TESIS DOCTORAL

"Técnicas Inteligentes para el Análisis de Condiciones Medioambientales"

Autor:

Ángel Arroyo Puente

Directores:

Dr Emilio S. Corchado Rodríguez

Dra Verónica Tricio Gómez

Dr Álvaro Herrero Cosío

Doctorado en Ingeniería Informática

Departamento de Ciencias de la Computación y Automática

Marzo, 2017

VNiVERSiDAD
D SALAMANCA

La memoria titulada "Técnicas Inteligentes para el Análisis de Condiciones Medioambientales" que presenta D. Ángel Arroyo Puente para optar al Grado de Doctor por la Universidad de Salamanca ha sido realizada bajo la dirección del Dr. Emilio Santiago Corchado Rodríguez de la Universidad de Salamanca y de los Dres. Verónica Tricio Gómez y Álvaro Herrero Cosío, ambos de la Universidad de Burgos.

Salamanca, marzo de 2017

Los Directores

Dr. Emilio S. Corchado Rodríguez
Universidad de Salamanca

Dra. Verónica Tricio Gómez
Universidad de Burgos

Dr. Álvaro Herrero Cosío
Universidad de Burgos

El Doctorando

D. Ángel Arroyo Puente

A mi padre Ángel,

Formato de la Tesis

La presente tesis doctoral ha sido elaborada con el formato de compendio de artículos/publicaciones, de acuerdo con lo contemplado en el artículo 14.1 del Capítulo II del Reglamento de Doctorado de la Universidad de Salamanca aprobado por la Comisión de Doctorado y Posgrado el 15 de febrero de 2013 que indica: "*los estudios de doctorado concluyen con la elaboración y defensa de una tesis doctoral, que consistirá en un trabajo original de investigación, elaborado por el doctorando, en cualquier campo del conocimiento, siguiendo el formato determinado por la Comisión Académica del Programa de Doctorado, entre los posibles formatos establecidos por la Comisión de Doctorado*". El presente apartado da cumplimiento al artículo 4.1 del Procedimiento para la Presentación de la Tesis Doctoral en la Universidad de Salamanca en el formato de Compendio de Artículos/Publicaciones, aprobado por la Comisión de Doctorado y Posgrado el 15 de febrero de 2013.

De ente las publicaciones que presentan los resultados obtenidos con esta tesis, se han seleccionado los siguientes artículos. Estos han sido publicados en revistas científicas del ámbito de la especialidad del trabajo desarrollado en la tesis e indexadas en el Journal Citation Report del Science Citation Reports (o equivalente según la CNEAI en los campos científicos en los que dicho criterio no sea aplicable).

1.- [1]Ángel Arroyo, [2]Emilio Corchado, [3]Verónica Tricio, [1]Álvaro Herrero. **Analysis of Meteorological Conditions in Spain by means of Clustering Techniques**. Journal of Applied Logic (2016). DOI: 10.1016/j.jal.2016.11.026. 2015 ISI JCR Science Edition: 0.524. 2015 ISI JCR Ranking: 13/22 (Logic) – Q3.

2.- [1]Ángel Arroyo, [2]Emilio Corchado, [3]Verónica Tricio. **Soft Computing Models to Analyze Atmospheric Pollution Issues**. Logic Journal of the IGPL, 20 (4) (2012) 699-711. DOI: 10.1093/jigpal/jzr023. 2012 ISI JCR Science Edition: 1.136. 2012 ISI JCR Ranking: 1/20 (Logic) – Q1.

---

[1] Departamento de Ingeniería Civil, Universidad de Burgos, Burgos, España

[2] Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España

[3] Departamento de Física, Universidad de Burgos, Burgos, España

3.- [2]Emilio Corchado, [1]Ángel Arroyo, [3]Verónica. Tricio. **Soft computing models to identify typical meteorological days.** Logic Journal of the IGPL, 19 (2) (2011) 373-383. DOI: 10.1093/jigpal/jzq035. 2011 ISI JCR Science Edition: 0.913. 2011 ISI JCR Ranking: 1/19 (Logic) – Q1

Adicionalmente, con la presente tesis se pretende optar a la mención de "Doctor internacional", por lo que esta se ajusta también a las indicaciones establecidas en el Procedimiento para la Obtención de la Mención de "Doctor Internacional" en el Título de Doctor por la Universidad de Salamanca, aprobado por la Comisión de Doctorado, en su sesión de fecha 10 de noviembre de 2011.

De acuerdo con lo anterior y considerando la motivación y objetivos de esta tesis, el resto del presente documento queda estructurado como se expone a continuación:

- Parte I. Introducción y Objetivos: En esta primera parte se presenta la motivación y objetivos generales de la tesis, así como la estructura del propio documento.
- Parte II. Trabajos seleccionados: Esta segunda parte presenta los artículos más representativos publicados por el autor y directores de la tesis en diversos congresos y revistas de ámbito internacional, listados anteriormente.

Índice

Agradecimientos

En primer lugar, agradecer a la Universidad de Salamanca por posibilitar la presentación de este trabajo de Tesis Doctoral en tan honorable universidad.

Mi más sincero agradecimiento al Dr. Emilio Corchado Rodríguez de la Universidad de Salamanca, a la Dra. Verónica Tricio Gómez y al Dr. Álvaro Herrero Cosío de la Universidad de Burgos por su infatigable trabajo de asesoramiento y guiado durante todos estos años. A todos los miembros del grupo de investigación GICAP y del área de Lenguajes y Sistemas Informáticos de la Universidad de Burgos por su compañía y amistad.

A las personas que han dedicado parte de su tiempo a revisar mi trabajo, incluyendo publicaciones, artículos en congresos y al tribunal evaluador de esta tesis doctoral.

A mi madre, hermana y familia por preocuparse tanto de mí.

Ángel Arroyo Puente

Salamanca, marzo de 2017

Resumen

Como es bien sabido, la calidad del aire es un tema importante y preocupante en la actualidad que afecta no solamente a la salud humana sino a otros muchos aspectos como el cambio climático o la supervivencia de la biosfera. En los últimos años, numerosas entidades públicas se han ido adaptando a las restrictivas medidas de contaminación ambiental impuestas por las diversas normativas europeas, siendo España uno de los países obligados a cumplir estas normativas. Tanto en España como en otros países existen diversas redes de monitorización de la calidad del aire y de adquisición de valores meteorológicos de una forma continua. Estas redes de estaciones de medida no sólo están presentes en las grandes ciudades sino también en zonas periféricas, polígonos industriales y en zonas donde la preservación de la naturaleza es fundamental. Además, están sometidas a constantes procesos de reordenación para mejorar su función.

En la presente Tesis Doctoral se han aplicado diversas técnicas inteligentes (*Soft Computing* más específicamente) a conjuntos de datos públicos con información meteorológica y/o de calidad del aire. Las técnicas aplicadas llevan a cabo fundamentalmente dos tareas: reducción de la dimensionalidad y agrupamiento (*clustering*). Estas se han aplicado de forma aislada y de forma combinada para mejorar los resultados obtenidos en el análisis de la información medioambiental. Las técnicas de reducción de la dimensionalidad aplicadas son: *Principal Component Analysis* (PCA) como técnica aplicada en primer lugar para obtener una primera aproximación a la estructura del conjunto de datos, *Locally Linear Embedding* (LLE) como técnica no lineal local, *Maximum Likelihood Hebbian Learning* (MLHL) y *Cooperative Maximum Likelihood Hebbian Learning* (CMLHL) como modelos neuronales que implementan *Exploratory Projection Pursuit*, *Curvilinear Component Analysis* (CCA) como modelo no lineal que intenta preservar la distancia entre los puntos en la salida, *Multidimensional Scalling* (MDS) como técnica global no lineal basada en la matriz de distancias, *Isometric Mapping* (ISOMAP) como técnica derivada de MDS y los *Self-Organizing Maps* (SOM), un importante modelo neuronal que implementa aprendizaje competitivo.

Las técnicas de agrupamiento aplicadas han sido por una lado particionales: *k-means* como primer método a aplicar en agrupamiento y que busca la asignación de muestras a grupos aplicando métricas de distancia, *SOM k-means* que utiliza los algoritmos de SOM para la actualización de los pesos, *k-medoids* como técnica derivada de *k-means* y que asigna el centroide de cada grupo a uno de los puntos del mismo y *fuzzy c-means,* técnica que aplica lógica difusa para tareas de agrupamiento. Por otro lado, también se ha empleado el método aglomerativo jerárquico en el que se van

formando los grupos de forma ascendente, junto con diversos métodos de evaluación de agrupamiento que sirven para determinar el posible número de grupos existentes en un conjunto de datos y dendrogramas para obtener una representación gráfica de la agrupación de los datos en forma de árbol.

Los casos de estudio han sido cuidadosamente seleccionados y se extienden desde el ámbito local, regional hasta el nacional. Por otra parte, también se ha dado importancia a los periodos de tiempo seleccionados. En alguno de los estudios se analizan periodos de tiempo tan cortos como un día para el análisis de la meteorología/calidad del aire en un breve periodo de tiempo en un lugar determinado, mientras que en otros se emplean ventanas temporales próximas a una década y en los puntos más representativos climatológicamente en España. Partiendo de uno o más conjuntos de datos públicos con la información más completa posible acerca de las condiciones medioambientales (meteorológica, de calidad del aire o ambas), pero siempre analizando variables determinantes en la caracterización de las condiciones medioambientales, el objetivo es extraer la información fundamental almacenada en los conjuntos de datos mediante las técnicas inteligentes. De esta forma es posible analizar las condiciones medioambientales en los casos de estudio seleccionados. En cada uno de los casos de estudio se hace un análisis de la situación meteorológica o de calidad del aire en las localizaciones y periodos seleccionados, buscando semejanzas y diferencias en las muestras de datos analizadas y haciendo énfasis en aquellas situaciones anómalas detectadas y tratando de dar explicación a las mismas. También se hace un análisis comparativo de los resultados obtenidos con las distintas técnicas empleadas, planteando las ventajas e inconvenientes del uso de cada uno de ellas en cada caso de estudio.

Las técnicas de reducción de la dimensionalidad resultan de gran utilidad para analizar gráficamente conjuntos de datos multidimensionales, encontrar relaciones en los datos y detectar situaciones anómalas. De manera complementaria, las técnicas de agrupamiento revelan la estructura de un conjunto de datos asignando las muestras de datos a los distintos grupos en función de las medidas de distancias y similitud aplicadas. Esto resulta de gran utilidad en el presente trabajo para entender las semejanzas y diferencias en la meteorología y/o calidad del aire de los distintos puntos seleccionados en cada caso de estudio.

Abstract

It is well known that air quality is an important and worrying issue nowadays, affecting not only human health but also many other aspects such as climate change or the survival of the biosphere. In recent years, many public institutions have been adapted to the restrictive normative about environmental pollution imposed by European regulations, being Spain one of the countries that must comply with these regulations. Both in Spain and in other countries there are various air-quality networks and stations for the continuous acquisition of meteorological parameters. These networks are not only present in big cities, but also in peripheral and industrial areas, as well as in places where the preservation of nature is fundamental key issue. Furthermore, they are constantly rearranged to improve their function.

In present PhD Thesis, different intelligent techniques (more specifically, *Soft Computing* techniques) have been applied to publicly available databases with air quality and/or meteorological information. The applied techniques perform two fundamental tasks: dimensionality reduction and clustering. They have been applied in isolation and in conjunction in order to improve the results in the analysis of environmental conditions. The applied dimensionality reductions techniques are: *Principal Component Analysis* (PCA) as the technique firstly applied to obtain an approximation to the dataset structure, *Locally Linear Embedding* (LLE) as a non-linear local technique, *Maximum Likelihood Hebbian Learning* (MLHL) and *Cooperative Maximum Likelihood Hebbian Learning* (CMLHL) as neural models which implement *Exploratory Projection Pursuit*, *Curvilinear Component Analysis* (CCA) as a non-linear technique which tries to preserve the interpoint distance in the output space, *Multidimensional Scalling* (MDS) as a non-linear global technique operating with the distance matrix, *Isometric Mapping* (ISOMAP) as a technique derived from MDS and *Self-Organizing Maps* (SOM), as a competitive learning neural model.

The applied clustering techniques are, on the one hand partitional techniques*: k-means* as the clustering technique firstly applied, which assigns samples to groups using distance metrics, *SOM k-means* which use the SOM algorithm for the weight updating process, *k-medoids* as a *k-means* derived technique which assigns the centroid of each cluster to one of the belonging samples, and *fuzzy c-means* as a fuzzy-logic based technique for grouping samples. On the other hand, hierarchical agglomerative techniques have also been applied (where groups are formed in an ascending way) together with different clustering evaluation indexes, used to determine the possible

number of existing groups in a dataset, and finally dendrograms for a tree-form graphical representation of clustering.

Case studies have been carefully selected and range from local, regional to national contexts. Similarly, the selected periods of time have also been a priority. In some of the studies, the analyzed period of time is one day long, considered for the analysis of meteorological / air quality in a short time interval in a certain place, while in other cases, long periods of time (close to a decade), are used to analyze some of the most climatological representative places in Spain. From one or more public datasets comprising all the information about environmental conditions (weather, air quality, or both), but always analyzing key variables in the characterization of environmental conditions, the goal is to extract the meaningfully information in the datasets by applying intelligent techniques. This leads to an analysis of the environmental conditions in the selected case studies. In each case study, an analysis of the weather or air quality conditions is carried out in the selected places and periods of time, searching for similarities and differences in the analyzed data samples, emphasizing those detected anomalous situations and trying to give an explanation to these phenomena's. A comparative analysis of the results obtained with the different techniques applied is also performed, considering the advantages and disadvantages of using each of them in each case study

Dimensionality reduction techniques are useful for graphically analyzing high-dimensional data sets, find relationships in datasets and detect anomalous situations. Complementarily, clustering techniques reveal the structure of datasets by assigning the data samples to different clusters depending on the applied distance and similarity measures. This is useful in present work to understand the similarities and differences in the meteorological and / or air quality conditions of the different locations selected in each case study.

Parte I. Introducción y Objetivos

Capítulo I.Introducción

Antecedentes

El término 'Condiciones Medioambientales' es realmente amplio, pero se puede resumir como la suma de factores o circunstancias que influyen en la vida de cualquier organismo y por lo tanto también en la vida humana. Estas circunstancias son muy variadas y entre ellas figuran con gran importancia la climatología y la calidad del aire. A continuación, se describen estos dos aspectos.

En los últimos años, el conocimiento acerca de la contaminación atmosférica y la comprensión de sus efectos han avanzado mucho. Se ha aceptado que la contaminación ambiental no sólo representa un riesgo para la salud, sino que también, por ejemplo, reduce la producción de alimentos y el crecimiento vegetativo debido a sus efectos sobre la fotosíntesis. Otras consecuencias graves que pueden mencionarse son la lluvia ácida, la corrosión, el cambio climático o el calentamiento global. Por lo tanto, los esfuerzos que se dirigen hacia el estudio de estos fenómenos [1][2] pueden mejorar nuestro conocimiento y ayudan a mejorar la grave problemática de la contaminación atmosférica.

Hace varios años, diversas entidades públicas adoptaron los principios del V Programa Marco de la Unión Europea [3] y de la Conferencia de Río concernientes a la aplicación de directivas comunes internacionales; en particular, las relativas a la calidad del aire ambiental en nuestras ciudades. Las directivas mencionadas anteriormente se basaban en la legislación europea en vigor con lo que todos los estados miembros de la Unión Europea estaban sujetos a ellas.

La búsqueda de soluciones a los problemas actuales del medio ambiente constituye un paso fundamental para el sostenimiento de la vida. El cumplimiento de este deseo viene en gran medida determinado por la necesidad de preservar un aire limpio, teniendo en cuenta su impacto sobre la biosfera. La comprensión de los mecanismos por los cuales se emiten los contaminantes en el aire es indispensable, como lo es el conocimiento de sus ciclos de vida atmosféricos, posibles reacciones derivadas de su combinación y las pautas de eliminación entre otros puntos, teniendo en cuenta que los enfoques del problema varían de acuerdo a sus contextos espaciales y temporales.

Las mediciones sistemáticas en España [4], tomadas por lo general dentro de las grandes ciudades, son fundamentales debido a los riesgos que para la salud suponen los altos niveles de contaminación atmosférica. Las últimas tendencias apuntan a los beneficios de continuar la extensión de la red de estaciones de medición de la contaminación atmosférica. La legislación europea, además de fijar ciertos valores límite y objetivo en relación con los diferentes contaminantes, establece el cómo y dónde se deben medir estos contaminantes a largo plazo.

Por otro lado, la meteorología y la climatología son campos de estudio diferentes, a pesar de que generalmente se perciben como algo similar. La meteorología consiste en el estudio de la atmósfera, el estudio científico de los fenómenos y procesos físicos que ocurren en la misma, así como los efectos atmosféricos sobre el clima. Los meteorólogos realizan previsiones meteorológicas destinadas a predecir las condiciones meteorológicas a corto plazo. La climatología es el estudio de los cambios atmosféricos que definen los climas y su evolución en el tiempo, debido tanto a la variabilidad climática natural como antropogénica. Los estudios climatológicos utilizan los mismos parámetros que los meteorológicos, pero su propósito es diferente; estudiar las características climáticas a largo plazo. Los periodos de tiempo analizados no son los suficientemente amplios como para sacar conclusiones acerca de la climatología en las localizaciones seleccionadas. Por otra parte, el estudio de la climatología necesitaría disponer de un conjunto de parámetros mayor de los disponibles y analizados en los distintos casos de estudio. La presente Tesis Doctoral se centra en el estudio de la meteorología y de la calidad de aire en las localizaciones seleccionadas.

Con la aplicación de las técnicas inteligentes a los conjuntos de datos con información meteorológica, de calidad del aire o ambas, se pretende demostrar la validez y eficiencia de estas técnicas para analizar gráficamente y numéricamente grandes conjuntos de datos de alta dimensionalidad, pudiendo analizar las diferencias y semejanzas en la caracterización de la meteorología o calidad del. En el caso de la calidad del aire, se puede verificar si los niveles de contaminación aérea se corresponden con el tipo de estación analizada (urbana, suburbana, de vegetación protegida, industrial …), cómo evoluciona la misma en el periodo de tiempo seleccionado o la influencia de días festivos en la contaminación aérea en un punto seleccionado.

En el caso de los estudios meteorológicos, la aplicación de las técnicas inteligentes permitirá determinar las características de la meteorología a lo largo de un día en un punto concreto de una ciudad o determinar las semejanzas y diferencias en la

meteorología de distintos puntos seleccionados con cuatro climas diferentes en España.

Previamente, otros autores han aplicado técnicas de reducción de la dimensionalidad en el campo de la calidad del aire y de la meteorología. Tal es el caso de [5], donde por primera vez se aplica PCA para analizar la calidad de los sistemas de control del aire de 15 estaciones de medida europeas. Estas 15 estaciones de medida recogen datos sobre los contaminantes más habituales y el trabajo trata de analizar las ventajas y desventajas de cada una de ellas, así como el interés público que tienen. En [6] se aplica la técnica no lineal de ISOMAP para el estudio de la temperatura del Pacífico de cara a predecir las oscilaciones del fenómeno conocido como "El Niño". En [7] se analizan las concentraciones de ozono en 4 ciudades de la India: Kolkata, Mumbai, Chennai y New Delhi. Para ello se propone PCA como preprocesado para posteriormente aplica un modelo neuronal encargado de predecir los niveles de Ozono en las 4 ciudades mencionadas. En [8] se presenta un método basado en PCA para reducir grandes conjuntos de datos con información meteorológica a conjuntos de datos fácilmente interpretables, a partir de los datos mensuales provenientes de los Satélites SAT y SLP desde 1948 hasta la actualidad. Las nuevas componentes espaciales son utilizadas, por ejemplo, para ser usadas como nodos en complejas redes de análisis de la evolución climática a gran escala.

En la mayoría de los trabajos previos analizados se utilizan las técnicas de reducción de la dimensionalidad como preprocesado de los datos, pero no para visualizar gráficamente la estructura y comportamiento de los conjuntos de datos con el objetivo de analizar estos, como se propone en esta tesis. Además, en el trabajo actual se realiza un estudio comparativo muy completo, empleando un gran número de técnicas.

Las técnicas inteligentes de agrupamiento han sido aplicadas en numerosas ocasiones en el campo de estudio de la meteorología y calidad del aire, como por ejemplo en [9] donde se aplican dos técnicas de agrupamiento y una red neuronal con el fin de analizar la calidad del aire en Grecia y el impacto de la meteorología en la calidad del aire considerando núcleo urbano y en un plazo de cinco años. Un cubo con datos multidimensionales se analiza en [10], tratando de analizar coincidencias en el etiquetado de datos climatológicos y de vegetación. Se aplican técnicas de agrupamiento y se estudian las diferencias entre la agrupación de las variables climáticas en comparación con las variables de vegetación. En [11], las trayectorias del aire por periodos de 24 horas se calculan durante tres años en España. Estos cálculos se hacen mediante la aplicación de técnicas de agrupamiento con trigonometría esférica, junto

con el método de "*kernel regression*". A diferencia de los trabajos previos analizados, en esta Tesis Doctoral se combinan las técnicas de agrupamiento con las técnicas de reducción de la dimensionalidad, además se utilizan las técnicas de agrupamiento para confrontar conjuntos de datos con información de diferentes localizaciones y analizar la distancia existente entre estas muestras de diferentes localizaciones.

Solución Propuesta

Para resolver la problemática anteriormente presentada, este trabajo propone fundamentalmente la aplicación de técnicas de Inteligencia Artificial [12] (más específicamente de *Soft Computing* [13][14]. El propósito de estas técnicas es la investigación, análisis y resolución de problemas complejos en los que puede aparecer información incompleta o inexacta.

Las técnicas de reducción de la dimensionalidad [15] se pueden definir como aquellas que pretenden transformar un conjunto de datos de alta dimensionalidad para que este tenga un número de dimensiones menor, extrayendo la información más útil o representativa del conjunto de datos original con el objeto de encontrar posibles patrones, relaciones o direcciones en los datos, permitiendo la visualización de los mismos. Entre las técnicas de reducción de la dimensionalidad aplicadas en los trabajos expuestos en esta Tesis Doctoral destacan el Análisis de Componentes Principales (ACP) (en inglés *Principal Component Analysis* - PCA) [16], *Locally Linear Embedding* (LLE) [17], *Maximum Likelihood Hebbian Learning* (MLHL) [18], *Cooperative Maximum Likelihood Hebbian Learning* (CMLHL) [19], *Isometric Mapping* (ISOMAP) [20], *Curvilinear Component Analysis* (CCA) [21], *Multidimensional Scalling* (MDS) [22] y Mapas Auto-organizados (en inglés Self-Organizing Maps - SOM) [23]. Como resultado de aplicar estas técnicas de reducción de la dimensionalidad al conjunto de datos analizado, se obtiene un nuevo conjunto de datos de dimensionalidad más baja (normalmente dos dimensiones), el cual contiene la información más interesante del conjunto de datos original. Estos datos en el espacio con menor dimensionalidad son representados, con el fin de poder analizar gráficamente el conjunto de datos, visualizando su estructura, principales rasgos de los mismos y detectar posibles situaciones anómalas en los datos ('*outliers*' o valores atípicos).

El otro tipo de técnicas empleadas en el presente trabajo son las denominadas como técnicas de agrupamiento. La tarea que llevan a cabo se puede definir como la clasificación no supervisada de muestras de datos en grupos [24]. Se consigue gracias a que las técnicas de agrupamiento dividen un determinado conjunto de datos en grupos

de características similares, de acuerdo con la aplicación de diferentes medidas de "similitud". Algunas de las técnicas de agrupamiento aplicadas en esta Tesis Doctoral se enumeran a continuación. Por una parte técnicas de agrupamiento particionales [24]: *k*-means [25], SOM *k*-means [26], *k*-medoids [27] y fuzzy c-means [28]. Por otra parte, también se ha empleado el método aglomerativo jerárquico [30], diversos métodos de evaluación de agrupamiento [31] y dendrogramas [32]. Mediante la aplicación de estas técnicas de agrupamiento se consigue el reparto de las muestras del conjunto de datos original en distintos grupos de datos. El número de grupos de datos deseados se establece de dos formas: aplicando previamente las técnicas de reducción de la dimensionalidad comentadas anteriormente o aplicando las técnicas de evaluación de agrupamiento, las cuales devuelven el número de grupos óptimo para un cierto conjunto de datos. Mediante la asignación de las muestras a los grupos de datos se consigue evaluar de forma numérica las semejanzas y diferencias en las muestras de datos pertenecientes a diferentes localizaciones o climas. De esta manera es posible evaluar cómo las muestras pertenecientes a una misma localización permanecen juntas en un mismo grupo de datos, se distribuyen en distintos grupos o se mezclan con muestras de otras localizaciones, lo cual es una forma de comprobar las semejanzas y diferencias entre los distintos climas analizados o respecto a la calidad del aire en diversas estaciones de medida.

Metodología

La metodología empleada en cada uno de los trabajos que forman parte de esta tesis está explicada en los Capítulos 2 a 6, en el correspondiente apartado "Objetivos, metodología, resultados y conclusiones". Se puede definir la siguiente metodología común a todos estos trabajos:

1.- En un primer lugar se establece el caso de estudio a analizar, de acuerdo con los objetivos de la investigación. El caso de estudio puede ser de ámbito local, regional, nacional o internacional, y por otra parte puede ser de carácter meteorológico o de calidad del aire. Además, debe de ser interesante tanto por la ubicación geográfica del mismo como por el periodo de estudio llevado a cabo.

2.- Recopilación de datos. Para la aplicación de las técnicas al caso de estudio seleccionado es necesario disponer de datos a analizar, que son

preprocesados. Se ha establecido que estos deben de tener las siguientes características:

a) Deben de ser datos públicos, disponibles para cualquier usuario. De esta forma se facilita la reproducción de los experimentos por los evaluadores de los distintos trabajos.

b) Debe de ser un conjunto de datos con un tamaño suficiente, tanto en cuanto a la dimensionalidad como en cuanto a la cardinalidad.

c) Deben de recoger las variables más importantes que determinan las características meteorológicas y/o de calidad del aire que se quieren estudiar. Existen múltiples estudios que definen las variables más importantes en cuanto a calidad del aire [33] y meteorología [34]. Finalmente, los datos son normalizados entre 0 y 1 para su estudio.

d) Debe de tener una calidad mínima, fundamentalmente en relación con el número de datos nulos o corruptos. En el caso de que estos existan, se considera su estimación/corrección con distintas técnicas. Algunas bases de datos públicas identifican estos tipos de valores mediante códigos de error, mientras que en otros casos no ocurre así y ha sido necesario procesar los datos para identificar estas situaciones.

3.- Aplicación de las técnicas de *Soft Computing*, con el objetivo de llevar a cabo la visualización neuronal del conjunto de datos (técnicas de reducción de la dimensionalidad) y/o agrupamiento de los mismos (técnicas de clustering). Se aplican el mayor número posible de técnicas adecuadas a la naturaleza de los datos, para poder hacer un estudio comparativo lo más exhaustivo posible. En este paso se presta especial atención al valor de los parámetros requeridos por cada método, con una intensiva experimentación destinada a conseguir resultados coherentes con el caso de estudio analizado.

4.- En los artículos incluidos en el presente compendio sólo se muestran los mejores resultados obtenidos, dadas las restricciones de espacio. Para estos resultados, se realiza un análisis de los mismos desde el punto de vista meteorológico o de calidad del aire identificando aquellas situaciones

anómalas y dando respuesta a sus causas, así como al comportamiento general de la meteorología o calidad del aire en los periodos de tiempo seleccionados, justificando estos resultados con información pública y de calidad proporcionada por el organismo público competente.

Se lleva a cabo un análisis pormenorizado del comportamiento de las técnicas aplicadas en el caso de estudio analizado, extrayendo conclusiones asociadas tanto a la problemática medioambiental considerada como a las técnicas empleadas.

Objetivos

En el desarrollo de esta Tesis Doctoral se ha pretendido avanzar en la aplicación de técnicas inteligentes a los campos de estudio de la meteorología y de la calidad del aire.

El objetivo principal de esta tesis es demostrar la validez de las técnicas de *Soft Computing* en el análisis de conjuntos de datos multidimensionales con información medioambiental. Este objetivo general incluye los siguientes puntos:

1.- Revisar el estado del arte acerca de los trabajos existentes en el campo de la AI aplicada al análisis de condiciones medioambientales y de una forma más exhaustiva a los que utilizan las mismas técnicas de *Soft Computing* aplicadas en los casos de estudio de este trabajo. Mediante este estudio se verifica la falta de trabajos existentes similares a los expuestos en esta tesis.

2.- Buscar posibles casos de estudio interesantes tanto en el ámbito local, nacional e internacional, así como el análisis de breves periodos de tiempo y largos periodos de tiempo.

3.- Buscar las bases de datos públicas con información meteorológica y de contaminación de calidad para poder llevar a cabo estos estudios. Es importante que estos conjuntos de datos sean públicos, con el fin de que los experimentos puedan ser repetidos por los evaluadores con el fin de verificar su autenticidad.

4.- Analizar las variables meteorológicas y de contaminación más importantes que deben estar presentes en los conjuntos de datos, para obtener unos

resultados que reflejen fielmente la realidad de las condiciones medioambientales en los entornos seleccionados.

5.- Para aquellos trabajos en los que se dispone de información tanto de meteorología como de contaminación, conocer la influencia que las condiciones meteorológicas tienen en el comportamiento de los contaminantes.

6.- Aplicar el mayor número de técnicas de *Soft Computing* a los conjuntos de datos y seleccionar aquellas que mejores resultados proporcionen. Finalmente demostrar la validez de las técnicas de reducción de la dimensionalidad y agrupamiento aplicadas para cada caso de estudio, comparando detalladamente las diferencias de comportamiento de cada una de ellas.

Para cumplir con este objetivo se han aplicado las técnicas mencionadas a una serie de casos de estudio que abarquen un abanico integral de situaciones:

- Ámbito de estudio: tanto asociado a la meteorología (Capítulo II y Capítulo IV) como a la contaminación atmosférica (Capítulo III y Capítulo V)
- Intervalo temporal analizado: desde breves periodos de tiempo (Capítulo III y Capítulo IV) a largos periodo de tiempo (Capítulo II y Capítulo V).
- Ámbito geográfico: abarcando el nivel local (Capítulo III y Capítulo IV), regional (Capítulo V) y nacional (Capítulo II).
- Situación económica: en los casos de estudio que cubren un largo periodo de tiempo y se analiza la calidad del aire (Capítulo V), se han tenido en cuenta los años analizados debido a que el impacto de la crisis económica en los últimos años influye tanto en la actividad industrial como en la densidad del tráfico rodado, todo ello determinante en la calidad del aire.

En el apartado Listado de Publicaciones se recoge una serie de publicaciones vinculadas a estudios similares, no compiladas en esta tesis y que cubren otras situaciones complementarias a las expuestas.

Listado de Publicaciones

En esta apartado se listan todas las publicaciones relacionadas con el trabajo de

investigación de la presente tesis. Estas publicaciones, todas ellas internacionales, son clasificadas en cuanto a su tipo (revista científica o conferencia internacional). Aquellas publicaciones marcadas con el símbolo (*) son las contenidas en esta tesis, seleccionadas debido a su relevancia e interés.

1.- Revistas científicas indexadas en Journal Citation Reports (*):

1.- Ángel Arroyo, Emilio Corchado, Verónica Tricio, Álvaro Herrero. **Analysis of Meteorological Conditions in Spain by means of Clustering Techniques**. Journal of Applied Logic (2016). DOI: 10.1016/j.jal.2016.11.026. 2015 ISI JCR Science Edition: 0.524. 2015 ISI JCR Ranking: 13/22 (Logic) – Q3.

2.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **Soft Computing Models to Analyze Atmospheric Pollution Issues**. Logic Journal of the IGPL, 20 (4) (2012) 699-711. DOI: 10.1093/jigpal/jzr023. 2012 ISI JCR Science Edition: 1.136. 2012 ISI JCR Ranking: 1/20 (Logic) – Q1.

3.- Emilio Corchado, Ángel Arroyo, Verónica. Tricio. **Soft computing models to identify typical meteorological days.** Logic Journal of the IGPL, 19 (2) (2011) 373-383. DOI: 10.1093/jigpal/jzq035. 2011 ISI JCR Science Edition: 0.913. 2011 ISI JCR Ranking: 1/19 (Logic) – Q1.

2.- Publicaciones en congresos internacionales:

1.- Angel Arroyo, Verónica Tricio, Álvaro Herrero, Emilio Corchado. **Time Analysis of Air Pollution in a Spanish Region through k-means**. 11th International Conference on Soft Computing Models in Industrial and Environmental Applications. Advances in Intelligent Systems and Computing. Vol. 527 p. 63-72 (2016) DOI: 10.1007/978-3-319-47364-2_7.

2.- Angel Arroyo, Verónica Tricio, Álvaro Herrero, Emilio Corchado. **Neural networks for the visual analysis of regional pollution**. EUROCON 2015 - International Conference on Computer as a Tool

(EUROCON), IEEE. (2015). DOI: 10.1109/EUROCON.2015.7313793.

3.- (*) Ángel Arroyo, Emilio Corchado, Verónica Tricio, Álvaro Herrero. **Neuro-Fuzzy Analysis of Atmospheric Pollution**. 10th International Conference on Hybrid Artificial Intelligent Systems. Lecture Notes in Computer Science. Vol. 9121 p. 382-392 (2015) DOI: 10.1007/978-3-319-19644-2_32.

4.- Ángel Arroyo, Emilio Corchado, Verónica Tricio, Álvaro Herrero. **A comparison of Clustering Techniques for Meteorological Analisys**. 10th International Conference on Soft Computing Models in Industrial and Environmental Applications. Advances in Intelligent Systems and Computing. Vol. 368 p. 117-130 (2015). DOI: 10.1007/978-3-319-19719-7_11.

5.- Ángel Arroyo, Emilio Corchado, Verónica Tricio, Laura García-Hernández, Václav Snášel. **Soft Computing techniques applied to a case study of air quality in industrial areas in the Czech Republic**. 7th International Conference on Soft Computing Models in Industrial and Environmental Applications. Advances in Intelligent Systems and Computing. Vol. 188 p. 537-546 (2012). DOI: 10.1007/978-3-642-32922-7_55.

6.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **Soft Computing Models for Enviromental DataVisualization and Analysis**. 8th International Conference on Air Quality Science and Application. ISBN: 978-1907396-80-9. (2012).

7.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **A Climatologycal Analysis by means of Soft Computing Models**. 6th International Conference on Soft Computing Models in Industrial and Environmental Applications. Advances in Intelligent and Soft Computing. Vol. 87 p. 551-559 (2011). DOI: 10.1007/978-3-642-19644-7_58.

8.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **A Study of Meteorological Conditions by means of Soft Computing Models**. CMMSE 2010: Proceedings of the 10th International Conference on Mathematical Methods in Science and Engineering. ISBN: 978-84-613-5510-5. Vol. 4 p. 1050-1057 (2010).

9.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **Soft Computing Models for an Environmental Application.** 5th International Conference on Soft Computing Models in Industrial and Environmental Applications. Advances in Intelligent and Soft Computing. Vol. 73 p. 127-135 (2010). DOI: 10.1007/978-3-642-13161-5_17.

10.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **Atmospheric Pollution Analysis by Unsupervised Learning**. 10th International Conference on Intelligent Data Engineering and Automated Learning .Springer - Lecture Notes in Computer Science. Vol. 5788 p. 767-772 (2009). DOI: 10.1007/978-3-642-04394-9_94.

11.- Ángel Arroyo, Emilio Corchado, Verónica Tricio. **Computational Methods for Immision Analysis of Urban Atmospheric Pollution**. CMMSE 2009: Proceedings of the 9th International Conference on Mathematical Methods in Science and Engineering. (CMMSE 2009). ISBN:978-84-612-9727-6. Vol. 4 p. 1169-1176 (2009).

De entre todas estas publicaciones, se han seleccionado las indicadas debido tanto a su impacto como a la coherencia entre ellas. Las tres publicaciones que forman la parte principal de esta Tesis Doctoral (Parte II) cumplen con los requisitos para presentar la misma en la modalidad de compendio de publicaciones, considerando que versan plenamente en el tema recogido en el título de la misma y en el resumen de la presente Tesis Doctoral. En la primera de ellas, "***Soft computing models to identify typical meteorological days*"**, se presenta un análisis de la meteorología en un punto concreto de la población de Burgos (España). La segunda publicación, titulada "***Soft Computing Models to Analyze Atmospheric Pollution Issues*"** tiene unos objetivos más ambiciosos, combinando parámetros de calidad del aire y

meteorológicos para determinar las condiciones de calidad del aire en la ciudad de Burgos durante una semana anómala por la presencia de un largo puente festivo. Estas dos primeras publicaciones aplican únicamente técnicas inteligentes de reducción de la dimensionalidad. En la tercera publicación, titulada "***Analysis of Meteorological Conditions in Spain by means of Clustering Techniques*"** se incorpora la aplicación de técnicas de agrupamiento a un conjunto de datos más amplio, tanto en el periodo de tiempo analizado (casi una década), como en la localización geográfica de los lugares seleccionados (cuatro lugares representativos de los cuatro principales climas existentes en España).

También se ha incluido la publicación titulada ***"Neuro-Fuzzy Analysis of Atmospheric Pollution",*** no indexada por el íncide Journal Citation Reports pero publicada en la prestigiosa serie Lecture Notes in Computer Science de la editorial Springer. Se ha incluido esta publicación por suponer un importante avance en el análisis de la calidad del aire en un lugar tan representativo como la Región de Madrid y la capital de la misma, así como por combinar el uso de las técnicas de reducción de la dimensionalidad y agrupamiento, consiguiendo un análisis más completo de la información contenida en los conjuntos de datos. De esta manera se complementan todos los trabajos seleccionados.

Conclusiones y trabajo futuro

La principal conclusión extraída del trabajo asociado a esta tesis es la validez de la aplicación de técnicas inteligentes para la visualización y análisis de conjuntos de datos multidimensionales con información medioambiental.

Con respecto a las técnicas de reducción de la dimensionalidad aplicadas, la técnica de PCA se muestra como una técnica ideal para ser empleada como primer paso en el descubrimiento de una posible estructura en el conjunto de datos bajo análisis. Una vez que se tiene conocimiento de la posible estructura de los datos mediante la aplicación de la citada técnica, se pasa a aplicar el resto de técnicas de reducción de la dimensionalidad. Las técnicas que en la mayoría de los casos han ofrecido un mejor comportamiento han sido LLE y CMLHL, como se explica en la parte II. En la aplicación de estas técnicas es muy importante la adecuada selección de los parámetros, especialmente el número de vecinos en el caso de LLE y el ratio de aprendizaje y el parámetro p (parámetro asociado a la función de energía) para CMLHL. La correcta aplicación de estas técnicas permite la visualización de un mayor número de grupos en los datos de entrada, así como una mayor separación entre los mismos, respecto a los resultados obtenidos mediante PCA. Esto permite realizar un análisis mejor y más profundo del comportamiento medioambiental del conjunto de datos en el caso de estudio seleccionado.

Respecto a la aplicación de técnicas de agrupamiento a los conjuntos de datos, estas se han mostrado como una eficaz forma de analizar los conjuntos de datos de una forma numérica, en contraposición a las técnicas de reducción de la dimensionalidad que ofrecen resultados en una forma gráfica. Estas técnicas también permiten conocer exactamente el número de muestras asignadas a cada grupo e identificarlas individualmente. Para estas técnicas se hace indispensable la adecuada selección del parámetro k o número de grupos a formar. Tanto las medidas de evaluación de grupos como la aplicación de técnicas de reducción de la dimensionalidad resultan útiles para determinar el parámetro k o número de grupos deseados. La técnica de evaluación de grupos que mejores resultados ofrece es el índice de Calinski-Harabask, mientas que las técnicas de reducción de la dimensionalidad que mejores resultados ofrecen en este sentido son LLE y CMLHL. El número de grupos devuelto por estos dos métodos suele ser muy similar, pero es importante aplicar el mayor número posible de técnicas de para aproximar este valor de la forma más precisa posible. El otro parámetro clave en la aplicación de las técnicas de agrupamiento para obtener unos resultados óptimos es la adecuada selección de la medida de la distancia a aplicar. En los casos de estudio

analizados, las medidas de la distancia que mejores resultados han ofrecidos han sido las medidas euclideas, mientras que otras medidas de la distancia como la 'cosine' o 'correlation' intentan hacer un ajuste más fino en la asignación de muestras de datos a grupos pero su fiabilidad en los resultados es menor.

Entre el conjunto de técnicas de agrupamiento aplicadas, la más eficaz ha sido probablemente la clásica técnica de *k*-means, tanto en su tiempo de ejecución como en los resultados obtenidos. También las técnicas de agrupamiento difuso se pueden considerar útiles ya que proporcionan el porcentaje de asignación de cada muestra a cada grupo, permitiendo de esta manera analizar el nivel de compactación de cada grupo y el nivel de filiación de cada muestra a cada grupo de datos.

En términos generales, las técnicas jerárquicas han sido las que peores resultados han ofrecido ya que apenas subdividen el conjunto de muestras en grupos de datos. Los dendrogramas son una herramienta útil ya que permiten observar gráficamente la subdivisión del conjunto de datos en grupos, representando la separación o diferenciación existente entre los grupos y el nivel de compactación de los mismos, dado este nivel de compactación por el número de muestras que se aglutinan en la misma hoja.

Cabe destacar los similares resultados ofrecidos por ambos grupos de técnicas en los casos de estudio en los que han sido aplicadas (Capítulo II y Capítulo V), la reducción de la dimensionalidad de forma visual y el agrupamiento de forma numérica. Ambos grupos de técnicas revelan una estructura similar de los datos analizados.

Respecto al trabajo futuro, se está trabajando en un sistema híbrido que combina en varios pasos estos dos tipos de técnicas inteligentes. Mediante este sistema híbrido, se pretende conseguir una forma eficaz de identificar la estructura de datos medioambientales, de manera no supervisada, a partir de conjuntos de datos multidimensionales.

Conclusions and future work

The main conclusion derived from present thesis is the validity of the application of intelligent techniques for visualization and analysis of high-dimensional data sets with environmental information.

Regarding to the dimensionality reduction techniques applied, PCA is revealed as an ideal technique to be applied as a first step in the discovery of a possible structure in a dataset under analysis. Once a possible structure in the dataset is discovered by applying PCA, the other dimensionality reduction techniques are subsequently applied. In most case studies LLE and CMLHL have outperformed the rest of techniques, as explained in Part II. When applying these techniques, it is very important to fine tune the input parameters, especially the number of neighbors in the case of LLE and the learning rate and the p parameter (associated to the energy function) in the case of CMLHL. The proper application of these techniques let us display a larger number of groups from the input dataset, as well as wider gaps between them, improving the results obtained by PCA. This enables a better and deeper analysis of the dataset with environmental values in the selected case study.

Regarding to the application of clustering techniques to the datasets, these techniques have proved to be an effective method to analyze datasets in a numerical form, as opposed to dimensionality reduction techniques that offer graphically results. These techniques also let us know the precise number of samples assigned to each group and identify them individually. For these set of techniques, it is essential the proper selection of values for the k parameter or the number of groups to be defined. Clustering evaluation indices and the application of dimensionality reduction techniques are useful determining the k parameter or number of desired groups. Calinski-Harabask outperforms the rest of indices that have been compared, and on the other hand, the dimensionality reduction techniques with best results are LLE and CMLHL. The number of groups obtained by these two methods usually is very similar, but it is important to apply as many techniques as possible in order to obtain an accurate value. The other key parameter when applying clustering techniques for optimal results is the proper selection of the distance measure to be applied. In the analyzed case studies, the distance measure which best results were the Euclidean distance measures. Other measures such as 'cosine' or 'correlation' are aimed at a fine-grained adjustment in the process of samples allocation to groups but the reliability of the results is lower.

Among the set of applied clustering techniques, the most effective was the well-known *k*-means technique, by taking into account both the runtime and the obtained results. In addition, fuzzy clustering techniques can be considered useful as they provide the percentage of allocation of each sample to each group, thus enabling the analysis of the level of compaction of each cluster, and the level of affiliation of each sample to each cluster.

In general terms, hierarchical technique achieved the worst results as samples are not grouped into clusters. Dendrograms are useful tools as lead to graphical observation of the subdivision of the dataset into groups, representing the distance between the groups and the level of compaction between them, given this level of compaction by the number of samples gathered on the same leaf.

It is worth highlighting the similar results offered by both groups of techniques in the case studies in which they have been applied (Chapter II and Chapter V), the dimensionality reduction in a visual way and the clustering in numerical form. Both sets of techniques reveal a similar structure in the datasets.

Regarding the future work, a hybrid system which combines these two types of intelligent techniques in several steps is being developed at present time. With this hybrid system, it is expected to propose an effective process to identify the internal structure of environmental high-dimensional datasets, always in an unsupervised-learning way.

Referencias

1. Intergovernmental Panel on Climate Change. Climate Change 2014: Impacts, Adaptation and Vulnerability: Regional Aspects. Cambridge University Press. (2014). [Online]. URL: http://ipcc-wg2.gov/AR5/report/. [Último acceso: 12-Ene-2017].
2. T. Godish, W. T. Davis, S. Fu. Air quality. CRC Press. ISBN: 9781466584440 (2014).
3. CORDIS Archive: Fith Framework-Fifth RTD Framework Programme (1998-2002)-European Commission. [Online]. URL: http://cordis.europa.eu/fp5/. [Último acceso: 12-Ene-2017].
4. Gobierno de España. Redes de Calidad del aire autonómicas y locales. [Online]. URL: http://www.mapama.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/redes/ [Último Acceso: 12-Ene-2017].
5. K. Voigt, G. Welzl, R. Brüggemann. Data analysis of environmental air pollutant monitoring systems in Europe. Environmetrics. Vol. 15(6) p. 577-596 (2004).
6. I. Ross, P. J. Valdes, S. Wiggins. ENSO dynamics in current climate models: an investigation using nonlinear dimensionality reduction. Nonlinear Processes in Geophysics. Vol. 15 p. 339-363 (2008).
7. G. Chattopadhyay, S. Chattopadhyay, P. Chakraborthy. Principal component analysis and neurocomputing-based models for total ozone concentration over different urban regions of India. Theoretical and Applied Climatology. Vol. 109 (1) p. 221-231 (2011).
8. M. Vejmelka, L. Pokomá, J. Hlinka, D. Hartman, N. Jackay, M. Palus. Non-random correlation structures and dimensionality reduction in multivariate climate data. Climate Dynamics. Vol. 44 (9) p. 2663-2682 (2014).
9. I. Hokenko. On clustering of non-stationary meteorological time series. Dynamics of Atmospheres and Oceans. Vol. 49(2) p. 164-187 (2010).
10. J. Zscheischle, M. D. Mahecha, S. Harmeling. Climate Classifications: The Value of Unsupervised ClusteringProcedia Computer Science. Vol. 9 p. 897–906 (2012).
11. I. A. Pérez, M. L. Sanchez, M. A. Garcia, N. Pardo. Analysis of air mass trajectories in the northern plateau of the Iberian Peninsula. Journal of Atmospheric and Solar-Terrestrial Physics. Vol. 134 p. 9-21 (2015).
12. S. Rusell, P. Norvig. Artificial Intelligence: A modern Approach. Prentice Hall, 3rd Edition. ISBN: 978-0136042594 (2010).
13. A. G. Tettamanzi, M. Tomassini. Soft computing: integrating evolutionary, neural, and fuzzy systems. Springer Science & Business Media. ISBN 978-3-662-04335-6 (2013).
14. D. K. Pratihar. Soft computing: fundamentals and applications. Oxford, U.K.: Alpha Science International, Ltd. ISBN: 978-1842658635 (2013).
15. L. J. P. Van der Maaten, E. O. Postma, H. J. van den Herik. Dimensionality Reduction: A Comparative Review. Technical Report. Tilburg Centre for Creative Computing (2009).
16. H. Abdi, L. J. Williams. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. Vol. 2 p. 433-459 (2010).
17. X. Li, S. Lin, S. Yan, D. Xu. Discriminant locally linear embedding with high-order tensor data. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. Vol. 38 p. 342-352 (2008).
18. E. Corchado, D. MacDonald, C. Fyfe. Maximum and minimum likelihood Hebbian learning for exploratory projection pursuit. Data Mining and Knowledge Discovery. Vol. 8 (3) p. 203-225 (2004).
19. E. Corchado, Y. Han, C. Fyfe. Structuring global responses of local filters using lateral connections. Journal of Experimental & Theoretical Artificial Intelligence. Vol. 15 p. 473-487 (2003).

20. C. Shao, H. Hu. Extension of ISOMAP for Imperfect Manifolds, Journal of Computers. Vol. 7 p. 1780-1785 (2012).
21. P. Demartines. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks. Vol. 8 (1) p. 148-154 (1997).
22. F. W. Young. Multidimensional scaling: History, theory, and applications. Psychology Press. Taylor & Francis Press. ISBN-13: 978-0898596632 (2013).
23. T. Oja, M., Kaski, S., Kohonen. Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. Neural Computing Surveys. Vol. 3 (1) p. 1-156 (2003).
24. K. Aparna, M. K. Nair. Comprehensive Study and Analysis of Partitional Data Clustering Techniques. International Journal of Business Analytics. Vol. 2 p. 23-38 (2015).
25. C. Ding, X. He. K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning. ACM. p. 193-146 (2004).
26. F. Bação, V. Lobo, M. Painho. Self-organizing maps as substitutes for k-means clustering. International Conference on Computational Science. p. 476-483 (2005).
27. H. S. Park, C. H. Jun. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. Vol. 36 (2) p. 3336–3341 (2009).
28. J. C. Bezdek, R. Ehrlich, W. Full. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences. Vol. 10(2) p. 191-203 (1984).
29. S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M. P. Eccles, J. Cane, C. E. Wood. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Journal of behavioral medicine. Vol. 46 (1) (2013).
30. Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu. Understanding of internal clustering validation measures. IEEE International Conference on Data Mining. p. 911-916 (2010).
31. R. R. Sokal, F. J. Rohlf. The Comparison of Dendrograms by Objective Methods. Taxon. Vol. 11 (2) p 33-40 (1962).
32. D. Nychka, N. Saltzman. Design of Air-Quality Monitoring Networks. Case Studies in Environmental Statistics. Vol. 132 p. 51-76 (1998).
33. M. de Castro, J. Martín-Vide, S. Alonso. The climate of Spain: Past, present and Scenarios for the 21th Century. A preliminary assessment of the impacts in Spain due to the effects of climate change. Spanish Ministry of Environment. Vol. 162 (2005).

Parte II. Artículos Seleccionados

Capítulo II.

Analysis of Meteorological Conditions in Spain by means of Clustering Techniques

Autores: Ángel Arroyo[1], Emilio Corchado[2], Verónica Tricio[3] y Álvaro Herrero[1]

Afiliaciones:

[1]Departamento de Ingeniería Civil, Universidad de Burgos, Burgos, España

[2]Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España

[3]Departamento of Física, Universidad de Burgos, Burgos, España

Índice

Resumen

En este trabajo se aplican y analizan técnicas de agrupamiento para el análisis de las condiciones meteorológicas en cuatro puntos de España. Cuatro técnicas partitionales (*k*-means, *k*-medoids, SOM *k*-means y modelo de mezclas de Gaussian), junto con una técnica de agrupamiento jerárquico, con diferentes criterios de medida de la distancia, y cuatro medidas de evaluación del número óptimo de grupos (Calinski-Harabasz, Davies-Bouldin, Gap y Silhouette) son aplicadas para analizar condiciones meteorológicas. Han sido analizados datos proporcionados por la Agencia Española de Meteorología (AEMET) provenientes de cuatro estaciones de adquisición de datos. Algunas de las principales variables meteorológicas de recogida diaria por estas estaciones son estudiadas con el fin de analizar la variabilidad de las condiciones ambientales en los lugares seleccionados. Además, se pretende caracterizar las estaciones de acuerdo con su ubicación.

Respecto a la metodología empleada, el primer paso llevado a cabo ha sido la selección de cuatro localidades (Burgos, Santiago de Compostela, Almería y Palma de Mallorca) representativas de los cuatro climas típicos que se dan en España (Continental, Atlántico, Mediterráneo seco y típico Mediterráneo). Posteriormente se ha tratado de conseguir un conjunto de datos de calidad, para que los resultados sean lo más fiables y significativos posible. En este sentido, se ha tratado de que estén presentes las variables de contaminación más importantes para caracterizar la calidad del aire y que el conjunto de datos posea el menor número posible de valores nulos o corruptos. Esto llevó a tener que cambiar la elección de algunas de las localizaciones por no cumplir con esta premisa. Posteriormente se han aplicado las diferentes técnicas de agrupamiento ya mencionadas, aplicando en primer lugar las técnicas de evaluación del número de grupos. Esto ha servido de orientación para, a continuación, aplicar el resto de técnicas con los valores para el parámetro *k* devuelto por las mencionadas técnicas de evaluación de grupos. Tanto las técnicas particionales como jerárquicas se han aplicado variando las distintas funciones de medición de la distancia para los distintos valores del parámetro *k*. Finalmente se han llevado a cabo unos cálculos para obtener los porcentajes de asignación de muestras a cada una de las localizaciones y así facilitar el análisis de los resultados.

En cuanto a las condiciones meteorológicas en los cuatro lugares seleccionados y en el período de tiempo analizado, un hecho evidente que se puede destacar es la gran diferencia entre la meteorología en Burgos respecto a la existente en los otros tres lugares. En Santiago de Compostela se aprecia una meteorología diferente a de los

otros tres lugares, pero no es una diferencia tan notable como en el caso de Burgos respecto a las otras tres localizaciones. La meteorología de Palma de Mallorca y Almería son muy similares entre sí, como se evidencia en el hecho de que ninguno de los métodos aplicados ha separado las muestras de ambas localizaciones en diferentes grupos de datos. Las muestras de estos dos lugares, con diferentes climas mediterráneos, tienden a mantenerse juntas en los mismos grupos de datos con las distintas técnicas.

En cuanto al comportamiento de las técnicas de agrupamiento aplicadas, hay que destacar la conveniencia de aplicar las medidas de evaluación de grupos como un primer paso. La aplicación de las cuatro medidas principales y la comprobación de cómo tres de ellos sugieren el mismo valor para el parámetro *k*, es un resultado muy valioso de cara a aplicar posteriormente el resto de técnicas con un valor muy aproximado para este parámetro. En cuanto a la comparación de los resultados ofrecidos por las técnicas de agrupamiento, se puede concluir que *k*-means, *k*-medoids y SOM *k*-means ofrecen resultados similares, siendo el criterio de distancia seleccionado un factor clave. También se puede concluir que *k*-means es la mejor técnica en cuanto a la carga computacional. En cuanto a las distintas medidas de distancia aplicadas, las distancias euclideas resultan ser las que mejores resultados obtienen, mientras que *'cosine'* y *'correlation'* demuestran una tendencia a dividir las muestras de la misma ubicación (y clima) en más de un grupo de datos, pero no siempre de una forma óptima. GMM genera resultados que son similares a los obtenidos por las tres técnicas mencionadas anteriormente, aunque en algunos casos se han obtenido resultados inconsistentes con esta técnica, repartiendo muestras de datos del mismo clima en grupos diferentes. Cabe hacer hincapié en los resultados ofrecidos por la técnica jerárquica en comparación con las técnicas particionales. En muchos casos la técnica de agrupamiento jerárquico no muestra una respuesta satisfactoria, al no asignar correctamente muestras de datos de diferentes climas a grupos diferentes. Ninguna de las técnicas ha sido capaz de separar con precisión las muestras de los cuatro lugares en grupos separados, los cual corrobora los resultados ofrecidos por PCA en el estudio inicial de este trabajo.

Capítulo III.

Soft Computing Models to Analyze Atmospheric Pollution Issues

Autores: Ángel Arroyo[1], Emilio Corchado[2], Verónica Tricio[3]

Afiliaciones:

[1]Departamento de Ingeniería Civil, Universidad de Burgos, Burgos, España

[2]Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España

[3]Departamento of Física, Universidad de Burgos, Burgos, España

Índice

Resumen

El impacto sobre la contaminación atmosférica de un fin de semana festivo en la población de Burgos (comunidad autónoma de Castilla y León) es analizado en este artículo. El trabajo trata de diferenciar la contaminación de los días laborables de la de los días no laborables, teniendo en cuenta aspectos tales como los niveles de actividad industrial y el tráfico rodado. Los datos recogidos pertenecen a una estación de medida situada en la ciudad de Burgos, que forma parte de la red de estaciones de medición de la contaminación dentro de la región de Castilla y León.

Respecto a la metodología empleada en este trabajo, el primer paso ha sido la obtención de un conjunto de datos con información sobre la calidad del aire en un punto de la ciudad de Burgos durante un periodo de tiempo corto del año 2007. Se analiza y diferencia la calidad del aire en una semana atípica por la presencia de un importante puente festivo, respecto a otras semanas en las que no existe esta circunstancia. En primer lugar, se ha aplicado PCA a este conjunto de datos con el fin de identificar la estructura interna en los datos bajo análisis. Posteriormente se ha comparado este resultado con los obtenidos al aplicar PCA a otras semanas del mismo año y en la misma localización, semanas en las que no existe circunstancia especial, con el fin de identificar alguna diferencia significativa. Una vez analizados los resultados de esta comparación, se han aplicado el resto de técnicas de reducción de la dimensionalidad descritas en la publicación con el fin de obtener unos resultados más detallados y completar de esa manera la comparativa planteado. En la aplicación del resto de técnicas de reducción de la dimensionalidad se llevan a cabo distintos experimentos, modificando los parámetros requeridos por cada una de ellas hasta llegar a los resultados más óptimos.

Después de la aplicación de nueve técnicas de reducción de la dimensionalidad diferentes a los conjuntos de datos (sólo cinco resultados se incluyeron en el artículo por limitación en el espacio a emplear), se ha demostrado la existencia de una estructura interna en el conjunto de datos.

Como conclusiones se puede indicar que ACP proporciona una primera aproximación a la estructura interna de los datos, pero no ofrece información detallada sobre el comportamiento de la calidad del aire a lo período de tiempo analizado. Por otra parte, las técnicas MLHL, Isomap, y CCA proporcionan una interesante representación gráfica que permite la identificación de agrupaciones de los datos,

proporcionando resultados muy similares entre ellas. Estos positivos resultados se deben al hecho de que estos métodos ofrecen una subdivisión de los grupos de datos; siendo capaz de agrupar los resultados por días de la semana. CMLH es el método que ofrece unos resultados más óptimos y con mayor nivel de detalle, siendo posible identificar que los niveles más bajos de contaminación corresponden a los días incluidos en el puente festivo, en los que la actividad industrial y el tráfico rodado se reduce: jueves, viernes, sábado y domingo. Se puede concluir que la baja actividad industrial reduce el nivel de contaminación en el aire, a pesar de que pudiera haber un alto tráfico rodado por el movimiento de población durante esos días festivo.

Capítulo IV.

Soft Computing Models to Identify Typical Meteorological Days

Autores: Emilio Corchado[2] , Ángel Arroyo[1], Verónica Tricio[3]

Afiliaciones:

[1]Departamento de Ingeniería Civil, Universidad de Burgos, Burgos, España

[2]Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España

[3]Departamento of Física, Universidad de Burgos, Burgos, España

Índice

Resumen

En el presente trabajo se pretende demostrar como las técnicas de reducción de la dimensionalidad son capaces de identificar patrones para caracterizar un "día típico" en términos de sus condiciones meteorológicas. Este estudio se analizan los datos de seis variables meteorológicas observadas en la ciudad de Burgos. Los datos analizados pertenecen a una estación de medida que forma parte de la red de estaciones de meteorología de la comunidad autónoma de Castilla y León. Las muestras analizadas tienen una frecuencia de muestreo quinceminutal y abarcan un periodo de más de seis meses durante el año 2007. Mediante la aplicación de las técnicas más representativas de reducción de la dimensionalidad a un número importante de días, se trata de identificar un día típico en la ciudad de Burgos durante las estaciones de verano y otoño. La principal razón para la selección de estas dos estaciones ha sido principalmente la calidad de los datos disponibles. La fuente de datos utilizada en su versión de datos quinceminutales presenta gran cantidad de datos nulos, omitidos o corruptos. Las estaciones de otoño y verano de 2007 es el subconjunto de datos con un mayor porcentaje de datos válidos entre los inspeccionados.

Respecto a la metodología empleada, el primer paso y determinante para obtener unos resultados satisfactorios fue conseguir un conjunto de datos lo más completo posible, con el menor número de valores corruptos o inexistentes. Posteriormente ese conjunto de datos se dividió en tantos subconjuntos de datos como días disponibles. A continuación se ha aplicado la técnica de PCA a un mínimo de 3 días a la semana, durante los seis meses analizados para obtener el día típico en verano y otoño. Mediante la observación de los resultados ofrecidos por PCA, se ha estimado un patrón de comportamiento de la meteorología durante estas dos estaciones del año. Posteriormente se han aplicado las técnicas LLE y CMLHL para intentar obtener unos resultados más precisos. A la hora de aplicar LLE y CMLHL es clave la selección de los parámetros de entrada, especialmente el ratio de aprendizaje, el parámetro p asociado a la función de energía y la fuerza de las conexiones laterales en CMLHL. Los mejores resultados son aquellos que ofrecen una similitud con PCA, pero consiguiendo una mayor claridad o subdivisión en la formación de grupos de datos y en la detección de situaciones anómalas.

Después de aplicar tres métodos diferentes de reducción de la dimensionalidad para los dos casos de estudio, identificar el día típico en Burgos en las estaciones de verano y otoño, se ha demostrado la existencia de una estructura interna en los datos

analizados. CMLHL es el método más sensible, capaz de obtener una agrupación mayor y más clara de las muestras que forman los diferentes grupos, ayudando a realizar un mejor análisis de los resultados. Esta técnica subdivide los grandes grupos identificados por PCA en nuevos subgrupos, lo cual nos proporciona más información acerca del comportamiento del día típico en Burgos en las estaciones de otoño y verano.

Los "días típicos" en verano y en otoño en la ciudad de Burgos, determinados en este estudio, tienen en común la gran importancia de la radiación solar. Por el contrario, lo que difiere entre estos dos días típicos es la gran variabilidad en las condiciones ambientales que se dan en el otoño y su dificultad para encontrar el "día típico" en esa temporada. Mientras que el "día típico" en verano se caracteriza por un gran contraste entre las horas de radiación solar y las horas de noche, en el caso del otoño la radiación solar no tiene una influencia tan grande en el comportamiento de la meteorología y la identificación del llamado "día típico".

Capítulo V.

Neuro-Fuzzy Analysis of Atmospheric Pollution

Autores: Ángel Arroyo[1], Emilio Corchado[2], Verónica Tricio[3] y Álvaro Herrero[1]

Afiliaciones:

[1]Departamento de Ingeniería Civil, Universidad de Burgos, Burgos, España

[2]Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España

[3]Departamento of Física, Universidad de Burgos, Burgos, España

Índice

Resumen

En este trabajo se propone la aplicación de diferentes técnicas de reducción de la dimensionalidad y agrupamiento para la visualización y análisis de la calidad del aire en una región determinada de España (Madrid), caracterizando la calidad del aire y su evolución en el tiempo. Para ello se analizan los datos de tres estaciones de medida tanto de la red de calidad del aire de la ciudad de Madrid como de la red de control de calidad del aire de la comunidad autónoma de Madrid. Los principales contaminantes registrados por estas estaciones son estudiados con el fin de investigar cómo la ubicación geográfica y las diferentes estaciones del año son determinantes en el comportamiento de la contaminación del aire.

Respecto a la metodología empleada, primeramente se han seleccionado tres localizaciones con diferentes características (zona urbana, extraurbana y de vegetación protegida), de acuerdo con la zonificación realizada tanto para Madrid ciudad como para la región de Madrid por las administraciones competentes. Estas tres localizaciones además deben de compartir el mayor número de variables con información acerca de la calidad del aire posible, lo que representa una dificultad ya que las variables comunes a las distintas estaciones de medidas no suelen superar el número de cuatro, entre los datos disponibles. Con los datos disponibles, se han omitido o aproximado los datos con valores nulos o corruptos. Una vez los datos se encuentran en su estado final, se les ha aplicado en primer lugar ACP para hacer una primera exploración de la estructura de los mismos. Posteriormente se han aplicado otras técnicas más habituales de reducción de la dimensionalidad. Sólo se muestran en este artículo los resultados de aquellas que proporcionan unos resultados más claros y concluyentes, es decir, aquellas que permiten visualizar las características de calidad del aire de las dos estaciones con mayor claridad. Una vez que se dispone de esta visión de la calidad del aire en las tres estaciones, se han aplicado las diferentes técnicas de agrupamiento, utilizando como valores para el parámetro k el número de grupos identificado gracias a las técnicas de reducción de la dimensionalidad. Posteriormente se ha realizado un análisis estadístico del reparto de muestras en grupos obtenido por las diferentes técnicas de agrupamiento para los distintos valores de k y los distintos criterios de medición de la distancia utilizadas.

Las conclusiones sobre la calidad de la contaminación del aire en Madrid en este caso de estudio reflejan el hecho de que hay valores muy altos de NO en el centro de Madrid, especialmente durante las estaciones de otoño e invierno. La concentración

de O_3 y la velocidad del viento en el centro de Madrid son más bajos que en las otras dos estaciones, lo cual afecta a la calidad del aire de una manera negativa. Las localizaciones fuera del centro de la ciudad presentan niveles mucho más bajos de NO y mayores niveles de O_3 en verano, especialmente en las áreas de vegetación protegida. También es posible observar una evolución en los niveles de contaminación del aire a lo largo del año, mostrando valores más altos en parámetros como el NO durante las estaciones de otoño e invierno. La contaminación del aire observada en estas tres localizaciones no sufre apenas variación entre 2007 y 2008.

En cuanto a la aplicación de técnicas de reducción de dimensionalidad, CMLH ofrece un excelente resultado en la identificación de patrones internos en los datos, siendo capaz de identificar tres grupos de datos perfectamente diferenciados en el caso de proyectar acorde a la localización de la estación de medida.

Las técnicas de agrupamiento permiten encontrar relaciones entre los datos de una manera más precisa. Mediante la comparación de los métodos de agrupamiento aplicados, se puede concluir que ofrecen resultados similares. Sin embargo, fuzzy c-means, gracias a sus porcentajes de asignación de cada muestra de datos a cada grupo, da una respuesta más precisa sobre el nivel de pertenencia de cada muestra a cada grupo, lo cual proporciona una idea del nivel de compactación de los grupos.