necesario usar el comparador de resultados, ya que como bien se ha explicado anteriormente hay que realizar diversas iteraciones, realizar diversas comparaciones y comprobar varios resultados de cada método de estimación como son el tiempo computacional y el ECM.

Haciendo uso del comparador, se obtienen los resultados mostrados en la Tabla 6.1. Los resultados aportan la información citada anteriormente, el tiempo de cómputo medido en milisegundos y el error cuadrático medio (ECM), que es una medida tanto para la insesgabilidad como para la eficiencia.

De manera externa al comparador se deberá realizar un cálculo de la consistencia y bajo el propio criterio de los desarrolladores se debe estudiar la suficiencia.

Estos resultados se han aplicado sobre cada variable en un contexto con 1000 iteraciones, sin valores ausentes en el resto de variables y sin usar ningún orden de estimación específico ya que al tener datos ausentes en una única variable da igual el método que se utilice, se estiman en el mismo orden. Los métodos de estimación utilizados han sido testados en la base de datos completa más grande de la que sea dispone, entre la de diesel y la de gasolina95, dicha base de datos es la de Gasolina95 con un total de 80 registros siendo una cifra considerablemente baja de registros. Además a dichos datos se le ha aplicado una pérdida de información del 10 %.

En términos de suficiencia, se cosidera que los datos han sido utilizados teniendo en cuenta todas las posibles relaciones entre variables, ya que se extrae información de 3 tipos de relaciones en un entorno que contiene solamente 4 variables, que permitiría un máximo de 6 relaciones. Además no es solamente que se hayan utilizado 3 relaciones, sino que las

[1] De hecho la vía para determinar el consumo medio usada anteriormente devuelve un valor NA, lo que verifica lo citado anteriormente

Variable	Tiempo	ECM
Litros [Euros]	8.688	1.135
Litros [Consumo]	27.534	32.23
Euros [Litros]	8.688	1.036
KM.totales [Consumo]	18.114	1226

Cuadro 6.1: *Resultados medios obtenidos con el comparador para 1000 pruebas.*

otras 3 han sido descartadas por ser menos fiables o innecesarias que las 3 existentes, como puede ser el caso de la relación entre los euros y los kilómetros, no tiene sentido tener en cuenta dicha relación cuando tendría que hacer uso de las relaciones de los litros con ambas variables.

Los resultados obtenidos reflejan que la estimación de los litros utilizando los euros es con mucha diferencia mejor que la estimación de los litros mediante el uso del consumo, esto es debido al ajuste tan fuerte que tienen los datos del INE y los datos reales de la base de datos de la gasolina, ya que son prácticamente iguales.

En el caso de los euros también obtiene buenos resultados, esto es evidente dado que se utiliza la misma relación que en el caso anterior, de los litros con los euros y se sigue ajustando a los datos que se tienen del INE.

En el caso de los kilómetros totales, se puede apreciar una subida considerable, ya que tarda más del doble que en los casos anteriores, pasando de 8.688 en los dos anteriores a 18.114 en este caso. Esto es debido al mayor uso de datos, ya que en el caso del cálculo de los litros mediante el consumo se puede observar un incremento en el tiempo de más de 3 veces más (27.534), y los resultados obtenidos son peores, por lo que hace que sea mayor la diferencia entre esta forma de estimar los litros y estimarlos mediante los euros.

Al aplicar una única función de estimación a cada valor puede darse el caso de que, tal y como sucede en la Figura 6.5 no pueda hallarse una estimación de los litros utilizando los euros como única referencia y viceversa. Es el caso de los registros 94 y 102, que tienen ambas variables con valores NA. En ese caso para asegurar el correcto funcionamiento del estimador, es necesario hacer uso de otras estimaciones, de modo que aseguren que siempre

	Litros	Euros	KM	Fecha
90	37,25	51,52	48032	2012-01-28
91	36,27	49,47	48329	2012-02-10
92	46,19	66,61	48623	2012-02-24
93	27,65	40,01	48998	2012-03-09
94	NA	NA	49197	2012-03-25
95	21,21	31,86	49519	2012-03-30
96	28,21	41,74	49694	2012-04-01
97	30,70	38,03	49929	2012-04-03
98	30,21	44,01	50186	2012-04-06
99	18,30	28,16	50439	2012-04-08
100	32,67	50,08	50592	2012-04-09
101	27,08	33,42	50827	2012-04-17
102	NA	NA	51055	2012-04-28
103	44,40	61,98	51440	2012-05-19
104	52,94	65,34	51816	2012-06-03
105	38,33	47,49	52215	2012-06-22
106	36,67	51,23	52544	2012-07-08
107	43,03	53,11	52861	2012-07-23
108	39,39	48,61	56871	2012-08-24
109	18,41	22,81	57151	2012-09-07

Figura 6.5: *Ejemplo de un caso en el que una estimación euros litro no puede hallarse*

se pueda conseguir todos los datos aunque falten algunos con los que se pudieran conseguir muy buenas estimaciones, de este modo el uso de estimaciones algo peores es necesario.

Pero aún así en ciertas circunstancias puede suceder que los resultados devueltos sean valores NA, esto no tiene por qué ser un fallo del estimador en sí, sino que puede darse el caso de que no sea posible hallar una estimación suficientemente coherente para cierto dato. Este es el caso de la última línea en el caso de que no se tenga el valor ni de los euros ni de los litros, sean datos NA. En esa circunstancia además de no tener los euros gastados en el repostaje no se sabe cuantos kilómetros se han recorrido con ese depósito, ya que no se tiene el dato del siguiente cuentakilómetros.

Para asegurar que no se dan estas situaciones, se va a evitar eliminar la última fila, de tal modo que siempre haya datos finales para tomar las referencias.

Por ello se incluyen dos modificaciones agregando las dos funciones restantes a los estimadores, de modo que se usan las 5 Fórmulas del punto 3.4, además para asegurar que se pueden estimar todos los datos, se realizará una modificación y es que se alterará el orden de las columnas, de modo que se hallará primero la columna de los kilómetros totales, a continuación la de los litros y finalmente los euros. De este modo se asegura que las esti-

Variable / Orden estimación	Por variables	Cronológicamente
Litros	104.60	374.5
Euros	137.10	557.3
KM.totales	5285000	1125000

Cuadro 6.2: *Resultados medios obtenidos con el comparador para cada tipo de orden de estimación.*

maciones se van a llevar a cabo ya que al menos estarán las fechas a la hora de calcular los kilómetros totales y por estimarse después también es seguro que estan los kilómetros totales al estimar los euros.

Aplicando esto, los resultados obtenidos para los 3 órdenes de estimación son los mostrados en la Tabla 6.2.

Estos resultados muestran las diversas medias de ECM de cada uno de los métodos usados para estimar los datos, se puede apreciar que los resultados en general son muy elevados, ya que son medias de las 1000 iteraciones y no la acumulación de las mismas. Siempre hay que tener en cuenta que el ECM toma valores en función de la magnitud del error que es proporcional a la magnitud de la variable con la que se está trabajando. Por eso en el caso de los kilómetros totales son tan elevadas, pero la estimación por variables obtiene una estimación muy lejos de ser buena, ya que tiene un ECM medio de 5285000, mientras que al estimar en el orden cronológico se puede apreciar que el error es de 1125000, que es algo más de una quinta parte del error anterior, esta diferencia puede ser debida a que en el primer caso lo primero que se estima son los kilómetros y por lo tanto es sensible a fallo, ya que se estima aunque no haya buenos datos para ello.

Los litros estimados por contraposición son mejores para la estimación por variables (104.60), ya que la estimación cronológica (374.5) no ofrece buenos datos en este punto. También en los euros se puede apreciar como siguen siendo considerablemente mejores las estimaciones por variables, ya que supone una estimación mucho más precisa, que la cronológica, ya que la primera ofrece un ECM de 137.10 mientras que la segunda aporta uno de 557.3, siendo al igual que en el caso anterior una diferencia notable.

Se esperaban tiempos similares entre las estimaciones halladas por variable (49.255) y las estimaciones halladas cronológicamente (116.244), los cuales no se cumplen ya que el del orden cronológico asciende hasta más del doble del tiempo empleado para la estimación por variables. Esto puede deberse a que las estimaciones cronológicas puedan necesitar un mayor acceso a los datos, ya que en el caso de las variables con un solo acceso se estiman tantos elementos como NAs tenga la columna mientras que calculándolo por filas es necesario entrar en tantos métodos como NAs existan.

Por lo tanto se puede afirmar que las estimaciones por variables aportan mejores resultados, pese a que puntualmente en una de las variables consigue mejores resultados las estimaciones cronológicas. Este resultado es un dato exclusivo de la base de datos y depende de la naturaleza de la misma, por lo que los usuarios deberán determinar cuál utilizar en función de su base de datos.

Los resultados obtenidos se pueden aplicar sobre la base de datos de diesel, pero no es posible realizar evaluaciones sobre esta base de datos ya que el número de registros totalmente completos se reduce drásticamente a 4, ya que en total la tabla de diesel tiene una pérdida de datos porcentual de un 64 %, cifra que si se toma la referencia de lo citado anteriormente, es demasiado elevada como para poder verificar los datos estimados.

6.3. Guardar los datos

Finalmente se pueden almacenar los resultados obtenidos utilizando la siguiente función.
$guardaDB(input, output = "default.csv", formato)$
La cual permite almacenar un determinado *input*, que es la base de datos a almacenar, en el archivo dado por el *output* que en el caso de no introducir ningún valor toma como valor por defecto $default.csv$.

Se puede apreciar un ejemplo del guardado de datos en la Figura 6.6, en la cual al introducir un archivo con extensión *.csv*, hace que no sea necesario que se introduzca el parámetro $formato$.

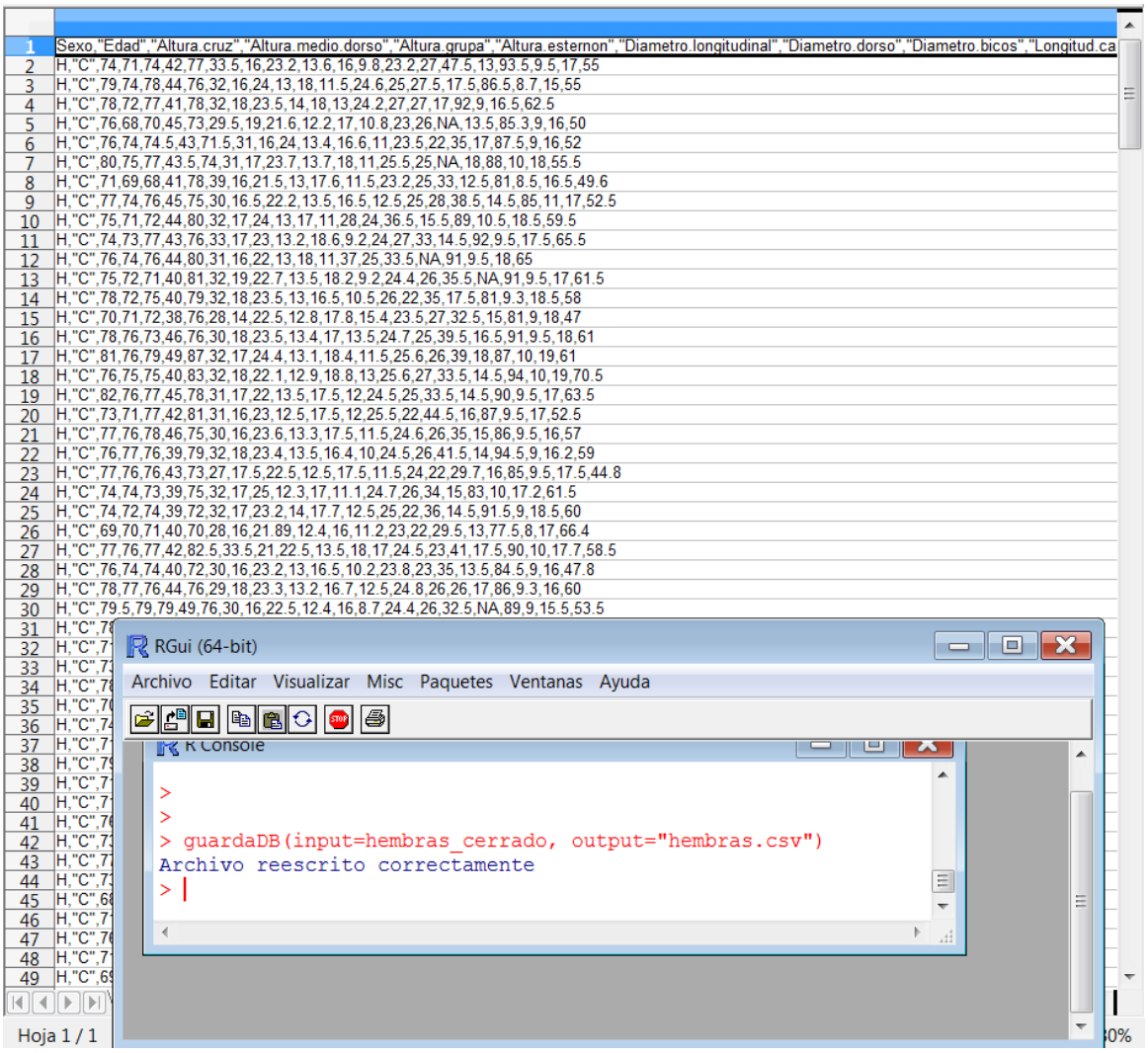


Figura 6.6: *Muestra del resultado de guardar un archivo en formato CSV.*

Capítulo 7

Conclusiones y trabajos futuros

El sistema de carga desarrollado en este trabajo tiene un funcionamiento correcto si bien es cierto que puede no ser todo lo versátil que se desearía.

En las modificaciones realizadas sobre las caras de Chernoff, se obtienen muy buenos resultados, ya que se consiguen localizar los outliers y agregar a las caras de una nueva funcionalidad que anteriormente no tenía.

Las correlaciones obtenidas son considerablemente buenas y permiten pensar que si hubiera un mayor número de elementos se podrían conseguir mejores correlaciones, ya que al aumentar el número de individuos se ha conseguido mejores correlaciones, se confirma que en caprinos también existe una correlación entre el peso y el perímetro torácico al igual que en los vacunos. Las correlaciones de la anchura de la caña con el perímetro de la caña, la altura de la cruz con la altura del medio dorso y esta última con la altura de la grupa, pese a no ser tan relevantes como el peso con el perímetro torácico, tienen un coeficiente de correlación alto y obtienen unos resultados considerablemente buenos.

En el caso de las bases de datos de carburantes se han descartado posibles correlaciones, así como la aplicación de un sistema experto para estimar datos. Para evaluar dicho sistema experto se ha necesitado usar el comparador de estimadores. En la gasolina los resultados obtenidos son considerablemente peores de lo esperado, siendo la estimación por variables la que ofrece mejores resultados, mientras que en el diesel el tener muy pocos valores de la variable *km.totales* lastra sus resultados, haciendo que no sea posible evaluarlos de manera

correcta.

El comparador de estimadores ha sido probado con la estimación del experto de los carburantes, ofreciendo resultados satisfactorios y suficientemente completos como para que un usuario pueda determinar qué comprador es mejor.

Para finalizar el trabajo y con perspectiva de futuro, este trabajo no tiene por qué representar un trabajo aislado, ya que tiene muchas variantes y cada una de ellas tiene ramas por las cuales se puede continuar con la investigación Manteniendo el orden citado en los capítulos anteriores, lo idóneo es comenzar hablando de la carga de archivos.

Como trabajo futuro existe la opción de ampliar las extensiones de bases de datos que aceptamos, así como aceptar un mayor número de formatos de datos, ya que en estos momentos se cubren necesidades básicas para que los usuarios puedan utilizar la biblioteca, permitiendo cargar tres tipos de bases de datos, pero no cubre todas las posibilidades. El objetivo de esto es ganar mayor versatilidad y así poder abarcar un mayor espectro del mercado con el fin de que la biblioteca sea más utilizada en el sector. Para ello, la librería de carga de datos XLConnect permite también cargar datos de tipo SPSS y Stata, también sería interesante la carga de PSPP que es la versión de código libre de SPSS.

En el ámbito de las caras de Chernoff, lo más interesante es encontrar la mejor combinación de 31 colores posibles, que ofrezcan la mejor precisión, para lo cual habrá que reordenarlos y en algunos casos en los que se confundan con frecuencia modificarlos. Una vez conseguido esto, puede ser interesante ampliar el número de variables, incluyendo categorías como pueden ser las orejas o el pelo. También puede ser interesante evaluar si una reordenación de las variables de las caras puede ayudar a obtener mejores resultados, esto sería que las primeras 5 variables sean referentes al contorno de la cara, boca, nariz, ojos y cejas pretendiendo que sea más sencilla la interpretación de las caras, consiguiendo así que las combinaciones más complejas de valores atípicos se de solamente cuando el número de variables es muy elevado.

Continuando con las correlaciones, es necesario una base de datos mayor para poder

confirmar las correlaciones vistas en este trabajo, ya que el número de individuos en este caso puede ser escaso. También puede ser interesante evaluar si existe una correlación que usando dos variables entre altura cruz, altura medio dorso y altura grupa consiga buenos resultados estimando la otra restante.

Antes de cerrar con la base de datos de las cabras sería interesante obtener registros periódicos de estas, ya que en función de la evolución de las mismas se podrían localizar si la cabra está preñada o incluso enfermedades como la teniasis, enfermedad que se representa con una progresiva pérdida de peso en caprinos acompañada de una pérdida de apetito, por lo que el sistema podría notificar al ganadero de posibles casos de teniasis para que el ganadero haga un seguimiento a la cabra en cuestión y así ayudar a localizar los casos antes. Además en el caso de tener una base de datos de cabras con un número elevado de registros, es planteable continuar con un desarrollo específico de las caras de Chernoff, de tal modo que en vez de caras dibuje cabras.

En las tablas de datos de carburantes, es interesante ampliar los datos, al igual que en el caso anterior, para lo cual se puede hacer uso de una aplicación web que aporte varios servicios a los usuarios y almacene los repostajes. Con una mayor cantidad de datos de un vehículo se podría intentar buscar consumos anómalos y avisar a los propietarios con el fin de localizar las averías con mayor antelación y por lo tanto reducir el gasto de las reparaciones.

En el comparador de estimadores se podrían incluir estimaciones por intervalo y utilizarlo para medir el máximo de pérdida de datos posible que mantenga los datos en cierto intervalo de confianza. Así como agregar más métodos para comparar ambos estimadores, como puede ser en función del porcentaje de datos perdidos ya que uno con pequeñas pérdidas de datos puede aportar resultados muy buenos y con pérdidas mayores aportar resultados malos.

También se planteado ampliar el orden de las estimaciones, con un orden que que consistiría en estimar los datos solamente cuando se pueda realizar la mejor estimación posible y en caso contrario seguir estimando otros datos, con la esperanza de que esos datos conseguidos

permitan realizar dicha estimación sobre los datos aún no estimados.

Finalmente y como punto más importante, este trabajo debe incluirse en la revisión final de la librería BioSeq ya que es el objetivo principal desde un primer momento.

Bibliografía

[Adler et al., 2010] Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J., and Zucchini, W. (2010). ff: Memory-efficient storage of large data on disk and fast access functions. *R package version*, pages 2–2.

[Brazma et al., 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., et al. (2001). Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371.

[Chernoff, 1973] Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

[DDBJ,] DDBJ. Ddbj website.

[EBI,] EBI. Ebi website.

[ENSEMBL,] ENSEMBL. Ensembl website.

[García Lara et al., 2009] García Lara, I., Ferreño, V., Fernández Calviño, E., Vidal Galego, L., Lara, G., and de Jesús, M. T. (2009). Ecuaciones de predicción del peso vivo de hembras holstein. *Frisona española*, 29(171):90–95.

[Gardner et al., 2011] Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., et al. (2011). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic acids research*, 39(suppl 1):D141–D145.

[Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open

software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

[Gilbert, 2005] Gilbert, D. (2005). Biomolecular interaction network database. *Briefings in bioinformatics*, 6(2):194–198.

[GNULicense,] GNULicense. Gnulicense.

[Golemis and Adams, 2002] Golemis, E. and Adams, P. D. (2002). *Protein-protein interactions: a molecular cloning manual*. Cold Spring Harbor Laboratory Press New York.

[INE,] INE. Precios carburante.

[MacBeath and Schreiber, 2000] MacBeath, G. and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763.

[MetaBase,] MetaBase. Metabase website.

[MINSEQE,] MINSEQE. Minseqe website.

[NCBI,] NCBI. Ncbi website.

[Pearson, 1896] Pearson, K. (1896). Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60(359-367):489–498.

[PRIDE,] PRIDE. Pride website.

[PubMed,] PubMed. Pubmed website.

[Rustici et al., 2013] Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., et al. (2013). Arrayexpress update—trends in database growth and links to data analysis tools. *Nucleic acids research*, 41(D1):D987–D990.

[Salazar Mendoza, 2011] Salazar Mendoza, M. I. (2011). *Aproximación bayesiana a los contrastes de hipótesis múltiples con aplicaciones a los microarrays*. PhD thesis, Universidad Complutense de Madrid, Servicio de Publicaciones.

[Sánchez, 2013] Sánchez, Ó. (2013). Algoritmo de programación dinámica con r para resolver problemas de alineamientos de secuencias.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

[UKBiobank,] UKBiobank. Ukbiobank website.

[Uniprot,] Uniprot. Uniprot website.

[XLConnect,] XLConnect. Xlconnect website.