

UNIVERSIDAD DE CANTABRIA

Escuela Técnica Superior de Ingenieros Industriales y de Telecomunicación
Departamento de Ingenierías Química y Biomolecular

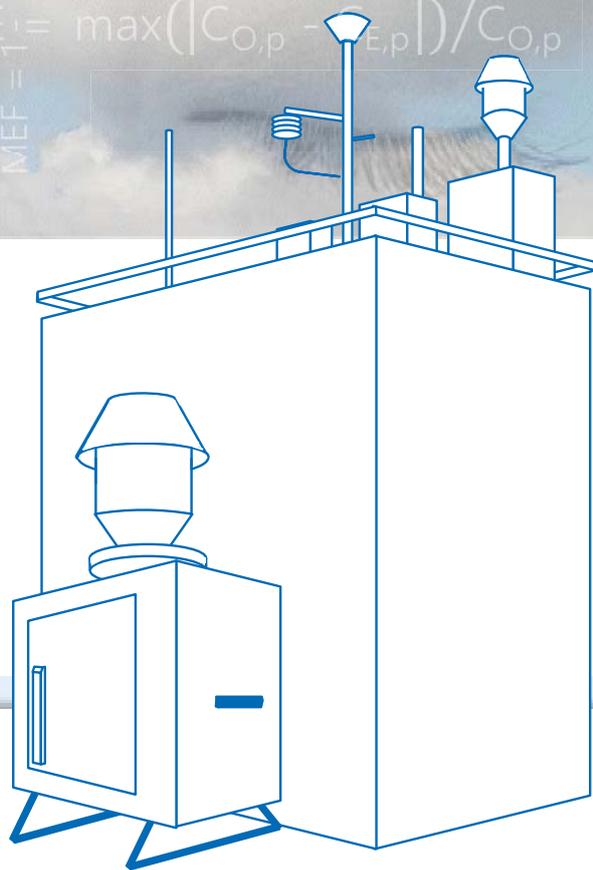


EVALUACIÓN DE LOS NIVELES DE METALES EN AIRE AMBIENTE MEDIANTE MODELOS ESTADÍSTICOS

ASSESSMENT OF METAL LEVELS IN AMBIENT AIR BY STATISTICAL MODELS

1.6	23.9	5.0	51.2	3.15	30.93
3.3	20.7	1.5	59.2	3.91	7.78
0.0	59.3	1.5	19.1	3.90	5.37
0.3	33.3	2.4	41.2	3.98	10.54
31.6	7.3	2.9	63.2	3.06	1.76
5.1	24.7	3.5	47.9	3.60	3.99
0.0	15.8	5.0	65.0	3.37	3.04
0.7	24.6	2.8	64.3	2.78	1.01
0.2	22.9	6.1	74.4	2.78	1.30
0.0	37.9	4.0	53.6	3.44	6.03
1.1	16.4	5.6	88.4	3.10	0.34
0.4	14.2	4.9	89.7	3.45	2.91
0.0	14.3	4.7	73.8	2.52	3.68
5.0	30.5	5.1	81.5	2.95	0.34
14.8	20.5	3.7	57.7	3.12	2.91
0.0	17.8	4.2	65.3	3.16	1.33
0.3	19.3	2.8	58.5	3.42	4.09
0.3	15.1	4.1	41.6	3.08	3.03

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (C_{O_i} - C_{E_i})^2}$$
$$MEF = 1 - \frac{RMSE}{\sigma_0}$$
$$MEF = 1 - \frac{\max(|C_{O,p} - C_{E,p}|)}{C_{O,p}}$$



Director de tesis:
Dr. Ignacio Fernández Olmo

Germán Santos Bregel



UNIVERSIDAD DE CANTABRIA

Escuela Técnica Superior de Ingenieros Industriales y de Telecomunicación

Departamento de Ingenierías Química y Biomolecular



Evaluación de los niveles de metales en aire ambiente mediante modelos estadísticos

Assessment of metal levels in ambient air by statistical models

Memoria de Tesis Doctoral presentada para optar al Título de
Doctor por la Universidad de Cantabria

Programa Oficial de Doctorado en Ingeniería Química y de Procesos (BOE núm. 36, de 10 de febrero de 2010.
RUCT: 5311209) con Mención hacia la Excelencia (BOE núm. 253, de 20 de Octubre de 2011. Referencia: MEE2011-0031)

Germán Santos Bregel

Director de tesis:

Dr. Ignacio Fernández Olmo

Santander, mayo 2016

Este trabajo ha sido realizado en el grupo de investigación Desarrollo de Procesos Químicos y Control de Contaminantes (DePRO) perteneciente al Departamento de Ingenierías Química y Biomolecular de la Universidad de Cantabria. La investigación ha sido financiada por el Ministerio de Economía y Competitividad de España (MINECO) a través del proyecto *Herramientas para la evaluación de los metales pesados y compuestos orgánicos persistentes (COPs) en aire ambiente a escala local/regional* con referencia CTM2010-16068.

Durante la ejecución del presente trabajo, su autor, Germán Santos Bregel, ha disfrutado de una ayuda predoctoral de Formación de Personal Investigador (FPI) del Ministerio de Economía y Competitividad de España concedida por resolución de 21 de julio de 2011, con referencia BES-2011-047908. Asimismo, Germán Santos Bregel ha disfrutado también de una ayuda económica a la movilidad predoctoral del Ministerio de Economía y Competitividad de España concedida por resolución de 23 de abril de 2013, con referencia EEBB-I-13-07691, para la realización de una estancia breve de investigación de tres meses de duración en el centro de investigación Maison de la Recherche en Environnement Industriel (MREI) adscrito a la Université du Littoral Côte d'Opale en la ciudad de Dunkerque (Francia).

Mi más sincero agradecimiento a dichas instituciones.

Agradecimientos

Dice el refrán que "es de bien nacido ser agradecido" y aún más cuando hay motivo. Por ello, me gustaría dejar a un lado la objetividad que se le presupone a un documento como este, y aprovechar estas líneas para dar las gracias a todo aquel que de un modo u otro, en mayor o menor medida, haya contribuido a este trabajo.

En primer lugar, me gustaría expresar mi más sincera gratitud a mi director de tesis, el Dr. Ignacio Fernández Olmo, por la oportunidad que me ha brindado al permitirme incorporarme a su equipo de trabajo, por la confianza que ha depositado en mí para llevar a cabo esta tesis y por orientarme e instruirme durante toda esta etapa. Quisiera agradecer especialmente su paciencia, comprensión, esfuerzo, entrega e implicación en el proyecto. Ha sido un honor.

En segundo lugar, he de dar las gracias a mis compañeros de trabajo de la línea de investigación, Mariano Puente y Sara Ruiz, tanto por la acogida que me ofrecieron, como por sus innumerables enseñanzas y consejos, que me han acompañado durante todo el camino y me han servido de gran ayuda. Una buena parte del trabajo que he desarrollado ha sido gracias a vosotros. Por otra parte, también me gustaría dar las gracias a Ana Hernández, que aunque ha llegado algo más tarde, ha demostrado ser una magnífica incorporación al equipo.

También me gustaría expresar mi gratitud a los técnicos de laboratorio, en especial a Beatriz Arce, por su inestimable ayuda.

Un agradecimiento especial a mis compañeros de despacho, los hermanos Abejón, Azucena y Ricardo, por estar siempre dispuestos a echar una mano en todo lo necesario. La convivencia con vosotros ha sido inmejorable. Ha sido un placer.

Asimismo, me gustaría expresar mi agradecimiento al Prof. Dr. Dominique Courcot y al Dr. Frédéric Ledoux de la Maison de la Recherche en Environnement Industriel por su amabilidad al acogerme en sus instalaciones durante la realización de mi estancia, en especial en un momento tan complicado para el centro de investigación, y por estar abiertos a la colaboración en todo momento. No puedo olvidarme de todos los chicos de la UCEIV, que me facilitaron la adaptación y la integración (a nivel personal y laboral) en un nuevo entorno, tratándome como a uno más desde el primer momento.

También, quisiera extender mi agradecimiento a todos los miembros del grupo de investigación DePRO y del Departamento de Ingenierías Química y Biomolecular, los que están y los que estuvieron, por haberme permitido formar parte de su particular familia durante este tiempo. En especial, a mis compañeros de doctorado, algunos de los cuales se han convertido en verdaderos amigos, y que, con su apoyo, sentido del humor y ayuda desinteresada, han hecho que cada día mereciera la pena.

Por último, pero no por ello menos importante (al contrario), me gustaría dar las gracias a mi familia, en particular a mis padres, quienes han vivido este proceso conmigo casi en primera persona, tanto los momentos buenos como los no tan buenos, mostrándome siempre su apoyo incondicional.

A todos, ¡MUCHAS GRACIAS!

Índice de contenidos

Resumen Abstract	i
Nomenclatura	vii
Listado de figuras	ix
Listado de tablas	xi
Capítulo 1. Planteamiento	1
1.1. Contaminación atmosférica por metales y metaloides	3
1.1.1. Fuentes atmosféricas	4
1.1.2. Niveles en aire ambiente	6
1.1.3. Efectos sobre la salud	9
1.1.3.1. <i>Arsénico</i>	9
1.1.3.2. <i>Cadmio</i>	9
1.1.3.3. <i>Níquel</i>	10
1.1.3.4. <i>Plomo</i>	10
1.1.3.5. <i>Cromo</i>	10
1.1.3.6. <i>Manganeso</i>	10
1.1.3.7. <i>Vanadio</i>	11
1.2. Regulación de los niveles de metales y metaloides en aire ambiente	11
1.3. Niveles de metales y metaloides en aire ambiente en Cantabria en el marco de la regulación europea	13
1.4. Modelización de la calidad del aire	15
1.4.1. Modelos determinísticos	16
1.4.2. Modelos estadísticos	19
1.5. Modelos estadísticos empleados en el modelado de contaminantes atmosféricos	20
1.5.1. Modelos estadísticos para contaminantes mayoritarios	21
1.5.1.1. <i>Óxidos de Nitrógeno</i>	21
1.5.1.2. <i>Dióxido de azufre</i>	21
1.5.1.3. <i>Monóxido de carbono</i>	21
1.5.1.4. <i>Ozono Troposférico</i>	22
1.5.1.5. <i>Partículas en suspensión</i>	22
1.5.2. Modelos estadísticos para contaminantes minoritarios	24
1.6. Hipótesis de partida y objetivos	25
1.7. Bibliografía	27

Capítulo 2. Bases teóricas de regresión basada en mínimos cuadrados parciales (PLSR) y redes neuronales artificiales (ANNs)	33
2.1. Regresión de mínimos cuadrados parciales	35
2.1.1. Concepto y fundamentos teóricos	35
2.1.2. Aplicaciones	39
2.2. Redes neuronales artificiales	42
2.2.1. Concepto y fundamentos teóricos	42
2.2.2. Aplicaciones	46
2.3. Bibliografía	49
Capítulo 3. Metodología	53
3.1. Descripción de las áreas de estudio	55
3.2. Bases de datos ambientales: variables de influencia en niveles en inmisión	59
3.2.1. Contaminantes mayoritarios	59
3.2.2. Variables meteorológicas	59
3.2.2.1. <i>El viento</i>	60
3.2.2.2. <i>Temperatura y radiación solar</i>	60
3.2.2.3. <i>Humedad relativa</i>	60
3.2.2.4. <i>Precipitación</i>	61
3.2.3. Variables meteorológicas relacionadas con la estabilidad atmosférica	61
3.2.3.1. <i>Altura de la capa de mezcla</i>	62
3.2.3.2. <i>Inversión térmica</i>	66
3.2.3.3. <i>Número de Richardson</i>	66
3.2.3.4. <i>Clases de estabilidad de Pasquill-Gifford</i>	67
3.3. Selección de las bases de datos de partida	69
3.3.1. Variables dependientes	69
3.3.2. Variables independientes	70
3.3.2.1. <i>Variables continuas</i>	70
3.3.2.2. <i>Variables nominales</i>	72
3.4. Ampliación de las bases de partida mediante muestreos y determinaciones adicionales	73
3.4.1. Muestreos de PM ₁₀	73
3.4.2. Determinación de metales/metaloideos en PM ₁₀	74
3.5. Reconciliación/procesado de datos	75
3.5.1. Detección de valores atípicos multivariantes	75
3.5.2. Normalización de variables	76
3.6. Procedimiento de división de las bases de datos	77
3.7. Árboles de clasificación para estimación de muestras por debajo del límite de detección	79
3.8. Desarrollo de modelos estadísticos	80
3.8.1. Estrategia seguida en la elaboración de los modelos PLSR	80
3.8.2. Estrategia seguida en la elaboración de los modelos ANN	81
3.9. Criterios de calidad/desempeño de los modelos	82
3.9.1. Incertidumbre de acuerdo a la UE	82
3.9.2. Otros criterios de desempeño de los modelos	83
3.10. Metodología para la identificación y evaluación del desempeño de los mejores modelos	85
3.11. Bibliografía	88

Capítulo 4. Estimación de niveles de metales/metaloideos en áreas urbanas mediante PLSR y ANNs	91
4.1. Estimación de los niveles de As, Cd, Ni y Pb en áreas urbanas de Cantabria (Castro Urdiales y Reinosa) a partir de campañas extensivas	93
4.1.1. Niveles de As, Cd, Ni y Pb en Castro Urdiales y Reinosa	93
4.1.2. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: media y error de acuerdo a las Directivas de calidad del aire	95
4.1.3. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: otros estadísticos de desempeño de los modelos	98
4.2. Estimación de los niveles de metales regulados (Ni y Pb) y no regulados (Cr, Mn y V) en Dunkerque (Francia) a partir de una campaña intensiva	105
4.2.1. Niveles de los metales estudiados	105
4.2.2. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: media y error de acuerdo a las Directivas de calidad del aire	107
4.2.2.1. <i>Estimación de metales regulados: Ni y Pb</i>	107
4.2.2.2. <i>Estimación de metales no regulados: Cr, Mn y V</i>	110
4.2.3. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: otros estadísticos de desempeño de los modelos	112
4.2.3.1. <i>Estimación de metales regulados: Ni y Pb</i>	112
4.2.3.2. <i>Estimación de metales no regulados: Cr, Mn y V</i>	114
4.3. Influencia de variables meteorológicas relacionadas con la estabilidad atmosférica en el desempeño de modelos PLSR y ANN para la estimación de los niveles de As, Cd, Ni y Pb en un área urbana: Santander	116
4.3.1. Niveles en inmisión de As, Cd, Ni y Pb	116
4.3.2. Evaluación de los modelos estadísticos desarrollados para la estimación de los niveles en inmisión de As, Cd, Ni y Pb como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire	116
4.3.3. Influencia de la consideración de variables meteorológicas relacionadas con la estabilidad atmosférica	123
4.3.3.1. <i>Clasificación de las variables independientes de acuerdo a su coeficiente de correlación</i>	123
4.3.3.2. <i>Estimación de los niveles en inmisión de As, Cd, Ni y Pb considerando nuevas variables meteorológicas</i>	124
4.4. Bibliografía	129
Capítulo 5. Conclusiones Chapter 5. Conclusions	131
Anexo I. Difusión de resultados	137
AI.1. Comunicaciones a congresos	139
AI.2. Publicaciones en revistas científicas	140

Resumen

La Unión Europea (UE) a través de la Directiva Marco sobre calidad del aire (Directiva 2008/50/CE) y la cuarta directiva hija (Directiva 2004/107/CE) establece unos valores límite/objetivo para metales y semimetales en aire como valores medios anuales: As (6 ng/m^3), Cd (5 ng/m^3), Ni (20 ng/m^3) y Pb (500 ng/m^3). Asimismo, se permite que la evaluación de la calidad del aire se pueda realizar únicamente mediante técnicas de modelización o de estimación objetiva o de ambas en aquellas zonas en las que los niveles de contaminantes se encuentren por debajo del correspondiente umbral inferior de evaluación, de acuerdo a la nomenclatura empleada por las Directivas de calidad del aire de la UE, que viene definido en términos relativos con respecto al valor límite/objetivo (40% del valor límite/objetivo para As y Cd y 50 %, para Ni y Pb).

El objetivo principal de este trabajo se ha centrado en el desarrollo de una metodología para la estimación de los niveles en aire ambiente de algunos metales y semimetales presentes en el material particulado en suspensión y la evaluación de la adecuación de dicha metodología como herramienta de evaluación de la calidad del aire con respecto a As, Cd, Ni y Pb de acuerdo con los requerimientos de incertidumbre establecidos por la legislación europea como criterios de calidad para las técnicas de modelización y de estimación objetiva destinadas a este fin. Esta metodología está basada en el empleo de modelos estadísticos desarrollados a partir de las técnicas de regresión por mínimos cuadrados parciales (PLSR, por sus siglas en inglés) y de redes neuronales artificiales (ANNs, por sus siglas en inglés), que permiten determinar la concentración en inmisión de los contaminantes atmosféricos objeto de estudio en zonas urbanas a escala local en función de los niveles en aire ambiente de otros contaminantes atmosféricos mayoritarios y de las condiciones meteorológicas, lo que reduciría la necesidad de la realización de medidas experimentales para su posterior análisis en laboratorio.

La técnica basada en PLSR presenta ventajas frente a otras técnicas como regresión lineal múltiple o regresión basada en componentes principales. Si bien ha sido empleada en numerosos trabajos de distintas disciplinas, destacando principalmente la quimiometría, no se han encontrado estudios que empleen PLSR para la estimación de los niveles de metales en aire ambiente. Por otra parte, los modelos basados en ANNs han sido ampliamente utilizados en la literatura de calidad del aire para la estimación de la concentración de contaminantes atmosféricos mayoritarios tales como óxidos de nitrógeno, dióxido de azufre, ozono troposférico y material particulado. No obstante, pocos estudios que emplean esta técnica se centran en la estimación de la concentración de microcontaminantes en aire.

El punto de partida de esta Tesis Doctoral ha consistido en la selección de bases de datos de los niveles de los contaminantes minoritarios a evaluar, así como de las variables independientes que podrían influir a priori en los niveles de dichos contaminantes. Para ello se ha elegido en primer lugar la información disponible en el grupo de investigación DePRO (Desarrollo de Procesos Químicos y Control de Contaminantes) sobre los niveles de As, Cd, Ni y Pb en material particulado menor de 10 μm (PM_{10}) durante los años 2008 y 2009 en tres áreas urbanas de Cantabria (Castro Urdiales, Reinosa y Santander). Para estas tres áreas se ha creído conveniente ampliar la base de datos por medio de la determinación adicional de los niveles de dichos elementos en los años 2010 (Castro Urdiales y Reinosa) y 2011 (Santander). En segundo lugar, y en el marco de una colaboración con l'Unité de Chimie Environnementale et Interactions sur le Vivant (UCEIV), de l'Université du Littoral Côte d'Opale, se ha elegido una segunda base de datos con los niveles de Ni, Pb, Cr, Mn y V en la ciudad de Dunkerque, en la costa norte de Francia.

Así, la metodología desarrollada se ha aplicado a tres situaciones diferenciadas: (i) pequeñas áreas urbanas de Cantabria en las que se han realizado campañas extensivas de muestreo de PM_{10} , Castro Urdiales (litoral) y Reinosa (interior); (ii) una zona urbana costera con fuerte contribución industrial en la que se ha realizado una campaña intensiva de muestreo (Dunkerque, Francia); y (iii) un área urbana más compleja a partir de una campaña de muestreo extensiva, Santander. En esta última, se ha evaluado además la influencia de la consideración de variables meteorológicas relacionadas con la estabilidad atmosférica como valores de entrada a los modelos.

Los resultados obtenidos en esta tesis indican que la metodología desarrollada representa una herramienta válida para la evaluación de la calidad del aire con respecto a As, Cd, Ni y Pb en el marco de la legislación europea, como complemento a la determinación cuantitativa de los niveles de estos contaminantes en aire ambiente mediante la toma de medidas experimentales, lo que permitiría reducir significativamente el número de determinaciones necesarias, con el consecuente ahorro de recursos.

Abstract

The European Union (EU) through the Air Quality Framework Directive (Directive 2008/50/EC) and the 4th Daughter Directive (Directive 2004/107/EC) establishes limit/target values for metals and metalloids in ambient air as annual mean values: As (6 ng/m³), Cd (5 ng/m³), Ni (20 ng/m³) y Pb (500 ng/m³). Furthermore, the air quality assessment by modelling and objective estimation techniques or both is permitted in those zones and agglomerations where the levels of pollutants are below the corresponding lower assessment threshold, according to the nomenclature used in the EU Air Quality Directives, which is defined in relative terms with respect to the limit/target value (40% of the limit/target value for As and Cd, and 50% for Ni and Pb).

The main goal of this work is focused on the development of a methodology to estimate the ambient air levels of some metals and metalloids that are present in suspended particulate matter and the evaluation of the adequacy of that methodology as an air quality assessment tool with respect to As, Cd, Ni y Pb, according to the uncertainty requirements established by the European legislation as quality criteria for modelling and objective estimation techniques to that end. This methodology involves the use of statistical models, which are developed based on Partial Least Squares Regression (PLSR) and Artificial Neural Networks (ANNs), and allows determining the immission concentration of the studied atmospheric pollutants at a local scale as a function of the ambient air levels of major air pollutants and the meteorological conditions, what would reduce the necessity of taking experimental measurements for their subsequent analysis in laboratory.

PLSR presents advantages over other techniques such as multiple linear regression or principal component regression. Although it has been applied in numerous studies from different fields, principally chemometrics, no studies on the application of PLSR for the estimations of metal levels in ambient air have been found in the literature. Models based on ANNs have been widely applied in the air quality literature for the estimation of the immission concentration of major air pollutants such as nitrogen oxides, sulphur dioxide, tropospheric ozone and particulate matter. However, few studies employing this technique are focused on the estimation of the ambient concentration of micropollutants.

The starting point of this doctoral thesis consisted in the selection of the databases of the levels of minor air pollutants and the independent variables that in principle may affect their levels. To achieve this goal, firstly, a screening of the available information gathered by the research group DePRO (Development of Chemical Processes and Pollution Control) was conducted, selecting

that related to the levels of As, Cd, Ni and Pb in particulate matter lower than 10 μm (PM_{10}) for 2008 and 2009 in three urban areas in Cantabria (Castro Urdiales, Reinosa and Santander). In these areas it was considered convenient to implement an enlargement of the databases by means of additional determinations of the levels of the aforementioned elements in 2010 (Castro Urdiales and Reinosa) and 2011 (Santander). Secondly, within a collaboration framework with l'Unité de Chimie Environnementale et Interactions sur le Vivant (UCEIV), from l'Université du Littoral Côte d'Opale, a second database comprising the levels of Ni, Pb, Cr, Mn and V in the city of Dunkerque (Northern coast of France) was also selected.

The developed methodology has been applied to three different situations: (i) small urban areas of Cantabria where extensive PM_{10} sampling campaigns were performed, Castro Urdiales (coastal site) and Reinosa (inland site); (ii) an urban coastal zone with a strong industrial contribution to atmospheric pollution where an intensive sampling campaign was performed (Dunkerque, France); and (iii) a more complex urban area where an extensive PM_{10} sampling campaign was conducted, Santander. In the latter, the influence on model performance of the consideration of meteorological variables related to atmospheric stability as additional model inputs was also evaluated.

The results obtained in this thesis indicate that the developed methodology represents a valid tool to perform the air quality assessment with respect to As, Cd, Ni and Pb in the framework of the European legislation, as a complement to the quantitative determination of the levels of these pollutants in ambient air by experimental measurements, which would allow reducing significantly the number of necessary determinations, with the consequent saving of resources.

Nomenclatura

Acrónimos y abreviaturas

AEMA	Agencia Europea de Medio Ambiente	OMS	Organización Mundial de la Salud
AEMET	Agencia Estatal de Meteorología	Pb	Plomo
ANFIS	Sistema adaptativo de inferencia neuro-difusa	PCA	Análisis de componentes principales
ANNs	Redes neuronales artificiales	PCR	Regresión por componentes principales
ARIMA	Modelo auto-regresivo de media móvil	PLS	Mínimos cuadrados parciales
ARX	Modelo auto-regresivo de entrada exógena	PLSR	Regresión por mínimos cuadrados parciales
As	Arsénico	PM ₁₀	Material particulado de diámetro aerodinámico igual o inferior a 10 µm
Cd	Cadmio	PM _{2,5}	Material particulado de diámetro aerodinámico igual o inferior a 2,5 µm
CE	Comisión Europea	RSPM	Material particulado respirable en suspensión
CIMA	Centro de Investigación del Medio Ambiente	RVCCA	Red Automática de Vigilancia y Control de la Calidad del Aire de Cantabria
CLP	Capa límite planetaria	SIMPLS	Statistically Inspired Modification of Partial Least Squares
CMT	Centro Meteorológico Territorial de la AEMET en Santander	SO ₂	Dióxido de azufre
CO	Monóxido de carbono	SODAR	Sonic Detection and Ranging
Cr	Cromo	SOM	Mapas auto-organizados
CSIC	Consejo Superior de Investigaciones Científicas	SPXY	Método de división de datos basado en distancias x-y
DePRO	Grupo de Investigación Desarrollo de Procesos Químicos y Control de Contaminantes	TSP	Partículas totales en suspensión
ETSIIT	Escuela Técnica Superior de Ingenieros Industriales y de Telecomunicación. Lugar de muestreo en Santander	UCEIV	Unité de Chimie Environnementale et Interactions sur le Vivant
FAIRMODE	Foro de la modelización de calidad del aire en Europa	UE	Unión Europea
ICP-MS	Espectrometría de masas con plasma de acoplamiento inductivo	UIE	Umbral inferior de evaluación
IMD	Intensidad media diaria de tráfico	USE	Umbral superior de evaluación
JRC	Joint Research Centre	USEPA	Agencia de Protección Ambiental de los Estados Unidos
LD	Límite de detección	VL/VO	Valor límite/objetivo
LSR	Regresión por mínimos cuadrados		
MAGRAMA	Ministerio de Agricultura, Alimentación y Medio Ambiente del Gobierno de España		
MD	Distancia de Mahalanobis		
MDKS	Modificación del algoritmo de Kennard-Stone basada en la distancia de Mahalanobis		
MLP	Perceptrón Multicapa		
MLR	Regresión lineal múltiple		
Mn	Manganeso		
Ni	Níquel		
NIPALS	Non-Iterative Partial Least Squares		
NO	Monóxido de nitrógeno		
NO ₂	Dióxido de nitrógeno		
NO _x	Óxidos de nitrógeno		
O ₃	Ozono		

Símbolos

b_0	Intercepto en una regresión lineal	u	Componente horizontal del viento (m/s)
b_1, \dots, b_n	Coefficientes de regresión lineal	u^*	Velocidad de fricción
B	Matriz de coeficientes de regresión en un modelo PLSR	W	Matriz de pesos en un modelo PLSR
Bias	Sesgo (ng/m^3)	WD	Dirección de viento predominante ($^\circ$)
$C_{O,i}, C_{E,i}$	Concentración en inmisión de contaminante observada (O) y estimada (E) asociada a la muestra i (ng/m^3)	WE	Fin de semana (-)
CRMSE	Error cuadrático medio cruzado (ng/m^3)	WS	Velocidad de viento predominante (m/s)
E	Matriz de error de la descomposición de X en un modelo PLSR	X	Matriz de variables independientes o predictoras
F	Matriz de error de la regresión en un modelo PLSR	\bar{x}	Valor medio de variable de entrada
f	Parámetro de Coriolis	x_1, \dots, x_n	Variables independientes
FB	Sesgo fraccional (-)	x_i	Valor observado de variable de entrada
FV	Varianza fraccional (-)	Y	Matriz de variables dependientes o respuesta
G	Matriz de error de la descomposición de Y en un modelo PLSR	y	Variable dependiente o respuesta
g	Aceleración de la gravedad (m/s^2)	z	Altura geopotencial (m)
I	Inversión térmica (-)	z_i	Valor tipificado de variable de entrada asociado a la muestra i
k	Número de variables X en un modelo PLSR	Γ	Gradiente adiabático del aire seco ($^\circ\text{C}/\text{m}$)
l	Número de factores o variables latentes en un modelo PLSR	σ	Desviación estándar
L	Longitud de Monin-Obukhov	σ_O, σ_E	Desviación estándar observada (O) y estimada (E) (ng/m^3)
LnPM_{10}	Media del logaritmo neperiano de la concentración de PM_{10} ($\mu\text{g m}^{-3}$)		
LV	Valor límite (ng/m^3)		
m	Número de variables Y en un modelo PLSR		
MEF	Eficiencia del modelo (-)		
MH	Altura de la capa de mezcla (m)		
N, n	Número total de observaciones en el conjunto de entrenamiento		
NMSE	Error cuadrático normalizado (-)		
NO_x	Concentración media de óxidos de nitrógeno, calculada como el sumatorio de las concentraciones medias de NO y NO_2 ($\mu\text{g}/\text{m}^3$)		
O_3	Concentración media de ozono troposférico ($\mu\text{g}/\text{m}^3$)		
P	Matriz de <i>loadings</i> de X en un modelo PLSR (Capítulo 2); Presión media (mbar) (Tabla 3.7., Capítulo 3)		
PG	Clase de estabilidad de Pasquill-Gifford (-)		
PP	Precipitación acumulada (L/m^2)		
Q	Matriz de <i>loadings</i> de Y en un modelo PLSR		
r	Coefficiente de correlación (-)		
RDE	Error relativo de la Directiva (%)		
RH	Humedad relativa media (%)		
R_i	Número adimensional de Richardson (-)		
R_{iB}	Número de Richardson <i>Bulk</i> (-)		
R_{iF}	Número de Richardson <i>Flux</i> (-)		
RME	Error relativo máximo (%)		
RMSE	Error cuadrático medio (ng/m^3)		
SD	Intrusión de polvo sahariano (-)		
SE	Estación del año (-)		
SO_2	Concentración media de dióxido de azufre ($\mu\text{g}/\text{m}^3$)		
T	Matriz de <i>scores</i> de X en un modelo PLSR (Capítulo 2); Temperatura (K) (Ecuación 3.1, Capítulo 3); Temperatura media ($^\circ\text{C}$) (Tabla 3.7., Capítulo 3)		
U	Matriz de <i>scores</i> de Y en un modelo PLSR		

Listado de figuras

Capítulo 1. Planteamiento

Figura 1.1. Métodos de evaluación de la calidad del aire según la Unión Europea.	14
Figura 1.2. Niveles de As, Cd, Ni y Pb en tres zonas urbanas de Cantabria.	15
Figura 1.3. Esquema de modelo de calidad del aire.	17

Capítulo 2. Bases teóricas de regresión basada en mínimos cuadrados parciales (PLSR) y redes neuronales artificiales (ANNs)

Figura 2.1. Representación esquemática de la descomposición matricial de un modelo PLSR.	37
Figura 2.2. Representación esquemática de una neurona biológica.	42
Figura 2.3. Esquema básico de una neurona artificial.	43
Figura 2.4. Esquema general de una red neuronal hacia adelante y una única capa oculta	45

Capítulo 3. Metodología

Figura 3.1. Localización de los lugares de muestreo en Cantabria (España).	57
Figura 3.2. Localización del lugar de muestreo en Dunkerque (Francia).	58
Figura 3.3. Determinación gráfica de la altura de capa de mezcla según el método propuesto por Holzworth (1967).	65
Figura 3.4. Perfil de temperaturas que establece las categorías de estabilidad atmosférica (Heinsohn y Kabel, 1999).	68
Figura 3.5. Representación esquemática del diagrama objetivo (Pederzoli y col., 2011).	85
Figura 3.6. Diagrama de flujo de la metodología propuesta para la evaluación de la calidad del aire mediante modelos estadísticos (Santos y col., 2016).	87

Capítulo 4. Estimación de niveles de metales/metaloideos en áreas urbanas mediante PLSR y ANNs

Figura 4.1. Diagrama de cajas de los niveles en aire ambiente de As, Cd, Ni y Pb en los años 2008 ('08), 2009 ('09) y 2010 ('10) en los lugares de muestreo de: a) Castro Urdiales y b) Reinosa. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE. UIE = $2,4 \text{ ng/m}^3$ (As), 2 ng/m^3 (Cd), 10 ng/m^3 (Ni) y 250 ng/m^3 (Pb). Los valores medios correspondientes al año 2008 se han obtenido de Arruti y col. (2011).	94
Figura 4.2. Comparación entre los niveles medios de concentración estimados y observados obtenidos en Castro Urdiales para el subconjunto de validación externa con respecto a los correspondientes umbrales de evaluación y valores objetivo/límite: a) As, b) Cd, c) Ni y d) Pb. VO/VL = 6 ng/m^3 (As), 5 ng/m^3 (Cd), 20 ng/m^3 (Ni), 500 ng/m^3 (Pb); USE = 60 % (As y Cd) y 70% (Ni y Pb); UIE = 40% (As y Cd) y 50% (Ni y Pb).	99

Capítulo 4. Estimación de niveles de metales/metaloides en áreas urbanas mediante PLSR y ANNs

- Figura 4.3. Comparación entre los niveles medios de concentración estimados y observados obtenidos en Reinosa para el subconjunto de validación externa con respecto a los correspondientes umbrales de evaluación y valores objetivo/límite: a) As, b) Cd, c) Ni y d) Pb. VO/VL = 6 ng/m³ (As), 5 ng/m³ (Cd), 20 ng/m³ (Ni), 500 ng/m³ (Pb); USE = 60 % (As y Cd) y 70% (Ni y Pb); UIE = 40% (As y Cd) y 50% (Ni y Pb). 99
- Figura 4.4. Ejemplificación gráfica de subestimación a pesar del ajuste prácticamente lineal de los datos ($r = 0,97$). 102
- Figura 4.5. Ajuste de los datos del subconjunto de entrenamiento de modelos PLSR y ANN desarrollados en Castro Urdiales para la estimación de los niveles de: a) Ni y b) Pb. 104
- Figura 4.6. Diagrama de cajas de los niveles de Ni, Pb, Cr, Mn y V medidos en Dunkerque en el periodo de estudio. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE o UIE-equivalente. UIE = 10 ng/m³ (Ni) y 250 ng/m³ (Pb); UIE-equivalente = 150 ng/m³ (Mn) y valor máximo observado (V y Cr). 107
- Figura 4.7. Comparación entre las concentraciones normalizadas observadas y modeladas de Ni y Pb en Dunkerque. 114
- Figura 4.8. Diagrama de cajas de los niveles en aire ambiente de As, Cd, Ni y Pb en los años 2008 ('08), 2009 ('09) y 2011 ('11) en el punto de muestreo de Santander. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE. UIE = 2,4 ng/m³ (As), 2 ng/m³ (Cd), 10 ng/m³ (Ni) y 250 ng/m³ (Pb). Los valores medios correspondientes a los años 2008 y 2009 se han obtenido de Arruti y col. (2011). 117
- Figura 4.9. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el As (azul), Cd (amarillo), Ni (verde) y Pb (rojo). Los puntos de datos relativos a la PLSR se representan por medio de un círculo y los relativos a las ANNs por medio de un rombo. Los marcadores rellenos y huecos se utilizan para representar los resultados referentes al entrenamiento (E) y a la validación externa (V), respectivamente. 122
- Figura 4.10. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el Pb para la determinación de la influencia de la PG_{am} en el desempeño de los mismos. 127
- Figura 4.11. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el Ni para la determinación de la influencia de la PG_{am} en el desempeño de los mismos. 128

Listado de tablas

Capítulo 1. Planteamiento

Tabla 1.1.	Concentraciones medias (ng/m^3) de algunos metales y semimetales en PM_{10} de los principales estudios a nivel regional, nacional y europeo reportados en la bibliografía. Tabla adaptada de Fernández-Olmo y col., 2016.	7
Tabla 1.2.	Objetivos de calidad para As, Cd, Ni y Pb en la regulación europea.	12
Tabla 1.3.	Valores guía de la calidad del aire de la OMS para los metales y semimetales considerados en este trabajo.	13
Tabla 1.4.	Umbral de evaluación de la calidad del aire con respecto a As, Cd, Ni y Pb según las Directivas europeas.	14

Capítulo 2. Bases teóricas de regresión basada en mínimos cuadrados parciales (PLSR) y redes neuronales artificiales (ANNs)

Tabla 2.1.	Aplicaciones de la técnica PLS con fines de regresión.	40
Tabla 2.2.	Aplicaciones de la técnica PLS con fines de clasificación.	41
Tabla 2.3.	Analogía entre una neurona biológica y una neurona artificial (Flórez y Fernández, 2008)	44
Tabla 2.4.	Aplicaciones de las ANNs	47

Capítulo 3. Metodología

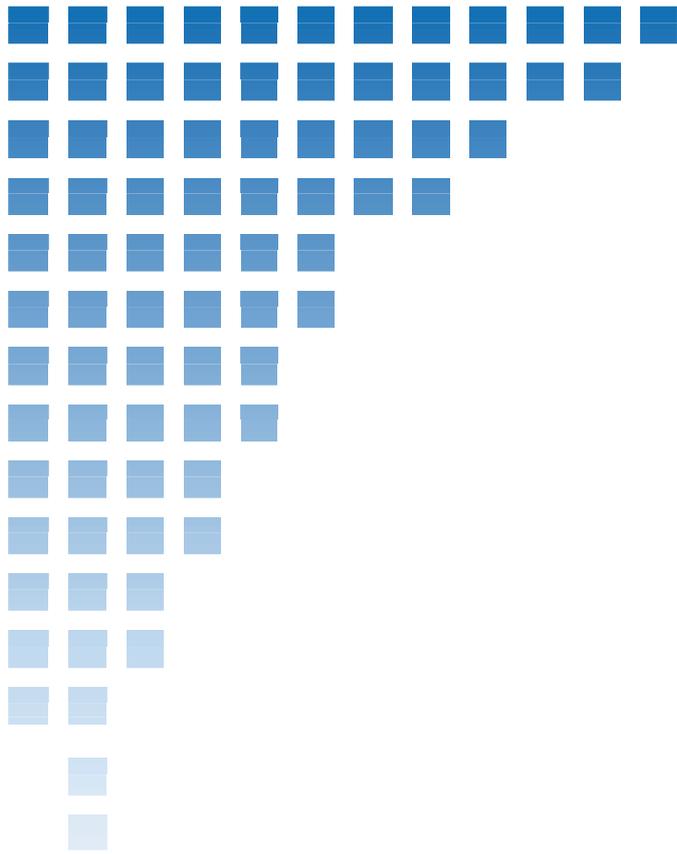
Tabla 3.1.	Criterio de estimación de la altura de la capa de mezcla a partir de las categorías de estabilidad atmosférica de Pasquill (Spadaro, 1999).	63
Tabla 3.2.	Diagnóstico de algunas ecuaciones desarrolladas para el cálculo de la altura de la capa de mezcla en función de la clase de estabilidad que presenta la atmósfera. Tabla adaptada de Burzynski y col. (2004).	64
Tabla 3.3.	Equivalencia nominal para la introducción de la inversión térmica como variable de entrada a los modelos.	66
Tabla 3.4.	Categorías de estabilidad atmosférica según el criterio Pasquill-Gifford.	67
Tabla 3.5.	Categorías de radiación solar (Turner, 1969).	68
Tabla 3.6.	Equivalencia entre valores cuantitativos del R_i y las clases de estabilidad de Pasquill-Gifford (Woodward, 1998).	69
Tabla 3.7.	Variables continuas utilizadas como variables de entrada a los modelos.	72
Tabla 3.8.	Variables meteorológicas relacionadas con la estabilidad atmosférica para su uso como variables de entrada adicionales a los modelos a desarrollar en Santander.	72
Tabla 3.9.	Variables nominales utilizadas como variables de entrada a los modelos.	73
Tabla 3.10.	Condiciones del método de digestión.	75

Capítulo 4. Estimación de niveles de metales/metaloides en áreas urbanas mediante PLSR y ANNs

Tabla 4.1.	Evaluación de los modelos desarrollados en Castro Urdiales y Reinososa como herramientas de evaluación de la calidad del aire.	96
Tabla 4.2.	Estadísticos para la evaluación del desempeño de los modelos desarrollados en Castro Urdiales y Reinososa.	101

Capítulo 4. Estimación de niveles de metales/metaloides en áreas urbanas mediante PLSR y ANNs

Tabla 4.3.	Evaluación de los resultados de los modelos desarrollados en Dunkerque para los metales regulados (Ni y Pb) desde el punto de vista de una herramienta de evaluación de la calidad del aire.	108
Tabla 4.4.	Evaluación de los resultados de los modelos desarrollados en Dunkerque para los metales no regulados (Cr, Mn y V) desde el punto de vista de una herramienta de evaluación de la calidad del aire.	111
Tabla 4.5.	Índices estadísticos para la evaluación del desempeño de los modelos de estimación de Ni y Pb desarrollados en Dunkerque.	112
Tabla 4.6.	Índices estadísticos para la evaluación del desempeño de los modelos de estimación de Cr, Mn y V desarrollados en Dunkerque.	115
Tabla 4.7.	Índices de incertidumbre y error en la estimación del valor medio de los niveles de As, Cd, Ni y Pb mediante modelos estadísticos desarrollados en Santander basados en PLSR y ANNs bajo los enfoques A y B.	118
Tabla 4.8.	Otros índices de desempeño de los modelos desarrollados en Santander basados en PLSR y ANNs para la estimación de los niveles de As, Cd, Ni y Pb bajo los enfoques A y B.	119
Tabla 4.9.	Clasificación de las variables independientes consideradas en los modelos desarrollados para Santander de acuerdo al coeficiente de correlación para el periodo 2008-2009 (Santos y col., 2016).	124
Tabla 4.10.	Clasificación de las variables independientes consideradas en los modelos desarrollados para Santander de acuerdo al coeficiente de correlación para el periodo 2008-2009 y 2011.	125
Tabla 4.11.	Comparativa de los resultados de los modelos PLSR y ANN desarrollados en Santander considerando PG_{am} como variable de entrada adicional bajo el enfoque B.	126



Capítulo 1. Planteamiento

1.1. Contaminación atmosférica por metales y metaloides

Los metales y los metaloides o semimetales se encuentran presentes de forma natural tanto en el medio ambiente como en organismos vegetales, animales y en el propio cuerpo humano puesto que algunos de ellos actúan como oligoelementos, resultando indispensables para el correcto funcionamiento del metabolismo de los organismos vivos. No obstante, tanto una carencia como un exceso de estos micronutrientes esenciales pueden resultar altamente perjudiciales para la salud. Asimismo, otros metales y semimetales no pueden ser metabolizados por el organismo, por lo que resultan tóxicos, suponiendo un riesgo para la salud. Aunque a menudo están presentes en el medio ambiente con baja abundancia, en grandes dosis resultan peligrosos. Las principales vías de exposición de los seres humanos a los metales se dan con frecuencia por ingestión (generalmente se trata de metales con efectos bioacumulativos que se transmiten a los humanos a través de la cadena trófica o por ingesta de aguas contaminadas) e inhalación (incorporación de metales al organismo a través de las vías respiratorias por su presencia en el aire ambiente), siendo esta última la que atañe a este trabajo ya que constituye la ruta principal de entrada al cuerpo.

En la atmósfera, los metales se encuentran principalmente vinculados al material particulado, que se define como una mezcla compleja de partículas sólidas y líquidas de pequeño tamaño en suspensión en la atmósfera, que presentan diferentes propiedades físicas, químicas y biológicas. El tamaño de las partículas que componen el material particulado determina su denominación, así como el grado de penetración en el cuerpo. En relación a esto, las partículas que presentan un diámetro aerodinámico igual o inferior a $10\ \mu\text{m}$ se conocen como PM_{10} y forman parte de la fracción inhalable y torácica del material particulado, porque pueden penetrar en el tracto respiratorio superior e incluso más allá de la laringe; mientras que las partículas de diámetro aerodinámico igual o inferior a $2,5\ \mu\text{m}$ se conocen como $\text{PM}_{2,5}$ y constituyen la fracción respirable, la de aquellas partículas inhaladas que pueden penetrar en las vías respiratorias no ciliadas. La composición del material particulado, así como la concentración relativa de los diferentes componentes en el mismo vienen condicionadas por las fuentes de emisión y por factores geográficos, topográficos y meteorológicos que determinan su estabilidad y que condicionan los procesos de dilución, dispersión y deposición. Por este motivo, los componentes del material particulado -y en el caso que nos ocupa, los metales- se caracterizan por presentar una gran variabilidad espacial y temporal (Querol y col., 2012).

1.1.1. Fuentes atmosféricas

Como ocurre con los demás contaminantes atmosféricos, las fuentes de emisión de metales a la atmósfera pueden tener un origen natural o antropogénico. En cuanto a las fuentes naturales, destacan la resuspensión del polvo de los desiertos, el mar y los océanos o los volcanes. En cuanto a las fuentes antropogénicas, destacan el tráfico y la industria (Arruti, 2011). En caso de la presencia de un bajo número de fuentes de emisión a escala local, existe una correlación entre ciertas actividades antropogénicas y los niveles de metales en la atmósfera. En caso de una mayor complejidad en el entramado industrial de una zona, es más difícil observar una relación directa entre fuentes de emisión específicas y aquellas sustancias o elementos que sirven como marcadores de las mismas debido a que muchas especies no proceden únicamente de una sola fuente de emisión, sino que con frecuencia se encuentran presentes por causa de las emisiones procedentes de distintas fuentes. Además, hay que tener en cuenta que los perfiles de emisión pueden cambiar con el tiempo como consecuencia de cambios en las tecnologías utilizadas (Querol y col., 2007). Por tanto, es preciso recurrir a técnicas de identificación de fuentes para determinar la contribución de cada tipo de fuente de emisión a los niveles de metales en inmisión. Los modelos de contribución de fuentes se aplican con el objetivo de relacionar los impactos de las emisiones de diversos contaminantes atmosféricos con los niveles de dichos contaminantes en el sitio o lugar receptor estudiado (Viana y col., 2008).

La aplicación de modelos de receptor a las bases de datos de especiación química de PM₁₀ y PM_{2.5} para la determinación de las contribuciones de las fuentes de emisión a los niveles de sus componentes en aire ambiente permite identificar las principales fuentes. Así, en cuanto al contenido de metales y semimetales en el material particulado, en la literatura se encuentran disponibles distintos estudios que describen trazadores de fuentes específicas (Pacyna, 1986; Pacyna, 1987; Maenhaut, 2001; Sternbeck y col., 2002; Adachi y Tainosho, 2004; Win Lee y col., 2004; Arditsoglou y Samara, 2005; Uexküll y col., 2005; Wåhlin y col., 2006). La mayoría de los estudios se centran en emisiones de tráfico, mientras que otros presentan trazadores de emisiones industriales, principalmente de procesos de combustión.

En cuanto al tráfico, se distinguen principalmente las siguiente fuentes de emisión: los catalizadores, fuente de Pd, Pt y Rh (Bocca y col., 2006; Grieshop y col., 2006), el desgaste de los frenos, fuente de Ba, Ca, Cu, Fe, Sb, Sr, Ti y Zn (Sternbeck y col., 2002; Grieshop y col., 2006; Querol y col., 2007; Dongarrá y col., 2009), el polvo de la carretera, fuente de Al, Ca, Fe y Mg (Grieshop y col., 2006) y el desgaste de los neumáticos, fuente de Ba y Zn (Grieshop y col., 2006; Querol y col., 2007).

En lo que respecta a las emisiones industriales, la comparación de la contribución entre zonas es difícil, ya que en muchas zonas con bajo desarrollo industrial la contribución de la industria

se encuentra mezclada con los aportes de aerosoles secundarios de origen regional, mientras que en zonas industriales es posible identificar la contribución industrial de forma independiente (Querol y col., 2012). A pesar de esta dificultad, en diversos estudios se han identificado, entre otras, las siguientes fuentes de carácter industrial con trazadores característicos: metalurgia del cobre, fuente de As, Bi, Cu y Ga (Morawska y Zhang, 2002; Querol y col., 2007); metalurgia del zinc, fuente de Cd y Zn (Querol y col., 2007); producción de acero, fuente de Cr, Cd, Mn, Mo, Ni, Pb, Se, Sn y Zn (Querol y col., 2007); producción de acero inoxidable, fuente de Cr, Mo, Mn y Ni (Querol y col., 2007); producción de aleaciones de ferromanganeso, fuente de Mn (Fernández-Espinosa y col., 2001; Morawska y Zhang, 2002; Boudissa y col., 2006); industria petroquímica, fuente de Ni y V (Morawska y Zhang, 2002; Bosco y col., 2005; Querol y col., 2007); procesos de combustión de aceites, fuente de Cu, Cr, Ni y V (Morawska y Zhang, 2002; Almeida y col., 2005; Viana y col., 2008); y procesos de combustión de carbón, fuente de Al, As, Co, Cu, Cr y Se (Morawska y Zhang, 2002).

A continuación se abordan con mayor profundidad las fuentes de los cuatro elementos regulados por la legislación europea y que son objeto de estudio en esta Tesis Doctoral: As, Cd, Ni y Pb. La mayor parte de las emisiones de As proceden de los procesos de fundición para la producción de metales no ferrosos como Cu (Pacyna, 1987; Hedberg y col., 2005; Thurston y col., 2011; Fernández-Camacho y col., 2012; González-Castanedo y col., 2015), Pb y Zn, así como de los procesos de combustión de carbón (Pacyna, 1987; Keegan y col., 2006; Thurston y col., 2011) y de petróleo. Los pesticidas solían ser una fuente considerable de As, pero las restricciones establecidas en varios países han reducido su contribución. Las fuentes antropogénicas de Cd incluyen también la producción de metales no ferrosos (Pacyna y col., 1987; Pacyna y col., 2009) como Zn (Batonneau y col., 2004; Shao y col., 2013) o Al (Kuo y col., 2007), la quema estacionaria de combustibles fósiles, la incineración de residuos (Gao y col., 2002; Wang y col., 2006), la producción de hierro y acero (Dai y col., 2015) y la producción de cemento. La combustión de gasolina constituyó la principal fuente de emisiones atmosféricas de Pb en Europa (Pacyna y Pacyna, 2001; Pacyna y col., 2007) hasta la entrada en vigor de la Directiva 98/70/CE (EC, 1998) relativa a la calidad de la gasolina y el gasóleo y sobre el uso de combustible sin plomo en los países europeos desde el 1 de enero de 2000, momento a partir del cual las emisiones de Pb se redujeron drásticamente. En la actualidad, la industria es la fuente con mayor influencia sobre los niveles de Pb, representando un 46% de las emisiones totales de Pb en 2012 (EEA, 2014). En este sentido, las fuentes de emisión de Pb incluyen los procesos de combustión, la incineración de residuos y la producción de metales no ferrosos, hierro, acero y cemento. Hay varias fuentes de emisión de Ni a la atmósfera; entre ellas: la combustión para calefacción residencial (Celo y Dabek-Zlotorzynska, 2010; Vuković y col., 2015), tráfico marítimo o generación de energía (Becagli y col., 2012), explotación minera y producción

de Ni, incineración de residuos y lodos de aguas residuales, producción de acero, galvanoplastia y procesos de combustión (López y col., 2005; EEA, 2014).

1.1.2. Niveles en aire ambiente

La composición química del material particulado varía ampliamente en función de las principales fuentes de emisión y de las reacciones químicas que tienen lugar en la atmósfera (Viana y col., 2006). En consecuencia, la magnitud de los niveles en aire ambiente de dichos componentes, entre los que se encuentran los metales y semimetales, presenta una gran variabilidad dependiendo del lugar en el que se realice el estudio para su determinación.

En la Tabla 1.1. se recogen los niveles medios de concentración en aire ambiente de algunos metales y semimetales medidos en diversas zonas urbanas en diferentes periodos de tiempo. Estos valores corresponden a los principales estudios de caracterización de la composición de material particulado llevados a cabo tanto a nivel nacional como europeo.

Por su alcance y envergadura, el principal estudio de ámbito nacional en relación a los niveles de concentración en aire ambiente de metales y semimetales es el llevado a cabo por Querol y col. (2008), en el que se determinaron, en un periodo de 6 años (1999-2005), los niveles de elementos traza en las fracciones PM₁₀ y PM_{2.5} del material particulado en 21 estaciones de control de la calidad del aire distribuidas por toda España. En los entornos con menor grado de contaminación los niveles de concentración de la mayoría de los elementos se encontraban dentro del intervalo 0,1-10 ng/m³, y únicamente Zr, Mo, Ni, V, Ti, Ba, Cu, Pb, Zn (en orden creciente) superaban el valor de 1 ng/m³. Los niveles de los elementos traza aumentaban según el grado de contaminación de origen antropogénico, hasta llegar a registrarse en algunas ocasiones niveles de Ti, Cr, Mn, Cu, Zn, As, Sn, W, Mn y Pb 10 veces superiores a los registrados en estaciones de fondo rural (Querol y col., 2012).

Como se observa en la Tabla 1.1. y de acuerdo con la Agencia Europea de Medio Ambiente, AEMA, (EEA, 2014), los niveles de concentración en aire ambiente en Europa de los contaminantes en los que se centra principalmente este trabajo, es decir, As, Cd, Ni y Pb, no son elevados. En general, presentan pocas superaciones de los respectivos valores límite/objetivo establecidos por la legislación europea correspondiente (se describen con más detalle en el apartado 1.2.) e, incluso, a menudo se encuentran por debajo del umbral inferior de evaluación. Aún más, en el período 2003-2012 se registró una disminución en las emisiones de los 28 estados que conforman la Unión Europea (UE) y de los 33 estados miembros de la AEMA como sigue: As en un 9%, Cd en un 27-26%, Ni en un 44% y Pb en un 19% (EEA, 2014). No obstante, también se ha de notar que la presencia de estos contaminantes en la atmósfera no carece de importancia pues contribuye, tras su deposición, al aumento de los niveles de metales pesados en suelos, sedimentos y organismos.

Tabla 1.1. Concentraciones medias (ng/m^3) de algunos metales y semimetales en PM_{10} de los principales estudios a nivel regional, nacional y europeo reportados en la bibliografía. Tabla adaptada de Fernández-Olmo y col., 2016.

Ámbito	País	Área de estudio	Estación ^a	Periodo	Contaminante											Referencia			
					As	Cd	Cr	Cu	Fe	Mn	Mo	Ni	Pb	Ti	V		Zn		
Regional	España	Barreda	T/I	2008-2010	1,11	0,2	9,48	19,6	601	17,6	1,2	3,37	10	5,7	1,98	306	Fernández-Olmo y col. (2016)		
			I	2008-2009	0,16	0,18	6,91	9,38	426	139	0,51	4,32	10,5	5,05	2,28	85,4	Fernández-Olmo y col. (2016)		
		Los Corrales de Buelna	U/I	2008-2010	0,38	0,17	5,08	7,29	574	25,2	0,65	3,7	14	6,55	1,58	226	Fernández-Olmo y col. (2016)		
			U	2008	0,2	0,1	2,3	3,5	-	9,3	0,5	3,0	8,0	5,2	3,1	-	Arruti y col. (2011a)		
		Reinosa	U	2008	0,3	0,2	5,8	3,4	-	9,5	2,1	2,2	11,2	4,1	1,5	-	Arruti y col. (2011a)		
			U	2008-2009	0,4	0,3	5	4,8	-	41,1	0,4	1,2	6,5	2,5	1,2	-	Arruti y col. (2011a)		
		Torrelavega	U/I	2007-2008	0,6	0,3	3,8	46,9	400	28,7	3,2	1,5	19,9	15,8	2,6	1,48	Moreno y col. (2011)		
			I	2010	0,63	0,81	3,09	10	477	1071	1,2	5,01	28,6	-	1,85	341	CIMA (2010)		
		Nacional	España	Huelva	I	2008-2009	6,2	0,7	2,3	45,3	600	9,6	12,2	3,7	14,4	46,5	5,3	47,4	Fernández-Camacho y col. (2012)
					U	2001	0,7	0,3	2	31	-	10	3	2	24	34	4	84	Querol y col. (2008)
Alcobendas	U/I			2002-2006	5,0	1,4	5	5	-	6	1,5	3	156	54	4	196	Querol y col. (2008)		
	U			2004	0,8	0,2	4	6	-	12	7,9	2	7	37	4	24	Querol y col. (2008)		
Barcelona	U/I			2002-2006	1,0	0,4	5	81	-	17	2,5	6	25	34	13	106	Querol y col. (2008)		
	U			2001	0,3	0,2	2	23	-	11	2	4	15	52	8	14	Querol y col. (2008)		
Puertollano	U/I			2004	1,9	0,1	4	27	-	11	3	4	12	62	10	54	Querol y col. (2008)		
	I			2001	1,8	1,2	24,6	32,6	-	86,5	-	33	102	24,2	8,3	417	Moreno y col. (2006)		
Tarragona	U/I			2001	0,8	0,3	2,9	32,9	-	9,2	4,2	25,5	22,5	7,7	35	35	Moreno y col. (2006)		
	U			2001-2002	-	-	7,7	22,8	666	24,7	0,8	18,7	-	6,6	212	-	López y col. (2005)		
Europeo	Europa	Budapest	U	2008-2011	0,65	0,26	-	-	-	-	-	0,70	9,81	-	-	-	Szabó y col. (2015)		
			U	2008-2012	0,64	0,6	-	-	-	-	-	3,73	14,47	-	-	-	Szabó y col. (2015)		
		Copenhague	U	2008-2011	-	-	-	15,1	361,8	-	-	1,0	-	-	2,7	19,9	Tsai y col. (2015)		
			U	2008-2011	-	-	-	7,3	363,1	-	-	1,6	-	-	3,9	22,9	Tsai y col. (2015)		
		París	U	2008-2011	-	-	-	33,8	747,2	-	-	1,3	-	-	1,8	40,2	Tsai y col. (2015)		
			U	2008-2011	-	-	-	42,3	1214,0	-	-	4,7	-	-	7,6	68,3	Tsai y col. (2015)		
		Turín	U	2008-2011	-	-	-	53,5	1343,4	-	-	5,7	-	-	3,7	61,8	Tsai y col. (2015)		
			U	2008-2011	-	-	-	8,1	357,2	-	-	0,3	-	-	1,6	58,0	Tsai y col. (2015)		
		Oslo	U	2008-2011	-	-	-	15,8	552,0	-	-	1,2	-	-	2,4	22,4	Tsai y col. (2015)		
			U	2008-2011	-	-	-	10,4	306,1	-	-	0,9	-	-	1,6	18,0	Tsai y col. (2015)		
Londres	U	2008-2011	-	-	-	18,7	520,6	-	-	1,0	-	-	2,5	23,5	Tsai y col. (2015)				

a): Industrial; T: Tráfico; U: Urbana

Tabla 1.1. Concentraciones medias (ng/m^3) de algunos metales y semimetales en PM_{10} de los principales estudios a nivel regional, nacional y europeo reportados en la bibliografía. Tabla adaptada de Fernández-Olmo y col., 2016 (Continuación).

Ámbito	País	Área de estudio	Estación ^a	Periodo	Contaminante											Referencia		
					As	Cd	Cr	Cu	Fe	Mh	Mo	Ni	Pb	Ti	V		Zn	
	Suecia	Estocolmo	U	2008-2011	-	-	-	20,1	816,9	-	-	-	1,7	-	-	2,4	24,5	Tsai y col. (2015)
	Suiza	Lugano	U	2008-2011	-	-	-	28,0	741,1	-	-	1,4	-	-	0,9	35,4	Tsai y col. (2015)	
	Italia	Roma	U	2010-2011	2,28	-	-	85,3	-	57,1	-	-	-	177	27,7	-	Avino y col. (2014)	
	Portugal	Lisboa	U	2001	0,47	-	-	11,1	400	6,5	-	4,2	14,6	29,5	11,2	36	Almeida y col. (2013)	
Europeo	Italia	Venecia	U	2007-2010	2,0	2,2	5,2	9,8	357,0	-	-	3,9	21,3	-	4,4	57,4	Contini y col. (2012)	
	Francia	Dunkerque	I	2005-2006	2,49	1,27	4,99	15,6	1752	99	2,14	9,43	34	17	16	131	Gaudry y col. (2008)	
	Italia	Palermo	T	2005	1,4	-	9,3	83	827	18	7	8	17	-	22	60	Dongarrà y col. (2007)	
	Grecia	Atenas	I	2001-2002	6,1	3,7	14,5	43,2	-	21,1	-	15,9	71,1	-	8,6	-	Manalis y col., (2005)	
	Reino Unido	Dundee	U	1999-2000	-	-	-	24	-	-	-	25	21	-	-	28	Qin y Oduyemi (2003)	

^aI: Industrial; T: Tráfico; U: Urbana

1.1.3. Efectos sobre la salud

Un gran número de trabajos ponen de manifiesto los impactos de la contaminación atmosférica sobre la salud humana. En especial, hay un claro surgimiento de indicios que relacionan la inhalación de material particulado, que contiene metales, con efectos adversos sobre la salud humana a concentraciones cercanas (o no mucho mayores) a las de los actuales niveles ambientales. En este sentido, [Chen y Lippmann \(2009\)](#) realizaron una revisión de los principales efectos sobre la salud humana de los metales asociados al material particulado presente en aire ambiente. No obstante, al tratarse de una mezcla compleja de sustancias, hay cierta incertidumbre a la hora de atribuir un efecto específico a un constituyente en concreto.

Por tanto, la mayoría de indicios en relación a los efectos biológicos de los metales sobre los seres humanos proceden tanto de estudios derivados de exposiciones ocupacionales a dichas sustancias de trabajadores durante su actividad laboral, como de exposiciones en laboratorio de animales (*in vivo*) o de células (*in vitro*). Así, varios estudios han determinado que la acumulación de metales pesados en el cuerpo, por inhalación o ingestión, puede ser responsable de un amplio espectro de efectos de diversa índole sobre la salud humana –entre los más comunes: problemas respiratorios, efectos neurotóxicos, alteraciones del sistema inmunitario, patologías cardíacas e, incluso, cáncer– que conducen a un incremento de la tasa de mortalidad/morbilidad en la población ([Dockery y col., 1993](#); [Pope y col., 1995](#); [Silbergeld, 1995](#); [Silbergeld, 1996](#); [USEPA, 1996](#); [Vijayanand y col., 2008](#)).

Los siguientes apartados recogen una breve descripción de los principales efectos que normalmente se atribuyen a cada uno de los metales objeto de estudio en esta Tesis Doctoral.

1.1.3.1. Arsénico

El arsénico inorgánico es identificado como agente carcinogénico por diversas instituciones internacionales. La exposición a largo plazo por inhalación de bajos niveles de arsénico, aun no siendo la vía principal de exposición de la mayoría de la población, está asociada con un aumento del riesgo de padecimiento de cáncer de piel y de pulmón ([ATSDR, 2007a](#); [EEA, 2014](#)). Además, también puede tener efectos cutáneos adversos y causar trastornos circulatorios y nerviosos menores ([ATSDR, 2007a](#)).

1.1.3.2. Cadmio

El cadmio y los compuestos de cadmio son igualmente considerados carcinógenos, siendo capaces de causar cáncer de pulmón. Por otra parte, una exposición a largo plazo también está asociada con la aparición de patologías renales y pulmonares, así como con el desarrollo de fragilidad ósea ([ATSDR, 2012a](#); [EEA, 2014](#)). La inhalación constituye una pequeña parte de la exposición total a cadmio; no obstante, los niveles en aire ambiente tienen importancia por su

consiguiente deposición sobre el suelo y, en consecuencia, por su posible ingesta a través de la dieta, vía más común de exposición para la mayor parte de la población (ATSDR, 2012a).

1.1.3.3. *Níquel*

En cuanto al níquel, su inhalación se asocia principalmente con trastornos cutáneos, como la dermatitis por contacto con níquel (ATSDR, 2005), y alteraciones del sistema endocrino (EEA, 2014). Para exposiciones a niveles más elevados, normalmente ocupacionales, destacan los efectos respiratorios como inflamación pulmonar, bronquitis crónica, daños sobre el epitelio olfatorio, etc., e incluso cáncer; no en vano, varios compuestos de níquel han sido clasificados como carcinógenos para los seres humanos (ATSDR, 2005; EEA, 2014).

1.1.3.4. *Plomo*

El plomo, puede afectar a casi cualquier órgano o sistema, en especial a los sistemas nervioso y cardiovascular. Igualmente, puede tener efectos negativos sobre el desarrollo cognitivo y neuroconductual en niños y puede producir un incremento en la presión sanguínea y anemia en adultos (EEA, 2014). La exposición a niveles elevados puede causar daños neurológicos y renales, puede provocar abortos en mujeres embarazadas y dañar los órganos responsables de la producción de esperma en hombres (ATSDR, 2007b; Martin y Griswold, 2009). Además, si bien no hay pruebas concluyentes que indiquen que produce cáncer, diversas instituciones apuntan a que el plomo inorgánico es sospechoso de tener efectos carcinogénicos en los humanos (ATSDR, 2007b).

1.1.3.5. *Cromo*

Mientras que el cromo (III) es un nutriente esencial, el cromo (VI) es tóxico y un reconocido agente carcinogénico. La inhalación de altos niveles de cromo también puede causar irritación del revestimiento epitelial de la nariz, úlceras nasales o problemas respiratorios como asma, tos, rinorrea, dificultad para respirar o sibilancias. Una exposición a largo plazo puede causar daños hepáticos, renales, circulatorios, daños a los tejidos nerviosos e irritación cutánea (Martin y Griswold, 2009; ATSDR, 2012b).

1.1.3.6. *Manganeso*

Se ha demostrado que la inhalación de manganeso, incluso en bajos niveles de concentración (cerca de 50 ng/m^3), produce efectos neurotóxicos que varían desde alteraciones neuropsicológicas y de las funciones motoras (Mergler y col., 1999; ATSDR, 2012c), alteraciones en la estabilidad postural (Hernández-Bonilla y col., 2011) y un aumento del riesgo de padecimiento de la enfermedad de Parkinson (Finkelstein y Jerret, 2007; ATSDR, 2012c). A concentraciones superiores a 1 mg/m^3 puede producir manganismo, un trastorno de las funciones motoras similar a la enfermedad de Parkinson (Aschner y col., 2005).

1.1.3.7. Vanadio

Con respecto al vanadio, sus efectos tóxicos por inhalación dependen de su grado de oxidación y pueden incluir irritación del tracto respiratorio, cambios hematológicos y bioquímicos y lesiones funcionales en ciertos órganos (ATSDR, 2012d; Ghosh y col., 2015).

1.2. Regulación de los niveles de metales y metaloides en aire ambiente

El principal objetivo de los estándares y valores guía de calidad del aire es la protección de la salud humana. Los valores guía de referencia para contaminantes atmosféricos empleados con mayor frecuencia son aquellos que proceden de la Organización Mundial de la Salud (OMS), de la UE y de la Agencia de Protección Ambiental de los Estados Unidos (USEPA, por sus siglas en inglés). Estos valores guía se basan en pruebas clínicas, toxicológicas y epidemiológicas y se establecen mediante la determinación de aquellas concentraciones con las que se observan los mínimos efectos adversos sobre la salud humana –implícitamente, se acepta que existe un umbral inferior por debajo del cual no se pueden detectar dichos efectos-, aplicando un factor de seguridad para así tener en cuenta posibles incertidumbres en la extrapolación de los resultados de las pruebas de animales a humanos y de pequeños grupos de gente a grandes poblaciones. Los valores guía de la OMS se basan únicamente en consideraciones en materia de salud, mientras que los estándares de la UE y de la USEPA tienen en cuenta también la viabilidad técnica para cumplir dichos estándares.

Los principales instrumentos normativos sobre contaminación atmosférica dentro de la UE, diseñados para evitar, prevenir o reducir los efectos de dichos contaminantes sobre la salud humana y el medio ambiente, incluyen las Directivas de Calidad del Aire y los Techos Nacionales de Emisión. Las primeras, regulan las concentraciones ambientales de los contaminantes atmosféricos más importantes implementado una serie de valores objetivo o límite, mientras que los segundos establecen límites en las emisiones de alguno de estos contaminantes a la atmósfera de carácter antropogénico.

En este sentido, la normativa europea vigente en materia de calidad del aire viene establecida por la Directiva Marco de Calidad del Aire, Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa (EC, 2008). Por medio de dicha directiva, que sustituye a la Directiva 96/62/CE, anterior Directiva Marco de Calidad del Aire, así como a las tres Directivas Hijas (Directiva 1999/30/CE, Directiva 2000/69/CE y Directiva 2002/3/CE), se amplía el número de sustancias

bajo regulación, introduciendo otros contaminantes como el PM_{2,5} y, además, se establecen nuevos requerimientos en lo referente a la evaluación, gestión y objetivos de la calidad del aire, teniendo en cuenta además las directrices procedentes de la OMS. La Directiva 2008/50/CE establece un valor límite para el Pb, es decir, un nivel fijado con el fin de evitar, prevenir o reducir los efectos nocivos para la salud humana y el medio ambiente, y que no debe superarse una vez alcanzado. No obstante, la nueva Directiva Marco no sustituye a la Directiva 2004/107/CE del Parlamento Europeo y del Consejo, de 15 de diciembre de 2004, relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente (EC, 2004), a la que también se hace referencia como cuarta Directiva Hija, que permanece en vigor. Esta Directiva Hija establece valores objetivo para el As, Cd, Ni y el benzo(a)pireno en la fracción PM₁₀ del material particulado. Estos valores objetivo se definen como las concentraciones en aire ambiente fijadas para evitar, prevenir o reducir los efectos perjudiciales de los contaminantes en la salud humana y en el medio ambiente en su conjunto, que deben alcanzarse en lo posible durante un determinado período de tiempo. Por otra parte, aún sin establecer un valor objetivo específico, propone una serie de hidrocarburos aromáticos policíclicos relevantes cuyos niveles se deben controlar: benzo(a)antraceno, benzo(b)fluoranteno, dibenzo(a,h)antraceno, benzo(j)fluoranteno, benzo(k)fluoranteno e indeno(1,2,3-c,d)pireno. La Tabla 1.2. recoge los objetivos de calidad establecidos por la UE para As, Cd, Ni y Pb en aire ambiente.

Tabla 1.2. Objetivos de calidad para As, Cd, Ni y Pb en la regulación europea.

Contaminante	Valor límite u objetivo (ng/m ³)	Promedio	Directiva reguladora
As	6	Anual	2004/107/CE
Cd	5	Anual	2004/107/CE
Ni	20	Anual	2004/107/CE
Pb	500	Anual	2008/50/CE

La trasposición al ordenamiento jurídico español de la regulación europea en materia de calidad del aire, Directiva 2008/50/CE y Directiva 2004/107/CE, se llevó a cabo por medio del Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire. Este real decreto ha sido modificado por el Real Decreto 678/2014.

Otros metales y metaloides, como Mn, V y Cr, carecen de regulación específica en materia de calidad del aire en el ámbito europeo. En ausencia de esta normativa, a la hora de evaluar la calidad del aire con respecto a dichas sustancias, únicamente se cuenta con los valores guía propuestos por la OMS, que se recogen en la Tabla 1.3.

Tabla 1.3. Valores guía de la calidad del aire de la OMS para los metales y semimetales considerados en este trabajo.

Contaminante	Valor guía de la calidad del aire (ng/m ³)	Promedio
As	-	-
Cd	5 ^(a)	Anual
Ni	-	-
Pb	500	Anual
Cr	1000	Diario
Mn	150	Anual
V	1000	Diario

^(a) Valor guía de la calidad del aire para prevenir incremento de cadmio en suelo agrícola, que pueda aumentar la ingesta de dicho contaminante a través de la alimentación en futuras generaciones.

1.3. Niveles de metales y metaloides en aire ambiente en Cantabria en el marco de la regulación europea

Además de valores límite/objetivo para los contaminantes atmosféricos, la Directiva europea de Calidad del Aire también establece los métodos para la evaluación de la calidad del aire con respecto a dichos contaminantes en función de los niveles de concentración de los mismos. Para ello, se establecen dos umbrales de evaluación, umbral inferior de evaluación o UIE (LAT, por sus siglas en inglés) y umbral superior de evaluación o USE (UAT, por sus siglas en inglés), expresados como un porcentaje del valor límite/objetivo (VL/VO), de tal manera que dependiendo de la posición relativa de los niveles de concentración de los contaminantes con respecto a estos umbrales se permiten diferentes métodos de evaluación de la calidad del aire. Así, como se muestra en la Figura 1.1., en aquellas zonas o aglomeraciones en las que los niveles de contaminantes se encuentran por encima del USE, la evaluación de la calidad del aire se ha de efectuar obligatoriamente mediante medidas experimentales fijas, que se podrán complementar con técnicas de modelización y/o medidas indicativas; en caso de que los niveles se encuentren entre el USE y el UIE, se puede utilizar una combinación de mediciones fijas y técnicas de modelización y/o mediciones indicativas; y por debajo del UIE, las técnicas de modelización o de estimación objetiva o ambas son suficientes para evaluar la calidad del aire ambiente. La aplicación de las técnicas de modelización y de estimación objetiva es de especial interés para la evaluación de la calidad del aire con respecto a algunos microcontaminantes como los metales y metaloides, que a diferencia de otros contaminantes no cuentan con sensores e instrumentos de medida automáticos, de tal manera que su aplicación podría ayudar a la reducción de las necesidades de determinación analítica de sus niveles para verificar el

cumplimiento con los estándares de calidad del aire de la UE, con el consecuente ahorro de recursos económicos y de tiempo que ello conllevaría.

Los UIE y USE establecidos por la UE para As, Cd, Ni y Pb se recogen en la Tabla 1.4.

Tabla 1.4. Umbrales de evaluación de la calidad del aire con respecto a As, Cd, Ni y Pb según las Directivas europeas.

Contaminante	Valor límite u objetivo (ng/m ³)	UIE (%)/(ng/m ³)	USE (%)/(ng/m ³)
As	6	40/2,4	60/3,6
Cd	5	40/2	60/3
Ni	20	50/10	70/14
Pb	500	50/250	70/350

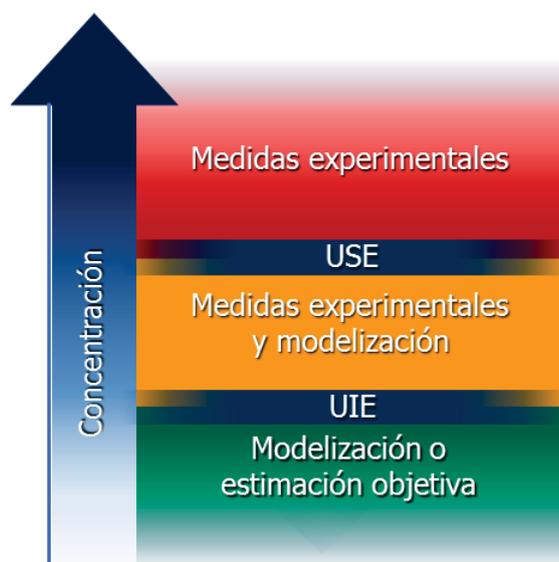


Figura 1.1. Métodos de evaluación de la calidad del aire según la Unión Europea.

En la Figura 1.2. se muestran los niveles de As, Cd, Ni y Pb en tres zonas urbanas de Cantabria (véase apartado 3.1.) en los años 2008, 2009 y 2011 en comparación con los correspondientes valores límite/objetivo y con los umbrales de evaluación. Se puede observar que, a excepción del valor medio de arsénico en 2011 en un punto de muestreo, los niveles de As, Cd, Ni y Pb se encuentran muy por debajo del umbral inferior de evaluación y, por tanto, de acuerdo con la Directiva europea de Calidad del Aire, en estas zonas sería posible efectuar la evaluación de la calidad del aire con respecto a dichos contaminantes únicamente mediante técnicas de modelización y/o estimación objetiva.

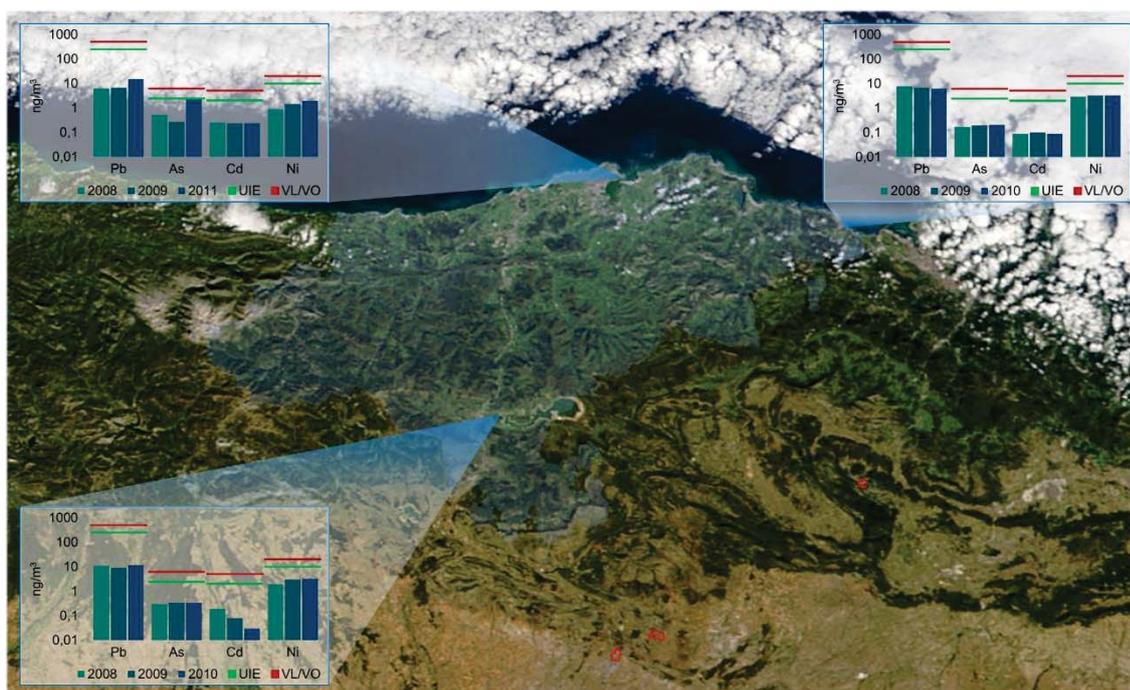


Figura 1.2. Niveles de As, Cd, Ni y Pb en tres zonas urbanas de Cantabria.

1.4. Modelización de la calidad del aire

Un modelo es una descripción de un sistema concreto expresada en términos matemáticos que, si bien no considera todas las características del sistema real, sí toma en consideración los aspectos y parámetros de interés fundamentales para la resolución del problema objeto de estudio (El-Harbawi, 2013).

Los modelos se utilizan ampliamente en campos de muy diversa índole para la resolución de problemas, para la generación de predicciones, así como para determinar las mejores alternativas para la gestión de problemas medioambientales (El-Harbawi, 2013).

En lo referente a la calidad del aire, los modelos constituyen el medio para llevar a cabo una estimación en base a medidas experimentales de las concentraciones de sustancias contaminantes en la atmósfera a través del estudio de las interacciones de dichos contaminantes con el medio, así como de la actuación de los procesos meteorológicos en la dispersión de los mismos (USEPA, 1996; Seinfeld y Pandis, 2006).

Una primera clasificación de modelos de la calidad del aire se establece distinguiendo entre modelos físicos y matemáticos. Los primeros consisten en una representación de la realidad a escala reducida, suponiendo unas condiciones meteorológicas determinadas, mientras que los segundos consisten en una descripción de un sistema mediante relaciones matemáticas

(Seinfeld y Pandis, 2006). Para llevar a cabo el modelado físico se construyen maquetas de los edificios y del terreno objeto de estudio y se sumergen en un fluido en movimiento. Con ayuda de un trazador que sea visible y tenga un comportamiento similar al fluido, se estudia la dispersión de éste al ser liberado desde diferentes alturas o ubicaciones. Asimismo, se observa la pluma desde que el fluido es emitido por la chimenea; de esta manera se comprueba cómo disminuye la intensidad de ésta a medida que disminuye la concentración. Las localizaciones de los edificios, su forma y los parámetros del fluido son algunos de los factores que determinan las relaciones empíricas del estudio. Una vez concluido el estudio se puede aplicar para establecer la mejor ubicación posible de las fuentes y la altura máxima de estas (Griffin, 2007). En cambio, los modelos matemáticos permiten realizar varios escenarios tipo bajo diferentes condiciones atmosféricas y a diferentes alturas. Por tanto, el método convencional de diseño de modelos físicos, que consiste en la réplica de un sistema real, es un proceso engorroso y que requiere mucho tiempo. Por otra parte, los modelos matemáticos son capaces de describir razonablemente bien los procesos o sistemas reales mediante el mero planteamiento de un sistema de ecuaciones matemáticas. La principal ventaja de los modelos matemáticos sobre los modelos físicos radica, en consecuencia, en su menor coste, puesto que los últimos requieren de una inversión inicial asociada a la construcción a escala reducida del escenario de estudio. Por esta razón, se espera que la aplicación de los modelos matemáticos continúe aumentando en el futuro (Griffin, 2007).

Los modelos matemáticos se clasifican a su vez en dos grandes grupos diferenciados: modelos determinísticos y modelos estadísticos. Los modelos determinísticos están basados en relaciones causa-efecto a través de descripciones matemáticas de los procesos atmosféricos que controlan la dispersión y/o transformación de los contaminantes en la atmósfera desde las fuentes de emisión hasta el lugar receptor. Por su parte, los modelos estadísticos se basan en relaciones estadísticas semiempíricas derivadas de los datos y medidas experimentales disponibles. Esta clasificación se describe con más detalle en los apartados 1.4.1 y 1.4.2.

1.4.1. Modelos determinísticos

Los modelos determinísticos proporcionan estimaciones de las concentraciones de las sustancias contaminantes en una ubicación y momento determinados. Dichas estimaciones se realizan a partir de inventarios de emisiones y variables meteorológicas (véase Figura 1.3.) –que incluyen velocidad y dirección del viento, temperatura y altura de la capa de mezcla– en base a la resolución de varias ecuaciones que representan los procesos físicos relevantes, como la ecuación de difusión-reacción, que es en esencia el principio de conservación de la masa para especies contaminantes. Se basan, por tanto, en el establecimiento de relaciones matemáticas entre las emisiones y la calidad del aire, incorporando una descripción del transporte, dispersión y transformación de los compuestos emitidos a la atmósfera.

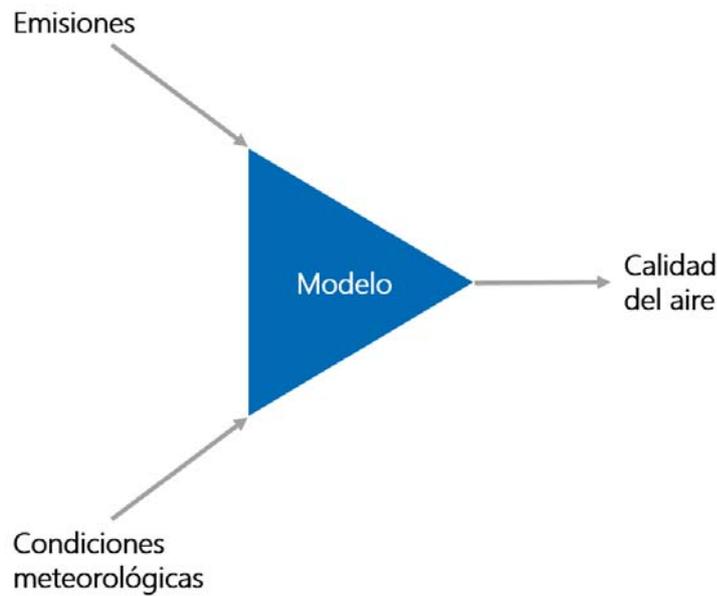


Figura 1.3. Esquema de modelo de calidad del aire.

Existen numerosos modelos determinísticos de dispersión de contaminantes en la atmósfera. No obstante, desde el punto de vista de la aproximación al problema, es decir, en función de la base algorítmica empleada para la descripción del comportamiento de los contaminantes en la atmósfera, se establece la siguiente clasificación fundamental de los modelos más utilizados (Colls, 2002):

- Modelos de caja, que asumen una distribución homogénea de los contaminantes en todo el volumen de una caja tridimensional y constituyen la alternativa más sencilla pero también la más limitada.
- Modelos gaussianos, que asumiendo condiciones meteorológicas estacionarias y turbulencia homogénea, establecen que la distribución de la concentración de contaminantes se ajusta a una función de distribución normal (gausiana).
- Modelos lagrangianos, que estudian los procesos que sufre un elemento de volumen o parcela de aire (también llamadas partículas) durante su desplazamiento en la atmósfera, cambiando su posición por acción de las corrientes atmosféricas. Utilizan un sistema de coordenadas que sigue el movimiento de la atmósfera,
- Modelos eulerianos, que describen el movimiento y las transformaciones que sufren los contaminantes en los elementos de volumen o parcelas (celdas) fijas en las que se divide la región a modelar. Utilizan, por tanto, un sistema de coordenadas fijo con respecto a la tierra, al contrario que la aproximación lagrangiana, que lo considera fijo con respecto a la partícula que se dispersa.

La implementación de estas aproximaciones fundamentales en sistemas de modelización ha dado lugar a la aparición de numerosos modelos determinísticos en todo el mundo, la mayoría de los cuales han sido revisados en profundidad por [El-Harwabi \(2013\)](#).

Los modelos determinísticos presentan claras ventajas frente a los modelos estadísticos, como la cobertura de un dominio tridimensional. No obstante, también presentan desventajas en relación a su complejidad y a la precisión e incertidumbre en cuanto a los datos de entrada al modelo. Según [Hanna \(1989\)](#), cuanto mayor sea el número de entradas a un modelo, menor es la incertidumbre y el error de las predicciones generadas por el mismo. Sin embargo, al incrementar el número de parámetros de entrada se incrementa el error y la incertidumbre asociada a éstos. En este sentido, conviene destacar la existencia de cierta incertidumbre asociada a los inventarios de emisiones, la cual se debe tanto al incumplimiento, por omisión o falta de exactitud, en las obligaciones legales de declaración de emisiones en el Registro Europeo de Emisiones y Transferencia de Contaminantes (E-PRTR) por parte de las actividades industriales, como a la estimación a través de factores de emisión de aquellas emisiones de contaminantes de las que no se dispone de toda la información necesaria, o al variado tipo de fuentes que originan dichas emisiones –industria (fuentes fijas), tráfico (fuentes móviles), fuentes puntuales, fuentes de área, fuentes naturales, etc.–, que redundan en una consecuente variabilidad de las mismas en tiempo y en espacio. Una forma de abordar este último problema, consiste en la realización de una desagregación espacial de los inventarios de emisiones. Sin embargo, existe también una incertidumbre inherente al propio método de desagregación empleado. Por consiguiente, dado que no es posible controlar cada una de las fuentes de emisión de manera individual, la generación de los inventarios de emisiones está siempre supeditada a la asunción de aproximaciones en la interpolación y extrapolación de los limitados conjuntos de datos y tienen, por tanto, una incertidumbre e imprecisión asociadas. En consecuencia, resulta difícil determinar con exactitud las tasas de emisión de los contaminantes, que constituye parte de la información de entrada necesaria para el desarrollo de modelos determinísticos. No obstante, en la actualidad se están realizando esfuerzos en cuanto a la definición de requisitos básicos de calidad de la información referente a las emisiones y su aplicación para garantizar la aceptabilidad de los resultados proporcionados por los modelos de dispersión en el contexto de la Directiva de calidad del aire ([Borge y col., 2011](#)).

En este sentido, se puede decir que los modelos determinísticos funcionan correctamente cuando se satisfacen los requerimientos de datos de entrada; sin embargo, esto raramente ocurre con las emisiones de contaminantes como el As, Cd, Ni y Pb. Por esta razón, el desempeño de modelos complejos para la estimación de los niveles de estos contaminantes se puede equiparar al de metodologías más simples. De ahí surge el interés por los modelos estadísticos para estimar las concentraciones en inmisión de microcontaminantes atmosféricos.

1.4.2. Modelos estadísticos

Los modelos estadísticos de calidad del aire se definen como técnicas basadas esencialmente en el análisis estadístico de los datos de concentraciones ambientales observadas. En concreto, se basan en el establecimiento de relaciones empíricas entre los datos históricos de calidad del aire y variables específicas de influencia, como las emisiones o las condiciones meteorológicas. Se trata, por tanto, de modelos no determinísticos, en el sentido de que no establecen ni simulan relaciones físicas causa-efecto entre emisiones y concentraciones ambientales (Daly y Zanetti, 2007). Se distinguen dos tipos: modelos de predicción de calidad del aire y sistemas de alarma, y modelos de receptor.

Los primeros se basan en técnicas estadísticas que se utilizan para predecir tendencias de contaminación con anticipación con el propósito de alertar a la población o, por ejemplo, bloquear el tráfico. En Gilbert (1987) se puede encontrar información general básica acerca de modelos estadísticos de calidad del aire. No obstante, Finzi y Nunnari (2005) han llevado a cabo una revisión más actualizada incluyendo las técnicas más recientes. Entre los métodos empleados en el desarrollo de modelos estadísticos de calidad del aire se destacan los siguientes:

- Análisis espectral. Permite la identificación de ciclos en series temporales de datos meteorológicos y de calidad del aire.
- Análisis de series temporales, el método clásico de Box-Jenkins. Se aplica para evaluar patrones de evolución en series temporales de medidas meteorológicas y de concentración de contaminantes.
- Distribuciones estadísticas de datos de concentración. El estudio de las propiedades estadísticas de la distribución de los valores medidos de concentración de contaminantes del aire permite inferir fácilmente la frecuencia de la ocurrencia de elevadas concentraciones esporádicas.
- Técnicas de regresión. Constituyen un tipo particular de análisis de series temporales múltiples, que consiste, por lo general, en el establecimiento de relaciones estadísticas entre medidas simultáneas de variables meteorológicas (variables predictoras) y de concentración de contaminantes en aire (variables respuesta) observados en un área y en un periodo de tiempo (pasado) determinados. Posteriormente, las ecuaciones de regresión obtenidas se aplican a los valores de las variables predictoras que se observan en un periodo de tiempo posterior para calcular las concentraciones aproximadas de contaminantes que resultan a partir de dichas condiciones meteorológicas.
- Modelos de receptor multivariables. Relacionan la composición química de aerosoles en un lugar determinado (receptor) con la de los aerosoles emitidos desde distintas fuentes de contaminación atmosférica. El objetivo de estos modelos consiste en utilizar

las medidas de concentración para predecir el número de fuentes emisoras que afectan al receptor y con qué fuentes está asociado un determinado contaminante.

- Técnicas de inteligencia artificial o *soft-computing*. En este grupo se engloban los modelos neuronales, los modelos de lógica difusa o los algoritmos evolutivos, entre otros. De forma general, se trata de técnicas inspiradas en los procedimientos de procesamiento de la información de los sistemas biológicos para la solución de problemas que se caracterizan por la necesidad de interactuar de manera eficiente con sistemas complejos, cuando la información es insuficiente, incierta o inexacta.

En base a la gran cantidad de estudios publicados en la literatura, los modelos estadísticos son objeto de interés creciente entre los investigadores de modelización de calidad del aire. Esto se debe no solo al incremento gradual de la disponibilidad de datos de calidad del aire en los lugares de interés, sino porque además, al contrario que los modelos determinísticos, requieren menor información de entrada (que en muchos casos cuenta con una gran incertidumbre asociada y no se encuentra disponible por redes de calidad del aire en tiempo real) o, alternativamente, introducen dicha información en el análisis de forma indirecta y más sencilla por medio de otra información relacionada más accesible. Por tanto, los modelos estadísticos representan un método sencillo y atractivo para su uso como alternativa o complemento a los modelos determinísticos.

1.5. Modelos estadísticos empleados en el modelado de contaminantes atmosféricos

Un amplio número de clasificaciones de los agentes químicos en la atmósfera es posible en función del criterio elegido para ello. Así, en función de su concentración, abundancia relativa o importancia cuantitativa en la atmósfera, se distinguen contaminantes mayoritarios y minoritarios. Los contaminantes atmosféricos mayoritarios son aquellas sustancias producidas y vertidas a la atmósfera en grandes cantidades y que son causantes de gran parte de la contaminación atmosférica, por lo que representan un peligro para la salud y el medioambiente. Las sustancias que componen el grupo de los contaminantes mayoritarios son fundamentalmente los óxidos de nitrógeno (NO_x) -principalmente, el monóxido de nitrógeno (NO) y el dióxido de nitrógeno (NO_2)-, el dióxido de azufre (SO_2), las partículas en suspensión, el monóxido de carbono (CO) y los hidrocarburos. Esta clasificación suele aplicarse a contaminantes primarios, es decir, a aquellos emitidos directamente desde las fuentes, por lo que el ozono troposférico, que se trata de un contaminante secundario, no suele ser incluido en esta categoría. Sin embargo, en numerosos estudios el ozono es igualmente considerado

contaminante mayoritario. Por el contrario, los contaminantes minoritarios son aquellos cuya presencia a nivel de trazas puede originar efectos contaminantes aun siendo emitidos a la atmósfera en pequeñas cantidades. Una gran cantidad de sustancias se clasifican dentro del grupo de contaminantes minoritarios; entre ellos: los metales y compuestos metálicos, el cloro molecular y el cloruro de hidrógeno, los compuestos fluorados, los compuestos reducidos de azufre, y compuestos orgánicos diversos, como hidrocarburos aromáticos policíclicos, dioxinas y furanos.

1.5.1. Modelos estadísticos para contaminantes mayoritarios

En la literatura se pueden encontrar numerosos estudios en relación a la estimación de la concentración en aire ambiente de contaminantes mayoritarios mediante modelos estadísticos. A continuación, se presenta una breve descripción de una selección de algunos de los trabajos más recientes y relevantes sobre la aplicación de modelos estadísticos para la estimación de las concentraciones en inmisión de contaminantes mayoritarios.

1.5.1.1. Óxidos de Nitrógeno

En lo que respecta a los óxidos de nitrógeno, [Gardner y Dorling \(1999\)](#) desarrollaron modelos basados en redes neuronales artificiales de tipo perceptrón multicapa (MLP, por sus siglas en inglés) para la predicción de concentraciones horarias de NO_x y NO_2 en Londres (Reino Unido) a partir únicamente de datos meteorológicos. Por su parte, [Kolehmaninen y col. \(2001\)](#) aplicaron igualmente redes neuronales de tipo MLP y de tipo mapa auto-organizados (SOM, por sus siglas en inglés) para la estimación de la concentración horaria de NO_2 en Estocolmo (Suecia) en el periodo 1994-1998.

1.5.1.2. Dióxido de azufre

[Chelani y col. \(2002a\)](#) emplearon redes neuronales artificiales para la predicción de concentraciones diarias de dióxido de azufre en Delhi (India) correspondientes al periodo 1997-1998. Asimismo, desarrollaron modelos de regresión multivariante y compararon su desempeño con el de las redes neuronales, concluyendo que estas últimas generan mejores predicciones presentando un error cuadrático medio más bajo. [Denby y col. \(2010\)](#) utilizaron la regresión lineal múltiple para confeccionar un mapa de calidad del aire con respecto a valores anuales de SO_2 en el periodo 1996-2005.

1.5.1.3. Monóxido de carbono

[Noori y col. \(2010\)](#) predijeron la concentración diaria de monóxido de carbono en la atmósfera de Teherán (Irán) mediante modelos de redes neuronales y un sistema adaptativo de inferencia neuro-difusa (ANFIS, por sus siglas en inglés). [Yeganeh y col. \(2012\)](#) desarrollaron un modelo híbrido de mínimos cuadrados parciales y máquinas de vectores de soporte (SVM, por sus siglas en inglés) para la predicción de concentraciones horarias de CO en Teherán (Irán).

1.5.1.4. Ozono Troposférico

Se han realizado diversos estudios empleando redes neuronales artificiales en relación a la predicción de niveles medios de concentración de ozono troposférico (Abdul-Wahab y Al-Alawi, 2002; Chaloulakou y col., 2003; Yazdanpanah y col., 2009; Inal, 2010), y niveles máximos diarios (Yi y Prybutok, 1996; Wang y col., 2003). Asimismo, se han llevado a cabo estudios sobre la comparación de redes neuronales con modelos de regresión lineal (Comrie, 1997; Demuzere y col., 2010), y con modelos de programación genética y series de tiempo (Khatibi y col., 2013) y con algoritmos genéticos y máquinas de vectores soporte (Feng y col. 2011). Por su parte, Gómez-Sanchis y col. (2006), llevaron un análisis de relevancia de variables de entrada a modelos de redes neuronales para la predicción de ozono y Pires y Martins (2011) propusieron métodos de corrección para mejorar el desempeño tanto de modelos de regresión lineal múltiple como de redes neuronales. Sousa y col. (2007) avanzaron en la estimación de las concentraciones en inmisión de ozono desarrollando redes neuronales basadas en componentes principales.

Por otra parte, Sun y col. (2013) llevaron a cabo un estudio para predecir la concentración media de ozono correspondiente a 8h mediante un modelo oculto de Markov combinado con técnicas lineales en San Francisco (Estados Unidos).

1.5.1.5. Partículas en suspensión

Goyal y col. (2006) emplearon tanto modelos de regresión lineal múltiple (MLR, por sus siglas en inglés) y modelos de series de tiempo de tipo auto-regresivo de media móvil (ARIMA, por sus siglas en inglés), como una combinación de ellos, para estimar las concentraciones medias diarias de material particulado respirable en suspensión (RSPM, por sus siglas en inglés) en zonas urbanas de Delhi (India) y Hong Kong (China). Los resultados mostraron que las estimaciones procedentes del modelo combinado se ajustan mejor a los datos observados.

Pires y col. (2008) comprobaron la capacidad de diferentes modelos lineales, incluyendo la regresión por mínimos cuadrados parciales (PLSR, por sus siglas en inglés), para la estimación de los niveles medios diarios de PM₁₀ en la ciudad de Oporto (Portugal). Si bien todos los modelos ajustaban los datos de forma similar, PLSR mostró una mejor capacidad de generalización que las demás técnicas. Por otra parte, Polat y Durduran (2012) emplearon modelos de regresión por mínimos cuadrados (LSR, por sus siglas en inglés), regresión por mínimos cuadrados parciales y regresión lineal múltiple para la estimación de las concentraciones medias diarias del material particulado en Iconio (Turquía). Singh y col. (2012) compararon la regresión por mínimos cuadrados parciales con técnicas no lineales de modelización para la predicción de RSPM, SO₂ y NO₂ en Lucknow (India). Si bien ambos tipos de técnicas, lineales y no lineales, proporcionaron estimaciones adecuadas (en especial para el

RSPM, con un coeficiente de correlación de hasta 0,9), las técnicas no lineales presentaron un desempeño ligeramente superior.

Según [Grivas y Chaloulakou \(2006\)](#), la modelización de las concentraciones de material particulado resulta más difícil que la de los contaminantes gaseosos habituales debido a la complejidad de los procesos que controlan la formación, transporte y eliminación de los aerosoles en la atmósfera. Ello ha favorecido el empleo de técnicas como las redes neuronales, capaces de capturar el carácter altamente no lineal de estos procesos. Como resultado, se han llevado a cabo numerosos estudios en relación con la aplicación de modelos estadísticos basados en redes neuronales para la predicción en exclusiva de niveles de partículas. [Pérez y col. \(2000\)](#) compararon las estimaciones, generadas mediante modelos de redes neuronales multicapa, de regresión lineal y de persistencia, de la concentración horaria de PM_{2.5} en Santiago (Chile) hasta 24 h por adelantado. En [Pérez y Reyes \(2002\)](#) y [Pérez y Reyes \(2006\)](#) extendieron su estudio a la fracción PM₁₀ del material particulado, prediciendo las concentraciones máximas con un día de antelación. Obtuvieron que las redes neuronales presentaban una ligera ventaja frente a los modelos lineales y de persistencia. No obstante, en cómputo global, ambas técnicas eran suficientemente precisas como para ser consideradas herramientas de evaluación de la calidad del aire. Estudios similares en relación con el uso de redes neuronales artificiales para la estimación de valores diarios de PM₁₀ y la comparación de su desempeño frente al de técnicas de regresión lineal múltiple han sido llevados a cabo por [Grivas y Chaloulakou \(2006\)](#) en Atenas (Grecia) y por [Papanastasiou y col. \(2007\)](#) en Volos (Grecia). En ambos trabajos, se obtuvo que el desempeño de las redes neuronales era superior al de las técnicas lineales, en especial en términos de coeficiente de correlación. Igualmente, [Caselli y col. \(2009\)](#) desarrollaron un modelo de redes neuronales de tipo perceptrón multicapa para la predicción de valores diarios de PM₁₀ en Bari (Italia) y compararon los resultados con una red neuronal de tipo función de base radial y un modelo de regresión lineal múltiple. Asimismo, [Hoi y col. \(2009\)](#) compararon las redes neuronales con un modelo de tiempo variable. Por su parte, [Kim y col. \(2009\)](#) llevaron a cabo un estudio con respecto a la calidad de aire interior. Desarrollaron modelos de redes neuronales y de regresión lineal múltiple para la predicción de PM₁₀ y PM_{2.5} dentro de una estación de metro de Seúl (Corea). [Paschalidou y col. \(2011\)](#) estudiaron a través de un caso de estudio en Chipre la aplicación de redes neuronales para la predicción de valores de PM₁₀ con mayor resolución temporal (horaria). Por último, [Voukantsis y col. \(2011\)](#) evaluaron la consideración de técnicas de análisis factorial como el análisis de componentes principales para su utilización en el desarrollo de modelos de redes neuronales para la estimación de PM₁₀ y PM_{2.5} en Salónica (Grecia) y Helsinki (Finlandia).

Además, hay un gran número de trabajos adicionales que tratan la aplicación de redes neuronales artificiales para el desarrollo de modelos de calidad del aire no solo para la predicción de concentraciones de material particulado, sino también la de otros contaminantes

atmosféricos gaseosos. Entre ellos se encuentran [Kukkonen y col. \(2003\)](#), [Lu y col. \(2003\)](#), [Jiang y col. \(2004\)](#), [Lu y col. \(2004\)](#), [Niska y col. \(2005\)](#), [Brunelli y col. \(2007\)](#), [Kurt y col. \(2008\)](#), [Turias y col. \(2008\)](#), [Cai y col. \(2009\)](#) y [Hrust y col. \(2009\)](#).

1.5.2. Modelos estadísticos para contaminantes minoritarios

En vista del cuantioso número de trabajos con resultados favorables para la estimación de niveles de contaminantes atmosféricos mayoritarios mediante modelos estadísticos, cabría esperar que estas técnicas fuesen igualmente válidas para la estimación de calidad del aire en relación a contaminantes minoritarios, en especial en cuanto a la estimación de los niveles en aire ambiente de los componentes del material particulado, como pueden ser los metales pesados y algunos microcontaminantes orgánicos. No obstante, para este tipo de contaminantes se encuentra un reducido número de publicaciones en la literatura. Con respecto al contenido de sustancias orgánicas, distintos autores han estudiado la aplicación de la regresión lineal múltiple para la estimación de concentraciones de benzo(a)pireno (BaP) contenido en el material particulado empleando como predictores variables meteorológicas como temperatura, velocidad del viento, humedad relativa o las precipitaciones ([Lobscheid y col. 2007](#)), la concentración del propio material particulado y variables meteorológicas como temperatura, radiación solar, y velocidad del viento ([Callén y col., 2010](#)), así como concentraciones de contaminantes mayoritarios (NO_x , O_3 , $\text{PM}_{2.5}$) y otras variables de influencia, como temperatura o estación del año ([Zalel y col. 2015](#)).

En cuanto a los metales, [Hernández y col. \(1992\)](#) aplicaron modelos de espacio-estado con filtro Kalman y modelos de series de tiempo de Box-Jenkins, de función de transferencia y auto-regresivos de media móvil (ARIMA), para la predicción de las concentraciones diarias de Fe y Pb en Madrid (España) a partir de datos de velocidad media diaria de viento del día precedente. Las estimaciones para Fe fueron mejores que las de Pb. Para ambos metales no se encontraron diferencias entre los modelos de espacio-estado y de función de transferencia; sin embargo, su desempeño fue mejor que el del modelo ARIMA en términos de error cuadrático medio, coeficiente de correlación y eficiencia del modelo. Un modelo auto-regresivo con entrada exógena (ARX, por sus siglas en inglés) junto con un modelo de espacio-estado con filtro Kalman fueron desarrollados por [Chelani y col. \(2001\)](#) para predecir la concentración diaria de RSPM, Pb, Fe y Zn a partir de la velocidad del viento en Delhi (India). El modelo de espacio-estado exhibió un desempeño superior al modelo ARX. [Chelani y col. \(2002b\)](#) emplearon redes neuronales artificiales para la estimación de PM_{10} y Cd, Cr, Fe, Ni, Pb y Zn en Jaipur (India), empleando variables meteorológicas como *inputs*, en concreto, las siguientes: velocidad y dirección del viento, temperatura y humedad relativa. Se observó que las redes neuronales eran capaces de predecir la concentración de todos los contaminantes con valores de error cuadrático medio relativamente bajos. Por su parte, [Li y col. \(2009\)](#) desarrollaron modelos de

redes neuronales y regresión lineal múltiple para su aplicación en la reconstrucción de exposición de manganeso ocupacional empleando como variables independientes factores de influencia como intensidad y frecuencia de la exposición, circunstancia (en interior o exterior), distancia a la fuente de emisión y duración de la exposición. Por otra parte, en un estudio previo llevado a cabo por [Arruti y col. \(2011\)](#) se estimaron los niveles de As, Cd, Ni y Pb en Cantabria (España) mediante regresión lineal múltiple y regresión por componentes principales en base a concentraciones en inmisión de contaminantes atmosféricos mayoritarios (NO_x , SO_2 , O_3 , PM_{10}), condiciones meteorológicas y variables nominales que reflejan los efectos estacionales, del fin de semana y la ocurrencia de episodios de intrusión de polvo sahariano. Se concluyó que ambas técnicas representan alternativas válidas para la evaluación de la calidad del aire con respecto a los contaminantes estudiados. Más aún, [Vicente y col. \(2012\)](#) desarrollaron modelos predictivos basados en análisis de regresión múltiple y series de tiempo para predecir la concentración de partículas totales en suspensión (TSP, por sus siglas en inglés), PM_{10} , As, Cd, Ni y Pb en el aire ambiente de Castellón (España) a partir de variables meteorológicas tales como las velocidades máxima y media del viento, la dirección predominante del viento, la temperatura media, la humedad relativa y la presión.

1.6. Hipótesis de partida y objetivos

En este trabajo se plantea como hipótesis la existencia de una relación entre las concentraciones en inmisión de metales y semimetales que están presentes en el material particulado y la presencia de contaminantes atmosféricos mayoritarios en el aire ambiente en una zona urbana determinada bajo unas condiciones meteorológicas dadas. Es sabido que los niveles en inmisión de dichos contaminantes están fuertemente influenciados por actividades antropogénicas como la industria, la combustión residencial e industrial o el tráfico rodado y, por tanto, reflejarían indirectamente el impacto en la calidad del aire de dichas actividades así como la manera en la que se han dispersado los contaminantes una vez liberados a la atmósfera.

El objetivo general de la presente Tesis Doctoral es el establecimiento de las relaciones estadísticas entre la concentración en inmisión de contaminantes minoritarios (metales y metaloides regulados por la UE, o bien con valores guía recomendados por la OMS) y la concentración en inmisión de contaminantes mayoritarios, así como las condiciones meteorológicas a escala local en zonas urbanas mediante el uso de distintas técnicas matemáticas, contribuyendo al avance en el conocimiento de las herramientas para la estimación de los niveles en aire ambiente de metales basadas en técnicas de estimación objetiva para su aplicación en procedimientos de evaluación de la calidad del aire. Para alcanzar el objetivo general es necesario abordar una serie de objetivos específicos:

1. Desarrollar una metodología para la estimación de los niveles de As, Cd, Ni y Pb cuando estos se encuentren por debajo del umbral inferior de evaluación, de acuerdo a la nomenclatura empleada por las Directivas de calidad del aire de la UE; esta metodología estará basada en el empleo de modelos estadísticos desarrollados a partir de regresión por mínimos cuadrados parciales (PLSR) y de redes neuronales artificiales (ANNs).
2. Preparar una base de datos con niveles de los contaminantes minoritarios a evaluar, y con el conjunto de variables de entrada que puedan tener relación con los mismos. Para ello, partiendo de la experiencia propia del grupo de investigación se partirá de datos obtenidos en diferentes estaciones urbanas de Cantabria (Reinosa, Castro Urdiales y Santander) durante 2008 y 2009. Para completar esta base de datos, se planteará también como objetivo específico la determinación de los niveles de metales traza en la fracción PM_{10} del material particulado en zonas urbanas de Cantabria en los años 2010 (Reinosa, Castro Urdiales) y 2011 (Santander).
3. Aplicación de la metodología desarrollada a pequeñas áreas urbanas y que proceden de una campaña extensiva de muestreo de PM_{10} : Reinosa y Castro Urdiales. En este tipo de áreas, las relaciones entre las fuentes de contaminantes minoritarios y sus niveles en aire ambiente se asumen como simples.
4. Aplicación de la metodología desarrollada a un caso de estudio que procede de una campaña intensiva de muestreo de PM_{10} : Dunkerque. Para este caso de estudio se aplicará la metodología también a metales no regulados por la UE pero que sí son recogidos en las guías de la OMS (Mn, V y Cr).
5. Aplicación de la metodología desarrollada a un área urbana más compleja a partir de una campaña extensiva de muestreo de PM_{10} : Santander. En este caso particular, se trata la mayor área urbana de Cantabria y la más compleja desde el punto de vista del número y variedad de fuentes de los metales estudiados.

1.7. Bibliografía

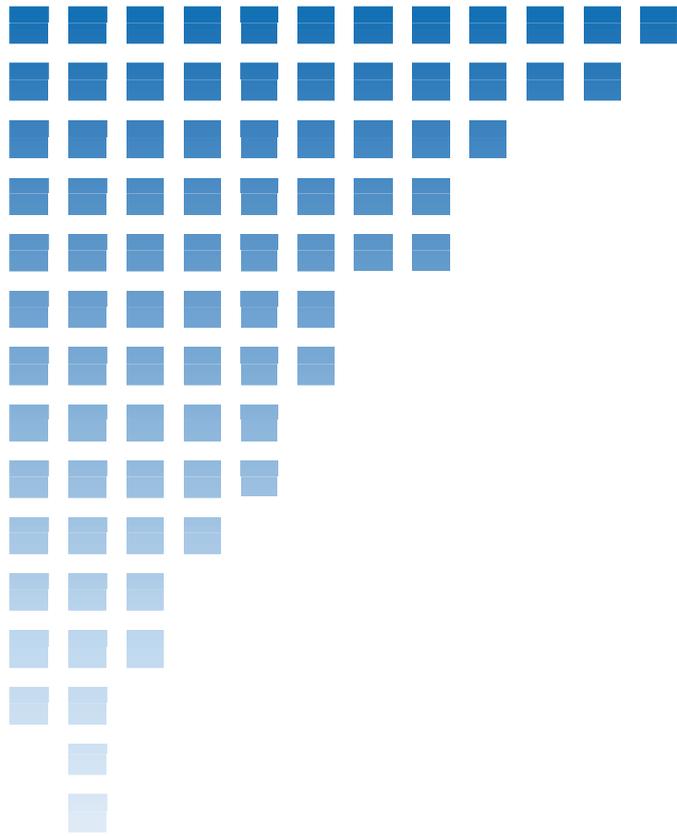
- Abdul-Wahab, S.A., Al-Alawi, S.M., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software* 17, 219-228.
- Adachi, K., Tainosho, Y., 2004. Characterization of heavy metal particles embedded in tire dust. *Environment International* 30, 1009-1017.
- Almeida, S.M., Freitas, M.C., Reis, M., Pinheiro, T., Felix, P.M., Pio, C.A., 2013. Fifteen years of nuclear techniques application to suspended particulate matter studies. *Journal of Radioanalytical and Nuclear Chemistry* 297, 347-356.
- Almeida, S.M., Pio, C.A., Freitas, M.C., Reis, M.A., Trancoso, M.A., 2005. Source apportionment of the fine and coarse particulate matter in a sub-urban area at Western European Coast. *Atmospheric Environment* 39, 3127-3138.
- Arditsoglou, A., Samara, C., 2005. Levels of total suspended particulate matter and major trace elements in Kosovo: a source identification and apportionment study. *Chemosphere* (59), 669-678.
- Arruti, A., Fernández-Olmo, I., Irabien, A., 2011. Assessment of regional metal levels in ambient air by statistical regression models. *Journal of Environmental Monitoring* 13(7), 1991-2000.
- Arruti, A., 2011. Evaluación de la composición del material particulado atmosférico en medio urbano y rural en Cantabria. Tesis doctoral de la Universidad de Cantabria.
- Aschner, M., Erikson, K.M., Dorman, D.C., 2005. Manganese dosimetry: species differences and implications for neurotoxicity. *Critical Reviews in Toxicology* 35(1), 1-32.
- ATSDR, 2012a. Toxicological profile for cadmium. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2012b. Toxicological profile for chromium. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2012c. Toxicological profile for manganese. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2012d. Toxicological profile for vanadium. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2007a. Toxicological profile for arsenic. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2007b. Toxicological profile for lead. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- ATSDR, 2005. Toxicological profile for nickel. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Public Health Service. Atlanta, Georgia.
- Avino, P., Capannesi, G., Rosada, A., 2014. Source identification of inorganic airborne particle fraction (PM₁₀) at ultratrace levels by means of INAA short irradiation. *Environmental Science and Pollution Research* 21, 4527-4538.
- Batonneau, Y., Bremard, C., Gengembre, L., Laureys, J., Le Maguer, A., Le Maguer, D., Perdrix, E., Sobanska, S., 2004. Speciation of PM₁₀ sources of airborne nonferrous metals within the 3-km zone of lead/zinc smelters. *Environmental Science and Technology* 38(2), 5281-5289.
- Becagli, S., Sferlazzo, D.M., Pace, G., di Sarra, A., Bommarito, C., Calzolari, G., Ghedini, C., Lucarelli, F., Meloni, D., Monteleone, F., Seeri, M., Traversi, R., Udisti, R., 2012. Evidence for heavy fuel oil combustion aerosols from chemical analyses at the island of Lampedusa: a possible large role of ships emissions in the Mediterranean. *Atmospheric Chemistry and Physics* 12, 3479-3492.
- Bocca, B., Caimi, S., Smichowski, P., Gómez, D., Caroli, S., 2006. Monitoring Pt and Rh in urban aerosols from Buenos Aires, Argentina. *Science of the Total Environment* 358, 255-264.
- Borge, R., Lumbreras, J., Tarrason, L., Rodríguez, E., 2011. Emission needs at local scale for air quality modelling. 14th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO14).
- Bosco, M.L., Varrica, D., Dongarrá, G., 2005. Case study: Inorganic pollutants associated with particulate matter from an area near a petrochemical plant. *Environmental Research* 99, 18-30.
- Boudissa, S.M., Lambert, J., Müller, C., Kennedy, G., Gareau, L., Zayed, J., 2006. Manganese concentrations in the soil and air in the vicinity of a closed manganese alloy production plant. *Science of the Total Environment* 361, 67-72.
- Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., Vitabile, S., 2007. Two-days ahead prediction of daily maximum concentrations of SO₂, O₃, PM₁₀, NO₂, CO in the urban area of Palermo, Italy. *Atmospheric Environment* 41, 2967-2995.
- Cai, M., Yin, Y., Xie, M., 2009. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment* 14(1), 32-41.
- Callén, M.S., López, J.M., Mastral, A.M., 2010. Seasonal variation of benzo(a)pyrene in the Spanish airborne PM₁₀. Multivariate linear regression model applied to estimate BaP concentrations. *Journal of Hazardous Materials* 180, 648-655.
- Caselli, M., Trizio, L., De Gennaro, G., Ielpo, P., 2009. A simple feedforward neural network for the PM₁₀ forecasting: Comparison with a radial basis function network and a multivariate linear regression model. *Water, Air, & Soil Pollution* 201(1-49), 365-377.
- Celo, V., Dabek-Zlotorzynska, E., 2010. Concentration and source origin of trace metals in PM_{2.5} collected at selected Canadian sites within the Canadian National Air Pollution Surveillance Program. En: Zereini, F., Wiseman, C.L.S. (Eds.) *Urban Airborne Particulate Matter*, Environmental Science and Engineering. Springer-Verlag Berlin Heidelberg.

- Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment* 313, 1-13.
- Chelani, A.B., Chalapati Rao, C.V., Phadke, K.M., Hasan, M.Z., 2002a. Prediction of sulphur dioxide concentration using artificial neural networks. *Environmental Modelling & Software* 17, 161-168.
- Chelani, A.B., Gajghate, D.G., Hasan, M.Z., 2002b. Prediction of ambient PM₁₀ and toxic metals using artificial neural networks. *Journal of the Air and Waste Management Association* 52(7), 805-810.
- Chelani, A.B., Gajghate, D.G., Tamhane, S.M., Hasan, M.Z., 2001. Statistical modeling of ambient air pollutants in Delhi. *Water, Air, and Soil Pollution* 132, 315-331.
- Chen, L.C., Lippmann, M., 2009. Effects of metals within ambient air particulate matter (PM) on human health. *Inhalation Toxicology* 21, 1-31.
- Colls, J., 2002. *Air Pollution*, 2nd Edition. London: Spon Press (Taylor & Francis).
- Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. *Air & Waste Management Association* 47, 653-663.
- Contini, D., Belosi, F., Gambaro, A., Cesari, D., Stortini, A.M., Bove, M.C., 2012. Comparison of PM₁₀ concentrations and metal content in three different sites of the Venice Lagoon: An analysis of possible aerosol sources. *Journal of Environmental Sciences* 24(11), 1954-1965.
- Dai, Q.L., Bi, X.H., Wu, J.H., Zhang, Y.F., Wang, J., Xu, H., Yao, L., Jiao, L., Feng, Y.C., 2015. Characterization and source identification of heavy metals in ambient PM₁₀ and PM_{2.5} in an integrated iron and steel industry zone compared with a background site. *Aerosol and air Quality Research* 15(3), 875-887.
- Daly, A., Zannetti, P., 2007. Air Pollution Modeling – an overview. En: Zannetti, P., Al-Ajmi, D., Al-Rashied, S. (Eds.) *Ambient air pollution*, Arab School for Science and Technology (ASST) and The EnviroComp Institute.
- Demuzere, M., Lipzig, N.P.M., 2010. A new method to estimate air-quality levels using a synoptic-regression approach. Part I: present-day O₃ and PM₁₀ analysis. *Atmospheric Environment* 44, 1341-1355.
- Denby, B., Sundvor, I., Cassiani, M., Smet, P., Leeuw, F., Horálek, J., 2010. Spatial mapping of ozone and SO₂ trends in Europe. *Science of the Total Environment* 408, 4795-4806.
- Dockery, D.W., Pope III, C.A., Spengler, J.D., Ware, H.J., Fay, M.E., Ferris Jr, B.G., Speizer, F.E., 1993. An association between air pollution and mortality in six US cities. *The New England Journal of Medicine* 329, 1753-1759.
- Dongarrá, G., Manno, E., Varrica, D., 2009. Possible markers of traffic related emissions. *Environmental Monitoring and Assessment* 154, 117-125.
- Dongarrá, G., Manno, E., Varrica, D., Vultaggio, M., 2007. Mass levels, crustal component and trace elements in PM₁₀ in Palermo, Italy. *Atmospheric Environment* 41, 7977-7986.
- EEA, 2014. *Air quality in Europe — 2014 report*, EEA Report No 5/2014, European Environment Agency, Copenhagen, Denmark.
- El-Harabawi, M., 2013. Air quality modelling, simulation, and computational methods: a review. *Environmental Reviews* 21, 149-179.
- EC, 2008. Council Directive 2008/50/EC Directive of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. The European Parliament and the Council of the European Union. *Off J L* 52 1-44.
- EC, 2004. Council Directive 2004/107/EC Directive of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air. The European Parliament and the Council of the European Union. *Off J L* 23 3-16.
- EC, 1998. Council Directive 98/70/EC. Directive of the European Parliament and of the Council of 13 October 1998 relating to the quality of petrol and diesel fuels and amending Council Directive 93/12/EEC. The European Parliament and the Council of the European Union. *Off J L* 350 58-67.
- Fernández-Camacho, R., Rodríguez, S., de la Rosa, J., Sánchez de la Campa, A.M., Alastuey, A., Querol, X., Gonzalez-Castanedo, Y., García-Orellana, I., Nava, S., 2012. Ultrafine particle and fine trace metal (As, Cd, Cu, Pb and Zn) pollution episodes induced by industrial emissions in Huelva, SW Spain. *Atmospheric Environment* 61, 507-517.
- Fernández-Espinosa, A.J., 2001. Especiación química y física de metales en la materia particulada atmosférica: aplicación al estudio de la contaminación ambiental de la ciudad de Sevilla. Universidad de Sevilla, Fundación Focus-Abengoa.
- Fernández-Olmo, I., Andecochea, C., Ruiz, S., Fernández-Ferreras, J.A., Irabien, A., 2016. Local source identification of trace metals in urban/industrial mixed land-use areas with daily PM₁₀ limit value exceedances. *Atmospheric Research* 171, 92-106.
- Finkelstein, M.M., Jerret, M., 2007. A study of the relationships between Parkinson's disease and markers of traffic-derived and environmental manganese air pollution in two Canadian cities. *Environmental Research* 104(3), 420-432.
- Finzi, G., Nunari, G., 2005. Air quality forecast and alarm systems. En: Zannetti, P. (Ed.) *Air Quality Modeling: Theories, methodologies, computational techniques, and available databases Vol. II – Advanced topics*, The EnviroComp Institute and Air & Waste Management Association.
- Gao, Y., Nelson, E.D., Field, M.P., Ding, Q., Li, H., Sherrell, R.M., 2002. Characterization of atmospheric trace elements on PM_{2.5} particulate matter over the New York-New Jersey harbor estuary. *Atmospheric Environment* 36, 1077-1086.
- Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NO_x and NO₂ concentrations in urban air in London. *Atmospheric Environment* 33, 709-719.
- Gaudry, A., Moskura, M., Mariet, C., Ayrault, S., Denayer, F., Bernard, N., 2008. Inorganic pollution in PM₁₀ particles collected over three French sites under various influences: rural conditions, traffic and industry. *Water, Air, and Soil Pollution* 193, 91-106.
- Ghosh, S.K., Saha, R., Saha, B., 2015. Toxicity of inorganic vanadium compounds. *Research on Chemical Intermediates* 41(7), 4873-4897.

- Gilbert, R.O., 1987. Statistical methods for environmental pollution monitoring. Nueva York: Van Nostrand Reinhold.
- Gómez-Sanchis, J., Martín-Guerrero, J.D., Soroa-Olivas, E., Vila-Francés, J., Carrasco, J.L., Valle-Tascón, S., 2006. Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. *Atmospheric Environment* 40, 6173-6180.
- González-Castanedo, Y., Sánchez-Rodas, D., Sánchez de la Campa, A.M., Pandolfi, M., Alastuey, A., Cachorro, V.E., Querol, X., de la Rosa, J.D., 2015. Arsenic species in the atmospheric particulate matter as tracer of the air quality of Doñana Park (SW Spain). *Chemosphere* 119, 1296-1303.
- Goyal, P., Chan, A.T., Jasiwal, N., 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment* 40, 2068-2077.
- Grieshop, A.P., Lipsky, E.M., Pekney, N.J., Takahama, S., Robinson, A.L., 2006. Fine particle emission factors from vehicles in a highway tunnel effects of fleet composition an season. *Atmospheric Environment* 40, S287-S298.
- Griffin, R.D., 2007. Principles of Air Quality Management. Florida: Taylor & Francis.
- Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of PM₁₀ hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment* 40(7), 1216-1229.
- Hanna, S.R., 1989. Plume dispersion and concentration fluctuations in the atmosphere. En: Cheremisinoff, P.N. (Ed.), *Encyclopedia of Environmental Control Technology Vol. 2, Air Pollution Control*. Houston, Texas: Gulf Publishing Co.
- Hedberg, E., Gidhagen, L., Johansson, C., 2005. Source contributions to PM₁₀ and arsenic concentrations in Central Chile using positive matrix factorization. *Atmospheric Environment* 39, 549-561.
- Hernández, E., Martín, F., Valero, F., 1992. Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain. *Atmospheric Environment* 26B, 107-116.
- Hernández-Bonilla, D., Schilman, A., Montes, S., Rodríguez-Agudelo, Y., Rodríguez-Dozal, S., Solís-Vivanco, R., Ríos, C., Riojas-Rodríguez, H., 2011. Environmental exposure to manganese and motor function of children in Mexico. *Neurotoxicology* 32(5), 615-621.
- Hoi, K.I., Yuen, K.V., Mok, K.M., 2009. Prediction of daily averages PM₁₀ concentrations by statistical time-varying model. *Atmospheric Environment* 43(16), 2579-2581.
- Hrust, L., Klaić, Z.B., Križan, J., Antonić, O., Hercog, P., 2009. Neural network forecasting of air pollutants hourly concentrations using optimized temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment* 43(35), 5588-5596.
- Inal, F., 2010. Artificial neural network prediction of tropospheric ozone concentrations in Istanbul, Turkey. *Clean – Soil, Air, Water* 38, 897-908.
- Jiang, D., Zahng, Y., Hu, X., Zeng, Y., Tan, J., Shao, D., 2004. Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment* 38(40 SPEC. ISS.), 7055-7064.
- Keegan, T.J., Farago, M.E., Thornton, I., Hong, B., Colvile, R.N., Pesch, B., Jakubis, P., Nieuwenhuijsen, M.J., 2006. Dispersion of As and selected heavy metals around a coal-burning power station in central Slovakia. *Science of the Total Environment* 358, 61-71.
- Khatibi, R., Naghipour, L., Ghorbani, M.A., Smith, M.S., Karimi, V., Farhoudi, R., Delafrouz, H., Arvanaghi, H., 2013. Developing a predictive tropospheric ozone model for Tabriz. *Atmospheric Environment* 68, 286-294.
- Kim, M., Kim, Y., Sung, S., Yoo, C., 2009. Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks. ICCAS-SICE 2009 – ICROS-SICE International Joint Conference 2009, Proceedings, 1688-1692.
- Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment* 35, 816-825.
- Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural network models for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment* 37(32), 4539-4550.
- Kuo, S.C., Hsieh, L.Y., Tsai, C.H., Tsai, Y.I., 2007. Characterization of PM_{2.5} fugitive metal in the workplaces and the surrounding environment of a secondary aluminum smelter. *Atmospheric Environment* 41(32), 6884-6900.
- Kurt, A., Gulbagci, B., Karaca, F., Alagha, O., 2008. An online air pollution forecasting system using neural networks. *Environment International* 34(5), 592-598.
- Li, Y., Luo, F., Jiang, Y., Lu, Y., Huang, J., Zhang, Z., 2009. A prediction model of occupational manganese exposure based on artificial neural network. *Toxicology Mechanisms and Methods* 19(5), 337-345.
- Lobscheid, B.A., McKone, E.T., Vallero, A.D., 2007. Exploring relationships between outdoor air particulate associated polycyclic aromatic hydrocarbon and PM_{2.5}: a case study of benzo(a)pyrene in California metropolitan regions. *Atmospheric Environment* 41, 5659-5672.
- López, J.M., Callén, M.S., Murillo, R., García, T., Navarro, M.V., de la Cruz, M.T., Mastral, A.M., 2005. Levels of selected metals in ambient air PM₁₀ in an urban site of Zaragoza (Spain). *Environmental Research* 99, 58-67.
- Lu, W., Wang, W., Wang, X., Yan, S., Lam, J.C., 2004. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. *Environmental research* 96(1), 79-87.
- Lu, W.Z., Wang, W.J., Wang, X.K., Xu, Z.B., Leung, A.Y.T., 2003. Using improved neural network model to analyze RSP, NO_x and NO₂ levels in urban air in Mong Kok, Hong Kong. *Environmental monitoring and assessment* 87(3), 235-254.
- Maenhaut, W., 2001. Inorganic tracers for source apportionment: analytical techniques and selected results. En: Proceedings of the European Aerosol Conference, 3-7 de septiembre de 2001, Leipzig, Alemania.
- Manalis, N., Grivas, G., Protonotarios, V., Moutsatsou, A., Samara, C., Chaloulakou, A., 2005. Toxic metal content of particulate matter (PM₁₀), within the Greater Area of Athens. *Chemosphere* 60, 557-566.

- Martin, S., Griswold, W., 2009. Human Health Effects of Heavy Metals. Environmental science and technology briefs for citizens. Issue 15. Center for Hazardous Substance Research (CHSR), Manhattan, Kansas.
- Mergler, D., Baldwin, M., Bélanger, S., Larribe, F., Beuter, A., Bowler, R., Panisset, M., Edwards, R., de Geoffroy, A., Sassine, M.P., Hudnell, K., 1999. Manganese neurotoxicity, a continuum of dysfunction: results from a community based study. *Neurotoxicology* 20(2-3), 327-342.
- Morawska, L., Zhang, J., 2002. Combustion sources of particles. Health relevance and source signatures. *Chemosphere* 49, 1045-1058.
- Moreno, T., Pandolfi, M., Querol, X., Lavín, J., Alastuey, A., Viana, M., Gibbons, W., 2011. Manganese in the urban atmosphere: identifying anomalous concentrations and sources. *Environmental Science and Pollution Research* 18, 173-183.
- Moreno, T., Querol, X., Alastuey, A., Viana, M., Salvador, P., Sanchez de la Campa, A., Artiñano, B., de la Rosa, J., Gibbons, W., 2006. Variation in atmospheric PM trace metal content in Spanish towns: illustrating the chemical complexity of the inorganic urban aerosol cocktail. *Atmospheric Environment* 40, 6791-6803.
- Niska, H., Rantamäki, M., Hiltunen, T., Karppinen, A., Kukkonen, J., Ruuskanen, J., Kolehmainen, M., 2005. Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. *Atmospheric Environment* 39(35), 6524-6536.
- Noori, R., Hoshyaripour, G., Ashrafi, K., Araabi, B.N., 2010. Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. *Atmospheric Environment* 44, 476-482.
- Pacyna, J.M., Pacyna, E.G., Aas, W., 2009. Changes of emissions and atmospheric deposition of mercury, lead, and cadmium. *Atmospheric Environment* 43, 117-127.
- Pacyna, E.G., Pacyna, J.M., Fudala, J., Strzelecka-Jastrzab, E., Hlawiczka, S., Panasiuk, D., Nitter, S., Pregger, T., Pfeiffer, H., Friedrich, R., 2007. Current and future emissions of selected heavy metals to the atmosphere from anthropogenic sources in Europe. *Atmospheric Environment* 41, 8557-8566.
- Pacyna, J.M., Pacyna, E.G., 2001. An assessment of global and regional emissions of trace metals to the atmosphere from anthropogenic sources worldwide. *Environmental Reviews* 9(4), 269-298.
- Pacyna, J.M., 1987. Atmospheric emissions of arsenic, cadmium, lead and mercury from high temperature processes in power generation and industry. En: Hutchinson, T.C., Meema, K.M. (Eds.), *Lead, Mercury and Arsenic in the Environment*. Wiley, pp. 69-87.
- Pacyna, J.M., 1986. Emission factors of atmospheric elements, En: Nriagu, J.O., Davidson, C.I. (Eds.), *Toxic Metals in the Atmosphere*. Wiley, pp. 33-52.
- Papanastasiou, D.K., Melas, D., Kioutsoukis, I., 2007. Development and assessment of neural network and multiple regression models in order to predict PM₁₀ levels in a medium-sized Mediterranean city. *Water, Air, & Soil Pollution* 182 (1-4), 325-334.
- Paschalidou, A.K., Karakitsios, S., Kleanthous, S., Kassomenos, P.A., 2011. Forecasting hourly PM₁₀ concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. *Environmental Science and Pollution Research*, 18(2), 316-327.
- Pérez, P., Reyes, J., 2006. An integrated neural network model for PM₁₀ forecasting. *Atmospheric Environment* 40(16), 2845-2851.
- Pérez, P., Reyes, J., 2002. Prediction of maximum of 24-h average of PM₁₀ concentrations 30h in advance in Santiago, Chile. *Atmospheric Environment* 36(28), 4555-4561.
- Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment* 34(8), 1189-1196.
- Pires, J.C.M., Martins, F.G., 2011. Correction methods for statistical models in tropospheric ozone forecasting. *Atmospheric Environment* 45, 2413-2417.
- Pires, J.C.M., Martins, F.G., Sousa, S.I.V., Alvim-Ferraz, M.C.M., Pereira, M.C., 2008. Prediction of the daily mean PM₁₀ concentrations using linear models. *American Journal of Environmental Sciences* 4(5), 445-453.
- Polat, K., Durduran, S.S., 2012. Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM₁₀) in the city of Konya. *Neural Computing and Applications* 21(8), 2153-2162.
- Pope II, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E., Heath, C.W., 1995. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American Journal of Respiratory and Critical Care Medicine* 151, 669.
- Qin, Y., Oduyemi, K., 2003. Chemical composition of atmospheric aerosol in Dundee, UK. *Atmospheric Environment* 37, 93-104.
- Querol, X., Viana, M., Moreno, T., Alastuey, A., Pey, J., Amato, F., Pandolfi, M., Minguillón, M.C., Reche, C., Pérez, N., González, A., Pallarés, M., Moral, A., Monfort, E., Escrig, A., Cristóbal, A., Hernández, I., Miró, J.V., Jiménez, S., Reina-Velázquez, F., Jabato-Martín, R., Ballester, F., Boldo, E., Bellido, J., 2012. Bases científico-técnicas para un Plan Nacional de Mejora de la Calidad del Aire. Madrid: Consejo Superior de Investigaciones Científicas (CSIC).
- Querol, X., Alastuey, A., Moreno, T., Viana, M., Castillo, S., Pey, J., Rodríguez, S., Artiñano, B., Salvador, P., Sánchez, M., García Dos Santos, S., Herce-Garraleta, M.D., Fernández-Patier, R., Moreno-Grau, S., Minguillón, M.C., Monfort, E., Sanz, M.J., Palomo-Martín, R., Pinilla-Gil, E., Cuevas, E., 2008. Spatial and temporal variations in airborne particulate matter (PM₁₀ and PM_{2.5}) across Spain 1999-2005. *Atmospheric Environment* 42, 3964-3979.
- Querol, X., Viana, M., Alastuey, A., Amato, F., Moreno, T., Castillo, S., Pey, J., de la Rosa, J., Sánchez de la Campa, A., Artiñano, B., Salvador, P., García Dos Santos, S., Gernández-Patier, R., Moreno-Grau, S., Negral, L., Minguillón, M.C., Monfort, E., Gil, J.I., Inza, A., Ortega, L.A., Santamaría, J.M., Zabalza, J., 2007. Source origin of trace elements in PM form regional background, urban and industrial sites of Spain. *Atmospheric Environment* 41, 7219-7231.

- Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric chemistry and physics: from air pollution to climate change, 2nd Edition. Nueva York: John Wiley & Sons, Inc.
- Shao, X., Cheng, H., Li, Q., Lin, C., 2013. Anthropogenic atmospheric emissions of cadmium in China. *Atmospheric Environment* 79, 155-160.
- Silbergeld, E.K., 1996. The elimination of lead from gasoline: impacts of lead in gasoline on human health and the costs and benefits of eliminating lead additives, draft paper, The World Bank, Washington, DC, 3.
- Silbergeld, E.K., 1995. The international dimensions of lead exposure. *International Journal of Occupational and Environmental Health* 1(4), 338-340.
- Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment* 426, 244-255.
- Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, *Environmental Modelling & Software* 22, 97-103.
- Sternbeck, J., Sjödin, A., Andréasson, K., 2002. Metal emissions from road traffic and the influence of resuspension - results from two tunnel studies. *Atmospheric Environment* 36, 4735-4744.
- Sun, W., Zhang, H., Palazoglu, A., 2013. Prediction of 8h-average ozone concentration using a supervised hidden Markov model combined with generalized linear models. *Atmospheric Environment* 81, 199-208.
- Szabó, J., Szabó, A., Erdős, J., 2015. Ambient concentrations of PM₁₀, PM₁₀-bound polycyclic aromatic hydrocarbons and heavy metals in an urban site of Győr, Hungary. *Air Quality, Atmosphere and Health* 8, 229-241.
- Thurston, G.D., Ito, K., Lall, R., 2011. A source apportionment of U.S. fine particulate matter air pollution. *Atmospheric Environment* 45, 3924-3936.
- Tsai, M.Y., Hoek, G., Eeftens, M., de Hoogh, K., Beelen, R., Beregszászi, T., Cesaroni, G., Cirach, M., Cyrus, J., De Nazelle, A., de Vocht, F., Ducret-Stich, R., Eriksen, K., Galassi, C., Gražulevičius, T., Grivas, G., Gryparis, A., Heinrich, J., Hoffmann, B., Iakovides, M., Keuken, M., Krämer, U., Künzli, N., Lanki, T., Madsen, C., Meliefste, K., Merritt, A.S., Mölter, A., Mosler, G., Nieuwenhuijsen, M.J., Pershagen, G., Phuleria, H., Quass, U., Ranzi, A., Schaffner, E., Sokhi, R., Stempfelet, M., Stephanou, E., Sugiri, D., Taimisto, P., Tewis, M., Udvardy, O., Wang, M., Brunekreef, B., 2015. Spatial variation of PM elemental composition between and within 20 European study areas – Results of the ESCAPE project. *Environmental International* 84, 181-192.
- Turias, I.J., González, F.J., Martín, M.L., Galindo, P.L., 2008. Prediction models of CO, SPM and SO₂ concentrations in the Campo de Gibraltar Region, Spain: A multiple comparison strategy. *Environmental monitoring and assessment* 143(1-3), 131-146.
- Uexküll, O., Skerfving, S., Doyle, R., Braungar, M., 2005. Antimony in brake pads: a carcinogenic component? *Journal of Cleaner Production* 13, 19-31.
- USEPA, 1996. Air quality criteria for particulate matter, United States Environmental Protection Agency, Research triangle, NC, Environmental Criteria and Assessment Office, Rep. EPA/600/P-95/001aF.
- Viana, M., Kuhlbusch, T.A.J., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A.S.H., Hueglin, C., Bloemen, H., Wählin, P., Vecchi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitenberger, R., 2008. Source apportionment of particulate matter in Europe: A review of methods and results. *Journal of Aerosol Science* 39, 827-849.
- Viana, M., Querol, X., Alastuey, A., 2006. Chemical characterisation of PM episodes in NE Spain. *Chemosphere* 62, 947-956.
- Vicente, A.B., Jordán, M.M., Sanfeliu, T., Sánchez, A., Esteban, M.D., 2012. Air pollution prediction models of particles, As, Cd, Ni and Pb in a highly industrialized area in Castellón (NE, Spain). *Environmental Earth Sciences* 66(3), 879-888.
- Vijayanand, C., Rajaguru, P., Kalaiselvi, K., Panneer Selvam, K., Palanivel, M., 2008. Assessment of heavy metal contents in the ambient air of the Coimbatore city, Tamilnadu, India. *Journal of Hazardous Materials* 160, 548-553.
- Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis and forecasting of PM₁₀ and PM_{2.5} concentrations using artificial neural networks in Thessaloniki and Helsinki. *Science of the Total Environment* 409(7), 1266-1276.
- Vuković, G., Urošević, M.A., Pergal, M., Janković, M., Goryainova, Z., Tomašević, M., Popović, A., 2015. Residential heating contribution to level of air pollutants (PAHs, major, trace, and rare earth elements): a moss bag case of study. *Environmental Science and Pollution Research* 22, 18956-18966.
- Wählin, P., Berkowicz, R., Plamgren, F., 2006. Characterisation of traffic-generated particulate matter in Copenhagen. *Atmospheric Environment* 40, 2151-2159.
- Wang, W., Lu, W., Wang, X., Leung, A.Y.T., 2003. Prediction of maximum daily ozone level using combined neural network and statistical characteristics. *Environment International* 29, 555-562.
- Wang, X., Sato, T., Xing, B., 2006. Size distribution and anthropogenic sources apportionment of airborne trace metals in Kanazawa, Japan. *Chemosphere* 65, 2440-2448.
- Win Lee, S., He, I., Young, B., 2004. Important aspects in source PM_{2.5} emissions measurement and characterization from stationary combustion systems. *Fuel Processing Technology* 85(6-7), 687-699.
- Yazdanpanah, H., Karimi, M., Hejazizadeh, Z., 2009. Forecasting of daily total atmospheric ozone in Isfahan. *Environmental Monitoring & Assessment* 157, 235-241.
- Yeganeh, B., Shafie Pour Motlagh, M., Rashidi, Y., Kamalan, H., 2012. Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. *Atmospheric Environment* 55, 357-365.
- Yi, J., Prybutok, V.R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution* 92(3), 649-357.
- Zalel, A., Yuval, Svecova, V., Sram, R.J., Bartonova, A., Broday, D.M., 2015. Modeling airborne benzo(a)pyrene concentrations in the Czech Republic. *Atmospheric Environment* 101, 166-176.



Capítulo 2. Bases teóricas de regresión basada en mínimos cuadrados parciales (PLSR) y redes neuronales artificiales (ANNs)

2.1. Regresión de mínimos cuadrados parciales

2.1.1. Concepto y fundamentos teóricos

La aplicación de la técnica estadística de análisis multivariante mediante mínimos cuadrados parciales (PLS, por sus siglas en inglés) en combinación con una etapa posterior de regresión es lo que se conoce como regresión basada en mínimos cuadrados parciales (PLSR, por sus siglas en inglés). Este método constituye una extensión de la regresión lineal múltiple (MLR, por sus siglas en inglés) y, por tanto, se utiliza con el objetivo de predecir un conjunto de variables dependientes o respuesta, denotado por una matriz Y , a partir de un conjunto de variables independientes o predictoras, denotado por una matriz X . En su forma más sencilla, un modelo lineal especifica la relación (lineal) entre una variable dependiente (respuesta, Y) y un conjunto de variables predictoras, X de tal forma que:

$$y = b_0 + b_1x_1 + \dots + b_nx_n \quad (2.1)$$

Siendo b_0 el intercepto (término análogo a la ordenada en el origen) y b_1, \dots, b_n los valores de los coeficientes de regresión computados a partir de los datos x_1, \dots, x_n e y .

No obstante, el método de regresión PLSR no está relacionado únicamente con la regresión lineal múltiple, sino también con la regresión basada en componentes principales (PCR, por sus siglas en inglés), pues combina características de ambas técnicas (Geladi y Kowalski, 1986; Lorber y col., 1987; Wold y col., 1987; Martens y Naes, 1989; Höskuldsson, 1996a). Mientras que MLR busca encontrar un único factor que maximice la correlación entre las variables predictoras y las variables respuesta, PCR pretende encontrar factores que capturen la máxima varianza en las variables predictoras. Por su parte, PLSR intenta satisfacer ambos objetivos de forma simultánea, encontrando factores que capturen varianza y consigan correlación. Para ello, en PLSR se construyen vectores estableciendo un compromiso razonable entre ambas situaciones, es decir, se intenta encontrar factores (llamados variables latentes) que maximicen la cantidad de varianza explicada en X que sea relevante para predecir Y . Esto contrasta con PCR, donde los factores (llamados componentes principales) se seleccionan únicamente en base a la cantidad de varianza que explican en X , sin tener en cuenta la variable respuesta, Y . Es decir, la idea básica del método PLS es que la variable respuesta ha de ser tenida en cuenta en la construcción de los componentes. Por tanto, comúnmente se dice que PLS intenta maximizar la covarianza. En consecuencia, en PLSR se utilizan como variables predictoras, vectores que explican menos varianza que los componentes principales, pero que a cambio están más correlacionados con las respuestas. Tanto los componentes principales como las variables latentes se obtienen en orden decreciente de máxima varianza explicada, por lo que no es

posible aumentar la correlación de los vectores con la respuesta sin perder varianza. No obstante, PLSR se señala como buena alternativa a métodos de regresión clásicos, como MLR y PCR, porque es más robusta (Wold y col., 1984; Otto y Wegscheider, 1985; Geladi y Kowalski, 1986).

El propósito general del método de regresión PLSR es explicar las relaciones entre las variables originales a través de un conjunto de variables latentes de forma que, dadas éstas, las otras sean independientes. Desde un punto de vista más intuitivo, se trata de extraer qué información tienen en común las variables originales y resumirlo en un conjunto reducido de variables latentes. Por tanto, uno de los puntos a favor de PLSR reside en su capacidad para abordar el problema de la multicolinealidad, el cual se presenta cuando una o más variables predictoras del modelo están correlacionadas con otras variables predictoras, es decir, cuando dichas variables no son independientes entre sí. Además, simultáneamente, se consigue una reducción de la dimensionalidad del problema. Estas características hacen de la técnica PLSR una herramienta muy versátil ya que potencian su efectividad frente a la de los habituales métodos de regresión cuando se manejan conjuntos de datos de grandes dimensiones (Wold y col., 2001; Abdi y Williams, 2013). La suposición básica de este método es que el modelo depende de un número pequeño de variables (latentes), que se estiman como combinación lineal de las originales y que se utilizan en su lugar como predictoras de Y . Se trata de descomponer la información de los predictores que corresponden con observaciones correlacionadas en un nuevo subconjunto de variables ortogonales (no correlacionadas). La descomposición de cada una de las matrices de datos de entrada se alcanza a través del establecimiento de dos matrices, que se conocen como matriz de cargas o *loadings* y matriz de puntuaciones o *scores*, que corresponden a la proyección de las variables sobre los componentes.

En base a lo anteriormente expuesto, dado un conjunto de datos de entrenamiento de n observaciones (objetos, casos, compuestos, etc.) con k variables independientes y m variables dependientes -formado, por tanto, por dos matrices X e Y de dimensiones $(n \times k)$ y $(n \times m)$, respectivamente- un modelo PLSR se desarrolla para hallar una descomposición en l factores latentes T tal que:

$$X = TP' + E \tag{2.2}$$

$$Y = UQ' + G \tag{2.3}$$

donde T y U , de dimensiones $(n \times l)$, son las matrices de *scores* de X e Y , respectivamente. P , de dimensiones $(k \times l)$, y Q , de dimensiones $(m \times l)$, son las matrices de *loadings* de X e Y , respectivamente. Por último, E , de dimensiones $(n \times k)$, y G , de dimensiones $(n \times m)$, son las matrices de error de X e Y , respectivamente.

Para la construcción de la matriz T , transformación lineal de X , se requiere además un conjunto adicional de vectores conocidos como pesos, W , de dimensiones $(k \times l)$, necesarios para asegurar la ortogonalidad de los *scores*:

$$T = XW \tag{2.4}$$

Una vez obtenida, esta transformación se utiliza en la regresión en lugar de la matriz original:

$$Y = TQ' + F \tag{2.5}$$

donde F es la matriz de error.

De acuerdo a la ecuación 2.4, la ecuación 2.5 se puede reescribir de tal manera que adquiera la formulación de una ecuación de un modelo de regresión:

$$Y = TQ' + F = XWQ' + F = XB + F \tag{2.6}$$

Por consiguiente, la matriz de los coeficientes asociados a las variables originales, B , de dimensiones $(m \times k)$ se calcula como:

$$B = WQ' \tag{2.7}$$

En base a la ecuación 2.6, se puede decir que PLSR corresponde a una técnica de estimación básicamente lineal, aunque hay casos de modelos PLSR no lineales (Höskuldsson, 1996b; Berglund y Wold, 1997).

La Figura 2.1. muestra una representación gráfica de las matrices y vectores empleados en PLSR.

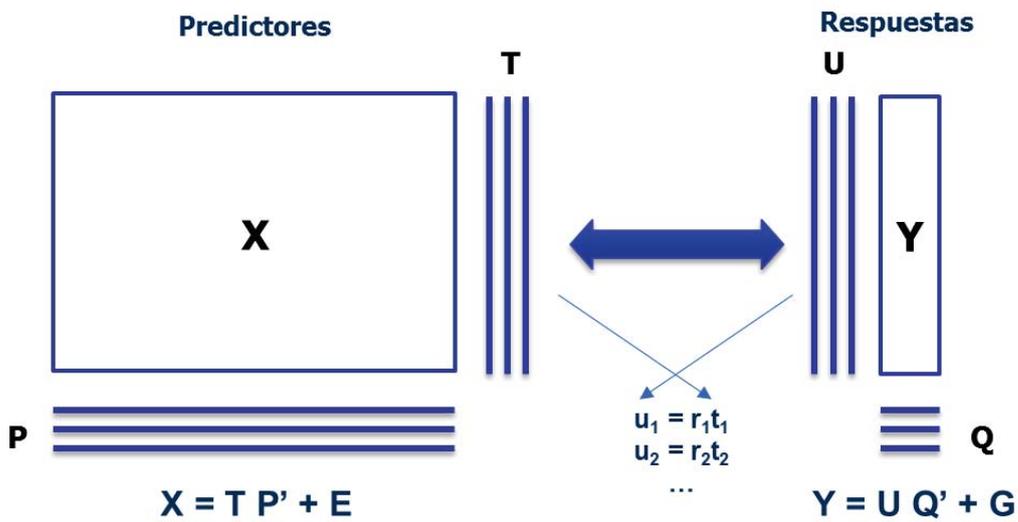


Figura 2.1. Representación esquemática de la descomposición matricial de un modelo PLSR.

Para la interpretación de un modelo PLSR, la matriz de *scores* contiene información acerca de los objetos o muestras y sus similitudes o diferencias. Los pesos proporcionan información sobre cómo las variables se combinan para formar la relación cuantitativa entre X e Y , proporcionando

por tanto una interpretación de los *scores*. Por otro lado, la parte de los datos que no son explicados por el modelo, los residuales, son de interés para su diagnóstico. Así, grandes valores residuales indican que el modelo es pobre. Asimismo, el gráfico de normalidad de los residuales, en problemas que involucran una única variable dependiente, sirve para identificar valores. Además, como en PLSR hay residuales para X , correspondientes a la parte no usada en la modelización de Y , éstos también sirven para identificar valores atípicos en el espacio X (Wold y col., 2001).

Uno de los aspectos más importantes a considerar en el desarrollo de modelos estadísticos PLSR es la determinación de la complejidad del modelo, que viene condicionada por el número de componentes o variables latentes a considerar. En este sentido, se han de tener en cuenta únicamente tantas nuevas variables como se requieran, esto es, tantas como sean significativas para la predictibilidad del modelo. La estrategia estándar para la determinación del número óptimo de variables latentes consiste en aplicar un método de validación cruzada, que constituye una forma práctica y fiable de evaluar la significancia predictiva. Como criterio de selección se utiliza el error cuadrático medio correspondiente a las estimaciones de los subconjuntos de datos considerados para efectuar la validación cruzada, de tal manera que se acepta como óptimo el número de variables latentes que consiga que se minimice dicho error.

Los parámetros del modelo se determinan por medio de un algoritmo. Aunque el cálculo de los parámetros de un modelo PLSR se puede efectuar de varias formas, los algoritmos principales más conocidos son el algoritmo NIPALS (*Non-Iterative Partial Least Squares*) y el algoritmo SIMPLS (*Statistically Inspired Modification of Partial Least Squares*). Una descripción detallada del algoritmo NIPALS se encuentra en Wold y col. (1984). La formulación del algoritmo SIMPLS se presenta en Jong (1993). Ambos algoritmos resultan equivalentes, proporcionando el mismo resultado, cuando se aplican para la resolución de problemas en los que interviene una variable respuesta univariante, es decir, una única variable dependiente. Sin embargo, proporcionan un resultado ligeramente diferente cuando se aplican para la resolución de problemas en los que interviene una variable respuesta multivariante, es decir, formada por más de una variable dependiente, ya que aunque la función objetivo de ambos algoritmos es la misma, las restricciones son ligeramente diferentes (Braak y Jong, 1998). En este sentido, hay evidencias de que la solución proporcionada por el algoritmo SIMPLS en estos casos es en realidad ligeramente superior a la del algoritmo NIPALS puesto que el primero, a diferencia de este último, maximiza el criterio de covarianza. Otra diferencia entre ambos algoritmos radica en la mayor velocidad computacional del algoritmo SIMPLS, la cual aumenta con el tamaño de X . No obstante, en cualquier caso, las implicaciones prácticas son mínimas, por lo que los dos algoritmos se consideran en general igualmente válidos.

2.1.2. Aplicaciones

La técnica de análisis multivariante basada en mínimos cuadrados parciales (PLS) fue desarrollada originalmente para abordar problemas en el campo de la quimiometría y de la economía (Wold y col., 1983). No obstante, desde entonces la implementación de esta técnica se ha extendido ampliamente a otros campos de conocimiento como la medicina y farmacología, el estudio de los genomas, la inteligencia artificial, la informática, sistemas de información, marketing, gestión, o medioambiente, entre otros.

Por otra parte, dado que en su forma general PLS es un método para la predicción de distintos tipos de variable respuesta, el uso de esta técnica no se limita exclusivamente a la resolución de problemas de regresión, en los que la variable respuesta sería continua, sino que también se aplica para abordar otros problemas tales como clasificación (en los que la variable respuesta es una clase o categoría), selección de variables, y análisis de supervivencia, entre otros. En las Tablas 2.1. y 2.2. se muestra una breve revisión de algunos de los trabajos en los que se ha empleado esta técnica en función del objetivo de la aplicación (dichas tablas se centran en aplicaciones de regresión/predicción y clasificación/análisis discriminante, por ser los más habituales) y de la disciplina en la que se enmarca el estudio.

Las aplicaciones de PLSR en el campo de medioambiente para la predicción de los niveles de contaminantes atmosféricos, ya descritas en el apartado 1.5. de esta Tesis no se incluyen en las tablas.

Tabla 2.1. Aplicaciones de la técnica PLS con fines de regresión.

Aplicación	Campo	Descripción	Referencia
Predicción	Genómica	Predicción del contenido de la estructura secundaria de proteínas	Clementi y col., 1997
		Predicción de la interrelación de los niveles de expresión entre varios genes para la formación de levaduras	Johansson y col., 2003
	Quimiometría	Modelización de los descriptores moleculares para estimación de propiedades de compuestos: relaciones cuantitativas estructura-actividad (QSAR)	Alves y col., 2001
		Modelización de datos espectrales	Alsberg y col., 1998; Hubert y Branden, 2003; Gutkin y col., 2009
	Ingeniería y control de procesos	Análisis de fallos y estimación de desempeño de procesos y calidad de productos	Hidajat y Chong, 2000; Lennox y col., 2003; Rosipal y col., 2003; Zamprogna y col., 2004; Zhang y Zhang, 2009
	Medicina y farmacología	Determinación de compuestos en fármacos	Martos y col., 2000
	Economía y contabilidad	Estimación de índices de valores	Huang y Wu, 2008
	Marketing	Predicción de índices de satisfacción de consumidores	Bayol y col., 2000
	Medioambiente	Predicción del contenido de contaminantes en sedimentos	Aulinger y col., 2004
	Agricultura	Predicción de presencia de compuestos en productos agrícolas	Cozzolino y col., 2008; Niu y col., 2008
Inteligencia artificial	Visión artificial	O'Toole y col., 2007	

Tabla 2.2. Aplicaciones de la técnica PLS con fines de clasificación.

Aplicación	Campo	Descripción	Referencia	
Clasificación y análisis discriminante	Genómica	Clasificación de tumores y tipos de cáncer a partir de patrones de expresión de genes	Alaiya y col., 2000; Nguyen y Rocke, 2002; Wu, 2005	
	Quimiometría	Clasificación de niveles de expresión de genes	Boulesteix, 2004	
	Ingeniería y control de procesos	Clasificación basada en datos espectrales	Indhal, 2005; Filzmoser y col., 2012	
	Medicina y farmacología	Control de calidad y de procesos	Charlton y col., 2002; Serranti y col., 2012	
	Economía y contabilidad	Diagnos de presencia y severidad de patologías	Brindle y col., 2002	
	Agricultura		Predicción de bancarrota	Serrano-Cinca y Gutiérrez-Nieto, 2013
			Estimación de índices de valores	Huang y Wu, 2008
			Detección de daños en cultivos	Delalieux y col., 2007
	Inteligencia artificial		Caracterización y clasificación de productos agrícolas para alimentación	Sinelli y col., 2008
			Reconocimiento facial	Schwartz y col., 2009

2.2. Redes neuronales artificiales

2.2.1. Concepto y fundamentos teóricos

Las redes neuronales artificiales son modelos computacionales inspirados en los sistemas nerviosos biológicos de los vertebrados que consisten en redes complejas de unidades básicas de procesamiento, adaptables y altamente interconectadas entre sí. Cada una de estas unidades tiene por objetivo proporcionar una respuesta única o salida en base a un vector de entrada externo (Bishop, 1995; Hassoun, 1995; Bishop, 2006; Haykin, 2009).

Las neuronas biológicas son células del sistema nervioso de los seres vivos encargadas de la recepción de estímulos y de la conducción de los impulsos nerviosos generados en respuesta a dichos estímulos, ya sea entre ellas o con otros tipos celulares. Su morfología se caracteriza por tres elementos principales: soma, dendritas y axón. El cuerpo de la célula, conocido como soma, contiene el núcleo y en él tienen lugar las transformaciones bioquímicas necesarias para la vida de la célula. Del soma parten unas ramificaciones denominadas dendritas, que reciben y transfieren al mismo, vía sinapsis, la información de las células circundantes o señales de entrada. Por su parte, una prolongación del soma neuronal, conocida como axón, lleva a la estructura a la que apunta el impulso nervioso generado por el soma como resultado del procesamiento de los estímulos (Figura 2.2.). Las dendritas reciben la entrada a través de las sinapsis de otras neuronas. El soma procesa las señales de entrada y convierte ese valor procesado en una salida que se envía a otras neuronas a través del axón y de las sinapsis (Flórez y Fernández, 2008).

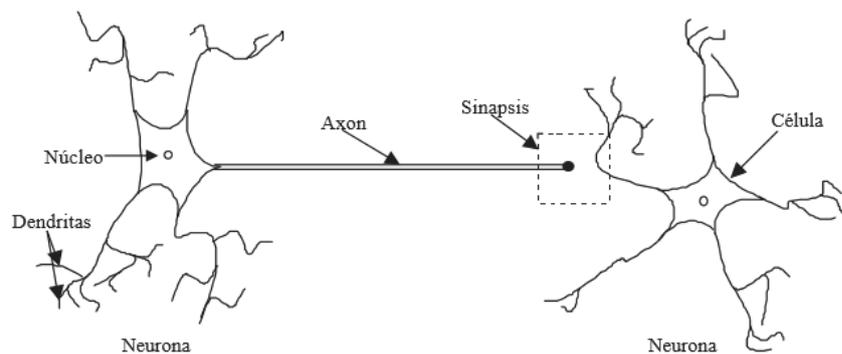


Figura 2.2. Representación esquemática de una neurona biológica.

De manera análoga, como modelo simplificado de la estructura biológica, una neurona artificial contiene como elementos básicos: entradas, elementos de procesamiento y salidas (Figura 2.3.). Las entradas al elemento procesador (equivalente al soma) constituyen el análogo artificial de las dendritas biológicas con las sinapsis. Cada una de estas entradas tiene un peso de conexión asociado, simulando la fuerza de una conexión biológica determinada. El elemento procesador

multiplica cada entrada por el correspondiente peso de la conexión y, generalmente, suma estos productos, que pasan entonces a la función de la transferencia (función de activación) para generar un resultado que se transmite por la vía de salida (axón) (Flórez y Fernández, 2008).

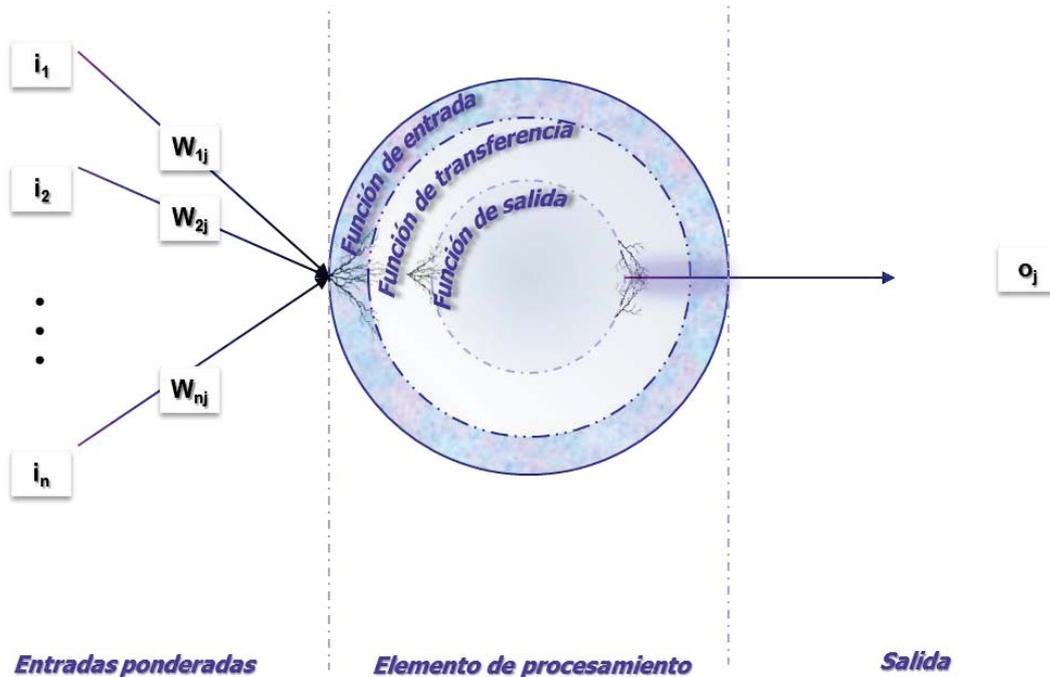


Figura 2.3. Esquema básico de una neurona artificial.

Entre las principales ventajas de las redes neuronales artificiales destacan su capacidad de aprendizaje adaptativo, amplia tolerancia a fallos, así como su capacidad de autoorganización.

En estas redes, una única neurona procesa múltiples entradas aplicando una función de activación sobre una entrada global, calculada como una combinación de cada una de las entradas simples o individuales. Por lo general, para el cálculo de esta entrada global se suele llevar a cabo el sumatorio de las entradas ponderadas, es decir, teniendo en cuenta los pesos correspondientes a cada entrada (Hassoun, 1995; Haykin, 2009).

Al igual que las neuronas biológicas, las neuronas artificiales presentan diferentes estados de activación, que se obtienen por medio de la correspondiente función de activación o de transferencia de la neurona, encargada de la transformación de la entrada global. Entre las funciones de activación más frecuentes se encuentran la función lineal, la función sigmoidea y la función tangente hiperbólica (Hassoun, 1995; Flórez y Fernández, 2008).

El último componente del elemento de procesamiento es la función de salida, cuyo valor resultante determina la salida de la neurona, que se transfiere a las neuronas vinculadas. Las funciones de salida más habituales son la función identidad o la función binaria.

Finalmente, a modo de resumen y al objeto de establecer una analogía directa entre la actividad sináptica biológica y las redes neuronales artificiales, la Tabla 2.3. recoge una serie de relaciones fundamentales.

Tabla 2.3. Analogía entre una neurona biológica y una neurona artificial (Flórez y Fernández, 2008).

Neurona biológica	Neurona artificial
Señales que llegan a las sinapsis	Entradas a la neurona
Carácter excitador o inhibitor de las sinapsis de entrada	Pesos de entrada
Estímulo total de la neurona	Entradas ponderadas
Activación o no de la neurona	Función de activación
Respuesta de la neurona	Función de salida

La forma de organizar las neuronas en una red neuronal artificial se realiza formando niveles, comúnmente conocidos como capas, con un número determinado de dichas neuronas en cada una. Así, en función de la situación dentro de la red se distinguen tres tipos de capas:

- De entrada, que recibe directamente la información procedente de fuentes externas a la red.
- Ocultas, que no tienen contacto directo con el entorno exterior pues son internas a la red. El número de capas ocultas es variable y éste, junto con las distintas maneras en las que pueden estar interconectadas, determina las distintas tipologías de las redes neuronales.
- De salida, que transfieren la información desde la red hacia el exterior.

La representación típica se basa en la estructura de grafo dirigido cuyos nodos son neuronas artificiales. Cuando el grafo que representa la red es acíclico, lo que implica una transferencia unidireccional de la información con sentido de la entrada a la salida, la red se denomina hacia adelante. En una red hacia adelante cada capa recibe sus entradas de unidades de la capa inmediatamente anterior (Figura 2.4.). Otras arquitecturas corresponden a las de redes recurrentes, en las que las unidades de salida retroalimentan a las de entrada (Hassoun, 1995).

Por consiguiente, un sistema neuronal al completo está compuesto y queda definido por los siguientes elementos básicos (Flórez y Fernández, 2008):

- Un conjunto de elementos de procesamiento o neuronas artificiales
- Un patrón de conectividad o arquitectura
- Una dinámica de activaciones
- Una regla o dinámica de aprendizaje
- El entorno donde opera

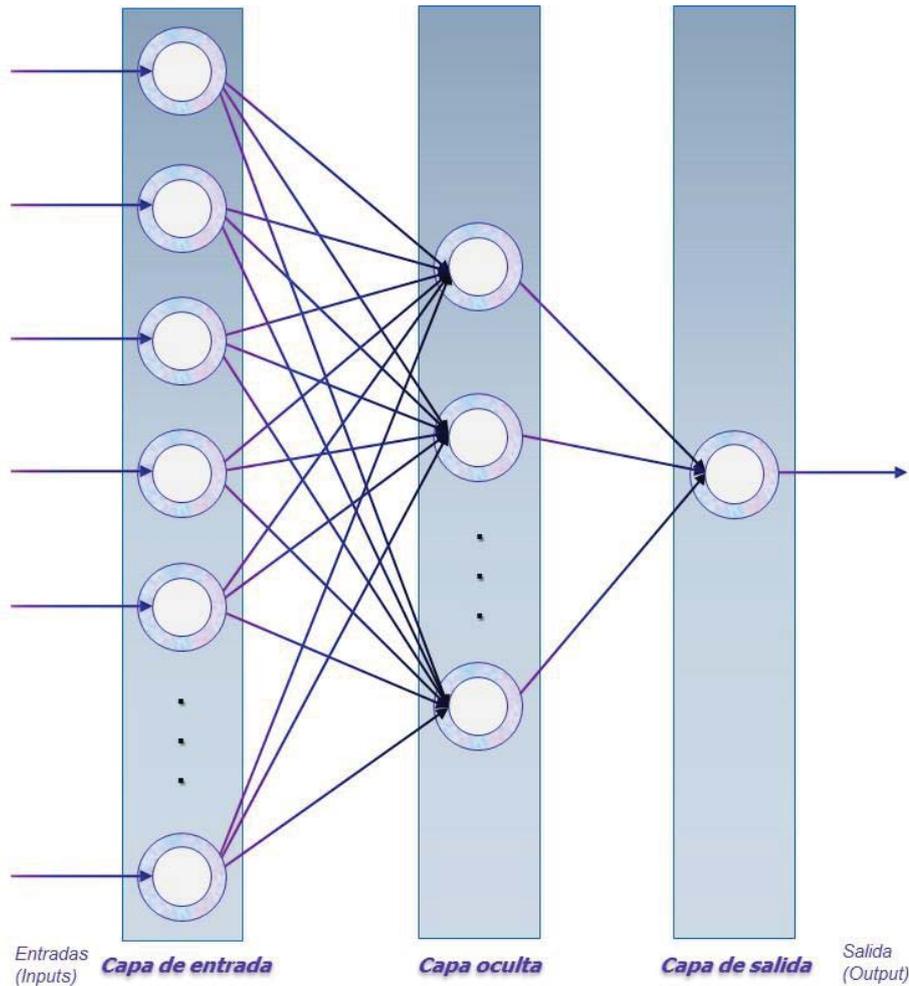


Figura 2.4. Esquema general de una red neuronal hacia adelante y una única capa oculta.

El tipo más común de red neuronal artificial para aplicaciones de regresión es el perceptrón multicapa. Se trata de una red neuronal con conexiones hacia adelante, función de propagación hacia atrás o retropropagación del error y al menos una capa oculta.

El aprendizaje o entrenamiento de las redes neuronales artificiales es el proceso mediante el cual se modifican los pesos de las conexiones entre unidades en respuesta a la información de entrada, de manera que la red se comporte de una determinada manera, descrita por un conjunto de entrenamiento. Los cambios que se producen durante dicho proceso se reducen a la destrucción, modificación y creación de conexiones entre las neuronas. El proceso de aprendizaje finaliza cuando los valores de los pesos permanecen estables.

Un aspecto importante respecto al aprendizaje de las redes neuronales es conocer cómo se modifican los valores de los pesos, es decir, cuáles son los criterios que se siguen para cambiar el valor asignado a las conexiones cuando se pretende que la red aprenda una nueva información. Hay dos métodos de aprendizaje importantes que pueden distinguirse:

aprendizaje supervisado y aprendizaje no supervisado. El primero se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor controla la salida de la red y en caso de que no coincida con la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la deseada. El método de aprendizaje supervisado más utilizado es el aprendizaje por corrección de error.

2.2.2. Aplicaciones

Las características de las redes neuronales las convierten en una técnica especialmente atractiva en dominios de aplicación en los que se tiene poco o nulo conocimiento del problema a resolver. En concreto, entre las principales aplicaciones se encuentran tareas complejas como clasificación, reconocimiento de patrones, optimización, predicción (climatológica, de propiedades, etc.), interpretación de datos sensoriales del mundo real (reconocimiento de voz, visión artificial, reconocimiento de imágenes,...), satisfacción de restricciones, compresión de datos, establecimiento de diagnósticos, etc. Por consiguiente, el desarrollo y aplicación de las redes neuronales artificiales no se limita a una única área de conocimiento, sino que abarca una amplia variedad de campos, tales como economía y finanzas, salud y medicina, ingeniería y producción o marketing, entre otros. Algunas de estas aplicaciones se recogen en a Tabla 2.4.

Las aplicaciones medioambientales de las redes neuronales para la predicción de los niveles de concentración en inmisión de contaminantes atmosféricos han sido revisadas en el apartado 1.5. de la presente Tesis.

Tabla 2.4. Aplicaciones de las ANNs

Campo	Descripción	Referencia
Economía, contabilidad y finanzas	Detección de fraude	Fanning y col., 1995; Lenard y col., 1995
	Predicción de quiebra y probabilidad de bancarrota empresarial	Odom y Sharda, 1990; Pendharkar, 2005
	Predicción de precios en la bolsa de valores	Yoon y col., 1993
	Insolvencia de entidades aseguradoras	Brockett y col., 1994
Medicina y farmacología	Modelos de admisión de crédito	Desai y col., 1996; Lee y Jung, 2000; Limsombunchai y col., 2005
	Diagnóstico de cardiopatías	Fujita y col., 1992
	Detección de tumores cancerígenos mediante análisis de imágenes	Fogel y col., 1998
	Detección de patologías tiroideas	Zhang y Berardi, 1998
	Predicción de resistencia a antibióticos	Shang y col., 2000
	Predicción de reincidencia de ataques cerebrovasculares	Ottenbacher y col., 2001
	Predicción de efectos adversos derivados de infecciones bacterianas	Nguyen y col., 2002
Ingeniería y producción	Estimación de supervivencia a cáncer de pecho	Denlen y col., 2005
	Predicción en estudios epidemiológicos	Gaudart y col., 2004; Ottenbacher y col., 2004
	Control de procesos industriales	Willis y col., 1991; Meireles y col., 2003
	Optimización de procesos y control de calidad	Nascimento y col., 2000; Uraikul y col., 2007

Tabla 2.4. Aplicaciones de las ANNs (Continuación)

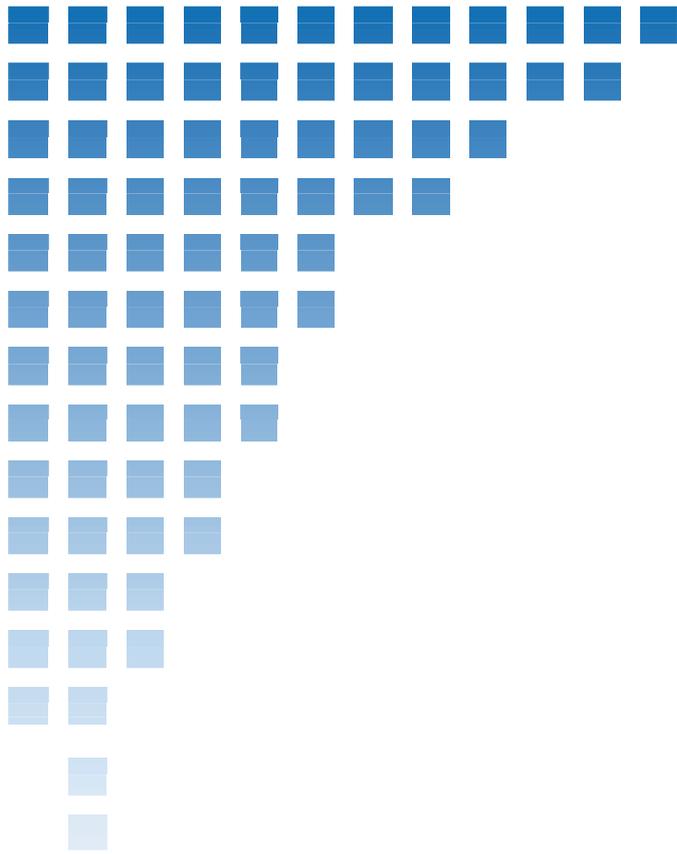
Campo	Descripción	Referencia
	<i>Gestión de residuos</i>	
	Predicción de generación de residuos sólidos urbanos	Zade y Noori, 2008; Noori y col., 2009
	Predicción de generación de residuos médicos	Jahandideh y col., 2009
	Predicción de la composición de residuos sólidos	Dong y col., 2003; Chen y col., 2008; Srivastava y Nema, 2008; Bayar y col., 2009
	Predicción de la lixiviación de metales en residuos estabilizados	Irabien y col., 2002
	Predicción de las propiedades mecánicas de sistemas cenizas volantes/cemento	Sebastiá y col., 2003
	<i>Tratamientos de aguas residuales</i>	
	Tratamiento de efluentes	Al-Mutairi y col., 2004
	<i>Evaluación de contaminación de sedimentos</i>	
	Clasificación de la calidad de sedimentos	Álvarez-Guerra y col., 2008
	Predicción de la calidad de sedimentos	Álvarez-Guerra y col., 2010
	<i>Contaminación atmosférica</i>	
	Clasificación del origen de episodios de ozono	Fontes y col., 2014
	Predicción de respuesta de mercado	Hruschka, 1993; Dasgupta y col., 1994
	Predicción de frecuencia de compra	Dutta y col., 1994
	Predicción de porcentaje de ventas	Ainslie y Dreze, 1996
	Segmentación de mercado	Fish y col., 1995
	Toma de decisiones en el diseño de nuevos productos	Thieme y col., 2000
Marketing		

2.3. Bibliografía

- Abdi, H., Williams, L.J., 2013. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. En: Reisfeld, B., Mayeno, A.N. (Eds.) *Computational Toxicology: Volume II, Methods in Molecular Biology*, vol. 930, Springer Science + Business Media.
- Ainslie, A., Dreze, X., 1996. Data-mining and choice classic models: neural networks. *Decisions Marketing* 7, 77–86.
- Alaiya, A.A., Franzén, B., Hagman, A., Silfverswärd, C., Moberger, B., Linder, S., Auer, G., 2000. Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. *International Journal of Cancer* 86, 731–736.
- Al-Mutairi, N., Kartam, N., Koushki, P., Al-Mutaire, M., 2004. Modeling and predicting biological performance of contact stabilization process using artificial neural networks. *Journal of Computing in Civil Engineering* 18, 341–349.
- Alsberg, B., Woodward, A., Winson, M., Rowland, J., Kell, D., 1998. Variable selection in wavelet regression models. *Analytica Chimica Acta* 368, 29–44.
- Álvarez-Guerra, M., Ballabio, D., Amigo, J.M., Bro, R., Viguri, J.R., 2010. Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks. *Environmental Pollution* 158, 607–614.
- Álvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., Viguri, J.R., 2008. Assessment of Self-Organizing Map artificial neural networks for the classification of sediment quality. *Environmental International* 34, 782–790.
- Alves, C., Pinheiro, J., Camargo, A., Ferreira, M., Romero, R., Da Silva, A., 2001. A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *Journal of Molecular Structure: THEOCHEM* 541, 81–88.
- Aulinger, A., Einax, J.W., Prange, A., 2004. Setup and optimization of a PLS regression model for predicting element contents in river sediments. *Chemometrics and Intelligent Laboratory Systems* 72, 35–41.
- Bayar, S., Demir, I., Onkal Engin, G., 2009. Modeling leaching behavior of solidified wastes using back-propagation neural networks. *Ecotoxicology and Environmental Safety* 72, 843–850.
- Bayol, M.P., de la Foye, A., Tellier, C., Tenenhaus, M., 2000. Use of PLS path modelling to estimate the European Consumer Satisfaction Index (ECSI) model. *Statistica Applicata - Italian Journal of Applied Statistics* 12, 361–375.
- Berglund, A., Wold, S., 1997. INLR, implicit non-linear latent variable regression. *Journal of Chemometrics* 11, 141–156.
- Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Singapur: Springer Science + Business Media.
- Bishop, C. M., 1995. *Neural Networks for Pattern Recognition and Machine Learning*. Oxford: Clarendon Press.
- Boulesteix, A.L., 2004. PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology* 3, 1–30.
- Braak, C.J.F., Jong, S., 1998. The objective function of partial least squares. *Journal of Chemometrics* 12, 41–54.
- Brindle, J.T., Antti, H., Holmes, E., Tranter, G., Nicholson, J.K., Bethell, H.W., Clarke, S., Schofield, P.M., McKilligin, E., Mosedale, D.E., Grainger, D.J., 2002. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ¹H-NMR-based metabonomics. *Nature Medicine* 8, 1439–1445.
- Brockett, P.L., Cooper, W.W., Golden, L.L., Pitaktong, U., 1994. A neural network method for obtaining an early warning of insurer insolvency. *The Journal of Risk and Insurance* 61(3), 402–424.
- Charlton, A.J., Farrington, W.H., Brereton, P., 2002. Application of ¹H NMR and multivariate statistics for screening complex mixtures: quality control and authenticity of instant coffee. *Journal of Agricultural and Food Chemistry* 50, 3098–3103.
- Chen, L.J., Cui, L.Y., Xing, L., Han, L.J., 2008. Prediction of the nutrient content in dairy manure using artificial neural network modeling. *Journal of Dairy Science* 91, 4822–4829.
- Clementi, M., Clementi, S., Cruciani, G., Pastor, M., Davis, A.M., Flower, D.R., 1997. Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Engineering* 10, 747–749.
- Cozzolino, D., Cynkar, W.U., Damberg, R.G., Mercurio, M.D., Smith, P.A., 2008. Measurement of condensed tannins and dry matter in red grape homogenates using near infrared spectroscopy and partial least squares. *Journal of Agricultural and Food Chemistry* 56, 7631–7636.
- Dasgupta, C.G., Dispensa, G.S., Ghose, S., 1994. Comparing the predictive performance of a neural network model with some traditional market response models. *International Journal of Forecasting* 10, 235–244.
- Delalieux, S., Van Aardt, J., Keulemans, W., Schrevels, E., Coppin, P., 2007. Detection of biotic stress (*Venturia inaequalis*) in apple trees using hyperspectral data: non-parametric statistical approaches and physiological implications. *European Journal of Agronomy* 27, 130–143.
- Delen, D., Walker, G., Kadam, A., 2005. Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine* 34(2), 113–127.
- Desai, V.S., Crook, J.N., Overstreet, G.A.Jr., 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research* 95(1), 24–37, 22.
- Dong, C., Jin, B., Li, D., 2003. Predicting the heating value of MSW with a feed-forward neural network. *Waste Management* 23, 103–106.
- Dutta, S., Shekhar, S., Wong, W.Y., 1994. Decision support in non-conservative domains: Generalization with neural networks. *Decision Support Systems* 11, 527–544.
- Fanning, K., Cogger, K. O., Srivastava, R., 1995. Detection of management fraud: A neural network approach. *Intelligent Systems in Accounting Finance and Management* 4, 113–126.
- Filzmoser, P., Gschwandtner, M., Todorov, V., 2012. Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics* 26, 42–51.

- Fish, K.E., Barnes, J.H., Aiken, M.W., 1995. Artificial neural networks – A new methodology for industrial market segmentation. *Industrial Marketing Management* 24, 431–438.
- Flórez, R., Fernández, J.M., 2008. Las redes neuronales artificiales. Fundamentos teóricos y aplicaciones prácticas. La Coruña: Netbiblo.
- Fogel, D.B., Wasson, E.C., Boughton, E.M., Porto, V.W., Angeline, P.J., 1998. Linear and neural models for Classifying Breast Masses. *IEEE Transactions on Medical Imaging* 17(3), 485–488.
- Fontes, T., Silva, L.M., Silva, M.P., Barros, N., Carvalho, A.C., 2014. Can artificial neural networks be used to predict the origin of ozone episodes? *Science of the Total Environment* 488–489, 197–207.
- Fujita, H., Katafuchi, T., Uehara, T., Nishimura, T., 1992. Application of artificial neural network to computer-aided diagnosis of coronary artery disease in myocardial spect Bull's-eye images. *Journal of Nuclear Medicine* 33(2), 272–276.
- Gaudart, J., Giusiano, B., Huiart, L., 2004. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. *Computational Statistics and Data Analysis* 44, 547–570.
- Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 185, 1–17.
- Gutkin, M., Shamir, R., Dror, G., 2009. SlimPLS: a method for feature selection in gene expression-based disease classification. *PLoS ONE* 4, 416–422.
- Haykin, S., 2009. *Neural Networks and Learning Machines* (3rd Edition). New Jersey: Prentice Hall (Pearson).
- Hassoun, M. H., 1995. *Fundamentals of Artificial Neural Networks*. Londres: MIT Press.
- Hidajat, K., Chong, S., 2000. Quality characterisation of crude oils by partial least square calibration of NIR spectral profiles. *Journal of Near Infrared Spectroscopy* 8, 53–60.
- Höskuldsson, A., 1996a. Experimental design and priority PLS regression. *Journal of Chemometrics* 10(5-6), 637–668.
- Höskuldsson, A., 1996b. *Prediction Methods in Science and Technology*, Vol. 1. Copenhagen: Thor Publishing.
- Hruschka, H., 1993. Determining market response functions by neural network modelling: A comparison to econometric techniques. *European Journal of Operational Research* 66, 27–35.
- Huang, S.C., Wu, T.K., 2008. Forecasting stock indices with wavelet-based kernel partial least square regressions. *IEEE International Joint Conference on Neural Networks* 1, 1910–1916.
- Hubert, M., Branden, K.V., 2003. Robust methods for partial least squares regression. *Journal of Chemometrics* 17, 537–549.
- Indahl, U., 2005. A twist to partial least squares regression. *Journal of Chemometrics* 19, 32–44.
- Irbien, A., Fernandez Olmo, I., Andres, A., Sebastián, M., 2002. Prediction of TCLP leachates of electric arc furnace dust/cement products using neural network analysis. *Environmental Progress* 21(2), 95–104.
- Jahandideh, S., Jahandideh, S., Asadabadi, E.B., Askarian, M., Movahedi, M., Mehdi, M., Hos-seini, S., Jahandideh, M., 2009. The use of artificial neural networks and multiple linear regression to predict rate of medical waste generation. *Waste Management* 29, 2874–2879.
- Johansson, D., Lindgren, P., Berglund, A., 2003. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics* 19, 467–473.
- Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 18, 251–263.
- Lee, T.H., Jung, S., 2000. Forecasting creditworthiness: Logistic vs. artificial neural net. *The Journal of Business Forecasting Methods and Systems* 18(4), 28–30.
- Lenard, M.J., Alam, P., Madey, G.R., 1995. The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences* 26(2), 209–227.
- Lennox, B., Wang, X., Kruger, U., 2003. Recursive partial least squares algorithms for monitoring complex industrial processes. *Control Engineering Practice* 11, 613–632.
- Limsombunchai, V., Gan, C., Lee, M., 2005. An analysis of credit scoring for agricultural loans in Thailand. *American Journal of Applied Sciences* 2(8), 1198–1205.
- Lorber, A., Wangen, L.E., Kowalski, B.R., 1987. A theoretical foundation for the PLS algorithm. *Journal of Chemometrics* 1(1), 19–31.
- Martens, H., Naes, T., 1989. *Multivariate Calibration*. London: Wiley.
- Martos, N.R., Díaz, A.M., Navalón, A., Payá, I.D.O., Vallvey, L.C., 2000. Simultaneous spectrofluorimetric determination of (acetyl) salicylic acid, codeine and pyridoxine in pharmaceutical preparations using partial least squares multivariate calibration. *Journal of Pharmaceutical and Biomedical Analysis* 23, 837–844.
- Meireles, M.R.G., Almeida, P.E.M., Simões, M.G., 2003. A comprehensive review for industrial applicability of artificial neural networks. *IEEE Transactions on Industrial Electronics* 50(3), 585–601.
- Nascimento, C.A.O., Giudici, R., Guardani, R., 2000. Neural network based approach for optimization of industrial chemical processes. *Computers and Chemical Engineering* 24(9-10), 2303–2314.
- Nguyen, D.V., Roche, D.M., 2002. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18, 1216–1226.
- Nguyen, T., Malley, R., Inkelis, S.H., Kuppermann, N., 2002. Comparison of prediction models for adverse outcome in pediatric meningococcal disease using artificial neural network and logistic regression analyses. *Journal of Clinical Epidemiology* 55(7), 687–695.
- Niu, X., Shen, F., Yu, Y., Yan, Z., Xu, K., Yu, H., Ying, Y., 2008. Analysis of sugars in Chinese rice wine by Fourier transform near-infrared spectroscopy with partial least-squares regression. *Journal of Agricultural and Food Chemistry* 56, 7271–7278.
- Noori, R., Abdoli, M.A., Farokhnia, A., Abbasi, M., 2009. Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. *Expert Systems with Applications: An International Journal* 36, 9991–9999.

- O'Toole, A.J., Abdi, H., Jiang, F., Phillips, P.J., 2007. Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 1149–1155.
- Odom, M., Sharda, R., 1990. A neural network model for bankruptcy prediction. En: *Proceedings of the international joint conference on neural networks Vol. 2*, IEEE Press, Alamitos, CA, 163–168.
- Ottenbacher, K.J., Smith, P.M., Illig, S.B., Linn, R.T., Mancuso, M., Granger, C.V., 2004. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. *Annals of Epidemiology* 14(8), 551–559.
- Ottenbacher, K.J., Smith, P.M., Illig, S.B., Linn, R.T., Fiedler, R.C., Granger, C.V., 2001. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology* 54(11), 1159–1165.
- Otto, M., Wegscheider, W., 1985. Spectrophotometric multicomponent analysis applied to trace metal determinations. *Analytical Chemistry* 57(1), 63–69.
- Pendharkar, P.C., 2005. A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers and Operations Research* 32, 2561–2582.
- Rosipal, R., Trejo, L.J., Matthews, B., 2003. Kernel PLS-SVC for linear and nonlinear classification. *International Conference on Machine Learning* 1, 640–647.
- Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S., 2009. Human detection using partial least squares analysis. *IEE Computer Visualization* 1, 24–31.
- Sebastiá, M., Fernández Olmo, I., Irabien, A., 2003. Neural network prediction of unconfined compressive strength of coal fly ash-cement mixtures. *Cement and Concrete Research* 33, 1137–1146.
- Serrano-Cinca, C., Gutiérrez-Nieto, B., 2013. Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems* 54, 1245–1255.
- Serranti, S., Gargiulo, A., Bonifazi, G., 2012. Hyperspectral imaging for process and quality control in recycling plants of polyolefin flakes. *Journal of Near Infrared Spectroscopy* 20, 573–545.
- Shang, J.S., Lin, Y.E., Goetz, A.M., 2000. Diagnosis of MRSA with neural networks and logistic regression approach. *Health Care Management Science* 3(4), 287–297.
- Sinelli, N., Casiraghi, E., Tura, D., Downey, G., 2008. Characterisation and classification of Italian virgin olive oils by near-and mid-infrared spectroscopy. *Journal of Near Infrared Spectroscopy* 16, 335–342.
- Srivastava, A.K., Nema, A. K., 2008. Forecasting of solid waste composition using fuzzy regression approach: a case of Delhi. *International Journal of Environment and Waste Management* 2, 65–74.
- Thieme, R.J., Song, M., Calantone, R.J., 2000. Artificial neural network decision support systems for new product development project selection. *Journal of Marketing research* 37(4), 499–507.
- Uraikul, V., Chan, C.W., Tontiwachwuthikul, P., 2007. Artificial intelligence for monitoring and supervisory control of process systems. *Engineering Applications of Artificial Intelligence* 20(2), 115–131.
- Willis, M.J., Di Massimo, C., Montague, G.A., Tham, M.T., Morris, A.J., 1991. Artificial neural networks in process engineering. *IEE Proceedings S: Control Theory and Applications* 138(3), 256–266.
- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58, 109–130.
- Wold, S., Geladi, P., Esbensen, K., Öhman, J., 1987. Multi-way principal components and PLS analysis. *Journal of Chemometrics* 1(1), 41–56.
- Wold, S., Rube, A., Wold, H., Dunn, W.J., 1984. The collinearity problem in linear regression: the partial least squares approach to generalized inverses. *SIAM Journal of Scientific and Statistical Computing* 5, 735–743.
- Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. *Conference Proceeding, Matrix, Pencils, Piteå, Suecia*, 286–293.
- Wu, B., 2005. Predictive model building for microarray data using generalized partial least squares model. Technical report 2005-023, Division of Biostatistics, University of Minnesota.
- Yoon, Y., Swales, G.Jr., Margavio, T.M., 1993. A comparison of discriminant analysis versus artificial neural networks. *The Journal of the Operational Research Society* 44(1), 51–60.
- Zade, J.G., Noori, R., 2008. Prediction of municipal solid waste generation by use of artificial neural network: A case study of Mashhad. *International Journal of Environmental Research* 2, 13–22.
- Zamprogna, E., Barolo, M., Seborg, D.E., 2004. Estimating product composition profiles in batch distillation via partial least squares regression. *Control Engineering Practice* 2, 917–929.
- Zhang, Y., Zhang, Y., 2009. Complex process monitoring using modified partial least squares method of independent component regression. *Chemometrics and Intelligent Laboratory Systems* 98, 143–148.
- Zhang, G., Berardi, V., 1998. An investigation of neural networks in thyroid function diagnosis. *Health Care Management Science* 1, 29–37.



Capítulo 3. Metodología

3.1. Descripción de las áreas de estudio

Para la realización de este trabajo se han seleccionado tres áreas urbanas de la Comunidad Autónoma de Cantabria cuya calidad del aire puede estar influenciada por la presencia en sus proximidades de actividades industriales, intentando cubrir las distintas partes de la región: litoral (Castro Urdiales), interior (Reinosa) y el núcleo urbano de mayor población (Santander).

Castro Urdiales es una localidad situada en el litoral de Cantabria, al NE. Constituye la principal zona urbana y capital del municipio del mismo nombre, que comprende un área de aproximadamente 97 km², con algo más de 30000 habitantes de media en el periodo 2008-2010 (INE, 2016). La contaminación atmosférica en esta zona tiene un origen antropogénico claro causado por tráfico, no en vano Castro Urdiales está rodeado por una de las autovías nacionales de mayor intensidad de tráfico en el norte de España, la A-8: en particular, en el punto kilométrico 146,1 próximo a Castro Urdiales, la intensidad media diaria (IMD) fue de 46311 vehículos en 2010 (Ministerio de Fomento, 2011). Asimismo, la contaminación atmosférica también procede de actividades industriales como plantas químicas y metalúrgicas y una refinería de petróleo, ubicadas a unos 10-30 km al SE (cerca de la ciudad de Bilbao). La estación de muestreo se encuentra ubicada en las coordenadas 43°22'53"N, 3°13'22"O y a unos 20 m sobre el nivel del mar, en el núcleo del área urbana.

Reinosa es una localidad que comprende un área de aproximadamente 4 km², con una población de aproximadamente 10000 habitantes de media en el periodo 2008-2010 (INE, 2016). Dicha ciudad está ubicada en el interior, a unos 50 km de la costa, en la parte sur de la Región de Cantabria. El lugar de muestreo, ubicado en las coordenadas 43°00'01"N, 4°08'13"O a unos 850 m sobre el nivel del mar, se encuentra en las proximidades de una planta de fabricación de acero, así como de una autovía nacional, la A-67, principal ruta de salida de Cantabria, que establece conexión con la parte central de la Península Ibérica: en particular, en el punto kilométrico 135,3 próximo a Reinosa el IMD en 2010 fue de 9145 vehículos (Ministerio de Fomento, 2011).

Santander es la capital de la región de Cantabria, con aproximadamente 180000 habitantes de media en el periodo 2008-2011 (INE, 2016). La ciudad se extiende a lo largo de su bahía y a unos 5-10 km, en la periferia, se encuentra un área industrial relacionada entre otros con la producción de acero y ferroaleaciones, que convierte a Santander en el área urbana más compleja de la región. El punto de muestreo se localizó en el tejado del edificio de la Escuela de Ingenieros Industriales y de Telecomunicación (ETSIT) de la Universidad de Cantabria, a 23 m sobre el nivel del mar, con coordenadas 43°28'24"N, 3°47'54"O, aproximadamente a 1 km de la costa.

La Figura 3.1. muestra la localización de los puntos de muestreo y de las áreas industriales de mayor relevancia más próximas a los mismos en las tres áreas urbanas consideradas.

Por otra parte, para comprobar la aplicabilidad de las herramientas propuestas en este trabajo en otros lugares, se seleccionó un área urbana en Francia, la ciudad de Dunkerque, en vista de las similitudes existentes con la ciudad de Santander, en cuanto a la localización geográfica (ciudad costera de población similar altamente influenciada por el aerosol marino) y presencia cercana de actividad industrial común (acerías y fabricación de ferroaleaciones de manganeso). Además, un motivo adicional para la selección de esta área urbana radicó en el hecho de contar con datos relativos a campañas de muestreo llevadas a cabo de forma intensiva, contando con un gran número de muestras en un corto periodo de tiempo, lo que permitiría evaluar la efectividad de las herramientas para estos casos.

La ciudad de Dunkerque, con una población de aproximadamente 68000 habitantes en 2008, se encuentra ubicada en la costa noreste de Francia, junto a la frontera con Bélgica. El núcleo urbano está rodeado en su parte norte por el puerto de Dunkerque, que ha sido clasificado como el tercer puerto más importante de Francia debido al transporte marítimo y de mercancías (que incluyen minerales, carbón, cobre y otros bienes) y como el séptimo en orden de importancia en el norte de Europa. La ciudad se encuentra asimismo muy próxima al Canal de la Mancha, que conecta el Mar del Norte con el Océano Atlántico, y que constituye la vía marítima más transitada, con aproximadamente 500 embarcaciones al día. También cuenta con un área altamente industrializada en los alrededores de la ciudad, donde destaca la industria siderúrgica puesto que cuenta con una planta siderúrgica integral (a unos 4 km NO) y una no integral, de horno de arco eléctrico, (a unos 6 km NE) y una planta de producción de aleaciones de ferromanganeso a unos 6 km O.

La Figura 3.2. muestra la localización del lugar de muestreo ($51^{\circ}02'07''\text{N}$, $02^{\circ}22'05''\text{E}$ y aproximadamente a 10 m sobre el nivel del mar), que se encontraba ubicado en el tejado del edificio Les Darses en la línea de separación entre la zona industrial de la ciudad, de tal manera que durante la campaña de muestreo fuesen registrados tanto las emisiones urbanas como industriales (correspondientes a los sectores OSO y NNO).



Figura 3.1. Localización de los lugares de muestreo en Cantabria (España).

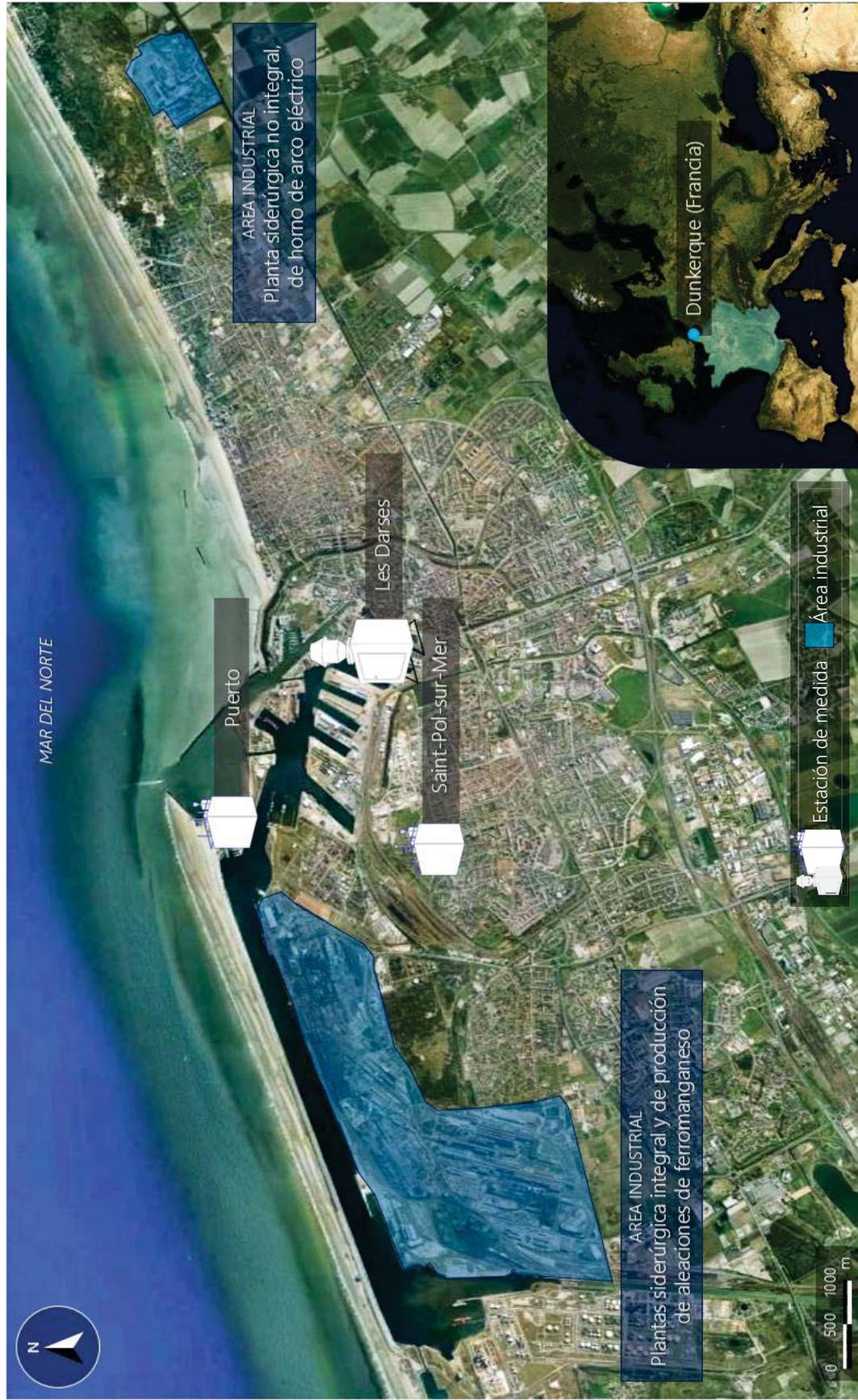


Figura 3.2. Localización del lugar de muestreo en Dunkerque (Francia).

3.2. Bases de datos ambientales: variables de influencia en niveles en inmisión

3.2.1. Contaminantes mayoritarios

Dentro de los contaminantes presentes de forma mayoritaria en la atmósfera se encuentran contaminantes gaseosos como el dióxido de azufre, el ozono troposférico y los óxidos de nitrógeno, así como otros no gaseosos como el material particulado.

Los niveles de estos contaminantes tienen una gran influencia sobre los niveles de microcontaminantes en inmisión porque se generan por el desarrollo de actividades antropogénicas variadas como la industria, el tráfico rodado o la combustión doméstica y reflejan el impacto e incidencia de las mismas. Indudablemente, la relación más directa es la que se establece entre los niveles de material particulado y los metales y semimetales, ya que estos contaminantes están presentes mayoritariamente en la fracción particulada.

3.2.2. Variables meteorológicas

Los procesos meteorológicos, responsables de los movimientos de las masas de aire, tanto a nivel superficial como en altura afectan indiscutiblemente al destino de los contaminantes en la atmósfera, de tal manera, que si el aire se encuentra en condiciones de calma los contaminantes presentan dificultades para dispersarse correctamente y, en consecuencia, su concentración en el aire ambiente se incrementa, mientras que cuando se producen fuertes turbulencias, los contaminantes se dispersan rápidamente, favoreciendo la reducción de sus niveles en inmisión.

Las observaciones meteorológicas de superficie (a nivel de suelo) se componen generalmente de datos cuantitativos en relación a las siguientes variables: temperatura de bulbo seco, temperatura de bulbo húmedo, que junto con la anterior permite determinar el contenido de vapor de agua o humedad, dirección y velocidad del viento, presión, evaporación, radiación solar y cantidad de precipitación. No obstante, la región de la atmósfera que gobierna el transporte y dispersión de los contaminantes, la capa límite planetaria (CLP), comprende aproximadamente los primeros 1000 m, por lo que también se efectúan observaciones en estratos superiores. La observación más común es el radiosondeo, que consiste en un globo lleno de un gas ligero (hidrógeno o helio) que asciende verticalmente al mismo tiempo que es arrastrado horizontalmente por el viento. Este procedimiento permite determinar el viento a distintas alturas, así como la distribución vertical de temperatura y humedad en la atmósfera, de tal manera que quedan registrados tanto el transporte convectivo horizontal, que depende

fundamentalmente del viento, como el transporte convectivo vertical, que depende de la estabilidad atmosférica y del fenómeno de inversión térmica.

En base a lo anteriormente expuesto, cuando se estudia la calidad del aire, es importante tener en cuenta, entre otros, los siguientes factores meteorológicos de gran influencia sobre los procesos que tienen lugar en la atmósfera y que condicionan la dispersión de los contaminantes:

3.2.2.1. *El viento*

El viento es una magnitud vectorial caracterizada por dos valores que representan su dirección y velocidad, respectivamente. La dirección y velocidad de viento son dos variables de gran influencia en los modelos estadísticos. Desafortunadamente, presentan una gran variabilidad puesto que el viento en superficie raramente es constante durante un periodo de tiempo determinado, sino que, en general, varía rápidamente, siendo estas variaciones irregulares tanto en velocidad como en dirección. Por esta razón, hay una gran dificultad para obtener un valor promedio representativo. El método de estimación de la velocidad y dirección de viento promedio tiene asociada una gran incertidumbre debido a la variabilidad propia de dichas variables meteorológicas. Distintos métodos de determinación del valor promedio conducen a resultados diferentes, que afectan positiva o negativamente al desempeño del modelo según se ajusten más o menos a la realidad.

3.2.2.2. *Temperatura y radiación solar*

La temperatura y la radiación solar desempeñan un papel importante en las reacciones químicas que tienen lugar en la atmósfera para formar *smog* fotoquímico a partir de otros contaminantes. Además, algunos elementos como el mercurio o el arsénico son muy volátiles y se ven afectados por la temperatura: el equilibrio vapor-aerosol depende de la temperatura, viéndose favorecida la fase aerosol a bajas temperaturas y la fase vapor a altas temperaturas, por lo que su concentración depende inversamente de la temperatura (Pallarés y col. 2007). Por otra parte, la emisión de metales y semimetales a la atmósfera también puede aumentar en los periodos fríos como consecuencia de los procesos de combustión para calefacción residencial.

3.2.2.3. *Humedad relativa*

La humedad favorece la acumulación de contaminantes. Además, al igual que la temperatura y la radiación solar, el vapor de agua juega un papel importante en muchas reacciones que tienen lugar en la atmósfera. Las moléculas de agua, por ser pequeñas y muy polares, se pueden unir a muchas sustancias. Por tanto, la humedad puede conducir a la formación de contaminantes secundarios y a la formación de lluvia ácida previa reacción con compuestos de azufre y nitrógeno.

3.2.2.4. Precipitación

La lluvia tiene un efecto purificador favoreciendo la disolución de los contaminantes gaseosos y el arrastre de la materia particulada de la atmósfera para su deposición sobre suelos, aguas, vegetación y otras estructuras no bióticas, fenómeno conocido como deposición húmeda. En algunos estudios de deposición atmosférica, se ha observado la existencia de una buena correlación entre la cantidad de precipitación y los flujos de deposición húmeda de metales y semimetales, la cual constituye uno de los destinos más importantes de los metales presentes en las partículas (Tate y Bates, 1984; Sakata y col., 2006; Okubo y col., 2013). Por tanto, en general, existe una clara relación entre la ocurrencia de precipitaciones y la calidad del aire ambiente con respecto a este tipo de contaminantes.

3.2.3. Variables meteorológicas relacionadas con la estabilidad atmosférica

La estabilidad atmosférica, descrita en detalle en Hanna y col. (1982), se define en términos de la tendencia que presenta una hipotética masa de aire a realizar un movimiento ascendente o descendente tras haber sido sometida a un desplazamiento vertical (Seinfeld y Pandis, 2006). Por tanto, conceptualmente, es una expresión de la tendencia que presenta la capa superficial de la atmósfera para promover o impedir el movimiento vertical del aire por convección. Dicha tendencia está controlada principalmente por los gradientes verticales de la densidad del aire, que son función esencialmente de la temperatura del aire (Hernández-Ramírez y col., 2011). En consecuencia, el concepto de estabilidad atmosférica es útil para determinar si los contaminantes atmosféricos se elevarán y diluirán o si por el contrario se acumularán y concentrarán próximos al nivel del suelo: una atmósfera inestable favorece el movimiento vertical de los contaminantes y, por tanto, menores concentraciones en inmisión, puesto que los contaminantes se dispersan en un volumen de aire mucho mayor; mientras que una atmósfera estable impide el movimiento vertical de los contaminantes y conduce a unos niveles de concentración más elevados.

En la literatura se encuentran propuestas de indicadores de uso común como medida de la estabilidad atmosférica, entre los que se encuentran la inversión térmica, la capa estable, la altura de la capa de mezcla o el rango de temperatura diaria (este último, en caso de ausencia de datos de temperatura vertical, actúa como medida alternativa en relación a la posibilidad de desarrollo de inversión térmica). Sin embargo, no existe consenso en cuanto a cuáles resultan idóneos o más recomendables, en especial cuando se trata de incorporar dichos indicadores a modelos estadísticos de estimación.

En base a la disponibilidad de datos procedentes de sondeos termodinámicos, en este trabajo la estabilidad atmosférica ha sido evaluada de forma cuantitativa, por medio de la altura de la capa de mezcla (m) y del número de Richardson (-); y de forma cualitativa, a través de la

inversión térmica y de las clases de estabilidad de Pasquill-Gifford. Dichas variables, así como su procedimiento de cálculo, se describen en los siguientes apartados.

3.2.3.1. *Altura de la capa de mezcla*

Se denomina capa de mezcla o capa límite atmosférica a la capa de aire adyacente al suelo sobre la cual los contaminantes o cualquier constituyente emitido dentro de esta capa se dispersa verticalmente por convección o turbulencia mecánica dentro de una escala temporal de alrededor de una hora (Seibert y col., 1997), a través de la cual los contaminantes liberados desde la superficie terrestre son transportados y difundidos a través de la atmósfera. En dicho estrato las sustancias contaminantes emitidas se dispersan gradualmente en dirección horizontal y vertical por acción de la turbulencia, mezclándose finalmente si permanecen el tiempo suficiente y no hay ningún sumidero significativo, transformaciones químicas u otros fenómenos meteorológicos que perturben el proceso de homogeneización. Por consiguiente, la altura de esta capa constituye un parámetro fundamental que caracteriza la estructura de la baja troposfera. Asimismo, existe una estrecha relación entre los índices de contaminación atmosférica y el espesor de este estrato del aire. No obstante, es conveniente destacar la diferencia entre el espesor de la capa de mezcla y el espesor real del aire que puede denominarse estrato contaminado. La capa de mezcla es considerada como el volumen total en el que una masa de aire contaminado puede ser potencialmente dispersada y transportada. El estrato en el que se mezclan los contaminantes coincidirá con el espesor límite de la capa de mezcla cuando haya transcurrido el tiempo suficiente para que se produzca la dispersión vertical y horizontal completa de dicha contaminación (Seibert y col., 1997).

La capa límite en la atmósfera presenta una escala temporal característica. Dicha escala recoge la importante dependencia de la altura de la capa de mezcla con la actividad solar de forma que su crecimiento, desarrollo y decrecimiento están condicionados por la aportación energética del sol, lo que se manifiesta en una clara componente diurna. Esta dependencia temporal marca una enorme diferencia respecto de la capa límite en condiciones mecánicas, ya que se traduce en un comportamiento dinámico y variable del espesor del aire, condicionado por el proceso de convección. De esta forma, la capa límite presenta un carácter nocturno y diurno claramente diferenciados. Mientras que durante la noche, normalmente la capa límite viene definida por el estrato estable representado por la inversión radiativa superficial, durante el día, la actividad turbulenta provoca el desarrollo de la capa de mezcla.

Puesto que la altura de la capa de mezcla determina el volumen total disponible para la dispersión de las sustancias contaminantes, se trata de un parámetro de gran importancia para los modelos de contaminación atmosférica. La altura de la capa de mezcla es un parámetro clave en los modelos de simulación del transporte y difusión turbulenta de los contaminantes en la atmósfera (Seibert y col., 1997), incidiendo en la determinación del dominio turbulento, en

las características de los perfiles verticales de turbulencia y en los cálculos de los perfiles de concentración con límite superior reflectivo (Seibert y col., 1997). Por este motivo, muchos métodos y modelos predictivos y de diagnóstico para evaluar las concentraciones de los contaminantes tienen en consideración este parámetro.

Típicamente, las alturas de la capa de mezcla matutinas varía entre 300 y 900 m, mientras que las alturas de la capa de mezcla de la tarde varía entre 600 y 4000 m en verano y entre 600 y 1400 m en invierno. La proximidad a grandes masas de agua también influye en el espesor de esta capa.

A pesar de su gran importancia e influencia, se trata de un parámetro cuyo cálculo a menudo no resulta sencillo puesto que no se mide con las prácticas meteorológicas habituales. En la literatura se encuentran distintos métodos alternativos para la estimación o determinación de la altura de la capa de mezcla, cuya aplicación depende de la disponibilidad de datos meteorológicos. Uno de los métodos más simples consiste en la estimación de la altura de la capa de mezcla a partir de la estabilidad atmosférica en función de las categorías de estabilidad atmosférica de Pasquill (Spadaro, 1999), como se muestra en la Tabla 3.1.

Tabla 3.1. Criterio de estimación de la altura de la capa de mezcla a partir de las categorías de estabilidad atmosférica de Pasquill (Spadaro, 1999).

	Velocidad del viento (m/s)	Clase de estabilidad atmosférica de Pasquill	Altura de la capa de mezcla (m)
Día de alta insolación*	0-2	A (extremadamente inestable)	1600
	2-3	B (moderadamente inestable)	1200
	3-5	C (ligeramente inestable)	800
	> 5	D (neutra)	560
Día de baja insolación	0-2	B	200
	2-3	C	800
	3-5	D	560
	> 5	D	560
Noche	0-2	F (moderadamente estable)	200
	2-3	F	200
	3-5	E (ligeramente estable)	320
	> 5	D	560
	Nublado	D	560

*Se considera insolación alta en las horas del día cuando $\alpha > 45^\circ$

Por otra parte, la estimación de la altura de la capa de mezcla también se puede llevar a cabo mediante procedimientos de parametrización, basados en distintas formulaciones de la altura de la capa de mezcla (h) en función de la velocidad de fricción (u^*), la longitud de Monin-Obukhov (L) y el parámetro de Coriolis (f). Algunas de estas formulaciones se recogen en la Tabla 3.2. Sin embargo, en un estudio llevado a cabo por Burzynski y col. (2004) en el que se comparan datos experimentales de SODAR con formulaciones de este tipo propuestas por

diferentes autores se obtuvieron coeficientes de correlación muy bajos, inferiores a 0,3, lo que indica que quizá no constituyan el procedimiento más adecuado para una estimación precisa de la altura de la capa de mezcla.

Tabla 3.2. Diagnóstico de algunas ecuaciones desarrolladas para el cálculo de la altura de la capa de mezcla en función de la clase de estabilidad que presenta la atmósfera. Tabla adaptada de Burzynski y col. (2004).

Referencia	Ecuación de altura de capa de mezcla	Coefficientes	Estabilidad
Deardorff, 1972	$h = \left[\frac{1}{30L} + \frac{f}{0,4 u^*} \right]^{-1}$		D-F
Zilitinkevich, 1972	$h = c_2 \left(\frac{u^* L}{f} \right)^{\frac{1}{2}}$	$c_2 \approx 0,4 (0,13 \div 0,72)$	E,F
Zilitinkevich, 1972	$h = \alpha \frac{u^*}{f}$	$\alpha \in (0,07; 0,3)$	D
Benkley y Schulman, 1979	$h = 125 u_{10}$		D-F
Arya, 1981	$h = 0,089 \frac{u^*}{f} + 85,1$		D-F
Arya, 1981	$h = a \left(\frac{u^* L}{f} \right)^{\frac{1}{2}} + b$	$a = 0,43; b = 29,3$	D-F
Mahrt y col., 1982	$h = 0,06 \frac{u^*}{f}$		D-F
Nieuwstadt, 1984	$h = 28 u_{10}^{\frac{3}{2}}$		D-F
Batchvarova y Gryning, 1991	$\frac{dh}{dt} = (1+2A) \frac{S}{\rho C_p \gamma_{\theta} h} + 2B \frac{u^{*3} T}{g \gamma_{\theta} h^2}$		A-C
Middleton y Thomson, 2001	$h = \left(\frac{u^* L}{f} \right)^{\alpha}$	$\alpha \in (0,3; 0,7)$	E,F

No obstante, la metodología más destacable y de aplicación más extendida es la propuesta por Holzworth (1967). Por ello, en este trabajo se ha utilizado el método de Holzworth para la estimación de la altura de la capa de mezcla. Se trata de un método gráfico en el que a partir de la información de un radiosondeo llevado a cabo por la mañana, suponiendo que las condiciones en la vertical permanecen estacionarias durante las horas diurnas, y a partir de la información horaria de temperatura registrada en superficie, se puede estimar la evolución horaria del crecimiento de la capa límite durante las horas diurnas. El procedimiento de determinación gráfica de la altura de la capa de mezcla según el método de Holzworth se muestra en la Figura 3.3.

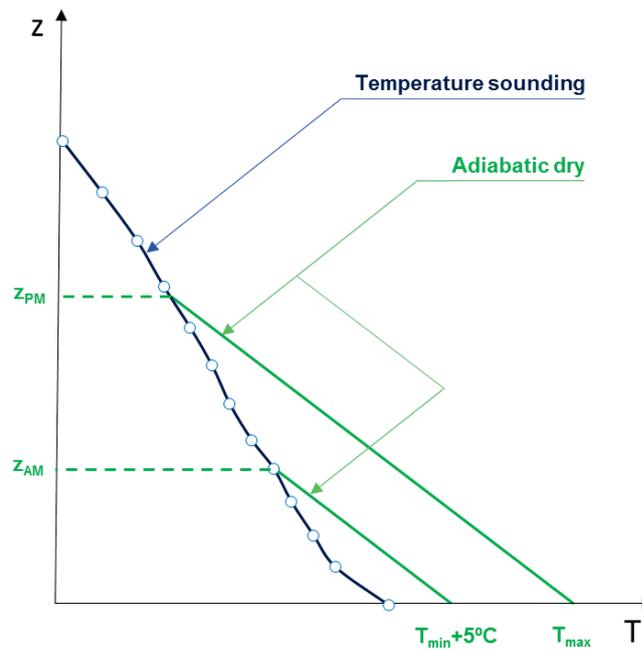


Figura 3.3. Determinación gráfica de la altura de capa de mezcla según el método propuesto por Holzworth (1967).

La altura de la capa de mezcla se calcula como la altura correspondiente a la del punto de intersección entre el perfil vertical de temperatura observado durante la realización del sondeo y la recta adiabática seca, que representa el perfil vertical de temperatura que presentaría un volumen de aire seco que ascendiese o descendiese adiabáticamente. Se trata, por tanto, de una recta de pendiente igual al gradiente adiabático del aire seco, $\Gamma = 0,0098 \text{ }^{\circ}C/m$. Tradicionalmente este método se aplica solamente a un único sondeo realizado por la mañana. No obstante, ante la disponibilidad de datos referentes a dos sondeos termodinámicos por día, a las 0 h y a las 12 h, en este trabajo se ha decidido aplicar este procedimiento a ambos, de manera que se ha determinado una altura de capa de mezcla matutina (MH_{am}) y una vespertina (MH_{pm}), respectivamente. La temperatura inicial a partir de la cual trazar la recta adiabática seca se obtiene de los valores horarios de temperatura a nivel de suelo: la mínima temperatura registrada entre las 0 h y las 6 h (más un incremento de $5^{\circ}C$ para tener en cuenta el efecto isla de calor) para el sondeo de las 0 h, y la máxima temperatura registrada entre las 12 h y las 23 h para el sondeo de las 12 h. Para los casos en los que se estima la altura de la capa de mezcla sobre un centro urbano, Miller y Holzworth (1967) consideran que la altura de la capa de mezcla de la mañana es la que se obtiene de considerar la temperatura mínima observada en superficie en la estación meteorológica urbana más un incremento de temperatura de $5^{\circ}C$. Se asume que el efecto de isla de calor (situación urbana de acumulación de calor por la enorme cantidad de cemento y demás materiales absorbentes de calor en una ciudad) se contabiliza considerando un incremento de temperatura de $5^{\circ}C$ (Trejo y Villagómez, 2011). Una desventaja de este método es que es muy sensible al valor de la temperatura en superficie: un error de $1^{\circ}C$ puede producir cambios en la altura de la capa de mezcla de cientos de metros (Myrick y col., 1994).

3.2.3.2. *Inversión térmica*

El fenómeno de inversión térmica se define como el comportamiento contrario en la variación vertical habitual de la temperatura, es decir, un aumento térmico con la altura. Un estrato de aire que presente dicho fenómeno no permite que se produzcan movimientos ascendentes de las masas de aire, puesto que el aire que se encuentra en la parte inferior del mismo es más frío y, por tanto, más denso y pesado. En consecuencia, constituye una zona de máxima estabilidad atmosférica que dificulta la dispersión vertical de los contaminantes en la atmósfera. No obstante, la principal desventaja de utilizar la inversión como indicador de estabilidad atmosférica es que se trata de una forma extrema de estabilidad, de tal forma que una atmósfera puede ser estable en ausencia de inversión térmica.

A partir de los datos de los sondeos termodinámicos, se han identificado aquellos casos en los que se produce inversión térmica y se han distinguido tres situaciones en función de si hay inversión o no, y en caso afirmativo, de la magnitud de esta. Se ha asignado un valor numérico a cada una de estas situaciones para evaluarlas cualitativamente y así introducirlas como variables de entrada nominales a los modelos (Tabla 3.3.).

Tabla 3.3. Equivalencia nominal para la introducción de la inversión térmica como variable de entrada a los modelos.

Inversión	Equivalencia nominal	Puntos con pendiente positiva
Ausencia	1	0
Moderada	2	≤ 20 %
Completa	3	> 20%

3.2.3.3. *Número de Richardson*

La turbulencia es el principal proceso físico que hace posible la mezcla, caracterizando, por tanto, la capa límite atmosférica. Encontrar un parámetro para aplicaciones prácticas que ayude de forma sencilla a determinar si el flujo es laminar o turbulento no es baladí (Jeričević y Grisogono, 2006).

Para describir numéricamente la dinámica de los movimientos de aire dentro de la capa superficial atmosférica se ha empleado el número adimensional de Richardson (Ri), que determina la estabilidad a través de la relación entre las fuerzas convectivas y mecánicas que intervienen en los movimientos atmosféricos (Viloria y col., 2012). El número de Richardson es la relación entre el gradiente de temperatura vertical y el gradiente vertical de velocidad de viento al cuadrado. El numerador se relaciona con las fuerzas de desestabilización que generan corrientes de aire ascendentes. En cambio, el denominador está relacionado con la energía

cinética que destruye esos ascensos. Por tanto, un mayor valor de Ri indica una mayor estabilidad (Woodward, 1988).

El valor de Ri puede evaluarse por distintos métodos, como el Richardson *Flux* (Ri_f) y el Richardson *Bulk* (Ri_B) (Viloria y col., 2012), y desafortunadamente varía con la altura. En este trabajo, conocidas las temperaturas y la velocidad del viento en dos niveles, se ha calculado el Ri según Hernández-Ramírez y col. (2011) mediante la ecuación 3.1:

$$Ri = \frac{g (T_f - T_i) (z_f - z_i)}{T_b (u_f - u_i)^2} \quad (3.1)$$

donde g es la aceleración de la gravedad (m/s^2), T es la temperatura del aire (K), z es la altura geopotencial (m) y u es la componente horizontal de la velocidad del viento (m/s). Los subíndices f e i indican las posiciones final e inicial de los perfiles verticales, que en este trabajo corresponden a la altura de capa de mezcla y al nivel de suelo, respectivamente.

El valor que puede tomar el Ri da cuenta del grado de turbulencia. El signo del Ri viene condicionado por el signo del gradiente de temperatura con la altura y éste por el tipo de estabilidad, de tal forma que si el Ri es positivo, la atmósfera es estable, si es negativo, inestable, y si es cero, neutra.

3.2.3.4. Clases de estabilidad de Pasquill-Gifford

Las clases de estabilidad de Pasquill (1961) constituyen la primera propuesta de clasificación de estabilidad atmosférica. Dicha clasificación establece seis categorías de estabilidad, designadas por una letra de A a F, donde la categoría A corresponde a una atmósfera extremadamente inestable y la F a una atmósfera moderadamente estable. De acuerdo a esta clasificación, como se muestra en la Tabla 3.4., la estabilidad atmosférica depende principalmente de dos parámetros: (i) la turbulencia mecánica, considerada por medio de la velocidad del viento en superficie, y (ii) la flotabilidad, que se tiene en cuenta indirectamente a través de la radiación solar durante el día y del grado de nubosidad durante la noche.

Tabla 3.4. Categorías de estabilidad atmosférica según el criterio Pasquill-Gifford.

Velocidad del viento en superficie (a 10 m) m/s	Radiación solar			Noche	
	Fuerte	Moderada	Débil	Nubosidad	
				$\geq 4/8$	$\leq 3/8$
< 2	A	A-B	B	-	-
2-3	A-B	B	C	E	F
3-5	B	B-C	D	D	E
5-6	C	C-D	D	D	D
> 6	C	D	D	D	D

El criterio tenido en cuenta para la clasificación de la radiación solar en las tres categorías consideradas en la Tabla 3.4. se recoge en la Tabla 3.5.

Tabla 3.5. Categorías de radiación solar (Turner, 1969).

Radiación recibida (categoría)	Radiación solar (W/m ²)
Fuerte	$I > 700$
Moderada	$350 \leq I \leq 700$
Débil	$I < 350$

Asimismo, como se muestra en la Figura 3.4., también existe una relación entre las clases de estabilidad de Pasquill-Gifford y el perfil de temperatura.

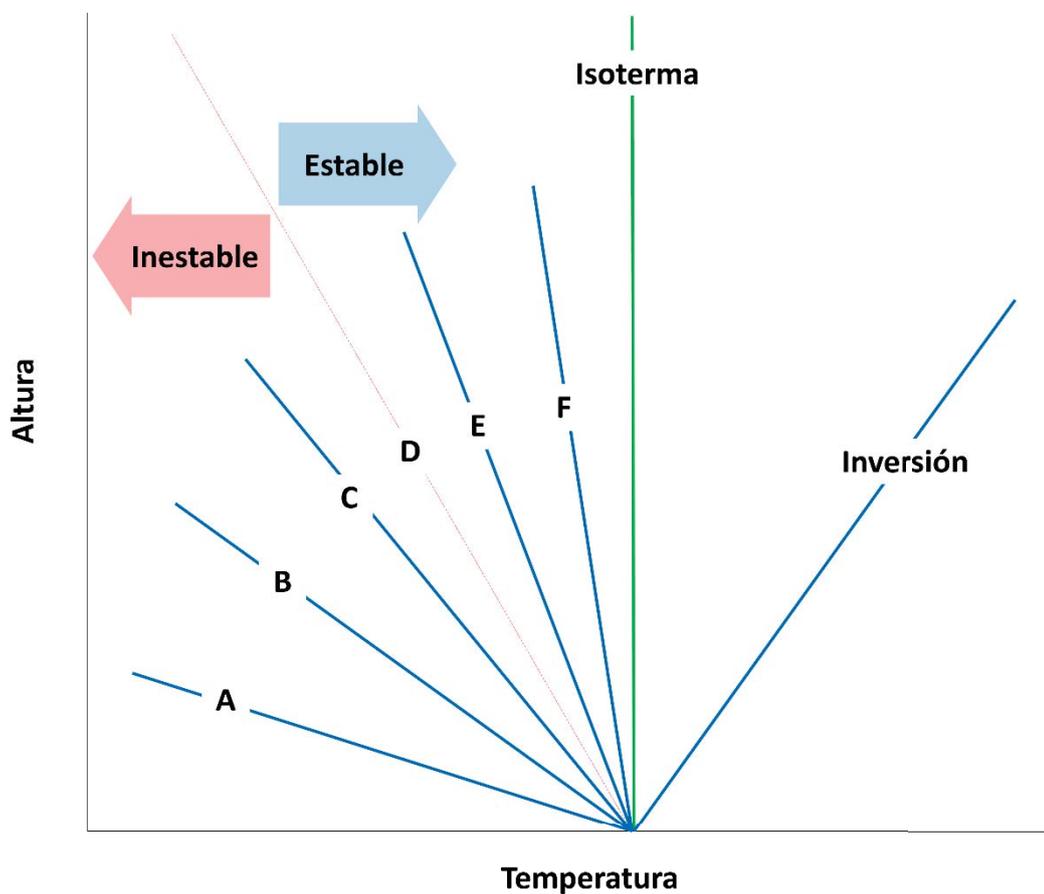


Figura 3.4. Perfil de temperaturas que establece las categorías de estabilidad atmosférica (Heinsohn y Kabel, 1999).

En este trabajo la clase de estabilidad de Pasquill-Gifford ha sido calculada en base a una equivalencia entre dichas clases y el valor del número de Richardson propuesta por Woodward (1998), que requiere únicamente información de los gradientes verticales de temperatura y velocidad del viento. Dicha equivalencia se recoge en la Tabla 3.6.

Tabla 3.6. Equivalencia entre valores cuantitativos del Ri y las clases de estabilidad de Pasquill-Gifford (Woodward, 1998).

Case de estabilidad de Pasquill	Equivalencia	Equivalencia nominal
A	$Ri < -0,86$	1
B	$-0,86 \leq Ri < -0,37$	2
C	$-0,37 \leq Ri < -0,10$	3
D	$-0,10 \leq Ri < 0,053$	4
E	$0,053 \leq Ri < 0,134$	5
F	$0,134 \leq Ri$	6

3.3. Selección de las bases de datos de partida

3.3.1. Variables dependientes

Las variables dependientes o respuesta que forman parte de la base de datos de partida empleada en este trabajo consisten en las concentraciones en inmisión (ng/m^3) en los cuatro puntos de muestreo seleccionados (véase el apartado 3.1.) de los siguientes contaminantes atmosféricos vinculados a la fracción PM_{10} del material particulado cuyos niveles en aire ambiente se encuentran regulados por la UE: As, Cd, Ni y Pb. Adicionalmente, únicamente en el punto de muestreo de Dunkerque se incluyen también en el estudio las concentraciones de Mn, V y Cr. La determinación cuantitativa de los niveles de dichos contaminantes se realizó como sigue:

- En los puntos de muestreo de Castro Urdiales y Reinoso, el organismo encargado de llevar a cabo las campañas de muestreo de material particulado fue el Centro de Investigación del Medio Ambiente (CIMA), organismo autónomo dependiente de la Consejería de Medio Ambiente, Ordenación del Territorio y Urbanismo de Cantabria. En los años 2008 y 2009 se tomaron muestras de PM_{10} con periodos de muestreo de 48 horas de duración y en 2010, muestras diarias. No obstante, el grupo de investigación Desarrollo de Procesos Químicos y Control de Contaminantes (DePRO) fue el encargado de determinar el contenido de As, Cd, Ni y Pb en las muestras de PM_{10} . Los niveles de metales observados en el año 2008 se encuentran en [Arruti y col. \(2011\)](#).
- En el punto de muestreo de Santander, el citado grupo de investigación llevó a cabo una campaña de muestreo en los años 2008 y 2009 para determinar los valores diarios de PM_{10} , así como los niveles de metales y semimetales asociados, que se encuentran publicados en [Arruti y col. \(2011\)](#). Como resultado, se obtuvo un número total de 94 muestras entre ambos años.
- En el punto de muestreo de Dunkerque, la *Unité de Chimie Environnementale et Interactions sur le Vivant* (UCEIV) llevó a cabo una campaña intensiva de muestreo de PM_{10} entre los meses de febrero y mayo de 2008, para la posterior determinación de

los niveles de los elementos Al, Ca, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Sn, Ti, V y Zn, así como de los iones Cl^- , NO_3^- , SO_4^{2-} y NH_4^+ . Un total de 78 muestras fueron recogidas en dicha campaña. Los valores medios observados de estos contaminantes se recogen en [Hleis \(2010\)](#).

Esta base de datos ha sido ampliada por medio de la adición de nuevas determinaciones de acuerdo a lo descrito en el apartado 3.4.

3.3.2. Variables independientes

Uno de los objetivos que se persigue con el planteamiento de las herramientas a desarrollar en este estudio es que éstas sean capaces de realizar estimaciones de los niveles de contaminantes en inmisión en base a variables de influencia (véase apartado 3.2.), que no supongan un coste adicional y que sean de fácil acceso. En este sentido, los datos registrados por los organismos e instituciones dedicados al estudio de la calidad del aire y de la meteorología, que cumplen con estos requisitos, resultan de gran interés y utilidad.

A continuación, se describen las variables independientes consideradas en este estudio, que se dividen en dos grupos: variables continuas y variables nominales.

3.3.2.1. Variables continuas

Las variables continuas consideradas para el desarrollo de modelos de estimación de niveles de metales en aire ambiente se componen de datos en relación a los niveles de concentración de contaminantes atmosféricos mayoritarios y a las condiciones meteorológicas registrados en los puntos de muestreo o en las estaciones de medida más próximas a los mismos.

Niveles de contaminantes atmosféricos mayoritarios

Los datos referentes a la concentración en inmisión de contaminantes atmosféricos mayoritarios utilizados en este trabajo como variables independientes para el desarrollo de modelos en Cantabria se obtuvieron a partir de los históricos de las estaciones de medida de la Red Automática de Vigilancia y Control de la Calidad del Aire de Cantabria (RVCCA), gestionada por el CIMA y que se encuentran accesibles al público en la página web de dicho organismo ([CIMA, 2016](#)). Las estaciones de medida de la RVCCA consideradas en este trabajo se encuentran localizadas en las siguientes ubicaciones:

- Castro Urdiales y Reinosa. Las estaciones de medida de la RVCCA se corresponden con los puntos de muestreo descritos en el apartado 3.1.
- Santander. La estación de medida más próxima al punto de muestreo en Santander es la estación de Tetuán ($43^{\circ}28'0''\text{N}$, $3^{\circ}47'29''\text{O}$ y a 24 m sobre el nivel del mar), situada aproximadamente a 1 km SSE del punto de muestreo (Figura 3.1).

Para el estudio llevado a cabo en Francia, los datos proceden de la estación de medida ubicada en la población de Saint-Pol-sur-Mer, (51°2'2"N, 02°20'7"E y 4 m sobre el nivel del mar) anexa a la ciudad de Dunkerque (Figura 3.2.), y gestionada por la red de calidad del aire de la región francesa Norte-Paso de Calais (Atmo-Nord-Pas-de-Calais). Dichos datos se encuentran disponibles a través de la página web del organismo (ATMO, 2016).

Datos meteorológicos

En cuanto a la información referente a la meteorología, en el caso de Castro Urdiales y Reinosa, las estaciones de medida de la RVCCA cuentan además con torre meteorológica, por lo que los datos referentes a las condiciones meteorológicas en los puntos de muestreo se obtuvieron igualmente en dichas ubicaciones. La disponibilidad de datos meteorológicos en el mismo punto de muestreo corresponde a una situación ideal que en la práctica no siempre es posible. En este sentido, en ausencia de torre meteorológica propia en la estación de Tetuán (Santander), fue necesario recurrir a otras fuentes para la obtención de datos meteorológicos representativos para el punto de muestreo de Santander. Así, en Santander, los datos meteorológicos han sido proporcionados por la Delegación Territorial en Cantabria de la Agencia Estatal de Meteorología (AEMET), perteneciente al Centro Meteorológico Territorial (CMT) de Cantabria y Asturias, cuya sede se encuentra en la localidad de Cueto del municipio de Santander. Los datos meteorológicos facilitados son los registrados por la estación de medida CMT I, Santander, que se encuentra ubicada en las coordenadas 43°29'28"N, 03°48'2"O y a 52 m sobre el nivel del mar. Se trata de la estación meteorológica más próxima al punto de muestreo correspondiente, situada a aproximadamente 2 km N (Figura 3.1.). Dicha estación ofrece además la posibilidad de disponer de datos de sondeos termodinámicos para el cálculo de variables relacionadas con la estabilidad atmosférica. Estos datos se componen de pares de datos de presión (hPa) y temperatura (°C) en función de la altura geopotencial (m). Asimismo, también incluyen otras variables meteorológicas de interés como la temperatura del punto de rocío (°C), la humedad relativa (%) y la dirección (°) y velocidad del viento (m/s). Los datos de sondeos termodinámicos del año 2011 fueron decodificados según la clave TEMP FM35. Habiendo disponibles dos sondeos termodinámicos por día en la estación CMT I, Santander, (a las 0h y a las 12h), se han considerado dos valores diferentes para cada variable de entrada relacionada con la estabilidad atmosférica: uno matutino, identificado mediante el subíndice *am*, y uno vespertino, identificado mediante el subíndice, *pm*.

Los datos meteorológicos empleados en el estudio llevado a cabo en Dunkerque proceden de la estación meteorológica perteneciente al Servicio Meteorológico Nacional de Francia, Météo-France, ubicada en el puerto de Dunkerque, con coordenadas 51°3'21"N, 02°20'22"E y a 1 m sobre el nivel del mar (Figura 3.2.). Estos datos pueden ser consultados en la página web de dicho servicio meteorológico (Météo-France, 2016).

Teniendo en cuenta los datos comunes de las cuatro estaciones de calidad del aire y meteorológicas seleccionadas (tres pertenecientes a la RVCCA y una a la AEMET), las variables continuas consideradas para el desarrollo de esta Tesis han sido las que se recogen en la Tabla 3.7.

Tabla 3.7. Variables continuas utilizadas como variables de entrada a los modelos.

Notación	Descripción ^a	Tipo	Unidades
LnPM ₁₀	Media del logaritmo neperiano de la concentración de PM ₁₀ (µg/m ³)	Contaminante mayoritario	-
SO ₂	Concentración media de dióxido de azufre	Contaminante mayoritario	µg/m ³
O ₃	Concentración media de ozono troposférico	Contaminante mayoritario	µg/m ³
NO _x	Concentración media de óxidos de nitrógeno, calculada como el sumatorio de las concentraciones medias de NO y NO ₂	Contaminante mayoritario	µg/m ³
T	Temperatura media	Meteorológica	°C
RH	Humedad relativa media	Meteorológica	%
WD	Dirección de viento predominante	Meteorológica	°
WS	Velocidad de viento predominante	Meteorológica	m/s
P	Presión media	Meteorológica	mbar
PP	Precipitación acumulada	Meteorológica	L/m ²

^aLos valores medios han sido calculados de acuerdo a la correspondiente duración de los periodos de muestreo de PM₁₀

Además, de manera particular en Santander, se han considerado de forma adicional variables asociadas a la estabilidad atmosférica (Tabla 3.8.), calculadas a partir de sondeos termodinámicos facilitados.

Tabla 3.8. Variables meteorológicas relacionadas con la estabilidad atmosférica para su uso como variables de entrada adicionales a los modelos a desarrollar en Santander.

Notación	Descripción	Tipo	Unidades
MH	Altura de la capa de mezcla	Meteorológica	m
Ri	Número de Richardson	Meteorológica	-

3.3.2.2. Variables nominales

Por otra parte, de acuerdo al apartado 3.2.4., se han considerado también variables cualitativas o nominales para tener en cuenta los efectos estacionales, los efectos de la posible intrusión de polvo sahariano y los efectos asociados al fin de semana. Dichas variables, así como la forma en la que han sido codificadas para su utilización, se recogen en la Tabla 3.9. Los datos de ocurrencia de episodios de intrusión de polvo sahariano se obtuvieron de los informes anuales de episodios de polvo africano en España (MAGRAMA, 2016), que son desarrollados por el

Consejo Superior de Investigaciones Científicas (CSIC) en colaboración con el Ministerio de Agricultura, Alimentación y Medio Ambiente del Gobierno de España (MAGRAMA). Además, en dicha tabla se recogen las variables meteorológicas cualitativas relacionadas con la estabilidad atmosférica que han sido calculadas de manera particular en Santander.

Tabla 3.9. Variables nominales utilizadas como variables de entrada a los modelos.

Notación	Descripción	Codificación
SE	Estación del año	1: Invierno; 2: Primavera; 3: Verano; 4: Otoño
SD	Intrusión de polvo sahariano	0: No intrusión; 1: Intrusión
WE	Fin de semana	0: Día laborable; 1: Fin de semana
I	Inversión térmica	1: Ausencia; 2: Moderada; 3: Completa
PG	Clase de estabilidad atmosférica de Pasquill-Gifford	1: A; 2: B; 3: C; 4: D; 5: E; 6: F

3.4. Ampliación de las bases de partida mediante muestreos y determinaciones adicionales

Ante el limitado número de muestras que conforman las bases de datos de partida disponibles (véase apartado 3.3.), ha sido necesario llevar a cabo una ampliación de las mismas mediante muestreos y determinaciones adicionales, con el fin de contar con un número de muestras apropiado como para efectuar no solo la calibración de los parámetros de los modelos estadísticos a desarrollar, sino también una evaluación de la capacidad predictiva de los mismos por medio de un procedimiento de validación externa, para lo cual es necesario reservar un subconjunto de muestras que no se utiliza en dicha etapa. En este sentido, se han llevado a cabo los siguientes trabajos en los puntos de muestreo de Cantabria:

- Castro Urdiales y Reinosa. Análisis del contenido de As, Cd, Ni y Pb en muestras de PM₁₀ procedentes de las campañas de muestreo llevadas a cabo por el CIMA en ambas ubicaciones en el año 2010.
- Santander. Campaña de muestreo de PM₁₀ en el año 2011 y determinación cuantitativa del contenido de As, Cd, Ni y Pb en las mismas.

Para ello, se han seguido los procedimientos de muestreo, preparación de muestra y análisis descritos a continuación.

3.4.1. Muestreos de PM₁₀

El muestreo de PM₁₀ efectuado durante la realización de este trabajo se ha llevado a cabo siguiendo el procedimiento descrito en la norma UNE-EN 12341:1999, para la determinación de la fracción PM₁₀ del material particulado en suspensión (AENOR, 1999), que constituye un

método de referencia, establecido como estándar por la normativa vigente en materia de calidad del aire. Dicho método consiste en la captación de materia particulada mediante aspiración, durante un período de tiempo determinado, de un caudal fijo de aire ambiente que se hace pasar por un soporte (filtro) sobre el que las partículas se depositan quedando retenidas. La selección de la fracción de materia particulada de interés (PM_{10}) se realiza con la ayuda de un cabezal de corte, diseñado para llevar a cabo la separación de las partículas en función de su tamaño por deposición inercial.

La carga másica del filtro se determina por gravimetría, esto es, mediante la diferencia entre la masa del filtro en un instante previo (en blanco) y posterior al muestreo, habiendo sido equilibrado en ambas ocasiones bajo las mismas condiciones de temperatura y humedad relativa. Finalmente, la concentración de materia particulada se obtiene dividiendo la carga másica del filtro entre el volumen total de aire ambiente muestreado.

La realización del muestreo se ha llevado a cabo empleando filtros de microfibras de cuarzo de 150 mm de diámetro de la empresa Sartorius AG, S.A. y un captador de alto volumen (modelo CAV-A/M) con un cabezal de corte PM_{10} (modelo PM1025-CAV), ambos de la empresa MCV, S.A., operando a un caudal de 30 m³/h durante 24 horas.

3.4.2. Determinación de metales/metaloides en PM_{10}

La determinación de los niveles de los elementos estudiados se ha llevado a cabo de acuerdo al método descrito en la Norma UNE-EN 14902:2006, para la medida de As, Cd, Ni y Pb en la fracción PM_{10} de la materia particulada en suspensión (AENOR, 2006), que es el método de referencia según el RD 102/2011. El procedimiento de determinación cuantitativa descrito en dicha norma consta de dos etapas, preparación de la muestra y análisis, que se describen a continuación. Este procedimiento ha sido ampliamente utilizado en el grupo DePRO en anteriores trabajos (Arruti, 2011; Ruiz, 2014).

La primera etapa, de extracción, tiene por objetivo la transferencia de los analitos de interés presentes en la muestra desde una fase sólida (materia particulada retenida en el filtro) a una fase líquida, de tal forma que se obtenga una disolución en medio ácido, que pueda ser utilizada por el equipo de análisis en la siguiente etapa. Dicho proceso se lleva a cabo por medio de una digestión ácida de los filtros, asistida por microondas (Ethos I), en vasos de digestión de teflón de 100 ml. Como reactivos se emplean, por cada muestra, 8 ml de ácido nítrico (HNO_3), 65% v/v, y 2 ml de peróxido de hidrógeno (H_2O_2), 30% v/v.

En primer lugar, se introduce todo o parte del filtro en un vaso de muestras del microondas, habiendo sido determinada previamente la masa de materia particulada por gravimetría. A continuación, se añaden los volúmenes de reactivo a cada vaso de digestión, 8 ml de HNO_3 y 2 ml de H_2O_2 , y se deja que tenga lugar una predigestión en la campana de extracción.

Posteriormente, se introduce el método de digestión en el microondas, que consta de las etapas que se muestran en la Tabla 3.10.

Tabla 3.10. Condiciones del método de digestión.

Etapa	Tiempo (min)	Potencia (W)	Temperatura final (°C)
1	3	900	100
2	10	900	170
3	5	900	190
4	15	800	190

Una vez finalizado el método y habiendo atemperado los vasos (hasta aproximadamente 50°C), para asegurar la recogida de los posibles restos de ácido que pudieran estar presentes en los mismos, se realizan tres lavados de la tapa y las paredes de cada vaso de digestión y de los restos de filtro, si parte permaneciese sin digerir, empleando para ello agua ultrapura Mili-Q. Estos lavados se incorporan a la muestra. A continuación, cada muestra se filtra con un filtro de jeringa de 0,45 µm y se determina por gravimetría su masa total. Por último, se almacenan en refrigeración hasta análisis.

La segunda etapa consiste en la determinación de la masa total de As, Cd, Ni y Pb, mediante espectrometría de masas con plasma de acoplamiento inductivo, ICP-MS, por sus siglas en inglés (Agilent, modelo 7500c). Asimismo, este procedimiento se emplea igualmente para la determinación de los niveles de otros contaminantes como Ti, V, Cr, Cu, Fe, Mn, Mo, Sb y Zn, cuya masa también se analiza. La calidad del método de análisis ha sido comprobada mediante material de referencia (NST, SRM 1649^a "Urban dust") con resultados satisfactorios, siguiendo el procedimiento empleado en el grupo DePRO en anteriores trabajos y que se explica con mayor detalle en los mismos (Arruti, 2011; Ruiz, 2014).

3.5. Reconciliación/procesado de datos

3.5.1. Detección de valores atípicos multivariantes

La presencia de valores atípicos (*outliers*) en conjuntos de datos muestrales es inevitable y puede proceder tanto de errores de medida (errores de transmisión o registro, fallos de equipos, o calibración deficiente de instrumentos) como de eventos excepcionales, que producen una respuesta única en los parámetros medidos (Alameddine y col., 2010).

Los *outliers* multivariantes son aquellas observaciones que se consideran atípicas por el valor que toman dentro del conjunto de variables, a diferencia de los *outliers* univariantes, que son aquellas observaciones que se consideran atípicas por el valor que toman en una determinada variable, con frecuencia valores extremos. Por este motivo, los primeros son más difíciles de

identificar que los segundos. Además, su presencia tiene efectos todavía más perjudiciales que en el caso unidimensional, porque distorsionan no sólo los valores de la medida de posición (media) o de dispersión (varianza), sino también las correlaciones entre las variables. Hasta hace poco, la mayoría de los métodos de identificación de *outliers* consistían en métodos de detección de una o dos variables, univariantes o bivariantes, que aunque bastante sencillos no se pueden utilizar para conjuntos de datos de múltiples variables. Jackson y Chen (2004) y Robinson y col. (2005) demostraron que el uso de aproximaciones univariantes en datos multivariantes puede conducir a una incorrecta identificación de valores atípicos, lo que conlleva la pérdida de observaciones válidas. Esto ocurre porque al considerar las variables individualmente se asume que cada variable es independiente del resto. Esta suposición raramente se cumple, por lo que las correlaciones entre variables han de ser tenidas en cuenta. Además, es importante que el procedimiento no esté afectado por los fenómenos de enmascaramiento y empantanamiento, que restringen considerablemente la utilidad de algunas técnicas clásicas. Ambos tienen lugar por la distorsión generada por la presencia de un conjunto de *outliers* que puede enmascarar la detección de otros, de manera que algunos valores atípicos no sean identificados como tal, fenómeno conocido como enmascaramiento (*masking*), o que, por el contrario, puede hacer que sean identificadas erróneamente como valores atípicos observaciones que no lo son, fenómeno conocido como empantanamiento (*swamping*) (Tordescini y col., 2013).

Existe una gran variedad de procedimientos para la identificación de valores atípicos en conjuntos de datos multivariantes. Las pruebas de discordancia, basadas en el conocimiento de la distribución de probabilidad generadora de los datos, se emplean ampliamente en la identificación de valores atípicos tanto univariantes como multivariantes. Mediante el uso de estas pruebas se examina si un determinado elemento es estadísticamente razonable o no bajo la distribución de probabilidad asumida, para poder determinar si se trata de un valor atípico. En el caso de estudios multivariantes, se debe asumir con anterioridad una medida que permita reducir la información de las variables a un escalar, introduciendo así una medida de orden en los datos. Esta medida suele ser una distancia que tenga en cuenta la localización de los datos (media) y la variabilidad de los mismos (matriz de varianzas y covarianzas). En este sentido, el método estadístico basado en la distancia de Mahalanobis (*MD*), constituye una estrategia clásica para la detección de valores atípicos multivariantes, ya que se trata del método más común, sencillo y de fácil implementación. En el mismo, se utiliza dicho parámetro como medida indicativa de la distancia existente entre la posición de cada punto (observación) y el centro de la nube multivariante creada por la dispersión de los datos, de tal manera que aquellas observaciones con grandes *MD* se identifican como valores atípicos. Se asume que los datos siguen una distribución normal y se considera que el cuadrado de las distancias de Mahalanobis calculadas sigue una distribución chi-cuadrado (χ^2). Por convenio, cualquier punto con una MD^2

que exceda al cuantil 97.5% de la distribución χ^2 con un determinado número de grados de libertad (el número de variables consideradas en el análisis), se considera atípico (Alameddine y col., 2010). En general, muchos estudios han estudiado la aplicación de métodos de detección de valores atípicos basados en la distancia de Mahalanobis (Giménez y col., 2012).

3.5.2. Normalización de variables

En ausencia de conocimiento en cuanto a la importancia relativa de cada variable de entrada y teniendo en cuenta que las mismas pueden presentar valores de diferentes órdenes de magnitud, lo cual afecta negativamente al ajuste –por ejemplo, en redes neuronales, el incremento de pesos de una neurona es proporcional a su entrada–, se suele recurrir a la tipificación, estandarización o normalización de variables. Un método de tipificación habitual consiste en transformar cada variable en una nueva que tenga media cero y varianza y desviación estándar unitarias. Para ello, se centra la variable restándole su media aritmética y se divide entre su desviación estándar (ecuación 3.2). De esta manera, de forma previa al análisis se asigna a cada variable el mismo peso, es decir, la misma importancia relativa.

$$Z_i = \frac{x_i - \bar{x}}{\sigma} \quad (3.2)$$

Donde z_i es el valor tipificado de la variable de entrada al modelo, x_i es el valor observado de dicha variable, y \bar{x} y σ son el valor medio y la desviación estándar de las observaciones, respectivamente.

3.6. Procedimiento de división de las bases de datos

Hay tres aspectos principales que tienen una gran influencia en el desempeño de los modelos, especialmente en los modelos de redes neuronales, y que deben ser tenidos en cuenta antes del desarrollo de los mismos: (i) la calidad de los datos; (ii) el número de datos, y (iii) la división de los datos. La calidad de los datos empleados para el desarrollo de los modelos afecta directamente a la precisión de las estimaciones de los mismos. Además, Flood y Kartam (1994) también señalaron que el desempeño de un modelo depende fuertemente del número de muestras disponibles para realizar el ajuste de los parámetros del mismo, de tal forma, que cuanto mayor sea el número de muestras disponible, más preciso es potencialmente el modelo. No obstante, desde un punto de vista práctico, no siempre es viable contar con tantas observaciones como sería deseable para el entrenamiento/calibración de un modelo, en especial cuando se destinan a aplicaciones como la que se plantea en este estudio, debido principalmente al coste y al tiempo asociados a la determinación experimental de los niveles de microcontaminantes en aire ambiente, como los metales y semimetales. Por tanto, el número y, por consiguiente, la proporción de muestras a incluir en cada uno de los tres subconjuntos de datos necesarios que se han de formar a partir del total de observaciones disponibles para

el desarrollo de los modelos, es decir, la división de los datos, se convierte en un asunto de suma importancia.

En general, para llevar a cabo el desarrollo de modelos estadísticos, las bases de datos han de ser divididas en tres diferentes subconjuntos de datos: (i) de ajuste, calibración o entrenamiento; (ii) de validación interna o cruzada; y (iii) de validación externa. El primero se utiliza para ajustar los parámetros del modelo; el segundo, para evitar el fenómeno de sobreajuste; y el tercero, para evaluar la capacidad de generalización del modelo, es decir, la capacidad de proporcionar un desempeño satisfactorio en cuanto a la estimación de muestras no utilizadas con anterioridad en la etapa de calibración. En cuanto a la validación externa de un modelo, lo recomendable es que se lleve a cabo utilizando un subconjunto de muestras cuyos valores se encuentren dentro del mismo rango que el de las muestras del subconjunto de entrenamiento. De lo contrario, la capacidad predictiva del modelo puede verse comprometida, porque normalmente los modelos de regresión presentan una mayor facilidad para resolver problemas de interpolación.

Varios estudios han destacado el papel decisivo que tiene la división de la base de datos en el desempeño de los modelos, especialmente en el de los modelos basados en redes neuronales (Tokar y Johnson, 1999; Maier y Dandy, 2000; Maier y col., 2010). Por tanto, una división óptima de los datos es un asunto crucial.

Desde un punto de vista práctico, hay dos formas de abordar este problema. Por un lado, un primer enfoque –de ahora en adelante, Enfoque A– correspondería al desarrollo de modelos empleando, en la etapa de calibración o entrenamiento, datos históricos procedentes de años precedentes, para posteriormente validar de forma externa los modelos desarrollados usando datos procedentes del año en el cual se pretende evaluar la calidad del aire con respecto a un contaminante atmosférico en concreto. De esta forma, la partición de la base de datos viene parcialmente condicionada por el problema. Además, esta representaría la situación real para las autoridades locales en caso de que empleasen este tipo de herramientas para la evaluación de la calidad del aire. Por otro lado, un segundo enfoque –de ahora en adelante, Enfoque B– consistiría en emplear el total de datos disponibles de forma conjunta sin tener en cuenta el periodo de tiempo al que pertenecen.

Varios estudios han tratado el problema de seleccionar subconjuntos representativos a partir de un conjunto superior de muestras (Wu y col., 1996; Tominaga, 1998; Sales y col., 2000; Bowden y col., 2002; Daszykowski y col., 2002). La división de datos al azar es un método común debido a su simplicidad y porque los datos extraídos al azar siguen la distribución estadística del conjunto de datos al completo. No obstante, este método no asegura la representatividad de las muestras extraídas ni que las muestras se encuentren dentro de los límites del subconjunto de entrenamiento/calibración, lo que ayudaría a minimizar problemas de

extrapolación. Por tanto, son preferibles otras técnicas de extracción de datos basadas en criterios con una base científica más rigurosa. Además, esto facilita la reproducibilidad en los análisis. En este contexto, [Kennard y Stone \(1969\)](#) propusieron un algoritmo que mide la similitud entre muestras en base a la distancia euclidiana. Este método, así como un grupo de otros métodos, como el algoritmo DUPLEX, han sido ampliamente utilizados desde entonces. En este contexto, [Sapto y col. \(2012\)](#) propusieron una modificación del algoritmo de Kennard-Stone basado en la distancia de Mahalanobis (MDKS, por sus siglas en inglés), para llevar a cabo la división de datos para el posterior desarrollo de modelos de redes neuronales. Se obtuvieron mejores resultados en comparación con el algoritmo de Kennard-Stone estándar y otros algoritmos relacionados como el método SPXY, un método de división de datos basado en las distancias x-y, propuesto por [Galvão y col. \(2005\)](#). El algoritmo modificado MDKS desarrollado por [Sapto y col. \(2012\)](#) ha sido el método elegido en este trabajo para llevar a cabo la partición de los datos, pues representa una mejora con respecto al algoritmo de Kennard-Stone tradicional, y además implica una implementación más sencilla que la de otros algoritmos más complejos.

3.7. Árboles de clasificación para estimación de muestras por debajo del límite de detección

En aquellas zonas y aglomeraciones donde el uso de técnicas de modelización y de estimación objetiva está permitido como herramienta para la evaluación de la calidad del aire, se puede dar el caso de que los niveles de los contaminantes objeto de estudio de un gran número de observaciones, se encuentren por debajo del límite de detección del método empleado para su análisis, puesto que estas técnicas están pensadas para su aplicación en zonas donde los niveles de dichos contaminantes sean relativamente bajos; al menos, donde se encuentren por debajo del correspondiente UIE (véase apartado 1.3.).

En caso de que la concentración de un determinado analito en una muestra se encuentre por debajo del límite de detección del método analítico empleado para la determinación cuantitativa de dicho analito, la estrategia más común para la consideración de este dato puntual en análisis y procesados subsiguientes de los datos es cuantificar la concentración del mismo como la mitad del límite de detección correspondiente. No obstante, en caso de utilizar estos datos para el desarrollo de modelos de regresión, el hecho de tener que efectuar esta sustitución en un número considerable de muestras, especialmente cuando se cuenta con una base de datos de tamaño modesto, puede afectar negativamente a la capacidad del modelo para llevar a cabo el ajuste de los datos, puesto que de esta manera a distintos valores de entrada les corresponde el mismo valor de salida. En este sentido, siguiendo el método

propuesto por Zalel y col. (2015) para estimar concentraciones de benzo(a)pireno (un microcontaminante al igual que los metales) en la República Checa, se ha aplicado un árbol de clasificación binario previo al desarrollo de los modelos para obtener, en base a los valores observados tanto de las variables independientes como de las dependientes, un conjunto de reglas que permita estimar si la concentración de metal de un muestra específica se encuentra por encima de un valor umbral establecido como el límite de detección. El árbol de clasificación permite identificar si una combinación de variables de entrada correspondiente a una muestra representa condiciones que se espera que estén asociadas a concentraciones de metales por encima del límite de detección. En ese caso, dicha muestra se incluye en la base de datos que se utilizará posteriormente para la estimación cuantitativa de las concentraciones de metales y semimetales por medio del desarrollo de modelos estadísticos. De lo contrario, se asigna un valor estimado preestablecido como la mitad del correspondiente límite de detección.

3.8. Desarrollo de modelos estadísticos

3.8.1. Estrategia seguida en la elaboración de los modelos PLSR

Con el propósito de desarrollar el modelo PLSR que permita realizar unas estimaciones más acertadas, se ha llevado a cabo un procedimiento iterativo que permitiese estudiar la influencia de la consideración de:

- Distintos porcentajes de división de la base de datos
 - En vista de la importancia del tamaño relativo de cada uno de los subconjuntos de datos necesarios para el desarrollo y validación de los modelos PLSR, se han tenido en cuenta distintos porcentajes.
- Distintos números de variables latentes
 - Se han considerado desde una hasta el número máximo de variables de entrada.

No obstante, el número óptimo de variables latentes ha sido determinado mediante un procedimiento de validación cruzada.

Los modelos PLSR han sido desarrollados en este trabajo empleando la versión R2014a MATLAB (MathWorks, Inc.).

3.8.2. Estrategia seguida en la elaboración de los modelos ANN

En este trabajo se han empleado redes neuronales de tipo perceptrón multicapa (MLP), que constituye el tipo de red neuronal de mayor aplicación en el campo de medioambiente para la predicción de la concentración de contaminantes atmosféricos. La arquitectura del perceptrón multicapa se caracteriza porque tiene sus neuronas agrupadas en tres tipos: de entrada, ocultas y de salida. Las neuronas de la capa de entrada no actúan como neuronas propiamente dichas, sino que se encargan únicamente de recibir las señales o patrones del exterior y propagar dichas señales a todas las neuronas de la siguiente capa. La última capa actúa como salida de la red, proporcionando al exterior la respuesta de la red para cada uno de los patrones de entrada. Las neuronas de las capas ocultas realizan un procesamiento no lineal de los patrones recibidos. Las conexiones del perceptrón multicapa siempre están dirigidas hacia adelante, es decir, las neuronas de una capa se conectan con las neuronas de la siguiente capa, de ahí que reciban el nombre de redes alimentadas hacia adelante o *feedforward*.

Dado un conjunto de datos de partida, el número de neuronas en las capas de entrada y de salida viene definido por el problema y coincide, respectivamente, con el número de variables independientes y de variables dependientes considerado. Con el propósito de encontrar la red neuronal que permita realizar estimaciones de forma óptima, se ha llevado a cabo un procedimiento iterativo que permitiese estudiar la influencia de la consideración de:

- Distintos porcentajes de división de la base de datos
 - En vista de la importancia del tamaño relativo de cada uno de los subconjuntos de datos necesarios para el desarrollo y validación de los modelos de redes neuronales artificiales, se han tenido en cuenta distintos porcentajes.
- Distintos números de neuronas en la capa oculta
 - Se han considerado desde una hasta el doble del número máximo de variables de entrada, para no aportar mayor complejidad a la red.

Asimismo, para cada iteración se han modificado los siguientes parámetros:

- Función de activación en las capas oculta y de salida
- Función de inicialización de pesos.
- Función de aprendizaje o actualización de los pesos
- Constante de adaptación
- Criterio de parada del aprendizaje
- Función de desempeño

Adicionalmente, en los dos primeros casos de estudio de la presente Tesis, se ha considerado un procedimiento alternativo que consiste en efectuar un análisis de componentes principales

(PCA) previo al desarrollo de los modelos de redes neuronales según el procedimiento descrito anteriormente.

Los modelos de redes neuronales artificiales se han desarrollado en este trabajo empleando la versión R2014a de MATLAB (MathWorks, Inc.).

3.9. Criterios de calidad/desempeño de los modelos

3.9.1. Incertidumbre de acuerdo a la UE

El nivel de calidad de las técnicas de modelización y estimación objetiva para su uso como métodos de evaluación de la calidad del aire ha de ser evaluado en el contexto de la Directiva europea sobre Calidad del Aire. En la misma, la incertidumbre se define como la máxima desviación entre los niveles de concentración calculados y medidos en un periodo de tiempo determinado –el que corresponda en cada caso– sin tener en cuenta la cronología de los acontecimientos (EC, 2008). Algunos autores (Wilks, 2006; Flemming y Stern, 2007; Heinke y Sokhi, 2007; Borrego y col., 2008) interpretan dicha incertidumbre como el error relativo máximo sin tener en cuenta los efectos del tiempo (*RME*, por sus siglas en inglés), que se define como la mayor diferencia entre las concentraciones observadas y estimadas de entre todos los percentiles ($C_{O,p}$ y $C_{E,p}$, respectivamente), normalizada por la concentración observada correspondiente al percentil (p) en el que tiene lugar ese máximo ($C_{O,p}$), y que se calcula de acuerdo a la ecuación 3.3. No obstante, el RD 102/2011, en base a la interpretación desarrollada en Denby (2010), propone como metodología para la cuantificación de la incertidumbre el cálculo del error relativo de la directiva (*RDE*, por sus siglas en inglés), que, con una formulación similar, se define como la diferencia entre la concentración observada más próxima al valor límite/objetivo ($C_{O,LV}$) y su valor estimado dado por el modelo dentro de la secuencia ordenada de menor a mayor de valores modelados correspondientes ($C_{E,LV}$), normalizando dicha diferencia por medio del propio valor límite/objetivo (LV) (Denby, 2010) y que se calcula de acuerdo a la ecuación 3.4. Por tanto, en este trabajo, el grado de cumplimiento con los requerimientos de incertidumbre para la evaluación de la calidad del aire con respecto a los metales y semimetales objeto de estudio por medio de técnicas de modelización y estimación objetiva ha sido evaluado a través de la consideración de ambos parámetros como índices de incertidumbre:

$$RME = \max(|C_{O,p} - C_{E,p}|) / C_{O,p} \quad (3.3)$$

$$RDE = |C_{O,LV} - C_{E,LV}| / LV \quad (3.4)$$

3.9.2. Otros criterios de desempeño de los modelos

La evaluación del desempeño de los modelos se lleva a cabo por medio del establecimiento de comparaciones de las estimaciones frente a las observaciones para lo cual es necesario un conjunto adicional de indicadores estadísticos. En general, se acepta que ningún indicador estadístico es suficiente por si solo para evaluar el desempeño de un modelo, de tal manera que habrá que definir un conjunto de estadísticos relevante, dependiendo del tipo de aplicación a la que se destine el modelo, para ser evaluados de manera conjunta. En este sentido, en este trabajo, se ha considerado la utilización de parámetros estadísticos incluidos en el kit de validación de modelos de la EURASAP y recomendados por [Borrego y col. \(2008\)](#) por ser considerados como los más importantes, en concreto: el sesgo fraccional (*FB*), el coeficiente de correlación (*r*), el error cuadrático medio (*RMSE*), el error cuadrático normalizado (*NMSE*) y la varianza fraccional (*FV*), que se calculan de acuerdo a las ecuaciones 3.5-3.9.

$$FB = \frac{\overline{C_O} - \overline{C_E}}{0.5 (\overline{C_O} + \overline{C_E})} \quad (3.5)$$

$$r = \left[\frac{\sum_{i=1}^n (C_{O,i} - \overline{C_O})(C_{E,i} - \overline{C_E})}{\sqrt{\sigma_O \sigma_E}} \right] \quad (3.6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (C_{O,i} - C_{E,i})^2} \quad (3.7)$$

$$NMSE = \frac{(\overline{C_O} - \overline{C_E})^2}{\overline{C_O} \overline{C_E}} \quad (3.8)$$

$$FV = 2 \frac{\sigma_O - \sigma_E}{\sigma_O + \sigma_E} \quad (3.9)$$

donde *N* es el número total de muestras, $C_{O,i}$ y $C_{E,i}$ son las concentraciones en inmisión de contaminante asociadas a la muestra *i* ($i = 1, \dots, N$), $\overline{C_O}$ y $\overline{C_E}$ son los valores medios de concentración, y σ_O y σ_E son las desviaciones estándar. Los subíndices *O* y *E* hacen referencia a las observaciones y a las estimaciones, respectivamente.

Teniendo en cuenta que los valores límite/objetivo en relación a los niveles en aire ambiente de As, Cd, Ni y Pb establecidos por la UE vienen dados en valores medios anuales, el *FB*, que aporta información acerca de la diferencia entre los valores medios observados y los estimados por los modelos, constituye un estadístico de especial interés en este estudio.

No obstante, el análisis del valor de estos indicadores, aunque proporciona información en cuanto al desempeño general del modelo en cuestión, no indica necesariamente si los resultados de dicho modelo han alcanzado un nivel de calidad suficiente para una aplicación

dada, como es en este caso la evaluación de la calidad del aire. Idealmente, los resultados de los modelos deberían conducir al cálculo de un coeficiente de correlación próximo a 1 y de unos FB , $RMSE$, $NMSE$ y FV próximos a 0. Desafortunadamente, en la práctica, debido a la incertidumbre inherente a los métodos de observación y de determinación analítica en laboratorio, estos valores rara vez se alcanzan. Por ello, es necesario el establecimiento de unos valores límite para estos parámetros, que sirvan como criterio para evaluar la calidad de los modelos para su uso orientado a una aplicación en concreto. En este sentido, [Kumar y col. \(1993\)](#) proponen valores para algunos de estos parámetros, que se asocian con unos requerimientos mínimos de calidad para los modelos: $NMSE \leq 0,5$ y $-0,5 \leq FB \leq 0,5$. No obstante, puesto que dependiendo de la aplicación estos requerimientos pueden resultar restrictivos, en este trabajo se propone la consideración del siguiente intervalo como rango de validez para el sesgo fraccional, $-2/3 \leq FB \leq 2/3$, que corresponde a la mitad y al doble de la concentración media observada, respectivamente.

Por otra parte, se ha empleado una herramienta gráfica para realizar una comparación sencilla entre los resultados de los modelos y para evaluarlos con respecto a su nivel de precisión para aplicaciones normativas. Esta representación, conocida como diagrama objetivo o *target diagram* fue propuesta por [Pederzoli y col. \(2011\)](#), habiendo sido desarrollada como una modificación del diagrama de Taylor ([Taylor, 2001](#)) por el Joint Reseach Centre (JRC) de la Comisión Europea (CE) dentro del marco del Foro de la modelización de calidad del aire en Europa (FAIRMODE, por sus siglas en inglés) e introducido en la herramienta DELTA ([Thunis y col., 2012](#)). En ella se compara el error en la estimación de valor medio, calculado según la ecuación 3.10 y representado en el eje de ordenadas, con el error cuadrático medio cruzado ($CRMSE$, por sus siglas en inglés), calculado según la ecuación 3.11 y representado en el eje de abscisas, siendo ambos normalizados por la desviación estándar de las observaciones. Además, incluye una circunferencia de radio unitario, que define el valor límite aceptable de la eficiencia del modelo, MEF (ecuación 3.12).

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N (C_{E,i} - C_{O,i}) = \bar{C}_E - \bar{C}_O \quad (3.10)$$

$$\text{CRMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N [(C_{E,i} - \bar{C}_E) - (C_{O,i} - \bar{C}_O)]^2} \quad (3.11)$$

$$\text{MEF} = 1 - \left(\frac{\text{RMSE}}{\sigma_O} \right)^2 \quad (3.12)$$

En la Figura 3.5. se muestra una diagrama objetivo de ejemplo. Aunque el $CRMSE$ es siempre positivo por su propia definición (ecuación 3.11), los valores negativos en el eje de abscisas se utilizan para proporcionar información adicional en relación a la varianza, de tal manera que si

las estimaciones presentan una varianza menor que las observaciones, el punto se localizará en el lado izquierdo del diagrama ($CRMSE$ negativo), ubicándose en el lado derecho en caso contrario ($CRMSE$ positivo). Por otra parte, los puntos que se localizan dentro de la circunferencia, presentan un MEF positivo, lo que indica que los resultados del modelo constituyen una mejor predicción de las observaciones que un valor constante igual al valor medio de las mismas (Jolliff y col., 2009). No obstante, de forma general, según esta representación, cuanto más cerca del centro de la circunferencia se encuentren los resultados de los modelos, mejor es el desempeño de los mismos.

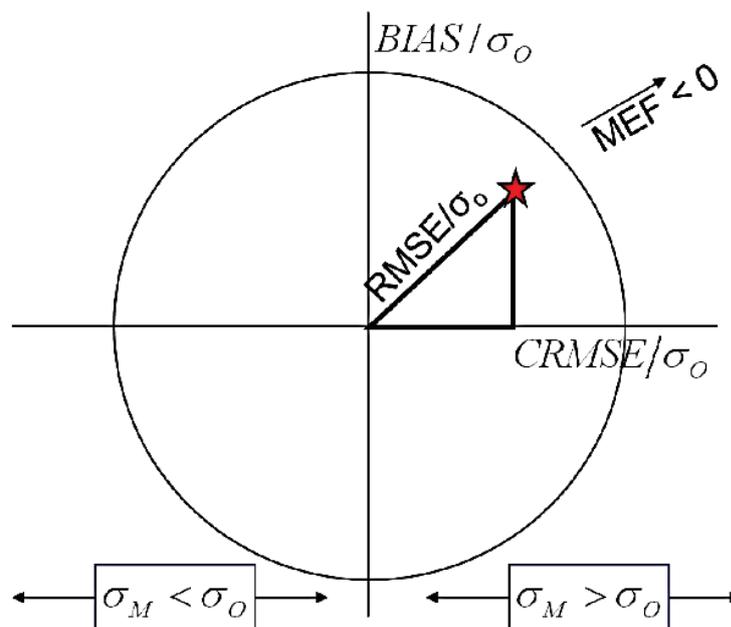


Figura 3.5. Representación esquemática del diagrama objetivo (Pederzoli y col., 2011).

3.10. Metodología para la identificación y evaluación del desempeño de los mejores modelos

La Figura 3.6., resume en forma de diagrama de flujo, la metodología propuesta para la evaluación de la calidad del aire con respecto a As, Cd, Ni y Pb mediante el desarrollo de modelos estadísticos basados en PLSR y ANNs. Dicha metodología, partiendo de datos meteorológicos y de contaminación registrados por los organismos encargados del estudio de la meteorología y de la evaluación de la calidad del aire, consta de cinco etapas principales: (i) eliminación de valores atípicos multivariantes mediante el método basado en la distancia de Mahalanobis; (ii) predicción de muestras por encima del límite de detección, de tal manera que estas son las únicas consideradas en los vectores de entrada a los modelos, mediante los árboles de decisión diseñados al efecto (véase apartado 3.7); (iii) partición de los datos en los

subconjuntos de datos requeridos mediante el método MDKS en función del enfoque empleado (véase apartado 3.6.), (iv) desarrollo de modelos estadísticos y (v) evaluación del desempeño.

Esta última etapa es la que permite identificar aquellos modelos que ofrecen un mejor desempeño. Así, en lo que respecta a la análisis de los resultados de los modelos desarrollados para su uso como aplicación con propósitos de regulación, en primer lugar, se determina si cumplen con los requerimientos de incertidumbre para técnicas de estimación objetiva establecidos por la UE; en segundo lugar, se evalúa la precisión en la estimación de los valores medios, que es la forma en la que los contaminantes objeto de estudio se encuentran regulados por la Directiva europea de calidad del aire y, por último, se evalúa si la eficiencia del modelo es positiva.

Estos requisitos mínimos de calidad son bastante permisivos, en el sentido de que de forma implícita se acepta como válido un amplio margen de error en las estimaciones. Por tanto, cabe la posibilidad de que más de un modelo desarrollado cumpla con los mismos, dando lugar a un conjunto de modelos a priori igualmente válidos. En ese caso, es preciso discernir cual es el modelo más apropiado, para lo cual se recurre al resto de indicadores estadísticos propuestos. Se seleccionan aquellos modelos que maximicen el coeficiente de correlación, r , y minimicen los errores, $RMSE$ y $NMSE$, de manera simultánea.

En caso de que ningún modelo desarrollado cumpla con los requisitos considerados, se plantea la consideración de variables de entrada relacionadas con la estabilidad atmosférica en aquellos lugares en los que se dispongan datos meteorológicos en altura.

Esta metodología en su totalidad solo se ha aplicado al caso de estudio de Santander, por las siguientes razones: (i) representa la mayor complejidad en cuanto a las fuentes de contaminantes atmosféricos; (ii) el contenido en el material particulado de alguno de los contaminantes estudiados se encuentra por debajo del correspondiente límite de detección del método analítico en numerosas muestras, lo que hace adecuado utilizar un árbol de decisión; (iii) los modelos preliminares desarrollados presentan un amplio margen de mejora; y (iv) se dispone de datos de sondeos termodinámicos en altura. En este sentido, en los casos de estudio Castro Urdiales, Reinoso y Dunkerque se ha prescindido del árbol de clasificación y del estudio de la influencia de las variables meteorológicas relacionadas con la estabilidad atmosférica.

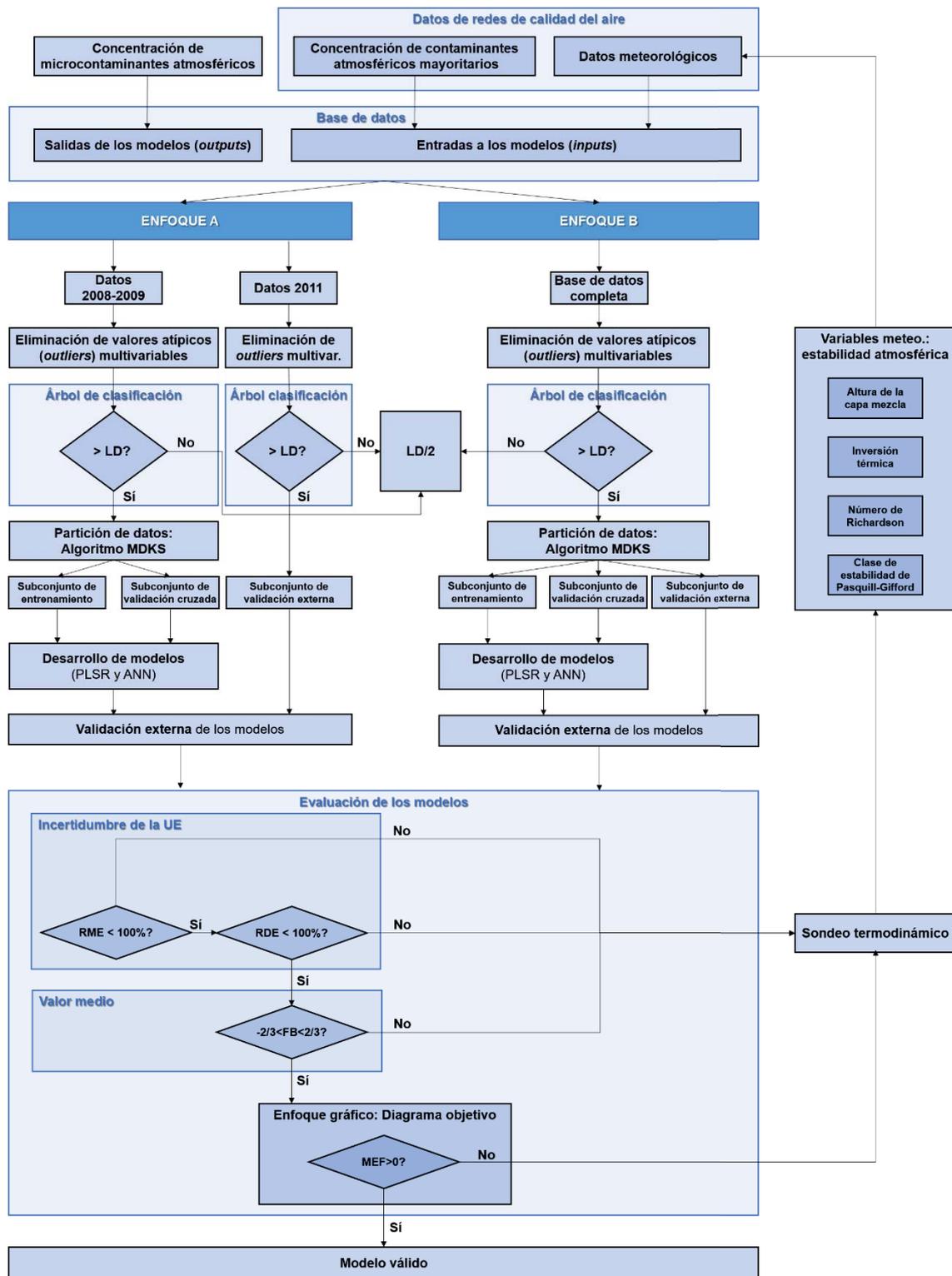


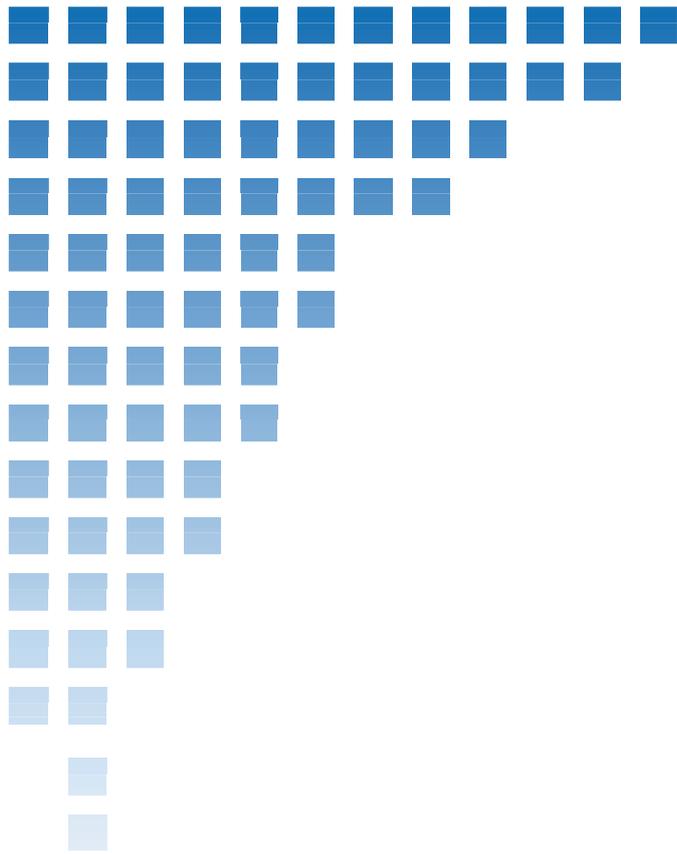
Figura 3.6. Diagrama de flujo de la metodología propuesta para la evaluación de la calidad del aire mediante modelos estadísticos (Santos y col., 2016).

3.11. Bibliografía

- AENOR, 2006. Norma UNE-EN 14902:2006. Método normalizado para la medida de plomo, cadmio, arsénico y níquel en la fracción PM₁₀ de la materia particulada en suspensión.
- AENOR, 1999. Norma UNE-EN 12341. Calidad del aire. Determinación de la fracción PM₁₀ de la materia particulada en suspensión: Método de referencia y procedimiento de ensayo de campo para demostrar la equivalencia de los métodos de medida al de referencia.
- Alameddine, I., Kenney, M.A., Gosnell, R.J., Reckhow, K.H., 2010. Robust multivariate outlier detection methods for environmental data. *Journal of Environmental Engineering* 136, 1299-1304.
- Arruti, A., Fernández-Olmo, I., Irabien, A., 2011. Regional evaluation of particulate matter composition in an Atlantic coastal area (Cantabria region, northern Spain): spatial variations in different urban and rural environments. *Atmospheric Research* 101, 208-293.
- Arruti, A., 2011. Evaluación de la composición del material particulado atmosférico en medio urbano y rural en Cantabria. Tesis doctoral de la Universidad de Cantabria.
- Arya, S.P.S., 1981. Parameterizing the height of the stable atmospheric boundary layer. *Journal of Applied Meteorology* 20, 1192-1202.
- ATMO, 2016. ATMO Nord-Pas-de-Calais: Surveillance et évaluation de l'atmosphère. <http://www.atmo-npdc.fr/>.
- Batchvarova, E., Gryning, S.E., 1991. Applied model for the growth of the daytime mixed layer. *Boundary-Layer Meteorology* 56, 261-274.
- Ben-Gal, I., 2005. Outlier detection. En: Maimon O., Rockach, L. (Eds.) *Data mining and knowledge discovery handbook: a complete guide for practitioners and researchers*. Kluwer Academic Publishers.
- Benkley, C.W., Schulman, L.L., 1979. Estimating hourly mixing depths from historical meteorological data. *Journal of Applied Meteorology* 18, 772-780.
- Borrego, C., Monteiro, A., Ferreira, J., Miranda, A.I., Costa, A.M., Carvalho, A.C., Lopes, M., 2008. Procedures for estimation of modelling uncertainty in air quality assessment. *Environmental International* 34, 613-620.
- Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. *Water Resources Research* 38(2), 1010.
- Burzynski, J., Godłowska, J., Tomazewska, A.M., Walczewski, J., 2004. The calculated mixing height in comparison with the measured data. 9th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO9), 24-28.
- CIMA: Centro de Investigación del Medio Ambiente, 2016. Calidad del Aire de Cantabria: Histórico de estaciones. Consejería de Medio Ambiente, Ordenación del Territorio y Urbanismo. Gobierno de Cantabria. http://www.medioambientecantabria.es/calidad_aire/.
- Chen, D., Cai, W., Shao, X., 2007. An adaptive strategy for selection representative calibration samples in the continuous wavelet domain for near-infrared spectral analysis. *Analytical and Bioanalytical Chemistry* 387, 1041-1048.
- Daszykowski, M., Walczak, B., Massart, D.L., 2002. Representative subset selection. *Analytica Chimica Acta* 468, 91-103.
- Deardorff, J.W., 1972. Parametrization of the planetary boundary layer for use in general circulation models. *Monthly Weather Review* 100, 93-106.
- Denby, B., 2010. Guidance on the use of the models for the European Air Quality Directive. A working document of the Forum for Air Quality Modelling in Europe, FAIRMODE. ETC/ACC Report.
- EC, 2008. Council Directive 2008/50/EC Directive of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. The European Parliament and the Council of the European Union. Off J L152 1-44.
- Flemming, J., Stern, R., 2007. Testing model accuracy measures according to the EU directives-examples using the chemical transport model REM-CALGRID. *Atmospheric Environment* 41, 9206-9216.
- Flood, I., Kartam, N., 1994. Neural networks in civil engineering. I: Principles and understanding. *Journal of Computing in Civil Engineering* 8(2), 131-148.
- Galvão, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., Saldanha, T.C.B., 2005. A method for calibration and validation subset partitioning. *Talanta* 67, 736-740.
- Giménez, E., Crespi, M., Garrido, M.S., Gil, A.J., 2012. Multivariate outlier detection based on robust computation of Mahalanobis distances. Application to positioning assisted by RTK GNSS Networks. *International Journal of Applied Earth Observation and Geoinformation* 16, 94-100.
- Hanna, S.R., Briggs, G.A., Hosker, R.P., 1982. *Handbook on Atmospheric Diffusion*. U.S. Dept. of Energy, Technical Information Center, Washington D.C.
- Heike, K., Sokhi, R.S., 2008. Overview of tools and methods for meteorological and air pollution mesoscale model evaluation and user training. Joint Report of COST Action 728.
- Heinsohn, R.J., Kabel, R.L., 1999. *Sources and Control of Air Pollution*. New Jersey: Prentice Hall.
- Hernández-Ramírez, G., Sauer, T.J., Hatfield, J.L., Prueger, J.H., 2011. Quantifying atmospheric stability conditions at a swine facility and an adjacent corn field in Iowa, USA. *Theoretical and Applied Climatology* 105, 495-503.
- Hleis, D. 2010. Evaluation de la contribution d'émissions sidérurgiques à la teneur en particules en suspension dans l'atmosphère à une échelle locale. Tesis doctoral de la Université du Littoral Côte d'Opale.
- Holzworth, G., 1967. Mixing depths, wind speeds and air pollution potential from selected locations in the United States. *Journal of Applied Meteorology* 6, 1039-1044.

- INE: Instituto Nacional de Estadística, 2016. INEbase: Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero. <http://www.ine.es/dynt3/inebase/index.htm?padre=517>.
- Jackson, D.A., Chen, Y., 2004. Robust principal component analysis and outlier detection with ecological data. *Environmetrics* 15(2), 129-139.
- Jeričević, A., Grisogono, B., 2006. The critical bulk Richardson number in urban areas: verification and application in a numerical weather prediction model. *Tellus* 58A, 19-27.
- Jolliff, J.K., Stow, C.A., McGillicuddy, D.J., Allen, J.I., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. *J. Marine Syst.* 76, 4-15.
- Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. *Technometrics* 11(1), 137-148.
- Kumar, A., Luo, J., Bennett, G., 1993. Statistical evaluation of lower flammability distance (LFD) using four hazardous release models. *Process. Safety Progress* 12, 1-11.
- MAGRAMA: Ministerio de Agricultura, Alimentación y Medio Ambiente, 2016. Histórico de informes de episodios naturales. Gobierno de España. <http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/gestion/anales.aspx>.
- Mahrt, L., Andre, J.C., Heald, R.C., 1982. On the depth of the nocturnal boundary layer. *Journal of Applied Meteorology* 21, 90-92.
- Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling and Software* 25, 891-909.
- Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software* 15, 101-124.
- Météo-France, 2016. METEO FRANCE: Données Publiques. <https://donneespubliques.meteofrance.fr/>.
- Middleton, D.R., Thomson, D.J., 2001. Uncertainties in met pre-processing for dispersion models. Atmospheric Dispersion Modelling Liaison Committee (ADMLC) Workshop at NRPB.
- Miller, M.E., Holzworth, G.C., 1967. An atmospheric diffusion model for metropolitan areas. *Journal of the Air Pollution Control Association* 17, 46-50.
- Ministerio de Fomento, 2011. Mapa de tráfico 2010. Gobierno de España.
- Myrick, R.H., Sakiyama, S.K., Angle, R.P., Sandhu, H.S., 1994. Seasonal mixing heights and inversions at Edmonton, Alberta. *Atmospheric Environment* 28(4), 723-729.
- Nieuwstadt, F.T.M., 1984. The turbulent structure of the stable, nocturnal boundary layer. *Journal of Atmospheric Sciences* 41, 2202-2216.
- Okubo, A., Takeda, S., Obata, H., 2013. Atmospheric deposition of trace metals to the western North Pacific Ocean observed at coastal station in Japan. *Atmospheric Research* 129-130, 20-32.
- Pallarés, S., Vicente, A.B., Jordán, M.M., Sanfeliu, T., 2007. Study of the levels of concentrations of As, Cd and Ni in a ceramic cluster. *Water, Air, and Soil Pollution* 180, 51-64.
- Papadimitriou, S., Kitawaga, H., Gibbons, P.G., Faloutsos, C., 2003. LOCI: Fast outlier detection using the local correlation integral. 19th International Conference on Data Engineering (ICDE03).
- Pasquill, F., 1961. The estimation of the dispersion of windborne material. *Meteorology Magazine* 90, 33-49.
- Pederzoli, A., Thunis, P., Georgieva, E., Borge, R., Carruthers, D., 2011. Performance criteria for the benchmarking of air quality model regulatory applications: the "target" approach. 14th Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO14), 297-301.
- Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire. (BOE núm. 25, 29 de enero de 2011).
- Robinson, R.B., Cox, C.D., Odom, K., 2005. Identifying outliers in correlated water quality data. *Journal of Environmental Engineering* 131(4), 651-657.
- Rodionova, O.Y., Pomerantsev, A.L., 2008. Subset selection strategy. *Journal of Chemometrics* 22, 674-685.
- Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85, 633-651.
- Ruiz, S., 2014. Evaluación de los niveles de metales pesados e hidrocarburos aromáticos policíclicos (HAPs) en aire ambiente en zonas afectadas por fuentes puntuales en Cantabria. Tesis doctoral de la Universidad de Cantabria.
- Sakata, M., Marumoto, K., Nerukawa, M., Asakura, K., 2006. Regional variations in wet and dry deposition fluxes of trace elements in Japan. *Atmospheric Environment* 40, 521-531.
- Sales, F., Rius, A., Callao, M.P., Rius, F.X., 2000. Standardization of a multivariate calibration model applied to the determination of chromium in tanning sewage. *Talanta* 52(2), 329-336.
- Santos, G., Fernández-Olmo, I., 2016. A proposed methodology for the assessment of arsenic, nickel, cadmium and lead levels in ambient air. *Science of the Total Environment* 554-555, 155-166.
- Sapoto, A., Tadé, M.O., Vuthaluru, H., 2012. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chemical Product and Process Modeling* 7(1), 13.
- Seibert, P., Beyrich, F., Gryning, S.E., Joffre, S., Rasmussen, A., Tercier, P., 1997. Mixing height determination for dispersion modelling. Report of Working Group 2, COST Action 710: Preprocessing of meteorological data for dispersion modelling.
- Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric chemistry and physics: from air pollution to climate change, 2nd Edition. Nueva York: John Wiley & Sons, Inc.
- Spadaro, J.V., 1999. Quantifying the effects of airborne pollution: Impact models, sensitivity analyses and applications. Tesis doctoral, École des Mines de Paris – ARMINES.
- Tate, M.B., Bates, M.H., 1984. Bulk deposition of metals in Tulsa, Oklahoma. *Water, Air, and Soil Pollution* 22, 15-26.

- Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research* 106, No. D7, 7183-7192.
- Thunis, P., Georgieva, E., Pederzoli, A., 2012. A tool to evaluate air quality model performances in regulatory applications. *Environmental Modelling and Software* 38, 220-230.
- Tokar, A.S., Johnson, P.A., 1999. Rainfall-Runoff modelling using artificial neural networks. *Journal of Hydrologic Engineering* (ASCE), 121(6), 499-508.
- Tominaga, Y., 1998. Representative subset selection using genetic algorithms. *Chemometrics and Intelligent Laboratory Systems* 43(1-2), 157-163.
- Tordeschini, R., Ballabio, D., Consonni, V., Sahigara, F., Filzmoser, P., 2013. Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection. *Analytica Chimica Acta* 787, 1-9.
- Trejo R., Villagómez, L., 2011. Método propuesto para estimar la altura de capa de mezcla en la atmósfera con ayuda de la aviación. *Conciencia Tecnológica* 44, 36-40.
- Turner, J.S., 1969. Buoyant plumes and thermals. *Annual Review of Fluid Mechanics* 1, 29-44.
- Viloria, R., Tricio, V., Oliveira-Júnior, J.F., Lyra, G.B., 2012. Análisis de la estabilidad atmosférica mediante distintas formulaciones del número de Richardson. 11º Congreso Nacional del Medio Ambiente (CONAMA 2012).
- Wilks, D.S., 2006. *Statistical methods in the atmospheric sciences*. International Geophysics Series, 2nd Edition.
- Woodward, J.L. 1998. Atmospheric stability classification schemes. En: *Estimating the Flammable Mass of a Vapor Cloud: A CCPS Concept Book*. American Institute of Chemical Engineers (AIChE). Nueva York.
- Wu, W., May, R., Dandy, G., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models. 2012 International Congress on Environmental Modelling and Software: Managing Resources of a Limited Planet, Leipzig, Alemania.
- Wu, W., Walczack, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R., Prebble, K.A., 1996. Artificial neural networks in classification of NIR spectral data: Design of the training set. *Chemometrics and Intelligent Laboratory Systems* 33, 35-46.
- Zalel, A., Yuval, Svecova, V., Sram, R.J., Bartonova, A., Broday, D.M., 2015. Modeling airborne benzo(a)pyrene concentrations in the Czech Republic. *Atmospheric Environment* 101, 166-176.
- Zilitinkevich, S.S., 1972. On the determination of the height of the Ekman boundary layer. *Boundary-Layer Meteorology* 3(2), 141-145.



Capítulo 4. Estimación de niveles de metales/metaloides en áreas urbanas mediante PLSR y ANNs

4.1. Estimación de los niveles de As, Cd, Ni y Pb en áreas urbanas de Cantabria (Castro Urdiales y Reinosa) a partir de campañas extensivas

El punto de partida de este trabajo consiste en el desarrollo de herramientas de evaluación de la calidad del aire para la estimación de los niveles de As, Cd, Ni y Pb en aire ambiente en pequeñas zonas urbanas en las que se han realizado campañas de muestreo extensivas. En concreto, se han seleccionado dos puntos de muestreo que permitan evaluar la aplicabilidad de estas herramientas en distintos puntos geográficos: en el litoral, Castro Urdiales, y en el interior, Reinosa. En ambas áreas urbanas las fuentes de emisión de contaminantes atmosféricos están bien identificadas, tanto en número como en variedad, por lo que las relaciones entre las fuentes de contaminantes minoritarios y sus niveles en aire ambiente se asumen a priori como simples.

4.1.1. Niveles de As, Cd, Ni y Pb en Castro Urdiales y Reinosa

Se ha caracterizado la composición de la fracción PM₁₀ del material particulado de muestras procedentes de campañas extensivas de muestreo llevadas a cabo en los años 2008, 2009 y 2010 en dos zonas urbanas de Cantabria: Castro Urdiales y Reinosa. A partir de esta caracterización, se han determinado los niveles en aire ambiente de As, Cd, Ni y Pb, que se muestran en la Figura 4.1. En la misma, se observa que los valores medios anuales de As, Cd, Ni y Pb en estas dos zonas durante el periodo considerado no solo no exceden el valor límite/objetivo (VL/VO), sino que se encuentran muy por debajo del umbral inferior de evaluación (UIE), y además no presentan variaciones significativas a lo largo del periodo de estudio. Por tanto, de acuerdo a la Directiva 2008/50/CE, se permite el uso de técnicas de modelización o de estimación objetiva para llevar a cabo la evaluación de la calidad del aire con respecto a dichos contaminantes en los lugares de muestreo como alternativa al procedimiento habitual basado en la toma de muestras experimentales para su medición posterior en laboratorio.

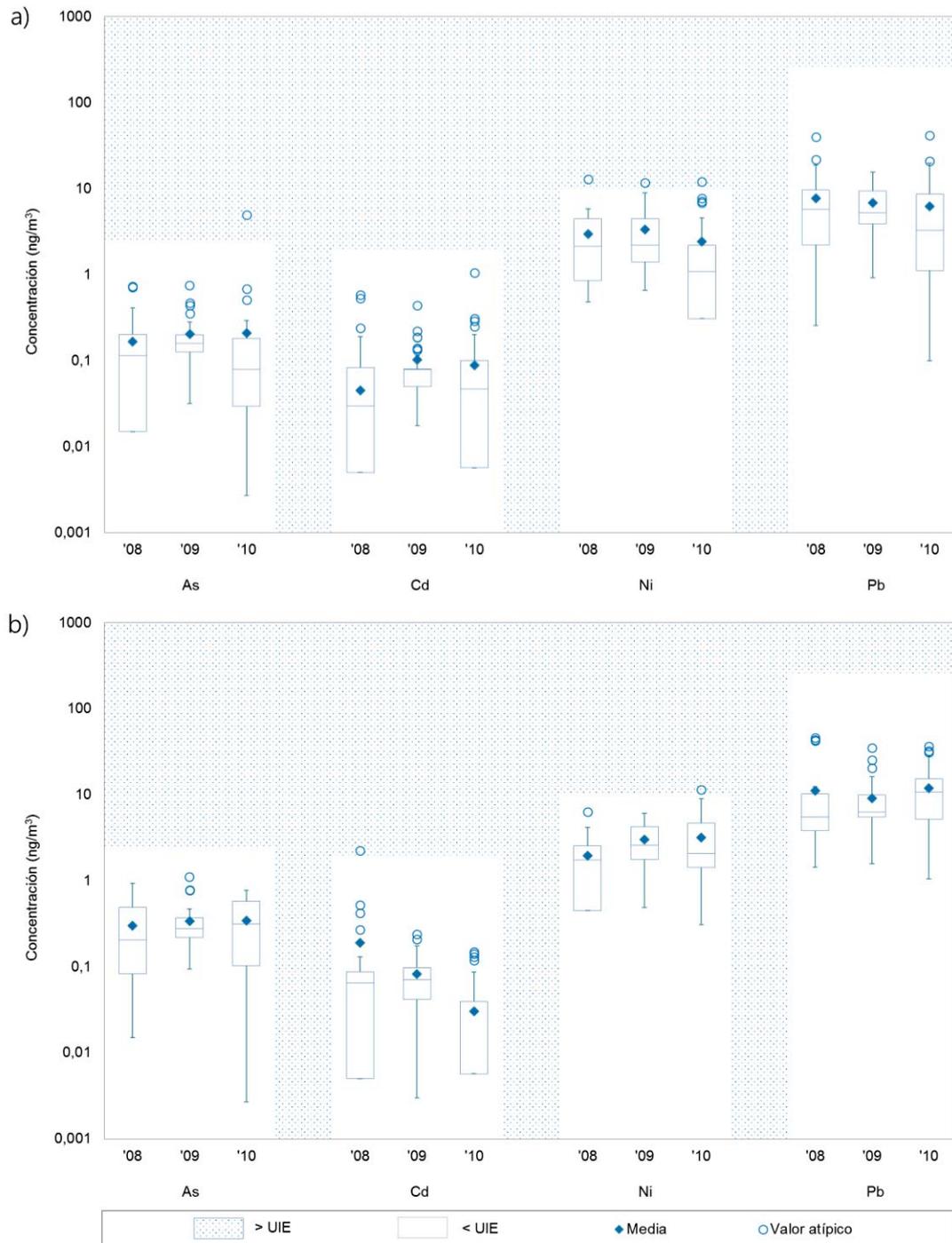


Figura 4.1. Diagrama de cajas de los niveles en aire ambiente de As, Cd, Ni y Pb en los años 2008 ('08), 2009 ('09) y 2010 ('10) en los lugares de muestreo de: a) Castro Urdiales y b) Reinos. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE. UIE = $2,4 \text{ ng/m}^3$ (As), 2 ng/m^3 (Cd), 10 ng/m^3 (Ni) y 250 ng/m^3 (Pb). Los valores medios correspondientes al año 2008 se han obtenido de Arruti y col. (2011a).

4.1.2. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: media y error de acuerdo a las Directivas de calidad del aire

En la Tabla 4.1. se recogen los índices estadísticos computados a partir de los resultados obtenidos por los mejores modelos desarrollados para la estimación de los niveles de As, Cd, Ni y Pb en Castro Urdiales y Reinosa para su evaluación como técnicas de estimación objetiva de acuerdo a las Directivas de calidad del aire. Dichos modelos han sido desarrollados en base a PLSR, ANNs y PCA previo a ANNs. Asimismo, se presentan tanto los resultados relativos a los subconjuntos de datos de entrenamiento (E) como de validación externa (V), para su comparación y para la evaluación de la habilidad de generalización de los modelos.

En primer lugar, es necesario validar las técnicas de estimación objetiva en el contexto de las Directivas europeas en términos de incertidumbre en la estimación. En este sentido, de acuerdo a [Arruti y col. \(2011b\)](#), se han considerado los índices descritos en el apartado 3.9.1: el *RME* y el *RDE*. Como se observa en la Tabla 4.1., para los cuatro contaminantes considerados, los valores de estos índices tanto para el subconjunto de entrenamiento/calibración como para el subconjunto de validación externa se encuentran por debajo del 100%, que es el máximo límite de incertidumbre en la estimación permitido para el uso de técnicas de estimación objetiva como método de evaluación de la calidad del aire con respecto a este tipo de contaminantes de acuerdo a la Directiva 2008/50/CE ([EC, 2008](#)). Aún más, en la mayoría de los casos, excepto para los modelos de ANNs desarrollados para la estimación de los niveles en aire ambiente de As y Cd, los valores de *RME* se encuentran por debajo del 50%, que constituye el requerimiento de incertidumbre para el uso de técnicas de modelización como método para la evaluación de la calidad del aire. No obstante, no se aprecian diferencias significativas entre los resultados obtenidos empleando las diferentes técnicas consideradas. Comparando con los *RME* obtenidos por [Arruti y col. \(2011b\)](#) para la estimación de los niveles en aire ambiente de As, Cd, Ni y Pb en las mismas ubicaciones por medio de regresión lineal múltiple, se observa un desempeño comparable en Castro Urdiales y un aumento generalizado en los valores de *RME* en Reinosa, de entre un 20 y un 30%. No obstante, conviene puntualizar que el estudio de [Arruti y col. \(2011b\)](#) se llevó a cabo únicamente con datos procedentes del año 2008 –a diferencia del presente trabajo en el que también se emplean datos de 2009 y 2010– y, además, se limitaba a su uso como subconjunto de datos de entrenamiento.

Por otra parte, en general, los valores de *RDE* se encuentran entre uno y dos órdenes de magnitud por debajo de los respectivos valores de *RME*. Esto concuerda con lo obtenido por [Arruti y col. \(2011b\)](#). En la formulación del *RDE* (ecuación 3.4) la diferencia entre observaciones y estimaciones se normaliza empleando el valor límite/objetivo en lugar de las propias observaciones, como ocurre en la formulación del *RME* (ecuación 3.3), por tanto, dado que en

Tabla 4.1. Evaluación de los modelos desarrollados en Castro Urdiales y Reinoso como herramientas de evaluación de la calidad del aire.

Contaminante	Modelo	Subconjunto ^a	Castro Urdiales						Reinoso								
			Incertidumbre de la UE			Concentración media ^b			Incertidumbre de la UE			Concentración media ^b					
			RME (%)	RDE (%)	\overline{C}_O (ng/m ³)	\overline{C}_E (ng/m ³)	FB 10 ²	RME (%)	RDE (%)	\overline{C}_O (ng/m ³)	\overline{C}_E (ng/m ³)	FB 10 ²	RME (%)	RDE (%)	\overline{C}_O (ng/m ³)	\overline{C}_E (ng/m ³)	FB 10 ²
As	PLSR	E	42,8	0,30	0,16	0,16	-4,2 10 ⁻¹²	28,2	1,25	0,33	0,33	4,1 10 ⁻¹⁴					
		V	34,8	1,74	0,13	0,13	-3,4	31,7	5,66	0,36	0,36	19,3					
	ANN	E	77,0	0,17	0,17	0,16	2,1	23,6	3,56	0,35	0,35	7,7					
		V	66,3	0,25	0,13	0,16	-24,6	35,4	6,01	0,36	0,36	24,5					
	PCA-ANN	E	54,6	0,86	0,17	0,15	9,1	25,0	1,41	0,38	0,37	2,7					
		V	33,5	0,96	0,13	0,15	-13,3	37,4	8,26	0,28	0,41	-38,3					
Cd	PLSR	E	40,9	0,14	0,08	0,08	-1,1 10 ⁻¹²	48,5	0,46	0,06	0,06	-3,8 10 ⁻¹³					
		V	46,2	0,86	0,09	0,07	25,6	46,2	0,59	0,04	0,07	-51,8					
	ANN	E	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	59,7	0,19	0,07	0,05	31,5					
		V	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	36,7	0,26	0,04	0,04	1,2					
	PCA-ANN	E	33,5	0,47	0,08	0,08	-1,1	67,5	7,02	0,06	0,07	-10,3					
		V	41,5	0,42	0,07	0,07	-1,7	34,3	0,40	0,07	0,07	-57,7					
Ni	PLSR	E	34,7	10,83	2,76	2,76	3,1 10 ⁻¹³	53,4	5,45	3,06	3,06	-3,9 10 ⁻⁶					
		V	22,1	2,64	1,89	2,27	-18,1	25,2	2,51	2,15	3,23	-39,9					
	ANN	E	45,0	6,19	2,83	2,33	19,3	38,9	22,21	3,36	3,46	-2,9					
		V	32,6	1,24	1,94	2,37	-20,0	28,2	2,51	2,15	3,61	-50,5					
	PCA-ANN	E	33,9	1,77	2,45	2,60	-5,8	26,9	9,01	3,08	2,84	8,3					
		V	25,6	2,06	2,36	2,34	0,9	42,5	20,85	2,60	2,09	21,4					
Pb	PLSR	E	26,7	0,09	6,98	6,98	-9,8 10 ⁻¹⁵	31,1	0,61	14,00	14,00	-6,2 10 ⁻¹⁴					
		V	49,6	0,71	10,33	7,53	31,4	14,0	0,21	12,78	15,03	-16,2					
	ANN	E	34,5	0,66	6,68	6,75	-0,9	30,3	0,83	15,03	14,28	5,2					
		V	32,0	1,76	10,85	8,70	22,0	35,9	1,07	12,78	14,88	-15,2					
	PCA-ANN	E	36,9	0,41	7,85	7,78	1,0	42,3	1,48	13,75	15,05	-8,9					
		V	30,6	0,07	7,70	8,58	-10,8	42,2	1,38	12,40	14,20	-13,4					

^aE: Entrenamiento; V: Validación externa^bO: Observada; E: Estimada^cn.c: No calculado

las zonas de estudio los niveles de concentración se encuentran muy por debajo del correspondiente valor límite/objetivo, resulta previsible obtener valores de *RDE* inferiores a los de *RME*. Aun teniendo en cuenta esta consideración, los resultados parecen indicar que la máxima diferencia de concentraciones no se obtiene para la observación con la concentración más próxima al valor límite/objetivo. Se observa un incremento en los valores de *RDE* en Reinosa, en especial para el As y el Ni, que son significativamente superiores a los obtenidos en Castro Urdiales. En cualquier caso, los valores de *RDE* se encuentran por debajo de un 15%, en Castro Urdiales, y de un 25%, en Reinosa. No obstante, estos índices presentan algunas limitaciones: es sabido que el *RME* es sensible a la presencia de valores atípicos, lo que redundaría en un incremento de los valores de incertidumbre (Flemming y Stern, 2007); además, ambos evalúan la incertidumbre sin considerar todas las muestras.

En vista de que los valores límite/objetivo establecidos en la legislación europea para los niveles en aire ambiente de los contaminantes objeto de estudio vienen dados como valores de concentración medios anuales, se ha de realizar especial hincapié en la evaluación del desempeño de los modelos de estimación en este sentido. Dentro de los índices estadísticos considerados en este trabajo (véase apartado 3.9.), el sesgo fraccional o *FB* es aquel que refleja el grado de precisión en la estimación de los valores medios. De forma general, el orden de magnitud obtenido para este índice es bajo, encontrándose por debajo de $2/3$ en valor absoluto en todos los casos. Por tanto, el error en la estimación del valor medio es inferior al 100%. Por otra parte, más de un 50% de los resultados obtenidos para este estadístico tiene un signo negativo, lo que indica un error por sobreestimación. En general, se observa que las estimaciones para el subconjunto de entrenamiento o calibración de los modelos presentan unos *FB* menores que los correspondientes al subconjunto de validación externa. Es decir, los modelos presentan un desempeño inferior en la estimación de la concentración en inmisión de As, Cd, Ni y Pb asociada a las muestras que no han sido empleadas en el desarrollo de los mismos que en el mero ajuste de los datos en la etapa de entrenamiento/calibración. Asimismo, dentro del subconjunto de entrenamiento, se observa que los modelos basados en PLSR proporcionan unos sesgos fraccionales muy inferiores a los proporcionados por las ANNs, ya sea de forma individual o precedidas por un PCA, ya que los valores medios de concentración de contaminantes estimados mediante los modelos PLSR son iguales —hasta dos cifras significativas— a los correspondientes valores observados, lo que según la ecuación 3.5 conduce a valores de *FB* más bajos. Eso indica que el valor medio es mejor ajustado mediante PLSR. Además, los sesgos fraccionales obtenidos mediante PLSR en este trabajo son comparables a los obtenidos mediante regresión lineal múltiple por Arruti y col. (2011b). No obstante, las diferencias entre las concentraciones medias estimadas empleando las tres técnicas consideradas en este trabajo y las observadas no son notablemente significativas. En lo que respecta a Reinosa, aunque hay diferencias evidentes entre las concentraciones medias

estimadas y observadas, el 90% de ellas no difiere más de un 50%. No obstante, en Reinososa se produce un incremento en los valores de *FB* en comparación con los de Castro Urdiales.

Desde un punto de vista práctico, para determinar la viabilidad del uso de estos modelos como técnicas de estimación objetiva para la evaluación de la calidad del aire con respecto a los contaminantes considerados, sería suficiente con que ofrezcan un desempeño apropiado en la validación externa de los mismos. En este sentido, de manera ilustrativa, en las Figuras 4.2. y 4.3. se representan los niveles de concentración media estimados en la validación externa de los modelos desarrollados, expresados en valor porcentual con respecto a los correspondientes valores límite/objetivo. El eje de ordenadas se presenta en escala logarítmica. El área verde representa la zona por debajo del UIE, el área amarilla representa la zona entre el UIE y el USE, y el área roja, la zona entre el valor límite/objetivo y el USE. Esta representación permite apreciar que si bien existen algunas diferencias entre los niveles estimados y observados, los mismos se encuentran muy próximos entre sí. Además, puesto que los niveles observados se encuentran dentro del área verde, muy por debajo del UIE, incluso si se obtuviera una mayor diferencia ésta podría ser aceptable. Por consiguiente, se puede decir que los modelos desarrollados proporcionan una estimación satisfactoria de los niveles medios de concentración para su uso como método de evaluación de la calidad del aire con respecto a los contaminantes en estudio.

4.1.3. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: otros estadísticos de desempeño de los modelos

Desde un punto de vista científico, además de una correcta estimación de los valores medios para cumplir con el marco normativo, es deseable que los modelos sean capaces de describir las variaciones temporales de las concentraciones en inmisión de los contaminantes estudiados. En este sentido, para evaluar el desempeño de los modelos, en este trabajo se ha utilizado un conjunto de índices estadísticos, detallados en el apartado 3.9.2. Los valores de dichos índices estadísticos computados a partir de los resultados obtenidos por los mejores modelos desarrollados para la estimación de los niveles de As, Cd, Ni y Pb en Castro Urdiales y Reinososa se muestran en la Tabla 4.2. El coeficiente de correlación, r , se ha empleado para cuantificar la linealidad del ajuste entre los valores estimados y observados. En Castro Urdiales, se observa que los coeficientes de correlación de los modelos PLSR, que se encuentran en el rango 0,6-0,7, son menos variables que los obtenidos por los modelos ANN y PCA-ANN. No obstante, en Reinososa presentan mayor variabilidad. Sin embargo, los coeficientes de correlación obtenidos con los modelos de ANNs son en general iguales o superiores a los obtenidos con PLSR y PCA-ANNs. Aún más, el mayor coeficiente de correlación ($r = 0,82$) se ha obtenido en el ajuste de los modelos de ANNs desarrollados para el Pb en ambas ubicaciones. Comparando estos resultados con los de otros estudios de índole similar reportados en la literatura (véase

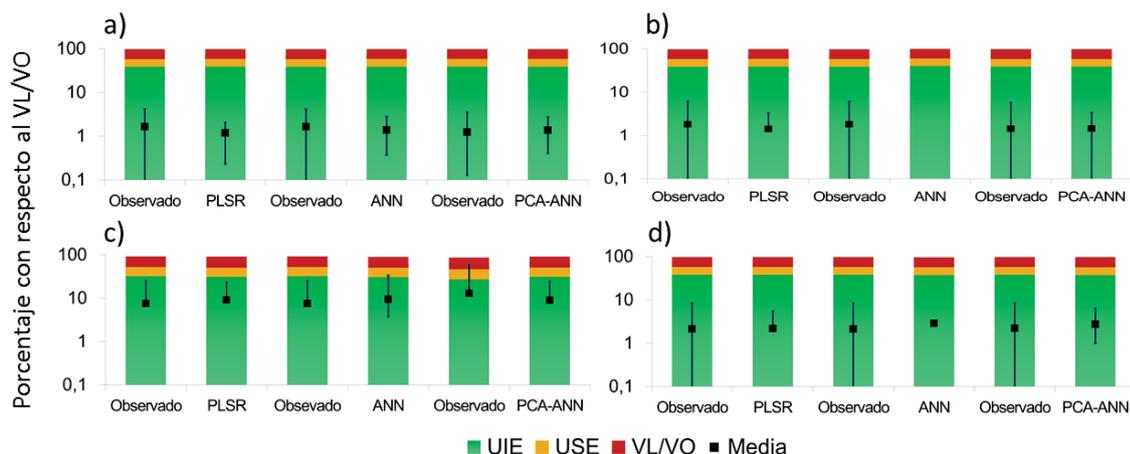


Figura 4.2. Comparación entre los niveles medios de concentración estimados y observados obtenidos en Castro Urdiales para el subconjunto de validación externa con respecto a los correspondientes umbrales de evaluación y valores objetivo/límite: a) As, b) Cd, c) Ni y d) Pb. VO/VL = 6 ng/m³ (As), 5 ng/m³ (Cd), 20 ng/m³ (Ni), 500 ng/m³ (Pb); USE = 60 % (As y Cd) y 70% (Ni y Pb); UIE = 40% (As y Cd) y 50% (Ni y Pb).

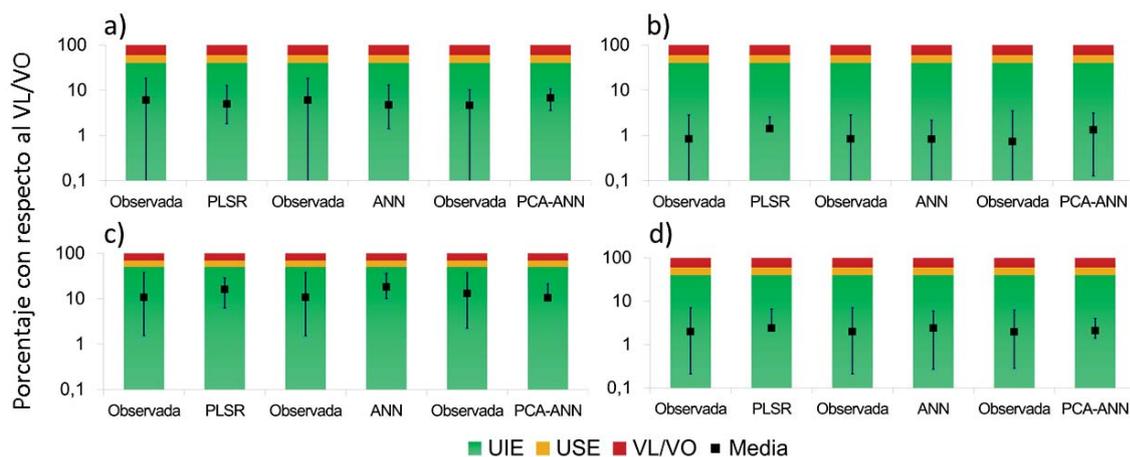


Figura 4.3. Comparación entre los niveles medios de concentración estimados y observados obtenidos en Reinosa para el subconjunto de validación externa con respecto a los correspondientes umbrales de evaluación y valores objetivo/límite: a) As, b) Cd, c) Ni y d) Pb. VO/VL = 6 ng/m³ (As), 5 ng/m³ (Cd), 20 ng/m³ (Ni), 500 ng/m³ (Pb); USE = 60 % (As y Cd) y 70% (Ni y Pb); UIE = 40% (As y Cd) y 50% (Ni y Pb).

apartado 1.5.2), se observa, en primer lugar, que los coeficientes de correlación obtenidos en este trabajo son en general inferiores a los obtenidos por los modelos estadísticos basados en técnicas lineales como la regresión lineal múltiple desarrollados por Arruti y col. (2011b) para la estimación de los niveles en aire ambiente de As, Cd, Ni y Pb en las mismas ubicaciones. Los coeficientes de correlación obtenidos en dicho estudio fueron: 0,6 (As), 0,8 (Cd y Ni) y 0,9 (Pb) en Castro Urdiales; y 0,9 (As y Pb) y 0,8 (Cd y Ni) en Reinosa. Por otro lado, En lo referente a los modelos de series de tiempo, Hernández y col. (1992) reportaron un coeficiente de correlación de 0,5 para la estimación de Pb en Madrid y Vicente y col. (2012) reportaron unos coeficientes de correlación de 0,7, 0,8, 0,6 y 0,8 para el As, Cd, Ni y Pb, respectivamente, en Castellón. No

obstante, como se ha mencionado en el apartado 4.1.2, el estudio de [Arruti y col. \(2011b\)](#) se llevó a cabo únicamente con datos procedentes del año 2008 y, además, al igual que los estudios de [Hernández y col. \(1992\)](#) y [Vicente y col. \(2012\)](#), se limitó únicamente a un entrenamiento, es decir, el desarrollo de los modelos se llevó a cabo sin emplear técnicas de validación cruzada o interna para evitar el sobreajuste y favorecer la capacidad de generalización de los mismos, la cual tampoco fue evaluada por medio de un procedimiento de validación externa. Si bien un procedimiento de validación cruzada o interna empobrece en cierta medida el ajuste, lo que redundará en un ligero empeoramiento de los estadísticos de desempeño del entrenamiento o calibración de los modelos, su consideración es recomendable ya que resulta conveniente para no comprometer la capacidad de generalización para una futura aplicación de los modelos. Además, los modelos de [Hernández y col. \(1992\)](#) y [Vicente y col. \(2012\)](#) fueron desarrollados a partir de bases de datos compuestas por 400 y 900 muestras, respectivamente, y, por tanto, de tamaño entre 3 y 6 veces mayor que el de las empleadas en el presente estudio.

Por otra parte, como era previsible, los coeficientes de correlación referentes al subconjunto de validación externa son inferiores a los del subconjunto de entrenamiento/calibración. Esto concuerda con lo obtenido en otros trabajos como el de [Chelani y col. \(2012\)](#), que obtuvieron unos coeficientes de correlación de 0,7 para Cd, Ni y Pb en la validación externa de modelos desarrollados en base a redes neuronales. No obstante, en dicho estudio, el coeficiente de correlación se obtuvo para evaluar el grado de asociación entre las concentraciones medias mensuales observadas y estimadas, ambas calculadas a partir de valores de concentración media diaria, pero no para evaluar la asociación de las concentraciones diarias directamente. Como se observa en la Tabla 4.1. y se ha discutido en el apartado 4.1.2., las redes neuronales tienden a proporcionar buenas estimaciones de los valores medios.

Por su parte, los coeficientes de correlación para los modelos PCA-ANN se encuentran sistemáticamente por debajo de 0,5.

El hecho de que los valores de coeficiente de correlación sean próximos a la unidad, no garantiza que los valores observados y los estimados por un modelo concuerden unos con otros, sino únicamente de que existe una relación de dependencia lineal entre ellos, independientemente de que los resultados del modelo tiendan a la subestimación o a la sobreestimación (Figura 4.4.). Por ello, aparte del coeficiente de correlación, la precisión en la estimación de las concentraciones de contaminante relativas a las muestras individuales se cuantifica por medio de los estadísticos *RMSE*, *NMSE* y *FV* (ecuaciones 3.7, 3.8 y 3.9). Los valores de *RMSE* proporcionan información en cuanto al error en la estimación de las concentraciones en inmisión observadas, es decir, en cuanto a las diferencias entre los valores observados y estimados. No obstante, los resultados obtenidos para distintos contaminantes no son

Tabla 4.2. Estadísticos para la evaluación del desempeño de los modelos desarrollados en Castro Urdiales y Reinosa.

Contaminante	Modelo	Subconjunto ^a	Desempeño Castro Urdiales			Desempeño Reinosa				
			r	RMSE 10 (ng/m ³)	NMSE 10	FV 10	r	RMSE 10 (ng/m ³)	NMSE 10	FV 10
As	PLSR	E	0,656	1,24	6,02	4,16	0,446	2,08	4,07	7,67
		V	0,629	1,11	7,34	0,88	0,441	3,04	8,51	7,23
	ANN	E	0,130	1,67	10,22	12,74	0,765	1,62	2,35	4,89
		V	0,193	1,37	8,99	11,26	0,393	3,16	9,71	6,78
	PCA-ANN	E	0,536	1,41	7,80	7,36	0,572	2,41	4,02	6,69
		V	0,190	1,37	9,30	2,77	0,132	2,42	5,21	2,75
Cd	PLSR	E	0,672	0,67	8,03	3,92	0,644	0,66	1,16	4,32
		V	0,628	0,68	7,30	4,69	0,338	0,55	1,01	3,26
	ANN	E	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	0,641	0,75	1,67	7,03
		V	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	0,521	0,39	8,70	4,30
	PCA-ANN	E	0,613	0,62	6,49	5,36	0,518	0,82	1,47	8,42
		V	0,534	0,62	7,48	5,05	0,579	0,50	1,01	2,74
Ni	PLSR	E	0,642	21,52	6,07	4,36	0,386	20,49	4,48	8,86
		V	0,663	12,27	3,51	-0,97	0,549	19,23	5,32	6,13
	ANN	E	0,676	21,59	7,08	6,18	0,460	20,92	3,78	8,54
		V	0,387	16,15	5,67	-0,58	0,455	22,71	6,63	4,10
	PCA-ANN	E	0,643	17,60	4,85	4,92	0,677	17,43	3,48	3,09
		V	0,216	21,19	8,11	1,64	0,304	22,50	9,32	5,53
Pb	PLSR	E	0,704	44,25	4,04	3,48	0,723	87,00	3,86	3,21
		V	0,620	67,75	5,90	7,44	0,553	94,00	4,60	0,17
	ANN	E	0,820	36,00	2,88	1,98	0,829	71,50	2,38	3,38
		V	0,676	60,25	3,85	2,47	0,563	89,50	4,22	1,05
	PCA-ANN	E	0,681	50,75	4,25	4,41	0,679	94,25	4,29	7,17
		V	0,269	65,00	6,41	3,60	0,374	95,00	5,13	9,70

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

^cn.c.: No calculado

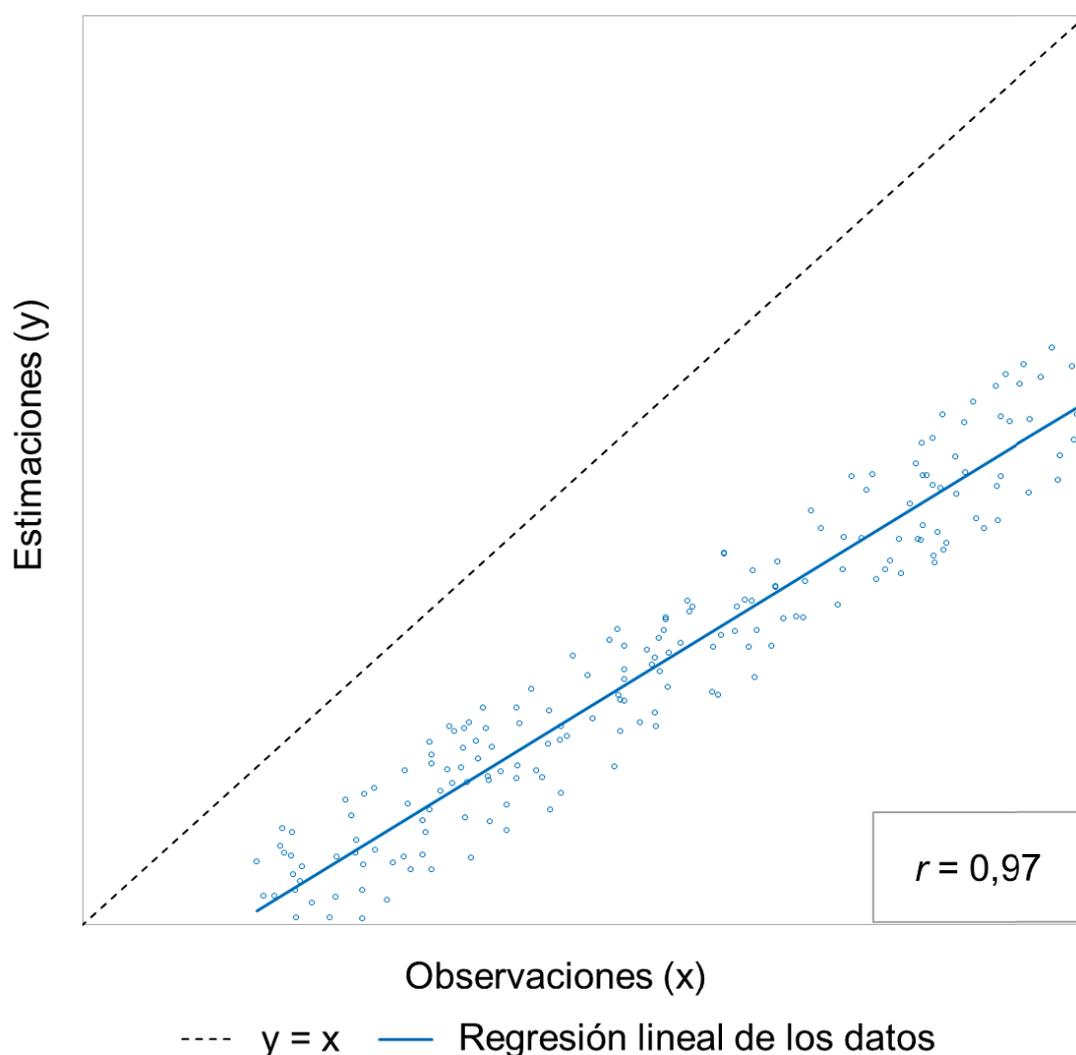


Figura 4.4. Ejemplificación gráfica de subestimación a pesar del ajuste prácticamente lineal de los datos ($r = 0,97$).

comparables entre sí de manera directa, pues dependen del rango en el que se encuentren las observaciones. Por ello, es preferible considerar una versión normalizada de este parámetro, el *NMSE*, que no tiene en cuenta el rango de la variable dependiente. En términos generales, los tres modelos considerados proporcionan valores bajos de *NMSE* en ambas ubicaciones, los cuales se encuentran dentro del mismo rango, del orden de 10^{-1} , en concordancia con los resultados de *NMSE* obtenidos por Arruti y col. (2011b) en las mismas ubicaciones. No obstante, los modelos desarrollados para la estimación de As y Cd en Castro Urdiales presentan unos *NMSE* ligeramente superiores a los obtenidos en Reinosá. Esto puede ser debido a una mayor complejidad en cuanto al número y variedad de fuentes próximas al área urbana de Castro Urdiales, así como al hecho de que los niveles de estos dos contaminantes, al contrario que los de Ni y Pb, se encuentran por debajo del límite de detección del método analítico en un gran número de muestras. En consecuencia, los modelos son entrenados para producir el mismo

valor de salida o *output* a partir de diferentes valores de entrada o *inputs*, una contradicción que puede afectar negativamente a la calibración de los modelos para la estimación del resto de muestras.

Con respecto al índice *FV*, que evalúa la adecuación en la estimación de la varianza, los valores positivos indican que la varianza estimada es inferior a la observada. De acuerdo a la Tabla 4.2, la mayoría de los modelos presentan valores de *FV* positivos, y por tanto los valores estimados presentan una menor dispersión que los observados, que tienden a estar más distanciados del valor medio. En general, los valores de *FV* son ligeramente superiores en Reinoso, aunque inferiores a 1, que representa el 50% de la varianza observada.

Este hecho, junto con unos *FB* mayoritariamente positivos, correspondientes a una ligera subestimación del valor medio, indica que hay ciertas limitaciones en la capacidad del modelo para describir todas las variaciones de concentración, especialmente los valores extremos. No obstante, a pesar de que no se encuentran diferencias sustanciales entre la PLSR y las ANNs, en general ambas técnicas son capaces de capturar la tendencia subyacente de los datos y proporcionar variaciones temporales con una forma similar a la de los valores observados, como se observa en la Figura 4.5. en la que se representa el ajuste de las concentraciones de Ni y Pb, normalizadas por medio del UIE, en la estación de Castro Urdiales.

En base a los resultados obtenidos, no se ha encontrado ninguna mejora asociada a la consideración de una técnica para la reducción de dimensiones como PCA antes del desarrollo de los modelos de ANNs. Esto puede ser debido al hecho de que las ANNs son menos sensibles que otras técnicas a los problemas asociados a la dimensionalidad, puesto que pueden concentrarse en una sección dimensional inferior dentro del espacio, lo que se puede lograr ignorando por completo un *input*, estableciendo los pesos correspondientes a cero. Por tanto, para esta aplicación en concreto la reducción de dimensiones no ha resultado ser una estrategia efectiva, puesto que la eliminación del análisis de variables de entrada conlleva una pérdida de capacidad predictiva del modelo.

Por otra parte, dado que estos modelos son concebidos para ser usados cuando los niveles de contaminantes son suficientemente bajos en una ubicación en particular, en principio una moderada imprecisión para estimar los valores máximos no debería suponer ningún inconveniente para considerar estos modelos como herramientas válidas, siempre que se cumpla con los requerimientos de la regulación. Sin embargo, es posible cierto refinamiento de los modelos, ya que como se ha mencionado, presentan ciertas dificultades a la hora de estimar los valores extremos de las concentraciones observadas, que tienden a la subestimación. En relación a esto, sería recomendable la consideración de variables de entrada adicionales o la ampliación de la base de datos a partir de la determinación de nuevas muestras correspondientes a otros periodos de tiempo.

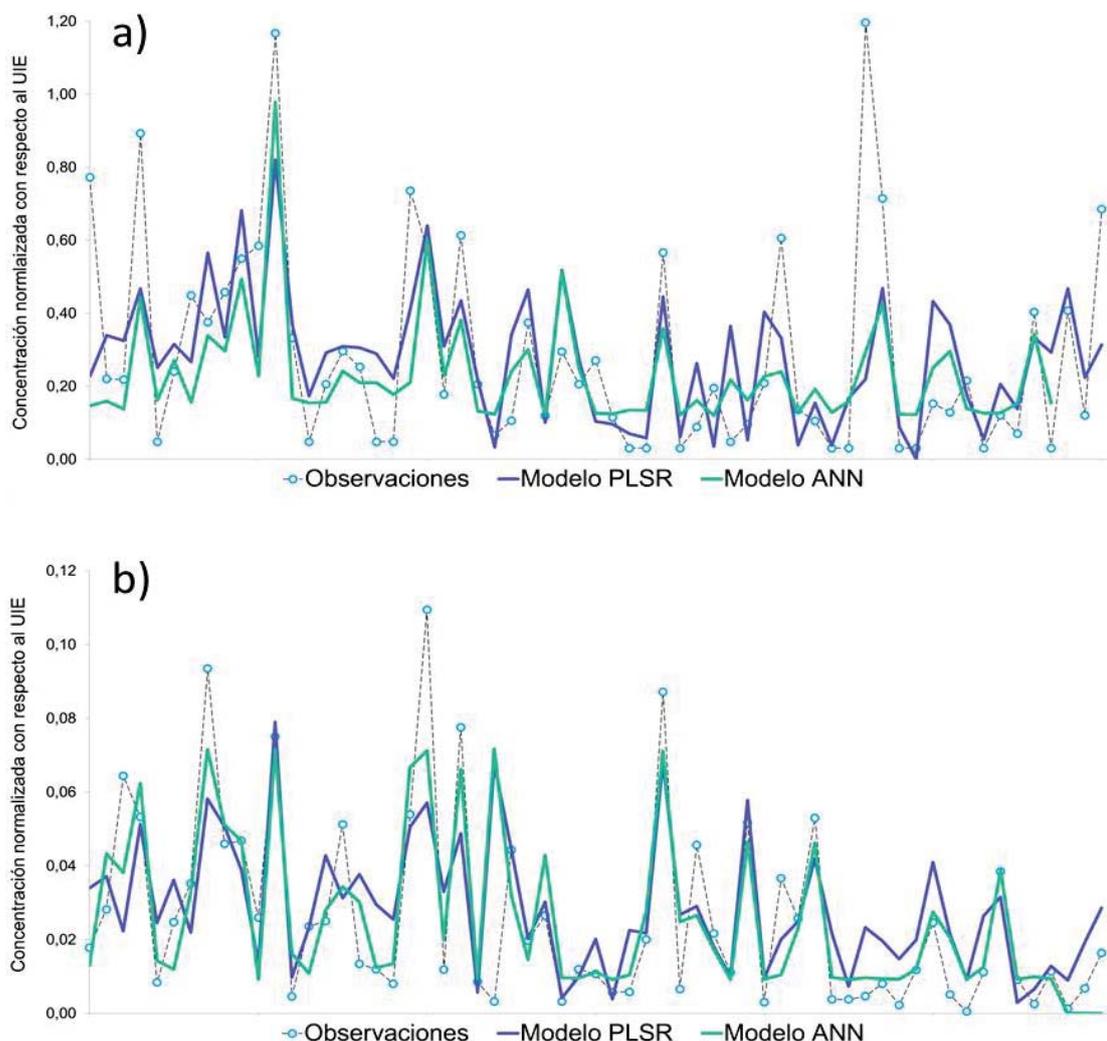


Figura 4.5. Ajuste de los datos del subconjunto de entrenamiento de modelos PLSR y ANN desarrollados en Castro Urdiales para la estimación de los niveles de: a) Ni y b) Pb.

Los resultados muestran que estos modelos presentan un desempeño aceptable en diferentes áreas de la región, incluso en Castro Urdiales, que presenta un marco de contaminación más complejo que Reinosa, en el que intervienen distintas fuentes de emisión. Sin embargo, puesto que los modelos han sido entrenados con datos de lugares concretos y habiéndose demostrado que la precisión en la estimación depende de la ubicación específica, estos modelos solo se pueden utilizar con confianza en los lugares en los que han sido desarrollados. Esta dependencia es especialmente pronunciada en los modelos de ANNs, que presentan una mayor variabilidad en los resultados que los modelos PLSR y PCA-ANN. Esto puede ser debido al hecho de que se ha utilizado un número de muestras limitado para el desarrollo de los modelos debido a la no disponibilidad de observaciones adicionales a consecuencia del coste y tiempo requerido para el muestreo. Por tanto, se puede concluir que para conjuntos de datos de dimensiones moderadas, las técnicas de regresión lineal pueden funcionar igual de bien que técnicas no lineales en cuanto a la estimación de los niveles de metales y semimetales en aire ambiente.

4.2. Estimación de los niveles de metales regulados (Ni y Pb) y no regulados (Cr, Mn y V) en Dunkerque (Francia) a partir de una campaña intensiva

Este apartado se centra en el desarrollo de herramientas de evaluación de la calidad del aire para la estimación de los niveles de Ni y Pb en aire ambiente en zonas urbanas en las que se han realizado campañas de muestreo intensivas, es decir, aquellas campañas de muestreo sin apenas intervalos de interrupción entre toma de muestras durante un breve periodo de tiempo. La evaluación del desempeño de estas herramientas en la estimación de los niveles relativos a campañas intensivas resulta de interés para determinar la capacidad de los mismos para predecir eventos puntuales de contaminación.

Por otra parte, en este estudio, se aplica la metodología no solo al conjunto de metales y semimetales cuyos niveles están regulados por la UE, sino también a otros metales no regulados por la UE pero que sí son recogidos en las guías de la OMS (Cr, Mn y V).

4.2.1. Niveles de los metales estudiados

Entre los meses de febrero y mayo de 2008 la *Unité de Chimie Environnementale et Interactions sur le Vivant* (UCEIV) llevó a cabo una campaña intensiva de muestreo de PM₁₀ en la ciudad de Dunkerque (Francia) para la posterior determinación de los niveles en aire ambiente de los elementos Al, Ca, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Sn, Ti, V y Zn. En la Figura 4.6. se muestra un diagrama de cajas de los niveles de los contaminantes objeto de estudio en la presente Tesis Doctoral, a excepción del As y el Cd, cuyos niveles no fueron determinados. En la misma, se observa que las concentraciones medias de Ni y Pb obtenidas durante el periodo de estudio en la estación de Dunkerque se encuentran por debajo de los umbrales inferiores de evaluación establecidos en la legislación. Asimismo, la concentración media de Mn en aire se encuentra por debajo de los valores guía de calidad del aire propuestos por la OMS. No obstante, la concentración de algunas muestras, en particular de Ni y Mn exceden ampliamente estos valores: como se muestra, las concentraciones observadas máximas de Ni y Mn exceden en 10 y 6 veces los correspondientes valores de UIE y UIE-equivalente. Por consiguiente, se debería prestar atención especial al desempeño de los modelos en cuanto a su capacidad para estimar dichos valores puesto que las exposiciones a niveles elevados de estos metales pueden tener efectos negativos sobre la salud humana. Se ha demostrado que la inhalación de manganeso puede producir efectos neurotóxicos que varían desde trastornos neurofísicos y motores (Mergler y col., 1999), trastornos en la estabilidad postural (Hernández-Bonilla y col., 2011) y un aumento del riesgo de padecimiento de la enfermedad de Parkinson (Finkelsteinn y Jerret, 2007) para exposiciones a concentraciones bajas (cercanas a los 50 ng/m³) hasta un trastorno del movimiento conocido como manganismo a concentraciones superiores a 1 mg/m³ (Aschner y

col., 2005). En lo que respecta al vanadio, sus efectos tóxicos dependen de su grado de oxidación y pueden incluir irritación del tracto respiratorio, cambios hematológicos y bioquímicos y lesiones funcionales en ciertos órganos (Ghosh y col., 2015).

Cabe destacar que los valores medios en el caso de estudio de Dunkerque no son en realidad concentraciones medias anuales puesto que los datos a partir de los cuales se calculan corresponden a una campaña intensiva de muestreo de PM₁₀, de únicamente tres meses de duración. En dicha campaña los periodos de muestreo variaron de 6 a 14 horas. En principio, esto habría de suponer una ventaja para el desarrollo de los modelos estadísticos de estimación puesto que el error en el promedio de los *inputs* (en especial de variables de valor tan cambiante como la dirección y velocidad del viento) es tanto menor cuanto más reducidos sean los periodos de muestreo y, por tanto, habría de esperar una mayor correspondencia entre los *inputs* y los *outputs*. No obstante, las concentraciones en inmisión de contaminantes obtenidas como resultado de esta campaña presentan una variabilidad acrecentada en comparación con la que presentan los niveles procedentes de las campañas de muestreo extensivas analizados en el apartado 4.1.1. Este hecho ha de ser tenido en cuenta a la hora de interpretar los resultados puesto que supone una dificultad adicional para la estimación.

Los estudios llevados a cabo por Hleis (2010) y Kfoury (2013) han puesto de manifiesto que los niveles de Ni, Pb, Mn, V, Cr y otros metales y semimetales en Dunkerque están asociados a la industria puesto que se han identificado como trazadores de las actividades de la industria local. Los resultados de las rosas de polución y de los métodos de contribución de fuentes por medio de la factorización matricial no negativa indican que las emisiones de Pb se deben principalmente a la siderurgia integral, que también es una fuente de emisión de Mn (Hleis, 2010; Kfoury, 2013). Asimismo, las emisiones de las plantas de producción de ferromanganeso, principales responsables de los niveles de Mn, también influyen los niveles de Pb (Hleis, 2010). Ni y V son trazadores de la combustión de petróleo puesto que explican el 72% y el 86% de las concentraciones observadas, respectivamente (Kfoury, 2013). Con respecto al Cr, se le considera como trazador de la actividad de una planta siderúrgica de horno eléctrico, aunque los niveles de Cr se deben también parcialmente a procesos de combustión. La fuerte presencia de actividades industriales en Dunkerque y la firme conexión entre los niveles de metales en aire ambiente y la industria local hacen de la ciudad un lugar adecuado para el desarrollo de modelos de estimación de niveles de metales puesto que las entradas a dichos modelos consisten parcialmente en concentraciones de contaminantes atmosféricos mayoritarios, que, indudablemente, están igualmente relacionados con la industria local.

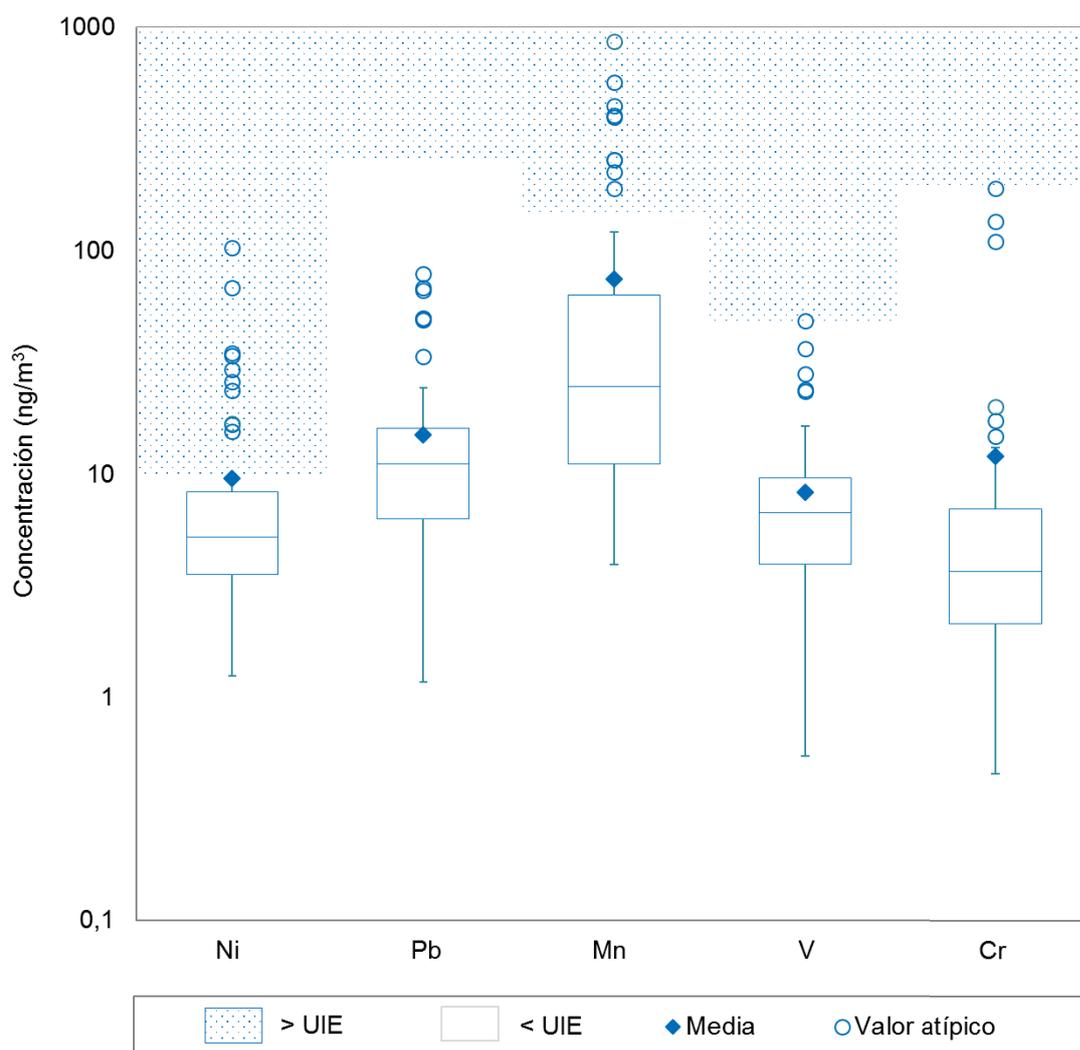


Figura 4.6. Diagrama de cajas de los niveles de Ni, Pb, Cr, Mn y V medidos en Dunkerque en el periodo de estudio. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE o UIE-equivalente. UIE = 10 ng/m^3 (Ni) y 250 ng/m^3 (Pb); UIE-equivalente = 150 ng/m^3 (Mn) y valor máximo observado (V y Cr).

4.2.2. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: media y error de acuerdo a las Directivas de calidad del aire

4.2.2.1. Estimación de metales regulados: Ni y Pb

En la Tabla 4.3. se presentan los índices estadísticos computados a partir de los resultados obtenidos por los mejores modelos desarrollados para la estimación de los niveles de los dos metales regulados por la UE considerados, Ni y Pb, en Dunkerque empleando las tres técnicas

Tabla 4.3. Evaluación de los resultados de los modelos desarrollados en Dunkerque para los metales regulados (Ni y Pb) desde el punto de vista de una herramienta de evaluación de la calidad del aire.

Contaminante	Modelo	Subconjunto ^a	Incertidumbre de la UE		Concentración media ^b		
			RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	FB 10 ²
Ni	PLSR	E	65,9	12,87	6,85	6,85	-7,510 ⁻¹¹
		V	83,6	11,70	15,66	9,82	45,8
	ANN	E	29,2	18,55	7,34	7,38	-0,5
		V	50,0	17,60	15,66	11,58	30,0
	PCA-ANN	E	64,9	24,86	9,59	9,58	0,1
		V	42,6	2,50	6,86	9,49	-32,2
Pb	PLSR	E	28,1	1,44	16,30	16,30	3,7 10 ⁻⁰⁸
		V	31,9	0,31	18,65	22,20	-17,4
	ANN	E	18,3	2,10	15,95	17,10	-7,0
		V	54,0	0,54	18,65	20,78	-10,8
	PCA-ANN	E	40,1	2,22	16,43	16,95	-3,2
		V	90,4	1,38	9,23	19,88	-73,1

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

PLSR, ANNs y PCA-ANN, para la evaluación de su uso como técnicas de estimación objetiva como método de evaluación de la calidad del aire de acuerdo a las Directivas europeas. Para cada modelo se presentan los resultados relativos a los subconjuntos de entrenamiento/calibración (E) y validación externa (V).

En primer lugar, para la validación de los modelos como métodos para la evaluación de la calidad del aire desde un punto de vista de la regulación en materia de calidad del aire, que constituye el objetivo principal de este trabajo, la estimación de los valores medios y la conformidad del cumplimiento de los requerimientos de incertidumbre constituyen dos aspectos fundamentales a tener en cuenta. En este sentido, al mismo tiempo que se cumplen con los requerimientos de incertidumbre, con valores de *RME* y *RDE* por debajo del 100%, los modelos desarrollados para la estimación de los niveles en aire ambiente de Ni y Pb en Dunkerque son capaces de proporcionar una estimación aproximada de las concentraciones medias en base a los bajos valores obtenidos para el índice *FB*. En la etapa de entrenamiento, los modelos PLSR presentan mejores *FB* que los modelos ANNs y PCA-ANN. La razón por la cual los *FB* de los modelos PLSR son tan bajos es que las concentraciones medias estimadas por dichos modelos y las observadas son prácticamente las mismas, lo que da lugar a que este índice esté próximo a su valor ideal, que de acuerdo a la ecuación 3.5, es cero. No obstante, se observa que en cualquier caso el resto de valores de *FB* no son significativamente más elevados, puesto que las concentraciones medias estimadas están próximas a las correspondientes concentraciones observadas, en todos los casos.

Los resultados de los modelos desarrollados para el Ni proporcionan unos índices *RME*, *RDE* y *FB* en general superiores a los que se obtienen mediante los modelos desarrollados para el Pb. Además, el signo del *FB* de los modelos desarrollados para el Ni, mayoritariamente positivo, indica que hay una cierta tendencia a la subestimación de los valores medios, al contrario que en el caso del Pb. Esto puede ser debido a la mayor variabilidad de los niveles de Ni en comparación con los del Pb. El error por defecto en la estimación de las concentraciones más elevadas de Ni contribuye a una disminución del valor medio del conjunto de estimaciones. Además, como estas se encuentran más alejadas de su valor medio que las respectivas concentraciones de Pb, contribuye igualmente al aumento de las diferencias entre las concentraciones observadas y estimadas consideradas en los índices de incertidumbre con respecto al caso del Pb, lo que eleva consecuentemente el valor del *RME* y del *RDE*.

Se ha de advertir que los datos señalados como valores atípicos en la Figura 4.6. no han de coincidir necesariamente con los valores atípicos identificados y no considerados finalmente para el desarrollo de los modelos. La consideración de valores extremos como atípicos en el diagrama de cajas responde a un análisis de los datos univariante. En este trabajo se ha

empleado un método de identificación de valores atípicos multivariante, teniendo en cuenta las interacciones entre las distintas variables consideradas en el problema.

4.2.2.2. Estimación de metales no regulados: Cr, Mn y V

Los índices estadísticos para la evaluación de los modelos basados en PLSR, ANNs y PCA previo a ANNs como métodos para la evaluación de la calidad del aire obtenidos a partir de los resultados de los mejores modelos desarrollados para Cr, Mn y V se recogen en la Tabla 4.4. A pesar de que los niveles en aire ambiente de estos contaminantes no se encuentran regulados por las Directivas de calidad del aire de la UE, el análisis de resultados para la evaluación del uso de los modelos como método de evaluación de la calidad del aire se ha llevado a cabo de la misma manera que para los metales regulados. Sin embargo, puesto que estos contaminantes carecen de un valor límite/objetivo normativo, se ha calculado un índice de incertidumbre equivalente al *RDE* (*RDE_{eq}*) en base a los valores equivalentes al UIE considerados en el apartado 1.2. Se ha de tener en cuenta además, que con estas suposiciones, los niveles medios de Cr, Mn y V, a diferencia de los de Ni y Pb, se encuentran mucho más próximos a su correspondiente UIE equivalente.

Se puede apreciar el cumplimiento de los requerimientos de incertidumbre en el contexto de las Directivas europeas, con valores de *RME* y *RDE_{eq}* inferiores a 100% en todos los casos, excepto el *RME* correspondiente a la validación externa del modelo PCA-ANN desarrollado para el Cr. No obstante, puesto que estos metales se encuentran en concentraciones más elevadas, e.g. el valor máximo observado del Mn supera en casi 6 veces el valor guía de calidad del aire establecido por la OMS y usado como equivalente al UIE (150 ng/m³) con un valor de 872,8 ng/m³ en el periodo de estudio (Figura 4.6.), quizá se deberían considerar requerimientos de incertidumbre más restrictivos para estos metales ya que permitir un 100% de incertidumbre en la estimación podría llevar a considerar como aceptable una subestimación de niveles de contaminación potencialmente peligrosos.

Con respecto a las concentraciones medias, valores bajos de *FB* tanto para los subconjuntos de entrenamiento/calibración como para el de validación externa indican que los modelos constituyen una herramienta aceptable. En general, estos valores se encuentran dentro del mismo orden de magnitud que el de los obtenidos para los metales regulados, por debajo de 0,5, a excepción del *FB* correspondiente a la validación externa del modelo PCA-ANN desarrollado para el Cr, con un valor de -1,46. Por tanto, excepto en este caso, se cumplen los requerimientos mínimos de calidad propuestos por Kumar y col. (1993).

Tabla 4.4. Evaluación de los resultados de los modelos desarrollados en Dunkerque para los metales no regulados (Cr, Mn y V) desde el punto de vista de una herramienta de evaluación de la calidad del aire.

Contaminante	Modelo	Subconjunto ^a	Incertidumbre de la UE		Concentración media ^b		FB 10 ²
			RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	
Cr	PLSR	E	88,8	1,01	8,74	7,55	14,6
		V	78,5	25,74	13,29	11,02	18,7
	ANN	E	50,0	39,17	10,66	6,20	53,1
		V	83,6	27,89	13,29	6,77	65,0
	PCA-ANN	E	79,2	0,43	13,64	15,07	-10,0
		V	489,9	0,43	2,77	17,76	-146,3
Mn	PLSR	E	53,9	4,71	48,83	50,15	-2,7
		V	53,1	50,23	97,43	87,36	10,9
	ANN	E	52,6	60,51	50,42	32,07	44,5
		V	48,2	68,86	97,43	95,66	1,8
	PCA-ANN	E	66,4	78,51	69,08	69,08	0,0
		V	35,3	11,62	44,24	45,50	-2,8
V	PLSR	E	42,7	1,20	6,43	6,58	-2,3
		V	31,5	4,74	8,86	9,06	-2,2
	ANN	E	41,6	4,28	6,99	6,86	1,9
		V	30,7	5,45	8,86	9,04	-2,1
	PCA-ANN	E	42,9	1,60	6,44	6,47	-0,5
		V	12,5	15,60	6,27	8,09	-25,4

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

4.2.3. Evaluación de los modelos estadísticos desarrollados como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire: otros estadísticos de desempeño de los modelos

4.2.3.1. Estimación de metales regulados: Ni y Pb

Los valores de los índices estadísticos para evaluar el desempeño de los modelos en cuanto a la estimación de concentraciones de muestras computados a partir de los resultados obtenidos por los mejores modelos desarrollados para la estimación de los niveles de Ni y Pb en Dunkerque se muestran en la Tabla 4.5.

De forma general, se observa un desempeño similar al obtenido en la estimación de los niveles de metales y semimetales regulados por la UE asociados a campañas extensivas de muestreo. Los coeficientes de correlación de los modelos PLSR y ANN se encuentran dentro del rango 0,5-0,9, lo que indica cierta tendencia de los valores estimados a variar en la misma medida que los observados. Por su parte, los índices *NMSE* y *FV* obtenidos para el Pb se encuentran dentro del mismo orden de magnitud que los obtenidos en Castro Urdiales y Reinososa (Tabla 4.2.). Los valores del índice *NMSE* obtenido por los modelos desarrollados para el Ni, son ligeramente superiores a los obtenidos para el Pb y a sus análogos del apartado 4.1., encontrándose por encima de 0,5. De nuevo, esto puede ser debido a la mayor variabilidad que presentan los niveles de Ni en Dunkerque durante el periodo de estudio.

Tabla 4.5. Índices estadísticos para la evaluación del desempeño de los modelos de estimación de Ni y Pb desarrollados en Dunkerque.

Metal	Modelo	Subconjunto ^a	Desempeño			
			r	RMSE (ng/m ³)	NMSE 10	FV 10
Ni	PLSR	E	0,560	8,05	13,81	5,64
		V	0,556	24,13	37,80	14,62
	ANN	E	0,873	5,42	5,37	1,44
		V	0,702	18,66	19,23	4,77
	PCA-ANN	E	0,470	16,11	28,27	7,20
		V	0,443	6,32	6,13	-3,04
Pb	PLSR	E	0,823	9,85	3,65	1,95
		V	0,837	11,20	3,03	-2,78
	ANN	E	0,932	6,80	1,77	-0,85
		V	0,861	12,25	3,88	-4,12
	PCA-ANN	E	0,663	9,75	5,82	5,82
		V	0,266	14,10	6,89	-2,56

^aE: Entrenamiento; V: Validación

La validación externa lleva asociada cierta pérdida de precisión en las estimaciones, lo que se refleja en un ligero descenso en los valores de *r* y un incremento en los valores de *RMSE*, *NMSE* y *FV* en comparación con los obtenidos para el subconjunto de datos de

entrenamiento/calibración. El coeficiente de correlación de la validación externa de los modelos PLSR y ANN se encuentra dentro del intervalo 0,5-0,8, similar al obtenido en el entrenamiento. Por tanto, en base a estos resultados, los modelos basados en PLSR y ANNs se pueden considerar apropiados para la estimación de los niveles en aire ambiente de Ni y Pb en el lugar de estudio. Sin embargo, como comentario general, el mejor par de estadísticos de entrenamiento y validación externa se han obtenido mediante un modelo de ANNs, lo que indica que esta técnica no solo ajusta los datos sino que también es capaz de proporcionar buenas estimaciones de aquellos datos que no han sido utilizados para el desarrollo de los modelos, lo que implica que las ANNs presentan una mejor capacidad de generalización que las otras técnicas consideradas para este caso de estudio.

En cuanto a los modelos PCA-ANN, estos proporcionan coeficientes de correlación muy bajos, entre 0,3 y 0,6. Esto, junto con unos índices de incertidumbre superiores a los de los modelos PLSR y ANN, indica que para esta aplicación específica, la aplicación de un análisis de factores como el PCA de forma previa al desarrollo de los modelos basados en ANNs no es una alternativa efectiva.

De forma complementaria, se evalúa gráficamente el desempeño de los modelos a través de la Figura 4.7., que muestra las concentraciones de Ni y Pb normalizadas por medio de su correspondiente UIE que han sido obtenidas empleando los modelos basados en PLSR y en ANNs, tanto para el subconjunto de entrenamiento/calibración como para el de validación externa. Como se adelantaba en el apartado 4.2.2. se comprueba que ambos modelos presentan dificultades para estimar con precisión los datos puntuales, con cierta tendencia a la subestimación de las concentraciones más elevadas, en especial en el caso del Ni, quizá porque estas se encuentran más alejadas del grueso de los datos. Esta limitación coincide con la de los resultados obtenidos en el apartado 4.1. para las campañas extensivas de muestreo llevadas a cabo en Castro Urdiales y Reinosa. No obstante, de nuevo, tanto los modelos basados en PLSR como los modelos de ANNs son capaces de capturar la tendencia subyacente de los datos de entrenamiento e incluso, en menor medida, la de los datos de validación externa. A este respecto, se observa que las ANNs presentan una mayor facilidad para el ajuste de los datos, como se observa en los picos de concentración del subconjunto de entrenamiento y validación externa del Ni (Figura 4.7.) y que redundan en unos valores de *NMSE* y *FV* inferiores en comparación con los obtenidos por los modelos PLSR.

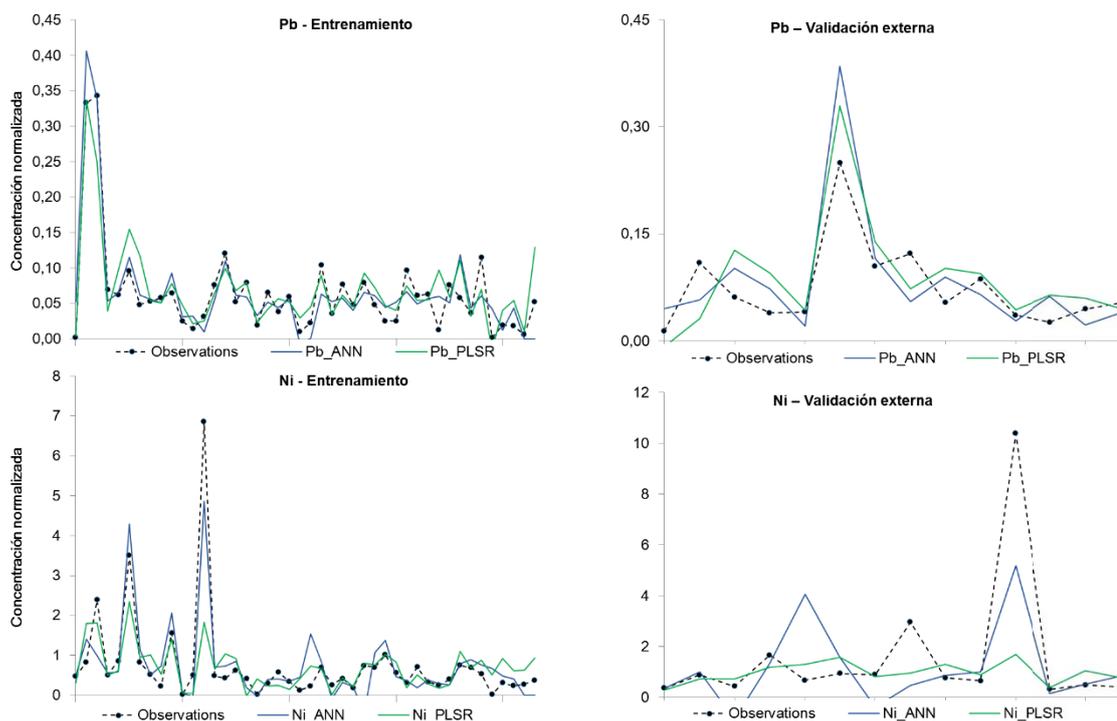


Figura 4.7. Comparación entre las concentraciones normalizadas observadas y modeladas de Ni y Pb en Dunkerque.

La evaluación del desempeño de los modelos se ha de llevar a cabo en el contexto del problema de estudio, prestando especial atención al limitado número de muestras disponibles para el desarrollo de los mismos. Las campañas de muestreo intensivas, por ser de una duración breve de tiempo dan lugar a un reducido número de muestras. Este hecho, unido a la variabilidad que presentan los niveles de contaminantes en el periodo de estudio supone una desventaja adicional de partida para el desarrollo de los modelos de estimación, cuyo desempeño, como se ha discutido en el apartado 3.6., es fuertemente dependiente del número de observaciones o muestras a emplear.

4.2.3.2. Estimación de metales no regulados: Cr, Mn y V

En la Tabla 4.6. se recogen los valores de los índices estadísticos para evaluar el desempeño de los modelos en cuanto a la estimación de los datos puntuales computados a partir de los resultados obtenidos por los mejores modelos desarrollados para la estimación de los niveles de Cr, Mn y V en Dunkerque.

Los coeficientes de correlación obtenidos a partir de los resultados de los modelos desarrollados para el Cr, inferiores a 0,3 en valor absoluto —en algunos casos incluso con signo negativo— junto a unos *NMSE* y *FV* hasta un orden de magnitud superiores a los de los obtenidos para el resto de contaminantes, indican una incapacidad de los modelos desarrollados tanto para ajustar como para predecir los niveles puntuales de Cr. Esto pone de manifiesto que el hecho de que los modelos representen una herramienta válida para la

evaluación de la calidad del aire desde el punto de vista normativo no es suficiente para garantizar un correcto desempeño de los modelos. En cuanto al modelo PCA-ANN desarrollado para el Cr, a pesar de que proporcione un r superior y unos $NMSE$ y FV inferiores a los obtenidos con los modelos PLSR y ANN, el hecho de que proporcione unos índices de incertidumbre superiores a 100% y una sobreestimación del valor medio hasta 6 veces superior a la concentración media observada (Tabla 4.4.), invalidarían directamente este modelo como método de evaluación de la calidad del aire con respecto al Cr.

Los coeficientes de correlación de los modelos desarrollados para el Mn y el V, con un valor máximo de 0,66 en la validación externa, son ligeramente inferiores a los de los modelos desarrollados para el Ni y el Pb, pero aceptables. Los valores del índice $NMSE$ obtenidos mediante los modelos desarrollados para la estimación del Mn, por encima de 1, son un orden de magnitud superiores a los obtenidos en la estimación del V y de los metales regulados. No obstante, estos indicadores pueden estar influenciados por una pequeña proporción de valores extremos que no reflejan el comportamiento del grueso de los datos, en especial teniendo en cuenta el amplio rango de variación de las concentraciones observadas de Mn.

De nuevo, las ANNs proporcionan unos r y $NMSE$ iguales o superiores a los de los modelos PLSR y PCA-ANN, lo que la sitúa como la alternativa más apropiada para este caso de estudio.

Tabla 4.6. Índices estadísticos para la evaluación del desempeño de los modelos de estimación de Cr, Mn y V desarrollados en Dunkerque.

Metal	Modelo	Subconjunto ^a	Desempeño			
			r	RMSE (ng/m ³)	NMSE 10	FV 10
Cr	PLSR	E	-0,031	29,14	128,78	11,34
		V	0,077	34,50	81,18	12,78
	ANN	E	-0,040	36,70	204,16	6,53
		V	-0,240	36,35	146,74	14,76
	PCA-ANN	E	0,331	34,96	59,45	9,70
		V	0,275	19,15	74,83	-13,57
Mn	PLSR	E	0,580	62,24	15,82	3,57
		V	0,184	139,20	22,76	6,99
	ANN	E	0,704	58,80	19,75	3,42
		V	0,457	122,34	16,06	4,64
	PCA-ANN	E	0,463	97,16	19,78	7,34
		V	0,431	64,08	20,40	1,74
V	PLSR	E	0,694	4,00	3,74	2,33
		V	0,590	5,49	3,76	3,16
	ANN	E	0,806	3,52	2,59	1,78
		V	0,663	5,11	3,26	2,11
	PCA-ANN	E	0,747	2,90	2,02	2,93
		V	0,366	6,85	9,17	0,56

^aE: Entrenamiento; V: Validación

4.3. Influencia de variables meteorológicas relacionadas con la estabilidad atmosférica en el desempeño de modelos PLSR y ANN para la estimación de los niveles de As, Cd, Ni y Pb en un área urbana: Santander

Por último, en este apartado se evalúa la aplicación en su totalidad de la metodología propuesta en el apartado 3.10 para la estimación de los niveles de As, Cd, Ni y Pb en aire ambiente en base a campañas de muestreo extensivas, como método de evaluación de la calidad del aire en zonas urbanas con una procedencia de la contaminación atmosférica más compleja. En concreto, se ha seleccionado la ciudad de Santander, que siendo la capital, constituye la mayor área urbana de Cantabria y presenta la mayor complejidad desde el punto de vista del número y variedad de fuentes de los metales estudiados. Por ello, las relaciones entre las fuentes de contaminantes minoritarios y sus niveles en aire ambiente se asumen a priori como más complejas que las de las zonas urbanas del apartado 4.1.

4.3.1. Niveles en inmisión de As, Cd, Ni y Pb

Se han llevado a cabo campañas extensivas de muestreo de la fracción PM₁₀ del material particulado en los años 2008, 2009 y 2011 en la zona urbana de Santander (Cantabria), caracterizando su composición. Los niveles en aire ambiente obtenidos de As, Cd, Ni y Pb se muestran en la Figura 4.8. En la misma, se observa que los valores medios anuales de estos contaminantes obtenidos en esta zona durante el periodo considerado se encuentran por debajo del umbral inferior de evaluación (UIE), y además no presentan variaciones significativas a lo largo del periodo de estudio. Por tanto, de acuerdo a la Directiva 2008/50/CE, se permite el uso de técnicas de modelización o de estimación objetiva para llevar a cabo la evaluación de la calidad del aire con respecto a dichos contaminantes en el lugar de muestreo como alternativa al procedimiento habitual basado en la toma de muestras experimentales para su medición posterior en laboratorio.

4.3.2. Evaluación de los modelos estadísticos desarrollados para la estimación de los niveles en inmisión de As, Cd, Ni y Pb como técnicas de estimación objetiva para su uso en la evaluación de la calidad del aire

La calidad del aire en relación al As, Cd Ni y Pb ha sido evaluada en Santander mediante modelos estadísticos basados únicamente en PLSR y ANNs siguiendo la metodología descrita en el apartado 3.10. Los resultados de los mejores modelos desarrollados considerando los enfoques A y B (véase el apartado 3.6.) se muestran en las Tablas 4.7. y 4.8. Al igual que en los demás casos de estudio de esta Tesis Doctoral, además de los resultados de validación externa,

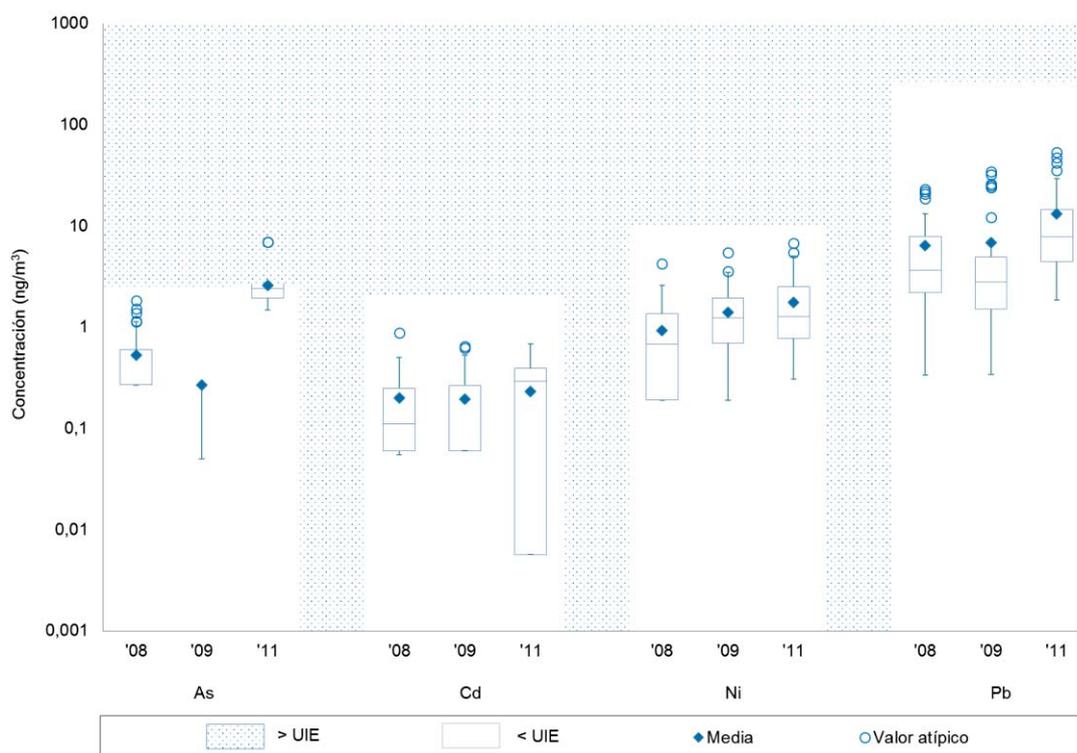


Figura 4.8. Diagrama de cajas de los niveles en aire ambiente de As, Cd, Ni y Pb en los años 2008 ('08), 2009 ('09) y 2011 ('11) en el punto de muestreo de Santander. La caja representa la distancia entre los cuartiles superior (Q3) e inferior (Q1). Por su parte, la línea interna representa el valor de la mediana y las barras indican los valores mínimo y máximo. Como es habitual en esta representación, los valores atípicos se expresan como aquellos datos que se sitúan por encima de $Q3+1,5RIC$ o por debajo de $Q1-1,5RIC$, siendo RIC el rango intercuartílico e igual a la diferencia entre los cuartiles superior e inferior ($Q3-Q1$). Asimismo, la zona no punteada del fondo indica la región de concentraciones por debajo del UIE. UIE = $2,4 \text{ ng/m}^3$ (As), 2 ng/m^3 (Cd), 10 ng/m^3 (Ni) y 250 ng/m^3 (Pb). Los valores medios correspondientes a los años 2008 y 2009 se han obtenido de Arruti y col. (2011a).

se incluyen igualmente los resultados relativos al entrenamiento/calibración de los modelos para su comparación y análisis de la capacidad de generalización de los mismos.

Matemáticamente, cada indicador de desempeño de los modelos considerado en este trabajo tiene un valor ideal de cero, a excepción del coeficiente de correlación y de la eficiencia del modelo, cuyo valor ideal es uno. En este sentido, excepto los modelos desarrollados para el Pb, que exhiben un desempeño similar independientemente de la técnica empleada, la validación externa es en general mejor en el caso de considerar un enfoque B en lugar de un enfoque A. Esto puede ser debido al hecho de que los modelos desarrollados bajo el enfoque A han sido entrenados con datos procedentes de los años 2008 y 2009 y posteriormente han sido validados con datos procedentes del año 2011. Por tanto, hay un lapso de un año entre

Tabla 4.7. Índices de incertidumbre y error en la estimación del valor medio de los niveles de As, Cd, Ni y Pb mediante modelos estadísticos desarrollados en Santander basados en PLSR y ANNs bajo los enfoques A y B.

Contaminante	Modelo	Subconjunto ^a	Enfoque A				Enfoque B				
			Incertidumbre de la UE		Concentración media ^b		Incertidumbre de la UE		Concentración media ^b		
			RME (%)	RDE (%)	\overline{C}_O (ng/m ³)	\overline{C}_E (ng/m ³)	FB	RME (%)	RDE (%)	\overline{C}_O (ng/m ³)	\overline{C}_E (ng/m ³)
As	PLSR	E	2,5	0,1	1,1	1,1	0,000	13,0	2,83	0,7	0,000
		V	61,4	71,1	2,7	1,3	0,733	47,8	16,01	1,0	0,268
	ANN	E	35,6	14,2	1,1	1,0	0,076	62,3	20,55	0,7	0,164
		V	79,8	96,2	2,7	0,1	0,971	59,2	14,82	1,0	0,461
Cd	PLSR	E	70,0	24,0	0,3	0,3	0,000	65,9	22,09	0,3	0,000
		V	60,7	5,0	0,2	0,4	-0,538	43,3	8,06	0,2	-0,492
	ANN	E	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	56,8	29,03	0,3	0,2
		V	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	24,5	10,85	0,2	-0,182
Ni	PLSR	E	46,9	15,7	1,1	1,1	0,000	45,2	13,75	1,1	0,000
		V	54,4	28,2	1,8	1,4	0,226	32,5	9,19	1,5	0,008
	ANN	E	58,7	12,4	0,9	1,3	-0,295	47,5	12,74	1,1	-0,010
		V	74,4	28,2	1,8	1,2	0,341	51,8	14,11	1,5	0,316
Pb	PLSR	E	50,9	5,0	6,7	6,7	0,000	62,0	7,58	7,6	0,000
		V	69,5	8,6	13,2	9,1	0,370	60,7	8,15	10,8	0,090
	ANN	E	55,3	3,6	6,7	6,7	-0,005	80,5	8,60	7,9	0,185
		V	74,6	8,2	13,2	8,7	0,407	78,1	7,42	10,8	6,0

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

^cn.c.: No calculado

Tabla 4.8. Otros índices de desempeño de los modelos desarrollados en Santander basados en PLSR y ANNs para la estimación de los niveles de As, Cd, Ni y Pb bajo los enfoques A y B.

Contaminante	Modelo	Subconjunto ^a	Desempeño Enfoque A			Desempeño Enfoque B				
			r	RMSE (ng/m ³)	NMSE	FV	r	RMSE (ng/m ³)	NMSE	FV
As	PLSR	E	1,000	0,02	0,00	0,00	0,971	0,1	0,04	0,12
		V	0,367	1,96	1,10	0,49	0,714	0,6	0,42	0,43
	ANN	E	0,801	0,50	0,23	1,07	0,878	0,5	0,56	1,33
		V	0,258	2,22	1,90	1,55	0,996	0,7	0,82	1,42
Cd	PLSR	E	0,354	0,29	0,93	0,81	0,199	0,3	1,22	0,91
		V	-0,071	0,30	1,14	0,52	0,160	0,2	0,97	1,38
	ANN	E	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	0,467	0,3	1,77	0,77
		V	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	-0,186	0,3	2,04	0,20
Ni	PLSR	E	0,553	0,82	0,51	0,51	0,488	0,8	0,58	0,36
		V	0,280	1,51	0,92	0,71	0,530	0,9	0,38	0,46
	ANN	E	0,584	0,80	0,55	1,28	0,564	0,7	0,44	0,68
		V	0,294	1,54	1,08	1,57	0,517	1,0	0,63	0,79
Pb	PLSR	E	0,424	6,77	1,02	0,73	0,366	9,4	1,55	0,71
		V	0,304	13,32	1,48	1,16	0,257	10,4	1,01	0,77
	ANN	E	0,636	5,77	0,74	0,46	0,523	9,2	1,64	1,03
		V	0,404	12,98	1,46	1,00	0,338	11,1	1,92	1,03

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

^cn.c.: No calculado

los datos empleados para el ajuste de los parámetros de los modelos y los datos usados para validarlos y, en consecuencia, puede que haya habido un cambio en los patrones de emisión especialmente considerando el periodo en el que se centra el estudio, que se caracteriza por un gran impacto de la crisis económica sobre las actividades industriales, lo que ha podido alterar su producción con importantes repercusiones sobre las emisiones, en particular a escala local. Por esta razón, se ha de asumir que considerando un enfoque A, los modelos no son calibrados correctamente para poder predecir la situación correspondiente al periodo utilizado para validarlos externamente, con un consecuente impacto negativo en la capacidad de generalización de los mismos. Por tanto, la razón por la cual el enfoque B proporciona mejores resultados radica en que el hecho de que considerar los tres años de los que se disponen datos de manera simultánea favorece hasta cierto punto una atenuación de las diferencias entre las relaciones de los datos de 2011 y las de los años precedentes. Esta constituye la principal limitación para el uso de estos modelos como herramienta rutinaria de evaluación de la calidad del aire por parte de las autoridades locales: esta aplicación se restringiría a aquellos casos en los que los patrones de emisión se mantengan relativamente estables en el tiempo, lo que es una suposición aceptable a corto plazo, pero que rara vez resulta válida a largo plazo. Por ello, los modelos deberían ser actualizados periódicamente con al menos un pequeño número de nuevas muestras, o muestras de control, para prevenir su desactualización.

Como se ha mencionado con anterioridad, en lo que respecta a la evaluación de las técnicas estudiadas como herramientas para la evaluación de la calidad del aire en el contexto de las Directivas europeas, estrictamente, basta con que las estimaciones obtenidas por los modelos cumplan con los requerimientos de incertidumbre y de calidad en la estimación del valor medio puesto que los valores objetivo/límite vienen dados en concentraciones medias a lo largo de un año natural. Como se muestra en la Tabla 4.7, todos los mejores modelos desarrollados considerando el enfoque B cumplen con estos requerimientos, puesto que ambos índices de incertidumbre, *RME* y *RDE*, se encuentran por debajo del 100% y además el índice *FB* se encuentra dentro del rango entre $-2/3$ y $2/3$, extremos que corresponden a la mitad y al doble de la concentración media observada, respectivamente. De hecho, los valores de *RDE* no superan en ningún caso el 30%, estando la mitad por debajo del 10%. Por otra parte, se observa que los valores del índice *FB* de los modelos PLSR son en general menores que los obtenidos por los modelos basados en ANNs tanto para el ajuste como para la validación externa.

Los valores de los índices *NMSE* y *FV* tienden a ser mayores de 0,5. Además, los valores de *FV* son todos positivos, lo que indica que las estimaciones presentan una menor varianza que las observaciones. Las estimaciones parecen variar dentro de un rango similar al de la variable dependiente y próximo al valor medio, por lo que hay una sobreestimación de las concentraciones observadas más bajas y, aún más, una subestimación de las más elevadas. Además, los coeficientes de correlación, excepto los del A5, son iguales o inferiores a 0,5 en la

mayoría de los casos. No obstante, no hay una clara evidencia de que una técnica tenga un desempeño superior al de la otra: tanto la PLSR como las ANNs muestran un desempeño equivalente, lo que se puede interpretar como que la relación entre los valores de entrada a los modelos y las concentraciones en inmisión de As, Cd, Ni y Pb tienden a ser lineales más que no lineales, al menos en lo que respecta a los datos de los que se dispone. Esta idea está reforzada por la sencillez de la mayoría de las arquitecturas de las redes neuronales que han aportado los mejores indicadores de desempeño, que cuentan con una única capa oculta con un bajo número de neuronas artificiales.

Un análisis adicional del desempeño de los modelos desde el punto de vista de su aplicación para la evaluación de la calidad del aire se puede llevar a cabo a partir de la Figura 4.9., que representa los resultados del entrenamiento y validación externa de los modelos PLSR y ANN desarrollados para As, Cd, Ni y Pb obtenidos bajo el enfoque B y siguiendo el método gráfico descrito en el apartado 3.9.2. Aunque el *CRMSE* es siempre positivo por su propia definición matemática (ecuación 3.11), en la representación gráfica se le asigna un valor negativo para distinguir aquellas situaciones en las que la desviación estándar de las estimaciones es inferior a la de las observaciones (Pederzoli y col., 2011). Los resultados de todos los modelos se encuentran en la parte izquierda del diagrama, que corresponde a valores negativos de *CRMSE*, lo que indica que las estimaciones varían en un rango más estrecho que las observaciones. Esto se corresponde asimismo con valores positivos del índice *FV*, como ya se ha mencionado (Tabla 4.8.). Además, como se muestra en la Figura 4.9., los resultados de la validación externa de los modelos desarrollados para el Pb y el Cd bajo el enfoque B tienen una eficiencia del modelo negativa y, en consecuencia, se encuentran fuera de la circunferencia. Según Jollif y col. (2009), el hecho de que los valores estimados se encuentren dentro de la circunferencia, es decir, que la eficiencia del modelo sea positiva, indica que el modelo es mejor predictor de las observaciones que un valor constante igual a la media de las observaciones. Esto, sumado al aumento en los errores de estimación *NMSE* y *FV* y al descenso en los coeficientes de correlación que se observan comparando los valores de los índices estadísticos obtenidos en Santander (Tablas 4.7. y 4.8.) con los obtenidos aplicando la misma metodología en otras zonas urbanas más pequeñas donde también se han llevado a cabo campañas extensivas de muestreo, como Castro Urdiales y Reinosa (Tablas 4.1. y 4.2.) en el apartado 4.1., evidencia que el desempeño de los modelos desarrollados en Santander presenta un margen de mejora. En este sentido, se plantea el uso de parámetros meteorológicos relacionados con la estabilidad atmosférica como variables meteorológicas adicionales de entrada a los modelos.

No obstante, este descenso generalizado en la adecuación del desempeño de los modelos desarrollados en Santander en comparación con el de los modelos desarrollados en Castro Urdiales y Reinosa concuerda con lo obtenido por Arruti y col. (2011) mediante modelos de regresión lineal múltiple en dichas ubicaciones en 2008. En dicho trabajo se obtuvieron unos

valores de *RME* de 33% (As), 72% (Cd), 59% (Ni) y 48% (Pb), superiores a los obtenidos en Castro Urdiales y Reinosa, y unos coeficientes de correlación de 0,8 (As), 0,4 (Cd), 0,5 (Ni) y 0,6 (Pb), inferiores a los obtenidos en Castro Urdiales y Reinosa. En la presente Tesis Doctoral, se ha conseguido mantener un desempeño equivalente de los modelos estadísticos para la estimación de los niveles de As, Cd, Ni y Pb en Santander con el añadido de haber realizado una validación externa de los mismo y a pesar de la consideración en el análisis de datos procedentes de 2009 y 2011, lo que supone una complicación adicional por los posibles cambios en los patrones de emisión, como se ha mencionado anteriormente.

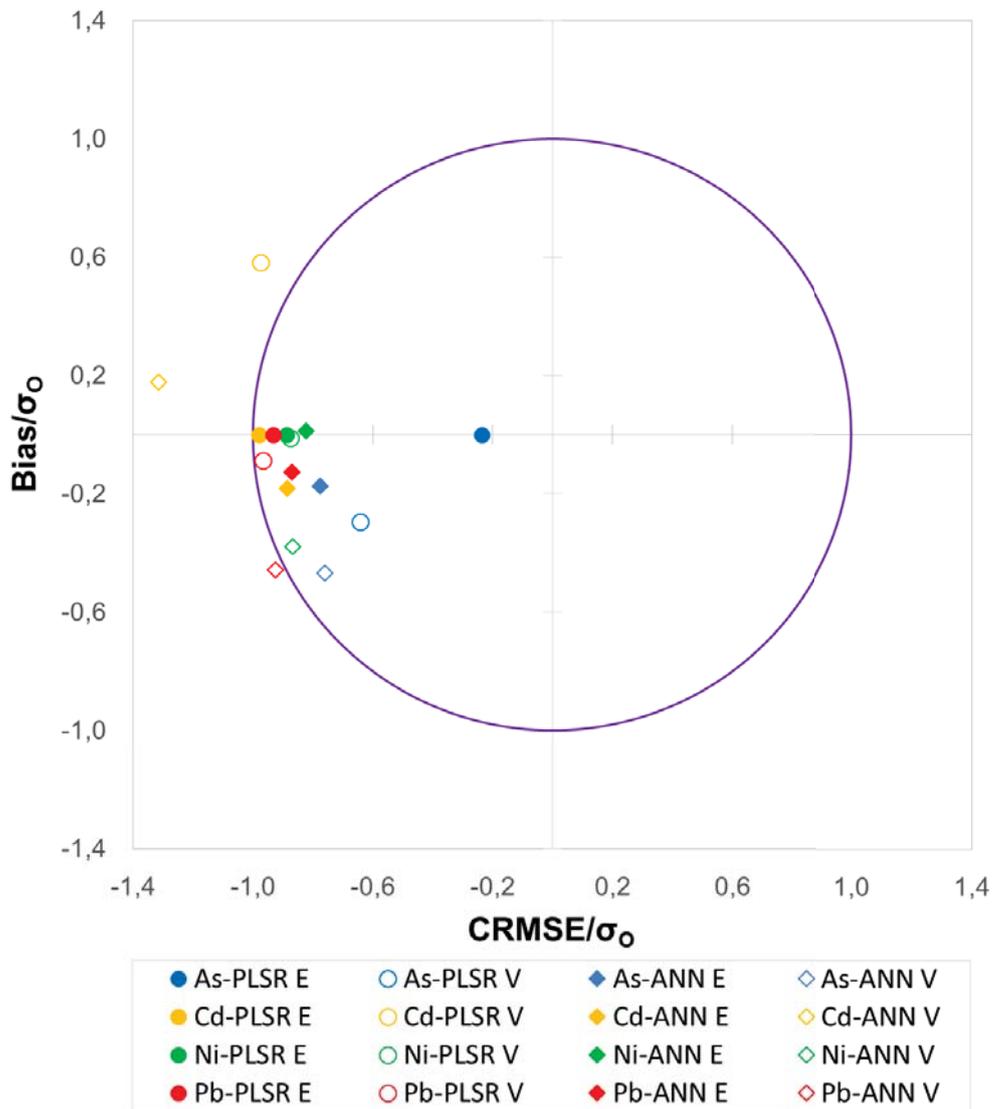


Figura 4.9. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el As (azul), Cd (amarillo), Ni (verde) y Pb (rojo). Los puntos de datos relativos a la PLSR se representan por medio de un círculo y los relativos a las ANNs por medio de un rombo. Los marcadores rellenos y huecos se utilizan para representar los resultados referentes al entrenamiento (E) y a la validación externa (V), respectivamente.

4.3.3. Influencia de la consideración de variables meteorológicas relacionadas con la estabilidad atmosférica

4.3.3.1. Clasificación de las variables independientes de acuerdo a su coeficiente de correlación

La importancia relativa de cada variable independiente para su uso como variable de entrada a los modelos se ha determinado mediante el coeficiente de correlación de Pearson que se establece entre estas variables y cada una de las variables de salida, de forma que se puedan identificar las entradas más significativas. Este coeficiente evalúa el grado de variación conjunta o similitud entre dos variables, cuantificando la relación lineal existente entre ambas.

Las Tablas 4.9. y 4.10. muestran una clasificación de las variables independientes –incluyendo las relacionadas con la estabilidad atmosférica– en orden descendiente según su coeficiente de correlación con respecto a cada variable dependiente, es decir, la concentración en inmisión de metal o semimetal. Asimismo, se incluye también el correspondiente *p-value*. Como se han considerado dos enfoques diferentes, enfoques A y B, la Tabla 4.9. hace referencia a los datos del periodo 2008-2009, mientras que la Tabla 4.10. presenta los resultados del conjunto de datos al completo, incluyendo los relativos al año 2011.

En primer lugar, se observa que no existe una fuerte correlación entre las variables independientes y las dependientes de manera individual. No obstante, la consideración del conjunto total de variables independientes de forma simultánea puede alterar la influencia individual debido a efectos sinérgicos.

Las concentraciones en inmisión de Pb y el Ni parecen estar principalmente relacionadas con las concentraciones de contaminantes atmosféricos mayoritarios, en especial con las concentraciones de PM₁₀, puesto que éstos ocupan las primeras posiciones en la clasificación. Sin embargo, esta tendencia no se observa en el caso del As y del Cd.

La consideración de muestras procedentes del año 2011 conduce a una modificación significativa de los coeficientes de correlación y afecta a la posición relativa de cada variable en la clasificación. Como se ha mencionado en el apartado 4.3.2., esto puede ser debido a un cambio en los patrones de emisión en la región, probablemente asociado a las fluctuaciones en la producción de las actividades industriales a consecuencia de la crisis económica y financiera. No en vano, el periodo de estudio comprende un intervalo de tiempo que se extiende desde el año 2008 hasta 2011, que coincide con un momento en el tiempo con importantes efectos adversos sobre la economía. No obstante, en cualquier caso, algunas de las variables meteorológicas relativas a la estabilidad atmosférica se encuentran en las posiciones superiores de las clasificaciones establecidas para cada contaminante. Aún más, para cada contaminante en el periodo 2008-2011, al menos una de las cuatro nuevas variables de entrada consideradas se encuentra entre aquellas estadísticamente significativas con un nivel

de confianza estadística del 90%. En términos generales, en base a las posiciones que ocupan estas nuevas variables en la clasificación, la clase de estabilidad de Pasquill-Gifford parece ser la variable más significativa.

Tabla 4.9. Clasificación de las variables independientes consideradas en los modelos desarrollados para Santander de acuerdo al coeficiente de correlación para el periodo 2008-2009 (Santos y col., 2016).

Clasificación	As		Cd		Ni		Pb	
	Variable	<i>p</i> -value						
1	Ri_{am}	0,002	MH_{pm}	0,014	Ln PM ₁₀	0,001	NO _x	0,003
2	SD	0,003	PP	0,194	I_{pm}	0,002	Ln PM ₁₀	0,010
3	T	0,053	SE	0,251	RH	0,012	O ₃	0,027
4	Ln PM ₁₀	0,081	I_{am}	0,314	PG_{am}	0,022	I_{pm}	0,031
5	RS	0,107	O ₃	0,366	I_M	0,025	MH_{am}	0,291
6	WE	0,134	I_{pm}	0,392	SO ₂	0,093	P	0,307
7	SE	0,174	NO _x	0,422	P	0,102	WE	0,361
8	PG_{pm}	0,190	P	0,481	MH_{am}	0,109	RS	0,365
9	PP	0,279	T	0,533	WS	0,141	T	0,419
10	I_v	0,310	WS	0,539	PG_{pm}	0,161	MH_{pm}	0,467
11	O ₃	0,432	PG_{pm}	0,549	NO _x	0,181	SO ₂	0,553
12	Ri_{pm}	0,437	WE	0,553	T	0,196	SD	0,555
13	MH_{pm}	0,446	MH_{am}	0,577	MH_{pm}	0,216	PP	0,612
14	NO _x	0,446	RS	0,654	SD	0,235	Ri_{pm}	0,642
15	P	0,526	Ln PM ₁₀	0,675	RS	0,236	Ri_{am}	0,705
16	MH_{am}	0,574	PG_{am}	0,681	WE	0,391	WD	0,710
17	RH	0,675	SO ₂	0,726	Ri_{pm}	0,793	SE	0,730
18	WD	0,679	RH	0,822	SE	0,798	PG_{pm}	0,738
19	PG_{am}	0,722	Ri_{pm}	0,832	O ₃	0,807	I_{am}	0,861
20	SO ₂	0,790	SD	0,857	Ri_{am}	0,883	WS	0,861
21	I_{am}	0,809	WD	0,871	WD	0,914	RH	0,886
22	WS	0,886	Ri_{am}	0,955	PP	0,988	PG_{am}	0,999

Las variables meteorológicas relativas a la estabilidad atmosférica y los *p*-values de aquellas variables estadísticamente significativas con un intervalo de confianza del 90% se presentan en negrita.

4.3.3.2. Estimación de los niveles en inmisión de As, Cd, Ni y Pb considerando nuevas variables meteorológicas

En este estudio se ha propuesto considerar variables meteorológicas relacionadas con la estabilidad atmosférica como variables adicionales de entrada a los modelos para mejorar su desempeño. En consecuencia, se han desarrollado modelos basados en PLSR y ANNs considerando como variables meteorológicas complementarias aquellas que, en base a los coeficientes de correlación individuales y de acuerdo al enfoque B (Tabla 4.10.), son más relevantes para cada contaminante. La Tabla 4.11. muestra los resultados para el Pb y el Ni. Se observa una ligera reducción en los valores de *RME* y *RDE* junto con un leve incremento de los valores de los coeficientes de correlación cuando se incluye la clase de estabilidad atmosférica

de Pasquill-Gifford matutina (PG_{am}) para el desarrollo de los modelos. Asimismo, se produce también una mejora de los valores del índice FB de la validación externa de los modelos, excepto para el caso del modelo PLSR desarrollado para el Ni. Los resultados obtenidos para el As y el Cd no se presentan puesto que no se obtuvo ninguna mejora significativa en el desempeño de los modelos.

Tabla 4.10. Clasificación de las variables independientes consideradas en los modelos desarrollados para Santander de acuerdo al coeficiente de correlación para el periodo 2008-2009 y 2011.

Clasificación	As		Cd		Ni		Pb	
	Variable	<i>p-value</i>	Variable	<i>p-value</i>	Variable	<i>p-value</i>	Variable	<i>p-value</i>
1	WS	0,000	PG_{pm}	0,005	WS	0,000	O ₃	0,001
2	WD	0,000	MH_{pm}	0,005	Ln PM ₁₀	0,001	NO _x	0,004
3	SO ₂	0,000	PG_{am}	0,016	T	0,015	WD	0,014
4	T	0,000	SE	0,021	WD	0,018	PG_{am}	0,049
5	NO _x	0,000	Ri_{pm}	0,121	RH	0,022	RH	0,052
6	SE	0,000	RH	0,123	PG_{am}	0,031	SE	0,064
7	I_{am}	0,001	I_{am}	0,125	MH_{am}	0,041	MH_{am}	0,085
8	O ₃	0,046	WD	0,163	I_{pm}	0,043	PG_{pm}	0,088
9	I_{pm}	0,076	P	0,182	PG_{pm}	0,097	I_{am}	0,097
10	Ln PM ₁₀	0,100	PP	0,212	SE	0,105	Ln PM ₁₀	0,130
11	Ri_{pm}	0,113	O ₃	0,226	RS	0,159	I_{pm}	0,131
12	PG_{pm}	0,154	MH_{am}	0,247	P	0,189	P	0,138
13	PG_{am}	0,170	SD	0,261	Ri_{pm}	0,269	SO ₂	0,181
14	MH_{am}	0,178	RS	0,293	O ₃	0,293	T	0,208
15	WE	0,315	WE	0,385	WE	0,323	MH_{pm}	0,239
16	RH	0,360	Ri_{am}	0,386	SO ₂	0,361	Ri_{pm}	0,242
17	PP	0,371	T	0,399	Ri_{am}	0,364	Ri_{am}	0,283
18	MH_{pm}	0,408	WS	0,447	I_{am}	0,387	PP	0,341
19	RS	0,409	I_{pm}	0,477	NO _x	0,432	SD	0,344
20	SD	0,441	SO ₂	0,480	PP	0,436	WS	0,352
21	Ri_{am}	0,459	Ln PM ₁₀	0,486	SD	0,445	WE	0,425
22	P	0,480	NO _x	0,494	MH_{pm}	0,486	RS	0,483

Las variables meteorológicas relativas a la estabilidad atmosférica y los *p-values* de aquellas variables estadísticamente significativas con un intervalo de confianza del 90% se presentan en negrita.

Las Figuras 4.10. y 4.11. muestran una comparación entre los diagramas objetivo (*target diagrams*) en los que se representan los resultados obtenidos con los modelos iniciales de Ni y Pb presentados en el apartado 4.3.2 y los obtenidos cuando se considera la PG_{am} como variable de entrada. En general, hay una ligera mejora en los resultados de los modelos. En particular, se observa que la consideración de la PG_{am} como variable meteorológica adicional tiene una mayor influencia en los resultados de los modelos basados en ANNs. Produce una disminución en el error de la estimación del valor medio y en los valores de $CRMSE$ de la validación externa del modelo ANN desarrollado para el Pb, que conduce a un cambio en el signo de la MEF , que se convierte en un valor positivo posicionándose en el interior de la circunferencia. En lo que

Tabla 4.11. Comparativa de los resultados de los modelos PLSR y ANN desarrollados en Santander considerando PG_{am} como variable de entrada adicional bajo el enfoque B.

Contaminante	Modelo	Subconjunto ^a	Incertidumbre de la UE		Concentración media			Desempeño			
			RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	FB	r	RMSE (ng/m ³)	NMSE	FV
Ni	PLSR	E	45,2	13,75	1,1	1,1	0,000	0,488	0,8	0,58	0,36
		V	32,5	9,19	1,5	1,5	0,008	0,530	0,9	0,38	0,46
	PLSR - PG_{am}	E	27,5	12,31	1,2	1,2	0,000	0,593	0,7	0,42	0,43
		V	41,6	21,36	1,5	1,4	0,092	0,410	1,2	0,68	0,52
	ANN	E	47,5	12,74	1,1	1,1	-0,010	0,564	0,7	0,44	0,68
		V	51,8	14,11	1,5	1,1	0,316	0,517	1,0	0,63	0,79
	ANN - PG_{am}	E	5,6	1,05	1,2	1,2	-0,025	0,958	0,2	0,05	0,06
		V	29,7	14,98	1,5	1,4	0,103	0,479	1,2	0,70	0,23
Pb	PLSR	E	62,0	7,58	7,6	7,6	0,000	0,366	9,4	1,55	0,71
		V	60,7	8,15	10,8	9,9	0,090	0,257	10,4	1,01	0,77
	PLSR - PG_{am}	E	57,3	6,70	7,6	7,6	0,000	0,574	8,0	1,11	0,56
		V	54,8	5,33	9,8	9,6	0,012	0,463	8,2	0,72	0,69
	ANN	E	80,5	8,60	7,9	6,5	0,185	0,523	9,2	1,64	1,03
		V	78,1	7,42	10,8	6,0	0,579	0,338	11,1	1,92	1,03
	ANN - PG_{am}	E	57,3	6,49	7,4	9,5	-0,247	0,642	7,6	0,82	0,71
		V	45,9	5,97	9,8	9,4	0,041	0,515	8,0	0,71	0,84

^aE: Entrenamiento; V: Validación externa

^bO: Observada; E: Estimada

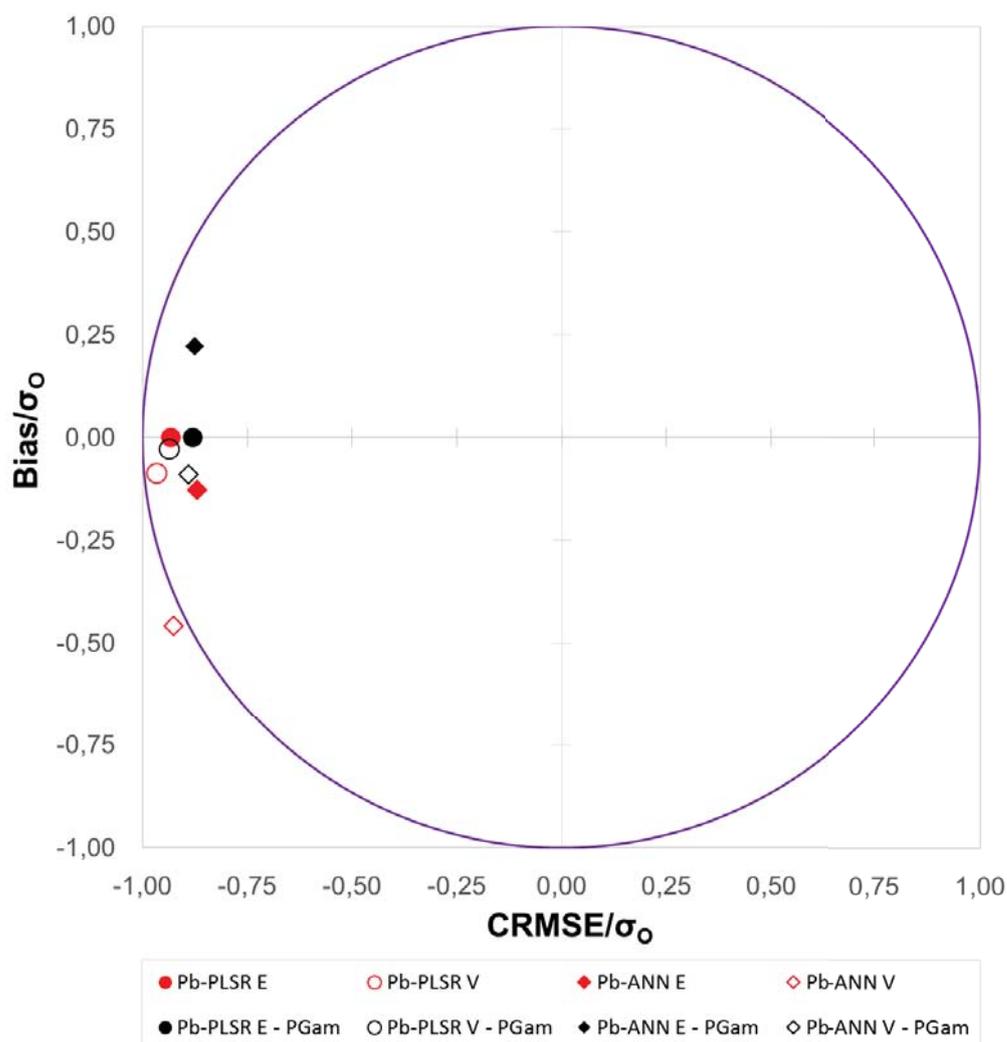


Figura 4.10. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el Pb para la determinación de la influencia de la PG_{am} en el desempeño de los mismos.

respecta a los resultados del modelo PLSR para el Pb, experimentan una mejora menor, moviéndose ligeramente hacia el centro.

Con respecto al Ni, los resultados del modelo ANN exhiben una mejora mayor que los del modelo PLSR, especialmente en términos de estimación de la concentración media en la validación externa del modelo, con un valor de *bias* considerablemente menor. En cuanto a los resultados del modelo PLSR, aunque permanecen en el interior de la circunferencia y, por tanto, aún se pueden considerar como válidos, experimentan un ligero empeoramiento a juzgar por los resultados de la validación externa.

Debido al considerable esfuerzo requerido por parte del usuario, puesto que conlleva la disponibilidad de datos de sondeos termodinámicos y una procesado de datos avanzado, esta propuesta solo debería ser tenida en cuenta en aquellas situaciones en la que haya un amplio

margen de mejora, es decir, en las que los valores de los parámetros estadísticos utilizados para evaluar el desempeño de los modelos no cumplan con los objetivos de calidad.

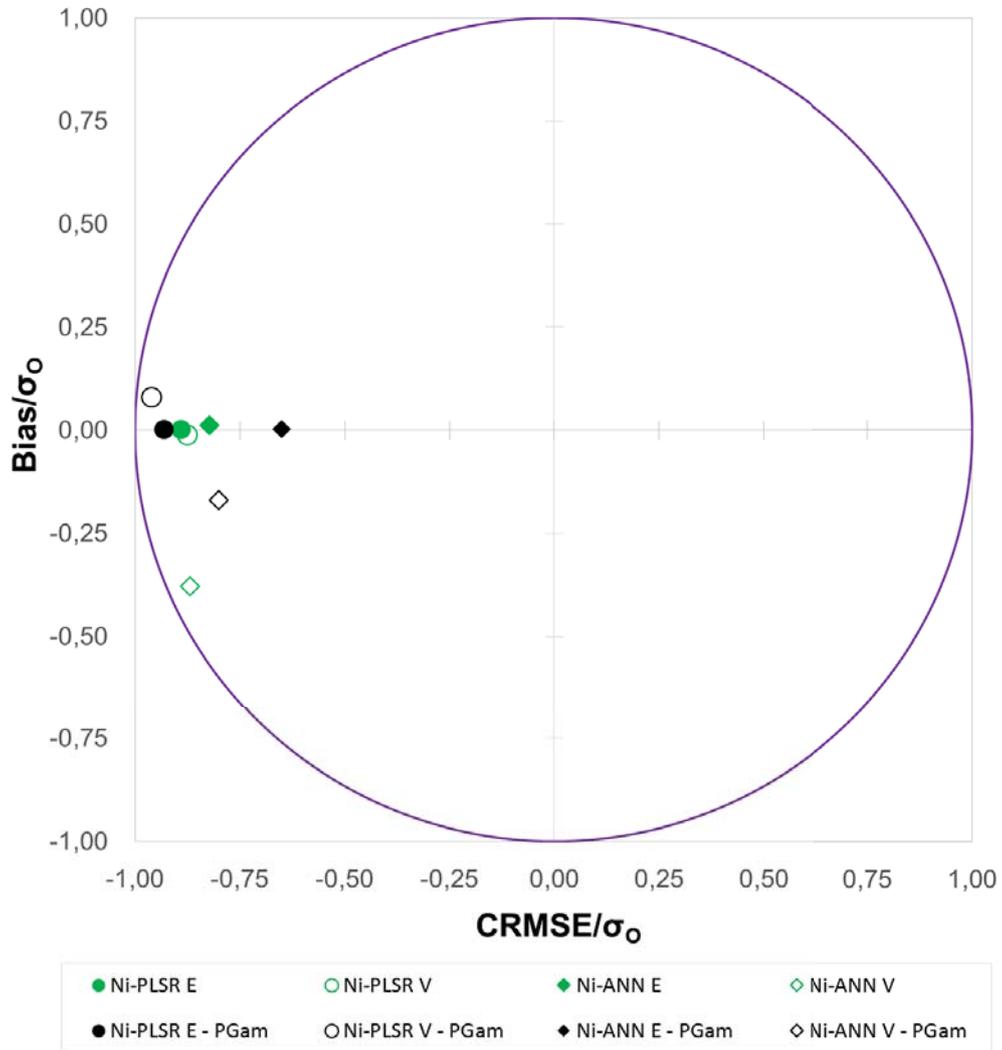
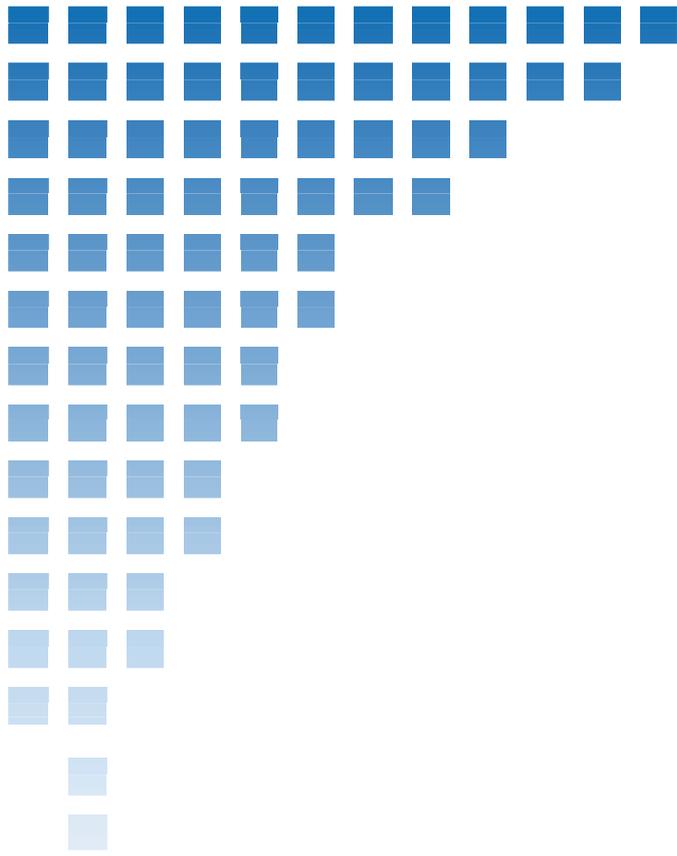


Figura 4.11. Diagrama objetivo de los resultados de los modelos PLSR y ANN desarrollados en Santander para el Ni para la determinación de la influencia de la PG_{am} en el desempeño de los mismos.

4.4. Bibliografía

- Arruti, A., Fernández-Olmo, I., Irabien, A., 2011a. Regional evaluation of particulate matter composition in an Atlantic coastal area (Cantabria region, northern Spain): spatial variations in different urban and rural environments. *Atmospheric Research* 101, 208-293.
- Arruti, A., Fernández-Olmo, I., Irabien, A., 2011b. Assessment of regional metal levels in ambient air by statistical regression models. *Journal of Environmental Monitoring* 13(7), 1991-2000.
- Aschner, M., Erikson, K.M., Dorman, D.C., 2005. Manganese dosimetry: species differences and implications for neurotoxicity. *Critical Reviews in Toxicology* 35(1), 1-32.
- Chelani, A.B., Gajghate, D.G., Hasan, M.Z., 2002. Prediction of ambient PM₁₀ and toxic metals using artificial neural networks. *Journal of the Air and Waste Management Association* 52(7), 805-810.
- EC, 2008. Council Directive 2008/50/EC Directive of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. The European Parliament and the Council of the European Union. Off J L152 1-44.
- Finkelstein, M.M., Jerret, M., 2007. A study of the relationships between Parkinson's disease and markers of traffic-derived and environmental manganese air pollution in two Canadian cities. *Environmental Research* 104(3), 420-432.
- Flemming, J., Stern, R., 2007. Testing model accuracy measures according to the EU directives-examples using the chemical transport model REM-CALGRID. *Atmospheric Environment* 41, 9206-9216.
- Ghosh, S.K., Saha, R., Saha, B., 2015. Toxicity of inorganic vanadium compounds. *Research on Chemical Intermediates* 41(7), 4873-4897.
- Hleis, D. 2010. Evaluation de la contribution d'émissions sidérurgiques à la teneur en particules en suspension dans l'atmosphère à une échelle locale. Tesis doctoral de la Université du Littoral Côte d'Opale.
- Hernández, E., Martín, F., Valero, F., 1992. Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain. *Atmospheric Environment* 26B, 107-116.
- Hernández-Bonilla, D., Schilman, A., Montes, S., Rodríguez-Agudelo, Y., Rodríguez-Dozal, S., Solís-Vivanco, R., Ríos, C., Riojas-Rodríguez, H., 2011. Environmental exposure to manganese and motor function of children in Mexico. *Neurotoxicology* 32(5), 615-621.
- Jolliff, J.K., Stow, C.A., McGillicuddy, D.J., Allen, J.I., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. *J. Marine Syst.* 76, 4-15.
- Kfoury, A. 2013. Origin and physicochemical behaviour of atmospheric PM_{2.5} in cities located in the littoral area of the Nord-Pas-de-Calais region, France. Tesis doctoral de la Université du Littoral Côte d'Opale.
- Kumar, A., Luo, J., Bennett, G., 1993. Statistical evaluation of lower flammability distance (LFD) using four hazardous release models. *Process. Safety Progress* 12, 1-11.
- Mergler, D., Baldwin, M., Bélanger, S., Larribe, F., Beuter, A., Bowler, R., Panisset, M., Edwards, R., de Geoffroy, A., Sassine, M.P., Hudnell, K., 1999. Manganese neurotoxicity, a continuum of dysfunction: results from a community based study. *Neurotoxicology* 20(2-3), 327-342.
- Pederzoli, A., Thunis, P., Georgieva, E., Borge, R., Carruthers, D., 2011. Performance criteria for the benchmarking of air quality model regulatory applications: the "target" approach. 14th Conference on harmonisation within atmospheric dispersion modelling for regulatory purposes (HARMO14), 297-301.
- Santos, G., Terán, S., Fernández-Olmo, I., 2016. Influence of atmospheric stability on the estimation of trace metal concentrations by means of linear statistical models. 10th International Conference on Air Quality – Science and Application (Air Quality 2016).
- Vicente, A.B., Jordán, M.M., Sanfeliu, T., Sánchez, A., Esteban, M.D., 2012. Air pollution prediction models of particles, As, Cd, Ni and Pb in a highly industrialized area in Castellón (NE, Spain). *Environmental Earth Sciences* 66(3), 879-888.



Capítulo 5. Conclusiones |

| Chapter 5. Conclusions

Conclusiones

Las principales conclusiones extraídas del conjunto de esta Tesis Doctoral son las siguientes:

- Se ha comprobado la viabilidad del uso de técnicas de estimación basadas en modelos empíricos para la evaluación de la calidad del aire con respecto a los metales y semimetales (As, Cd, Ni y Pb) regulados por la Directiva Marco de calidad el aire de la Unión Europea (Directiva 2008/50/CE) y la Directiva 2004/107/CE en tres zonas urbanas de Cantabria y una zona urbana con elevada influencia industrial del norte de Francia (Dunkerque). En esta última área también se ha podido comprobar la viabilidad de emplear este tipo de técnicas para estimar los niveles en aire de metales no regulados por la legislación europea, como el Cr, Mn y V.
- En particular, se ha demostrado la validez de las técnicas de modelización estadística basadas en regresión por mínimos cuadrados parciales y redes neuronales artificiales para la estimación de los niveles de As, Cd, Ni y Pb en aire ambiente en zonas urbanas en las que los niveles de estos contaminantes se encuentran por debajo de los correspondientes umbrales inferiores de evaluación establecidos por la Directiva Marco de calidad el aire de la Unión Europea (Directiva 2008/50/CE). Esta validez ha sido demostrada no solo en cuanto al ajuste de los datos sino también en cuanto a la capacidad de generalización de las técnicas de modelización, evaluada por medio de un procedimiento de validación externa.
- Se ha demostrado la aplicabilidad de la metodología propuesta como método de evaluación de la calidad del aire en zonas urbanas con respecto a dichos metales y semimetales regulados, en el contexto de la Directiva Europea de calidad del aire, en base a la aceptabilidad de los resultados con respecto a dos aspectos fundamentales: (i) el cumplimiento de los requerimientos de incertidumbre propuestos en dicha Directiva para las técnicas de estimación objetiva, y (ii) la adecuación en la estimación de los valores medios de series temporales de concentración en inmisión. Esto es así aun cuando las bases de datos de entrada para el desarrollo de los modelos se componen de un número limitado de observaciones, procedan bien de campañas extensivas de muestreo o bien de campañas intensivas, e incluso cuando los niveles de metales presentan una gran variabilidad.
- No obstante, se ha comprobado que las técnicas de modelización desarrolladas a partir de bases de datos de entrada que se componen de un número limitado de observaciones presentan ciertas dificultades en su capacidad para proporcionar

estimaciones adecuadas de las variaciones temporales de los niveles de As, Cd, Ni y Pb.

- En aquellas zonas urbanas en las que el origen de la contaminación atmosférica es atribuible a diversas y variadas fuentes de emisión, existe una correlación débil entre las variables independientes consideradas de influencia para el desarrollo de los modelos de estimación (concentración en inmisión de contaminantes mayoritarios y variables meteorológicas) y las concentraciones en inmisión de microcontaminantes como As, Cd, Ni y Pb. En este sentido, se ha comprobado que estrategias alternativas como la consideración de variables meteorológicas relacionadas con la estabilidad atmosférica puede dar lugar a una mejora moderada en el desempeño de los modelos estadísticos de calidad del aire con respecto a Ni y Pb. No obstante, debido a la complejidad en la computación de los datos a partir de sondeos termodinámicos, se propone su consideración únicamente en aquellas situaciones en las que la metodología simplificada falle o presente un amplio margen de mejora.
- Por otra parte, es necesario llevar a cabo una actualización periódica de los modelos desarrollados, en especial cuando se sospecha que se hayan podido producir cambios en los patrones de emisión con el tiempo, lo que supondría una modificación de las relaciones empíricas establecidas entre las variables estudiadas.

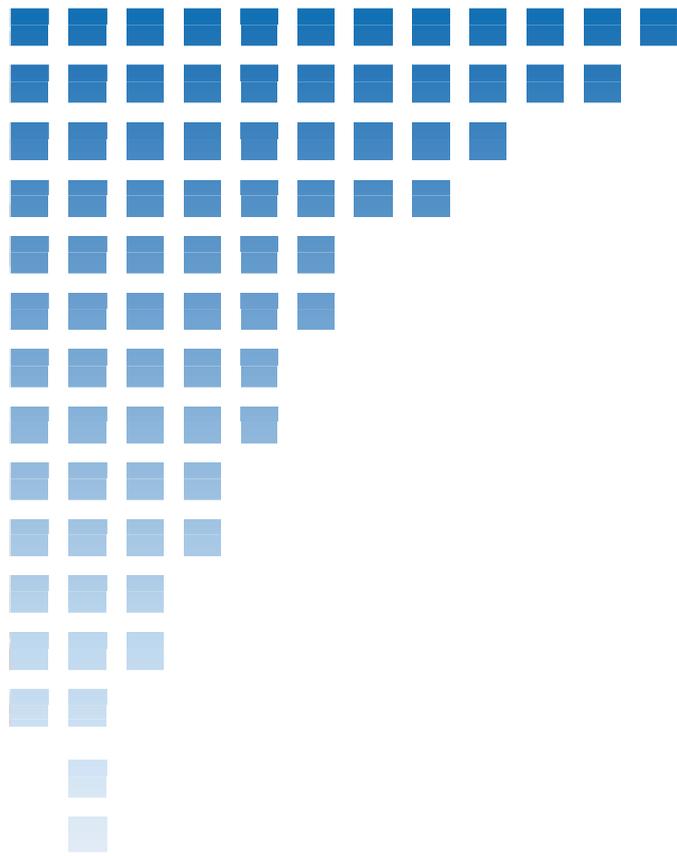
Conclusions

The following main conclusions have been drawn from this PhD Thesis:

- The feasibility of the use of estimation techniques based on empirical models for the air quality assessment in relation to metals and metalloids (As, Cd, Ni and Pb) regulated by the Air Quality Framework Directive of the European Union (Directive 2008/50/EC) and the Directive 2004/107/EC has been confirmed in three urban areas in Cantabria and an urban area with a strong industrial influence in the North of France (Dunkerque). In the latter, the feasibility of the use of this sort of techniques to estimate the ambient air levels of metals that are not regulated by the European legislation (Cr, Mn and V) has also been verified.
- Particularly, the validity of statistical modelling techniques based on Partial Least Squares Regression and Artificial Neural Networks to estimate the ambient air levels of As, Cd, Ni and Pb in urban areas where those levels are below the corresponding lower assessment thresholds established by the EU Air Quality Framework Directive (Directive 2008/50/EC) has been demonstrated, not only regarding the capability of these modelling techniques to fit the data but also regarding their generalization ability, which has been tested by means of an external validation procedure.
- The applicability of the proposed methodology as a method for air quality assessment with respect to the regulated metals and metalloids, in the context of the European Air Quality Directive, has been confirmed based on the acceptability of the results obtained regarding two main aspects: (i) the fulfilment of the uncertainty requirements proposed in the Directive for objective estimation techniques, and (ii) the adequacy of the estimation of mean values from time series of immission concentration. An acceptable quality level is achieved despite the fact that the input databases consist of a limited number of observations, disregarding if the data come from extensive or intensive sampling campaigns, and even when the levels of the studied pollutants present a significant variability.
- However, the studied modelling techniques, developed using input databases with a limited number of observations, present difficulties in providing with accurate estimations of the time series of levels of As, Cd, Ni and Pb.
- In those urban areas where the origin of air pollution is attributable to various and varied emission sources, there is a weak correlation between the independent variables considered to be influential for model development (major air pollutant concentrations and meteorological variables) and immission concentrations of

micropollutants such as As, Cd, Ni and Pb. In this sense, it has been observed that alternative strategies such as the consideration of meteorological variables related to atmospheric stability as additional inputs may lead to a moderate improvement in the performance of the air quality statistical models with respect to Ni and Pb. Nevertheless, given the complexity associated with the computation of data from atmospheric soundings, this strategy should only be considered in those situations in which the simplified methodology fails or there is plenty of scope for improvement.

- Finally, it will be required to conduct periodic updates of the developed models, particularly if there is evidence that emission patterns may have changed in time, what would imply a modification of the empirical relationships established between the studied variables.



Anexo I. Difusión de resultados

AI.1. Comunicaciones a congresos

A continuación se listan las contribuciones a congresos nacionales e internacionales a través de los cuales se ha realizado la difusión de los resultados de la presente tesis.

AI.1.1. Congresos nacionales

- **Santos, G.**, Fernández-Olmo, I., Irabien, A. Herramientas para la evaluación de la calidad del aire respecto a los metales pesados. XXXIV Reunión Bienal de la Real Sociedad Española de Química. Septiembre **2013**, Santander (España). Comunicación poster.

AI.1.2. Congresos internacionales

- **Santos, G.**, Terán, S., Fernández-Olmo, I., Influence of atmospheric stability on the estimation of trace metal concentrations by means of linear statistical models. 10th International Conference on Air Quality – Science and Application (Air Quality 2016). Marzo **2016**, Milán (Italia). Comunicación poster.
- **Santos, G.**, Fernández-Olmo, I., Irabien, A., Ledoux, F., Courcot, D., PLSR and ANN estimation models for PM₁₀-bound heavy metals in Dunkerque (Northern France). International Conference on Atmospheric Dust (DUST 2014). Junio **2014**, Castellaneta Marina (Italia). Comunicación oral.
- **Santos, G.**, Fernández-Olmo, I., Irabien, A. PLSR and ANN approaches to estimate heavy metal levels in airborne PM₁₀. 9th International Conference on Air Quality - Science and Application (Air Quality 2014). Marzo **2014**, Garmisch-Partenkirchen (Alemania). Comunicación oral.
- **Santos, G.**, Fernández-Olmo, I. Estimation of ambient air levels of regulated heavy metals by means of Partial Least Squares Regression (PLSR). 15th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO15). Mayo **2013**, Madrid (España). Comunicación oral.

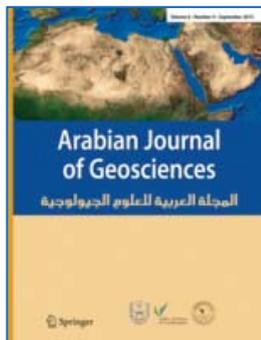
AI.2. Publicaciones en revistas científicas

A continuación se listan los trabajos que, en el marco de esta Tesis Doctoral, han sido publicados en revistas científicas incluidas en el *Journal of Citation Reports – Science Edition (JCR)*:



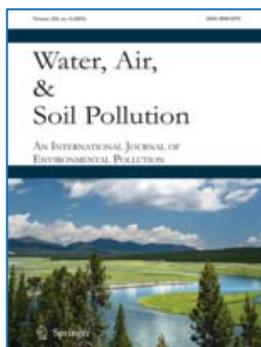
Índice de impacto: 4.099

- **Santos, G.,** Fernández-Olmo, I., **2016.** A proposed methodology for the assessment of arsenic, nickel, cadmium and lead levels in ambient air. *Science of the Total Environment* 554-555, 155-166.



Índice de impacto: 1.224

- **Santos, G.,** Fernández-Olmo, I., Irabien, A., Ledoux, F., Courcot, D., **2016.** Estimating airborne heavy metal concentrations in Dunkerque (northern France). *Arabian Journal of Geosciences* 9:231.



Índice de impacto: 1.554

- **Santos, G.,** Fernández-Olmo, I., Irabien, A., **2015.** Estimation of PM₁₀-bound As, Cd, Ni and Pb levels by means of statistical modelling: PLSR and ANN approaches. *Water, Air, and Soil Pollution* 226:275.

Estas publicaciones pueden ser consultadas en las siguientes páginas.

1 **Estimation of PM₁₀-bound As, Cd, Ni and Pb levels by means of statistical modelling: PLSR and ANN**
2 **approaches**

3

4 **Authors:** Germán Santos*, Ignacio Fernández-Olmo, Ángel Irabien

5

6 Universidad de Cantabria, Dep. Ingenierías Química y Biomolecular, Avda. Los Castros s/n, 39005

7 Santander (Cantabria), Spain

8

9 *Corresponding author:

10 Tel.: +34942201579; fax: +34942201591

11 E-mail address: santosg@unican.es

12 **1. Introduction**

13

14 Mathematical modelling for air quality assessment purposes has become increasingly important in recent
15 years. These models consist of a set of analytical/numerical algorithms that describe the physical and
16 chemical aspects of a problem and can be divided into two main groups: (i) deterministic models based on
17 fundamental mathematical descriptions of atmospheric processes, where emissions (causes) generate air
18 pollution (effects); (ii) statistical (empirical) models based on semiempirical statistical relations between
19 available data of input variables that are believed to be representative of the process behaviour and
20 measurements of the target parameters/properties of the system output. Moreover, the European Union Air
21 Quality Framework Directive establishes that in all zones and agglomerations where the level of pollutants
22 is below the lower assessment threshold (LAT), which is expressed as a percentage of the corresponding
23 target/limit value, modelling techniques or objective estimation techniques (or both) shall be sufficient for
24 the assessment of the ambient air quality (European Council Directive 2008/50/EC). Both statistical and
25 deterministic methods are currently used in regulatory air pollution forecasting by environmental
26 authorities.

27

28 Although deterministic models have some advantages over statistical models, such as a full-coverage 3D
29 domain, in some particular situations they may have some drawbacks in terms of accuracy and input data
30 uncertainty. According to Hanna (1989), generally, a larger number of input parameters corresponds to a
31 lower model uncertainty and smaller prediction errors, but unfortunately, by extending the number of input
32 parameters, the error and uncertainty attached to the input data also increase. Therefore, complex
33 deterministic models work well when their extensive input data requirements are satisfied, which rarely
34 occurs with some pollutants, such as As, Ni, Cd and Pb. This is due to the fact that the presence of these
35 pollutants in the atmosphere normally originates in a variety of pollution sources, not exclusively bound to
36 specific industrial activities at a certain location. As a consequence, the emission rates of these pollutants
37 from all the point or area sources are difficult to estimate. A solution to address this problem consists in
38 performing a spatial disaggregation of emission inventories (Maes et al. 2009). Notwithstanding, there is
39 an underlying uncertainty associated with the method of disaggregation together with the inherent
40 uncertainty of the emission inventories themselves. For that reason, for pollutants like the ones under study
41 in this work, the performance of complex models is often equal to that of simpler methodologies. This fact

42 highlights the interest of statistical models (e.g., linear regression techniques and non-linear modelling
43 techniques) to estimate the ambient air concentration of atmospheric pollutants even though a wide range
44 of deterministic models, as reviewed by El-Harbawi (2013), have been already developed and studied in
45 the literature. Nevertheless, techniques such as partial least squares regression (PLSR), which presents
46 advantages over other statistical linear regression techniques because it combines features from factor
47 analysis statistical methods, such as principal component analysis (PCA) and linear regression techniques,
48 as multiple linear regression (MLR), may potentially lead to more accurate estimations than those provided
49 by MLR or principal component regression (PCR). Furthermore, according to Wold et al. (2001), although
50 regression techniques such as MLR works reasonably well with problems involving fairly few uncorrelated
51 independent variables, PLSR is preferable when analysing more intricate problems because it is able to
52 manage simultaneously numerous and collinear predictor variables and responses. Despite the fact that it
53 has been widely applied in other disciplines, chemometrics in particular, and used in some works related to
54 atmospheric pollution (Ogulei et al. 2006; Wingfors et al. 2001), there are few studies on the application of
55 PLSR to predict atmospheric pollutant concentrations. Pires et al. (2008) tested the ability of different linear
56 models, including PLSR, to predict daily mean concentrations of particles with an aerodynamic diameter
57 of less than 10 μm (PM_{10}) in Oporto (Portugal). It was obtained that even though every model fitted the
58 data similarly well, PLSR shows higher generalization ability than other linear techniques. Polat and
59 Durduran (2012) used regression models such as least squares regression (LSR), PLSR and MLR to predict
60 daily particulate matter concentration values in the city of Konya (Turkey). PLSR performance, slightly
61 better than those of the other regression models, was remarkably improved by considering data pre-
62 processing methods such as output-dependent data scaling (ODDS). Singh et al. (2012) compared PLSR
63 with non-linear modelling approaches to predict respirable suspended particulate matter (RSPM), SO_2 and
64 NO_2 in Lucknow city (India). Both linear and non-linear approaches provided adequate estimations,
65 especially for the RSPM, with values of correlation coefficient up to 0.9. Nonetheless, non-linear models
66 performed relatively better than the linear PLSR models.

67

68 With respect to non-linear modelling approaches, artificial neural networks (ANNs) have been suggested
69 as fair alternatives to statistical linear regression methods because they usually provide equal or superior
70 results, especially when there is non-linear behaviour involved in the problem under analysis, i.e., cases in
71 the atmospheric sciences (Gardner and Dorling 1998). For this reason, ANNs are particularly expected to

72 produce good predictive results when modelling PM mass concentrations compared with common gaseous
73 pollutants based on their ability to capture the highly non-linear character of the complex processes that
74 control the formation, transportation and removal of aerosols in the atmosphere (Grivas and Chaloulakou
75 2006). Furthermore, ANNs have been extensively applied in the past in the atmospheric literature with
76 successful results regarding forecasting major gaseous air pollutant concentrations, such as nitrogen oxides
77 (Gardner and Dorling 1999; Kolehmainen et al. 2001; Lu et al. 2003), sulphur dioxide (Chelani et al.
78 2002a), and (commonly) ozone (Abdul-Wahab and Al-Alawi 2002; Chaloulakou et al. 2003; Comrie 1997;
79 Inal 2010; Sousa et al. 2007; Wang et al. 2003; Yi and Prybutok 1996). Moreover, a number of studies have
80 been conducted using ANN approaches to forecast airborne PM mass concentrations (Caselli et al. 2009;
81 Chelani 2005; Grivas and Chaloulakou 2006; Hoi et al. 2009; Kim et al. 2009; Papanastasiou et al. 2007;
82 Paschalidou et al. 2011; Perez and Reyes 2002; Perez and Reyes 2006; Pérez et al. 2000; Voukantsis et al.
83 2011), predict PM mass concentrations, and predict other gaseous pollutant concentrations (Brunelli et al.
84 2007; Cai et al. 2009; Hrust et al. 2009; Jiang et al. 2004; Kukkonen et al. 2003; Kurt et al. 2008; Lu et al.
85 2004; Lu et al. 2003; Niska et al. 2005; Turias et al. 2008). Nevertheless, regarding the PM composition
86 and estimation of PM constituents, few studies have been conducted. In particular, with respect to the metal
87 content in PM, Chelani et al. (2002b) used ANNs to predict ambient PM₁₀ and metals, such as Cd, Cr, Fe,
88 Ni, Pb and Zn, in the air of Jaipur, India, in 1999. It was observed that the ANN models were able to predict
89 all the pollutant concentrations with low values of root means square error (RMSE). Nonetheless, more
90 studies related to atmospheric metal concentration estimations by means of ANNs have been conducted,
91 such as the study performed by Li et al. (2009) in which statistical models based on back-propagation ANNs
92 and MLR are applied to reconstruct occupational manganese exposure. Apart from ANNs, some research
93 has been conducted to model metal concentrations in ambient air using other statistical approaches.
94 Hernández et al. (1992) applied state-space modelling, Box-Jenkins modelling and time series
95 autoregressive integrated moving average (ARIMA) models to estimate the daily concentrations of air-
96 particulate Fe and Pb in Madrid (Spain). Predictions of daily Fe were better than those of Pb. No difference
97 being found between State-space and Box-Jenkins models, their outcomes were better than those of
98 ARIMA models in terms of root mean squared error (RMSE), correlation coefficient and efficiency.
99 Chelani et al. (2001) used a state-space model coupled with Kalman filter and an autoregressive model with
100 external input (ARX model) to forecast Pb, Fe and Zn along with RSPM in Delhi (India). The state space
101 model performed better than the ARX model. On the other hand, Vicente et al. (2012) developed predictive

102 models based on multiple regression analysis together with time series (ARIMA) models to predict the
103 concentration of total suspended particles (TSP), PM₁₀, As, Cd, Ni and Pb in the ambient air of Castellón
104 (Spain). Furthermore, in a previous study conducted by Arruti et al. (2011), estimations of As, Cd, Ni and
105 Pb levels in Cantabria (Spain) by means of statistical MLR and PCR models have been conducted. It is
106 concluded that both represent valid approaches as objective estimation techniques.

107

108 This paper is focused on the development of PLSR and ANN statistical models to estimate the levels of As,
109 Cd, Ni and Pb in the ambient air of two urban areas: Castro Urdiales and Reinoso in the Cantabria region
110 (northern Spain). These models are evaluated according to the uncertainty requirements established by the
111 EU for objective estimation techniques as well as for their ability to estimate the mean concentration.
112 Additionally, an external validation of the models developed is performed.

113

114 **2. Materials and methods**

115

116 *2.1. Statistical model fundamentals*

117

118 *2.1.1. Partial least squares regression (PLSR)*

119

120 Partial least squares regression is a multivariate calibration technique whose aim is to investigate the
121 relationship between a set of dependent variables or responses and a set of independent variables known as
122 predictors. Firstly, in a similar manner to PCA, PLSR performs a decomposition of the original predictor
123 variables (X-matrix, which consists of environmental observations in this study) by projecting them to a
124 new space and extracts a set of orthogonal factors, called latent variables, which have the best predictive
125 ability. Simultaneously, a decomposition of the response variables (Y-matrix, composed of metal level
126 observations) is also performed. This decomposition step is made in a manner that the projections (scores)
127 of X have maximum covariance with the projections of Y. This procedure is followed by a regression stage,
128 where PLSR (just as MLR) creates a linear combination of the predictor variables in order to predict Y
129 (Abdi 2010).

130

131 In this work, cross-validation techniques were used to select the more suitable number of significant
132 components. PLS Toolbox (Eigenvector Research, Inc.) for MATLAB was used in the present study to
133 develop the PLSR models.

134

135 *2.1.2. Artificial neural networks (ANNs)*

136

137 Artificial neural networks are computational systems inspired by the biological central nervous system.
138 They consist of a number of simple process elements, commonly referred to as artificial neurons, which are
139 logically arranged into layers, highly interconnected, and interact with each other via weighted connections.
140 Through a supervised training process, in which they are successively presented with a series of input and
141 associated output data, ANNs are capable to learn to model highly non-linear relationships and, as a result,
142 to accurately generalise when previously unseen data are presented afterwards. The reader is referred the
143 handbooks of Bishop (1995) and Hassoun (1995) for a comprehensive description of the ANN technique.

144

145 Plenty of neural network architectures exist. In this work, based on the different ANN approaches found in
146 the air quality related literature, a multilayer perceptron (MLP) neural network architecture was selected;
147 details of the architecture are provided in Gardner and Dorling (1998).

148

149 Because the ratio of input variables/number of samples is relatively high in this work due to the number of
150 samples that were collected by the Regional Environmental Ministry, applying a dimension reduction
151 technique prior to the ANN models was expected to produce an improvement in the estimations as reported
152 in some studies (Lu et al. 2003; Sousa et al. 2007). Therefore, an alternative approach in which the PCA is
153 performed before the development of the ANN models (hereafter known as PCA-ANNs) is considered.

154

155 The ANN models in this study were developed using the Neural Network Toolbox for MATLAB
156 (MathWorks, Inc.).

157

158 *2.2. Study area*

159

160 Two urban areas in the Cantabria region (northern Spain) whose air quality may be influenced by the
161 presence of metallurgical and other industrial activities in their vicinity were selected: Castro Urdiales and
162 Reinosa (Fig. 1). The former area is a coastal urban site at the NE zone of Cantabria which has 32258
163 (2010) inhabitants and encompasses an area of approximately 97 km². Pollution in this area has a marked
164 anthropogenic origin which is caused by traffic, not in vain Castro Urdiales is surrounded by the main
165 national highway in the northern part of Spain. Pollution also proceeds from industrial activities, such as
166 chemical and metallurgical plants and an oil refinery, located 10-30 km SE (near the city of Bilbao). The
167 monitoring station is located at 43°22'53"N, 3°13'22" W and 20 m above sea level, in the core of the urban
168 area. In contrast, Reinosa, covering nearly 4 km² with approximately 10277 inhabitants (2010), is located
169 inland, at about 50 km off the shore, in the southern part of the region. The sampling station is located at
170 43°00'01"N, 4°08'13"W and 850 m above sea level. It is in close proximity to a steel manufacturing plant
171 and also to a national highway, main exit route from Cantabria, which establishes connection with the
172 central Iberian Peninsula.

173

174 *2.3. Input dataset*

175

176 The dataset used in this study is divided into response variables and predictor variables. The former data
177 consist of As, Cd, Ni and Pb concentrations (ng m⁻³) in airborne PM₁₀ for the period from 2008 to 2010 at
178 the two study sites. The PM₁₀ sampling was performed by the Cantabrian Regional Environmental Ministry
179 according to the reference method for the determination of the PM₁₀ fraction of suspended particulate matter
180 detailed in standard UNE-EN 12341:1999. 48h averaged samples of PM₁₀ were taken once every two weeks
181 for the period from 2008 to 2009 and 24h averaged samples of PM₁₀ were collected for 2010 with a weekly
182 sampling frequency. The content of a number of metals and metalloids in the PM₁₀ samples was determined
183 by our research group based on the standard method for the measurement of Pb, Cd, As and Ni in the PM₁₀
184 fraction of the suspended particulate matter described in standard UNE-EN 14902:2006. According to this,
185 after gravimetric determination of the particle concentration levels, the PM₁₀ filters were treated with
186 microwave-assisted acid digestion to extract the analytes into an aqueous solution prior to the analytical
187 determination of their concentration by inductively coupled plasma mass spectroscopy (ICP-MS). Further
188 details of this analytical method can be found in Arruti et al. (2010).

189

190 As a consequence of the high cost associated with the analytical determination of the content of this sort of
191 pollutants in particulate matter, a considerably low number of samples was selected for the analysis.
192 However, this number was sufficient to guarantee the minimum time coverage (14%) for indicative
193 measurements as European Council Directive 2004/107/EC requires.

194

195 The predictor variables are qualitative or nominal variables (Table 1) that take into account seasonal effects,
196 Saharan dust intrusion and weekend effects or quantitative or continuous variables, namely, meteorological
197 data and major atmospheric pollutant concentration, which are detailed in Table 2. With respect to the
198 nominal variables, the information regarding the occurrence of Saharan dust intrusion events has been
199 obtained from annual reports on African dust episodes over Spain (MAGRAMA 2015), which are
200 developed by the Spanish National Research Council (CSIC) in collaboration with the Spanish Ministry of
201 Agriculture, Food and Environment. In contrast, the continuous variables are measured automatically in
202 real time (maximum time resolution of fifteen minutes) at the monitoring stations of the Cantabrian
203 Regional Air Quality Monitoring Network located in the study sites and are available at the Regional
204 Environment Ministry website. Average values of continuous variables were calculated according to the
205 corresponding duration of the PM₁₀ sampling periods (48 hours for 2008-2009 samples and 24 hours for
206 2010 samples). Moreover, as regards to PM₁₀ concentration, it has been included as input variable in the
207 form of natural logarithm because of this transformation being reported to improve the performance of
208 regression models (Arruti et al. 2011).

209

210 Prior to model development it is always rather convenient to take account of the application of a data pre-
211 processing method, especially if there is lack of knowledge regarding the relative importance of the
212 variables. In this study, the following data pre-treatment procedure was applied:

213

- 214 1. Dependent variable normalisation by the respective LAT in order to minimise scale effects.
- 215 2. Input variable auto-scaling, subtracting the mean and dividing by the standard deviation, in an attempt
216 to make each variable a priori equally important.
- 217 3. Multivariate outlier identification and removal method based on Mahalanobis distance. It is a well-
218 known classical approach that computes the Mahalanobis distance (MD) of each observation as an
219 indicative measure of the distance of each data point from the centre of the multivariate data cloud. By

220 convention, this method identifies as outliers those observations with a large MD (exceeding the 99%
221 quantile of a chi-square distribution).

222

223 Apart from the data pre-processing treatment, over-fitting is another decisive matter that must be taken into
224 consideration beforehand so that it could be prevented. This term refers to the circumstance that occurs
225 when a model fit the data in such a manner that not only captures the underlying trend in the data but also
226 the unexplained variation or statistical noise and therefore it is unable to generalize properly - that is, to
227 correctly perform when new observations are presented. In order to overcome this phenomenon it is highly
228 recommended the consideration of an additional verification or cross-validation data subset, besides the
229 training or fitting dataset, to check the models performance during the model development stage (usually
230 known as calibration or fitting for PLSR and training for ANNs). Additionally, if the generalisation ability
231 of a model is to be tested, a subset of samples has to be kept in reserve to perform an external validation
232 with previously unused observations once the models have been developed. For that reason, the complete
233 dataset was divided into three different subsets: 60% for training/fitting, 20% for verification and 20% for
234 external validation. Data partition of the available data, often randomly conducted, was carried out in this
235 work by means of the Kennard-Stone algorithm (Kennard and Stone 1969) with the purpose that the
236 resulting subsets are statistically representative. This data division method, originally developed for design
237 of experiments, has been traditionally applied to select calibration samples extracting subsets, as much
238 diverse as possible, from a large set of candidate samples based on the Euclidean distance, which is
239 employed as a measure of similarity between samples (the lower the Euclidean distance, the higher the
240 similarity). Initially, the pair of samples with the largest Euclidean distance are selected. Subsequently, by
241 means of an iterative process that concludes when the number of required objects is reached, more samples
242 are selected, maximizing the minimal Euclidean distances between those already selected and the remaining
243 samples.

244

245 *2.4. Model evaluation*

246

247 The main criteria employed in this work to determine whether a model is suitable for air quality assessment
248 purposes is principally based on two aspects: (i) the fulfilment of the European Union uncertainty
249 requirements for objective estimation techniques, which are shown in Table 3 and (ii) the accuracy of

250 estimated mean values. Additionally, a number of statistical parameters has been considered to evaluate the
251 modelling performance and are also shown in Table 3.

252

253 **3. Results and discussion**

254

255 *3.1. As, Cd, Ni and Pb levels in Castro Urdiales and Reinosa*

256

257 Fig. 2 summarises the levels of As, Cd, Ni and Pb in PM₁₀ at Castro Urdiales and Reinosa for the period
258 from 2008 to 2010. According to the European Council Directive 2008/50/EC, because these levels did not
259 exceed their lower assessment threshold and did not present significant variations throughout the period of
260 study, modelling and objective estimation techniques are permitted as an alternative method to experimental
261 measurements for air quality assessment.

262

263 *3.2. Statistical estimation models for Castro Urdiales*

264

265 Table 4 shows the results relating to the best-developed models at the Castro Urdiales site for the four
266 pollutants under study using the three approaches: PLSR, ANNs and PCA coupled with ANNs. The results
267 obtained for both the training and the external validation subsets are presented.

268

269 Limit/target values for As, Cd, Ni and Pb in ambient air in the European regulations are given in annual
270 mean concentration values. Therefore, attention should be paid to the estimated mean concentrations in the
271 study period. The normalised mean concentrations are presented in Table 4. The accuracy in the estimation
272 of the mean concentration is evaluated by means of the fractional bias (FB) index. In this respect, the
273 estimations are more accurate for the training step. At this step, PLSR provides a FB index lower than those
274 obtained for ANNs and PCA-ANNs because the mean metal concentration estimated by the PLSR models
275 are equal—up to two significant figures—to the corresponding observed values and that, according to the
276 corresponding equation (Table 3), yields lower FB index values. However, the differences between
277 estimated and observed mean concentrations using the three considered techniques are not remarkably
278 significant. As for external validation, the precision is inferior to that of the training phase.

279

280 In a more illustrative way, Fig. 3 represents the mean metal concentration estimation expressed as a
281 percentage of the corresponding limit/target value. The vertical axis is presented in logarithmic scale. The
282 green area represents the zone below the LAT, the yellow area represents the zone between the UAT and
283 the LAT, and the red area is the zone between the limit/target value and the UAT. Fig. 3 shows that, even
284 though there are some differences between the estimated and the observed mean levels, they are similar.
285 Moreover, because the observed metal(loid) levels are within the green area, well below the LAT, even
286 higher discrepancy could be allowed. Therefore, the developed models provide satisfactory mean
287 concentration estimations.

288

289 It is necessary to validate objective estimation techniques in the context of the EU Directives in terms of
290 uncertainty. In this sense, according to Arruti et al. (2011), two indices have been considered: on the one
291 hand, the RME, which is defined as the largest concentration difference of all percentile differences
292 normalized by the respective observed value (Fleming and Stern 2007); on the other hand, the RDE, which
293 evaluates the accuracy in the estimation of the observation closest to the limit/target value (Denby 2009).
294 As observed in Table 4, the values of these indices for the four pollutants in question for the training and
295 the external validation are well below 100%, which is the maximum permissible uncertainty limit for using
296 objective estimation techniques as air quality assessment tools according to the European Council Directive
297 2008/50/EC. For As, Ni and Cd, ANNs provide higher RME values than PLSR and PCA-ANN. In the
298 majority of cases, except for the As and Cd ANN models, the RME values are below 50%, which is the
299 uncertainty requirement for modelling techniques. In all cases, the RDE values are below 10%. However,
300 these indices have some limitations: it has been discussed that RME is sensitive to the presence of outliers
301 resulting in an increase of the uncertainty values (Fleming and Stern 2007); RDE only evaluates the
302 uncertainty of just one sample, the closest to the limit/target value.

303

304 From a scientific point of view, apart from a precise estimation of mean values to comply with the policy
305 framework, a model should be able to correctly describe the temporal variations of dependent variables.
306 For this purpose, a set of statistics has been used in this work. In the first place, the correlation coefficient
307 is employed to measure the goodness of fit between the observed and the estimated values. The results
308 show that PLSR correlation coefficients, which are within the range of 0.6-0.7, are less variable than those
309 of ANNs and PCA-ANNs: whereas the highest correlation coefficient, an r value of 0.82, is found when

310 using ANNs for Pb training, the correlation coefficients for As and Cd ANN models are significantly low
311 and therefore unacceptable. This could be explained because in the area of study As and Cd tend to be in
312 lower concentration than Ni and Pb and consequently in the period of study a number of samples have
313 levels of As and Cd below their detection limits. As a result, models are trained to produce the same output
314 from different inputs, a detrimental contradiction that may negatively affect the estimation of the rest of the
315 samples. Moreover, as expected, the r values for external validation are often lower than those for training.
316 Nonetheless, the PCA-ANN external validation correlation coefficients are systematically below 0.5.

317

318 In addition to the correlation coefficient, the precision of the individual sample concentration estimation is
319 quantified by the RMSE, the NMSE and the FV, (see equations in Table 3). The RMSE values, which
320 provide information regarding the differences between the observed and estimated concentrations, are
321 shown in Table 4. However, to compare these differences for different approaches and pollutants, a
322 normalised version of this parameter (NMSE) is more preferable because it does not take into account the
323 range of the independent variable. In general, the three considered approaches provide low values of NMSE
324 in the order of 10^{-1} .

325

326 With respect to the FV index, positive values can be observed in Table 4; this indicates that the estimated
327 variance is lower than the observed variance. Therefore, estimated values are less dispersed than observed
328 values, which tend to be more distanced from the mean value. This fact, together with a positive FB
329 corresponding to a slight mean value underestimation, indicates that there are some shortcomings in the
330 model capacity to perfectly describe all the concentration variations, especially regarding peak values.
331 Nevertheless, despite no substantial differences being found when comparing PLSR and ANNs, in general
332 both models are able to capture the underlying trend and provide temporal variations with similar shape to
333 that of the observed values as depicted in Fig. 4 for Pb and Ni in the training stage.

334

335 Based on the results obtained, there is no improvement associated with considering a dimension reduction
336 technique such as PCA before the development of the ANNs. This could be accounted for the fact that most
337 ANNs suffer less from the curse of dimensionality than some other techniques, as they can concentrate on
338 a lower dimensional section of the high-dimensional space, which may be done, for instance, by
339 disregarding completely an input, setting the corresponding weights to zero. Hence, for this specific

340 application dimensionality reduction has been proven not to be effective because removing input variables
341 from the analysis entails a loss on the predictive ability of the model.

342

343 Furthermore, because these models are devised to be used when the pollutant levels are sufficiently lower
344 at a certain location, in principle the moderated inaccuracy to estimate peak values should not represent an
345 unacceptable drawback to acknowledge these models as proper approaches complying with regulatory
346 requirements: the uncertainty values obtained with the developed models and the accuracy in the estimation
347 of the mean values would be favourable enough from a regulatory perspective. Nonetheless, some
348 refinement is possible because, as mentioned, there are some difficulties in estimating the highest observed
349 concentrations, which are underestimated. In this regard, further work involving new additional input
350 variables and the enlargement of the database with additional samples from different periods of time would
351 be recommendable.

352

353 *3.3. Statistical estimation models for Reinosa*

354

355 Analogously to the results at the Castro Urdiales site, the statistical parameters corresponding to the best-
356 developed models at the Reinosa site are presented in Table 5.

357

358 Regarding the uncertainty indices, it is observed that, as in Castro Urdiales, the RME and RDE values at
359 the Reinosa site are below 100% for the estimations obtained with the three different models developed for
360 the four pollutants. Hence, the quality objectives for ambient air quality assessment by means of objective
361 estimation techniques are met. However, there is a general increase in the obtained RDE values, especially
362 for As and Ni, which are significantly greater than those obtained at the Castro Urdiales site.

363

364 In relation to the mean values, again, PLSR provides the lowest FB training values, but the FB external
365 validation values are greater than the training values. Although there are still evident differences between
366 the observed and estimated mean concentrations, 90% of the estimations do not differ by more than 50%.
367 Therefore, as shown in Fig. 5, the three developed models provide satisfactory estimations. Nonetheless, a
368 substantial increase in FB values is found in Reinosa compared with Castro Urdiales.

369

370 Results at the Reinosa site present more variability between the training and external validation correlation
371 coefficient values for each pollutant than the results at the Castro Urdiales site, which may be partially
372 accounted for the higher inherent variance of metal levels in Reinosa compared to those obtained in Castro
373 Urdiales. However, the ANN correlation coefficient values are generally equal or superior to those of PLSR
374 and PCA-ANNs. As for the errors in the individual sample concentration estimations, the NMSE values for
375 Reinosa and Castro Urdiales are within the same range. Nevertheless, the FV values are slightly greater in
376 Reinosa than in Castro Urdiales but still lower than 1.0, which represents 50% of the observed variance.

377

378 Results prove that these models provide an acceptable performance in varied areas of a region, even when
379 there is a complex pollution framework with diverse emission sources, as is the case of Castro Urdiales.
380 Nevertheless, because the models were trained on data for particular sites and having been demonstrated
381 that the precision in the estimation is dependent on the specific location, these models can therefore only
382 be used with confidence at those sites. This dependence is especially pronounced in the ANN models, which
383 produced a higher variability in the results than the PLSR or PCA-ANN models. This may be influenced
384 by the fact that a limited number of samples are used for developing the models due to the unavailability
385 of additional observations stemming from their costliness and time consumption. Thus, it could be inferred
386 that for small datasets, linear regression techniques can work as well as non-linear modelling approaches
387 in terms of the estimation of metal(loid) levels in ambient air.

388

389 **4. Conclusions**

390

391 Statistical models are developed as objective estimation techniques to estimate the As, Cd, Ni and Pb in
392 ambient air at a local scale in two urban areas in the Cantabria region (northern Spain): Castro Urdiales and
393 Reinosa. These models were built based on linear regression techniques, partial least squares regression
394 (PLSR), and the non-linear modelling technique of artificial neural networks (ANNs). Additionally, an
395 alternative approach is considered that performs principal component analysis (PCA) prior to the ANN
396 analysis (PCA-ANNs). Furthermore, these models were externally validated using previously unseen data.

397

398 The models are evaluated by means of a number of statistical parameters, including uncertainty indices, to
399 determine if they comply with the EU quality requirements for objective estimation techniques.

400 Additionally, the model performance in estimating the individual sample concentrations is evaluated by
401 means of a number of statistical parameters, including a correlation coefficient, RMSE, NMSE and FV.

402

403 Based on the results obtained, PLSR and ANN techniques are acceptable alternatives to estimate the mean
404 concentration of As, Cd, Ni and Pb for the period of study in the two considered sites while fulfilling the
405 uncertainty requirements for objective estimation techniques established in the EU Directives.
406 Consequently, PLSR and ANN-based statistical models represent a proper alternative to experimental
407 measurements for air quality assessment purposes in the area of study. However, ANNs have not
408 demonstrated to offer a clear superior performance over the linear regression technique, what may be
409 attributed to the modest size of the available database. Furthermore, the three considered approaches had
410 some difficulties providing accurate estimations of the levels of individual samples, particularly for the
411 external validation subset. Moreover, the application of PCA before the ANN model development did not
412 yield an improvement of the models.

413

414 **Acknowledgements**

415

416 The authors gratefully acknowledge the financial support from the Spanish Ministry of Economy and
417 Competitiveness through the Project CMT2010-16068. The authors also thank the Regional Environment
418 Ministry of the Cantabria Government for providing the PM₁₀ samples at the Castro Urdiales and Reinos
419 sites.

420

421 **References**

422

423 Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS
424 Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 97-106.

425

426 Abdul-Wahab, S. A., & Al-Alawi, S. M. (2002). Assessment and prediction of tropospheric ozone
427 concentration levels using artificial neural networks. *Environmental Modelling and Software*, 17(3), 219-
428 228.

429

430 Arruti, A., Fernández-Olmo, I., & Irabien, A. (2010). Evaluation of the contribution of local sources to
431 trace metals levels in urban PM_{2.5} and PM₁₀ in the Cantabria Region (Northern Spain). *Journal of*
432 *Environmental Monitoring*, 12, 1451-1458.

433

434 Arruti, A., Fernández-Olmo, I., & Irabien, A. (2011). Assessment of regional metal levels in ambient air
435 by statistical regression models. *Journal of Environmental Monitoring*, 13(7), 1991-2000.

436

437 Bishop, C. M. (1995). *Neural Networks for Pattern Recognition and Machine Learning*. Oxford: Clarendon
438 Press.

439

440 Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., & Vitabile, S. (2007). Two-days ahead prediction of daily
441 maximum concentrations of SO₂, O₃, PM₁₀, NO₂, CO in the urban area of Palermo, Italy. *Atmospheric*
442 *Environment*, 41(14), 2967-2995.

443

444 Cai, M., Yin, Y., & Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials
445 using artificial neural network approach. *Transportation Research Part D: Transport and Environment*,
446 14(1), 32-41.

447

448 Caselli, M., Trizio, L., De Gennaro, G., & Ielpo, P. (2009). A simple feedforward neural network for the
449 PM₁₀ forecasting: Comparison with a radial basis function network and a multivariate linear regression
450 model. *Water, air, and soil pollution*, 201(1-4), 365-377.

451

452 Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and
453 regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, 313(1-
454 3), 1-13.

455

456 Chelani, A. B., Gajghate, D. G., Tamhane, S. M., & Hasan, M. Z. (2001). Statistical modeling of ambient
457 air pollutants in Delhi. *Water, air, and soil pollution*, 132(3-4), 315-331.

458

459 Chelani, A. B., Chalapati Rao, C. V., Phadke, K. M., & Hasan, M. Z. (2002a). Prediction of sulphur dioxide
460 concentration using artificial neural networks. *Environmental Modelling and Software*, 17(2), 161-168.
461

462 Chelani, A. B., Gajghate, D. G., & Hasan, M. Z. (2002b). Prediction of ambient PM₁₀ and toxic metals
463 using artificial neural networks. *Journal of the Air and Waste Management Association*, 52(7), 805-810.
464

465 Chelani, A. B. (2005). Predicting chaotic time series of PM₁₀ concentration using artificial neural network.
466 *International Journal of Environmental Studies*, 62(2), 181-191.
467

468 Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of*
469 *the Air and Waste Management Association*, 47(6), 653-663.
470

471 Denby, B. (2009). Guidance on the use of the models for the European Air Quality Directive. A Working
472 document of the Forum for Air Quality Modelling in Europe, FAIRMODE. ETC/ACC Report.
473

474 El-Harbawi, M. (2013). Air quality modelling, simulation, and computational methods: A review.
475 *Environmental Reviews*, 21(3), 149-179.
476

477 Fleming, J., & Stern, R. (2007). Testing model accuracy measures according to the EU directives-examples
478 using the chemical transport model REM-CALGRID. *Atmospheric Environment*, 41, 9206-9216.
479

480 Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review
481 of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15), 2627-2636.
482

483 Gardner, M. W., & Dorling, S. R. (1999). Neural network modelling and prediction of hourly NO_x and NO₂
484 concentrations in urban air in London. *Atmospheric Environment*, 33(5), 709-719.
485

486 Grivas, G., & Chaloulakou, A. (2006). Artificial neural network models for prediction of PM₁₀ hourly
487 concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment*, 40(7), 1216-1229.
488

489 Hanna, S. R. (1989). Plume dispersion and concentration fluctuations in the atmosphere. In P.N.
490 Cheremisinoff (Ed.), *Encyclopedia of Environmental Control Technology, Vol. 2, Air Pollution Control*.
491 Houston, Texas: Gulf Publishing Co.

492

493 Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. London, England: MIT Press.

494 Hernández, E., Martín, F., & Valero, F. (1992). Statistical forecast models for daily air particulate iron and
495 lead concentrations for Madrid, Spain. *Atmospheric Environment*, 26B, 107-116.

496

497 Hoi, K. I., Yuen, K. V., & Mok, K. M. (2009). Prediction of daily averaged PM₁₀ concentrations by
498 statistical time-varying model. *Atmospheric Environment*, 43(16), 2579-2581.

499

500 Hrust, L., Klaic, Z. B., Križan, J., Antonic, O., & Hercog, P. (2009). Neural network forecasting of air
501 pollutants hourly concentrations using optimised temporal averages of meteorological variables and
502 pollutant concentrations. *Atmospheric Environment*, 43(35), 5588-5596.

503

504 Inal, F. (2010). Artificial Neural Network Prediction of Tropospheric Ozone Concentrations in Istanbul,
505 Turkey. *Clean - Soil, Air, Water*, 38(10), 981.

506

507 Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004). Progress in developing an ANN model
508 for air pollution index forecast. *Atmospheric Environment*, 38(40 SPEC.ISS.), 7055-7064.

509

510 Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137-
511 148.

512

513 Kim, M., Kim, Y., Sung, S., & Yoo, C. (2009). Data-driven prediction model of indoor air quality by the
514 preprocessed recurrent neural networks. *ICCAS-SICE 2009 - ICROS-SICE International Joint Conference*
515 *2009, Proceedings*, 1688-1692.

516

517 Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components
518 used in air quality forecasting. *Atmospheric Environment*, 35(5), 815-825.

519

520 Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H.,
521 Dorling, S., Chatterton, T., Foxall, R., & Cawley, G. (2003). Extensive evaluation of neural network models
522 for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modelling system and
523 measurements in central Helsinki. *Atmospheric Environment*, 37(32), 4539-4550.

524

525 Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using
526 neural networks. *Environment international*, 34(5), 592-598.

527

528 Li, Y., Luo, F., Jiang, Y., Lu, Y., Huang, J., & Zhang, Z. (2009). A prediction model of occupational
529 manganese exposure based on artificial neural network. *Toxicology Mechanisms and Methods*, 19(5), 337-
530 345.

531

532 Lu, W. Z., Wang, W. J., Wang, X. K., Xu, Z. B., & Leung, A. Y. T. (2003). Using improved neural network
533 model to analyze RSP, NO_x and NO₂ levels in urban air in Mong Kok, Hong Kong. *Environmental*
534 *monitoring and assessment*, 87(3), 235-254.

535

536 Lu, W., Wang, W., Wang, X., Yan, S., & Lam, J. C. (2004). Potential assessment of a neural network model
537 with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong.
538 *Environmental research*, 96(1), 79-87.

539

540 Maes, J., Vliegen, J., Van de Vel, K., Janssen, S., Deutsch, F., De Ridder, K., Mensink, C. (2009). Spatial
541 surrogates for the disaggregation of CORINAIR emission inventories. *Atmospheric Environment*, 43, 1246-
542 1254.

543

544 Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA), 2015. Histórico de informes
545 de episodios naturales. [http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-](http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/gestion/anuales.aspx)
546 [y-calidad-del-aire/calidad-del-aire/gestion/anuales.aspx](http://www.magrama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/gestion/anuales.aspx)

547

548 Niska, H., Rantamäki, M., Hiltunen, T., Karppinen, A., Kukkonen, J., Ruuskanen, J., & Kolehmainen, M.
549 (2005). Evaluation of an integrated modelling system containing a multi-layer perceptron model and the
550 numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant
551 concentrations. *Atmospheric Environment*, 39(35), 6524-6536.

552

553 Ogulei, D., Hopke, P. K., Zhou, L., Patrick Pancras, J., Nair, N., & Ondov, J. M. (2006). Source
554 apportionment of Baltimore aerosol from combined size distribution and chemical composition data.
555 *Atmospheric Environment*, 40(SUPPL. 2), 396-410.

556

557 Papanastasiou, D. K., Melas, D., & Kioutsioukis, I. (2007). Development and assessment of neural network
558 and multiple regression models in order to predict PM₁₀ levels in a medium-sized Mediterranean city.
559 *Water, air, and soil pollution*, 182(1-4), 325-334.

560

561 Paschalidou, A. K., Karakitsios, S., Kleanthous, S., & Kassomenos, P. A. (2011). Forecasting hourly PM₁₀
562 concentration in Cyprus through artificial neural networks and multiple regression models: Implications to
563 local environmental management. *Environmental Science and Pollution Research*, 18(2), 316-327.

564

565 Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of PM_{2.5} concentrations several hours in advance using
566 neural networks in Santiago, Chile. *Atmospheric Environment*, 34(8), 1189-1196.

567

568 Perez, P., & Reyes, J. (2002). Prediction of maximum of 24-h average of PM₁₀ concentrations 30h in
569 advance in Santiago, Chile. *Atmospheric Environment*, 36(28), 4555-4561.

570

571 Perez, P., & Reyes, J. (2006). An integrated neural network model for PM₁₀ forecasting. *Atmospheric*
572 *Environment*, 40(16), 2845-2851.

573

574 Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Alvim-Ferraz, M. C. M., & Pereira, M. C. (2008). Prediction
575 of the daily mean PM₁₀ concentrations using linear models. *American Journal of Environmental Sciences*,
576 4(5), 445-453.

577

578 Polat, K., & Durduran, S. S. (2012). Usage of output-dependent data scaling in modeling and prediction of
579 air pollution daily concentration values (PM₁₀) in the city of Konya. *Neural Computing and Applications*,
580 21(8), 2153-2162.

581

582 Singh, K. P., Gupta, S., Kumar, A., & Shukla, S. P. (2012). Linear and nonlinear modeling approaches for
583 urban air quality prediction. *Science of the Total Environment*, 426, 244-255.

584

585 Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M., & Pereira, M. C. (2007). Multiple linear regression
586 and artificial neural networks based on principal components to predict ozone concentrations.
587 *Environmental Modelling and Software*, 22(1), 97-103.

588

589 Turias, I. J., González, F. J., Martín, M. L., & Galindo, P. L. (2008). Prediction models of CO, SPM and
590 SO₂ concentrations in the Campo de Gibraltar Region, Spain: A multiple comparison strategy.
591 *Environmental monitoring and assessment*, 143(1-3), 131-146.

592

593 Vicente, A. B., Jordán, M. M., Sanfeliu, T., Sánchez, A., & Esteban, M. D. (2012). Air pollution prediction
594 models of particles, As, Cd, Ni and Pb in a highly industrialized area in Castellón (NE, Spain).
595 *Environmental Earth Sciences*, 66(3), 879-888.

596

597 Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., & Kolehmainen, M. (2011).
598 Intercomparison of air quality data using principal component analysis, and forecasting of PM₁₀ and PM_{2.5}
599 concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Science of the Total*
600 *Environment*, 409(7), 1266-1276.

601

602 Wang, W., Lu, W., Wang, X., & Leung, A. Y. T. (2003). Prediction of maximum daily ozone level using
603 combined neural network and statistical characteristics. *Environment international*, 29(5), 555-562.

604

605 Wingfors, H., Sjödin, A., Haglund, P., & Brorström-Lundén, E. (2001). Characterisation and determination
606 of profiles of polycyclic aromatic hydrocarbons in a traffic tunnel in Gothenburg, Sweden. *Atmospheric*
607 *Environment*, 35(36), 6361-6369.

608

609 Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics.

610 *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.

611

612 Yi, J., & Prybutok, V. R. (1996). A neural network model forecasting for prediction of daily maximum

613 ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3), 349-357.

Figure captions

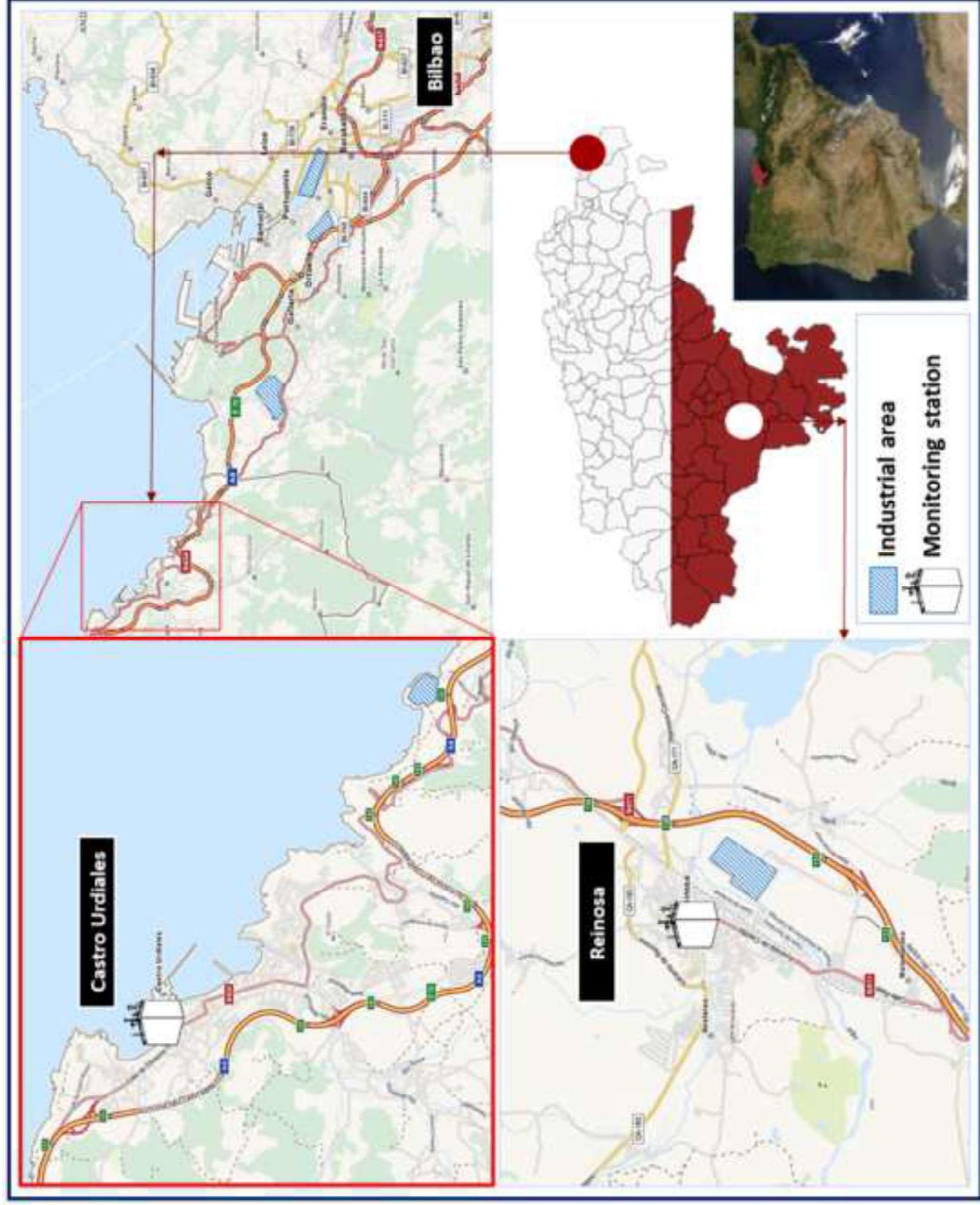
Fig. 1 Location of the monitoring stations

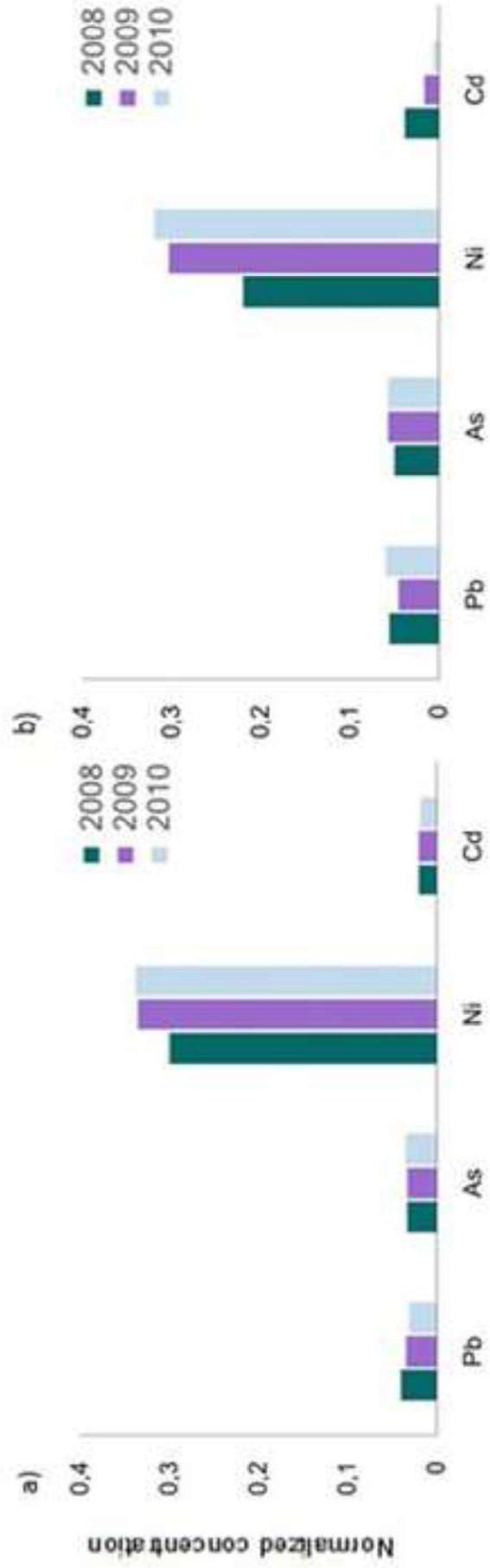
Fig. 2 As, Cd, Ni and Pb levels in PM₁₀, normalized with respect to their corresponding LAT, for the period of study at (a) the Castro Urdiales site and (b) the Reinoso site. The 2008 mean values are obtained from Arruti et al. (2011). LAT: 250 ng m⁻³ (Pb), 2.4 ng m⁻³ (As), 10 ng m⁻³ (Ni), 2 ng m⁻³ (Cd)

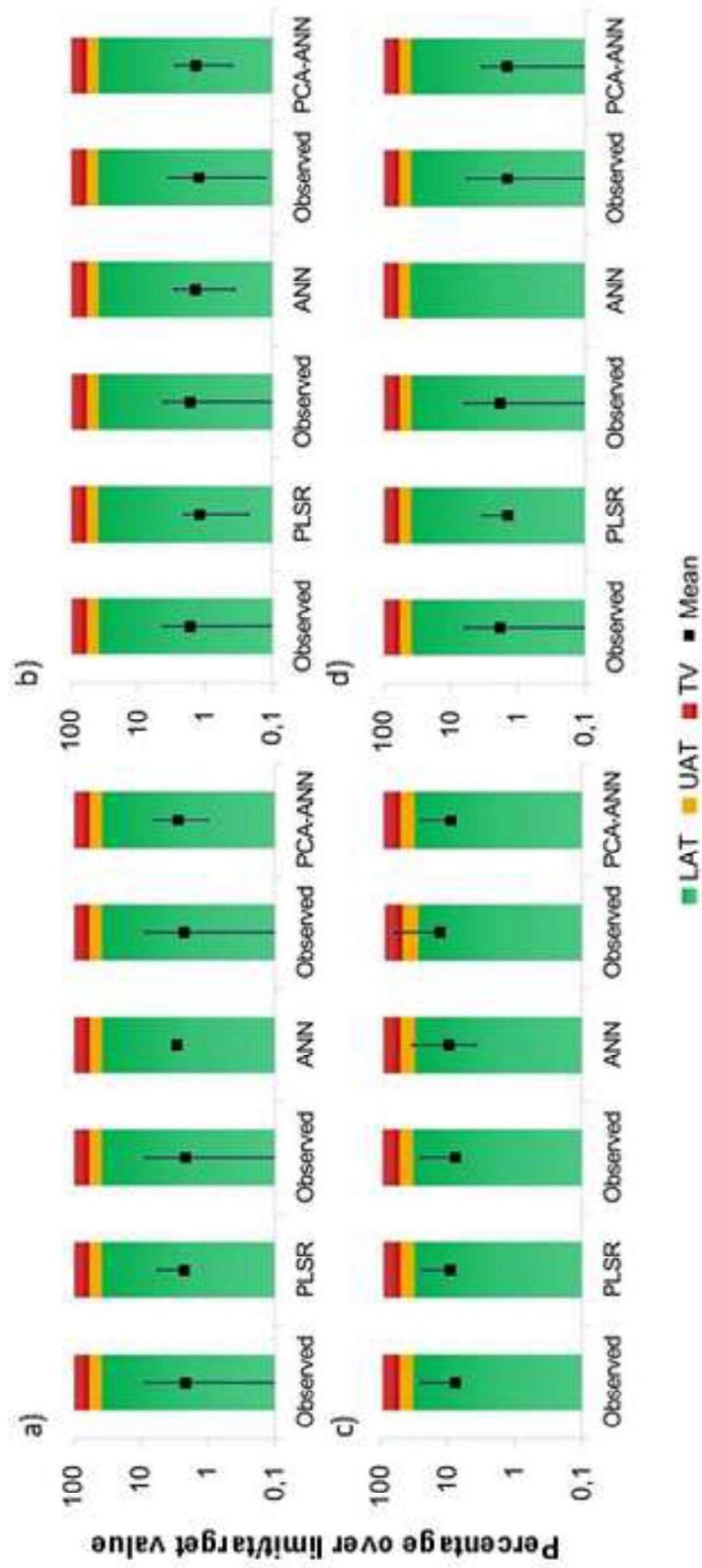
Fig. 3 Comparison between the observed and estimated mean concentrations at the Castro Urdiales site and their respective assessment thresholds and limit/target values. (a) Pb; (b) As; (c) Ni and (d) Cd. TV: 500 ng m⁻³ (Pb), 6 ng m⁻³ (As), 20 ng m⁻³ (Ni), 5 ng m⁻³ (Cd); UAT: 70% (Pb and Ni), 60% (As and Cd); LAT: 50% (Pb and Ni), 40% (As and Cd)

Fig. 4 Fitting of the Pb and Ni models for the training subset at the Castro Urdiales site

Fig. 5 Comparison between the observed and estimated mean concentrations at the Reinoso site and their respective assessment thresholds and limit/target values. (a) Pb; (b) As; (c) Ni and (d) Cd. TV: 500 ng m⁻³ (Pb), 6 ng m⁻³ (As), 20 ng m⁻³ (Ni), 5 ng m⁻³ (Cd); UAT: 70% (Pb and Ni), 60% (As and Cd); LAT: 50% (Pb and Ni), 40% (As and Cd)

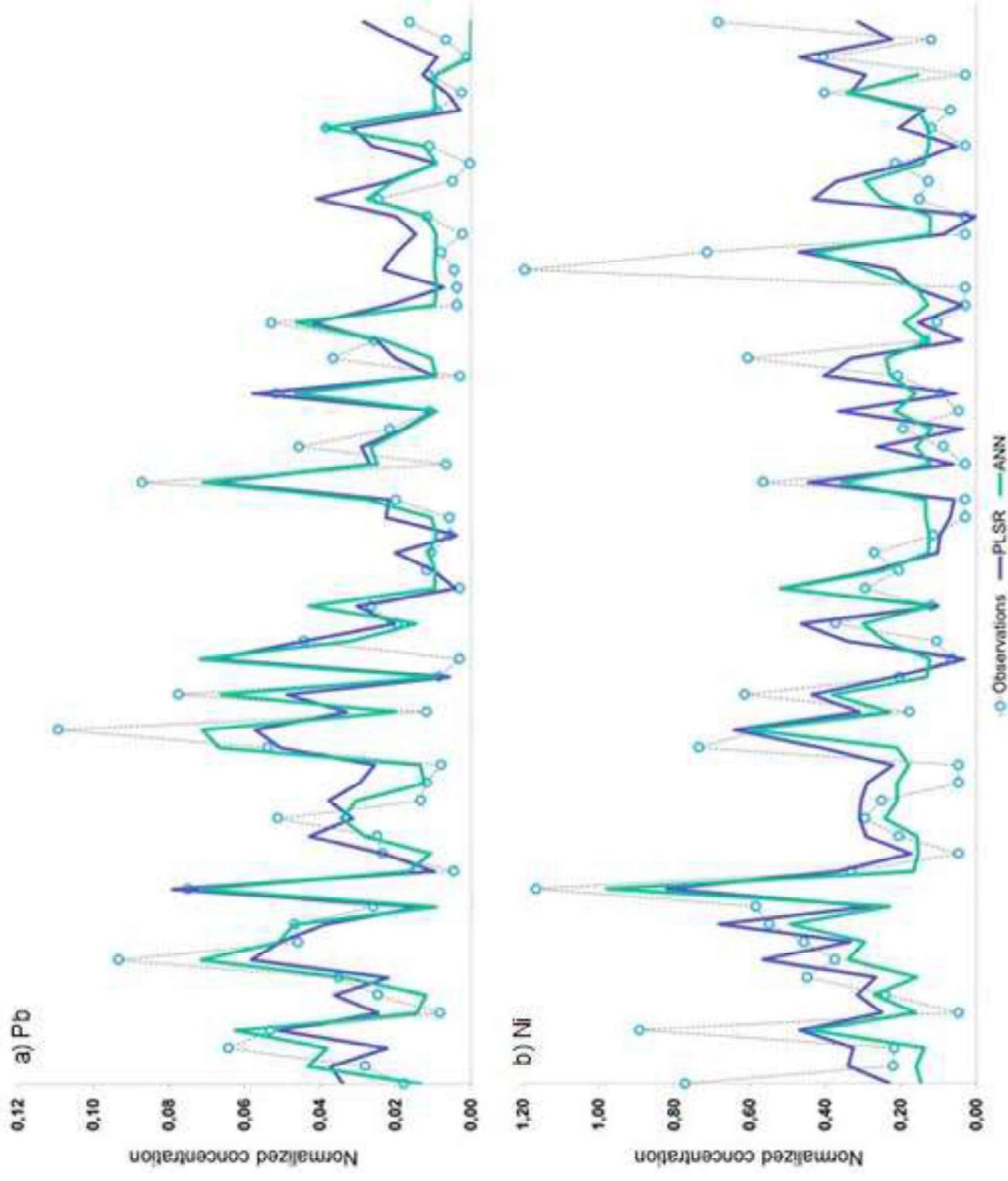






colour figure

[Click here to download colour figure: Fig 4.tif](#)



colour figure

[Click here to download colour figure: Fig 5.tif](#)

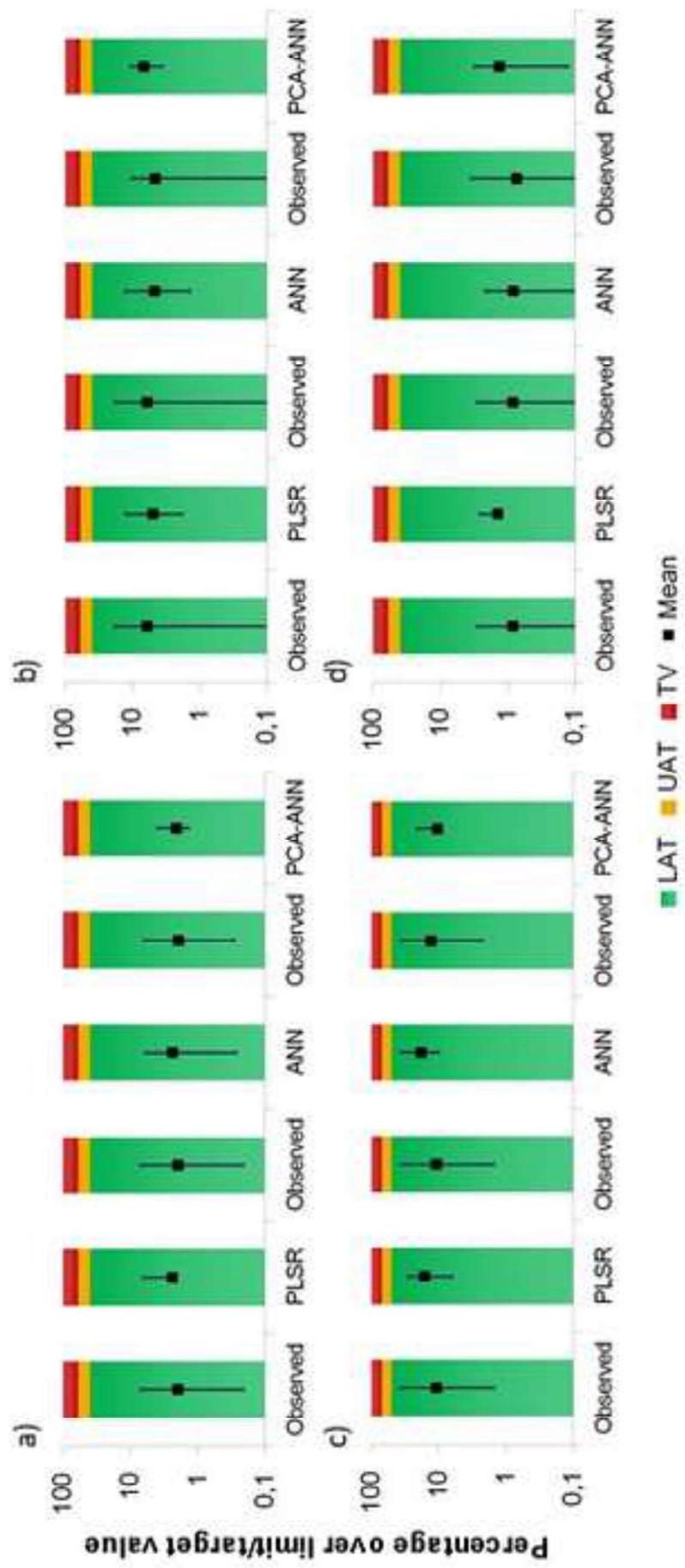


Table 1. List of nominal variables used as input for the models

Notation	Description	Codification
SE	Season	1: Winter; 2: Spring; 3: Summer; 4: Fall
SD	Saharan dust intrusion	0: No intrusion; 1: Intrusion
WE	Weekend	0: Working day; 1: Weekend

Table 2. List of continuous variables used as input for the models.

Notation	Description^a	Type	Units
LnPM ₁₀	Average natural logarithm of PM ₁₀ concentration ($\mu\text{g m}^{-3}$)	Major air pollutant	-
SO ₂	Average concentration of sulphur dioxide	Major air pollutant	$\mu\text{g m}^{-3}$
O ₃	Average concentration of ozone	Major air pollutant	$\mu\text{g m}^{-3}$
NO _x	Average concentration of nitrogen oxides	Major air pollutant	$\mu\text{g m}^{-3}$
T	Average temperature	Meteorological	°C
RH	Average relative humidity	Meteorological	%
WD	Prevailing wind direction	Meteorological	°
WS	Prevailing wind speed	Meteorological	ms ⁻¹
P	Average pressure	Meteorological	mbar
PP	Cumulative precipitation	Meteorological	L m ⁻²

^a Average values were calculated according to the corresponding duration of the PM₁₀ sampling periods

Table 3. Statistical parameters used for evaluating the model performance

Evaluation	Statistic	Equation
EU Uncertainty	Relative maximum error without timing	$RME = \max(C_{O,p} - C_{E,p}) / C_{O,p}$
	Relative directive error	$RDE = C_{O,LV} - C_{E,LV} / LV$
Mean concentration	Fractional bias	$FB = \frac{\overline{C_O} - \overline{C_E}}{0.5(\overline{C_O} + \overline{C_E})}$
	Correlation coefficient	$r = \left[\frac{\sum_{i=1}^n (C_{O,i} - \overline{C_O})(C_{E,i} - \overline{C_E})}{\sqrt{\sigma_O \sigma_E}} \right]$
Performance	Root mean square error	$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (C_{O,i} - C_{E,i})^2}$
	Normalised mean square error	$NMSE = \frac{(\overline{C_O} - \overline{C_E})^2}{\overline{C_O} \overline{C_E}}$
	Fractional variance	$FV = 2 \frac{\sigma_O - \sigma_E}{\sigma_O + \sigma_E}$

Table 4. Training and external validation performance indices of the various developed models at the Castro Urdiales site

Pollutant	Model	Subset ^a	EU Uncertainty		Mean Concentration ^b			Performance			
			RME (%)	RDE (%)	C _O 10 ²	C _E 10 ²	FB 10 ²	r	RMSE 10 ²	NMSE 10	FV 10
Pb	PLSR	T	26.7	0.09	2.79	2.79	-9.8 10 ⁻¹⁵	0.704	1.77	4.04	3.48
		V	49.6	0.71	4.13	3.01	31.4	0.620	2.71	5.90	7.44
	ANN	T	34.5	0.66	2.67	2.70	-0.9	0.820	1.44	2.88	1.98
		V	32.0	1.76	4.34	3.48	22.0	0.676	2.41	3.85	2.47
	PCA-ANN	T	36.9	0.41	3.14	3.11	1.0	0.681	2.03	4.25	4.41
		V	30.6	0.07	3.08	3.43	-10.8	0.269	2.60	6.41	3.60
As	PLSR	T	42.8	0.30	6.66	6.66	-4.2 10 ⁻¹²	0.656	5.17	6.02	4.16
		V	34.8	1.74	5.32	5.50	-3.4	0.629	4.63	7.34	0.88
	ANN	T	77.0	0.17	6.96	6.81	2.1	0.130	6.96	10.22	12.74
		V	66.3	0.25	5.32	6.81	-24.6	0.193	5.70	8.99	11.26
	PCA-ANN	T	54.6	0.86	6.96	6.35	9.1	0.536	5.87	7.80	7.36
		V	33.5	0.96	5.55	6.33	-13.3	0.190	5.72	9.30	2.77
Ni	PLSR	T	34.7	10.83	27.61	27.61	3.1 10 ⁻¹³	0.642	21.52	6.07	4.36
		V	22.1	2.64	18.91	22.67	-18.1	0.663	12.27	3.51	-0.97
	ANN	T	45.0	6.19	28.27	23.30	19.3	0.676	21.59	7.08	6.18
		V	32.6	1.24	19.36	23.71	-20.0	0.387	16.15	5.67	-0.58
	PCA-ANN	T	33.9	1.77	24.54	26.01	-5.8	0.643	17.60	4.85	4.92
		V	25.6	2.06	23.64	23.42	0.9	0.216	21.19	8.11	1.64
Cd	PLSR	T	40.9	0.14	3.75	3.75	-1.1 10 ⁻¹²	0.672	3.36	8.03	3.92
		V	46.2	0.86	4.55	3.52	25.6	0.628	3.42	7.30	4.69
	ANN	T	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c
		V	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c	n.c. ^c
	PCA-ANN	T	33.5	0.47	3.84	3.88	-1.1	0.613	3.11	6.49	5.36
		V	41.5	0.42	3.56	3.62	-1.7	0.534	3.11	7.48	5.05

^aT: Training; V: External validation^bO: Observed; E: Estimated^c Not calculated (n.c.)

Table 5. Training and external validation performance indices of the various developed models at the Reinososa site

Pollutant	Model	Subset ^a	EU Uncertainty		Mean Concentration ^b			Performance			
			RME (%)	RDE (%)	C _O 10 ²	C _E 10 ²	FB 10 ²	r	RMSE 10 ²	NMSE 10	FV 10
Pb	PLSR	T	31.1	0.61	5.60	5.60	-6.2 10 ⁻¹⁴	0.723	3.48	3.86	3.21
		V	14.0	0.21	5.11	6.01	-16.2	0.553	3.76	4.60	0.17
	ANN	T	30.3	0.83	6.01	5.71	5.2	0.829	2.86	2.38	3.38
		V	35.9	1.07	5.11	5.95	-15.2	0.563	3.58	4.22	1.05
	PCA-ANN	T	42.3	1.48	5.50	6.02	-8.9	0.679	3.77	4.29	7.17
		V	42.2	1.38	4.96	5.68	-13.4	0.374	3.80	5.13	9.70
As	PLSR	T	28.2	1.25	13.61	13.61	4.1 10 ⁻¹⁴	0.446	8.68	4.07	7.67
		V	31.7	5.66	15.12	12.46	19.3	0.441	12.66	8.51	7.23
	ANN	T	23.6	3.56	14.47	13.39	7.7	0.765	6.75	2.35	4.89
		V	35.4	6.01	15.12	11.82	24.5	0.393	13.17	9.71	6.78
	PCA-ANN	T	25.0	1.41	16.03	15.61	2.7	0.572	10.03	4.02	6.69
		V	37.4	8.26	11.52	16.96	-38.3	0.132	10.09	5.21	2.75
Ni	PLSR	T	53.4	5.45	30.61	30.61	-3.9 10 ⁻⁶	0.386	20.49	4.48	8.86
		V	25.2	2.51	21.54	32.26	-39.9	0.549	19.23	5.32	6.13
	ANN	T	38.9	22.21	33.60	34.60	-2.9	0.460	20.92	3.78	8.54
		V	28.2	2.51	21.54	36.09	-50.5	0.455	22.71	6.63	4.10
	PCA-ANN	T	26.9	9.01	30.81	28.36	8.3	0.677	17.43	3.48	3.09
		V	42.5	20.85	25.95	20.94	21.4	0.304	22.50	9.32	5.53
Cd	PLSR	T	48.5	0.46	3.06	3.06	-3.8 10 ⁻¹³	0.644	3.30	1.16	4.32
		V	46.2	0.59	2.09	3.56	-51.8	0.338	2.74	1.01	3.26
	ANN	T	59.7	0.19	3.40	2.47	31.5	0.641	3.74	1.67	7.03
		V	36.7	0.26	2.09	2.07	1.2	0.521	1.94	8.70	4.30
	PCA-ANN	T	67.5	7.02	3.21	3.56	-10.3	0.518	4.10	1.47	8.42
		V	34.3	0.40	1.83	3.32	-57.7	0.579	2.48	1.01	2.74

^aT: Training; V: External validation^bO: Observed; E: Estimated

Arabian Journal of Geosciences

Estimating airborne heavy metal concentrations in Dunkerque (Northern France) --Manuscript Draft--

Manuscript Number:	AJGS-D-15-01113R1	
Full Title:	Estimating airborne heavy metal concentrations in Dunkerque (Northern France)	
Article Type:	Dust Special Issue	
Corresponding Author:	Germán Santos SPAIN	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Germán Santos	
First Author Secondary Information:		
Order of Authors:	Germán Santos Ignacio Fernández-Olmo Ángel Irabien Frédéric Ledoux Dominique Courcot	
Order of Authors Secondary Information:		
Funding Information:	Ministerio de Economía y Competitividad (ES) (CTM2010-16068)	Not applicable
Abstract:	<p>This work aims to estimate the levels of lead (Pb), nickel (Ni), manganese (Mn), vanadium (V) and chromium (Cr) corresponding to a three-month PM10 sampling campaign conducted in 2008 in the city of Dunkerque (Northern France) by means of statistical models based on Partial Least Squares Regression (PLSR), Artificial Neural Networks (ANN) and Principal Component Analysis (PCA) coupled with ANN. According to the European Air Quality Directives, because the levels of these pollutants are sufficiently below the European Union (EU) limit/target values and other air quality guidelines, they may be used for air quality assessment purposes as an alternative to experimental measurements. An external validation of the models has been conducted, and the results indicate that PLSR and ANNs, with comparable performance, provide adequate mean concentration estimations for Pb, Ni, Mn and V, fulfilling the EU uncertainty requirements for objective estimation techniques, although ANNs seem to present better generalization ability. However, in accordance with the European regulation, both techniques can be considered acceptable air quality assessment tools for heavy metals in the studied area. Furthermore, the application of factor analysis prior to ANNs did not yield any improvements in the performance of the ANNs.</p>	
Response to Reviewers:	See attachment	

1 Estimating airborne heavy metal concentrations in Dunkerque (Northern France)

2
3
4 Germán Santos^{1*}, Ignacio Fernández-Olmo¹, Ángel Irabien¹, Frédéric Ledoux², Dominique Courcot²

5
6
7
8 ¹Departamento de Ingenierías Química y Biomolecular, Universidad de Cantabria, Santander, 39005, Spain

9
10
11
12 ²Unité de Chimie Environnementale et Interactions sur le Vivant (UCEIV) EA 4492, Université du Littoral
13 Côte d'Opale, Dunkerque, 59140, France

14
15
16
17
18 *Corresponding author. Tel.: +34 942201579; fax: +34 942201591

19
20 E-mail address: santosg@unican.es

21 Abstract

22
23
24 This work aims to estimate the levels of lead (Pb), nickel (Ni), manganese (Mn), vanadium (V) and
25 chromium (Cr) corresponding to a three-month PM₁₀ sampling campaign conducted in 2008 in the city of
26 Dunkerque (Northern France) by means of statistical models based on Partial Least Squares Regression
27 (PLSR), Artificial Neural Networks (ANN) and Principal Component Analysis (PCA) coupled with ANN.
28 According to the European Air Quality Directives, because the levels of these pollutants are sufficiently
29 below the European Union (EU) limit/target values and other air quality guidelines, they may be used for
30 air quality assessment purposes as an alternative to experimental measurements. An external validation of
31 the models has been conducted, and the results indicate that PLSR and ANNs, with comparable
32 performance, provide adequate mean concentration estimations for Pb, Ni, Mn and V, fulfilling the EU
33 uncertainty requirements for objective estimation techniques, although ANNs seem to present better
34 generalization ability. However, in accordance with the European regulation, both techniques can be
35 considered acceptable air quality assessment tools for heavy metals in the studied area. Furthermore, the
36 application of factor analysis prior to ANNs did not yield any improvements in the performance of the
37 ANNs.

38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
Keywords: harbour town; immission levels; PM10; heavy metals; statistical models (PLSR, ANN)

30 1. Introduction

31 Recent studies have shown a positive correlation between high concentrations of particles and public health
32 deterioration. Particulate Matter (PM) remains a concerning environmental problem in urban areas due to
33 its physical properties, such as mass distribution, particle size and shape, and chemical composition, which
34 may include various acidic and toxic species such as metals, metalloids and aromatic compounds (Karar
35 and Gupta 2006). In addition to industrial emissions, non-exhaust PM emissions from road traffic have
36 been identified as an important source of metals in urban environments (Thorpe and Harrison 2008).
37 Furthermore, long-term exposure to metals could cause severe toxic effects on human health (Chen and
38 Lippmann 2009).

39 In this context, the European Union, through the Air Quality Framework Directive (EC 2008) and the 4th
40 Daughter Directive (EC 2004), has established a set of air quality objectives for certain pollutants in PM₁₀:
41 a limit value of 500 ng m⁻³ for Pb (Directive 2008/50/EC) and target values of 6 ng m⁻³ for As, 20 ng m⁻³
42 for Ni and 5 ng m⁻³ for Cd (Directive 2004/107/EC) for the total content in the PM₁₀ fraction averaged over
43 a calendar year. Along with these limit/target values, an upper and lower assessment threshold (hereafter
44 known as UAT and LAT) are also specified, expressed as a percentage of the corresponding limit/target
45 value as follows: 70 and 50 % (Pb and Ni) and 60 and 40 % (As and Cd). Depending on the level of
46 pollutants with respect to these thresholds, different air quality assessment methods with respect to the
47 pollutants are permitted. Thus, in accordance with Directive 2008/50/EC, when the pollutant levels are
48 below the lower assessment threshold (LAT), the air quality may be assessed using solely modelling or
49 objective estimation techniques without the need for experimental measurements. Taking into account the
50 high cost and time consumption associated with analytical determination of the levels of these pollutants,
51 it may be interesting to try to find new alternatives for air quality assessment so that fewer experimental
52 measurements may be required.

53 According to the Guidance on Assessment under the EU Air Quality Directives, “objective estimation
54 technique” is a fairly broad term that includes mathematical methods to calculate concentrations from
55 values measured at other locations and/or times based on scientific knowledge of the concentration
56 distribution. Empirical data-based modelling or statistical modelling falls within this definition and
57 represents an attractive alternative to deterministic modelling (air dispersion modelling), given that it
58 requires less specific knowledge of the system under consideration as it attempts to find the existing
59 relationship between the immission concentrations of pollutants and other variables that may influence the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

processes that control the formation, transportation and removal of aerosols in the atmosphere, disregarding the physical principles in which the equations that describe these processes are based on, as well as other decisive information such as emission inventories.

Partial Least Squares Regression and Artificial Neural Networks have been proposed in this study to estimate PM₁₀-bound heavy metals because both have been used in the literature as mathematical techniques to forecast the air concentration of a number of pollutants. Pires et al. (2008), Polat and Durduran (2012) and Singh et al. (2012) applied PLSR to predict PM concentrations, and numerous authors over the years have investigated the development of ANN models to predict PM concentrations and gaseous pollutants (Gardner and Dorling 1999; Kukkonen et al. 2003; Perez and Reyes 2002), to cite but a few. Furthermore, Chelani et al. (2002) performed not only a prediction of PM₁₀ concentration but also of ambient air metal levels, namely Cd, Cr, Fe, Ni, Pb and Zn, with a low prediction error. Moreover, because the number of independent input variables is relatively high with respect to the number of samples, an alternative approach based on applying Principal Component Analysis (PCA) prior to ANN development was considered due to this technique being reported in the literature as an effective strategy to improve model performance (Lu et al. 2003, Sousa et al. 2007, UI-Saufie et al. 2013).

Despite having a relatively small contribution to the total content of PM in terms of mass, metals in Dunkerque have been reported as clear tracers of the local industrial activities in the city (Kfoury 2013). In this respect, the main objective of this work is to estimate the levels of some EU-regulated and non-regulated metals in airborne PM₁₀ in the urban area of Dunkerque. For this purpose, statistical models based on PLSR and ANNs have been developed as objective estimation techniques.

It is worth mentioning that because this work is devised as an air quality assessment tool at a later stage, it is about estimation instead of forecasting. Thus, it is intended to provide an estimation of the pollutant concentrations of the recent past as an alternative to experimental measurements instead of predicting future pollutant concentrations.

2. Description of the methodology and area of investigation

2.1 Partial least squares regression fundamentals

Partial least squares regression is a statistical method that, as with other multivariate regression techniques, seeks to find the relationship between two data matrices in order to predict a response or a set of response variables (Y) from a set of predictors (X). However, it differs from other multivariate calibration techniques

90 in that it aims to reach two goals simultaneously as follows: to capture variance and to achieve correlations,
91 i.e., maximize covariance (Abdi 2010). That is to say, PLSR attempts to find factors that maximize the
92 amount of variation explained in X that is relevant for predicting Y as a generalization of other related
93 techniques, e.g., principal component regression (PCR), which obtains factors based solely on the amount
94 of variance captured in X and disregards entirely the covariance, and multiple linear regression (MLR),
95 which tries to find a single factor that best correlates predictors with responses.

96 By performing a projection of the original predictor variables into a new space, PLSR creates a set of
97 orthogonal factors, referred to as *latent variables*, to be used to predict the output variable(s). This
98 projection is performed as follows: first, the X -matrix is decomposed as a product of a set of X -scores T
99 multiplied by a set of X -loadings P .

$$100 \quad X = TP' + E \quad (1)$$

101 X -scores are expressed as a linear combination of the original predictor variables by means of a set of
102 vectors of coefficients known as *weights*, which ensure the orthogonality of scores.

$$103 \quad T = XW^* \quad (2)$$

104 where

$$105 \quad W^* = W(P'W)^{-1} \quad (3)$$

106 In parallel, a similar decomposition is performed for the Y -matrix, which is expressed as a product of the
107 Y -scores U multiplied by the Y -loadings C .

$$108 \quad Y = UC' + G \quad (4)$$

109 As mentioned before, X -scores not only model X (Eq. (1)) but also predict Y . This prediction is achieved
110 using Eq. (5).

$$111 \quad Y = TC' + F = XW^*C' + F = XB + F \quad (5)$$

112 Therefore

$$113 \quad B = W^*C' \quad (6)$$

114 Further details of this technique can be found in Wold (2001). PLS Toolbox (Eigenvector Research, Inc.)
115 for MATLAB was used in the present study to develop the PLSR models.

116

117 2.2 Artificial neural network fundamentals

118 Artificial neural networks are computational systems based on biological nervous systems that attempt to
119 mimic the fault-tolerance and capacity to learn of biological neural systems. They are formed by a number

120 of highly interconnected simple processing elements, or artificial neurons (also known as nodes or units),
121 receiving a set of inputs, either from original data or from the output of other neurons in the neural network,
122 via weighted connections (or weights) that resemble synaptic connections in a biological neuron. These
123 nodes are arranged into three types of layers, i.e., input, hidden and output layers. Data are fed into the
124 nodes in the input layer and later transferred to the subsequent layers. Every node in the hidden and output
125 layers also has a single bias value known as the activation threshold value. Being the weighted sum of the
126 inputs computed, the corresponding threshold value is subtracted to compose the activation of the neuron.
127 The activation signal is passed through an activation function (also known as a transfer function) to produce
128 the output of the node. The relationship between the output and the inputs finally has the mathematical
129 representation, as presented in Eq. (7):

$$130 \quad y_t = w_0 + \sum_{j=1}^q w_j \cdot g(w_{0,j} + \sum_{i=1}^p w_{i,j} x_{t,i}) + \varepsilon_t \quad (7)$$

131 where p is the number of input nodes, q is the number of hidden nodes, $w_{i,j}$ ($i = 0,1,2, \dots, p, j = 1,2, \dots, q$)
132 and w_j ($j = 0,1,2, \dots, q$) are connection weights, and ε_t is a bias error.

133 A multitude of neural network architectures are possible. However, in practice, simple network structures
134 with a relatively small number of hidden nodes often work well in out-of-sample forecasting. In this work,
135 a multilayer perceptron (MLP) neural network with a sigmoid hidden transfer function and a linear output
136 transfer function has been selected, applying the Levenberg-Marquardt learning algorithm. A schematic
137 representation of the network structure is shown in Fig. 1.

138 The ANN models in this study were developed using the Neural Network Toolbox for MATLAB
139 (MathWorks, Inc.).

140

141 *2.3 Description of the area of study and sampling site*

142 The city of Dunkerque, with a population of approximately 68,000 inhabitants in 2008, is located on the
143 northern coastline of France, limited by the French-Belgian border. The main urban area is surrounded in
144 its northern part by the harbour of Dunkerque, which is classified as the third most important port in France
145 due to shipping and freight transport (including ore, coal and copper, among other goods) and as the seventh
146 port in order of importance of Northern Europe. The city is also in close proximity to the English Channel,
147 connecting the North Sea with the Atlantic Ocean, which is the world's busiest seaway, with approximately
148 500 vessels transiting daily. There is also a highly industrialized area in the city's vicinity for the metallurgic

149 industry, as it has an integrated steel manufacturing plant (nearly 4 km NW), an electric steel plant (6 km
150 NE) and a ferromanganese alloy production plant (at approximately 6 km W).

151 A total of 78 samples were measured throughout an *intensive* PM₁₀ sampling campaign performed from
152 February to May 2008 in Dunkerque by Hleis (2010). Fig. 2 shows the location of the sampling site
153 (51°02'07''N, 02°22'05''E and approximately 10 m above sea level), which was placed on the rooftop of
154 the Les Darses site (to prevent the sampling of punctual events at street / ground level) on the boundary line
155 between the industrial area and the city so that the effects of both urban and industrial emissions were
156 registered during sampling (under WSW and NNW wind sectors) (Kfoury 2013). Further details of the
157 sampling procedure are described in Hleis (2010). The composition of inorganic elements (Al, Ca, Cr, Cu,
158 Fe, K, Mg, Mn, Na, Ni, Pb, Sn, Ti, V and Zn) and ions (Cl⁻, NO₃⁻, SO₄²⁻ and NH₄⁺) in the particles was
159 determined. The mean values of these constituents are reported in Hleis (2010).

160

161 *2.4. Modelling database and pre-treatment*

162 As usual for this type of modelling, input variables consist of (i) meteorological data, namely average
163 temperature (°C), average relative humidity (%), prevailing wind direction (°), prevailing wind speed (ms⁻¹),
164 average pressure (mbar) and cumulative precipitation (L m⁻²), which are obtained at the meteorological
165 station in the harbour of Dunkerque, and (ii) major pollutant data, which are composed by average
166 concentrations (µg m⁻³) of sulphur dioxide (SO₂), tropospheric ozone (O₃) and nitrogen oxides (NO_x)
167 measured at the St. Pol sur Mer air quality monitoring station (the Atmo-Nord-Pas-de-Calais air quality
168 network) and PM₁₀ concentrations measured at the Les Darses site. Additionally, two nominal variables
169 were considered to account for the seasonal (1: Winter, 2: Spring, 3: Summer, 4: Fall) and weekend effects
170 (0: Working day, 1: Weekend).

171 Output variables in this study consisted of PM₁₀-bound Pb, Ni, Mn, V and Cr levels in ambient air (ng m⁻³)
172 at the sampling site. Among the EU regulated metals, Pb and Ni were determined. Additionally, three non-
173 regulated metals were also considered: Mn, V and Cr. These metals were tracers of various industrial
174 activities found in Dunkerque, where previous studies on trace metal levels have been developed: Mn, for
175 ferromanganese alloys manufacturing; V, for marine traffic and liquid fuel combustion; and Cr, for non-
176 integrated steel manufacturing and coal combustion (Kfoury 2013). Because these metals are not regulated
177 by the EU, they do not have a limit/target. Therefore, to normalize the metal concentration and calculate
178 the EU uncertainty indices, the following values were considered as equivalent to the LAT for non-

179 regulated metals: the annual air quality guideline for Mn (150 ng m⁻³) proposed by the World Health
180 Organization (WHO 2000) and the maximum observed concentration for V and Cr in the absence of a
181 standard value for a period of duration comparable with that of the period of study.

182 As shown in Fig. 3, the Pb and Ni mean values are below their respective LAT. Therefore, according to the
183 EU Air Quality Directives, objective estimation techniques can be applied for the air quality assessment in
184 relation to Pb and Ni.

185 A pre-treatment procedure for outlier identification and removal based on the statistical parameter of the
186 Mahalanobis distance was conducted. Additionally, as usual for this type of technique, the complete
187 database was divided into three subsets as a result of applying a data-splitting procedure, the Kennard-Stone
188 algorithm, which selects the more representative samples for each subset based on Euclidean distances.
189 Thus, 60 % of the total number of samples was used for model development, 20 % for verification to avoid
190 over-fitting and 20 % for external validation. Furthermore, to avoid scale effects, the dependent variables
191 were normalized by dividing the metal concentrations by their respective LAT.

192

193 *2.5 Model performance criteria*

194 In this study, the evaluation criteria to determine whether a model is suitable for air quality assessment
195 purposes is principally based on the following: (i) the fulfilment of the European Union uncertainty
196 requirements for objective estimation techniques, and (ii) the accuracy of estimated mean values because
197 the metal limit/target values correspond to annual mean concentrations. Two indices of uncertainty were
198 calculated: the relative maximum error without timing (RME) and the relative directive error (RDE). The
199 former is the largest concentration difference of all percentile (p) differences normalized by the respective
200 measured value (Borrego et al. 2008), as calculated by Eq. (8). The latter is the difference between the
201 closest observed concentration to the limit/target value and the correspondingly ranked modelled
202 concentration normalized by the limit/target value (Denby et al. 2010), as given by Eq. (9).

$$203 \text{ RME} = \max(|C_{O,p} - C_{E,p}|) / C_{O,p} \quad (8)$$

$$204 \text{ RDE} = |C_{O,LV} - C_{E,LV}| / LV \quad (9)$$

205 Additionally, a number of statistical parameters were considered to evaluate the model performance. These
206 performance indicators are the fractional bias (FB), the correlation coefficient (r), the root mean squared
207 error (RMSE) and the fractional variance (FV), as given by Eqs. (10-13):

$$208 r = \left[\frac{\sum_{i=1}^n (C_{O,i} - \bar{C}_O)(C_{E,i} - \bar{C}_E)}{\sqrt{\sigma_O \sigma_E}} \right] \quad (10)$$

$$209 \quad FB = \frac{\overline{C_O} - \overline{C_E}}{0.5(\overline{C_O} + \overline{C_E})} \quad (11)$$

$$210 \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (C_{O,i} - C_{E,i})^2} \quad (12)$$

$$211 \quad FV = 2 \frac{\sigma_O - \sigma_E}{\sigma_O + \sigma_E} \quad (13)$$

212 where n = the total number of observations, $C_{o,i}$ = the i th observed value, $C_{E,i}$ = the i th estimated value and
 213 $\overline{C_O}$ and $\overline{C_E}$ are the observation and estimation averages, respectively. These indicators were calculated in
 214 both development and validation steps.

215

216 **3. Results and discussion**

217 *3.1. Levels of the studied metals*

218 Fig. 3 shows that the Pb and Ni mean concentrations for the period of study are below the corresponding
 219 legislated objective/limit values for ambient air. Moreover, the Mn average concentration is also below the
 220 WHO air quality guideline in relation to manganese. Nevertheless, there are some particular cases, i.e., Ni
 221 and Mn, where the non-averaged concentrations (individual sample concentrations) of these pollutants
 222 amply exceed the corresponding objective/limit values as follows: as shown, the Ni and Mn maximum
 223 observed concentrations exceed by 10 and 6 times their LAT and LAT-equivalent values, respectively.
 224 Special attention should be paid to model performance in this sense because exposures to high levels of
 225 these metals may have detrimental effects on human health. It has been demonstrated that inhaled
 226 manganese produces neurotoxic effects that vary from neuropsychological and motor functions (Mergler
 227 et al. 1999), postural stability (Hernández-Bonilla et al. 2011) and increased risk of Parkinson's disease
 228 (Finkelsteinn and Jerret 2007) at lower concentration exposures (near 50 ng m⁻³) to a movement disorder
 229 known as Manganism at concentrations above 1 mg m⁻³ (Aschner et al. 2005). Regarding vanadium, its
 230 toxic effects depend on its degree of oxidation and may include irritation of the respiratory tract,
 231 haematological and biochemical changes and functional lesions in certain organs (Sumanta et al. 2015).
 232 The studies conducted by Hleis (2010) and Kfoury (2013) have shown that the levels of Pb, Ni, Mn, V, Cr
 233 and other metals and metalloids in Dunkerque are remarkably associated with industry as they have been
 234 reported to be tracers of local industrial activities. The results of pollution roses and receptor modelling for
 235 source apportionment by means of non-negative matrix factorization indicate that Pb emissions may be
 236 mainly attributed to integrated steelworks, which is an Mn emission source as well (Kfoury 2013; Hleis
 237 2010). However, the ferromanganese manufacturing plant emissions also influence the levels of Pb and
 238 certainly the levels of Mn (Hleis 2010). Ni and V are tracers of heavy oil combustion because they explain

239 72 % and 86 % of the observed concentrations, respectively (Kfoury 2013). With regard to Cr, it is
240 considered to be a marker of the activity of the electric steel plant, although Cr levels may also be partly
241 due to oil combustion. The strong presence of industrial activities in Dunkerque and the firm connection
242 between ambient air metal levels and the local industry makes the city a suitable location to develop
243 objective estimation models for metals because the inputs to these models partly consist of major pollutant
244 concentrations, which are undoubtedly related to industry as well.

245

246 *3.2. Estimation of Pb and Ni*

247 Table 1 presents the results obtained with the best developed models for the two considered EU-regulated
248 metals (Pb and Ni) using the three different considered approaches as follows: PLSR, ANNs and PCA-
249 ANN. Results related to training (T) and external validation (V) subsets are presented for each model.

250 In the first place, for the evaluation of these models as an air quality assessment tool from a regulatory point
251 of view, which is the main goal of the present study, the mean value estimation and the conformity of the
252 compliance with the uncertainty requirements are two key aspects to take into account. In this regard, while
253 complying with the uncertainty requirements (expressed in this study in terms of the RME and RDE indices)
254 because both uncertainty indices are well below 100 %, which is the maximum uncertainty percentage
255 allowed by the EU for objective estimation techniques to be used for air quality assessment, every model
256 is able to provide a good estimate of the mean concentration due to the lower values obtained for the FB
257 index, which is an indicative measure of the accuracy in estimating mean values. For the training stage, the
258 PLSR FB index values are lower than the FB values with ANNs and the PCA-ANN model. The reason why
259 the FB index of PLSR models is so small is that the PLSR-estimated and observed mean values are nearly
260 equal, resulting in an FB index very close to its ideal value, which, according to Eq. (11), is 0. However, if
261 attention is to be paid to the values of the rest of the FB indices, it is evident that they are not significantly
262 higher as the estimated mean values are close to the corresponding observed values in every case.

263 It is worth noting that the mean values in this work are not in fact annual mean concentrations because the
264 available data samples belonged to a period of study limited to three months, from mid-February to mid-
265 May 2008. The sampling period varied from 6 to 14 hours, and consequently, the levels of pollutants
266 presented significant variability. Therefore, there was an additional difficulty for the estimation.

267 Although a correct estimation of the mean value while fulfilling the EU uncertainty requirements is
268 sufficient for a model application in the frame of the EU Air Quality Directive, it would be greatly

269 preferable for the model to also be able to correctly estimate the individual sample concentrations. To
270 evaluate this capacity and provide a more comprehensive response of model performances, a series of
271 additional statistical indicators have been addressed. With regard to these performance indices, the
272 correlation coefficient values of the PLSR and ANN models are within the range of 0.5-0.9, indicating a
273 good tendency of the estimated and observed values to vary together. Nevertheless, even if the r values are
274 close to 1, there is no guarantee that the estimated and observed values match each other, only that they
275 may differ by a consistent factor. For this reason, other statistics must be taken into consideration.
276 As for PCA-ANN models, they provide lower values of the correlation coefficient - within the range of 0.3-
277 0.6- than PLSR and ANNs. This fact, together with an increase in uncertainty indices, indicates that, for
278 this specific application, performing PCA prior to the development of ANNs is not an effective alternative.
279 Models have been evaluated on the basis of comparisons against observations via a set of statistical
280 indicators, which, while providing insight on general model performance, do not necessarily indicate
281 whether model results have reached a sufficient quality level for a given application, e.g., for policy support.
282 Ideally, models should have a correlation coefficient close to 1 and FB, RMSE and FV values close to 0.
283 Unfortunately, in practice, due to the uncertainty of observation and the analytical determination in the
284 laboratory, these values will rarely be achieved. In this regard, Kumar et al. (1993) propose values for some
285 of these parameters associated with a minimum quality for the models: $NMSE \leq 0.5$ and $-0.5 \leq FB \leq 0.5$.
286 A supplementary manner to evaluate model performance is by means of a graphic representation. In this
287 regard, Fig. 4 depicts the estimated normalized concentrations of Pb and Ni obtained with the PLSR and
288 ANN models for the training and external validation subsets. As observed, both models exhibit difficulties
289 in accurately estimating the individual sample concentrations, leading to an underestimation of the highest
290 concentrations. Notwithstanding, PLSR and ANNs capture the underlying trend during training, although
291 there is a slightly better fitting when using ANNs, as reflected in the lower RMSE and FV values obtained
292 with ANNs with respect to those obtained with the PLSR model.
293 With respect to the external validation subset, as a result of a decrease in the accuracy of the estimations,
294 there is a general slight decrease in the correlation coefficient values and an increase in the values of the
295 RMSE and FV indices of every model compared with those obtained for the training subset. However, the
296 FB index values for PLSR and ANNs are below 0.5 and, therefore, within the acceptable range for FB for
297 an air quality model suggested by Kumar et al. (1993). Additionally, for the PLSR and ANN models, the
298 correlation coefficient of the external validation subset ranged from 0.5-0.8, which was similar to those

299 obtained for the training subset. Consequently, based on the performance results obtained for external
300 validation, PLSR and ANNs may be considered proper approaches to estimate ambient Pb and Ni levels in
301 the studied site. Nevertheless, as a general remark, the best pair of training and external validation statistics
302 are found when using ANNs, which indicates that the model is able to not only fit the data correctly but
303 also provide good estimates of data not used for the development of the models, which implies that ANNs
304 present better generalization ability than the other studied techniques.

305

306 *3.3. Estimation of Mn, V and Cr*

307 The results of the best developed models for Mn, V and Cr are presented in Table 2. Despite the fact that
308 the ambient air levels of these pollutants are not regulated by the European Directives, the model evaluation
309 analysis is performed in the same manner as in the case of Pb and Ni. Nevertheless, because these pollutants
310 are lacking a policy limit/target value, an RDE-equivalent (RDE_{eq}) uncertainty index has been calculated
311 based on the version equivalent to LAT values for regulated pollutants, as mentioned in section 2.4. Note
312 that, with this assumption, Mn, V and Cr mean values, unlike those of Pb and Ni, are closer to their
313 corresponding LAT.

314 As can be observed, the EU uncertainty requirements for objective estimation techniques are fulfilled with
315 an RME and an RDE_{eq} lower than 100 % in all cases, except for the external validation RME index of the
316 Cr PCA-ANN model. However, it could be argued that because some of these metals may present
317 considerably higher air concentrations, such as Mn, which exceeds by almost six times 150 ng m^{-3} (which
318 is the WHO air quality guideline used as LAT-equivalent) with a maximum observed value of 872.8 ng m^{-3}
319 for the period of study (Fig. 3), more restrictive uncertainty requirements should be addressed, considering
320 that an allowed 100 % uncertainty in the estimation may lead to erroneously regarding as acceptable an
321 underestimation of a potentially dangerous pollutant level.

322 With respect to the mean concentration, lower values of the FB index for the Mn, V and Cr PLSR and ANN
323 models indicate acceptable training and external validation estimations. These values are within the same
324 order of magnitude as those obtained for Pb and Ni and below 0.5, complying with the minimum quality
325 requirements proposed by Kumar et al. (1993). Again, that is not applicable to Cr PCA-ANN models with
326 an external validation FB index of -1.46.

327 As for the models' performance in relation to the estimation of individual sample concentrations,
328 correlation coefficient values lower than 0.66 for external validation indicate an unsatisfactory fitting.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

329 However, this measure can be dominated by a small proportion of extreme values that may not reflect the
330 behaviour of the bulk of the data. At any rate, the ANN model correlation coefficients are greater than those
331 obtained for the PLSR and PCA-ANN models. This, together with the fact that ANNs provide the lowest
332 FB index and adequate RME and RDE_{eq} , points to ANNs as the most suitable approach of the three studied.

333

334 **4. Conclusions**

335 In this work, statistical models based on PLSR and ANNs to estimate the levels of the considered EU-
336 regulated metals, i.e., Pb and Ni, have been developed and externally validated. Based on the results
337 obtained and according to the European Air Quality Framework Directive, these models can be taken into
338 consideration as valid approaches to be used as objective estimation techniques for air quality assessment
339 in relation to metals in the area of study because they are able to correctly estimate mean values within an
340 uncertainty range up to 100 %. Both linear (PLSR) and non-linear (ANNs) statistical models show a
341 comparable performance, although the latter exhibit an enhanced generalization ability. However, ANN
342 performance experienced no improvements by the application of factor analysis techniques, such as PCA,
343 before model development.

344 Additionally, in this study, some metals that lack a limit/target value in European legislation, namely Mn,
345 V and Cr, have also been considered due to the strong relationship that exists between their levels and the
346 local industry of the study area and due to the scientific evidence that suggests that some of these non-
347 regulated metals can also cause damage to human health. As with Pb and Ni, the PLSR and ANN models
348 for Mn and V work relatively well in terms of mean estimation within the EU Directive uncertainty limits.
349 Nevertheless, they are not able to properly describe variations of Cr.

350 Finally, the statistical models developed for every metal struggle with the estimation of the individual
351 sample concentrations and, as with many deterministic models, tend towards a slight underestimation.
352 Therefore, further work will focus on deepening knowledge regarding the interactions between the different
353 inputs and their relationship with the outputs to improve this specific matter.

354

355 **Acknowledgements**

356 This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through
357 the Projects CTM2010-16068/CTM2013-43904R and the FPI short stay EEBB-I-13-07691. Germán Santos
358 would also like to thank the Unité de Chimie Environnementale et Interactions sur le Vivant (UCEIV) at

359 La Maison de la Recherche en Environnement Industriel for welcoming him as a guest PhD student in their
360 facilities.

361

362 **References**

363 Abdi H (2010) Partial Least Squares regression and Projection on Latent Structure regression (PLS
364 regression). Wiley Interdiscip Rev Comput Stat 2:97-106

365 Aschner M, Erikson KM, Dorman DC (2005) Manganese dosimetry: species differences and implications
366 for neurotoxicity. Crit Rev Toxicol 35(1):1-32

367 Borrego C, Monteiro A, Ferreira J, Miranda AI, Costa AM, Carvalho AC, Lopes M (2008) Procedures for
368 estimation of modelling uncertainty in air quality assessment. Environ Int 34:613-620

369 Chelani AB, Gajghate DG, Hasan MZ (2002) Prediction of ambient PM₁₀ and toxic metals using artificial
370 neural networks. J Air Waste Manag Assoc 52:805-810

371 Chen LC, Lippmann M (2009) Effects of metals within ambient air particulate matter (PM) on human
372 health. Inhal Toxicol 21:1-31

373 Denby B (2010) Guidance on the use of models for the European Air Quality Directive: A working
374 document of the Forum for Air Quality Modelling in Europe (FAIRMODE). ETC/ACC report version
375 6.2

376 EC (2004) Council Directive 2004/107/EC Directive of the European Parliament and of the Council of 15
377 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons
378 in ambient air. The European Parliament and the Council of the European Union. Off J L23 3-16

379 EC (2008) Council Directive 2008/50/EC Directive of the European Parliament and of the Council of 21
380 May 2008 on ambient air quality and cleaner air for Europe. The European Parliament and the Council
381 of the European Union. Off J L152 1-44

382 Finkelstein MM, Jerrett M (2007) A study of the relationships between Parkinson's disease and markers of
383 traffic-derived and environmental manganese air pollution in two Canadian cities. Environ Res
384 104(3):420-432

385 Hernández-Bonilla D, Schilman A, Montes S, Rodríguez-Agudelo Y, Rodríguez-Dozal S, Solís-Vivanco
386 R, et al. (2011) Environmental exposure to manganese and motor function of children in Mexico.
387 Neurotoxicology 32(5):615-621

388 Gardner MW, Dorling SR (1999) Neural network modelling and prediction of hourly NO_x and NO₂
389 concentrations in urban air in London. *Atmos Environ* 33:709-719

390 Hleis D (2010) Evaluation de la contribution d'émissions sidérurgiques à la teneur en particules en
391 suspension dans l'atmosphère à une échelle locale. Dissertation, Université du Littoral Côte d'Opale

392 Karar K, Gupta AK (2006) Seasonal variations and chemical characterization of ambient PM₁₀ at residential
393 and industrial sites of an urban region of Kolkata. *Atmos Res* 81:36-53

394 Kfoury A (2013) Origin and physicochemical behaviour of atmospheric PM_{2.5} in cities located in the littoral
395 area of the Nord-Pas-de-Calais region, France. Dissertation, Université du Littoral Côte d'Opale

396 Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S,
397 Chatterton T, Foxall R, Cawley G (2003) Extensive evaluation of neural network models for the
398 prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modelling system and
399 measurements in central Helsinki. *Atmos Environ* 37:4539-4550

400 Kumar A, Luo J, Bennett G (1993) Statistical evaluation of lower flammability distance (LFD) using four
401 hazardous release models. *Process Saf Prog* 12:1-11

402 Lu WZ, Wang WJ, Wang XK, Xu ZB, Leung AYT (2003) Using improved neural network model to analyse
403 RSP, NO_x and NO₂ levels in urban air in Mong Kok, Hong Kong. *Environ Monit Assess* 87:235-254

404 Mergler D, Baldwin M, Belanger S, Larribe F, Beuter A, Bowler R et al. (1999) Manganese neurotoxicity,
405 a continuum of dysfunction: results from a community based study. *Neurotoxicology* 20(2-3):327-342

406 Perez P, Reyes J (2002) Prediction of maximum of 24-h average of PM₁₀ concentrations 30h in advance in
407 Santiago, Chile. *Atmos Environ* 36:4555-4561

408 Pires JCM, Martins FG, Sousa SIV, Alvim-Ferraz MCM, Pereira MC (2008) Prediction of the daily mean
409 PM₁₀ concentrations using linear models. *Am J Environ Sci* 4:445-453

410 Polat K, Durduran SS (2012) Usage of output-dependent data scaling in modelling and prediction of air
411 pollution daily concentration values (PM₁₀) in the city of Konya. *Neural Comput Appl* 21:2153-2162

412 Singh KP, Gupta S, Kumar A, Shukla SP (2012) Linear and nonlinear modelling approaches for urban air
413 quality prediction. *Sci Total Environ* 426:244-255

414 Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC (2007) Multiple linear regression and artificial
415 neural networks based on principal components to predict ozone concentrations. *Environ Modell Softw*
416 22:97-103

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

417 Sumanta KG, Rumpa S, Bidyut S (2015). Toxicity of inorganic vanadium compounds. Res Chem Intermed
418 41:4873-4897
419 Thorpe A, Harrison RM (2008) Sources and properties of non-exhaust particulate matter from road traffic:
420 a review. Sci Total Environ 400:270-282
421 Ul-Saufie AZ, Yahaya AS, Ramli NA, Rosaida N, Hamid HA (2013) Future daily PM₁₀ concentrations
422 prediction by combining regression models and feedforward backpropagation models with principle
423 component analysis (PCA). Atmos Environ 77:621-630
424 WHO (2000) Air quality guidelines for Europe 2nd edition. World Health Organisation, Geneva
425 Wold S, Sjöström M, Eriksson L (2001) PLS-regression: A basic tool of chemometrics. Chemom Intell Lab
426 Syst 58:109-130

427 **Figure captions**

1
2 428

3
4 429 Figure 1: Structure of the artificial neural network

5
6 430

7
8 431 Figure 2: Sampling site and monitoring and meteorological stations in Dunkerque

9
10 432

11 433 Figure 3: Box-plot of the levels of Pb, Ni, Mn, V and Cr for the period of study. The box extends between

12 434 the upper and lower quartiles with the inner line representing the median value. The whiskers indicate the

13
14
15 435 minimum and maximum values.

16
17
18 436

19
20 437 Figure 4: Comparison between the observed and modelled normalized Pb and Ni concentrations

21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Uncertainty, mean concentration and performance statistics for the best models developed for Pb and Ni

Metal	Model	Subset ^a	EU Uncertainty		Mean Concentration ^b			Performance		
			RME (%)	RDE (%)	C _O 10 ²	C _E 10 ²	FB 10 ²	r	RMSE 10 ²	FV 10
Pb	PLSR	T	28.1	1.44	6.52	6.52	3.7 10 ⁻⁰⁸	0.823	3.94	1.95
		V	31.9	0.31	7.46	8.88	-17.4	0.837	4.48	-2.78
	ANN	T	18.3	2.10	6.38	6.84	-7.0	0.932	2.72	-0.85
		V	54.0	0.54	7.46	8.31	-10.8	0.861	4.90	-4.12
	PCA-ANN	T	40.1	2.22	6.57	6.78	-3.2	0.663	3.90	5.82
		V	90.4	1.38	3.69	7.95	-73.1	0.266	5.64	-2.56
	PLSR	T	65.9	12.87	68.5	68.5	-7.510 ⁻¹¹	0.560	80.5	5.64
		V	83.6	11.70	156.6	98.2	45.8	0.556	241.3	14.62
Ni	ANN	T	29.2	18.55	73.4	73.8	-0.5	0.873	54.2	1.44
		V	50.0	17.60	156.6	115.8	30.0	0.702	186.6	4.77
	PCA-ANN	T	64.9	24.86	95.9	95.8	0.1	0.470	161.1	7.20
		V	42.6	2.50	68.6	94.9	-32.2	0.443	63.2	-3.04

^a T: Training; V: Validation

^b O: Observed; E: Estimated

Table 2. Uncertainty, mean value and performance statistics for the best models developed for Mn, V and Cr

Metal	Model	Subset ^a	EU Uncertainty		Mean Concentration ^b			Performance		
			RME (%)	RDE _{eq} (%)	C _O 10 ²	C _E 10 ²	FB 10 ²	r	RMSE 10 ²	FV 10
Mn	PLSR	T	53.9	4.71	32.55	33.43	-2.7	0.580	41.49	3.57
		V	53.1	50.23	64.95	58.24	10.9	0.184	92.80	6.99
	ANN	T	52.6	60.51	33.61	21.38	44.5	0.704	39.20	3.42
		V	48.2	68.86	64.95	63.77	1.8	0.457	81.56	4.64
	PCA-ANN	T	66.4	78.51	46.05	46.05	0.0	0.463	64.77	7.34
		V	35.3	11.62	29.49	30.33	-2.8	0.431	42.72	1.74
V	PLSR	T	42.7	1.20	13.12	13.42	-2.3	0.694	8.17	2.33
		V	31.5	4.74	18.07	18.48	-2.2	0.590	11.21	3.16
	ANN	T	41.6	4.28	14.27	14.00	1.9	0.806	7.19	1.78
		V	30.7	5.45	18.07	18.45	-2.1	0.663	10.43	2.11
	PCA-ANN	T	42.9	1.60	13.14	13.20	-0.5	0.747	5.91	2.93
		V	12.5	15.60	12.79	16.50	-25.4	0.366	13.97	0.56
Cr	PLSR	T	88.8	1.01	4.58	3.96	14.6	-0.031	15.28	11.34
		V	78.5	25.74	6.97	5.78	18.7	0.077	18.09	12.78
	ANN	T	50.0	39.17	5.59	3.25	53.1	-0.040	19.24	6.53
		V	83.6	27.89	6.97	3.55	65.0	-0.240	19.06	14.76
	PCA-ANN	T	79.2	0.43	7.15	7.90	-10.0	0.331	18.33	9.70
		V	489.9	0.43	1.45	9.31	-146.3	0.275	10.04	-13.57

^a T: Training; V: Validation^b O: Observed; E: Estimated

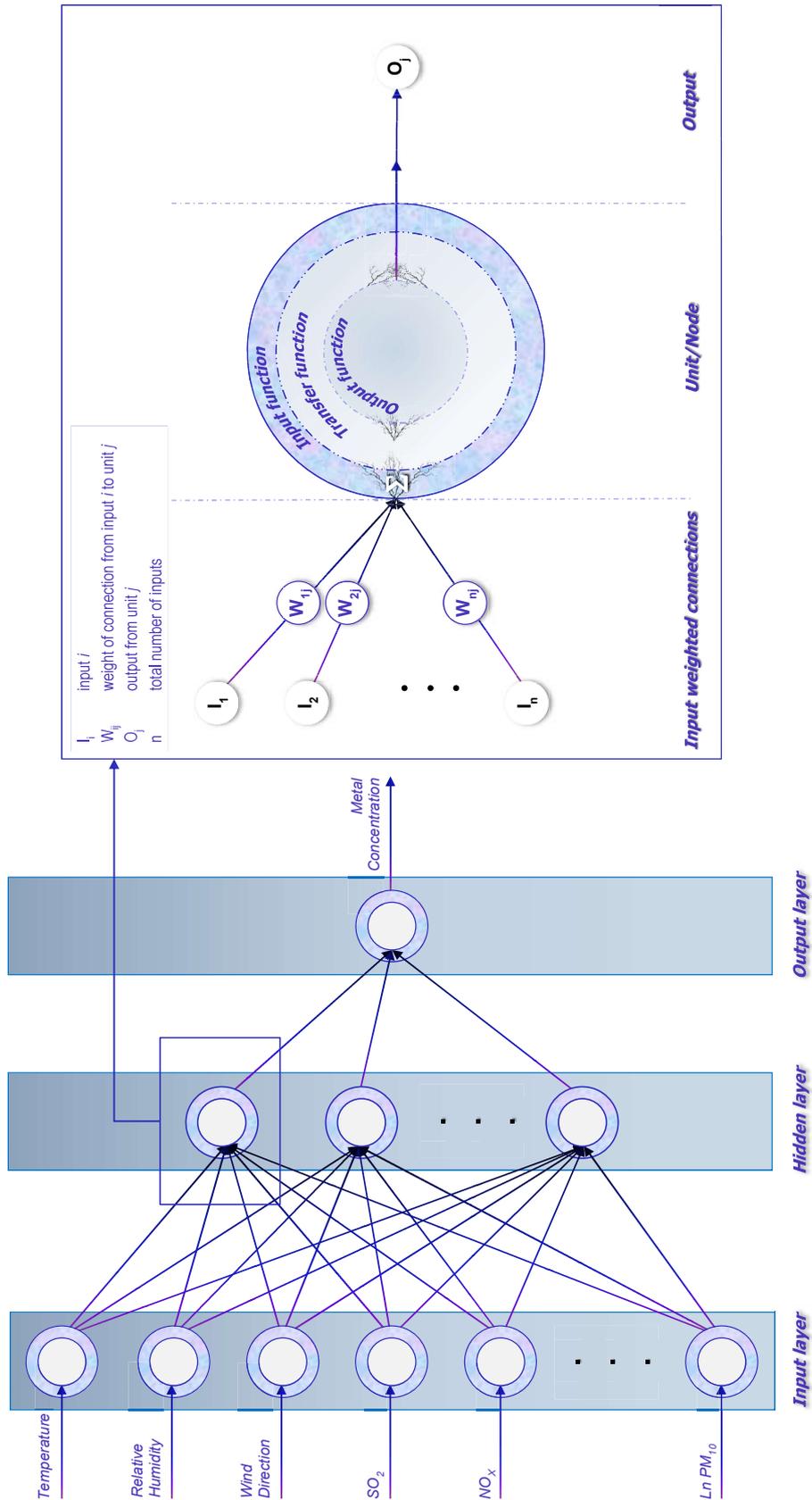
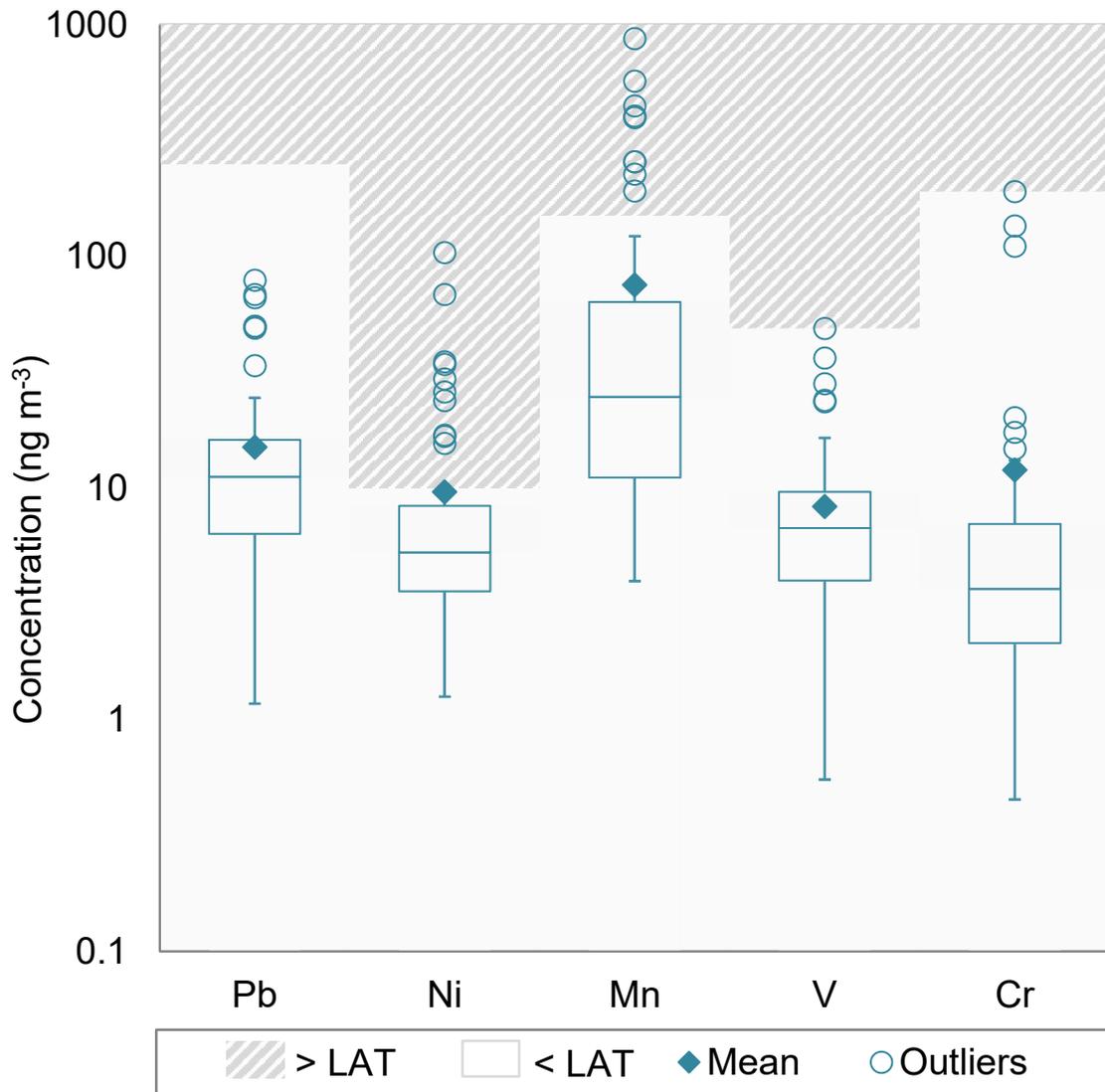
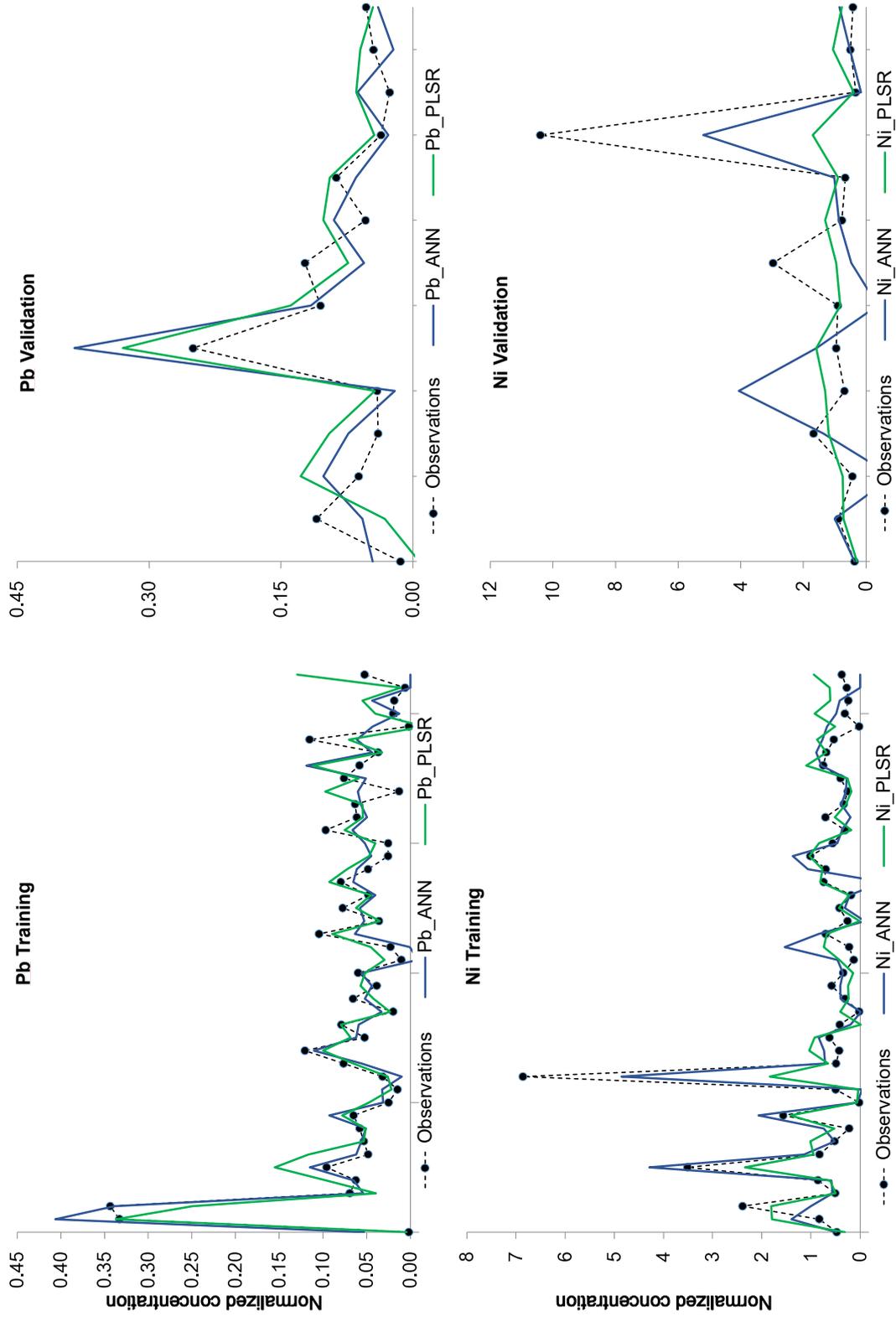


Figure 1 MS Word format

Figure 2 MS Word format



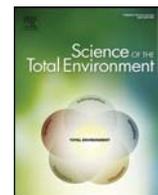






Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A proposed methodology for the assessment of arsenic, nickel, cadmium and lead levels in ambient air



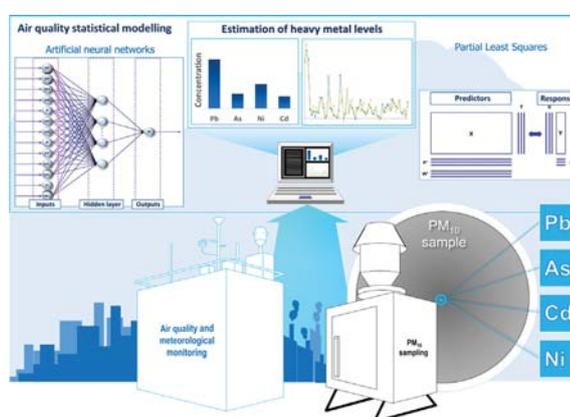
Germán Santos*, Ignacio Fernández-Olmo

Dpto. de Ingenierías Química y Biomolecular, Universidad de Cantabria, Santander 39005, Spain

HIGHLIGHTS

- EU encourages modelling techniques over measurements for air quality assessment.
- A methodology for minor pollutants assessment by statistical modelling is presented.
- PLSR and ANNs represent valid approaches fulfilling EU uncertainty requirements.
- Inputs related to atmospheric stability can slightly improve models performance.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 16 December 2015
 Received in revised form 24 February 2016
 Accepted 25 February 2016
 Available online xxxx

Editor: D. Barcelo

Keywords:

Air quality assessment
 Atmospheric stability
 Methodology
 Metals
 Statistical models

ABSTRACT

Air quality assessment, required by the European Union (EU) Air Quality Directive, Directive 2008/50/EC, is part of the functions attributed to Environmental Management authorities. Based on the cost and time consumption associated with the experimental works required for the air quality assessment in relation to the EU-regulated metal and metalloids, other methods such as modelling or objective estimation arise as competitive alternatives when, in accordance with the Air Quality Directive, the levels of pollutants permit their use at a specific location. This work investigates the possibility of using statistical models based on Partial Least Squares Regression (PLSR) and Artificial Neural Networks (ANNs) to estimate the levels of arsenic (As), cadmium (Cd), nickel (Ni) and lead (Pb) in ambient air and their application for policy purposes. A methodology comprising the main steps that should be taken into consideration to prepare the input database, develop the model and evaluate their performance is proposed and applied to a case of study in Santander (Spain). It was observed that even though these approaches present some difficulties in estimating the individual sample concentrations, having an equivalent performance they can be considered valid for the estimation of the mean values – those to be compared with the limit/target values – fulfilling the uncertainty requirements in the context of the Air Quality Directive. Additionally, the influence of the consideration of input variables related to atmospheric stability on the performance of the studied statistical models has been determined. Although the consideration of these variables as additional inputs had no effect on As and Cd models, they did yield an improvement for Pb and Ni, especially with regard to ANN models.

© 2016 Elsevier B.V. All rights reserved.

* Corresponding author.
 E-mail address: santosg@unican.es (G. Santos).

1. Introduction

The member states of the European Union are obliged by the Air Quality Framework Directive, Directive 2008/50/EC (EC, 2008), to carry out the air quality assessment of certain specified pollutants according to a group of common methods and criteria. These methods vary depending on the population within each air quality zone or agglomeration and the relative levels of the mentioned pollutants compared with the respective assessment thresholds, which are expressed as a percentage of the corresponding limit/target value (EC, 2004; EC, 2008). Thus, fixed measurements are required when it has been confirmed that the levels of an atmospheric pollutant exceed the upper assessment threshold. A combination of fixed measurements and modelling techniques and/or indicative measurements is possible when the pollutant concentration is between both assessment thresholds. Moreover, air quality assessment based exclusively on modelling and/or objective estimation is solely reserved for zones and agglomerations with relatively good air quality and no large conurbations, in which at a preliminary assessment stage the levels of those pollutants whose air quality is to be assessed consistently fall sufficiently below their respective lower assessment thresholds (EC, 2008). In this sense, the minimum quality requirements for the objective estimation techniques to be acceptable in the context of Directive 2008/50/EC, which include uncertainty and minimal data capture and time coverage, are less strict than those for modelling techniques and experimental measurements, either fixed or indicative. Specifically, the relative uncertainty should be lower than 100% (EC, 2008). Although no indications on the nature of these methods are provided in the European Directives, the European Commission in its report for guidance on assessment under the European Union Air Quality Directives interprets objective estimation techniques as “mathematical methods to calculate concentrations from values measured at other locations and/or times, based on scientific knowledge of the concentration distribution”. In this context, statistical models can be regarded as objective estimation techniques in the sense that they are based on statistical data analysis establishing empirical relationships between ambient concentrations and meteorological variables instead of simulating the relationship between emissions and immission concentrations through the description of the physical phenomena that rules the air transportation and dispersion of pollutants in the atmosphere, as with classical deterministic models (Daly and Zannetti, 2007). It consists of generating estimations of the European regulated pollutant concentrations directly from regional air quality information, which indirectly reflects atmospheric pollution in part due to local emission sources. Therefore, there is an interest in deepening the understanding of this sort of methods to be used as air quality assessment tools as they are cost-effective and simpler than other approaches, such as air dispersion modelling which can also be considered more affordable than experimental measurements, especially for certain pollutants, namely metals and polycyclic aromatic hydrocarbons, whose analytical determination, unlike other atmospheric pollutants, is expensive and time consuming.

The use of modelling techniques is encouraged even when continuous monitoring is compulsory because of its capacity of providing supplementary information to air quality monitoring and future projections regarding public exposure under different emission scenarios. Furthermore, modelling is becoming a prominent air quality assessment tool as it has been reported that different modelling approaches have already been incorporated as part of the routine air quality network monitoring and assessment procedures in a number of countries, e.g., Finland, Norway, Sweden or United Kingdom. For those reasons, the use of models and other alternative air quality assessment tools for policy support is expected to continue increasing in the future.

Although the use of statistical modelling techniques to predict immission concentrations of major atmospheric pollutants has been extensively investigated in the literature, there are not many studies addressing the estimation of the levels of compounds bound to

particulate matter; metals in particular. Artificial neural networks (ANNs), state-space modelling, time series autoregressive modelling and linear regression stand out among the approaches used to that end. With respect to ANNs, Chelani et al. (2002) estimated the ambient air levels of Cd, Cr, Fe, Ni, Pb, Zn as well as PM₁₀ in the city of Jaipur, India, by means of artificial neural network models obtaining low values of root mean square error (RMSE). In addition, Li et al. (2009) reconstructed occupational manganese exposure by means of back-propagation artificial neural networks and multiple linear regression. In a previous study, a state-space model coupled with Kalman filter and an autoregressive model with external input (ARX model) was used by Chelani et al. (2001) to predict the concentration of Pb, Fe and Zn and respirable suspended particulate matter in Delhi. Apart from ANNs, some research has been conducted to model metal concentrations in ambient air using other statistical approaches. Hernández et al. (1992) applied state-space modelling, Box-Jenkins modelling and time series autoregressive integrated moving average (ARIMA) models to estimate the daily concentrations of air-particulate Fe and Pb in Madrid (Spain). Predictions of daily Fe were better than those of Pb. No difference being found between State-space and Box-Jenkins models, their outcomes were better than those of ARIMA models in terms of root mean squared error (RMSE), correlation coefficient and efficiency. The state space model performed better than the ARX model. On the other hand, Vicente et al. (2012) developed predictive models based on multiple regression analysis together with time series (ARIMA) models to predict the concentration of total suspended particles (TSP), PM₁₀, As, Cd, Ni and Pb in the ambient air of Castellón (Spain). Arruti et al. (2011) developed statistical models based on multiple linear regression and principal component regression as objective estimation techniques to estimate the immission levels of As, Cd, Ni and Pb at four sampling sites in Cantabria (Northern Spain). Both techniques were found to be valid approaches; however, there was room for improvement with regard to their performance.

The implications of the use of objective estimation techniques by local air quality management authorities to assess air quality when it is possible are discussed in this work. For this purpose, we present a case of study focused on the estimation of ambient air levels of Pb, As, Ni and Cd in Santander (Spain), by means of statistical models based on partial least squares regression (PLSR) and artificial neural networks.

Additionally, given that normally deterministic models take into consideration directly or indirectly the atmospheric stability to simulate the dispersion of pollutants in the atmosphere, this work also aims at determining the influence of the use of input meteorological variables related to atmospheric stability on the performance of the studied statistical models.

2. Materials and methods

2.1. Case of study: European Union regulated metal(oids) in Santander, Spain

Santander, with approximately 176,000 inhabitants in 2014 (INE, 2014), is the capital of the Cantabria Region in the North of Spain. The city is extended over a bay and in its suburbs, at 5–10 km, there is an industrial area mainly related to steel and ferroalloys manufacturing plants that makes Santander as the most complex urban area in the region. A sampling campaign was conducted by our research group to determine daily values of PM₁₀ (particulate matter with aerodynamic diameter lower than 10 µm) in 2008, 2009 and 2011 in a coastal sampling site (ETSIT) located in the rooftop of the building “Escuela Superior de Ingenieros Industriales y de Telecomunicación” (43° 28′ 24″ N, 3° 47′ 54″ W at 23 m.a.s.l) and approximately 1 km far from the sea. Glass microfiber filters (Whatman, 150 mm diameter) were used for PM₁₀ sampling during 2008 and 2009, whereas the 2011 campaign was carried out with quartz microfiber filters (Sartorius AG, 47 mm diameter) because of their lower detection limits, mainly for As (0.03 vs.

0.54 ng/m³) and Cd (0.01 vs. 0.11 ng/m³). Metal levels of 2008 and 2009 are reported in Arruti et al. (2011). In that work, the content of Pb, As, Cd, Ni and other trace metals on PM₁₀ was determined by microwave assisted acidic digestion prior to ICP-MS analysis. Samples from 2011 campaign have been measured for metal concentration following the same experimental procedure given in Arruti et al. (2011) to be included in the whole dataset used in this work. As a result, a total number of 138 samples were available for model development after an outlier removal process – 89 from the period of 2008–2009 and 49 from 2011 –, from which the following percentages were below the method detection limit: 53% for As, 48% for Cd, 23% for Ni and a 6% for Pb.

This work is based on the hypothesis that at a certain location there is an intrinsic relationship between metal(loid) immission concentrations, highly influenced by the anthropogenic activities, such as local industry, road traffic, or residential combustion; and the meteorological conditions and the presence of other major air pollutants in the ambient air, which is mainly due to local anthropogenic activities as well and would indirectly reflect the impact of anthropogenic emissions and the fate of pollutants once they are released to the atmosphere. Therefore, the input variables for statistical model development in this study consist of major air pollutant immission concentrations and meteorological variables, whose values are monitored in real time by local authorities through air quality networks and consequently are available at no additional cost. Major air pollutant data, namely concentrations of nitrogen oxides (NO_x), Sulphur dioxide (SO₂), tropospheric ozone (O₃) and PM₁₀, were obtained from the records of the air quality monitoring station located 1 km far from the sampling site, in Tetuán (43° 28' 0" N, 3° 47' 29" W at 24 m above sea level), which belongs to the Regional Air Quality Network (CIMA, 2015). Meteorological data, namely temperature (T), relative humidity (RH), wind direction (WD), wind speed (WS), pressure (P), solar radiation (SR) and cumulative precipitation (PP), and atmospheric soundings were provided by the monitoring station (CMT) of the State Meteorological Agency in Cueto (43° 29' 28" N,

3° 48' 2" W and at 52 m above sea level), at 2 km from the sampling site. Fig. 1 shows a map with the location of the sampling sites.

2.2. Importance of data splitting on statistical modelling

There are three main aspects that strongly influence the statistical model performance, especially artificial neural network models, and should be taken into account prior to the development of the models: (i) the quality of data; (ii) the number of data, and (iii) the data splitting. The quality of data used for the training directly affects the accuracy of the model. In addition, Flood and Kartam (1994) also noted that the performance of a model strongly depends on the number of available training samples: the more number of samples available for model training/calibration, the more potentially accurate the model can be. Nevertheless, from a practical point of view, it is not always feasible having sufficient data for the training/calibration of a model, especially for certain applications such as the estimation of heavy metal levels in ambient air. Hence, the proportion of samples to include in each of the subsets becomes a very important issue.

Commonly, databases have to be divided into three different subsets to be used for statistical model development: training (calibration), test (cross-validation) and validation (external validation). As its name suggests, the former is used to adjust model parameters until an optimal set is found (train/calibrate the model), the following is used during training to avoid over-fitting, and the latter is employed to evaluate the model performance with respect to its capacity to provide estimations when using samples not previously considered in the training step, which is commonly referred to as testing the generalization ability of the model. The external validation of a model should be carried out by using an external validation subset of samples whose values are within the same range than those of the training subset. Otherwise, poor predictive ability is probable to be achieved, because normally a model is

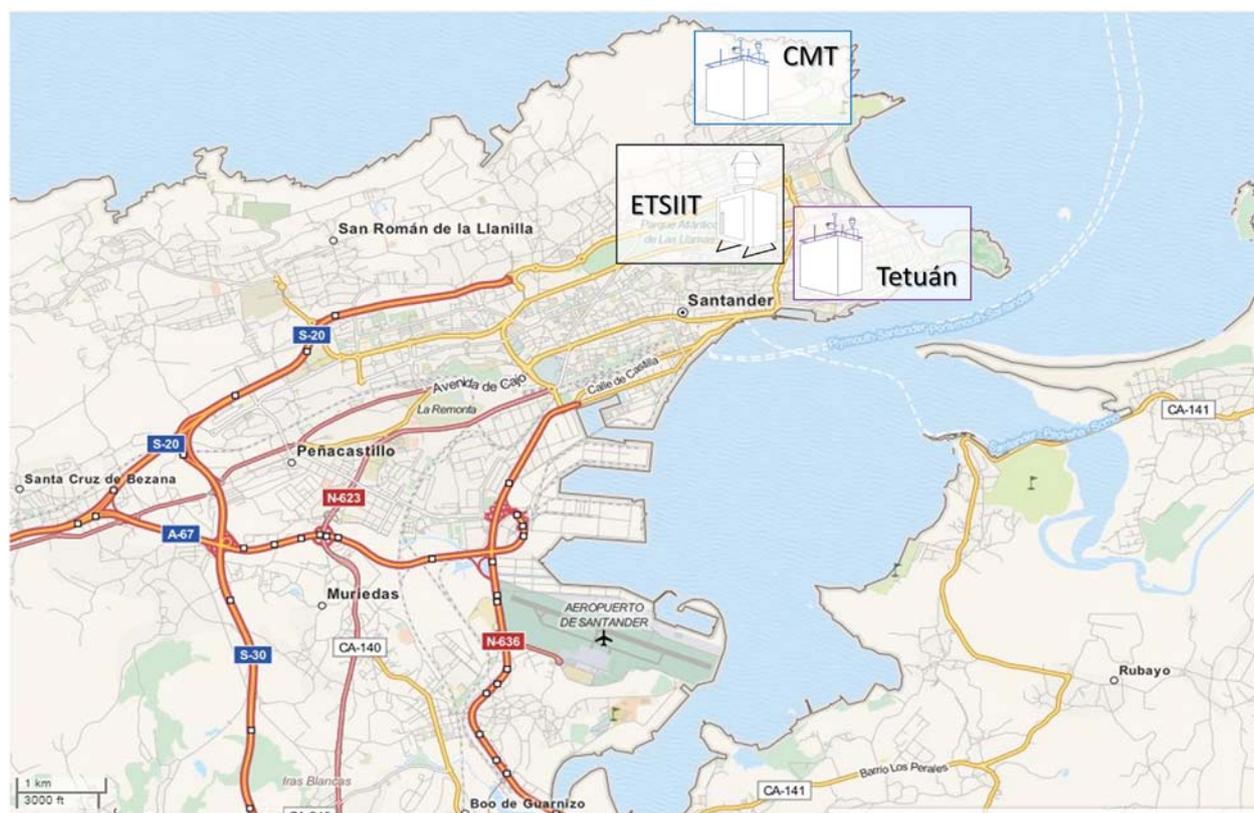


Fig. 1. Location of the sampling site and the air quality and meteorological monitoring stations.

unable to make an extrapolation beyond the range of the training dataset, only interpolation.

Furthermore, a number of studies have reported that the data splitting step can decisively affect the model performance, especially artificial neural network models (Tokar and Johnson, 1999; Maier and Dandy, 2000; Maier et al., 2010). Therefore, optimal data division into subsets is a crucial matter.

From a practical point of view, there are two different approaches of addressing this problem. On the one hand, approach A would correspond to use the data of previous years for model development, validating the model afterwards with the available data of the current year. This would be the real situation for authorities to use the developed model in the future. Authorities in practice are supposed to use the metal concentration data from precedent years along with major air pollutant concentration and meteorological information from the current year to develop models to estimate the pollutant concentrations for the current year. Afterwards, the resulting estimations should be compared with data from experimental measurements to test if those estimations are in fact acceptable. Whereas, on the other hand, approach B would consist of dividing the entire dataset into the aforementioned three sets disregarding the year to which they belong.

2.3. Description of methodology

Fig. 2 shows a flow diagram of the proposed methodology for the development of air quality models. It consists of five main steps: (i) removal of multivariate outliers, (ii) determination of samples that are above the method detection limit by means of a classification tree, (iii) data splitting into the required subsets for model calibration according to the two approaches described above, (iv) model development, and (v) model performance evaluation to check if air quality assessment is valid.

2.3.1. Multivariate outlier removal

Multivariate outlier identification and removal method was based on Mahalanobis distance. It is a well-known classical approach that computes the Mahalanobis distance (MD) of each observation as an indicative measure of the distance of each data point from the centre of the multivariate data cloud. By convention, this method identifies as outliers those observations with a large MD (exceeding the 99% quantile of a chi-square distribution).

2.3.2. Classification tree

The number of samples of metal concentration is generally limited because of the difficulty associated with their analytical determination. In addition, it may be possible that a great amount of these pollutant levels are below the method detection limit in those zones and agglomerations where objective estimation techniques are allowed to be used as air quality assessment tools. The common strategy for those samples below the method detection limit is to be quantified as half the method detection limit. Nevertheless, to conduct such substitution for a considerable number of samples especially in a database modest in size could make a negative impact on the subsequent model performance. Following the approach proposed by Zalel et al. (2015) to estimate airborne benzo(a)pyrene concentrations in the Czech Republic, a binary classification tree prior to model development has been applied to obtain, based on the observed independent and dependent variables, a set of rules that allow to estimate if the metal concentration of a specific sample is above or below a threshold value set as the method detection limit. The classification tree allow us identifying whether the combination of input variables corresponding to a certain sample represents conditions that are expected to be associated with a metal concentration above the method detection limit. In that case, the mentioned sample is included in the database that will be used afterwards for the quantitative estimation of the levels of that specific metal through

model development. Otherwise, a preset estimation value equal to half the value of the method detection limit is assigned.

2.3.3. Data splitting

Given the importance of data splitting, highlighted in Section 2.2, a number of studies have dealt with the problem of selecting representative subsets from a set of samples (Wu et al., 1996; Tominaga, 1998; Sales et al., 2000; Bowden et al., 2002; Daszykowski et al., 2002; Wu et al., 2012). Random data splitting is a common method due to its inherent simplicity and because the data randomly extracted from a larger set follows the statistical distribution of the entire set, however neither does this method ensure the representativity of the extracted subsets nor does it ensure that the samples on the boundaries are included in the training/calibration subset, what would help to minimize possible extrapolation problems. Therefore, other techniques to extract data based on criteria established under more rigorous scientific basis are preferable. In this context, Kennard and Stone (1969) proposed an algorithm that measures the similarity between samples on the basis of the Euclidean distance. This method, as well as a group of other methods, e.g. DUPLEX algorithm, has been widely applied since then. In this context, Saptoro et al. (2012) proposed a modification of the Kennard-Stone algorithm, namely Kennard-Stone algorithm based on Mahalanobis distance (MDKS), to perform the data partition to develop artificial neural network models. Better performance results were found for this method when compared with the standard Kennard-Stone algorithm and other related algorithms such as the SPXY method, a data division method based on joint x-y distances and proposed by (Galvão et al., 2005). In this work the MDKS algorithm has been applied.

2.3.4. Model development

PLSR is a statistical method that as multiple linear regression (MLR) performs a linear combination of the predictors that best correlates the outcome. However, these predictors, as with principal component analysis (PCA), are actually components or *latent variables* created as a combination of the original input variables, but with the constraint of describing as much as possible the covariance between the independent and the dependent variables, whereas PCA focuses on the variance of the independent variables. The number of latent variables to be considered has been determined by a cross-validation. The PLSR models in this study have been developed using MATLAB (MathWorks, Inc.).

ANNs are computational systems that emulate the behavior of neural nervous systems of animals. They are composed by simple processing elements known as artificial neurons or units connected to one another on a weighted basis and organized in layers. The network architecture for the ANN models in this work corresponds to a Multilayer perceptron (MLP) that consisted of an input layer (each input unit represents one input variable), followed by a single hidden layer before an output layer (with just one unit, containing the dependent variable or immission concentration of Pb, As, Cd or Ni). The ANNs have been trained using a Levenberg-Marquardt learning algorithm to update the set of connecting weights and with a sigmoid activation function in the hidden layer and a linear activation function in the output layer. Different numbers of units in the hidden layer have been tested and, finally, those that provided better results have been selected. The models have been developed using the Neural Network Toolbox for MATLAB (MathWorks, Inc.).

2.3.5. Performance criteria for model evaluation

Air quality model performance evaluation has become an important issue because the use of models for policy support has broadened. This is addressed in several documents published by policy-making authorities (ASTM standard D6589, 2005; EPA, 2009; Denby, 2010; Derwent et al., 2010). It is commonly conducted by means of a statistical performance analysis, which consists of comparing estimations against the corresponding observations using statistical indicators to evaluate the ability of the model to reproduce the measured concentrations. In this sense,

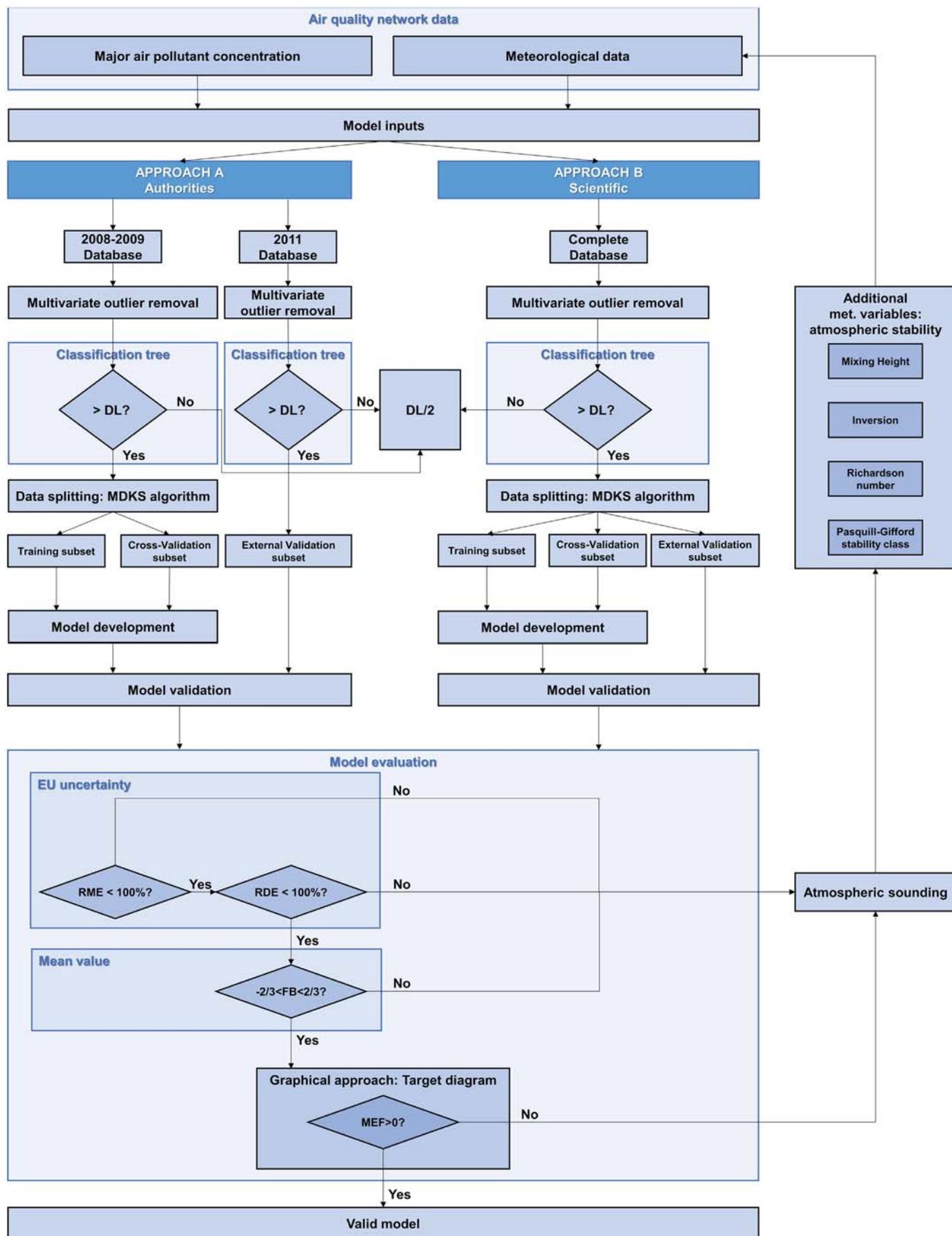


Fig. 2. Flow diagram of the proposed methodology for objective estimation air quality assessment.

Boylan and Russell (2006) define performance criteria as the minimum quality levels of the mentioned statistical parameters that a model has to comply with to be considered as acceptable for policy applications. These series of parameters allow making a general analysis of model performance. However, they provide insufficient insight on whether the models have satisfactorily reached an adequate quality level for a given application such as policy support (Boylan and Russell, 2006).

The quality level of the performance of objective estimation techniques has to be evaluated in the context of the European Air Quality Directives. The uncertainty for objective estimation is defined as the maximum deviation of the measured and calculated concentration levels over the period considered, by the limit value/target value, without taking into account the timing of the events. In this work, uncertainty requirements have been interpreted through the consideration of two indices: the relative maximum error without timing (RME), which is defined as the largest concentration difference of all percentile (p) differences between the observed concentration ($C_{O,p}$) and the estimated concentration ($C_{E,p}$) normalized by the respective measured value (Flemming and Stern, 2007; Borrego et al., 2008), as given by Eq. (1), and the relative directive error (RDE), the difference between the closest observed concentration to the limit/target value ($C_{O,LV}$) and the correspondingly ranked modelled concentration ($C_{E,LV}$) normalized by the limit/target value (LV/TV) (Denby, 2010), as calculated by Eq. (2).

$$RME = \max(|C_{O,p} - C_{E,p}|) / C_{O,p} \quad (1)$$

$$RDE = |C_{O,LV} - C_{E,LV}| / LV \quad (2)$$

According to Directive 2008/50/CE (EC, 2008), the data quality objective regarding objective estimation uncertainty is 100% and, therefore, that is the maximum value allowed for these indices.

Additionally, because limit/target values of metals in European Air Quality Directives are provided as annual mean values, the adequacy of mean value estimation is to be tested. This is done by means of the fractional bias:

$$FB = \frac{\bar{C}_O - \bar{C}_E}{0.5(\bar{C}_O + \bar{C}_E)} \quad (3)$$

where \bar{C}_O and \bar{C}_E are the mean values (ng/m^3) of the observations and the estimations, respectively.

Two limit situations have been considered depending on if the estimated mean value is double or half the observed mean value, which corresponds to a FB index values equal to $-2/3$ and $2/3$, respectively. Therefore, an FB index value outside the range of $-2/3$ and $2/3$ will be considered unacceptable.

As RDE is mainly related to the exceedance of limit values, it provides only a partial view of the strengths and weaknesses of a given model application (Pederzoli et al., 2011). Thus, in order to deepen the analysis of model performance, a set of additional statistical indicators was required. In this work, some statistical parameters included in the EURASAP model validation kit were considered. This includes: the correlation coefficient (r), the root mean squared error (RMSE), the normalized mean squared error (NMSE) and the fractional variance (FV).

However, up until now, not many minimum quality criteria for these statistical indicators have been proposed in the literature regarding model performance evaluation and not always in a harmonized way. In this context, Kumar et al. (1993) proposed that the normalized mean squared error (NMSE) should be lower or equal to 0.5 and the fractional variance (FV) should be within the range from -0.5 to 0.5 . These criteria have been applied in this work.

Additionally, in a more illustrative manner, a graphical approach has been applied to perform an easy comparison between model results and an evaluation with respect to their level of accuracy for policy

applications. This representation, known as the target diagram, was proposed by Pederzoli et al. (2011) and has been developed as a modification of the Taylor diagram (Taylor, 2001) by the Joint Research Centre (JRC) of the European Commission within the framework of the Forum for Air Quality Modelling in Europe (FAIRMODE) and introduced inside the DELTA Tool (Thunis et al., 2012). It compares the mean bias, y-axis, as given by Eq. (4), and the centered mean squared error (CRMSE), x-axis, as given by Eq. (5), both normalized by the standard deviation of the observations (σ_O). In addition, it includes a circumference of unit radius, which defines the acceptable limit value of model efficiency, MEF, (Eq. (6)). According to this representation, the closest the model results are to the center of the circumference the higher the model performance. However, the model results must be inside the circumference to be considered acceptable, which means in practice that the normalized bias and CRMSE must be lower than 1 and a MEF must have a positive value.

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N (C_{E,i} - C_{O,i}) = \bar{C}_E - \bar{C}_O \quad (4)$$

$$\text{CRMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N [(C_{E,i} - \bar{C}_E) - (C_{O,i} - \bar{C}_O)]^2} \quad (5)$$

where N is the total number of samples.

$$\text{MEF} = 1 - \left(\frac{\text{RMSE}}{\sigma_O} \right)^2 \quad (6)$$

2.4. Input variables related to atmospheric stability

Atmospheric stability plays an important role in air quality studies as it strongly affects the dispersion of pollutants in the atmosphere and it is indirectly introduced as an input in deterministic models. For these reasons, four meteorological variables regarded as good candidates to be introduced in the models as additional input variables to properly represent atmospheric stability have been considered: mixing height, the Richardson number, inversion and the Pasquill–Gifford atmospheric stability class. The value of these variables was determined from vertical temperature profiles of atmospheric soundings conducted by the Spanish State Meteorology Agency (AEMET) in Santander (Fig. 1).

The mixing height (MH) is the thickness of the air layer adjacent to the ground over which pollutants are mixed by convection or mechanical turbulence and, therefore, delimits the volume available for its dispersion (Beyrich, 1997; Seibert et al., 2000). In this study, the mixing height was determined applying the Holzworth method (Holzworth, 1972), a graphical approach of widespread use in the literature for this purpose. According to this method, the mixing height is given by the intersection between the dry adiabatic profile and the vertical temperature profile of an atmospheric sounding. In addition, those cases where inversion (I) occurred (i.e. temperature increased with height) were also identified (1: no inversion, 2: moderate inversion; 3: complete inversion).

Pasquill (1961) proposed a classification for the atmospheric stability in terms of increasing order from very unstable (A) to moderately stable (F). Thus, an atmosphere with an A stability class favors strong mixing whereas classes E or F impede the dispersion. In contrast, atmospheric stability can be numerically accounted for by means of the Richardson number (Ri), a dimensionless number that determines stability through the relationship of convective and mechanical forces that take part in the atmospheric movements. Specifically, it is the ratio of the vertical temperature gradient to the squared vertical gradient of the wind speed. The numerator is related to the destabilizing forces that generate updrafts. The denominator is related to the kinetic energy that destroys updrafts (Woodward, 1998). Therefore, the larger the

Richardson number, the more unstable is the atmosphere. In this study, the Richardson number was calculated in accordance to Woodward (1998) as follows:

$$Ri = \frac{g (T_t - T_b) (z_t - z_b)}{T_b (u_t - u_b)^2} \quad (7)$$

where g is the acceleration of gravity (m/s^2), T is the air temperature (K), z is sensor height (m) above ground level and u is the horizontal wind speed (m/s). The subscripts t and b are the top and bottom positions of vertical profiles, which in this work correspond to the mixing height and the ground level, respectively.

A number of equivalences between Pasquill-Gifford atmospheric stability classes and Ri values exist. In this work, the corresponding Pasquill-Gifford stability class (PG) was determined according to the equivalence proposed by Woodward (1998), and coded as 1–6, where 1 is very unstable (A class) and 6 is moderately stable (F class).

As two soundings a day (at 0 h and 12 h) were available in Santander station, two different values were considered for every input variable related to atmospheric stability: one in the morning and one in the evening, which are identified with the subscripts *am* and *pm*, respectively.

3. Results and discussion

3.1. Estimation of Pb, As, Cd and Ni immission levels in Santander

The air quality in relation to Pb, As, Cd and Ni has been assessed in Santander in this work by means of statistical models based on PLSR and ANNs following the methodology described in Fig. 2. The results of the best developed models considering approaches A and B (see Section 2.2) are shown in Tables 1 and 2, respectively. Aside from external validation results, training results are also included for comparison purposes in order to test the generalization ability of the models.

Mathematically, every model performance indicator considered in this work has an ideal value of zero, except for the correlation coefficient and the model efficiency, whose ideal value is one. In this sense, except for Pb models, which exhibit a similar performance with both approaches, the external validation is in general better when considering approach B instead of approach A. This is probably due to the fact that the models corresponding to Approach A have been trained with data from 2008 and 2009 and validated with data from 2011. Hence, there is a gap of one year between the data used to train/calibrate the models and the data used to test them and consequently there may have been a change in emission patterns especially if the period of time in which the study is focused is to be considered, because it corresponds to that of a great impact of the economic crisis on industrial activities that may

have altered their production, with severe repercussions on emissions, in a local scale in particular. For that reason, it is assumed that according to approach A the model is not correctly calibrated to be able to predict the external validation dataset (2011) situation, with the consequent detriment in the generalization ability of the model. Therefore, the reason why approach B provides better results resides in the fact that the consideration of the three studied years simultaneously favors to some extent the attenuation of the differences between data relationships from 2011 and those of the preceding years. This is the main limitation of the proposed approach to be used as a routine air quality assessment tool by local authorities: its application is restricted to those cases in which the emission patterns are maintained over time, which is an acceptable assumption in the short-term but rarely valid in the long-term. For that reason, the models should be periodically updated with at least a few number of new measurements – control measurements – to guarantee that they are not out to date.

Regarding the evaluation of the presented approaches as tools for air quality assessment in the context of the European Directives, it would be strictly sufficient that the model estimations comply with the uncertainty and the mean value quality requirements because limit/objective values of Pb, As, Cd and Ni are given in mean concentrations over a calendar year. As seen in Table 2, all the best developed models considering approach B fulfill these requirements because both uncertainty indices, RME and RDE, are below 100% and the FB index is within the range of $-2/3$ and $2/3$, which corresponds to half and double mean observed concentration, respectively. In fact, RDE values range up to 30%, with half of them below 10%. This means that, the difference between the observed value that is closest to the limit/target value and the respective estimation is lower than 30% of the actual limit/target value and, often, even lower than 10% of the limit/target value. Moreover, PLSR FB index values are in general below those of ANNs in both training and external validation steps.

Nevertheless, apart from that, local authorities would also be interested in being cognizant of the time variations of the studied pollutants or in being able to provide estimations with smaller time resolution different from an annual basis. Analyzing the values of the remaining statistical parameters, which in this case give an idea on how well the models are able to estimate the observed concentrations of the individual 24-h samples, it is observed that there are general difficulties in providing accurate estimations of the 24-h mean concentrations. The NMSE and FV tend to be greater than 0.5. In addition, FV index values are all positive, which indicates that the estimated values show less variation than the observations. The estimations seem to vary around a value within the range of the dependent variable and close to the average, therefore, there is an overestimation of the lower values and more importantly an underestimation of the higher values. In addition, except

Table 1
PLSR and ANN model results considering approach A.

Pollutant	Model	Subset ^a	RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	FB	r	RMSE (ng/m ³)	NMSE	FV
Pb	PLSR	T	50.9	5.0	6.7	6.7	0.000	0.424	6.77	1.02	0.73
		EV	69.5	8.6	13.2	9.1	0.370	0.304	13.32	1.48	1.16
	ANN	T	55.3	3.6	6.7	6.7	-0.005	0.636	5.77	0.74	0.46
		EV	74.6	8.2	13.2	8.7	0.407	0.404	12.98	1.46	1.00
As	PLSR	T	2.5	0.1	1.1	1.1	0.000	1.000	0.02	0.00	0.00
		EV	61.4	71.1	2.7	1.3	0.733	0.367	1.96	1.10	0.49
	ANN	T	35.6	14.2	1.1	1.0	0.076	0.801	0.50	0.23	1.07
		EV	79.8	96.2	2.7	0.1	0.971	0.258	2.22	1.90	1.55
Ni	PLSR	T	46.9	15.7	1.1	1.1	0.000	0.553	0.82	0.51	0.51
		EV	54.4	28.2	1.8	1.4	0.226	0.280	1.51	0.92	0.71
	ANN	T	58.7	12.4	0.9	1.3	-0.295	0.584	0.80	0.55	1.28
		EV	74.4	28.2	1.8	1.2	0.341	0.294	1.54	1.08	1.57
Cd	PLSR	T	70.0	24.0	0.3	0.3	0.000	0.354	0.29	0.93	0.81
		EV	60.7	5.0	0.2	0.4	-0.538	-0.071	0.30	1.14	0.52
	ANN	T	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.
		EV	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.

^a T: training; EV: external validation; n.c.: not calculated.

Table 2
PLSR and ANN model results for Pb, As, Cd and Ni considering approach B.

Pollutant	Model	Subset ^a	RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	FB	r	RMSE (ng/m ³)	NMSE	FV
Pb	PLSR	T	62.0	7.58	7.6	7.6	0.000	0.366	9.4	1.55	0.71
		EV	60.7	8.15	10.8	9.9	0.090	0.257	10.4	1.01	0.77
	ANN	T	80.5	8.60	7.9	6.5	0.185	0.523	9.2	1.64	1.03
		EV	78.1	7.42	10.8	6.0	0.579	0.338	11.1	1.92	1.03
As	PLSR	T	13.0	2.83	0.7	0.7	0.000	0.971	0.1	0.04	0.12
		EV	47.8	16.01	1.0	0.8	0.268	0.714	0.6	0.42	0.43
	ANN	T	62.3	20.55	0.7	0.6	0.164	0.878	0.5	0.56	1.33
		EV	59.2	14.82	1.0	0.6	0.461	0.996	0.7	0.82	1.42
Ni	PLSR	T	45.2	13.75	1.1	1.1	0.000	0.488	0.8	0.58	0.36
		EV	32.5	9.19	1.5	1.5	0.008	0.530	0.9	0.38	0.46
	ANN	T	47.5	12.74	1.1	1.1	-0.010	0.564	0.7	0.44	0.68
		EV	51.8	14.11	1.5	1.1	0.316	0.517	1.0	0.63	0.79
Cd	PLSR	T	65.9	22.09	0.3	0.3	0.000	0.199	0.3	1.22	0.91
		EV	43.3	8.06	0.2	0.3	-0.492	0.160	0.2	0.97	1.38
	ANN	T	56.8	29.03	0.3	0.2	0.269	0.467	0.3	1.77	0.77
		EV	24.5	10.85	0.2	0.2	-0.182	-0.186	0.3	2.04	0.20

^a T: training; EV: external validation.

for those of As, the correlation coefficient values are poor, lower than or equal to 0.5 in most cases. Nonetheless, there is no evidence of a superior performance of one technique over the other: PLSR and ANN models exhibit a comparable equivalent performance, what may be interpreted as that the relationship between model inputs and immission concentrations of Pb, As, Cd and Ni may tend to be linear rather than non-linear, at least as far as the available data are concerned. This idea is reinforced by the simplicity of most of the best neural network architectures developed, involving only a single hidden layer with just a few hidden neurons.

Fig. 3 represents the training and external validation results of the PLSR and ANN models developed for Pb, As, Cd and Ni taking into consideration approach B and the graphical approach described in Section 2.3.5. Although, the CRMSE is always positive by its own mathematical definition (Eq. (5)), a negative sign has been assigned to distinguish those situations when the standard deviation of the estimations is lower than the corresponding to the observations (Pederzoli et al., 2011). All the model results are located in the left side of the diagram, i.e. there are negative CRMSE values, indicating that the estimations vary within a narrower range than observations. This is also supported by the positive values of the FV index, as mentioned before (Table 2). Furthermore, as shown in Fig. 3, the external validation results for Pb and Cd models in approach B have a negative model efficiency and consequently they are out from the circumference. According to Jolliff et al. (2009) the fact that the estimated values are inside the circumference, i.e. the model efficiency is greater than zero, means that the model is a better predictor of the observations than a constant value set to the average of observations.

3.2. Influence of input meteorological variables related with atmospheric stability

3.2.1. Classification of independent variables according to their correlation coefficient

The relative importance of every independent variable to be used as input variable to the model has been determined through the corresponding Pearson correlation coefficient with respect to the output variables, so that the more significant inputs can be identified. This coefficient evaluates the joint variation degree or the similarity of two variables, quantifying the linear relationship existing between them. Tables 3 and 4 shows a classification of the independent variables – major pollutant concentrations and meteorological variables including those related to atmospheric stability – in descending order of their correlation coefficient in relation to each dependent variable, i.e. metal (loid) immission concentration. Additionally, the respective *p*-value is also presented. As two different approaches have been considered,

approaches A and B (see Section 2.2), Table 3 makes reference to the data of the period 2008–2009, whereas Table 4 presents results for the complete dataset, including 2011.

First of all, there is a non-strong correlation between the independent variables and the dependent variables individually. However, the consideration of the entire set of independent variables altogether may change the individual influence due to synergic effects.

Pb and Ni seem to be more related to major air pollutant concentration, especially PM₁₀ concentrations. However, this tendency is not observed for As and Cd.

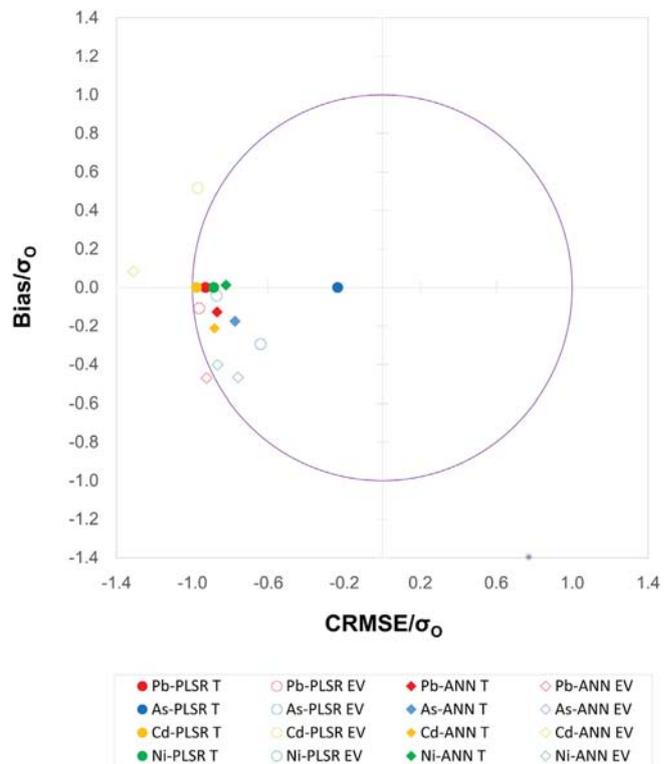


Fig. 3. Target diagram of PLSR and ANN model results for Pb (red), As (blue), Ni (green) and Cd (yellow). Data points related to PLSR are represented by a circle and those related to ANNs are represented by a rhombus. Filled and hollow markers have been used to depict training (T) and external validation (EV) results, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Classification of independent variables according to correlation coefficient for the period 2008–2009.

Ranking	Pb		As		Ni		Cd		
	Position	Variable	p-Value	Variable	p-Value	Variable	p-Value	Variable	p-Value
1		NO _x	0.003	Ri _{am}	0.002	Ln PM ₁₀	0.001	MH _{pm}	0.014
2		Ln PM ₁₀	0.010	SD	0.003	I _{pm}	0.002	PP	0.194
3		O ₃	0.027	T	0.053	RH	0.012	SE	0.251
4		I _{pm}	0.031	Ln PM ₁₀	0.081	PG _{am}	0.022	I _{am}	0.314
5		MH _{am}	0.291	RS	0.107	I _m	0.025	O ₃	0.366
6		P	0.307	WE	0.134	SO ₂	0.093	I _{pm}	0.392
7		WE	0.361	SE	0.174	P	0.102	NO _x	0.422
8		RS	0.365	PG _{pm}	0.190	MH _{am}	0.109	P	0.481
9		T	0.419	PP	0.279	WS	0.141	T	0.533
10		MH _{pm}	0.467	I _v	0.310	PG _{pm}	0.161	WS	0.539
11		SO ₂	0.553	O ₃	0.432	NO _x	0.181	PG _{pm}	0.549
12		SD	0.555	Ri _{pm}	0.437	T	0.196	WE	0.553
13		PP	0.612	MH _{pm}	0.446	MH _{pm}	0.216	MH _{am}	0.577
14		Ri _{pm}	0.642	NO _x	0.446	SD	0.235	RS	0.654
15		Ri _{am}	0.705	P	0.526	RS	0.236	Ln PM ₁₀	0.675
16		WD	0.710	MH _{am}	0.574	WE	0.391	PG _{am}	0.681
17		SE	0.730	RH	0.675	Ri _{pm}	0.793	SO ₂	0.726
18		PG _{pm}	0.738	WD	0.679	SE	0.798	RH	0.822
19		I _{am}	0.861	PG _{am}	0.722	O ₃	0.807	Ri _{pm}	0.832
20		WS	0.861	SO ₂	0.790	Ri _{am}	0.883	SD	0.857
21		RH	0.886	I _{am}	0.809	WD	0.914	WD	0.871
22		PG _{am}	0.999	WS	0.886	PP	0.988	Ri _{am}	0.955

The meteorological variables related to atmospheric stability and the *p*-values of those variables statistically significant with a 90% confident level are presented in bold.

The consideration of samples from 2011 leads to a significant modification in the correlation coefficients and affects the relative position of each variable in the classification. As mentioned before in Section 3.1, this may be caused by a change in the emission patterns in the Region, probably associated with the fluctuations in the production of industrial activities as a consequence of the economic and financial crisis. Not in vain, the period of study extends from 2008 to 2011, which is the period of time with more detrimental effects for the economy.

However, in any case, some of the meteorological variables related to atmospheric stability are ranked in the top positions for every pollutant. What is more, for every pollutant in the period 2008–2011, at least one of the four considered new input variables was among those statistically significant with a 90% confidence level. In general terms, based on

the ranking positions of these new variables, the Pasquill-Gifford stability class appears to be the most significant variable.

3.2.2. Estimation of Pb, Cd and Ni immission levels with the new meteorological variables

In this study, it is proposed to consider meteorological variables related with atmospheric stability as additional input variables to the models to further improve their performance. As a consequence, PLSR and ANN models have been developed considering as supplementary input meteorological variables those that, based on the individual correlation coefficient (Table 4), presented most relevance for each pollutant. Table 5 shows the results for Pb and Ni. A reduction in the RME and RDE values together with an increment of the correlation coefficient values is

Table 4
Classification of independent variables according to correlation coefficient for the period 2008–2009 and 2011.

Ranking	Pb		As		Ni		Cd		
	Position	Variable	p-Value	Variable	p-Value	Variable	p-Value	Variable	p-Value
1		O ₃	0.001	WS	0.000	WS	0.000	PG _{pm}	0.005
2		NO _x	0.004	WD	0.000	Ln PM ₁₀	0.001	MH _{pm}	0.005
3		WD	0.014	SO ₂	0.000	T	0.015	PG _{am}	0.016
4		PG _{am}	0.049	T	0.000	WD	0.018	SE	0.021
5		RH	0.052	NO _x	0.000	RH	0.022	Ri _{pm}	0.121
6		SE	0.064	SE	0.000	PG _{am}	0.031	RH	0.123
7		MH _{am}	0.085	I _{am}	0.001	MH _{am}	0.041	I _{am}	0.125
8		PG _{pm}	0.088	O ₃	0.046	I _{pm}	0.043	WD	0.163
9		I _{am}	0.097	I _{pm}	0.076	PG _{pm}	0.097	P	0.182
10		Ln PM ₁₀	0.130	Ln PM ₁₀	0.100	SE	0.105	PP	0.212
11		I _{pm}	0.131	Ri _{pm}	0.113	RS	0.159	O ₃	0.226
12		P	0.138	PG _{pm}	0.154	P	0.189	MH _{am}	0.247
13		SO ₂	0.181	PG _{am}	0.170	Ri _{pm}	0.269	SD	0.261
14		T	0.208	MH _{am}	0.178	O ₃	0.293	RS	0.293
15		MH _{pm}	0.239	WE	0.315	WE	0.323	WE	0.385
16		Ri _{pm}	0.242	RH	0.360	SO ₂	0.361	Ri _{am}	0.386
17		Ri _{am}	0.283	PP	0.371	Ri _{am}	0.364	T	0.399
18		PP	0.341	MH _{pm}	0.408	I _{am}	0.387	WS	0.447
19		SD	0.344	RS	0.409	NO _x	0.432	I _{pm}	0.477
20		WS	0.352	SD	0.441	PP	0.436	SO ₂	0.480
21		WE	0.425	Ri _{am}	0.459	SD	0.445	Ln PM ₁₀	0.486
22		RS	0.483	P	0.480	MH _{pm}	0.486	NO _x	0.494

The meteorological variables related to atmospheric stability and the *p*-values of those variables statistically significant with a 90% confident level are presented in bold.

observed when considering the morning Pasquill–Gifford stability class (PG_{am}). Except for the Ni PLSR model, there is also an enhancement of the FB values of the external validation subset. The results for Cd are not presented because there was no significant improvement of the model performance. Figs. 4 and 5 show a comparison between the target diagram obtained for regular Pb and Ni models presented in Section 3.1 and those obtained when considering the PG_{am} as an input. In general, there is a slight improvement of model results. In particular, it is observed that the consideration of the morning Pasquill–Gifford stability class as an additional input meteorological variable has a greater influence over the ANN model outcomes. It results in a decrease in the bias and the CRMSE values of the ANN external validation results for Pb what lead to a change in the MEF sign, which turn into a positive value positioned inside the circumference. Regarding Pb PLSR model results, they experienced a minor upgrade, moving slightly closer to the center.

With respect to Ni, the results of ANN models exhibit a more noticeable improvement than PLSR models, especially in terms of mean concentration estimation of the external validation subset, with a considerably reduced bias value. With respect to Ni PLSR results, although they still are inside the circumference and therefore they can still be considered valid, they experienced a slight deterioration regarding the external validation results.

Due to the considerable effort required for the user, as it entails the availability of atmospheric sounding data and an advanced data processing, this approach should only be taken into consideration in those situations in which there is room for improvement, i.e. an unsatisfactory model performance is obtained because the values of statistical parameters used to test it do not meet quality objectives.

4. Conclusions

In this study a proposal for a methodology to use statistical models based on PLSR and ANNs for air quality assessment in relation to EU regulated metal(loids) is presented. Two different approaches have been considered: (i) approach A, which would represent the actual situation for authorities regarding air quality assessment, and (ii) approach B, which would consist of using all the available information disregarding the period of time to which it belongs. These approaches have been tested using a dataset of Pb, As, Cd and Ni levels in Santander (Spain) in 2008, 2009 and 2011.

It is observed that statistical parameters used as indicators of model performance are in general closer to their ideal value when using approach B. This may be caused by a modification of the relationship between inputs and outputs of the models as a consequence of a change of emission patterns over time, because they have been trained/

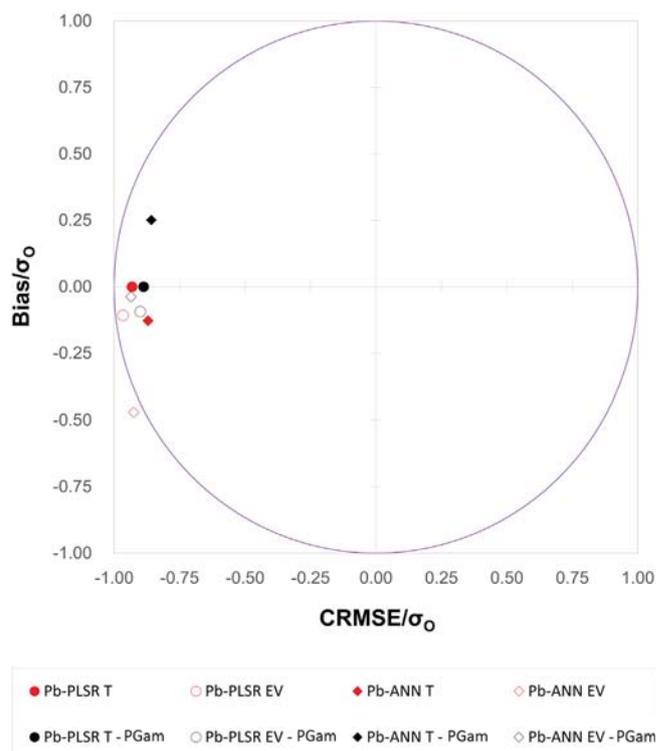


Fig. 4. Target diagram of PLSR and ANN model results for Pb to determine how considering PG_{am} as an additional input affects their performance.

calibrated with data from 2008 to 2009 and validated with data from 2011. Taking into account this consideration, in order to use approach A with confidence as a routine air quality assessment tool in the future, it would be indispensable to carry out brief control sampling campaigns to include a few new additional samples to keep the models up to date.

In the context of the European Framework Air Quality Directive, approach B would be considered a satisfactory air quality assessment tool in relation to Pb, As, Ni and Cd as it meets the quality requirements for estimations regarding uncertainty and accuracy in the mean value estimation. However, the models fail to estimate with precision daily concentrations. For that reason, the consideration of new input meteorological variables related to atmospheric stability to improve model performance has been tested with those variables that have shown a greater correlation with the levels of the EU regulated metals and metalloids in Santander. The consideration of the morning

Table 5
PLSR and ANN model results for Pb and Ni with and without considering PG_{am} as input variable under approach B.

Pollutant	Model	Subset ^a	RME (%)	RDE (%)	\bar{C}_O (ng/m ³)	\bar{C}_E (ng/m ³)	FB	r	RMSE (ng/m ³)	NMSE	FV
Pb	PLSR	T	62.0	7.58	7.6	7.6	0.000	0.366	9.4	1.55	0.71
		EV	60.7	8.15	10.8	9.9	0.090	0.257	10.4	1.01	0.77
	PLSR - PG_{am}	T	57.3	6.70	7.6	7.6	0.000	0.574	8.0	1.11	0.56
		EV	54.8	5.33	9.8	9.6	0.012	0.463	8.2	0.72	0.69
	ANN	T	80.5	8.60	7.9	6.5	0.185	0.523	9.2	1.64	1.03
		EV	78.1	7.42	10.8	6.0	0.579	0.338	11.1	1.92	1.03
	ANN - PG_{am}	T	57.3	6.49	7.4	9.5	-0.247	0.642	7.6	0.82	0.71
		EV	45.9	5.97	9.8	9.4	0.041	0.515	8.0	0.71	0.84
Ni	PLSR	T	45.2	13.75	1.1	1.1	0.000	0.488	0.8	0.58	0.36
		EV	32.5	9.19	1.5	1.5	0.008	0.530	0.9	0.38	0.46
	PLSR - PG_{am}	T	27.5	12.31	1.2	1.2	0.000	0.593	0.7	0.42	0.43
		EV	41.6	21.36	1.5	1.4	0.092	0.410	1.2	0.68	0.52
	ANN	T	47.5	12.74	1.1	1.1	-0.010	0.564	0.7	0.44	0.68
		EV	51.8	14.11	1.5	1.1	0.316	0.517	1.0	0.63	0.79
	ANN - PG_{am}	T	5.6	1.05	1.2	1.2	-0.025	0.958	0.2	0.05	0.06
		EV	29.7	14.98	1.5	1.4	0.103	0.479	1.2	0.70	0.23

^a T: training; EV: external validation.

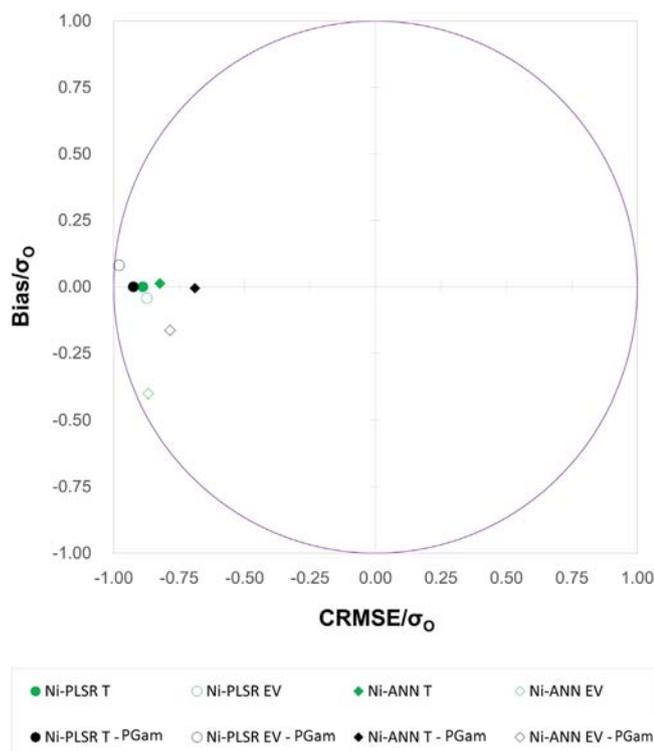


Fig. 5. Target diagram of PLSR and ANN model results for Ni to determine how considering PG_{am} as an additional input affects their performance.

Pasquill-Gifford stability class as an additional meteorological input was yield a slight improvement in the model performance for Pb and Ni.

On the other hand, the similarity between results obtained applying the two different techniques considered, PLSR and ANN, indicates that at least for the available data the relationship between inputs and outputs may have a linear character. This is something to take into account for future model development, given the relative simplicity of linear techniques in comparison to non-linear techniques.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Projects CTM2010-16068 and CTM2013-43904R. Germán Santos thanks MINECO for his FPI research fellowship (BES-2011-047908). The authors also acknowledge the Spanish State Meteorology Agency (AEMET) for providing meteorological and atmospheric sounding data for the period of study.

References

Arruti, A., Fernández-Olmo, I., Irabien, A., 2011. Assessment of regional metal levels in ambient air by statistical regression models. *J. Environ. Monit.* 13 (7), 1991–2000.

ASTM standard D6589, 2005. Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (No. D6589). ASTM International, West Conshohocken, PA.

Beyrich, F., 1997. Mixing height estimation from sodar data – a critical discussion. *Atmos. Environ.* 31 (23), 3941–3953.

Borrego, C., Monteiro, A., Ferreira, J., Miranda, A.I., Costa, A.M., Carvalho, A.C., Lopes, M., 2008. Procedures for estimation of modelling uncertainty in air quality assessment. *Environ. Int.* 34, 613–620.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. *Water Resour. Res.* 38 (2), 1–11.

Boylan, J.W., Russell, A.G., 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. *Atmos. Environ.* 40, 4946–4959.

Centro de Investigación del Medio Ambiente (CIMA), 2015. Datos on-line de la Red de Control y vigilancia de la Calidad del Aire de Cantabria. Consejería de Medio Ambiente, Ordenación del Territorio y urbanismo del Gobierno de Cantabria (http://www.medioambientecantabria.es/calidad_aire/).

Chelani, A.B., Gajghate, S.M., Tamhane, M., Hasan, M.Z., 2001. Statistical modelling of ambient air pollutants in Delhi. *Water Air Soil Pollut.* 132, 315–331.

Chelani, A.B., Gajghate, D.G., Hasan, M.Z., 2002. Prediction of ambient PM_{10} and toxic metals using artificial neural networks. *J. Air Waste Manage. Assoc.* 52 (7), 805–810.

Daly, A., Zannetti, P., 2007. Air Pollution Modeling – an overview. In: Zannetti, P., Al-Ajmi, D., Al-Rashied, S. (Eds.), *Ambient air pollution*, Arab School for Science and Technology (ASST) and The EnviroComp Institute.

Daszykowski, M., Walczak, B., Massart, D.L., 2002. Representative subset selection. *Anal. Chim. Acta* 468, 91–103.

Denby, B., 2010. Guidance on the use of the models for the European Air Quality Directive. A Working document of the Forum for Air Quality Modelling in Europe, FAIRMODE. ETC/ACC Report.

Derwent, D., Fraser, A., Abott, J., Willis, P., Murrells, T., 2010. Evaluating the performance of air quality models. (No issue 3). Department for Environment and Rural Affairs.

EPA, 2009. Guidance Document on the Development, Evaluation, and Application of Regulatory Environmental Models. No. EPA/100/K-09/003. U.S. Environmental Protection Agency.

European Commission (EC), 2004. Council Directive 2004/107/EC, Directive of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air. The European Parliament and the Council of the European Union. *Off. J. L23*, 3–16.

European Commission (EC), 2008. Council Directive 2008/50/EC, Directive of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. The European Parliament and the Council of the European Union. *Off. J. L152*, 1–44.

Flemming, J., Stern, R., 2007. Testing model accuracy measures according to the EU directives – examples using the chemical transport model REM-CALGRID. *Atmos. Environ.* 41, 9206–9216.

Flood, I., Kartam, N., 1994. Neural networks in civil engineering. I: principles and understanding. *J. Comput. Civ. Eng.* 8 (2), 131–148.

Galvão, R.K.H., Araújo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., Saldanha, E.C.B., 2005. A method for calibration and validation subset partitioning. *Talanta* 67, 736–740.

Hernández, E., Martín, F., Valero, F., 1992. Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain. *Atmos. Environ.* 26B, 107–116.

Holzworth, G.C., 1972. Mixing heights, wind speeds and potential for urban pollution throughout the contiguous United States. Office of Air Programs Ap-101, U.S.EPA (NTIS PB 207 103).

Instituto Nacional de Estadística (INE), 2014. INEbase. Cifras oficiales de población de los municipios españoles. <http://www.ine.es/dynt3/inebase/es/index.html?padre=517&dh=1>.

Jolliffe, J.K., Stow, C.A., McGillicuddy, D.J., Allen, J.I., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. *J. Mar. Syst.* 76, 4–15.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. *Technometrics* 11 (1), 137–148.

Kumar, A., Luo, J., Bennett, G., 1993. Statistical evaluation of lower flammability distance (LFD) using four hazardous release models. *Process. Saf. Prog.* 12, 1–11.

Li, Y., Luo, F., Jiang, Y., Lu, Y., Huang, J., Zhang, Z., 2009. A prediction model of occupational manganese exposure based on artificial neural network. *Toxicol. Mech. Methods* 19 (5), 337–345.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Softw.* 15, 101–124.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environ. Model. Softw.* 25, 891–909.

Pasquill, F., 1961. The estimation for the dispersion of wind borne material. *Meteorol. Mag.* 90, 34–49.

Pederzoli, A., Thunis, P., Georgieva, E., Borge, R., Carruthers, D., 2011. Performance criteria for the benchmarking of air quality model regulatory applications: the “target” approach. 14th Conference on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes, pp. 297–301.

Sales, F., Rius, A., Callao, M.P., Rius, F.X., 2000. Standardization of a multivariate calibration model applied to the determination of chromium in tanning sewage. *Talanta* 52, 329–336.

Saportó, A., Tadej, M.O., Vuthaluru, H., 2012. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chem. Prod. Process. Model.* 7 (1), 13.

Seibert, P., Beyrich, F., Gryning, S.E., Joffre, S., Rasmussen, A., Tercier, P., 2000. Review and intercomparison of operational methods for the determination of the mixing height. *Atmos. Environ.* 34, 1001–1027.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* 106 (No. D7), 7183–7192.

Thunis, P., Georgieva, E., Pederzoli, A., 2012. A tool to evaluate air quality model performances in regulatory applications. *Environ. Model. Softw.* 38, 220–230.

Tokar, A.S., Johnson, P.A., 1999. Rain fall-runoff modeling using artificial neural networks. *J. Hydrol. Eng.* 4 (3), 232–239.

Tominaga, Y., 1998. Representative subset selection using genetic algorithms. *Chemom. Intell. Lab. Syst. J.* 43, 157–163.

Vicente, A.B., Jordán, M.M., Sanfeliu, T., Sánchez, A., Esteban, M.D., 2012. Air pollution prediction models of particles, As, Cd, Ni and Pb in a highly industrialized area in Castellón (NE, Spain). *Environ. Earth Sci.* 66 (3), 879–888.

- Woodward, J.L., 1998. Improving the effect of atmospheric stability class for dispersion modeling. *Process. Saf. Prog.* 17 (1), 1–8.
- Wu, W., Walczak, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R., Prebble, K.A., 1996. Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom. Intell. Lab. Syst.* 33, 35–46.
- Wu, W., May, R., Dandy, G.C., Maier, H.R., 2012. A method for comparing data splitting approaches for developing hydrological ANN models. *Managing Resources of a Limited Planet: Proceedings of the 6th International Congress on Environmental Modelling and Software (iEMSs 2012)*, pp. 1620–1627.
- Zalel, A., Yuval, Svecova V., Sram, R.J., Bartonova, A., Broday, D.M., 2015. Modelling airborne benzo(a)pyrene concentrations in the Czech Republic. *Atmos. Environ.* 101, 166–176.



UC