Universidad de Valladolid

ESCUELA DE INGENIERÍAS INDUSTRIALES

UNIVERSIDAD DE VALLADOLID

ESCUELA DE INGENIERIAS INDUSTRIALES

Grado en Ingeniería Mecánica

Estimación de Residuos derivados de la Construcción y Demolición en Obra Civil

Autor:
Hernández González, Francisco Javier

Tutor:
Sainz Palmero, Gregorio I.
Dpto: Ingeniería de Sistemas y Automática

Valladolid, julio 2017.

RESUMEN

A partir de la información recopilada en distintas obras civiles referente a residuos derivados de construcción y demolición (RCD), en este trabajo se pretende analizar las relaciones existentes entre los diferentes tipos de residuo y formular modelos que permitan realizar estimaciones de los residuos que se pueden generar en una obra.

El análisis se inicia con una fase de preprocesamiento, que sirve para tratar de subsanar o al menos paliar los efectos de los datos faltantes y *Outliers* existentes en los datos iniciales. Posteriormente, se realiza la selección y extracción de variables mediante la técnica de análisis de componentes principales (PCA), determinando así el punto de partida para la aplicación de los modelos de estimación.

Como resultado de esta metodología se ha obtenido un conjunto de modelos que relacionan RCDs, permitiendo obtener la estimación de los mismos. Estas estimaciones resultan de gran interés, desde distintos puntos de vista, para las empresas de construcción y para los entes de gestión de residuos.

PALABRAS CLAVE

Residuos de Construcción y Demolición (RCD), Redes Neuronales artificiales (RNA), *Principal Analysis Component* (PCA), Correlación, Estimación.

Índice

Índice de Anexos

1. INTRODUCCIÓN

En este capítulo introductorio se expondrá la motivación por la cual se ha realizado este proyecto, los métodos existentes en la actualidad para tratar problemas similares, y los objetivos que se desean cumplir con el mismo.

1.1. Motivación del proyecto

La generación de residuos en la construcción y demolición (RCD) de edificaciones supone un gran impacto ecológico en nuestra sociedad, y una mala gestión de estos, deriva directamente en un alto cargo económico para los responsables de la obra.

Según datos del informe técnico realizado por la Federación Española de RCD, La producción de residuos derivados de la construcción y demolición de edificios, en España ha descendido progresivamente debido a la bajada generalizada del sector de la construcción con un descenso del 56% en el periodo 2009-2013.

Pese a esta bajada de actividad, el ratio de producción de estos residuos ha alcanzado una cota de 0,43 t/hab/año en 2013, siendo el vertido incontrolado de residuos el principal problema. Se estima que el 49% de la producción de RCD se ha depositado en lugares no autorizados durante los años 2009 a 2013. Debido a la presión ejercida por los ayuntamientos, la tendencia de estos vertidos es decreciente, la creación de nuevas instalaciones de reciclaje, junto a la concienciación social de la población ha logrado reducir el vertido incontrolado de estos residuos, situándose en 2013 en un 36% de la producción. (Fueyo, 2015)

Los principales problemas que resultan de la gestión de RCD son (Ayuso, 2011):

- Contaminación:
  Ya sea de tipo visual, acústica o atmosférica.

- Aumento de la siniestralidad en la obra:
  La necesidad de mantener un adecuado nivel de limpieza y orden en la ubicación donde se desarrolla la obra, afectando al ritmo de trabajo y a la seguridad de los trabajadores.

- Aumento de costes:
  Debidos a una mala planificación de escombreras, contenedores y zonas de almacenaje.

Con el fin de fomentar la prevención, reutilización, reciclado y, en definitiva, contribuir a un desarrollo sostenible de la actividad de construcción, en 2008 se aprobó el Real Decreto 105/2008. Esta norma, entre otra serie de obligaciones, impone tanto al productor como al poseedor del RCD, el incluir en el proyecto de obra un estudio de Gestión de Residuos de construcción y demolición en el cual deberá incluirse la estimación en metros cúbicos y toneladas de los RCD que se generarán en la obra.

Se considera que el primer paso para afrontar los problemas derivados de la generación de RCD es la estimación de estos. Una mala estimación de las cantidades y volúmenes de los RCD puede suponer un grave entorpecimiento de la obra, así como la pérdida de las fianzas, además del incremento de los trámites de y para las administraciones locales. Este es el punto de partida de este trabajo fin de grado, estudiar y analizar la metodología necesaria para a partir de datos recabados en obras, poder hacer estimaciones de RCD en una obra nueva.

1.2. Herramientas para estimar RCD

En la actualidad existe una gran cantidad de modelos y herramientas informáticas con las que estimar la producción de RCD, en general se tratan de aplicaciones enfocadas al cumplimiento de los requisitos legales del Real Decreto 105/2008. Así, podemos encontrar tres tipos de modelos. Cabe decir que los modelos listados a continuación no son los únicos, pero son los más eficaces y de más fácil acceso para el autor.

Modelo detallado

En este tipo de modelo se realiza una previsión del residuo que se va a generar en función de las mediciones del proyecto. Es el proyectista quien estima las cantidades y el tipo de RCD sirviéndose de las cantidades de materiales necesarias para la realización del proyecto.

Los modelos detallados se caracterizan por ser los más complejos y laboriosos, y por ello, los más fiables.

Para la realización de estos modelos, se propone el seguimiento de las siguientes pautas (Ramírez, 2015):

En primer lugar, se debe extraer dos listas de las mediciones del proyecto

- Relación de partidas generadoras de residuos
    - Partidas de demolición y desmontaje
    - Partidas de retiradas de sedimentos procedentes de excavación
    - Partidas que generan residuos de forma indirecta a consecuencia de la actividad de construcción.

- **Listado de necesidades**
  Consiste en la redacción detallada de los productos básicos y auxiliares que van a utilizarse en el proyecto.

Una vez recabada esta información el siguiente paso es combinar ambas listas de forma que se genere un nuevo listado en el que se detalle el residuo asociado a cada producto, como puede ser el embalaje de cada partida o los retales y chatarras producidas en cada etapa.

Por último, se expresa el tipo de RCD y las cantidades finales en metros cúbicos o toneladas de residuo. Para ello es necesario conocer la densidad del material, lo que en muchas situaciones se resuelve utilizando coeficientes de densidad empíricos obtenidos de pesar los contenedores de residuos de obras pasadas, de los cuales conocemos el volumen que ocupan.

Modelo de Predimensionado

Los modelos Predimensionados tratan de estimar las cantidades de RCD basándose en coeficientes de proporción(Gimenez, 2010).

Estos coeficientes se suelen generan en función de las siguientes variables:

- **Tipo de edificio:**
  Nave industrial, bloque residencial, local comercial...etc.

- **Tipo de Obra:**
  Nueva obra, operación de reforma, demolición...etc.

- **Tipo de cimentación:**
  Zapatas, losas, pilotes...etc.

- **Tipo de cerramiento:**
  Aluminio, PVC, madera...etc.

Este tipo de modelo es una versión simplificada del Modelo Detallado, por el contrario, para que estos modelos obtengan predicciones eficaces y fiables, debe cumplirse que la información de la que dispone el analista se adapte a los coeficientes listados anteriormente. Cuanto mayor sea la información disponible, mejor será la estimación.

Modelo Software

Existe una gran variedad de programas informáticos dedicados a facilitar el trabajo de estimar el RCD producido en una obra, ante la imposibilidad de citarlas todas, a continuación, se hace referencia a este tipo de software clasificándolo en función de la institución que lo ha desarrollado:

- **Los realizados por Colegios profesionales y agrupaciones técnicas del sector:**
  La mayoría de estos servicios son plantillas gratuitas de hojas de cálculo ofrecidas para facilitar el trabajo de los técnicos, como ejemplo de esto podemos citar las Fichas Técnicas del Colegio Oficial de Arquitectos de Cataluña(COAC, 2017).

- **Los realizados por instituciones Público-Privadas del sector:**
  Este tipo de aplicaciones han sido muy utilizadas en los inicios. Consisten en la obtención de relaciones entre las distintas variables involucradas en el estudio de generación de RCD. Los principales problemas que acarrean estas aplicaciones son: la escasa actualización de los métodos de cálculo, lo cual hace que estas herramientas se vuelvan obsoletas en poco tiempo y la generalización de los casos, lo cual les hace inservibles para la estimación de situaciones concretas(Comerón Graupera, 2017).

- **Los realizados por casas comerciales de software:**
  Estas aplicaciones como por ejemplo “Arquimedes” (de la compañía CYPE) o “Construbuit” ponen a disposición del usuario un potente generador de modelos de predicción. Las principales desventajas que encontraremos al utilizar estos son: el elevado coste de las licencias de compra de estos programas y la escasa información sobre los métodos de cálculo y las experiencias técnicas que fundamentan los resultados (CYPE , 2017).

Ante la necesidad de realizar estimaciones fiables para casos de obras concretas, en este trabajo evaluaremos las experiencias particulares de una empresa constructora para generar un modelo que se adapte a esta base de datos y poder estimar la cantidad y el tipo de residuo que se va a generar en obras futuras.
Este tipo de estudios conlleva una pérdida de generalidad, adaptando el modelo a la información proporcionada y a la forma de actuación de esta empresa en concreto.

1.3. Objetivos

Se desea obtener estimaciones de la cantidad y el tipo de residuo producido en una obra civil. Para ello disponemos de una base de datos real con información reducida acerca de residuos producidos en que pueden ser similares.

A partir de esto, los objetivos de este trabajo fin de grado son:

- Establecer una metodología experimental que permita el empleo de los datos disponibles en la generación de modelos (de regresión) basados en datos, para estimar los residuos de una obra.
- Analizar mediante estos modelos y otras técnicas clásicas, las relaciones existentes entre RCDs.
- Utilizar Redes Neuronales Artificiales para la computación del conjunto de modelos de estimación.
- Enumerar en base a lo anterior un mapa de relación entre RCDs.

1.4. Organización de memoria

Este trabajo está compuesto por 5 capítulos, a continuación, se detalla el contenido de cada uno:

- Capítulo 1: Introducción

  Este capítulo introductorio expone la motivación por la cual se ha desarrollado este trabajo, se describe el estado del arte y se comenta brevemente el contenido y la extensión del TFG.

- Capítulo 2: Aspectos Teóricos

  En este capítulo se detalla la documentación necesaria para comprender las técnicas utilizadas en la metodología del trabajo.

- Capítulo 3: Metodología

  Este capítulo describe la metodología utilizada, apoyada sobre las bases teóricas descritas en el capítulo anterior.

Se detallan técnicas de pretratamiento de datos, técnicas de análisis y técnicas de generación de modelos utilizadas para alcanzar una correcta estimación de RCD.

- Capítulo 4: Resultados

En este capítulo encontraremos los resultados numéricos obtenidos con la metodología expuesta anteriormente.

- Capítulo 5: Conclusiones

En este capítulo se explican los resultados obtenidos y se valora la utilidad de estos.

- Capítulo 6: Planificación

En este capítulo se muestra la organización del proyecto, los tiempos de ejecución necesarios para realizar cada una de las partes que lo componen y una estimación de los costes de este.

2. ASPECTOS TEÓRICOS

En este capítulo se describen las técnicas de procesamiento y análisis de datos, los métodos de selección y extracción de variables y los diferentes sistemas de generación de modelos con las que podamos alcanzar los objetivos descritos en el capítulo anterior.

Las técnicas descritas en este capítulo serán la base teórica sobre la que se apoyará la metodología seguida en la realización de este trabajo fin de grado.

2.1. Procesamiento de datos

Habitualmente el trabajo con datos reales implica afrontar una serie de problemas típicos de la recopilación de información mediante sondeos estadísticos, encuestas, o incluso mediante métodos automáticos de captura. Estos errores se introducen debido a varias razones, tales como procedimientos manuales de entrada de datos, errores de equipo y medidas incorrectas (Herrera, 2015).

La detección de estos errores es un trabajo que debe realizarse antes de comenzar con cualquier tipo de análisis o estudio, de no hacerlo se podría inferir en(Refaat, 2007):

- Pérdida de eficiencia.
- Complicaciones en la manipulación y análisis de los datos.
- Sesgos en los resultados debido a las diferencias entre los datos completos y los perdidos.

Para evitar que esto suceda y conseguir una base de datos homogénea y robusta sobre la que poder realizar los análisis posteriores, debemos subsanar previamente los siguientes supuestos (L. Hernández G, 2008):

- El problema de la no respuesta, *Missing Values.*
- El problema de los datos atípicos, *Outliers.*
- El problema de la disparidad en el origen de los datos.
- El problema de la disparidad en las unidades de medida de los datos.

El propósito del pretratamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería de datos, se pretende que los datos que van a ser utilizados en tareas de análisis o descubrimiento de conocimiento conserven su coherencia.

A continuación, se pasa a desarrollar las distintas técnicas para subsanar de forma óptima los problemas mencionados anteriormente.

2.1.1. *Missing Values*

Al realizar el análisis de cualquier muestra es muy probable que nos encontremos con variables en las que faltan datos. Esto puede ser debido a que la muestra ha sido construida por personas distintas, con métodos distintos, o simplemente que existen datos faltantes que no se han podido encontrar a la hora de formar la muestra. La discusión sobre el problema de la no respuesta y algunos métodos para manejarla se desarrollaron desde los años 1930-1940.
Para alcanzar un correcto análisis debemos identificar los valores perdidos, así clasificándolos según la causa de pérdida (Rubin, 1987), tenemos:

- MCAR (*Missing completely at random*):
  La probabilidad de pérdida es igual para todos los individuos y no dependen de la medida de otras variables.

- MAR (*Missing at random*):
  La probabilidad de pérdida de un individuo depende de la información observada.

- MNAR (*Missing not at random*):
  La probabilidad de pérdida viene condicionada por los valores perdidos.

Formas de abordar la pérdida de información

Una vez hemos identificado el tipo de pérdida que se ha producido (MCAR, MAR, MNAR) disponemos de una serie de métodos para corregir el sesgo que podría producir el espacio vacío que deja la no respuesta en cualquier variable:

- Análisis con los datos completos (*Listwise*):
  Este método consiste en analizar la muestra solo con los datos de los que disponemos toda la información, eliminando aquellos individuos que muestren signos de datos perdidos en alguna de sus variables.
  La ventaja de utilizar este método reside en su simplicidad a la hora de implementarlo (Allison, 2001).
  La utilización del método *Listwise* conlleva importantes pérdidas de información si el número de datos faltantes ese elevado, pudiendo acarrear sesgos en los resultados de los análisis que se realicen con este tipo de muestras.

Eliminar información de la muestra solo será viable cuando los datos sean homogéneos y la pérdida de información se haya producida de forma aleatoria, lo cual es poco habitual en la práctica, además se desprecia una gran cantidad de información conocida.

- Análisis con los datos disponibles (*Pairwise deletion*):
  Este método consiste en utilizar todos los datos de los que disponemos información en la muestra. Esto conlleva que existirán variables en la muestra con un tamaño inferior a otras, lo que acarreará problemas al analizar la muestra con métodos estadísticos como el análisis de correlaciones. Con este método se obtienen buenos resultados solo al tratar datos perdidos completamente aleatorios (MCAR).

- Ponderación:
  El método de ponderación se aplicará en aquellas muestras en las que encontremos individuos con datos faltantes en todas sus variables. La esencia de este método consiste en añadir más valor a los individuos de los que se conoce más información, haciendo que estos subsanen la ausencia del resto.

  El objetivo de este método es mejorar la precisión de las estimaciones posteriores y reducir lo máximo posible el sesgo producido por los individuos perdidos, ya que los resultados presuponen que todos los individuos contienen información.
  El principal problema de este método es que pueden darse estimaciones con una varianza muy alta (Berka, 1998).

- Imputación numérica:
  Una de las técnicas más utilizadas en los estudios estadísticos para el tratamiento de datos faltantes es el método de imputación. Consiste en sustituir los valores faltantes por datos de nuestra elección. Este método nace de la necesidad de obtener muestras completas para realizar los análisis estadísticos posteriores y consiste en completar los datos faltantes apoyándonos en el resto de información de la muestra. Haremos mayor hincapié en esta forma de abordar la no respuesta, ya que será la técnica utilizada en la metodología de este TFG (Rubin, 1987).

Imputación numérica

En el proceso de imputación se debe elegir cuidadosamente la variable objetivo y las variables auxiliares con las que completaremos a esta. Existen diversos métodos, el criterio de elegir uno u otro dependerá del problema en sí y de la decisión del analista. Una clasificación general de estos métodos puede ser según los siguientes criterios de calidad:

- **Manteniendo la distribución de la variable**:
  El objetivo de estos es conseguir que la distribución de la variable tras la imputación mantenga una tendencia lo más parecida a la variable original.
  Este criterio se impone a aquellas variables que se pueden modelar con una distribución de valores conocidos. Para asegurar esto, por ejemplo, ante una variable que responde a una distribución normal de media "M" y desviación típica "D", se debe elegir el conjunto de valores dispuestos a incorporarse a la muestra de manera que mantengan inalterables los parámetros "M" y "D".

- **Manteniendo la correlación entre variables:**
  Este tipo de métodos cambia los datos faltantes por otros, intentando mantener en todo momento la relación existente entre la variable objetivo y las variables auxiliares. Por ejemplo, se deberá tener en cuenta los coeficientes de correlación mostrados entre las variables auxiliares y la variable en la que se va a realizar la imputación, entonces para satisfacer este criterio se sustituirán los valores faltantes por estimaciones, generalmente relaciones lineales realizadas a partir de otras variables del muestreo. Estas estimaciones no tienen por qué ser certeras, el único requerimiento necesario es que mantengan un nivel de correlación lo más parecido al que había antes de sustituir los valores faltantes.

Otro criterio para clasificar los métodos de imputación puede ser en función del número de valores que vamos a utilizar para sustituir el dato faltante, según esto, podemos encontrar:

- **Imputación simple:**
  A cada valor perdido se le asigna un solo valor imputado. Por ejemplo, en una situación en la que dispongamos de un conjunto de datos dispuesto para sustituir los valores faltantes de una variable y el analista decide utilizar uno solo de ese conjunto, se trata de una imputación simple.

- **Imputación múltiple:**
  A cada valor perdido se le asigna un rango o un conjunto de datos con los que se trabajará de forma conjunta para obtener los parámetros de interés y combinar los resultados obtenidos.
  Por ejemplo, si se utiliza un método de imputación que conserve la correlación entre variables dispondremos de una gran cantidad de valores que satisfagan dicha condición. En una imputación múltiple se elegirán todos ellos, o un grupo formado por los más relevantes.

Otro criterio de clasificar los métodos de imputación puede ser en función de la repetitividad del método:

- **Determinísticos:**
  Estos métodos determinan un solo valor con el que sustituir el *Missing Value.* Independientemente de las veces que se repita el método de imputación o independientemente del analista que realice el proceso, se producirán las mismas respuestas. Esto ocurre cuando en una muestra, se requiere que el valor que va a sustituir al *Missing Value* cumpla una serie de condiciones estrictas, de tal forma que solo haya una solución posible.

- **Aleatorios**:
  La respuesta varia cada vez que se aplica el método, aun trabajando bajo las mismas hipótesis y condiciones, esto sucede, por ejemplo, al realizar una imputación simple cuando disponemos una serie de valores aptos para sustituir los *Missing Value*s y de forma aleatoria seleccionamos uno solo.

A lo largo de los años se ha desarrollado una gran cantidad de métodos de imputación, a continuación, se describen dos de los más utilizados y de los más antiguos, los cuales son de especial interés puesto que serán los métodos utilizados en la metodología de este trabajo.

Imputación por media

Se trata de un método de imputación simple, es uno de los métodos más sencillos de aplicación. En esencia, consiste en sustituir los *Missing Values* por el valor de la media de la variable, sin tener en cuenta para ello los propios *Missing Values* (Wilks, 1932).

Este método presenta un importante defecto al subestimar la variabilidad real de la muestra. Los principales inconvenientes que observamos en el método de imputación mediante la media son (Gomez García, 2006):

- Dificulta la estimación de la Varianza.
- Distorsiona la verdadera distribución de la variable.
- Distorsiona la correlación entre variables dado que añade valores constantes.

Existen dos variaciones de este método, una versión determinista y una aleatoria.

- Imputación por media no condicional:
  Este método consiste en estimar la media de los valores conocidos, sin tener en cuenta los datos faltantes, y se estima su valor con ello, manteniendo invariable la media de la variable tras la imputación. Esto se aplicará a las variables con datos faltantes del tipo MCAR.
  Aunque se preserve el valor de la media, este método puede acarrear sesgos en la varianza y percentiles, pudiendo llegar a distorsionar la relación con el resto de variables.

$$Valor\ de\ imputación = \frac{\sum X_i}{i} \qquad \textit{(Ec. 2.1)}$$

  La ecuación 2.1 representa el valor por el cual se va a sustituir a los *Missing Values* de la variable "X", siendo "i" el número de datos completos de una variable y "X_i" cada una de las muestras de dicha variable.

- Imputación por media condicionada:
  El método común de esta imputación consiste en observar la muestra en busca de patrones o grupos que se comporten de forma similar. Una vez definidos estos grupos o Casos de estudio, se le aplicará el método no condicional a cada uno de ellos.

$$Valor\ de\ imputación\ Caso\ A = \frac{\sum X_a}{a} \qquad \textit{(Ec. 2.2)}$$

  La ecuación 2.2 indica el valor por el cual se va a sustituir a los valores faltantes de la variable "X", siendo "a" el número de muestras pertenecientes al caso de estudio A y "X_a" el valor de cada muestra de la variable X perteneciente al caso de estudio A.

Imputación por regresión

Se trata de un método de imputación simple, para este método se utilizan modelos de regresión para imputar la variable *Y*, a partir de las covariables (X_1, …, X_k) correlacionadas con *Y* (Buck, 1960).

El método de imputación mediante regresión lineal consiste en encontrar una función lineal de tal forma que se consiga modelar la relación existente entre un par de variables. A esta línea se la conoce como Recta de Regresión (Ec. 2.3).

$$y = a \cdot x + b \qquad \textit{(Ec. 2.3)}$$

La finalidad de la Recta de regresión consiste en representar los datos de las variables X e Y. Las representaciones algebraicas de estas pueden ser muy variadas, de modo que existen varios criterios con los que formalizar la recta de regresión. El más utilizado, y al que recurriremos en este trabajo se trata del criterio de ajuste por mínimos cuadrados (Nieves Hurtado, 2014).

La recta de ajuste por mínimos cuadrados es aquella que pasa por entre los puntos de la muestra, de tal modo que produce el área total mínima. Este criterio da lugar a una recta única.

Frente a la imputación mediante la media, este método incorpora la información que tienen el resto de variables, lo cual tiene una parte positiva, evitando que los datos se centren en la media y manteniendo la dispersión de las variables y una parte negativa, pues puede acarrear sesgos al reforzar la tendencia lineal de las variables imputadas.

- Incrementa artificialmente las relaciones entre variables.
- Hace que se subestime la Varianza de las distribuciones.
- Asume que las variables con datos ausentes tienen relación de alta magnitud con las otras variables.

2.1.2. *Outliers*

Por las mismas razones por las que encontrábamos valores faltantes, encontramos valores que difieren de forma sustancial del resto de puntos de la muestra, a estos se les llama en inglés *Outliers.*

Los *Outliers* pueden ser elementos no representativos del conjunto de la muestra, los cuales pueden distorsionar el comportamiento de los estadísticos que se estudien en ella. Por otra parte, no se deben retirar los candidatos a O*utlier* sin realizar antes un correcto análisis, sería un error retirar de la muestra datos que

podrían representar la tendencia de un grupo de menor tamaño o con escasa representación.

Podemos realizar una clasificación de los posibles *Outliers* que podemos encontrar en función de su origen(Rodriguez, 2011):

- *Outliers* procedentes de un error en la entrada de datos en el sistema:
  Estos errores tipográficos o incluso ortográficos deben eliminarse siempre que sea posible.

- Observaciones de eventos puntuales o extraordinarios:
  En cuyo caso un solo *Outlier* no puede representar ningún segmento de población valido, y por ello debe ser eliminado del análisis.

- Datos atípicos con valores pertenecientes a un rango normal, pero atípicos al interaccionar con el resto de variables:
  Ante estos candidatos a *Outlier* se deben mantener en observación, comparando los resultados obtenidos al utilizarlos y al prescindir de ellos en la muestra.

- Datos atípicos sin explicación sobre su procedencia:
  Se debe recurrir a duplicar la muestra de manera que se pueda comprar el resultado obtenido al eliminar estos *Outliers* de la muestra.

Existe una gran variedad de métodos mediante los cuales poder determinar qué valores serán considerados *Outliers,* alguno de los más utilizados y reglados son (Murphy , 20017):

- Prueba de Grubbs:
  Utiliza una estadística de prueba, T, que es la diferencia absoluta entre el valor atípico, X_0, y el promedio de la muestra "$\overline{X}$" dividida por la desviación estándar de la muestra, "s".

$$T = |X_o - \bar{X}| / s \qquad \textit{(Ec. 2.6)}$$

- Prueba de Dixon:
  Utiliza relaciones de los espacios entre datos de diferentes modos según la cantidad de valores en el grupo de datos. Por lo tanto, el índice de Dixon se compara con un valor crítico de una tabla, y el valor se declara valor atípico si supera ese valor crítico. El valor crítico depende del tamaño de la muestra, n, y de un nivel de representatividad elegido, que es el riesgo de rechazar

una observación válida. La tabla por lo general utiliza niveles de baja representatividad tal como 1% o 5%.

- Diagrama Boxplot:
  El método clásico para considerar candidatos a valor atípico consiste en calcular la diferencia entre el primer y tercer cuartil (valor intercuartil), se considerará valor atípico leve aquel que se encuentre a 1,5 veces esta distancia y valores atípicos extremos aquellos que la superen 3 veces.

2.1.3. Transformación de datos: Normalización de variables

Independientemente del uso que vayamos a darle a la información, es conveniente normalizar los datos, de esta forma, al expresarlos en magnitudes normalizadas podremos compararlos y analizarlos con los procesos posteriores (Hoaglin, Mosteller, & Tukey, 1987).

Existen diversas formas de transformar las variables de una muestra, la finalidad de todas ellas consiste en escalar las variables a un rango de valores determinado para poder analizarlas independientemente de sus unidades de medida. Para este trabajo utilizaremos dos métodos básicos: El método Min-Max y el método Zscore.

- Método Min-Max:
  "Ejecuta una transformación lineal de los datos originales. Con base en los valores mínimo y máximo de un atributo, se calcula un valor de normalización v' con base en el valor v de acuerdo con la siguiente expresión"(L. Hernández G, 2008):

$$V' = \frac{V - Vmín}{Vmáx - Vmín} \qquad (Ec.2.7)$$

Este procedimiento corresponde a una linealización con ordenada en el origen, obtenemos un conjunto de valores comprendido en el intervalo [0,1].

- **Metodo Zscore:**

  Los valores para un atributo A son normalizados basados en la media y la desviación estándar de A. Un valor v de A es normalizado a v' con el cálculo de la siguiente expresión (Berka, 1998):

$$V' = \frac{V - \mu_A}{\sigma_A} \qquad \textit{(Ec. 2.8)}$$

  Con este procedimiento obtenemos una relación de valores cuya media será cero (μ=0) y su desviación típica será la unidad (σ =1).

  Este método es utilizado cuando el máximo y el mínimo del atributo A son desconocidos o cuando hay valores anómalos que predominan al utilizar la normalización min-max.

2.2. Análisis de Correlación

El análisis de correlaciones permite conocer la relación lineal existente entre cada par de variables de una muestra. Este análisis es un primer paso para comprender el comportamiento de las variables y servirá de ayuda para continuar con estudios estadísticos posteriores.

Los coeficientes que caracterizan esta matriz son los resultados de la correlación de Pearson, el cual se describe a continuación.

Coeficiente de Pearson

El coeficiente de correlación de Pearson es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente. En valor absoluto, este índice alcanza valores de cero a uno, si el coeficiente de Pearson de dos variables es igual a uno, quiere decir que dichas variables tienen una relación lineal perfecta, mientras que, si alcanza el valor cero, el coeficiente nos indica que dichas variables no mantienen relación lineal. Hay que tener en cuenta que este parámetro tan solo tiene en cuenta la relación lineal, una puntuación de cero en este coeficiente no indica que las variables sean independientes (Martinez-Vara, 2005).

$$0 \leq r_{xy} \leq 1 \qquad \textit{(Ec.2.9)}$$

Si contemplamos el signo del coeficiente, veremos que sus valores oscilan realmente entre menos uno y uno, la puntuación negativa del coeficiente de Pearson indica relación lineal inversa, pero independientemente del signo, una puntuación de “1” o “-1” indican relación lineal perfecta.

Siendo X e Y las variables participantes en el análisis de correlación, el coeficiente de correlación de Pearson viene definido por la siguiente expresión:

$$r_{xy} = \frac{\frac{\sum X \cdot Y}{N} - \overline{XY}}{S_x S_y} \qquad \textit{(Ec.2.10)}$$

Prueba de significación

Para asegurarnos de que el coeficiente de Pearson es un valor sobre el que poder apoyarnos, y la relación existente entre X e Y no es fruto del azar, recurriremos a la prueba de Student, con la que finalmente obtendremos un p-valor, el cual indicará si el índice de Pearson obtenido entre un par de variables tiene un grado de significación importante, o por el contrario no se puede demostrar tal relación y deberemos excluirlo del análisis(Triola, 2004).

Sin entrar en detalles del método para determinar la probabilidad de que un cierto coeficiente proceda de una población cuyo valor es cero, se realizan dos hipótesis de contraste:

- H_0: $r_{xy} = 0$
- H_1: $r_{xy} \neq 0$

Se demuestra que, con el supuesto de hipótesis nula, sigue una distribución de Student con N-2 grados de libertad, manteniendo la media del valor poblacional y siendo su desviación típica:

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{N - 2}} \qquad \textit{(Ec. 2.11)}$$

A efectos prácticos, se calculan las desviaciones tipo que se encuentra el coeficiente obtenido del centro de la distribución y se compara dicho valor con las tablas de Student, mediante las cuales podremos decidir, para un nivel de confianza determinado (generalmente se utiliza el 95%) si se rechaza o no la hipótesis nula.

2.3. Extracción y selección de variables: PCA

El PCA (*Principal Component Analysis*), técnica creada por Hotelling (1933) y originada por los ajustes ortogonales por mínimos cuadrados formulados por K.Pearson (1901), que consiste en transformar las variables originales de un muestreo, en general correladas, en nuevas variables incorreladas, pudiendo reducir la dimensión de los datos del problema al deshacernos de la información redundante del problema (Peña, 2002).

A estas nuevas variables incorreladas se las conoce como Componentes Principales.

Objetivo de esta técnica (Sainz, 2015):

- Reducción de la dimensión del espacio de entrada.
- Reducción del número de variables o características a emplear.
- Emplear las características más adecuadas y/o aquellas que contentan más información.

Estos componentes principales, además vendrán ordenados de forma decreciente según la variabilidad de los datos, es decir, el primer componente principal será el que más varianza recoge, el segundo componente principal será el segundo que más recoge respecto al resto y así sucesivamente. Dicho de otra forma, mediante el PCA transformaremos nuestra muestra inicial para obtener por orden decreciente las variables (componentes principales) más independientes del estudio.
Para nuestro estudio, el PCA no va a ser un fin, sino que va a ser un medio con el que poder identificar las variables de la muestra que más influyen sobre el resto.

Antes de adentrarnos en la formulación del PCA, cabe destacar que las variables de origen deben estar previamente normalizadas mediante alguno de los procesos descritos anteriormente, de tal forma que tengan media cero y por lo tanto los componentes principales también tengan media nula.

Cálculo de los Componentes Principales

A continuación, se describe de forma simplificada el proceso para obtener los componentes principales de un muestreo normalizado "**X**" (Jolliffe, 1986).

Los valores del PCA para cada observación vienen dados por:

$$X = \sum \alpha i \times \Phi i = \Phi \cdot \alpha \qquad \textit{(Ec. 2.12)}$$

Donde **X** es la matriz de datos normalizados, "Φ" contiene las (*i,k*) puntuaciones (*scores*) del *k* Componente Principal para la observación "*i*" y "**α**" es una matriz ($p \times p$) cuyas *K* columnas corresponden con los *K* autovectores de X'X. Calcular los componentes principales equivale a aplicar una transformación ortogonal "**α**" a las variables "**X**" (variables originales) para obtener variables nuevas "**Z**" incorreladas entre sí.

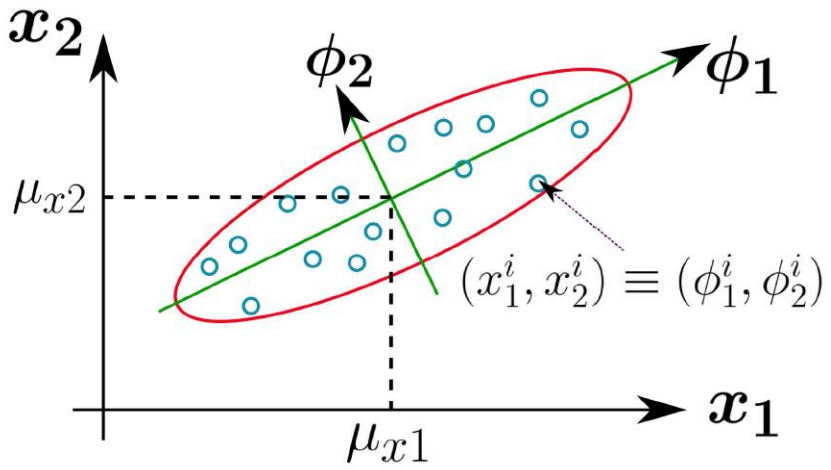


Fig.2.1. Representación bidimensional de los dos primeros componentes principales.

Propiedades de los componentes principales

Las nuevas variables a las que hemos llamado componentes principales, tienen las siguientes características:

- Cada λ_i (autovalor) mide la relevancia de la característica asociada: (Autovector Φ_1): $\lambda_1 > \ldots > \lambda_n > 0$.

- El error cometido al eliminar "n-m" características es conocido:

$$E(m) = f\left(\frac{\sum_{i=m+1}^{n} \lambda i}{\sum_{i=1}^{n} \lambda i}\right) \qquad \textit{(Ec.2.13)}$$

- Las Direcciones Principales son: $\Phi_1, \ldots, \Phi_m$.
- Los Componentes Principales son: $\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_m}$.
- Se pierde el significado de las variables originales.
- Se pierde la linealidad que mantenían las variables originales.

Elección del número de componentes principales

No se ha establecido una norma fija con la que guiarnos en la elección del número de componentes principales, nos es imposible generalizar, pues cada problema requerirá una solución concreta. Los criterios que suelen adoptarse son los siguientes:

- El codo del gráfico:
  Consiste en representar gráficamente λ_i frente a i. La idea consiste en encontrar un punto de inflexión en el cual la curva descrita por λ_i cambie de pendiente bruscamente.

  Este método de decisión, a pesar de la subjetividad que conlleva es el más utilizado, aunque en muchas situaciones nos encontraremos con que la variabilidad del muestreo se reparte de forma equitativa entre todas las componentes, haciendo inviable la utilización de esta técnica. Para estas situaciones se recurre a procesos de decisión más pragmáticos.

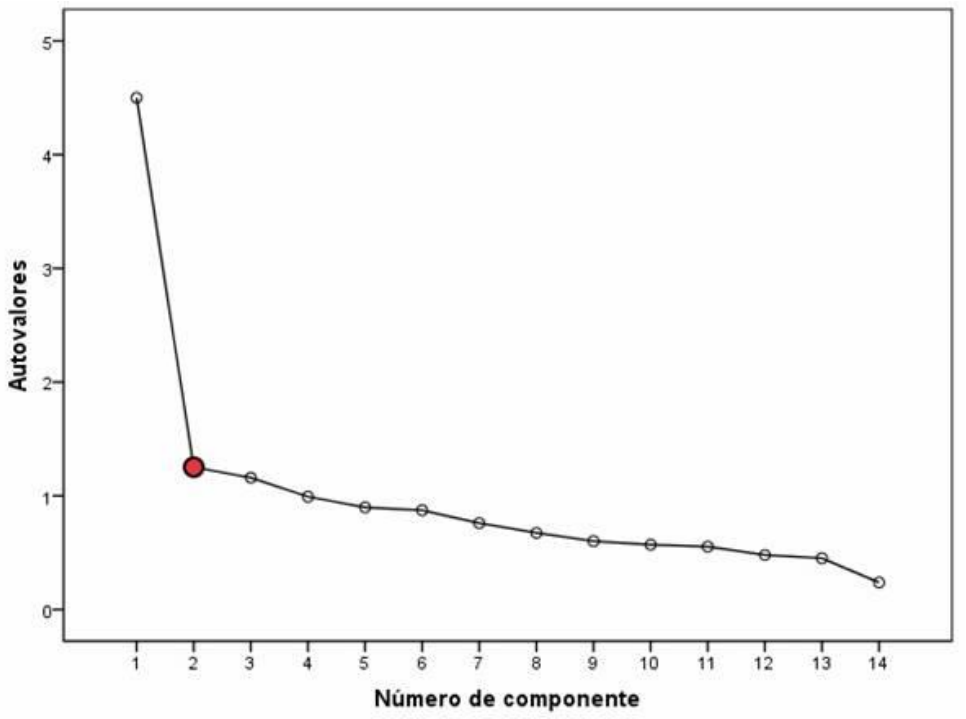


Fig. 2.2. Ejemplo punto de inflexión. Selección de componentes principales

- **Cubrir una proporción determinada de varianza**:
  Estableciendo una cantidad de varianza que deseamos recoger, obtendremos un método de decisión fijo para cualquier estudio. Generalmente se busca cubrir el 80% o el 90% de la varianza, aunque este criterio es de nuevo subjetivo, ya que será el analista quien decida el rango de varianza que desea cubrir para su análisis.

Selección de Variables

Como hemos visto, el PCA refleja la redundancia de las variables de un conjunto muestral, entonces es posible determinar qué variables son las que más información contienen o, dicho de otro modo, cuáles son las que menos información redundante contienen.

Para conocer el peso o la cantidad de información única que contiene una variable deberemos haber realizado en primer lugar la selección de las Componentes Principales (CP). Una vez sepamos cuantas CP son suficientes para el estudio, procedemos a sumar las puntuaciones obtenidas en la matriz de coeficientes "α" de cada variable en las CP seleccionados, así obtendremos una puntuación para

cada variable original, en función de la participación de cada una de estas variables originales en las componentes principales.

$$Peso_i = \alpha_{1,i} + \alpha_{2,i} + \ldots + \alpha_{n,i} \qquad \textit{(Ec. 2.13)}$$

Para comprender la magnitud de estas puntuaciones podemos expresarlo en forma porcentual de la siguiente forma:

$$\%Peso_i = 100 \cdot \frac{Peso_i}{Peso\ Total} \qquad \textit{(Ec. 2.14)}$$

Siendo "$Peso_i$" el nivel de importancia definido anteriormente de cada variable y "PesoTotal" la suma de los pesos de cada variable.

2.4. Redes Neuronales Artificiales

Las redes neuronales artificiales (RNA) son algoritmos matemáticos que tratan de emular el sistema neuronal del ser humano, como la capacidad de memorizar y asociar hechos, consiguiendo aprender mediante la experiencia, generalizando casos anteriores a nuevos casos realizando ensayos y calculando y corrigiendo el error producido(Tsoukalas & Uhrig, 1996).

Podemos enumerar las principales ventajas que caracterizan las redes neuronales como(Haykin, 1998):

- **Aprendizaje Adaptativo:**
  La capacidad de aprender en base a los procesos de entrenamiento es sin duda la característica principal de las redes neuronales. Durante la etapa de entrenamiento, las neuronas son capaces de adaptarse a cada intercambio de información, sin necesidad de algoritmos pre programados por parte del investigador. Una red neuronal no necesita un algoritmo de trabajo, es ella misma quien genera el algoritmo, variando los pesos de cada factor según las necesidades del problema hasta llegar a una solución óptima.
  La función del analista será la de escoger la arquitectura adecuada para su problema y seleccionar cuidadosamente los datos de entrada que recibirá la red neuronal.

- **Auto-organización:**
  La red neuronal puede organizar la información recibida durante la etapa de entrenamiento de forma automática, creando sus propios sistemas de

almacenamiento de información y generando sus propios patrones. Gracias a esta propiedad, las redes neuronales pueden determinar de forma autónoma qué datos son más relevantes para qué objetivos, pudiendo de esta forma enfrentarse a situaciones que para el ojo humano puedan parecer nuevas.

- **Tolerancia a fallos:**
  Las principales virtudes de la red neuronal se deben al trabajo en equipo de todas las neuronas, aun cuando algunas neuronas muestran fallos, el resto de la red se auto equilibra, haciendo que el resultado final se vea sesgado lo mínimo posible por esos fallos. Esto es posible debido a que en una red neuronal existen pedazos de información redundante, mientras que en un proceso algorítmico se almacenan datos en espacio reservados y únicos, las redes neuronales guardan información no localizada repartida entre todas las neuronas. El grueso de la información se encuentra en las interacciones de las neuronas, y no en la muestra de entrada propiamente dicha.

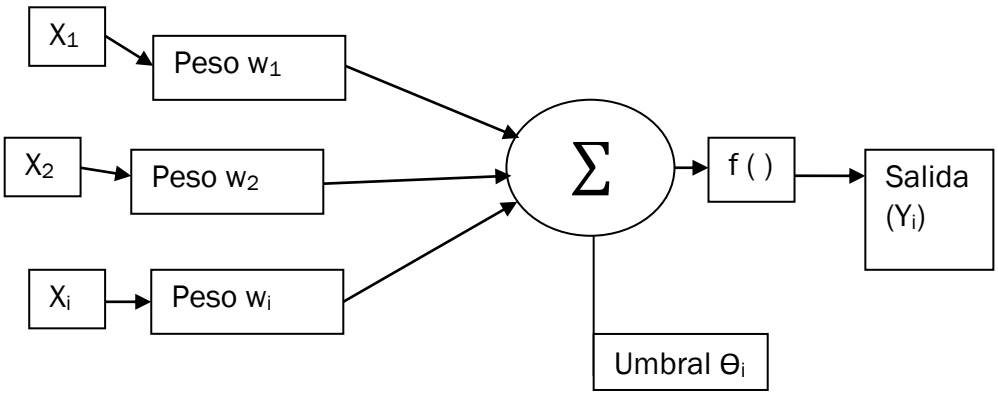
- **Operación en tiempo real:**
  Aunque las etapas de entrenamiento y validación requieren un alto tiempo de computación, existen máquinas con la capacidad de trabajar con redes en paralelo. Gracias a la evolución tecnológica de los componentes electrónicos se puede generar una red neuronal perfectamente funcional en cuestión de minutos.

- **Fácil inserción dentro de la tecnología existente**:
  Se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilitará la integración modular en los sistemas existentes.

Neurona

La unidad básica de la RNA es la neurona, por si sola una neurona no es muy eficaz para el tratamiento de información, por ello se agrupan en grandes estructuras o redes (Rodríguez-Bernal, 2012).

Las neuronas son las encargadas de recibir, procesar y enviar información a través de la red. Cada neurona tiene un peso (*Wi*) asociado a la información de entrada (*Xi*), si la combinación de estos supera el umbral determinado por la neurona (Θ_i), esta enviará un nuevo valor(*Yi*) a la siguiente neurona, propagándose de esta forma a través de toda la red.



(Fig. 2.3). Esquema de Neurona.

Como vemos en la figura 2.3, la neurona se puede describir mediante la función de entrada, la función de activación y la función de salida.

Función de entrada

En la entrada de la red neuronal se puede introducir una gran cantidad de información, para combinar estos datos se les dota a las neuronas con pesos (w_i) mediante los cuales se determinará qué datos son los más relevantes para llegar a la resolución del problema. La magnitud de estos pesos podrá ser determinada por el analista en caso de que así lo quiera, o por el contrario será la red quien irá variando estos valores de forma automática en el proceso de entrenamiento. De esta forma, la información de entrada se multiplicará por su peso correspondiente, formando así la función de activación.

Dependiendo de la arquitectura de red que estemos diseñando tendremos distintos tipos de funciones de entrada, siendo las siguientes las más habituales:

- **Sumatorio de los pesos**:
  Suma de todos los valores de entrada multiplicados por sus correspondientes pesos.

$$\sum_i n_i \cdot w_i \qquad \textit{(Ec. 2.15)}$$

  Siendo “n” el valor del dato de entrada y “w” su peso correspondiente

- **Productorio de los pesos**:
  Multiplicación de los valores de entrada multiplicados por sus correspondientes pesos.

$$\prod_i n_i \cdot w_i \qquad \textit{(Ec. 2.16)}$$

  Siendo "n" el valor del dato de entrada y "w" su peso correspondiente

- **Máximo de las entradas pesadas:**
  Tan solo se toma en cuenta el valor de entrada de mayor magnitud, previamente multiplicado por su peso.

$$Max_i(n_i \cdot w_i) \qquad \textit{(Ec. 2.17)}$$

  Siendo "n" el valor del dato de entrada y "w" su peso correspondiente

De forma general, podemos clasificar las funciones de entrada en dos grupos(Roger Jang, Sun, & Mizutani, 1997):

- **Función lineal o de tipo hiperplano**:
  Este grupo lo componen las funciones cuyo valor de red son combinación lineal de las entradas.

- **Función radial o de tipo hiperestésico**:
  Al contrario que las anteriores, esta, se trata de una función de segundo orden, el valor obtenido representa la distancia de ese punto a una referencia.

Función de activación

Las neuronas que componen la red pueden encontrarse en varios estados, activada, desactivada o en un punto intermedio.

La función de activación será la encargada de calcular el estado en el que se encuentran las neuronas. A partir de la información que entra en la red, la función de activación transformará esta señal en un valor de salida comprendido en el rango (0,1) o (-1,1), dependiendo del tipo de función de entrada utilizado. Esta

señal de salida adoptará el valor superior para indicar que la neurona está totalmente excitada y los inferiores (o ó -1) para expresar que la neurona se encuentra inactiva.

Existen multiples funciones de activación, sin adentrarnos en los fundamentos de cada tipo de función, diremos que las más habituales son (Roger Jang, Sun, & Mizutani, 1997):

- **Función lineal:**
  El rango de valores de esta función se encuentra entre -1 y 1.

$$\varphi(x) = m \cdot x \qquad \textit{(Ec. 2.18)}$$

  Dónde "m" es la pendiente de la recta, la representación de esta función es la siguiente:

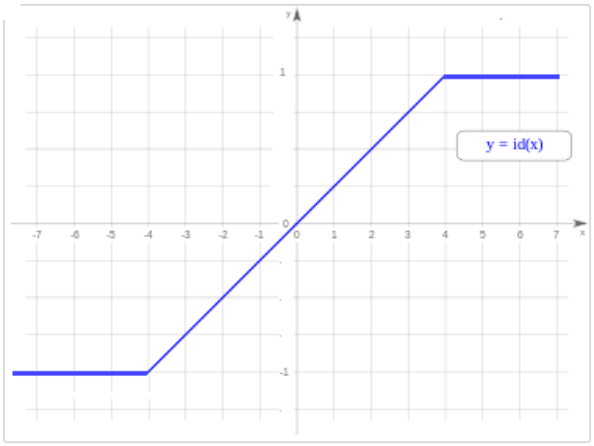


Fig.2.4. Función Lineal

- **Función Logaritmo Sigmoidea**:
  Rango de valores comprendidos en 0 y 1, del mismo modo que en la función lineal, la pendiente de esta se puede variar.

$$\varphi(v) = \frac{1}{1 + e^{-av}} \qquad \textit{(Ec. 2.19)}$$

  La representación de esta función es la siguiente:

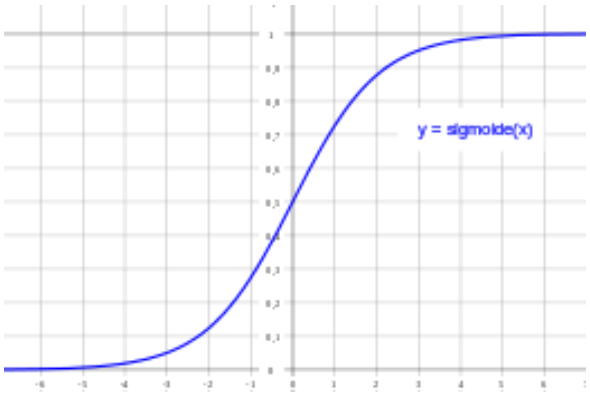


Fig. 2.5. Función Logaritmo Sigmoidea.

- **Función Tangente Sigmoidea:**
  Rango comprendido entre -1 y 1.

$$\varphi(v) = \tanh v = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \qquad \textit{(Ec. 2.20)}$$

La representación de esta función es la siguiente:

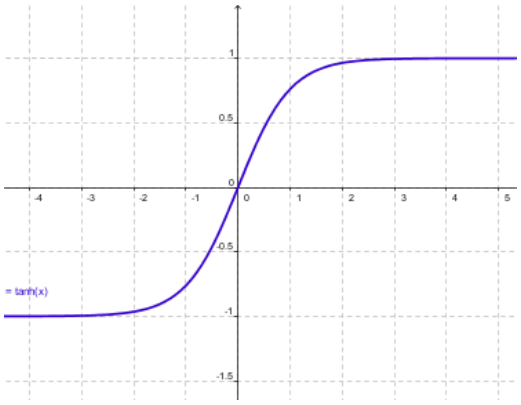


Fig. 2.6. Función Tangente Sigmoidea.

Función de salida

Por último, el valor resultante de aplicar la función de activación sobre su correspondiente u_k (la información proveniente de la función base) dará como resultado la señal de salida (y_k).

No toda la información de salida se transfiere a las neuronas siguientes, es la función de salida, quien se encarga de decidir qué datos pasan a qué neuronas. La función de salida se encarga de comprar el umbral (Θ_i) con la función de activación,

si la función de activación de una determinada neurona es inferior a su umbral, esta información no se transferirá a la siguiente. Con esto podemos intuir que la salida puede adoptar una forma binaria, 0 no pasa, 1 pasa, o adquirir valores en el rango (0,1) ó (-1,1), según su diseño.

Capas

La distribución de las neuronas dentro de la red se realiza a distintos niveles formando capas. En función de la situación que ocupa cada capa dentro de la red, podemos distinguir tres tipos de capas:

- **De entrada:** Reciben información del exterior de la red.
- **De salida:** Envían información al exterior de la red.
- **Oculta:** Procesan la información interna de la red y se comunican con el resto de las capas.

Estas capas pueden estar compuestas por una gran cantidad de neuronas, dependiendo de las necesidades del problema. La información de entrada (Inputs y Targets) se introduce en la primera capa, llegan hasta la capa oculta, la cual puede estar dividida en varias capas intermedias de neuronas interconectadas, en la capa oculta se producen las relaciones y algoritmos de reconocimiento de patrones que llegan hasta la capa de salida, donde se decodifica la información procesada y obtenemos los valores de salida (Outputs).

2.4.1. Clasificación de las Redes Neuronales

Podemos identificar distintos tipos de RNA en función de varios criterios, los dos más generales son(Ballesteros, 2017).

Clasificación según su topología

Esta clasificación distingue el número de capas de una red, el tipo de capas, pudiendo ser ocultas o visibles, de entrada, o de salida y la direccionalidad de las conexiones de las neuronas.

Hemos visto que las neuronas se agrupan formando capas, y estas a su vez forman redes. Dentro de esta clasificación podemos hacer una subdivisión, clasificando las redes en:

- **Monocapa:**
  Las neuronas crean conexiones laterales, manteniéndose todas interconectadas en una sola capa.

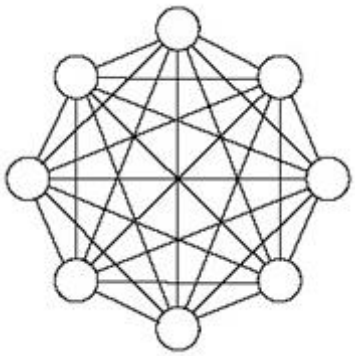


Fig. 2.7. Ejemplo de Red Monocapa: Red de Hopfield.

- **Multicapa:**
  Dentro de este tipo de arquitectura podemos diferenciar dos nuevos grupos, según la forma en que se conectan las capas. Generalmente el orden de las capas viene especificado por el orden en que reciben la señal desde la entrada hasta la salida, este tipo de conexiones se conocen como *feedforward*, es por ejemplo el caso de las redes Perceptron, Adaline, Madaline, Backpropagation...etc.

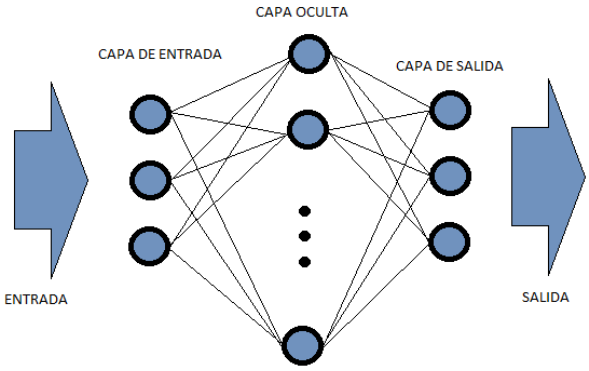


Fig. 2.8. Ejemplo de red multicapa: Perceptrón.

Las redes en las que la información puede regresar a capas anteriores, creando conexiones de capa hacia atrás, a diferencia de las anteriores, suelen ser bicapas, algunas de estas son las redes ART, Bidirectional Asociative Memory y Cognitron.

Clasificación según su algoritmo de aprendizaje

El proceso de trabajo de una red neuronal se puede dividir en dos etapas, la etapa de Test y la de Validación. Para ello, el analista deberá dedicar una parte de sus datos para entrenar a la red neuronal y el resto para realizar las comprobaciones necesarias para garantizar que el entrenamiento ha sido satisfactorio(Lehmann & Casella, 1998).

Existen varias formas de realizar la etapa de entrenamiento, así dependiendo del modo de entrenamiento de una red, nos podemos encontrar con dos tipos distintos:

- **Redes de entrenamiento supervisado**:
  Para este tipo de entrenamiento debemos proveer la red con valores de entrada "Inputs" y con los correspondientes valores esperados o "Targets". Al finalizar cada iteración en la fase de entrenamiento supervisado, se comparará la señal de salida con el target esperado, en función de la diferencia de valores ó error obtenido en la iteración anterior, se realizan cambios en los pesos para corregir o minimizar el error de salida.

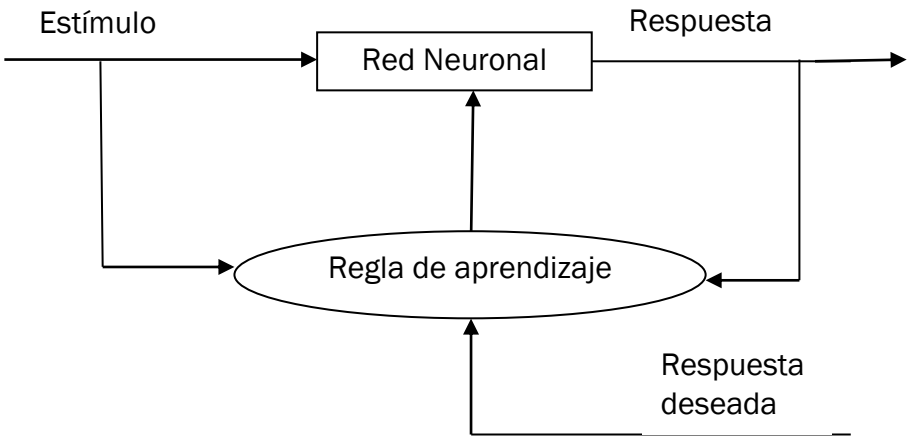


Fig. 2.9. Red Neuronal supervisada.

- **Redes de entrenamiento no supervisado**:
  Este tipo de entrenamiento basado en "data clustering" funcionan sin necesidad de introducir los datos target. Su procedimiento consiste en categorizar las entradas en distintos grupos según el patrón que muestren,

para así poder incluir dentro de una u otra categoría los futuros datos que se deseen estimar

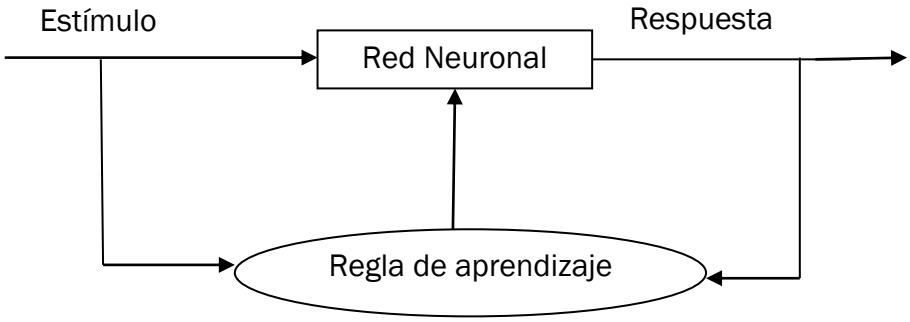


Fig. 2.10. Red No supervisada.

Finalización del entrenamiento

La etapa de entrenamiento finalizará cuando se cumpla una de las siguientes condiciones:

- **Número de iteraciones:**
  Al tratarse de un método iterativo, una condición de parada muy habitual es el número de iteraciones, cuando se sobrepasa un determinado límite de iteraciones, el entrenamiento concluirá.

- **Error aceptable:**
  La parada por error se puede aplicar en métodos de entrenamiento supervisado, cuando se consiga que el error entre la salida adquiera un determinado valor, se fijarán los pesos dados en esa situación y concluirá el entrenamiento.

- **Estabilidad del peso:**
  Durante el entrenamiento, se puede determinar un límite de estabilidad, los pesos de las funciones irán variando para adaptar la salida al target esperado, podemos cesar el entrenamiento cuando los pesos se mantengan estables durante un determinado número de iteraciones.

Fase de validación

Aunque la fase de entrenamiento haya conseguido uno valores de error bajos, no quiere decir que la red neuronal esté preparada para trabajar correctamente.

Puede ocurrir que durante el entrenamiento se produzca un sobreajuste, esto consiste en que la red es capaz de estimar correctamente los valores con los que se ha entrenado, pero produce fallos al enfrentarse a valores nuevos. Por este motivo hemos reservado una pequeña cantidad de la información, la cual no hemos utilizado en la fase de entrenamiento para comprobar que no se produce este sobreajuste en la red.

Tipos de RNA

Como hemos visto, en función de la forma en la que se organizan las neuronas, cómo aprenden o el número de capas que componen la red, tendremos distintos tipos de RNA, siendo algunas de las más comunes las que se listan a continuación:

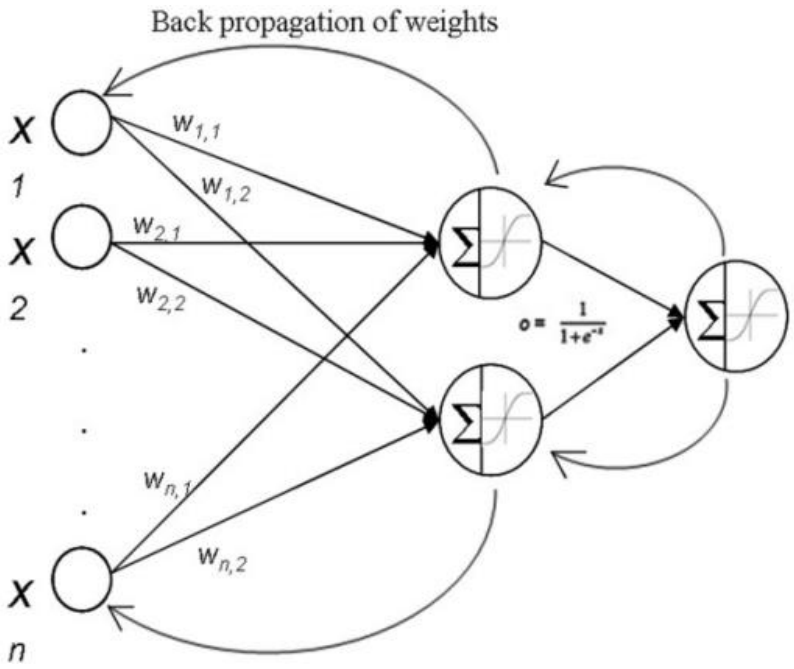
- El Perceptrón (Hebb, 1949).
- La Red de Hopfield (Hopfield, 1985).
- Red Neuronal Competitiva simple (McCulloch & Pitts, 1943).
- Redes Neuronales Online ART1 (Grossberg & Carpente,1980).
- Redes Neuronales Competitivas ART2 (Grossberg & Carpente,1980).
- Redes Neuronales Autoorganizadas: Mapas de Kohonen (Kohonen, 1982).

En los siguientes apartados se describen dos de los algoritmos de RNA más utilizados, el Perceptrón multicapa (MLP) usando Backpropagation y la red de base radial (RBF).

2.4.2. Perceptrón multicapa usando Backpropagation (MLP)

Una de las redes neuronales más utilizada en la actualidad, a la cual recurriremos en este trabajo, se trata del Perceptrón multicapa basada en el algoritmo de entrenamiento de propagación hacia atrás, *backpropagation (*Paul Werbos, 1974).

Se trata de un algoritmo de entrenamiento supervisado, el funcionamiento de este consiste en registrar el error producido entre la salida y el valor objetivo deseado, de forma que esta información regrese hacia atrás en la red para optimizar el proceso, variando los pesos asociados a la función de activación y corregir el error final producido.



(Fig. 2.11). Estructura del MLP Back propagation

Estructura del MLP usando Backpropagation

Generalmente este tipo de red multicapa está compuesta por tres capas: Capa de entrada, capa oculta y capa de salida, para calcular y minimizar el error producido en la estimación, el entrenamiento Backpropagation recurre a las siguientes ecuaciones(Valencia Reyes, Yáñez Márquez, & Sánchez Fernández, 2006):

El error en la salida se mide con el método del error cuadrático medio (ECM), el cual nos devuelve un estimador de la desviación al cuadrado existente entre el valor estimado y el valor real esperado:

$$E(\boldsymbol{x}; \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}) = \frac{1}{2}\sum_{i=1}^{m}(t_i - o_i^{(2)})^2$$

(Ec. 2.21)

Siendo $W^{(1)}$el peso que une la última capa oculta con la capa de salida y $W^{(2)}$ el peso que une la capa de entrada con la primera capa oculta.

A continuación se calculan las derivadas parciales del error con respecto a los pesos $W^{(1)}$ y respecto a los pesos $W^{(2)}$:

$$\Delta \boldsymbol{W} = -\gamma \frac{\partial E(\boldsymbol{W})}{\partial \boldsymbol{W}}$$

(Ec. 2.22)

Donde $0 < \gamma < 1$ es un parámetro conocido como factor de aprendizaje.

En proporción a la derivada del error, se ajustan los pesos de la función de salida. Por último, el error producido en la capa oculta dependerá de los términos de error de la capa de salida. Por esto recibe el nombre de "Propagación hacia atrás"

Como comentamos, para actualizar los pesos se utiliza un algoritmo recursivo, empezando por las neuronas pertenecientes a la capa de salida y trabajando hacia atrás hasta llegar a la capa de entrada, ajustando los pesos de la siguiente forma(Berzal Galiano, 2004):

- **En la capa de salida**:

$$W_{kj}^{0}(t+1) = W_{kj}^{0}(t) + \Delta W_{kj}^{0}(t+1) \qquad \textit{(Ec. 2.23)}$$

$$\Delta W_{kj}^{0}(t+1) = \alpha \cdot \delta_{pk}^{0} \cdot y_{pj} \qquad \textit{(Ec. 2.24)}$$

- **En la capa oculta:**

$$W_{ji}^{h}(t+1) = W_{ji}^{h}(t) + \Delta W_{ji}^{h}(t+1) \qquad \textit{(Ec. 2.25)}$$

$$\Delta W_{ji}^{h}(t+1) = \alpha \cdot \delta_{pj}^{h} \cdot x_{pi} \qquad \textit{(Ec. 2.26)}$$

El proceso de autoajustes se repetirá hasta que el error adquiera un valor aceptable, o se alcance alguno de los criterios de parada mencionados anteriormente.

2.4.3. Red neuronal artificial de base radial (RBF)

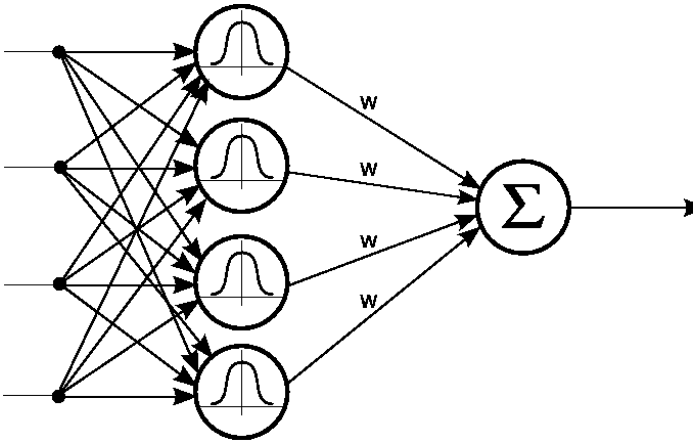
Las redes neuronales de base radial son sistemas multicapa de conexiones hacia delante. Están formadas por tan solo una capa oculta, en esta capa oculta las neuronas se activan cada una en una región distinta del espacio de entrada. Las neuronas de la capa de salida reaccionan a las activaciones de las neuronas ocultas realizando una combinación lineal de estas.

Las neuronas de esta red tienen carácter local, cada neurona crea una aproximación local en una región determinada, sin necesidad de ser lineales dichas

aproximaciones, las neuronas de la capa de salida realizan combinaciones lineales de estas.
El origen de estas redes se debe principalmente a Moody y Darken (1989), Renals (1898), Poggio y Girossi (1990).

Estructura de la red neuronal RBF

La arquitectura de estas redes concuerda con la estructura genérica de una red neuronal, estando compuesta por: Capa de entrada, Capa oculta y Capa de salida. A diferencia de la Backpropagation, que podía tener varias capas ocultas, la RBF tan solo dispone de una(Valls, 2007).



(Fig. 2.12). Estructura RBF

La activación de las neuronas de salida y el patrón de entrada n se puede expresar de la siguiente forma:

$$y_k(n) = \sum_{i=1}^{m} w_{ik} \cdot \emptyset_i(n) + u_k \qquad \textit{(Ec. 2.27)}$$

Para *k* = 1, 2, ..., *r*.

- w_{ik}: peso de la conexión de la neurona oculta i a la de salida k.
- ϕ i(n): activación de la neurona oculta i.
- u_k : umbral de la neurona de salida k.

Las funciones de base radial determinan las activaciones de las neuronas ocultas en función del vector de entrada:

$$\emptyset_i(n) = \emptyset\left(\frac{|X(n)-C_i|}{d_i}\right) \; para \; i = 1,2,\dots,m \qquad \textit{(Ec. 2.28)}$$

$$|X(n) - C_i| = \sqrt{\sum_{j=1}^{p}(x_j(n) - c_{ij})^2} \qquad \text{(Ec.2.29)}$$

- ϕ es una función de base radial.
- C_i=(ci1, c12, ...c_{ip}) son vectores: centros de las funciones de base radial.
- d_i son números reales: desviaciones de las funciones.
- || . || es la distancia euclídea desde el vector de entrada al centro de la función.

Método de aprendizaje

El aprendizaje en la RBF consiste en determinar los centros, desviaciones y pesos de la capa oculta a la capa de salida.

Para optimizar el proceso, el cálculo de estos parámetros se realiza dividiendo los parámetros en función de la capa a la que pertenezcan (Entrada, Oculta o Salida). De esta forma, los centros y las desviaciones se obtienen mediante un proceso guiado por la optimización del espacio de entrada, y los pesos se optimizan en función de las salidas que se desean obtener.

Los dos métodos de aprendizaje más utilizados son el método híbrido y el método totalmente supervisado (Valls, 2007), siendo el primero el más común. El aprendizaje híbrido se realiza en dos fases:

- **Fase no supervisada:**
  En esta fase se determinan los centros y amplitudes de las neuronas de la capa oculta con el objetivo de agrupar el espacio de entrada en diferentes clases. El representante de cada clase será el centro de la función de base radial y la desviación vendrá dada por la amplitud de cada clase.

- **Fase supervisada:**
  En esta, se determinan los pesos y los umbrales de la capa de salida

Para determinar los centros y poder minimizar las distancias euclídeas entre los patrones de entrada y el centro más cercano, se utiliza un algoritmo de clasigicación no supervisado conocido como “Algoritmo de K-medias”.

$$J = \sum_{i=1}^{K}\sum_{n=1}^{N} M_{in} \cdot |X(n) - C_i| \qquad \text{(Ec. 2.30)}$$

Donde N es el número de patrones, || . || es la distancia euclídea, X(n) es el patrón de entrada n y Min es la función de pertenencia, que vale 1 si el centro Ci es el más cercano al patrón X(n), y 0 en otro caso, es decir:

$$M_{in} = \begin{Bmatrix} 1 \; si \; |X(n) - C_i| < |X(n) - C_s| \\ 0 \; en \; otro \; caso \end{Bmatrix} \qquad (Ec.\ 2.31)$$

Siendo "s" distinto de "i", con valores comprendidos entre: s=1, 2, ..., K.

El algoritmo de K-medias suele ser bastante eficiente, convergiendo en pocas iteraciones hacia un mínimo de la función J. Su principal inconveniente, añade Valls, se trata de la dependencia de los valores iniciales asignados a cada centroide, ya que estos pueden tratarse de mínimos locales, llegando a resultados erroneos.

2.5. Cálculo del error de estimación

Existe una gran variedad de métodos para cuantificar el error producido entre un valor estimado y el valor real, alguno de los más utilizados son el *Mean Square Error* (MSE) y el *Mean Absolute Error* (MAE), cuyas expresiones se muestran a continuación (James & Witten, 2013):

Mean Square Error

Esta medida del error calcula el promedio de la diferencia al cuadrado producida entre el valor estimado y el valor esperado:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \qquad (Ec.2.32)$$

Mean Absolute Error

La diferencia con el MSE reside en que este método es menos penalizante con los casos en los que existe una gran diferencia entre el valor estimado y el valor real, es decir, este método suaviza el error adicional introducido por los valores atípicos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{Y}_i - Y_i|$$

(Ec. 2.33)

Cross Validation: K-fold

Con frecuencia, los procesos de estimación requieren una gran cantidad de datos con los que ensayar los modelos y comprobar su validez, para esto se recurre al método "*Cross Validation*"

Este método consiste en dividir el conjunto de datos en dos grupos, Grupo de Entrenamiento y Grupo de Validación. Del total de la muestra seleccionamos aleatoriamente dos tercios para el grupo de entrenamiento y reservamos el tercio restante para componer el grupo de test.

- Grupo de Entrenamiento (*Train*): 1/3 del total de los datos.
- Grupo de Validación (*Test*): 2/3 del total de los datos.

Al finalizar el entrenamiento y la validación, obtendremos un valor de error de entrenamiento (ErrorTrain1) y otro valor de error de test (ErrorTest1), archivamos estos valores y volvemos a repetir el proceso hasta *K* veces, utilizando otro grupo de entrenamiento y de validación, de nuevo aleatorios, pero distintos a los utilizados anteriormente.

Gracias a esto estamos aumentando K-veces el tamaño de la muestra. Finalmente se calcula la media de los errores obtenidos en la validación y el entrenamiento:

$$ErrorTrain = \frac{ErrorTrain1 + ErrorTrain2 + \ldots + ErrorTrainK}{K}$$

(Ec. 3.18)

$$ErrorTest = \frac{ErrorTest1 + ErrorTest2 + \ldots + ErrorTestK}{K}$$

(Ec.3.19)

En el capítulo siguiente podemos encontrar el listado con las configuraciones de red y el preprocesamiento utilizado para estimar cada una de las variables RCD con el menor MSE.

3. TRABAJO EXPERIMENTAL

Los métodos utilizados en este capítulo parten de la base teórica descrita en el capítulo anterior.

En este capítulo se va a describir la metodología que se ha seguido para realizar este estudio y alcanzar los resultados que se muestran en el capítulo 4.

3.1. Metodología

En este capítulo se describe con detalle las decisiones tomadas para llegar a los resultados finales del trabajo. La metodología seguida se puede resumir con el esquema mostrado en la Fig. 3.1.

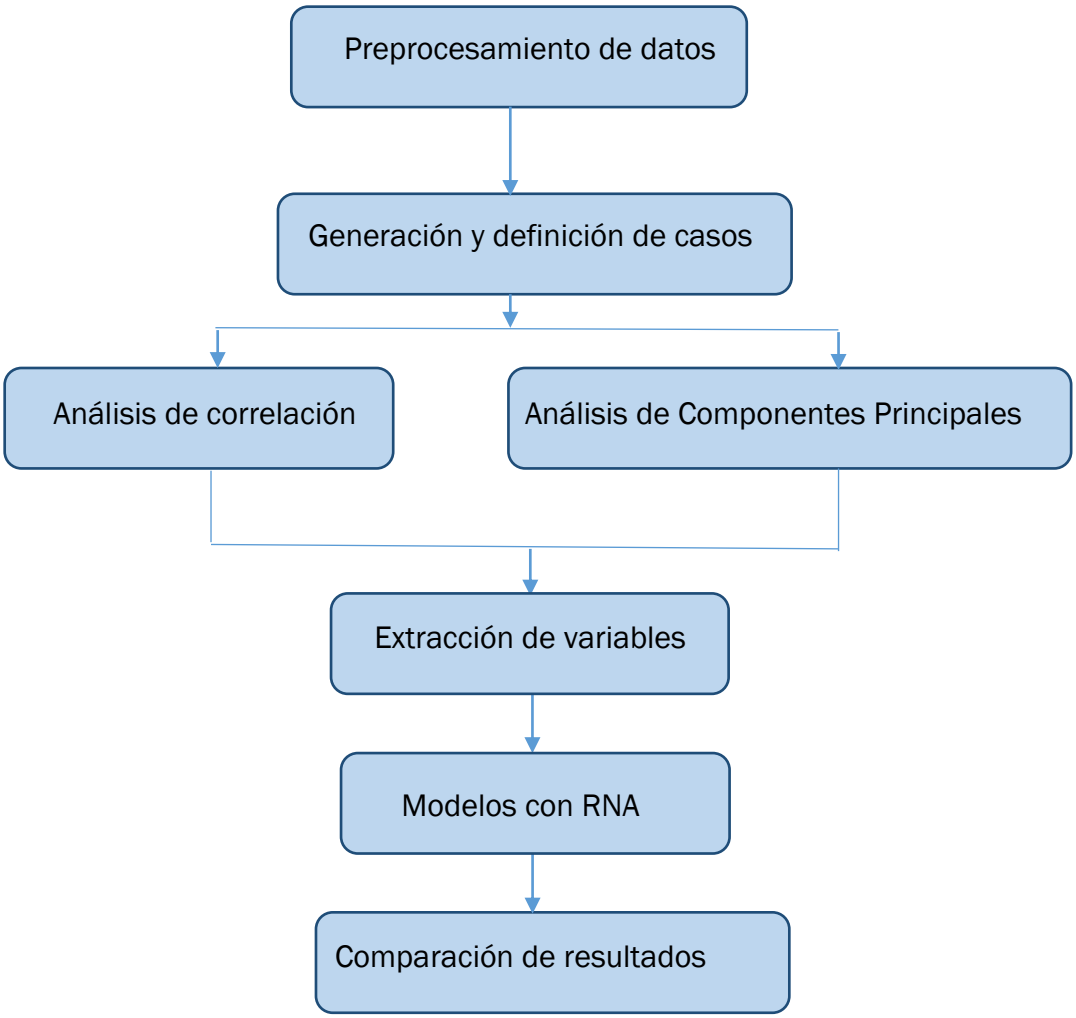


Fig. 3.1. Esquema metodológico.

En primer lugar, se debe atender a los problemas de datos faltantes, *Outliers*, normalización y dimensionalización, por ello someteremos el muestreo original a un preprocesamiento con el que se definirán distintos casos de estudio, en función del tipo de preprocesamiento que se haya aplicado a cada caso.

Una vez se hayan preparado los datos procedemos a estudiarlos mediante el análisis de correlación y el análisis de componentes principales, los cuales permitirán la selección y extracción de variables, y serán el punto de partida para poder realizar modelos de estimación con Redes Neuronales Artificiales.

Tras obtener los modelos de estimación de cada caso de estudio, se realizará una comparación de los resultados, en la que seleccionaremos las mejores opciones para estimar cada variable y se decidirá cuál ha sido el preprocesamiento idóneo para cada una de ellas.

Descripción de datos disponibles

Para realizar este trabajo disponemos de un muestreo de datos compuesto por 64 individuos y 92 variables, recogidos a mano desde distintas obras de construcción civil.

El conjunto de datos inicial se puede dividir en dos grupos en función de la información que aportan sus variables:

	Variables de Residuo	Variables de Características de la Obra
Muestreo: 64 obras	79	13

Tabla 3.2. Datos iniciales del muestreo.

- **Variables de Residuos:**
  Contiene información acerca de la cantidad y el tipo residuo que se ha generado en la construcción y demolición de cada obra, estas variables son, por ejemplo: cantidad de hormigón, residuo metálico, cubierta, cerramiento, pintura utilizada…etc.
  En total este muestreo está compuesto por 64 muestras y 79 variables.

- **Muestreo de características de obra:**
  Contiene información descriptiva de cada individuo de estudio: Localización, tipología, superficie de la parcela, número de edificios, viviendas, portales, trasteros, locales, plazas de garaje, duración, volumen de excavados. Utilizaremos la información aportada por estas variables para enriquecer las muestras de Residuos, analizando la correlación de estas variables con el muestreo de residuos y aplicando los métodos de enriquecimiento de datos que desarrollamos a continuación.
  En total este muestreo contiene 64 muestras y 13 variables.

Análisis de errores en los datos de partida

El conjunto de datos iniciales contiene una gran cantidad de errores debido a que la introducción de datos ha sido realizada por distintas personas y en distintas localizaciones, de modo que deben ser analizados y procesados para evitar sesgos en los resultados del trabajo. Los errores identificados en el muestreo inicial son los siguientes:

- Variables mudas:
  La mayor parte de las variables no contienen información o la información que contienen es insuficiente para aplicar métodos de imputación numérica.

- *Missing Values:*
  Existe una gran cantidad de datos perdidos, con mediciones NaN (*Not a Number*) o cero, en variables que deberían contener información para todas las obras de la muestra. Se debe analizar sobre qué variables es posible realizar tratamientos de imputación numérica de datos faltantes.

- *Outliers:*
  Se observan datos no concordantes con el resto de datos de la muestra.

- Dimensiones de medida poco uniformes:
  La alta variedad de obras registradas hace imposible la realización de una comparación entre estas. En el muestreo hay registros de obras de gran magnitud junto a obras ínfimas en comparación, se debe dimensionalizar el conjunto muestral para poder relacionar entre sí todas las obras.

- Datos no normalizados:
  Los análisis de datos que se van a utilizar en este trabajo requieren que los datos de partida estén acotados en una escala fija, para ello se debe normalizar el muestreo.

3.2. Preprocesamiento de datos

A continuación, se detalla la metodología a seguir para solventar los problemas derivados de la recopilación de datos mostrados en el apartado anterior.

3.2.1. Variables mudas

El primer problema con el que nos encontramos es el elevado número de datos faltantes.

Existe una gran variedad de métodos de enriquecimiento numérico, como se ha descrito en el apartado teórico, mediante métodos de imputación numérica trataremos de enriquecer la muestra, aun así, existen variables con muy poca información, para las cuales nos es imposible comprender su comportamiento. De tal forma que para evitar sesgos producidos por variables con una gran cantidad de datos faltantes y facilitar el estudio de las variables restantes se procede a eliminar todas aquellas variables que se consideren mudas.

Consideraremos variable muda, a todas aquellas que contenga un 50% de datos faltantes. Para esto, contamos el número de muestras no nulas de una variable y lo dividimos por el número total de muestras de la misma, si este cociente alcanza un índice superior o igual a 0,5, la variable contiene información suficiente para trabajar con ella, en caso contrario se considerará Variable Muda (VM).

$$VM\ si \quad \frac{Nº Datos\ faltantes\ de\ la\ variable}{Nº Obras\ (64)} > 0{,}5$$

(Ec. 3.1)

En total se eliminan por esta razón 65 variables de las 95 con las que partíamos. En la tabla 3.2 podemos ver la situación del muestreo tras eliminar las variables mudas (VM).

	Variables de Residuo	Variables de Características	Nº Obras
Muestreo original	79	13	64
Muestreo sin V.M.	15	11	64

Tabla 3.3

Tras eliminar las variables mudas, el muestreo se compone de 64 obras caracterizadas con las variables que se muestran en la tabla 3.3:

Muestreo RCD sin V.M.	Muestreo Características sin V.M.
RCD Total	Localización
CIM_ZapatasYMuros_m3	Sup_Parcela_m2
CIM_HormLimpieza_m3	Tipología
EST_Hormigon_m2	N_Edificios
CUB_PlanaTransit_m2	N_Viviendas
CUB_NoTransitable_m2	N_Portales
PART_Ladrillo_m2	N_Trasteros
REV_Aplacado_m2	N_Locales
REV_Enlucido_m2	N_Plz_Garaje
REV_Enfoscado_m2	Duración
PINT_Plastica_m2	Excavacion_m3
SOL_Ceramico_m2	
CINT_Madera_m2	
CEXT_Aluminio_m2	
CEXT_Vidrio_m2	

Tabla 3.4. Variables que componen el muestreo.

3.2.2. Detección de *Outliers*

En el muestreo, por las mismas razones por las que encontrábamos valores faltantes, encontramos valores que difieren de forma sustancial del resto de puntos de la muestra.

Los *Outliers* pueden ser elementos no representativos del conjunto de las muestras, los cuales pueden distorsionar el comportamiento de los estadísticos que se estudien en ella.

Puesto que no podemos recurrir a la fuente para comprobar y corregir estos valores, procedemos a analizar las muestras con el fin de aislar los *Outliers* y comparar posteriormente los resultados obtenidos al utilizar una muestra manteniendo los candidatos a *Outliers*, frente a la utilización de una muestra en la que los hemos eliminado.

Comenzaremos el estudio de los datos atípicos realizando una distinción de grupos según su tipología, del mismo modo que veíamos en el apartado Imputación Numérica, dividiendo el muestreo en dos casos:

- Caso A: Obras de tipología Residencial.
- Caso B: Obras de tipología Comercial.

Mediante el método *Box-plot* examinaremos las variables RCDtotal, Superficie de parcela y Excavados, con el fin de identificar aquellas obras que muestren una clara discordancia con el resto de valores del muestreo.

Método *Box-plot*

El método seguido para considerar candidatos a *Outlier* consiste en calcular la diferencia entre el primer y tercer cuartil Q3-Q1 (valor intercuartil), se considerará valor atípico leve aquel que se encuentre a 1,5 veces esta distancia y valores atípicos extremos aquellos que lo superen 3 veces.

$$Valor\ Intercuartil = Q_3 - Q_1$$

(Ec.3.2)

En este trabajo obviaremos los *Outliers* leves, manteniéndolos en el muestreo, y eliminaremos los *Outliers* graves, excluyendo de la muestra las obras que superen el límite superior o sean menores al límite inferior.

$$LimInf = Q_1 - 3 \cdot (Q_3 - Q_1)$$

(Ec. 3.3)

$$LimSup = Q_3 + 3 \cdot (Q_3 - Q_1)$$

(Ec. 3.4)

De nuevo realizaremos un estudio condicionado, dividiendo las muestras en dos grupos, el primero sin atender a la tipología de la obra y el segundo estará formado tan solo por obras de tipo residencial.

- Caso A: Sin distinguir las obras por su tipología.

- Caso B: Obras de tipo Residencial.

Caso A: Sin distinguir las obras por su tipología (64 obras).

	RCD	Superficie	Excavados
Q1	773	2232,2575	8164,25
Q2	1535	3596,105	13467,13
Q3	2259	6085,7025	23083,825
LimInf	-1456	-3547,91	-14215,1125
LimSup	4488	11865,87	45463,1875

Tabla. 3.5. Límites para definir Outliers Caso A.

Con este método analítico encontramos la obra Nº13 (RCDtotal=9905m^3) el cual excede el límite superior en la variable RCDTotal.

A continuación, repetimos el método, seleccionando esta vez sólo los individuos de tipología Residencial.

- **Caso B**: Aplicado solo a las obras de tipo residencial (49 obras).

Aplicando el método clásico, calculamos el rango intercuartil de cada variable. Determinamos el límite inferior y superior por el cual, fuera de esta región consideraremos cualquier dato como atípico:

	RCD	Superficie	Excavados
Q1	0	1705,44	8964,96
Q2	1161	3400	13988,79
Q3	1918,945	4334,505	23244,55
LimInferior	-2878,4175	-2238,1575	-12454,425
LimSuperior	4797,3625	8278,1025	66083,32

Tabla. 3.6. Outliers Caso B.

De esta forma tendremos 2 nuevos individuos atípicos:

- Nº52, con un valor de 8880 m^2 en la variable Superficie, sobrepasa el límite superior y se posiciona como Outlier grave.

- Nº54, con un valor de 9400m^2 en la variable Superficie, excede el límite superior impuesto para ello.

Casos de estudio: Muestreos Con o Sin *Outliers*

Hemos detectado un total de tres muestras atípicas, las obras número 13, 52 y 54. A continuación se genera una réplica del muestreo original, con la diferencia de que, en este nuevo caso de estudio, eliminaremos los *Outliers*. Del mismo modo que hacíamos en el apartado de imputación numérica reservaremos estos casos de estudio para su posterior comparación.

- Caso de estudio: Muestreo Con *Outliers.*
- Caso de estudio: Muestreo Sin *Outliers.*

3.2.3. Imputación numérica: *Missing Values*

Al realizar el análisis de cualquier muestreo es muy probable que nos encontremos con individuos en los que faltan valores de algunas variables. Esto puede ser debido a que la muestra ha sido construida por personas distintas, con métodos distintos, o simplemente que los datos faltantes no han sido encontrados a la hora de formar la muestra.

Puesto que existen obras en las que puede no haberse empleado algún material, se considera que someter a todas las variables del muestreo a imputación numérica, sería un grave error que condicionaría los resultados obtenidos a partir de esto. Por lo tanto, sólo se aplicarán los métodos de imputación a las variables en las que no se pueda justificar un valor de cero en cualquiera de las obras.
Dicho esto, las variables que podremos enriquecer mediante imputación numérica son:

- **RCDtotal:** Cantidad total (en volumen) de todos los residuos registrados en cada obra.
- **Superficie de Parcela:** Magnitud (en metros cuadrados) de la extensión de la parcela.
- **Volumen Excavado:** Magnitud (en metros cúbicos) del terreno removido de la parcela para realizar la edificación.

Valores de cero en estas variables solo indican que se trata de *Missing Values*, pues todas las obras generan residuo, disponen de una superficie de construcción limitada y requieren de extracción de tierras para su cimentación.

En este estudio desconocemos la fuente de la que proceden los datos y nos es imposible comprender el origen de los *Missing Values,* con lo que estamos sometidos a datos faltantes completamente aleatorios (MCAR), por lo que probablemente las técnicas de imputación multivariada desarrolladas por Rubin *et. al*. no serían determinantes.
Por solventar este problema se recurre a dos métodos clásicos:

- **Imputación por media no condicionada:**
  Se sustituyen todos los *Missing Values* de cada variable por el valor medio de dicha variable, sin hacer distinciones entre obras y sin tener en cuenta el resto de datos del muestreo.

- **Imputación por regresión lineal condicionada:**
  Antes de aplicar el método de imputación, se dividen las obras del muestreo según su tipología (Residencial o Comercial), a cada uno de estos grupos se le asigna una recta de regresión distinta.

Método de imputación por media no condicionada

Utilizaremos una imputación por media no condicionada, es decir, sin distinguir las obras por su tipología, ya sean de tipo Residencial o Comercial, calcularemos la media de cada variable y sustituiremos este valor en todos los *Missing Values* de cada una de ellas.

Método de imputación por regresión lineal condicionada

Este método consiste en calcular la recta de regresión lineal entre un par de variables (que estén relacionadas). Una vez se hayan determinado los coeficientes de la recta mediante mínimos cuadrados, podremos interpolar los valores faltantes de las variables objetivo.

Para realizar este proceso utilizaremos la información que aporta el resto de variables. Se observa que la variable RCDtotal está directamente relacionada con las variables Excavados y Superficie de parcela, por ejemplo, valores elevados en las variables “Excavados” y “Superficie” indican que se trata de una obra de grandes dimensiones, con lo que su valor de RCDtotal será acorde a ellos.

Además, en este método de imputación realizaremos un proceso condicionado, diferenciando las obras según su tipología, pudiendo ser obras de tipo Comercial o Residencial.

- Caso A: Obras de tipología Comercial.
- Caso B: Obras de tipología Residencial.

Las rectas de regresión lineal con las que calcular los valores de imputación, obtenidas por mínimos cuadrados, para cada uno de los casos son las siguientes:

Caso A. Obras de tipología Residencial:

$$Excavados = 4{,}6796 \cdot RCDtotal + 9341 \quad \textit{(Ec.3.5)}$$

$$Superficie = 0{,}4977 \cdot RCDtotal + 4091 \quad \textit{(Ec.3.6)}$$

Caso B. Obras de tipología Comercial:

$$Excavados = 1{,}2822 \cdot RCD + 6854 \quad \textit{(Ec.3.7)}$$

$$SUPERFICIE = 0{,}4853 \cdot RCD + 6088 \quad \textit{(Ec. 3.8)}$$

Casos de estudio: Imputación por Media o por Regresión

Aplicando estos métodos al muestreo original definiremos dos casos que estudiaremos posteriormente de forma independiente para comparar los resultados.

- Caso de estudio: Muestreo imputación por Media.
- Caso de estudio: Muestreo imputación por Regresión.

3.2.4. Dimensionalización

El siguiente paso será estandarizar los valores de las variables del muestreo de residuos, de tal forma que podamos comprender la extensión o la magnitud de cada uno de ellos y podamos relacionarles entre sí satisfactoriamente. Para esto nos vamos a servir del muestreo de características, pues contiene la información necesaria para comprender el tamaño de cada obra.

$$RCD_{ij} = \frac{Unidades\ de\ Residuo_{ij}}{Magnitud\ de\ Obra_j} \quad \textit{(Ec. 3.9)}$$

Los índices "i" y "j" indican la variable y la obra del muestreo, siendo i= 1, 2, …, 15. Y j=1, 2, …, 64.

La variable Volumen de Excavados (m3) va a ser nuestra "Magnitud de obra". De entre todas las variables características de la muestra, el volumen de excavados es la que mejor define el tamaño de la obra ya que contiene en sí misma a la variable Superficie de la obra y a una variable oculta a la que llamaremos "Profundidad (m)" por lo que será suficiente para determinar si una obra es grande o pequeña.

El proceso de estandarización consistirá en escalar las variables de residuos con su respectiva magnitud de obra, para ello dividimos el valor de cada residuo entre el volumen de excavados correspondiente en cada obra:

$$RCD_{ij} = \frac{[RCDij]}{[EXCAVADOSj]} \quad \textit{(Ec.3.10)}$$

Así las variables de toda la matriz de residuos tendrán valores estandarizados, lo cual será muy útil para comparar todas las obras entre sí, es decir, las obras grandes estarán divididas por un valor alto de "Excavados" y las obras pequeñas estarán divididas por un valor más bajo, así, independientemente de que se trate de obras grandes o pequeñas, todos los individuos tendrán valores comprendidos en un rango similar para cada obra.

Las unidades de medida originales de los residuos vienen dadas en metros cuadrados o metros cúbicos, dependiendo de cada variable, después de la estandarización las unidades serán:

$$\mathrm{RCD_{ij}} = \frac{Unidades\ superficiales\ de\ RCD\ [m2]}{Magnitud\ de\ obra\ [m3]}$$

(Ec.3.11)

$$\mathrm{RCD_{ij}} = \frac{Unidades\ volumétricas\ de\ RCD\ [m3]}{Magnitud\ de\ obra\ [m3]}$$

(Ec. 3.12)

3.2.5. Normalización

El último paso del pretratamiento, antes de proceder con los sucesivos análisis consiste en normalizar las variables del muestreo de Residuos. Los métodos de análisis que aplicaremos necesitan que las variables estén formadas por valores contenidos en un mismo rango, para ello utilizaremos el método "zscore" y el método "min-max". La utilización de un método y otro de nuevo hará que generemos nuevos casos de estudios independientes, los cuales utilizaremos para comparar los resultados obtenidos.

Zscore

Aplicamos el método a cada muestra del conjunto de datos de Residuos, siendo "*Y*" el valor de la variable iesima de la obra j, "μ", la media de la variable iesima y "σ" la desviación típica de la variable iesima

$$Zi = \frac{Yi - \mu i}{\sigma i}$$

(Ec. 3.13)

Como se ve en el capítulo de aspectos teóricos, con este procedimiento obtenemos una relación de valores cuya media será cero (μ=0) y su desviación típica será la unidad (σ=1)

Procedimiento min-max

Aplicamos la siguiente fórmula a los datos del muestreo de Residuos, siendo "*Ymin*" el valor mínimo de cada variable y "*Ymax*" el valor máximo.

$$Z = \frac{Y - Ymín}{Ymáx - Ymín}$$

(Ec.3.14)

Este procedimiento corresponde a una linealización con ordenada en el origen, obtenemos un conjunto de valores comprendido en el intervalo [0,1], siendo 0 el valor mínimo de cada variable y 1 el máximo, el resto de valores se repartirá en el intervalo según la distancia que tengan a estos.

Casos de estudio: Normalización Zscore o Min-Max

Con este proceso se definen dos nuevos casos de estudio, dependiendo del tipo de normalización que hayamos aplicado al muestreo original, obtenemos:

- Caso de estudio: Muestreo normalizado mediante Zscore.
- Caso de estudio: Muestreo normalizado mediante Min-Máx.

3.2.6. Casos de estudio definidos en el Preprocesamiento de datos.

Llegados a este punto disponemos de 8 casos a los que aplicaremos los sucesivos estudios y análisis, todos estos se encuentran Normalizados y se han enriquecido con los tratamientos antes mencionados. En la tabla 3.5 podemos ver la configuración de preprocesamiento que ha recibido cada caso.

Caso	Nombre	Tipo de Normalización	Imputación numérica	Tratamiento Outliers
1	ZscMedConOut	zscore	Media	No
2	ZscRegConOut	zscore	Regresión	No
3	ZscMedSinOut	zscore	Media	Si
4	ZscRegSinOut	zscore	Regresión	Si
5	MinMedConOut	min-max	Media	No
6	MinRegConOut	min-max	Regresión	No
7	MinMedSinOut	min-max	Media	Si
8	MinRegSinOut	min-max	Regresión	Si

Tabla 3.5. Casos de estudio.

Realizaremos los análisis sucesivos sobre los conjuntos de datos definidos con estos ocho casos de estudio, con el fin de determinar cuáles son los preprocesamientos óptimos para este trabajo.

3.3. Selección y extracción de variables

En este apartado se someten los casos de estudio a análisis de correlación y a Análisis de Componentes Principales (PCA). Mediante estas técnicas podremos comprobar la relación existente entre las variables de los muestreos, definiendo

aquellas que contienen información redundante y aquellas que muestran claros signos de independencia.

Mediante el análisis de correlación podremos identificar aquellos residuos que muestren relación lineal, los modelos de estimación que se calculen para estos serán más sencillos que para aquellos residuos que no muestren relación con otros. Para estos últimos utilizaremos el PCA, mediante el cual identificaremos las variables más importantes del muestreo.

Análisis de correlación

El primer análisis que vamos a realizar consiste en determinar la relación lineal existente entre cada par de variables. Para ello, mediante software de cálculo generaremos los correspondientes análisis de correlación basadas en el coeficiente de Pearson y la correspondiente prueba de Student, con la que, mediante el p-valor, corroboraremos la confianza de dicho coeficiente.

Gracias a este análisis podemos definir dos grupos de variables:

- Variables de RCD que muestran correlación lineal con algún otro RCD.
- Variables de RCD que no muestran correlación lineal con ningún otro RCD.

Para nuestro estudio consideraremos que los pares de variables con un coeficiente de Pearson superior a 0,75 mantienen una correlación lineal alta, siempre y cuando la prueba de Student apruebe su valía. Los resultados de este análisis los podemos encontrar en el capítulo 4, Resultados y en el anexo: “Análisis de correlación”.

Selección de variables. PCA

Para nuestro estudio, el Análisis de Componentes Principales (*PCA*) no es un fin, sino un medio con el que comprender las relaciones existentes entre la información disponible en los conjuntos muestrales y disminuir en la medida de lo posible los datos que puedan ser redundantes para obtener una selección con las variables más representativas de la muestra. Los objetivos que se pretenden alcanzar mediante PCA son los siguientes:

- **Generar nuevas variables. Componentes Principales:**
  Al transformar los datos de un muestreo mediante PCA, se ordena la información en Componentes Principales, deshaciéndonos del significado de las variables originales de dicho muestreo.

- **Reducir la dimensionalidad del muestreo:**
  Las nuevas variables (Componentes Principales) están ordenadas según su autovalor asociado, siendo la primera Componente Principal la de mayor autovalor, y la última la que contiene el menor autovalor. Como hemos visto en el capítulo 2, los autovalores asociados a cada componente determinan la variabilidad recogida por cada una de ellas, lo cual nos va a permitir prescindir de aquellos componentes principales que contengan información redundante.

- **Seleccionar las variables originales más interesantes:**
  Con esta técnica, además, podemos determinar qué variables originales son las que más información aportan a las primeras Componentes Principales (las que menos información redundante contienen).

Componentes principales

Existe una gran variedad de herramientas informáticas, que de manera trivial nos devuelve la matriz de coeficientes, la variabilidad de cada componente (Autovalores asociados) y la transformación de los datos originales en las nuevas variables (Componentes principales).

Al trabajar con los datos transformados mediante PCA, se pierde el significado original de las variables, pero se gana precisión, ya que disponer de los muestreos ordenados mediante componentes principales nos permite realizar una selección directa de las variables más representativas del muestreo, siendo el primer componente principal el más importante, después el segundo…y sucesivamente hasta el último.

Reducir la dimensionalidad del muestreo

En primer lugar, debemos prestar atención a la variabilidad recogida por cada Componente Principal, esto lo podemos calcular mediante los valores propios asociados a estos.

Para obtener el valor porcentual de la varianza recogida por cada Componente Principal podemos utilizar la siguiente expresión:

$$Varianza_i(\%) = \frac{\lambda_i}{\sum \lambda_n}$$

(Ec. 3.14)

Siendo "λ_i" el valor propio del Componente Principal *i*, y "*n*" el número de Componentes Principales del muestreo.

Quedará de la mano del usuario conformarse con un determinado nivel de varianza, teniendo en cuenta que el número de Componentes Principales debe ser, a la vez, pequeño para que el análisis sea eficaz y suficiente para absorber la mayor parte de la información de las variables originales.

Para este estudio se considera que para alcanzar un nivel aceptable de representación del total de la muestra se debe trabajar con tantos Componentes Principales como sean necesarios para conseguir albergar el 70% de la varianza.

Con este estudio se determina que para cubrir el 70% de la varianza tan solo son necesarias tres componentes principales. En el capítulo 4, Resultados y Análisis se mostrará la varianza acumulada recogida por cada Componente Principal.

Seleccionar las variables originales más representativas.

El PCA refleja la redundancia de los datos de las variables en cada muestra, una vez hayamos determinado el número de CP suficiente para representar el muestreo, determinaremos el peso que tiene cada variable.
Si una variable muestra un peso elevado en los primeros componentes principales, significa que dicha variable contiene una gran cantidad de información única respecto al resto. Del mismo modo podremos ver que las variables pertenecientes al grupo que mostraba una alta relación lineal, obtendrán un peso bajo en el PCA.

Para conocer el peso o la influencia de, por ejemplo, la variable RCD_{total} de la muestra ZscMedSinOut, sumaremos los valores de los *scores* obtenidos en el PCA correspondientes a los componentes principales elegidos para ese caso.

Realizaremos este proceso con las 15 variables de cada muestreo del estudio y dividiremos el peso de cada variable por el peso total (la suma de los 15 pesos) para poder determinar cuáles de estas son las más influyentes en cada muestreo.

$$\%Peso_i = 100 \cdot \frac{Peso_i}{Peso\ Total} \qquad (Ec.3.15)$$

De esta forma podremos identificar las variables más interesantes (con mayor porcentaje en peso) de cada muestreo.

3.4. Modelos de estimación

Una vez se han preparado los datos de partida, definido los distintos casos de estudio y analizado la relación existente entre las variables de estos, se procede a calcular los modelos de estimación mediante Redes neuronales Artificiales (RNA).

Configuración de Red Neuronal Artificial

A continuación, se generan modelos de predicción mediante RNA con dos de las arquitecturas de red más comunes:

- Arquitectura Backpropagation (BPN).
- Arquitectura Base Radial (RBF).

Para cada una de estas arquitecturas probaremos configuraciones diferentes, variando los campos que se detallan a continuación, de manera que podamos determinar posteriormente con qué arquitectura y configuración se han obtenido los mejores modelos de estimación.

		Parámetros	Rango de valores
Arquitectura	BPN	Algoritmo de entrenamiento	LM/GC
		Función de entrada	line/tan/log
		Función de salida	line/tan/log
		Nº Neuronas en capa oculta	10/20/30/40
	RBF	Spread	0,01/0,1/1/10
		Nº Neuronas en capa oculta	10/20/30/40

Tabla. 3.6. Parámetros y rango de pruebas de modelos de predicción.

Arquitectura de Red Backpropagation

Utilizaremos una red neuronal del tipo Perceptrón Multicapa de arquitectura Backpropagation sobre la que probaremos 72 configuraciones distintas como se detallan a continuación:

- **Método de entrenamiento:**
  Probaremos con dos algoritmos de entrenamiento: Levenberg Marquardt (LM) y Gradiente Conjugado (GC).

  Se determina que la etapa de entrenamiento se dará por concluida cuando se encuentre alguno de los siguientes casos:

  - Si se realizan 2000 iteraciones.
  - Si se detecta que el error entre la salida y el target es igual a cero.
  - Si se detecta que los pesos de la función se mantienen completamente estables durante un cierto número de iteraciones.

- **Función de entrada:**
  Probaremos tres funciones de activación: Función Lineal (line), Función Tangente Sigmoidea (tan) y Función Logaritmo sigmoidea (log).

- **Función de salida:**
  Al igual que en la entrada, probaremos tres funciones distintas en la salida: Función Lineal (line), Función Tangente Sigmoidea (tan) y Función Logaritmo sigmoidea (log).

- **Número de neuronas en la capa oculta:**
  Variaremos el número de neuronas utilizadas en la capa oculta, variando este rango en cuatro pruebas: 10, 20, 30 ó 40 Neuronas.

En función de estos parámetros obtenemos 72 posibles configuraciones para la arquitectura de red Backpropagation. En las tablas siguientes se describen los parámetros utilizados en cada una de estas configuraciones:

- Ensayos con 10 neuronas en la capa oculta.

Configuración	Función de activación en capa oculta	Función de activación en capa de salida	Algoritmo de entrenamiento	Número de neuronas
1-N10	Lineal	Lineal	Levenberg Marquardt	10
2-N10	Lineal	Lineal	Gradiente conjugado	10
3-N10	Lineal	Tangente Sigmoide	Levenberg Marquardt	10
4-N10	Lineal	Tangente Sigmoide	Gradiente conjugado	10
5-N10	Lineal	Logaritmo Sigmoide	Levenberg Marquardt	10
6-N10	Lineal	Logaritmo Sigmoide	Gradiente conjugado	10
7-N10	Tangente Sigmoide	Lineal	Levenberg Marquardt	10
8-N10	Tangente Sigmoide	Lineal	Gradiente conjugado	10
9-N10	Tangente Sigmoide	Tangente Sigmoide	Levenberg Marquardt	10
10-N10	Tangente Sigmoide	Tangente Sigmoide	Gradiente conjugado	10
11-N10	Tangente Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	10
12-N10	Tangente Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	10
13-N10	Logaritmo Sigmoide	Lineal	Levenberg Marquardt	10
14-N10	Logaritmo Sigmoide	Lineal	Gradiente conjugado	10
15-N10	Logaritmo Sigmoide	Tangente Sigmoide	Levenberg Marquardt	10
16-N10	Logaritmo Sigmoide	Tangente Sigmoide	Gradiente conjugado	10
17-N10	Logaritmo Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	10
18-N10	Logaritmo Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	10

Tabla 3.7. Configuraciones BPN

- Ensayos con 20 neuronas en la capa oculta.

Configuración	Función de activación en capa oculta	Función de activación en capa de salida	Algoritmo de entrenamiento	Número de neuronas
1-N20	Lineal	Lineal	Levenberg Marquardt	20
2-N20	Lineal	Lineal	Gradiente conjugado	20
3-N20	Lineal	Tangente Sigmoide	Levenberg Marquardt	20
4-N20	Lineal	Tangente Sigmoide	Gradiente conjugado	20
5-N20	Lineal	Logaritmo Sigmoide	Levenberg Marquardt	20
6-N20	Lineal	Logaritmo Sigmoide	Gradiente conjugado	20
7-N20	Tangente Sigmoide	Lineal	Levenberg Marquardt	20
8-N20	Tangente Sigmoide	Lineal	Gradiente conjugado	20
9-N20	Tangente Sigmoide	Tangente Sigmoide	Levenberg Marquardt	20
10-N20	Tangente Sigmoide	Tangente Sigmoide	Gradiente conjugado	20
11-N20	Tangente Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	20
12-N20	Tangente Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	20
13-N20	Logaritmo Sigmoide	Lineal	Levenberg Marquardt	20
14-N20	Logaritmo Sigmoide	Lineal	Gradiente conjugado	20
15-N20	Logaritmo Sigmoide	Tangente Sigmoide	Levenberg Marquardt	20
16-N20	Logaritmo Sigmoide	Tangente Sigmoide	Gradiente conjugado	20
17-N20	Logaritmo Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	20
18-N20	Logaritmo Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	20

Tabla.3.8. Configuraciones BPN.

- Ensayos con 30 neuronas en la capa oculta.

Configuración	Función de activación en capa oculta	Función de activación en capa de salida	Algoritmo de entrenamiento	Número de neuronas
1-N30	Lineal	Lineal	Levenberg Marquardt	30
2-N30	Lineal	Lineal	Gradiente conjugado	30
3-N30	Lineal	Tangente Sigmoide	Levenberg Marquardt	30
4-N30	Lineal	Tangente Sigmoide	Gradiente conjugado	30
5-N30	Lineal	Logaritmo Sigmoide	Levenberg Marquardt	30
6-N30	Lineal	Logaritmo Sigmoide	Gradiente conjugado	30
7-N30	Tangente Sigmoide	Lineal	Levenberg Marquardt	30
8-N30	Tangente Sigmoide	Lineal	Gradiente conjugado	30
9-N30	Tangente Sigmoide	Tangente Sigmoide	Levenberg Marquardt	30
10-N30	Tangente Sigmoide	Tangente Sigmoide	Gradiente conjugado	30
11-N30	Tangente Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	30
12-N30	Tangente Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	30
13-N30	Logaritmo Sigmoide	Lineal	Levenberg Marquardt	30
14-N30	Logaritmo Sigmoide	Lineal	Gradiente conjugado	30
15-N30	Logaritmo Sigmoide	Tangente Sigmoide	Levenberg Marquardt	30
16-N30	Logaritmo Sigmoide	Tangente Sigmoide	Gradiente conjugado	30
17-N30	Logaritmo Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	30
18-N30	Logaritmo Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	30

Tabla. 3.9. Configuraciones BPN.

- Ensayos con 40 neuronas en la capa oculta.

Configuración	Función de activación en capa oculta	Función de activación en capa de salida	Algoritmo de entrenamiento	Número de neuronas
1-N40	Lineal	Lineal	Levenberg Marquardt	40
2-N40	Lineal	Lineal	Gradiente conjugado	40
3-N40	Lineal	Tangente Sigmoide	Levenberg Marquardt	40
4-N40	Lineal	Tangente Sigmoide	Gradiente conjugado	40
5-N40	Lineal	Logaritmo Sigmoide	Levenberg Marquardt	40
6-N40	Lineal	Logaritmo Sigmoide	Gradiente conjugado	40
7-N40	Tangente Sigmoide	Lineal	Levenberg Marquardt	40
8-N40	Tangente Sigmoide	Lineal	Gradiente conjugado	40
9-N40	Tangente Sigmoide	Tangente Sigmoide	Levenberg Marquardt	40
10-N40	Tangente Sigmoide	Tangente Sigmoide	Gradiente conjugado	40
11-N40	Tangente Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	40
12-N40	Tangente Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	40
13-N40	Logaritmo Sigmoide	Lineal	Levenberg Marquardt	40
14-N40	Logaritmo Sigmoide	Lineal	Gradiente conjugado	40
15-N40	Logaritmo Sigmoide	Tangente Sigmoide	Levenberg Marquardt	40
16-N40	Logaritmo Sigmoide	Tangente Sigmoide	Gradiente conjugado	40
17-N40	Logaritmo Sigmoide	Logaritmo Sigmoide	Levenberg Marquardt	40
18-N40	Logaritmo Sigmoide	Logaritmo Sigmoide	Gradiente conjugado	40

Tabla. 3.10. Configuraciones BPN.

Arquitectura de Red Neuronal de Base Radial.

Utilizaremos una red neuronal de Base Radial (RBF) sobre la que probaremos 16 configuraciones distintas como se detallan a continuación:

- **Número máximo de neuronas:**
  Se prueba la red con un rango de neuronas en la capa oculta de: 10, 20, 30 ó 40 neuronas.

- **Coeficiente *Spread*:**
  Este coeficiente determina la suavidad con la que el modelo se ajusta a los datos de entrenamiento, como ejemplo podemos observar las figuras 3.2 y 3.3. Probaremos la red con 4 valores de *Spread:* 0,01 / 0,1 / 1 / 10.

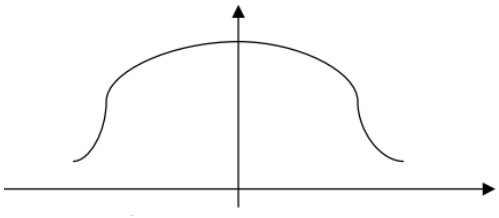


Fig.3.2 Spread alto

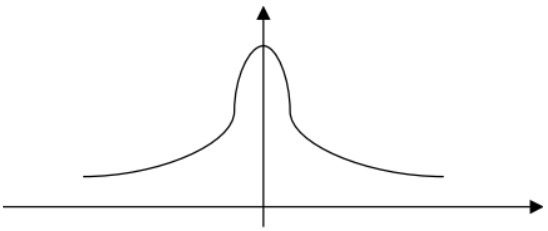


Fig. 3.3. Spread bajo

Es aconsejable mantener fijos el resto de parámetros de la red, manteniendo la siguiente configuración:

- Algoritmo de entrenamiento Levenberg Marquardt
- Función de entrada lineal
- Función de salida logarítmo sigmoide

En las siguientes tablas se describen las configuraciones que se van a ensayar con esta red:

Configuración	Spread	Nº Neuronas
RB-1	0,01	10
RB-2	0,01	20
RB-3	0,01	30
RB-4	0,01	40
RB-5	0,1	10
RB-6	0,1	20
RB-7	0,1	30
RB-8	0,1	40
RB-9	1	10
RB-10	1	20
RB-11	1	30
RB-12	1	40
RB-13	10	10
RB-14	10	20
RB-15	10	30
RB-16	10	40

Fig. 3.11. Configuraciones RBF.

Técnicas de estimación

El objetivo del modelado mediante RNA, es obtener predicciones de una variable objetivo (*Target)* en función de una o varias entradas (*Inputs*).

Además de utilizar las configuraciones descritas en el apartado anterior, realizaremos 3 tipos de pruebas distintas en función del *Input* seleccionado. Posteriormente se compararán los resultados para cada caso.

Las técnicas de estimación son las siguientes:

- La entrada será uno de los RCDs considerados y se estimará otro de dichos RCDs.

- La entrada serán las tres variables RCD con mayor relevancia y se estimará cada uno de los RCDs.

- La entrada serán las Componentes Principales que permitan alcanzar el 70% de la información disponible, en este caso será suficiente con las tres

primeras Componentes Principales. Con estas tres estimaremos cada una de las Componentes Principales restantes.

Elección del mejor modelo

Para decidir que configuración de Red ha sido la más fiable, se calcula el error producido en el entrenamiento y en la validación de cada una de ellas. Esto lo calcularemos con la fórmula del error cuadrático medio (*MSE* en inglés).

Esta medida del error calcula el promedio de la diferencia al cuadrado producida entre el valor estimado y el valor esperado:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2$$

(Ec.3.16)

A nivel didáctico se utilizará también la medida de error mediante la fórmula MAE (Mean Absolute Error). La diferencia con el MSE reside en que este método es menos penalizante con los casos en los que existe una gran diferencia entre el valor estimado y el valor real, es decir, este método suaviza el error adicional introducido por los valores atípicos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}_i - Y_i|$$

(Ec. 3.17)

Cross Validation

Ya que el tamaño de los muestreos es reducido (tan solo disponemos de 64 obras), para evitar que el proceso de entrenamiento-simulación de la red neuronal resulte escaso, realizaremos una técnica conocida como *Cross-Validation*.

Este método consiste en dividir el conjunto de datos en dos grupos, Grupo de Entrenamiento y Grupo de Validación. Del total de la muestra seleccionamos aleatoriamente dos tercios para el grupo de entrenamiento y reservamos el tercio restante para componer el grupo de test.

- Grupo de Entrenamiento (*Train*): 1/3 del total de los datos.
- Grupo de Validación (*Test*): 2/3 del total de los datos.

Al finalizar el entrenamiento y la validación, obtendremos un valor de error de entrenamiento (ErrorTrain1) y otro valor de error de test (ErrorTest1), archivamos estos valores y volvemos a repetir el proceso hasta 3 veces, utilizando otro grupo de entrenamiento y de validación, de nuevo aleatorios, pero distintos a los utilizados anteriormente.

Gracias a esto estamos aumentando al triple el tamaño de nuestra muestra, si realizamos 3 validaciones cruzadas será como entrenar una red con 135 (45+45+45) individuos y realizaremos el proceso de validación con 57 (19+19+19) individuos.

Finalmente se calcula la media de los errores obtenidos en la validación y el entrenamiento:

$$ErrorTrain = \frac{ErrorTrain1 + ErrorTrain2 + ErrorTrain3}{3}$$

(Ec. 3.18)

$$ErrorTest = \frac{ErrorTest1 + ErrorTest2 + ErrorTest3}{3}$$

(Ec.3.19)

En el capítulo siguiente podemos encontrar el listado con las configuraciones de red y el preprocesamiento utilizado para estimar cada una de las variables RCD con el menor MSE.

4. RESULTADOS Y ANÁLISIS

En este capítulo se exponen los resultados obtenidos con la metodología explicada en el capítulo anterior. Podremos ver en primer lugar los resultados del análisis de correlación y de componentes principales y continuaremos con los mejores modelos de estimación generados mediante Redes Neuronales al utilizar entradas individuales o de múltiples *Inputs*.

El capítulo 5 se compone de las conclusiones obtenidas a la vista de estos resultados.

4.1. Correlación

Del análisis de correlación, podemos identificar aquellos pares de variables que mantienen relación lineal entre sí.

Para nuestro estudio consideraremos que dos residuos están altamente relacionados cuando el coeficiente de Pearson entre estos dos sea mayor a 0,75 y la prueba del p-valor sea inferior a 0,05. En las tablas 4.1 y 4.2 se listan los residuos que han mostrado alta correlación con alguna otra.

	Conjunto muestral			
	ZscMedConOut	ZscRegConOut	ZscMedSinOut	ZscRegSinOut
Variables	RCDtotal	RCDtotal	EST_Hormigón	EST_Hormigón
	EST_Hormigón	EST_Hormigón	CUB_NoTransitable	REV_Aplacado
	CUB_PlanaTransitable	CUB_PlanaTransitable	Rev_Aplacado	CEXT_Aluminio
	REV_Aplacado	REV_Aplacado	CEXT_Aluminio	CEXT_Vidrio
	PINT_Plástica	PINT_Plástica		
	CINT_Madera	CINT_Madera		
	CEXT_Aluminio	CEXT_Aluminio		
	CEXT_Vidrio	CEXT_Vidrio		

Tabla.4.1. Variables con relación lineal alta.

	Conjunto muestral			
	MinMedConOut	MinRegConOut	MinMedSinOut	MinRegSinOut
Variables	RCDtotal	RCDtotal	EST_Hormigón	EST_Hormigón
	EST_Hormigón	EST_Hormigón	CUB_NoTransitable	REV_Aplacado
	CUB_PlanaTransitable	CUB_PlanaTransitable	REV_Aplacado	CEXT_Aluminio
	REV_Aplacado	REV_Aplacado	CEXT_Aluminio	CEXT_Vidrio
	PINT_Plástica	PINT_Plástica		
	CINT_Madera	CINT_Madera		
	CEXT_Aluminio	CEXT_Aluminio		
	CEXT_Vidrio	CEXT_Vidrio		

Tabla.4.2. Variables con relación lineal alta.

Observamos que dependiendo del preprocesamiento realizado, las correlaciones en residuos cambian, esto se hace notable en los muestreos en los que se han eliminado *Outliers*.

En los casos de estudio sin *Outliers* (ZscMedSinOut, ZscRegSinOut, MinMedSinOut y MinRegSinOut), las variables muestran menores correlaciones. Esto es interesante, ya que, como se adelantaba en el capítulo anterior, los residuos con altas correlaciones son más fáciles de estimar.

4.2. Análisis de Componentes Principales

Del Análisis de Componentes Principales se han obtenido los siguientes resultados:

- Selección del número de Componentes Principales suficiente para representar el total del muestreo en cada caso de estudio.
- Selección de los residuos más significativos de cada caso de estudio.

Número de Componentes Principales

Aplicando el PCA nos fijaremos en primer lugar en la variabilidad recogida por cada componente principal. A continuación, se muestran las tablas 4.3 y 4.4, en las que podremos comparar para cada conjunto muestral, el número de componentes principales necesarias para albergar el 70% de la varianza total.

A efectos prácticos, seleccionaremos tres Componentes Principales en todos los casos de estudio, de forma que todos los casos cumplen la condición de recoger al menos el 70% de la varianza de la muestra y para la generación de modelos mediante RNA dispondremos de tres componentes principales para estimar cada uno de los residuos.

Diferencias en los casos de estudio

Observamos que en el caso de estudio ZscMedConOut y ZscRegConOut, se alcanza el 70% de la varianza con tan solo dos Componentes Principales, mientras que para el resto de casos de estudio son necesarios tres.

Componente Principal	Casos de estudio			
	ZscMedConOut	ZscRegConOut	ZscMedSinOut	ZscRegSinOut
1	60,426	60,358	45,185	43,445
2	10,603	10,631	18,089	15,802
3	9,905	9,911	9,986	11,19
4	6,674	6,671	7,32	7,918
5	5,104	5,179	6,175	5,969
6	3,018	3,043	3,631	5,487
7	1,53	1,549	3,229	3,087
8	1,025	0,988	1,799	2,436
9	0,598	0,614	1,112	1,546
10	0,36	0,323	0,967	0,938
11	0,276	0,275	0,84	0,683
12	0,184	0,188	0,695	0,582
13	0,154	0,118	0,568	0,522
14	0,096	0,103	0,279	0,247
15	0,04	0,04	0,116	0,141

Tabla. 4.3. Elección de Componentes Principales (I)

Componente principal	Casos de estudio			
	MinMedConOut	MinRegConOut	MinMedSinOut	MinRegSinOut
1	53,513	53,446	45,953	45,422
2	15,414	15,447	17,057	15,082
3	11,723	11,641	9,932	10,65
4	6,416	6,579	7,528	7,255
5	5,634	5,672	5,599	6,138
6	3,225	3,234	3,97	4,401
7	1,522	1,547	3,383	3,476
8	1,026	0,927	2,341	2,824
9	0,615	0,632	1,221	1,936
10	0,295	0,264	0,918	0,987
11	0,226	0,238	0,748	0,632
12	0,147	0,156	0,555	0,481
13	0,137	0,093	0,497	0,407
14	0,074	0,088	0,204	0,176
15	0,028	0,028	0,085	0,127

Tabla. 4.4. Elección de Componentes Principales (II)

Los residuos más significativos

Como se ha explicado en capítulos anteriores, el PCA nos permite conocer entre otras cosas, qué variables son las más significativas.

Estas variables más representativas son dignas de observación, si una variable ha obtenido un peso alto en este análisis indica que son variables con poca información redundante, esto puede significar dos cosas:

- Se trata de variables con capacidad para representar y estimar una gran cantidad de datos de la muestra
- Se trata de variables que debido a su alto grado de independencia no pueden estimar ni ser estimadas por el resto de variables de la muestra

En las siguientes tablas (tabla 4.5 y tabla 4.6) encontramos el peso que ha obtenido cada residuo en cada caso de estudio. Será relevante conocer los residuos más representativos, pues, como se ha explicado en el capítulo anterior, serán utilizados como *Input* para estimar al resto de RCDs.

Residuos	Casos de estudio			
	ZscMedConOut	ZscRegConOut	ZscMedSinOut	ZscRegSinOut
RCDTotal	7,16	7,91	7,21	7,32
CIM_ZapatasYMuros	7,38	8,94	5,75	6,95
CIM_HormLimpieza	2,58	3,12	1,91	6,77
EST_Hormigon	6,93	7,65	7,82	6,72
CUB_PlanaTransit	4,35	6,36	6,01	6,84
CUB_NoTransitable	9,08	11,58	7,12	7,62
PART_Ladrillo	9,5	4,63	6,73	5,69
REV_Aplacado	6,66	5,37	8,28	7,64
REV_Enlucido	6,98	8,88	9,43	8,88
REV_Enfoscado	7,33	6,68	7,82	6,16
PINT_Plastica	4,51	5,2	4,87	4,54
SOL_Ceramico	8,07	4,13	8,57	7,54
CINT_Madera	6,64	6,43	5,96	4,87
CEXT_Aluminio	6,02	7,81	6,35	5,64
CEXT_Vidrio	6,74	5,24	6,1	6,77

Tabla. 4.5. Grado de independencia de las variables. Casos de estudio Zscore.

Residuos	Casos de estudio			
	MinMedConOut	MinRegConOut	MinMedSinOut	MinRegSinOut
RCDTotal	6,17	5,93	7,15	4,26
CIM_ZapatasYMuros	12,4	12,67	12,31	9,65
CIM_HormLimpieza	2,03	1,95	2,07	3,44
EST_Hormigon	6,72	6,68	5,25	6,8
CUB_PlanaTransit	4,04	3,95	7,37	7,82
CUB_NoTransitable	5,76	5,86	5,72	5,08
PART_Ladrillo	13,47	13,56	6,79	8,13
REV_Aplacado	5,99	6,04	7,25	8,19
REV_Enlucido	6,33	6,31	8,43	7,01
REV_Enfoscado	7,07	7,09	6,41	8,26
PINT_Plastica	4,92	4,91	6,05	7,15
SOL_Ceramico	7,67	7,73	6,81	4,39
CINT_Madera	5,23	5,12	6,15	7,18
CEXT_Aluminio	6	5,96	5,64	6,19
CEXT_Vidrio	6,12	6,18	6,54	6,38

Tabla. 4.6. Grado de independencia de las variables. Casos de estudio Min-Máx.

En las tablas siguientes (tabla 4.7 y 4.8) se encuentran listados los tres residuos más significativos de cada caso de estudio, los cuales serán los *Inputs* que utilizaremos en la generación de modelos mediante RNA con entradas múltiples para estimar cada uno de los residuos restantes.

Casos de estudio	Residuos		
ZscMedConOut	CIM_ZapatasYMuros	CUB_NoTransitable	SOL_Ceramico
ZscRegConOut	CIM_ZapatasYMuros	CUB_NoTransitable	REV_Enlucido
ZscMedSinOut	REV_Aplacado	REV_Enlucido	SOL_Ceramico
ZscRegSinOut	CUB_NoTransitable	REV_Aplacado	REV_Enlucido

Tabla.4.7 Selección de variables más significativas de los casos de estudio Zscore.

Casos de estudio	Residuos		
MinMedConOut	CIM_ZapatasYMuros	PART_Ladrillo	SOL_Ceramico
MinRegConOut	CIM_ZapatasYMuros	PART_Ladrillo	SOL_Ceramico
MinMedSinOut	CIM_ZapatasYMuros	REV_Enlucido	SOL_Ceramico
MinRegSinOut	CIM_ZapatasYMuros	REV_Aplacado	REV_Enfoscado

Tabla. 4.8. Selección de residuos más significativos los casos de estudio Min-Máx.

Diferencias en los casos de estudio

De nuevo encontramos diferencias entre los casos de estudio, se observa que en los casos de estudio normalizados mediante el método Min-Máx (tabla 4.6) se registran valores pico con puntuaciones de entre 12 y 13 grados de representatividad en los residuos CIM_ZapatasYMuros y PART_Ladrillo, mientras que en los casos de estudio normalizados mediante el método Zscore, las mayores puntuaciones no son superiores a 9 grados.

4.3. Mejores estimaciones

En este apartado, se muestran los mejores resultados obtenidos al estimar cada una de las quince variables del estudio, dividiremos este apartado en tres grupos en función del *Input* utilizado, como se describe en el capítulo anterior:

- La entrada es uno de los RCDs considerados y se estimará otro de dichos RCDs.
- La entrada es los tres residuos más representativos para estimar cada uno de los RCDs.
- La entrada es las Componentes Principales que permitan alcanzar el 70% de la información disponible, en este caso será suficiente con las tres primeras Componentes Principales. Con estas tres estimaremos cada una de las Componentes Principales restantes.

Estimación de un residuo con un solo *Input*

La estimación de un residuo mediante la entrada de otro se consigue realizar de forma fiable para aquellas variables que guardan una alta relación lineal.

A continuación, en la tabla 4.9, se muestran las variables *Target* que mantienen una alta relación lineal con la variable *Input* utilizada para su estimación.

Target	Input óptimo	Caso de estudio
RCDtotal	CEXT_Aluminio	MinMedSinOut
EST_Hormigon	CEXT_Vidrio	MinMedConOut
CUB_PlanaTransit	CEXT_Vidrio	MinMedConOut
CUB_NoTransitable	CEXT_Aluminio	MinRegSinOut
REV_Aplacado	CEXT_Vidrio	MinMedConOut
REV_Enlucido	CEXT_Vidrio	MinMedConOut
PINT_Plastica	CEXT_Vidrio	MinMedConOut
CINT_Madera	REV_Aplacado	MinRegConOut
CEXT_Aluminio	CEXT_Vidrio	MinMedConOut
CEXT_Vidrio	CINT_Madera	MinRegConOut

Tabla.4.9. Variables Target relacionadas linealmente con el Input utilizado

Estas variables *Target* se pueden estimar de manera fiable si conocemos la variable *Input* con la que mayor relación lineal guarde. En este trabajo se ha mostrado la manera de obtener la estimación de estas variables con el menor error, pero no es la única. En los anexos se pueden consultar los resultados de utilizar cada una de las variables del conjunto muestral como *Input.*

Los mejores resultados obtenidos con estas pruebas se han producido con la arquitectura Backpropagation con la siguiente configuración:

- **Arquitectura:** Backpropagation.
- **Neuronas en capa oculta:** No se puede determinar un valor fijo, se obtienen buenos resultados con un rango de entre 10 y 40 neuronas.
- **Función de entrada:** No se puede determinar una función óptima, depende del *Target* que se quiera estimar.
- **Función de salida:** No se puede determinar una función óptima, depende del *Target* que se quiera estimar.
- **Algoritmo de entrenamiento:** Los mejores resultados se han obtenido mediante el algoritmo Levenberg Marquardt.

En cuanto a los casos de estudio mediante los que se ha obtenido una mejor estimación, se puede determinar que han sido los casos normalizados mediante el método Mín-Máx:

- MinMedConOut.
- MinRegConOut.
- MinMedSinOut.
- MinRegSinOut.

En la tabla 4.10 se detalla el *Input* utilizado, si la variable *Target* ha mostrado relación lineal alta con el *Input* utilizado, los errores producidos en la estimación para obtener el mínimo error de validación según el criterio MSE (MSE Test), la configuración de red neuronal y el preprocesamiento de datos realizado para estimar cada variable *Target.*

Target	Input	¿Correladas linealmente?	MSE Test	MSE train	MAE test	MAE train	Configuración	Preprocesamiento
RCDtotal	CEXT_Aluminio	SI	0,00348	0,00129	0,00247	0,0009	17-N10	MinMedSinOut
CIM_ZapatasYMuros	CEXT_Aluminio	NO	0,00466	0,02597	0,0033	0,01814	4-N40	MinRegConOut
CIM_HormLimpieza	EST_Hormigon	NO	0,00364	0,02357	0,05099	0,07449	5-N20	MinRegSinOut
EST_Hormigon	CEXT_Vidrio	SI	0,00023	0,00004	0,00015	0,00003	17-N20	MinMedConOut
CUB_PlanaTransit	CEXT_Vidrio	SI	0,00017	0,00008	0,00011	0,00005	17-N10	MinMedConOut
CUB_NoTransitable	CEXT_Aluminio	SI	0,00443	0,02495	0,04229	0,06828	3-N10	MinRegSinOut
PART_Ladrillo	SOL_Ceramico	NO	0,00204	0,00075	0,03392	0,01795	12-N10	MinRegSinOut
REV_Aplacado	CEXT_Vidrio	SI	0,00016	0,00004	0,00011	0,00003	17-N20	MinMedConOut
REV_Enlucido	CEXT_Vidrio	SI	0,00025	0,00071	0,00017	0,00049	17-N20	MinMedConOut
REV_Enfoscado	CEXT_Vidrio	NO	0,0031	0,00074	0,00211	0,00052	18-N30	MinMedConOut
PINT_Plastica	CEXT_Vidrio	SI	0,00013	0,00004	0,00009	0,00003	18-N20	MinMedConOut
SOL_Ceramico	RCDTotal	NO	0,00184	0,00356	0,03036	0,0361	16-N10	MinRegSinOut
CINT_Madera	REV_Aplacado	SI	0,00083	0,00025	0,00058	0,00017	11-N40	MinRegConOut
CEXT_Aluminio	CEXT_Vidrio	SI	0,00022	0,00004	0,00015	0,00003	11-N40	MinMedConOut
CEXT_Vidrio	CINT_Madera	SI	0,0005	0,0006	0,00036	0,00042	15-N10	MinRegConOut

Tabla.4.10. Error mínimo obtenido al estimar cada una de las variables. Algoritmo Backpropagation.

Estimación de un residuo mediante las tres variables más representativas de cada caso de estudio

En este proceso hemos utilizado los 3 residuos más influyentes de los casos de estudio como *Inputs* múltiples para estimar al resto de estos (Tabla 4.11 y Tabla 4.12).

Casos de estudio	Residuos Input		
ZscMedConOut	CIM_ZapatasYMuros	CUB_NoTransitable	SOL_Ceramico
ZscRegConOut	CIM_ZapatasYMuros	CUB_NoTransitable	REV_Enlucido
ZscMedSinOut	REV_Aplacado	REV_Enlucido	SOL_Ceramico
ZscRegSinOut	CUB_NoTransitable	REV_Aplacado	REV_Enlucido

Tabla.4.11 Selección de variables más significativas de los casos de estudio Zscore.

Casos de estudio	Residuos Input		
MinMedConOut	CIM_ZapatasYMuros	PART_Ladrillo	SOL_Ceramico
MinRegConOut	CIM_ZapatasYMuros	PART_Ladrillo	SOL_Ceramico
MinMedSinOut	CIM_ZapatasYMuros	REV_Enlucido	SOL_Ceramico
MinRegSinOut	CIM_ZapatasYMuros	REV_Aplacado	REV_Enfoscado

Tabla. 4.12. Selección de residuos más significativos los casos de estudio Min-Máx.

Los casos de estudio y configuraciones de red óptimos para estimar cada residuo se muestran en la tabla siguiente (Tabla 4.13).

Los mejores resultados obtenidos con estas pruebas se han producido con la arquitectura Backpropagation con la siguiente configuración:

- **Arquitectura:** Backpropagation.
- **Neuronas en capa oculta:** No se puede determinar un valor fijo, se obtienen buenos resultados con un rango de entre 10 y 40 neuronas.
- **Función de entrada:** Lineal.
- **Función de salida:** Logaritmo Sigmoide.
- **Algoritmo de entrenamiento:** No se puede determinar un algoritmo de entrenamiento óptimo, se obtienen buenos resultados tanto con el método Levenberg Marquardt como con el método del Gradiente Conjugado.

En cuanto a los casos de estudio mediante los que se ha obtenido una mejor estimación, se puede determinar que han sido los casos normalizados mediante el método Mín-Máx.

- MinMedConOut.
- MinRegConOut.
- MinMedSinOut.
- MinRegSinOut.

Target	ECM Test	ECM train	MAE test	MAE train	Configuración	Pretratamiento
RCDtotal	0,00964	0,01128	0,06267	0,06849	2-N40	MinRegConOut
CIM_ZapatasYMuros	0,77798	0,56687	0,60471	0,549	11-N20	ZscMedSinOut
CIM_HormLimpieza	0,00413	0,02604	0,0532	0,07201	6-N30	MinRegConOut
EST_Hormigon	0,00336	0,00214	0,03913	0,03477	6-N30	MinRegSinOut
CUB_PlanaTransit	0,00711	0,00212	0,05019	0,03218	6-N40	MinMedConOut
CUB_NoTransitable	0,00389	0,00665	0,04228	0,05009	6-N10	MinRegSinOut
PART_Ladrillo	0,03842	0,01722	0,13604	0,10342	1-N10	MinMedSinOut
REV_Aplacado	0,02512	0,01507	0,08391	0,08201	3-N10	MinRegSinOut
REV_Enlucido	0,38013	0,88952	0,48422	0,60562	18-N10	ZscRegSinOut
REV_Enfoscado	0,01116	0,00165	0,04865	0,02515	5-N10	MinMedConOut
PINT_Plastica	0,01009	0,00714	0,0663	0,05275	3-N20	MinRegConOut
SOL_Ceramico	0,00239	0,00133	0,03072	0,02258	6-N20	MinRegConOut
CINT_Madera	0,00958	0,00605	0,06151	0,05087	4-N10	MinRegConOut
CEXT_Aluminio	0,00497	0,00128	0,04876	0,00135	18-N10	MinRegSinOut
CEXT_Vidrio	0,00382	0,00214	0,04076	0,00155	12-N10	MinRegSinOut

Tabla. 4.13. Configuración óptima para estimar residuos utilizando los tres más significativos de cada muestreo.

Estimación mediante las Componentes Principales más representativas.

En este proceso utilizamos los datos transformados con los coeficientes del PCA, al usar estos nos hemos desprendido del significado real asociado a cada variable, por lo tanto, con este método de estimación no estamos calculando "residuos" directamente.

A diferencia del proceso anterior, en el que utilizamos los 3 residuos más influyentes en, en este sistema utilizamos los tres primeros componentes, los cuales, como hemos visto en apartados anteriores son suficientes en todos los conjuntos muestrales para albergar el 70% de la varianza total de los datos.

Como podemos observar, se obtienen con este proceso errores entre el valor estimado y el real, mucho menores que al utilizar los 3 residuos más influyentes.

Puesto que el PCA ordena los componentes de mayor a menor varianza recogida, es lógico pensar que los últimos componentes principales se pueden estimar de forma más precisa que los primeros componentes, ya que en los últimos componentes se acumulan los datos más redundantes del muestreo.

Los mejores resultados obtenidos con estas pruebas se han producido con la arquitectura Backpropagation con la siguiente configuración:

- **Arquitectura:** Backpropagation.
- **Neuronas en capa oculta:** No se puede determinar un valor fijo, se obtienen buenos resultados con un rango de entre 10 y 40 neuronas.
- **Función de entrada:** No se puede determinar una función óptima, depende del *Target* que se quiera estimar.
- **Función de salida:** No se puede determinar una función óptima, depende del *Target* que se quiera estimar.
- **Algoritmo de entrenamiento:** Los mejores resultados se han obtenido mediante el algoritmo Levenberg Marquardt.

En cuanto a los casos de estudio mediante los que se ha obtenido una mejor estimación, se puede determinar que han sido los casos normalizados mediante el método Mín-Máx, y los métodos de imputación por regresión:

- MinRegConOut.
- MinRegSinOut.

Los casos de estudio y configuraciones de red óptimos para estimar cada variable se muestran en la tabla siguiente (Tabla 4.10).

Componente Principal	ECM Test	ECM train	MAE test	MAE train	Configuración	Pretratamiento
4	0,00325	0,01496	0,04697	0,07119	4-N40	MinRegSinOut
5	0,00336	0,00214	0,03913	0,03477	6-N30	MinRegSinOut
6	0,00382	0,00005	0,04076	0,00155	12-N10	MinRegSinOut
7	0,00389	0,00665	0,04228	0,05009	6-N10	MinRegSinOut
8	0,00285	0,00255	0,03049	0,02512	18-N30	MinMedConOut
9	0,00252	0,00129	0,0354	0,02414	5-N40	MinRegConOut
10	0,00093	0,00066	0,02538	0,01883	12-N20	MinRegConOut
11	0,00075	0,00073	0,0197	0,02001	17-N40	MinMedConOut
12	0,00065	0,00044	0,0189	0,01679	12-N40	MinRegConOut
13	0,00046	0,00033	0,01635	0,01347	18-N40	MinMedConOut
14	0,00021	0,00025	0,01125	0,00965	17-N10	MinMedConOut
15	4,00E-05	0,0001	0,00516	0,00731	5-N40	MinRegConOut

(Tabla. 4.10). Configuración óptima para estimar las componentes principales mediante las más representativas.

Representación de modelos

Puede ocurrir que, al estimar una variable, el error obtenido entre el valor estimado y el valor real sea bajo, y aun así el modelo de estimación no se ajuste al modelo real de forma satisfactoria.

Para comprobar el existo de cada modelo, se procede a representar los valores reales de cada variable *Target* frente a los mejores (menor MSE) valores estimados.

Modelo real de la variable RCDtotal (rojo) frente al modelo estimado con la variable CEXT_Aluminio (verde)

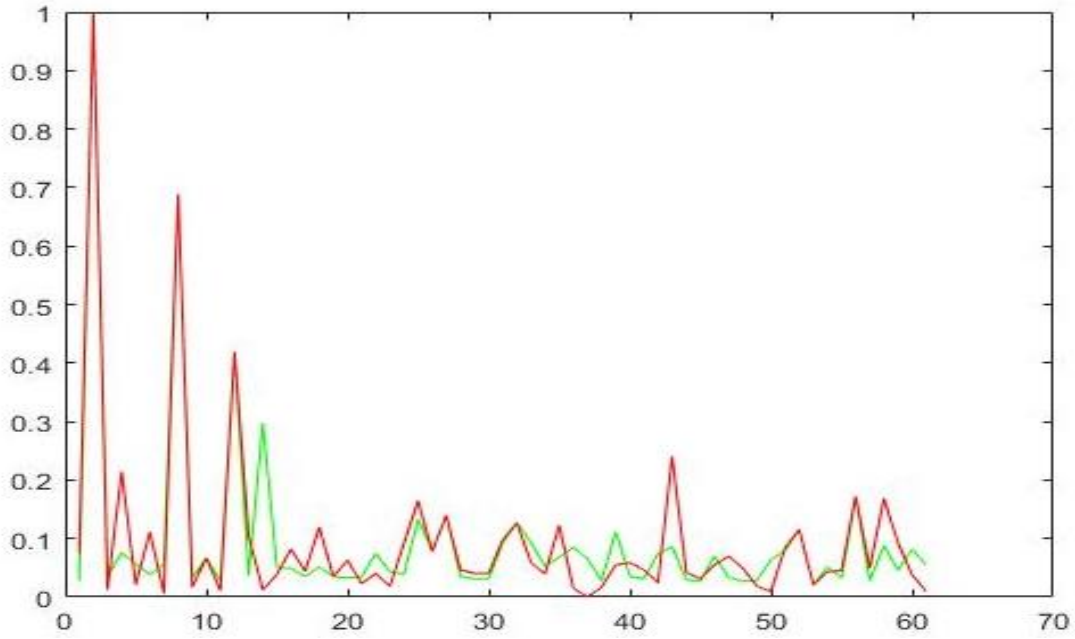
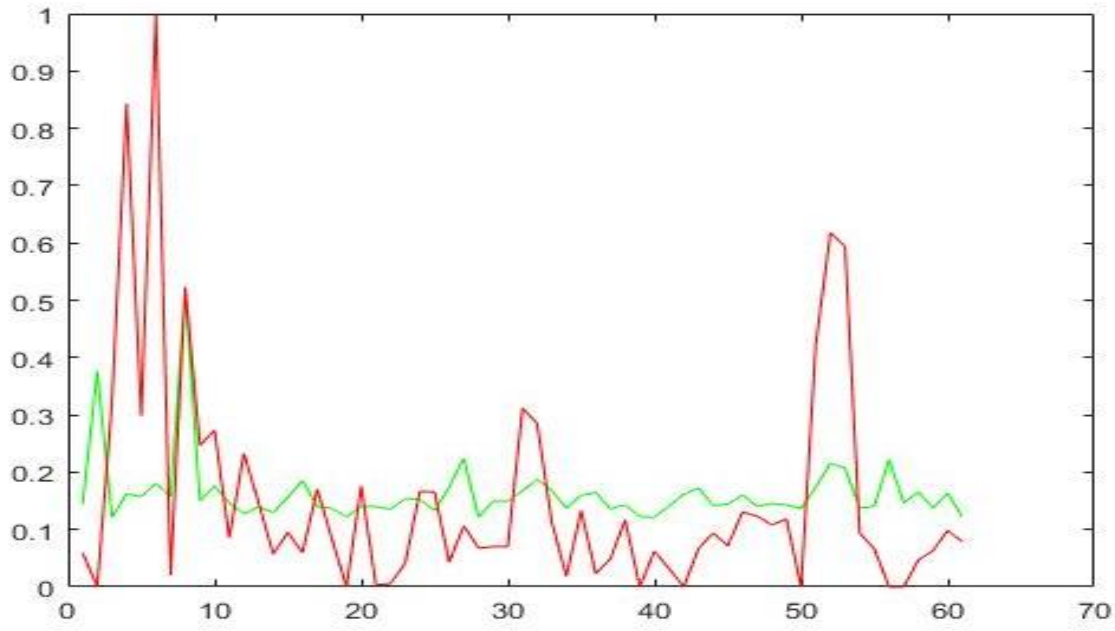


Fig.4.1. RCDtotal(R)-CEXT_Aluminio(V)

El modelo de estimación del residuo RCDtotal mediante el residuo CEXT_Aluminio, se ajusta de forma satisfactoria a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Media.
- Grado de importancia por PCA de la variable *Target:* Media.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable CIM_Zapatas y Muros (rojo) frente al modelo estimado con la variable CEXT_Aluminio (verde)

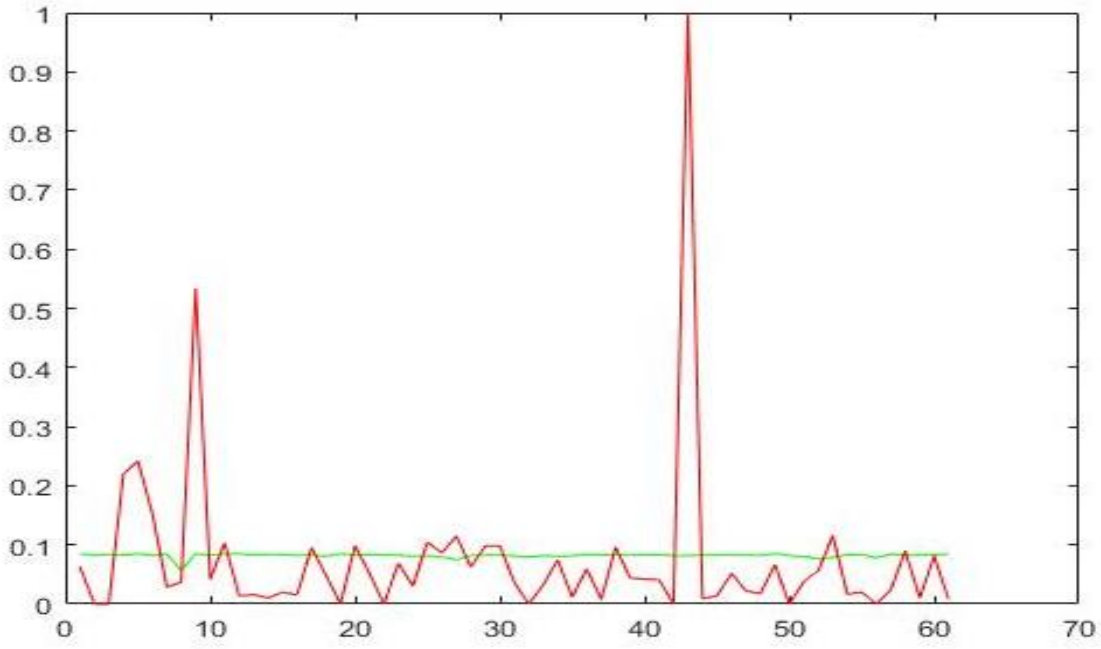


(Fig.4.2). CIM_Zapatas y Muros(R)-CEXT_Aluminio(V)

El modelo de estimación del residuo CIM_ZapatasYMuros mediante el residuo CEXT_Aluminio, NO se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Alto.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 40.
    - Función de entrada: Lineal.
    - Función de salida: Tangente Sigmoide.
    - Algoritmo de entrenamiento: Gradiente Conjugado.

Modelo real de la variable CIM_HormigónLimpieza (rojo) frente al modelo estimado con la variable EST_Hormigón (verde)

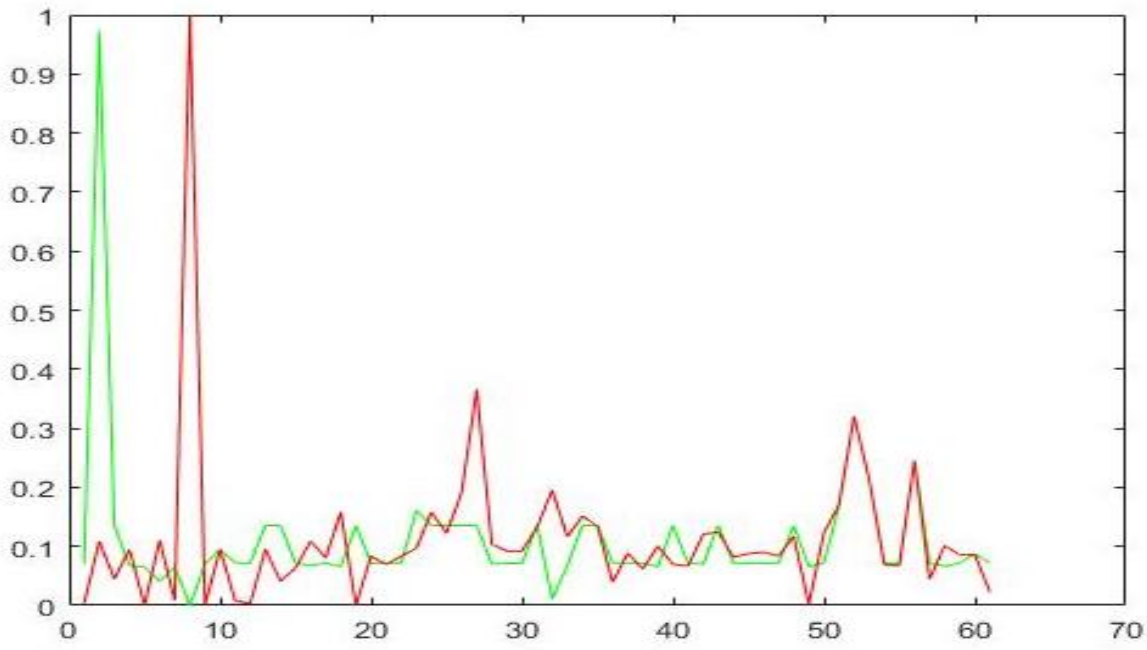


(Fig. 4.3). CIM_HormigónLimpieza(R)-EST_Hormigón(V)

El modelo de estimación del residuo CIM_HormigónLimpieza mediante el residuo EST_Hormigón, NO se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Bajo.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Lineal.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable EST_Hormigón (rojo) frente al modelo estimado con la variable CEXT_Vidrio (verde)

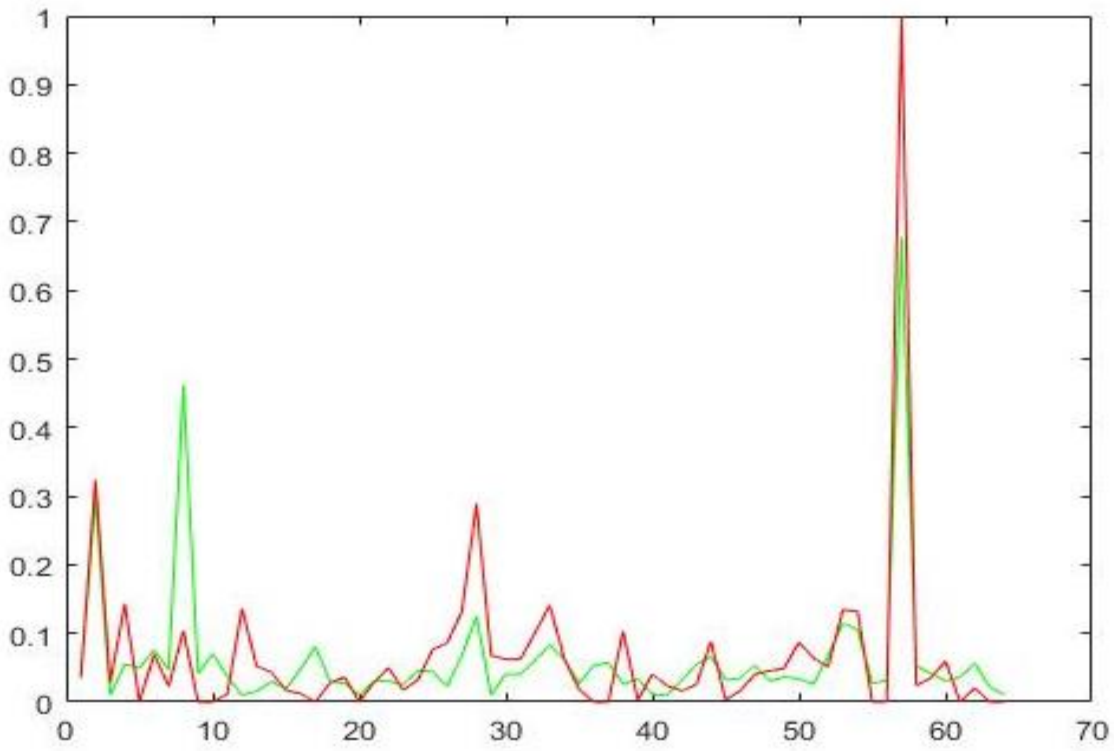


(Fig. 4.4). EST_Hormigón(R)_CEXT_Vidrio(V)

El modelo de estimación del residuo EST_Hormigón mediante el residuo CEXT_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* SI mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 20.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable CUB_PlanaTransitable (rojo) frente al modelo estimado con la variable CEX_Vidrio (verde)

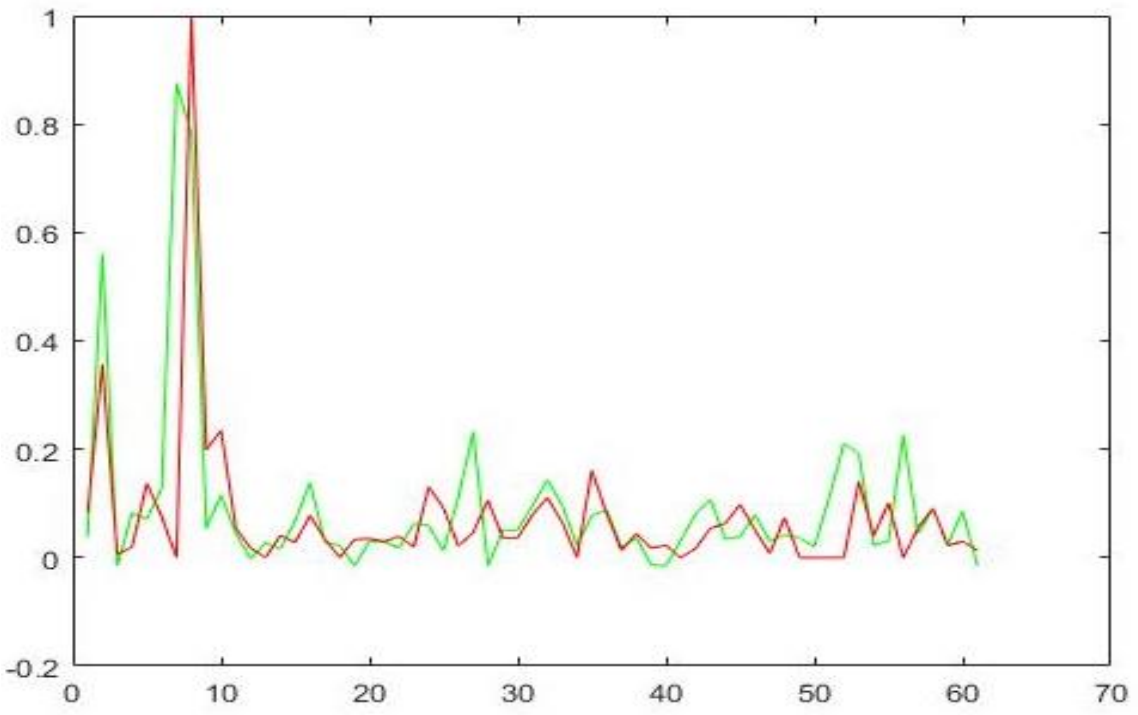


(Fig. 4.5). CUB_PlanaTransitable(R)-CEXT_Vidrio(V)

El modelo de estimación del residuo CUB_PlanaTransitable mediante el residuo CEX_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* SI mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Bajo.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable CUB_NoTransitable (rojo) frente al modelo estimado con la variable CEX_Aluminio (verde)

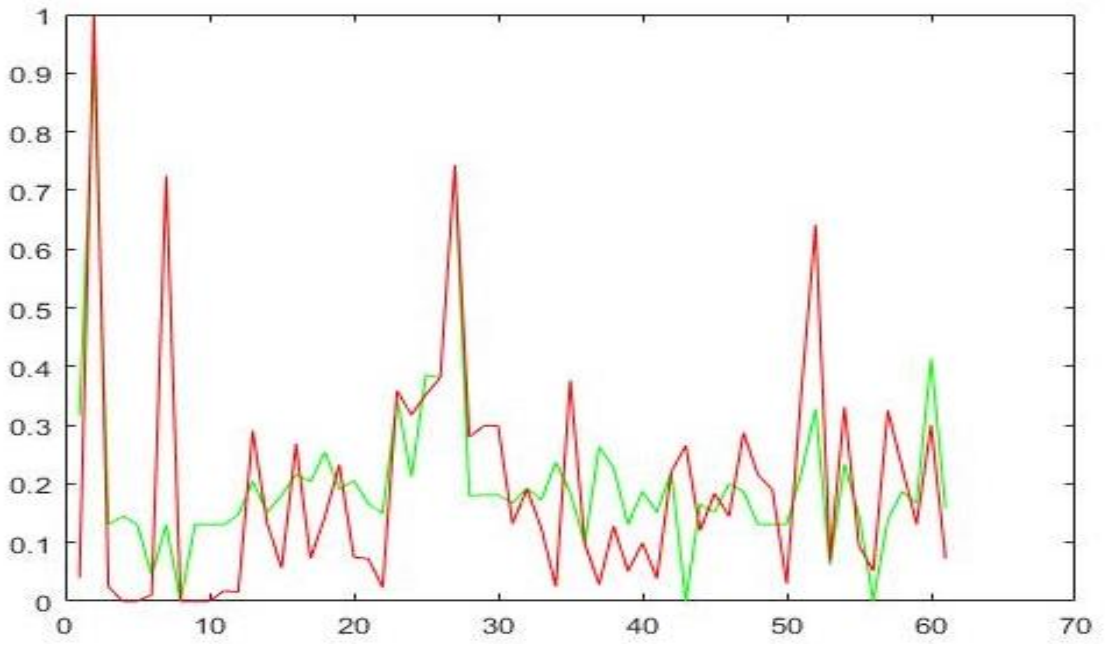


(Fig. 4.6). CUB_NoTransitable(R)-CEX_Aluminio(V)

El modelo de estimación del residuo CUB_NoTransitable mediante el residuo CEXT_Aluminio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input SI* mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Lineal.
    - Función de salida: Tangente Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable PART_Ladrillo (rojo) frente al modelo estimado con la variable SOL_Ceramico (verde)

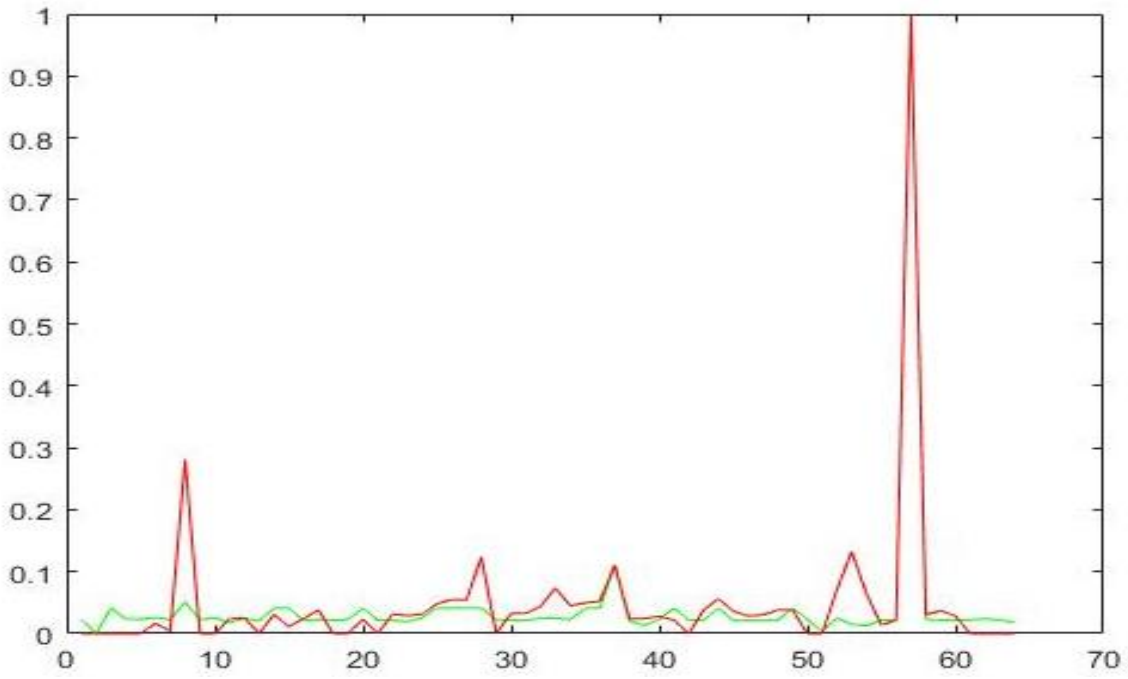


(Fig. 4.7). PART_Ladrillo(R)-SOL_Ceramico(V)

El modelo de estimación del residuo PART_Ladrillo mediante el residuo SOL_Ceramico, NO se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Bajo.
- Grado de importancia por PCA de la variable *Target:* Alto.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entradaTangente Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Gradiente Conjugado.

Modelo real de la variable REV_Aplacado (rojo) frente al modelo estimado con la variable CEXT_Vidrio (verde)

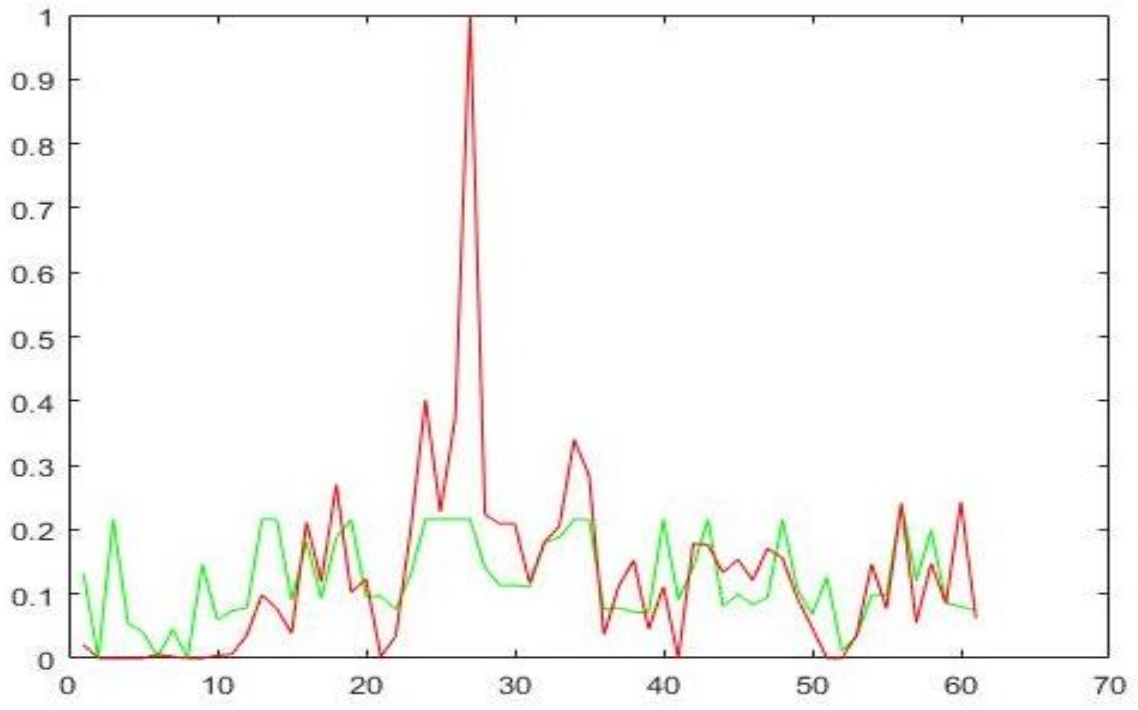


(Fig. 4.8). REV_aplacado(R)-CEXT_Vidrio(V)

El modelo de estimación del residuo REV_Aplacado mediante el residuo CEXT_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* SI mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 20.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable REV_Enlucido (rojo) frente al modelo estimado con la variable CEXT_Vidrio (verde)

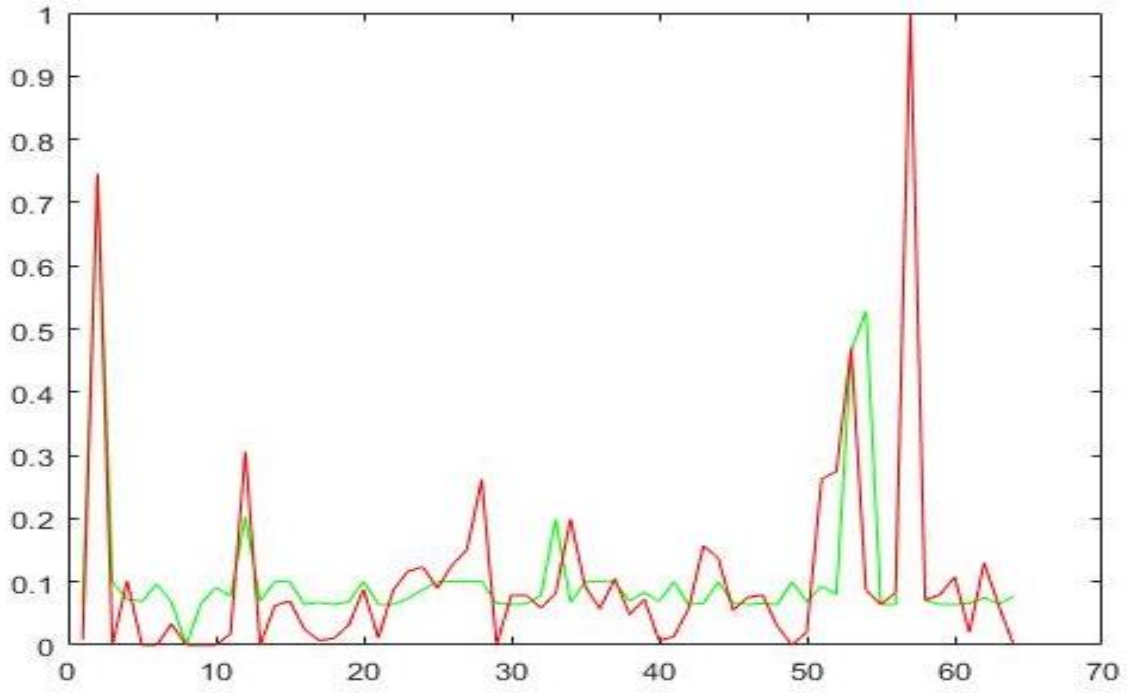


(Fig. 4.9). REV_Enlucido(R)-CEX_Vidrio(V)

El modelo de estimación del residuo REV_Enlucido mediante el residuo CEXT_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* SI mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 20.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable REV_Enfoscado (rojo) frente al modelo estimado con la variable CEX_Vidrio (verde)

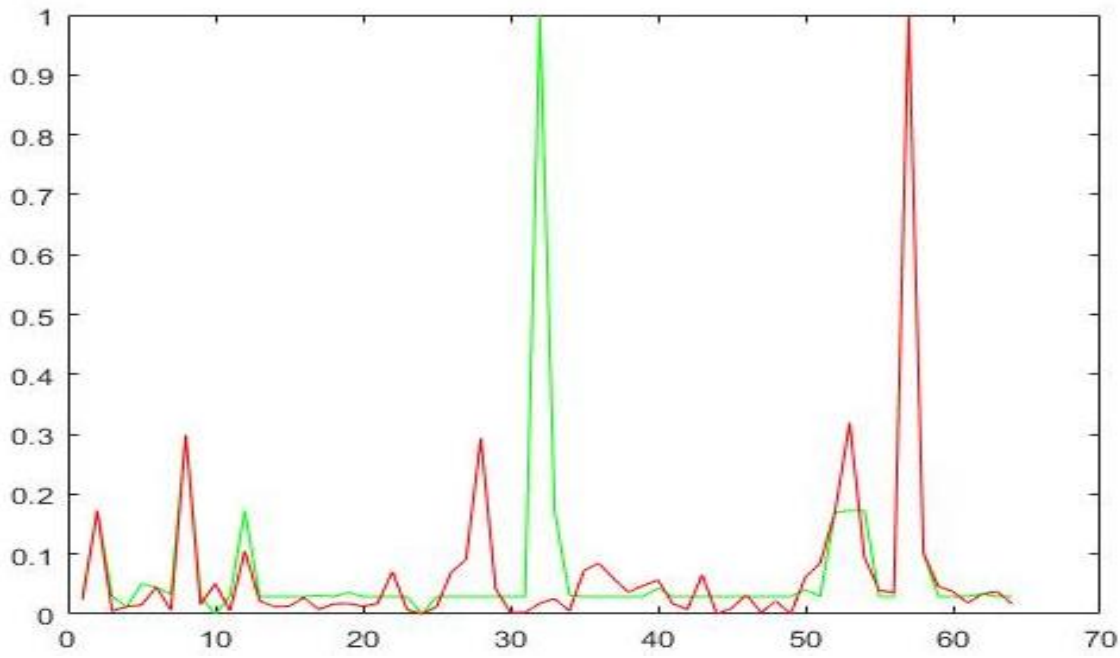


(Fig. 4.10). REV_Enfoscado(R)-CEX_Vidrio(V)

El modelo de estimación del residuo REV_Enfoscado mediante el residuo CEX_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable PINT_Plástica (rojo) frente al modelo estimado con la variable CEXT_Vidrio (verde)

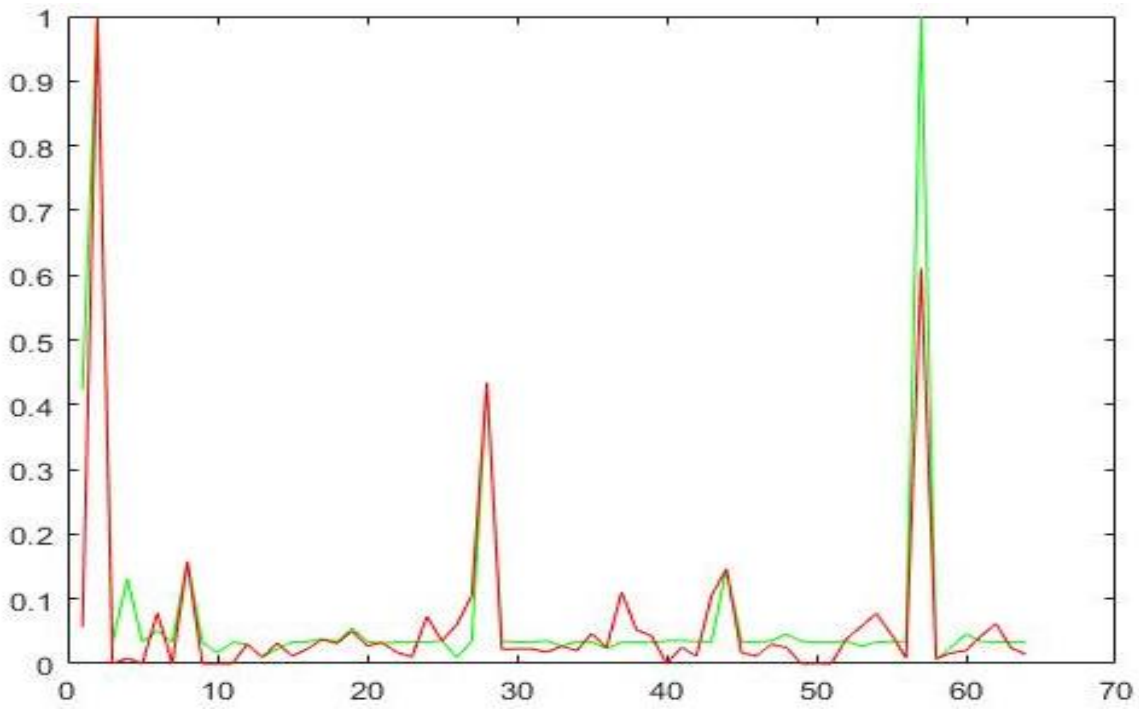


(Fig. 4.11). PINT_Plastica(R)-CEX_Vidrio(V)

El modelo de estimación del residuo PINT_Plástica mediante el residuo CEXT_Vidrio, NO se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* SI mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Bajo.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 20.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Logaritmo Sigmoide.
    - Algoritmo de entrenamiento: Gradiente Conjugado.

Modelo real de la variable SOL_Cerámico (rojo) frente al modelo estimado con la variable RCDtotal (verde)

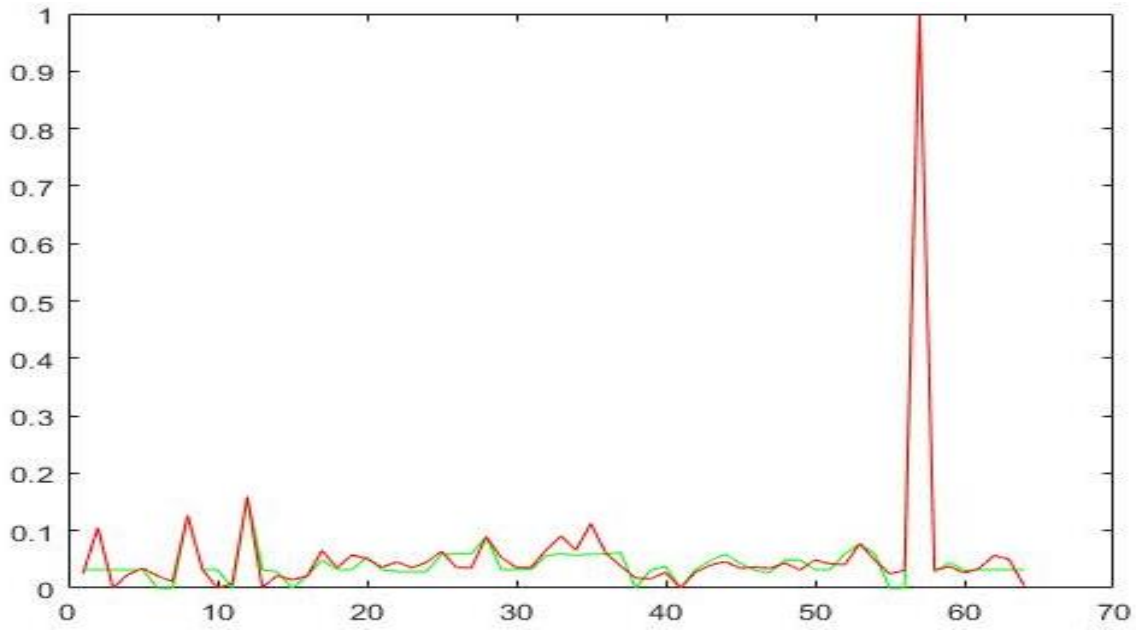


(Fig. 4.12). SOLCeramico(R)-RCDtotal(V)

El modelo de estimación del residuo SOL_Cerámico mediante el residuo RCDtotal, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Alto.
- Configuración de red óptima es la siguiente:
  - Arquitectura: Backpropagation.
  - Neuronas en capa oculta: 10.
  - Función de entrada: Logaritmo Sigmoide.
  - Función de salida: Tangente Sigmoide.
  - Algoritmo de entrenamiento: Gradiente Conjugado.

Modelo real de la variable CINT_Madera (rojo) frente al modelo estimado con la variable REV_Aplacado (verde)

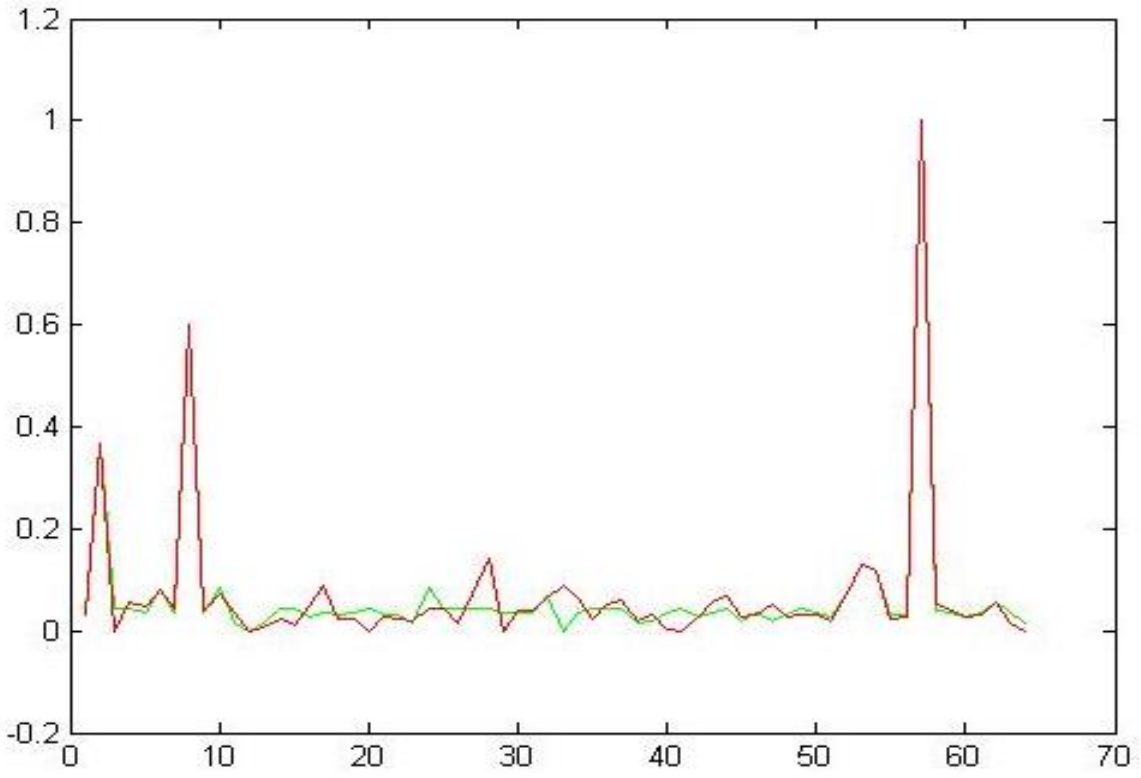


(Fig. 4.13). CINT_Madera(R)-REV_Aplacado(V)

El modelo de estimación del residuo CINT_Madera mediante el residuo REV_Aplacado, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input* NO mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Tangente Sigmoide.
    - Algoritmo de entrenamiento: Gradiente Conjugado.

Modelo real de la variable CEXT_Aluminio (rojo) frente al modelo estimado con la variable CEX_Vidrio (verde)

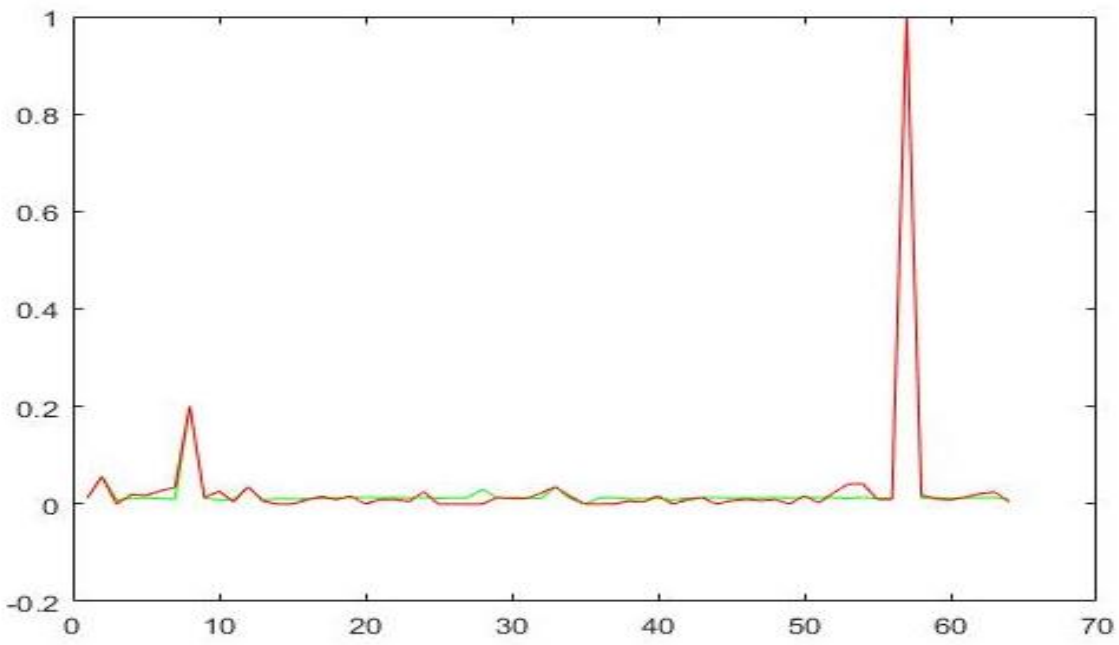


(Fig. 4.14). CEXT_Aluminio(R)_CEX_Vidrio(V)

El modelo de estimación del residuo CEXT_Aluminio mediante el residuo CEX_Vidrio, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input SI* mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
  - Arquitectura: Backpropagation.
  - Neuronas en capa oculta: 40.
  - Función de entrada: Tangente Sigmoide.
  - Función de salida: Logaritmo Sigmoide.
  - Algoritmo de entrenamiento: Levenberg Marquardt.

Modelo real de la variable CEXT_Vidrio (rojo) frente al modelo estimado con la variable CINT_Madera (verde)



(Fig. 4.15). CEXT_Vidrio(R)-CINT_Madera(V)

El modelo de estimación del residuo CEXT_Vidrio mediante el residuo CINT_Madera, SI se ajusta a la curva real. Las características de este modelo son las siguientes:

- La variable *Input SI* mantiene correlación lineal alta con la variable *Target.*
- Grado de importancia por PCA de la variable *Input:* Medio.
- Grado de importancia por PCA de la variable *Target:* Medio.
- Configuración de red óptima es la siguiente:
    - Arquitectura: Backpropagation.
    - Neuronas en capa oculta: 10.
    - Función de entrada: Logaritmo Sigmoide.
    - Función de salida: Tangente Sigmoide.
    - Algoritmo de entrenamiento: Levenberg Marquardt.

5. CONCLUSIONES

En este capítulo se comentan los resultados mostrados en el capítulo anterior.

A continuación, se exponen los resultados más significativos y las técnicas que en general han obtenido los mejores resultados.

5.1. Resumen

El problema al que nos hemos enfrentado consiste en analizar un conjunto de datos que de inicio se encuentra mal organizado e incompleto. Para ello se han llevado a cabo distintas técnicas de preprocesamiento numérico mediante las cuales hemos conseguido paliar el sesgo que producirían estos errores en los resultados.

Una vez se ha subsanado el problema de los datos faltantes y se han normalizado los datos de partida, se han generado ocho casos de estudio, que analizamos mediante PCA y análisis de correlaciones con el fin de seleccionar las variables más representativas de cada muestreo.

Mediante Redes Neuronales, apoyándonos en la selección de variables más significativas de cada muestra, formulamos distintos modelos de estimación, los cuales serán objeto de análisis con el fin de determinar cuál ha sido el procedimiento que arroja las mejores estimaciones.

Los resultados obtenidos en este trabajo muestran que, en obras de construcción civil, la generación de ciertos residuos está relacionada con otros. De esta forma, es posible obtener modelos matemáticos que permitan estimar RCDs en obras futuras.

Para realizar este trabajo se han considerado tres puntos clave:

- **Preprocesamiento:**
  Para la elaboración de este trabajo, se ha comprendido la importancia de tratar previamente los datos de partida, de tal manera que preprocesamientos distintos generan resultados distintos.

- **Selección de variables:**
  Otro punto importante en el cálculo de modelos de estimación, es la selección y extracción de variables. Conocer las relaciones existentes entre variables nos permite acotar las soluciones. El Análisis de Componentes Principales juega un papel clave en este punto, gracias a este, se pueden conocer las variables (residuos, para este trabajo en concreto) más significativas de la muestra. Así, a la hora de recopilar información para formar la base de datos, se deberá dedicar especial atención a estas variables, pues al conocer estas, se conoce una gran parte de la información de la muestra.

- **Configuración del modelo:**
  Por último, en el estudio de Redes Neuronales, existe una gran parte de trabajo de prueba y error. Debido a la gran variedad de arquitecturas y

configuraciones posibles, el analista deberá probar un gran número de ellas hasta obtener los modelos que mejor se adapten a los datos de partida.

Resumen de resultados

Con los resultados obtenidos en los procesos de selección y extracción de variables mediante análisis de correlación y PCA, podemos diferenciar dos grupos de residuos, en función de los resultados que ha obtenido cada residuo en la generación de modelos de estimación, se determina:

Residuos correlados linealmente

Como era de esperar, para aquellos residuos que muestran una alta correlación lineal con otros residuos del muestreo, es posible calcular modelos de estimación satisfactorios, los cuales se adaptan al modelo real, y mediante los cuales se obtienen medidas de error (MSE) bajas.

Los residuos pertenecientes a este grupo son los mostrados en la siguiente tabla:

Residuos Estimados	Correlado linealmente con:	Error en la estimación (MSE)
RCDtotal	CEXT_Aluminio	0,00348
EST_Hormigon	CEXT_Vidrio	0,00023
CUB_PlanaTransit	CEXT_Vidrio	0,00017
CUB_NoTransitable	CEXT_Aluminio	0,00443
REV_Aplacado	CEXT_Vidrio	0,00016
REV_Enlucido	CEXT_Vidrio	0,00025
CINT_Madera	REV_Aplacado	0,00083
CEXT_Aluminio	CEXT_Vidrio	0,00022
CEXT_Vidrio	CINT_Madera	0,0005

Tabla. 5.2. Residuos Correlados linealmente con algún otro residuo de la muestra.

Con esto se hace notable la importancia de los residuos CEXT_Aluminio y CEXT_Vidrio, ya que al conocer estos, se pueden realizar estimaciones fiables de la gran mayoría de residuos restantes.

Residuos correlados de forma no lineal

Estos residuos, a pesar de no haber mostrado correlaciones lineales significativas con el resto de la muestra, ha sido posible obtener modelos capaces de realizar estimaciones satisfactorias mediante el método de estimación de inputs múltiples.

Del mismo modo estos residuos pueden ser utilizados para estimar a otros. Los residuos pertenecientes a este grupo se listan a continuación:

Residuos relacionados de forma no lineal	Estimado con:	Error en la estimación (MSE)
REV_Enfoscado	CIM_ZapatasYMuros	0,01116
	PART_Ladrillo	
	SOL_Ceramico	
SOL_Ceramico	CIM_ZapatasYMuros	0,00239
	REV_Aplacado	
	REV_Enfoscado	

Tabla.5.3. Residuos correlados de forma no lineal, con capacidad de estimar y ser estimados.

Residuos independientes

Para estos residuos no se han obtenido modelos de estimación satisfactorios. Las estimaciones realizadas para estos arrojan resultados poco fiables, aunque el error en algún caso sea bajo, la representación del valor estimado frente al valor real, demuestra que el modelo no es fiable. Los residuos pertenecientes a este grupo se listan a continuación:

Residuos independientes
CIM_ZapatasYMuros
CIM_HormLimpieza
PART_Ladrillo
PINT_Plástica

Tabla 5.4. Residuos no correlados sin capacidad de estimar ni de ser estimados.

Influencia del Preprocesamiento

Podemos observar una gran diferencia al utilizar un pretratamiento u otro. Tras realizar los diferentes ensayos y determinar cuáles han sido los métodos que arrojan los mejores resultados, se observa lo siguiente:

	Preprocesamiento					
	Normalización		Imputación numérica		Tratamiento de Outliers	
	Zscore	Min-Max	Media	Regresión	Sin Outliers	Con Outliers
Casos óptimos	2/30	28/30	12/30	18/30	13/30	17/30

Tabla. 5.1. Número de casos óptimos obtenido con cada tipo de preprocesamiento.

Imputación numérica

Los dos procesos de tratamiento de datos faltantes (Media y Regresión) han determinado resultados similares. Al realizar estas imputaciones se ha reforzado la correlación existente entre algunos pares de variables, lo cual ha ayudado indirectamente a la estimación de variables mediante redes neuronales artificiales.

Normalización

En todos los casos, se han obtenido menores tasas de error al utilizar la normalización Min-Máx. Este tipo de normalización de variables transforma los datos originales haciendo que se adapten a un rango de valores comprendido entre cero y uno, lo cual ha ayudado notablemente a que el proceso de predicción de la RNA obtenga mejores resultados respecto a las muestras normalizadas con el proceso Zscore.

Tratamiento de *Outliers*

Se observa que el tratamiento de *Outliers*, modifica la correlación entre variables, siendo los conjuntos muestrales sin Outliers los que mayores tasas de correlación han obtenido, mientras que los conjuntos en los que se han eliminado *Outliers* parece haber debilitado la correlación existente entre algunos pares de variables.

Aunque la eliminación de estos ha reforzado las correlaciones, a la vista de los resultados de los ensayos con Redes Neuronales, no podemos inferir ninguna norma de utilización de este tratamiento, pues depende de la variable que se vaya a estimar, le favorece o perjudica la eliminación de *Outliers.*

Influencia de la configuración del modelo

Respecto a los ensayos realizados con Redes Neuronales, las diferencias mostradas al utilizar una configuración de Red determinada, son notables.

Las mejores estimaciones se han obtenido mediante los modelos generados con la Red Neuronal de tipo Perceptrón de arquitectura Backpropagation. Dentro de esta, las configuraciones que mejores resultados han arrojado son las siguientes:

Número de neuronas en la capa oculta: 10, 20, 30 o 40

Con los ensayos realizados no es posible determinar un valor óptimo de neuronas, se han obtenido buenos resultados al emplear un rango de neuronas de entre 10 y 40.

	Configuración Backpropagation			
	Neuronas en capa oculta			
	10	20	30	40
Casos óptimos	14/30	8/30	3/30	5/30

Tabla 5.5 Número de estimaciones óptimas en función de las neuronas de la capa oculta

Algoritmo de entrenamiento: Levenberg-M o Gradiente Conjugado

Respecto al algoritmo de entrenamiento, se han obtenido resultados similares utilizando ambos métodos.

	Configuración Backpropagation	
	Algoritmo de entrenamiento	
	Levenberg	Gradiente Conjugado
Casos óptimos	16/30	14/30

Tabla 5.6 Número de estimaciones óptimas en función del algoritmo de entrenamiento.

Función de activación: Logsig, Tansig, Line

De forma general podemos decir que se obtienen mejores resultados utilizando la función Logaritmo Sigmoidea, pero se debe examinar cada caso para obtener el mejor resultado.

	Función de activación Entrada		
	Lineal	Logaritmo	Tangente
Casos óptimos	9/30	13/30	8/30
	Función de activación Salida		
	Lineal	Logaritmo	Tangente
Casos óptimos	9/30	18/30	3/30

Tabla 5.7. Número de estimaciones óptimas en función de la función de activación utilizada.

5.2. Lecciones aprendidas

La realización de este trabajo fin de grado ha supuesto un crecimiento personal constante, mejorando como profesional en los siguientes ámbitos:

- Determinar y acotar el problema de partida y los objetivos del trabajo.
- Búsqueda de documentación y análisis de trabajos similares.
- Diseño de propuestas para solucionar un problema real.
- Optimización de procesos informáticos para ahorrar tiempos de ejecución.
- Inferir conclusiones a partir de resultados.

A nivel didáctico, para la realización de este trabajo ha sido necesario aprender técnicas de análisis de datos que no se han desarrollado durante el curso académico y se han afianzado otros métodos de los que se disponían nociones básicas:

- Preprocesamiento de muestreos reales:
  - Tratamiento de datos faltantes.
  - Detección de *Outliers.*
  - Normalización de muestreos.
  - Dimensionalización de variables
- Análisis estadísticos:
  - Análisis de Correlación.
  - Análisis de Componentes Principales.

- Modelado de redes neuronales
  - Nociones sobre los tipos de Redes Neuronales existentes.
  - Generación de ensayos Prueba-error para alcanzar una correcta configuración de red.

- Automatización de procesos repetitivos mediante el diseño de *Scripts* en Matlab.

6. Referencias bibliográficas

En este capítulo se encuentra el listado de la bibliografía y el material de consulta utilizado durante la elaboración de este trabajo fin de grado.

Bibliografía consultada

Allison, P. *Missing Data.* Pennsylvania, Ed. SAGE, 2001. *(*págs. 155-175*).*

Berka, P. *Discretization and grouping: Preprocessing steps for data mining.* Berlín, Ed. LNCS, 1998. (págs. 1-7).

Gomez García, Juan. «*Metodos de inferencia estadística con datos faltantes.*» Dialnet, 2006. Vol 48, núm 162. (págs 241-271)

Haykin S. *NEURAL NETWORKS: A Comprehensive Fundation.* Ontario: Ed. Pretice Hall, 1998. (págs 320-347).

Hernández, C, y Rodríguez, J.E. «Preprocesamiento de datos Estructurados.» Revista *VINCULOS*, 2008: Vol 4 núm 2. (págs. 1-15).

Herrera, F. *Data Preprocessing in Data Mining.* Polonia. Ed. Springer, 2015. (págs. 1-91).

Hoaglin D.C, Mosteller F y Tukey J.W. *Mathematical Aspects of Transformation. Undestanding Robust and Exploratory Data Analysis.* New York, Ed. John Wiley and Sons Ltd, 1987. (págs. 305-367).

James, G, y Witten, D. *An Introduction to Statistical Learning.* California. Ed. Springer, 2013. (págs. 1-441)

Jolliffe, I.T. *Principal Component Analysis.* Kent. Ed. Springer, 1986. (págs. 170-300).

Lehmann, E.L., y Casella, G. *Theory of Point Estimation.* New York. Ed. Springer, 1998. (págs. 1-193).

Nieves Hurtado, A. *Probabilidad y Estadistica para Ingenieria.* México. Ed. McGrawHill, 2014.

Peña, D. *Analisis de datos Multivariantes.* Madrid. Ed. McGrawHill, 2002. (págs. 200-529).

Refaat, M. *Data Preparation for Data Mining using SAS.* Toronto. Ed. Morgan Kaufman, 2006. (págs. 1-424).

Rodriguez, L.J. *Métodos Estadisticos para Ingenieria.* Madrid. Ed. Garceta, 2011. (págs. 317-518).

Rodríguez-Bernal, M.T. *Multiple hypothesis testing an clustering with mixtures of non-central t-distributions applied in microarray data analysis.* Ámsterdam. Ed. Elsevier, 2012. (págs. 1898-1907).

Roger J., Shing, J., Chuen-Tsai, S., y Mizutani, E. *Neuro Fuzzy and soft computing.* United Estates of America. Ed. Pearson Education, 1997. (págs. 93-126).

Rubin, D. *Inference and Missing Data.* Cambridge. Ed. Biometrika, 1987. (págs. 1-13).

Triola, M. F. *ESTADISTICA.* México. Ed. Pearson Education, 2004. (págs 635-814).

Tsoukalas, Lefteri, y Uhrig, R. *Fuzzy and Neural Approaches in Engineering.* New York. Ed. John Wiley & Sons Inc, 1996. (págs. 1-600).

Valencia Reyes, M.A., Yáñez Márquez, C., y Sánchez Fernández, L. «*Algoritmo Backpropagation para redes neuronales: Conceptos y aplicaciones.*» Instituto Politécnico Nacional, 2006: Vol 125. (págs 1-14).

Legislación consultada

Real Decreto 105/2008, de 1 de febrero, por el que se regula la producción y gestión de los residuos de construcción y demolición. (BOE núm. 38, de 13/02/2008).

Páginas web consultadas

Ballesteros, Alfonso. *Redes Neuronales.* Disponible en: http://www.redes-neuronales.com.es/tutorial-redes-neuronales/clasificacion-de-las-redes-neuronales-artificiales.htm [Consulta: 14 de Junio de 2017].

COAC. *Fichas Técnicas del Colegio Oficial de Arquitectos de Cataluña.*Disponible en: https://www.arquitectes.cat [Consulta: 14 de Junio de 2017].

Comerón Graupera, Lluís Xavier . *Programa para los proyectos y obras de construcción.* Disponible en: https://itec.es/programas/tcq [Consulta: 14 de Junio de 2017].

CYPE . *Arquimedes.* Disponible en: http://arquimedes.cype.es [Consulta: 14 de Junio de 2017].

Fueyo Editores. *Informe de Producción y Gestión de los Residuos de Construcción y Demolición (RCD) en España.*Disponible en: *http://www.rcdasociacion.es/documentacion/publicaciones* [Consulta: 15 de Junio de 2017].

Mathworks. *Guia de ayuda al usuario Matlab.* Disponible en: https://es.mathworks.com/store [Consulta: 8 de Junio de 2017].

Murphy , Thomas. *Metodos de detección de Outliers*. Disponible en: https://www.astm.org/SNEWS/SPANISH/SPND08/datapoints_spnd08.html. [Consulta: 10 de Junio de 20017].

Working Papers

Ayuso, Jesus. *Gestión y tratamiento de Residuos de Construcción y Demolición.* Córdoba: Universidad de Córdoba, 2011. Disponible en: http://www.aridosrcdandalucia.es/rcd/wp-content/uploads/2016/03/Gestion-y-Tratamiento-de-Residuos-de-Construccion-y-Demolicion-RCD-Guia-de-Buenas-Practicas.pdf [Consulta: 28 de Junio de 2017].

Berzal Galiano, Fernando. *Backpropagation.* Granada: Universidad de Granada, 2004. Disponible en: http://elvex.ugr.es/decsai/computational-intelligence/slides/N2%20Backpropagation.pdf [Consulta: 28 de Junio de 2017].

Gimenez, Blanca. *Introducción al estudio de la generación de residuos de construcción y estimación de cantidades generadas en obras.* Valencia: Universidad Politécnica de Valencia, 2010. Disponible en: http://bibing.us.es/proyectos/abreproy/30186/fichero/Cap%C3%ADtulo+16pdf.pdf [Consulta: 28 de Junio de 2017].

Martinez-Vara, Carlos. *Coeficiente de correlación lineal de Pearson.* Sevilla: Universidad de Sevilla, 2005. Disponible en: https://personal.us.es/vararey/adatos2/correlacion.pdf [Consulta: 28 de Junio de 2017].

Ramírez, Antonio. *Retirada Selectiva de Residuos.* Sevilla: Deposito de Investigación Universidad de Sevilla, 2015. Disponible en: http://bibing.us.es/proyectos/abreproy/30186/fichero/Cap%C3%ADtulo+16pdf.pdf [Consulta: 28 de Junio de 2017].

Valls, Jose María. *Redes de Neuronas. Redes de Base Radial.* Madrid: Universidad Carlos III, 2007. Disponible en: http://eva.evannai.inf.uc3m.es/et/docencia/rn-inf/documentacion/Tema4-RNBR.pdf [Consulta: 28 de Junio de 2017].

Anexos

Anexo 1: Descripción de tareas

Las tareas de aprendizaje, análisis y experimentación llevadas a cabo durante la elaboración de este trabajo son las siguientes:

A. **Estudio y análisis del problema**
   Durante esta primera etapa se ha examinado la base de datos de partida, recopilando los problemas y dificultades que se deben solventar y visualizando los objetivos que se desean conseguir.

B. **Búsqueda de bibliografía**
   Una vez se ha delimitado el problema y se han marcado los objetivos, se debe aprender a alcanzarlos, para ello nos serviremos de trabajos anteriores, mediante los cuales se aprenderán métodos nuevos y se depurarán las técnicas ya conocidas.

C. **Diseño de propuesta**
   Una vez conocida la extensión del problema y comprendidos los mecanismos de resolución existentes, comienza el trabajo creativo, apoyándonos en fundamentos teóricos universalmente aceptados para generar una metodología capaz de resolver el problema.

D. **Implementación del método**
   Esta etapa puede comenzar poco antes de terminar el diseño de propuesta, al mismo tiempo que se comienza a redactar la memoria del proyecto. Tratándose de un trabajo experimental, se realizan los ensayos necesarios sobre el sistema, de manera que los resultados obtenidos sean suficientes como para llegar a conclusiones satisfactorias.

E. **Análisis de resultados**
   En esta etapa final se ordenan los resultados, estudiándolos hasta obtener conclusiones capaces de solventar el problema de partida.

F. **Redacción de memoria**
   Finalmente se recopila toda la información utilizada durante la creación del trabajo y se redacta bajo una serie de normas formales con el fin de transmitir los conocimientos adquiridos a todo aquel que le pueda interesar.
   Esta etapa puede comenzar en el momento en que se crea la propuesta, aprovechando los tiempos muertos de la fase de experimentación y concluye en último lugar, tras haber obtenido y analizado los resultados.

Anexo 2: Diagrama de Gantt

Tarea ID	Actividad	Después de:	Duración
A	Estudio y Análisis del problema	-	3 Semanas
B	Búsqueda de bibliografía	-	5 Semanas
C	Diseño de propuesta	A	3 Semanas
D	Experimentación	C	4 Semanas
E	Análisis de resultados	D	3 Semanas
F	Redacción de memoria	A	6 Semanas

Tabla. A.2. Tareas planificadas

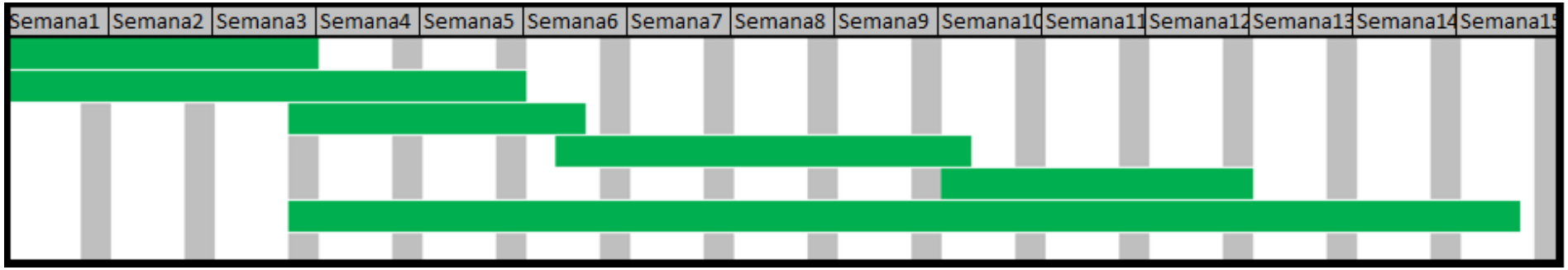


Fig. A.2. Diagrama de Gantt

Anexo 3: Material utilizado

Los recursos empleados para el desarrollo de este trabajo son los siguientes:

- **Hardware:**
  Para la realización de este trabajo ha sido necesaria la utilización de un ordenador de uso común.

- **Software:**
  Este trabajo ha sido desarrollado mediante software ofimático y software de cálculo y procesamiento de datos. Las licencias comerciales necesarias son:
  - Windows 7, incluyendo pack Microsoft Office.
  - Matlab R2016a, incluyendo pack de expansión para modelado de Redes Neuronales.

- **Material bibliográfico:**
  El material de consulta al que se ha recurrido es:
  - Libros de autores nacionales e internacionales.
  - Informes y publicaciones en revistas técnicas.
  - Apuntes y material didáctico de asignaturas relacionadas.
  - Sitios web especializados.