Introducció a l'estudi dels contaminats atmosfèrics mitjançant mineria de dades

Fede Saseta Ibáñez
Grau d'Enginyeria Informàtica especialitat Computació
Intel·ligència Artificial

David Isern Alarcón
Carles Ventura Royo

14/06/2016

FITXA DEL TREBALL FINAL

Títol del treball :	Introducció a l'estudi dels contaminats atmosfèrics mitjançant mineria de dades
Nom de l'autor :	*Fede Saseta Ibáñez*
Nom del consultor/a :	*David Isern Alarcón*
Nom del PRA :	*Carles Ventura Royo*
Data de lliurament (mm/aaaa) :	*06/2016*
Titulació o programa :	*Grau Enginyeria Informàtica*
Àrea del Treball Final :	*Intel·ligència Artificial*
Idioma del treball :	*Català*
Paraules clau :	*Mineria de dades, Contaminants, Weka*

Resum del Treball :

L'objectiu d'aquest treball és poder establir les bases per realitzar mineria de dades amb l'eina de programari lliure Weka i aprofitar-lo per estudiar la problemàtica de la contaminació atmosfèrica, un dels temes més preocupants en les societats desenvolupades avui en dia.

Per fer-ho possible s'han seguit les fases d'un projecte de mineria de dades, seleccionant primer les dades provinents de les mesures públiques dels contaminants en l'aire obtingudes per la Xarxa de Vigilància i Previsió de la Qualitat de l'Aire, a continuació preparant les dades per poder abastar el màxim d'algorismes possibles, per després realitzar la mineria de dades aplicant els algorismes de Weka i finalment realitzar l'anàlisi / interpretació / avaluació de resultats i l'assimilació del coneixement extret.

Entre els resultats obtinguts destacar que s'ha realitzat una regressió lineal en la que és possible obtenir el valor d'un contaminant en funció de la resta, que amb els algorismes M5P i M5Rules s'evidencia consistència entre les dades, que amb l'aplicació de l'algorisme de Naive Bayes s'observa que per a poder aproximar un model acceptable, són necessaris tres atributs com a mínim i que amb els algorismes SimpleKMeans i MakeDensityBasedClustered es possible observar dos clústers amb valors d'ozó molt diferenciats el màxim pot ésser atribuït a moments d'estabilitat atmosfèrica i el mínim en èpoques amb moviment meteorològic.

Finalment concloure que els resultats obtinguts són un bon punt de partida per a donar llum al problema dels contaminants atmosfèrics ja sigui establint les seves relacions, així com observant la seva evolució temporal.

Abstract :

The aim of this study is to establish the bases of data mining tool Weka software and use it to study the problems of air pollution, one of the most critical issues in developed societies today.

To make this possible we have followed the stages of a mining project, first we selecting the data from public measures of pollutants in the air obtained by the Xarxa de Monitorització i Previsió de Qualitat de l'Aire. Then we preparing the data to cover the maximum possible algorithms, then we perform data mining algorithms using Weka and finally we do the analysis / interpretation / evaluation of results and the assimilation of the knowledge extracted.

Among the results highlighted that we has made a linear regression in which it is possible to get the value of a pollutant according to the other. With algorithms and M5P M5Rules demonstrates consistency between the data. With the application Naive Bayes algorithm notes that in order to approximate an acceptable model are necessary at least three attributes. With the SimpleKMeans and MakeDensityBasedClustered algorithms is possible to observe two clusters with very different ozone values, we can be attributed the maximum to moments of atmospheric stability and the minimum to weather movement ages.

Finally, we concluded that the results are a good starting point to shed light on the problem of air pollutants both establishing their relations and observing its evolution.

Índex

1. Introducció ... 1
   1.1. Context i justificació del Treball : ... 1
      1.1.1. Motivació i referències a treballs anteriors ... 6
   1.2. Objectius del Treball ... 8
      1.2.1. Objectius generals ... 8
      1.2.2. Objectius específics ... 9
   1.3. Enfocament i mètode seguit ... 9
      1.3.1. Introducció ... 9
      1.3.2. Selecció i exploració de les dades ... 10
      1.3.3. Preparació de les dades (neteja i transformació) ... 11
      1.3.4. Mineria de dades ... 12
      1.3.5. Anàlisi de resultats i assimilació del coneixement ... 14
   1.4. Planificació del treball ... 15
      1.4.1. Fites ... 15
      1.4.2. Cronograma ... 15
      1.4.3. Diagrama de Gantt ... 16
   1.5. Breu sumari de productes obtinguts ... 16
   1.6. Breu descripció dels altres capítols de la memòria ... 17
2. Selecció i exploració de les dades ... 18
   2.1. Recol·lecció de dades ... 18
   2.2. Descripció de les dades ... 21
   2.3. Exploració de les dades ... 22
   2.4. Verificació de les dades ... 23
3. Preparació de les dades ... 24
   3.1. Estadístiques i gràfics de les dades ... 24
   3.2. Neteja de dades ... 34
   3.3. Integració de les dades ... 35
   3.4. Transformació de les dades ... 36
   3.5. Reducció de les dades ... 37
   3.6. Construcció de l'arxiu .arff ... 37

4. Mineria de dades 40
4.1. El programari lliure Weka 3.8.0 40
4.2. Millora de l'arxiu .arff 42
4.3. Sub-entorns Explorer de Weka 44
4.3.1. Sub-entorn Cluster 44
4.3.2. Sub-entorn Associate 44
4.3.3. Sub-entorn Classify 44
4.4. Algorismes escollits 45
4.4.1. Algorismes de predicció 46
4.4.2. Algorismes de classificació 48
4.4.3. Algorismes de Clustering (classificació) 50
4.4.4. Algorismes d'associació 51
4.5. Mineria 52
4.5.1. Significat dels paràmetres en els resultats 52
4.5.2. Presentació dels resultats obtinguts 54
5. Conclusions 63
5.1. Conclusions dels algorismes 63
5.2. Conclusions generals 64
5.3. Recomanacions 65
6. Línies futures 66
6.1. Sobre el programari Weka 66
6.2. Sobre Medi ambient 68
6.3. Salut 69
7. Glossari 70
8. Bibliografia / Webgrafia 73
8.1. Bibliografia 73
8.2. Webgrafia 73
9. Annexos 75
9.1. Aspectes legals : Dates d'interès 75
9.2. Passos detallats de la mineria de dades amb Weka 77
9.2.1. Captures Rellevants 80

Taula d'Il·lustracions

Il·lustració 1 : El procés de la contaminació atmosfèrica. 2
Il·lustració 2 : Esquema general de la contaminació. 3
Il·lustració 3 : Influència del medi ambient en la salut humana. 8
Il·lustració 4 : El procés de la mineria de dades. 10
Il·lustració 5 : Enteniment de les dades. 10
Il·lustració 6 : Preparació de les dades. 11
Il·lustració 7 : Reducció de les dades. 12
Il·lustració 8 :Tècniques i objectius. 13
Il·lustració 9 : Models i objectius. 13
Il·lustració 10 : Conocimiento extendido y predictivo. 14
Il·lustració 11 : El procés de la mineria de dades. 14
Il·lustració 12 : Entrega de producte. 17
Il·lustració 13 : Captura de pantalla de la web de la qualitat de l'aire de la Generalitat de Catalunya. 18
Il·lustració 14 : Captura de pantalla d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya. 18
Il·lustració 15 : Captura de pantalla de la pestanya Dades històriques d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya. 19
Il·lustració 16 : Captura de pantalla de la pestanya Dades històriques amb les opcions i les dates escollides d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya. 19
Il·lustració 17 : Captura de pantalla del mapa de les zones de qualitat de l'aire de (XVPCA). 20
Il·lustració 18 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 21
Il·lustració 19 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 21
Il·lustració 20 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 21
Il·lustració 21 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 22
Il·lustració 22 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 22
Il·lustració 23 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 22
Il·lustració 24 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 22
Il·lustració 25 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 22
Il·lustració 26 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. 23
Il·lustració 27 : Captura de pantalla de la capçalera amb unes poques files de l'arxiu Excel creat. 24
Il·lustració 28 : Captura de pantalla taula resum general de l'arxiu Excel creat. 25
Il·lustració 29 Il·lustració 30 25
Il·lustració 31 Il·lustració 32 26
Il·lustració 33 Il·lustració 34 26
Il·lustració 35 Il·lustració 36 26
Il·lustració 37 : Captura de pantalla arxiu .arff. 39
Il·lustració 38 : Captura de pantalla entorns Weka. 41
Il·lustració 39 : Captura de pantalla sub-entorns Weka. 41

Il·lustració 40 : Etapes del procés d'extracció de coneixement. 42
Il·lustració 41 : Captura de pantalla segon arxiu .arff. 42
Il·lustració 42 : Captura de pantalla segon arxiu .arff. 43
Il·lustració 43 : Captura de pantalla Weka amb la majoria d'algorismes actius. 43
Il·lustració 44 : Regressió lineal múltiple. 47
Il·lustració 45 : Regressió lineal múltiple, coeficient de determinació. 47
Il·lustració 46 : Teorema de Bayes. 48
Il·lustració 47 : MAP, màxim a posteriori hipòtesi. 48
Il·lustració 48 : ML, màximum likelihood. 49
Il·lustració 49 : Simplificació. 49
Il·lustració 50 : Distància Euclidiana. 50
Il·lustració 51 : Distància Euclidiana Normalitzada. 51
Il·lustració 52 : Distància entre instàncies properes. 51
Il·lustració 53 : Mean Absolute Error. 52
Il·lustració 54 : Root Mean Squared Error. 52
Il·lustració 55 : Relative Absolute Error. 52
Il·lustració 56 : Root Relative Squared Error. 53
Il·lustració 57 : Precision. 53
Il·lustració 58 : Recall. 53
Il·lustració 59 : F-Measure. 53
Il·lustració 60 : MCC. 53
Il·lustració 61 : Captura de pantalla algorisme Apriori. 61
Il·lustració 62 : Captura de pantalla algorisme FilteredAssociator sense discretitzar. 61
Il·lustració 63 : Captura de pantalla algorisme FilteredAssociator sense discretitzar amb 10 atributs. 62
Il·lustració 64 : Captura de pantalla algorisme FilteredAssociator amb CO discretitzat. 62
Il·lustració 65 : Els diferents efectes de la contaminació atmosfèrica sobre la salut. 69
Il·lustració 66 : Captura de pantalla LinearRegression Ozó per Mes 80
Il·lustració 67 : Captura de pantalla Apriori de NO2 discretitzat amb comarca i zona redefinida 80
Il·lustració 68 : Captura de pantalla MakeDensityBasedClusterer amb les dades dels sis contaminants 81
Il·lustració 69 : Captura de pantalla NaiveBayes amb les dades dels 6 contaminants comarca i zona redefinida 82

Índex de Taules

Taula 1 : Característiques i fonts principals dels contaminants (part 1) 4
Taula 2 : Característiques i fonts principals dels contaminants (part 2) 5
Taula 3 : Principals fonts emissores antropogèniques 6
Taula 4 : Estacions d'estudi escollides segons les zones redefinides incorporant la zona oficial 20
Taula 5 : Rang de valors numèrics dels contaminants 35
Taula 6 : SimpleLinearRegression 54
Taula 7 : LinearRegression 54
Taula 8 : M5P 55
Taula 9 : M5Rules 55
Taula 10 : NaiveBayes amb dos atributs (1) 55
Taula 11 : NaiveBayes amb dos atributs (2) 55
Taula 12 : NaiveBayes amb dos atributs (3) 56
Taula 13 : NaiveBayes amb tres atributs (1) 56
Taula 14 : NaiveBayes amb tres atributs (2) 56
Taula 15 : NaiveBayes amb tres atributs (3) 56
Taula 16 : NaiveBayes amb tres atributs (4) 56
Taula 17 : NaiveBayes amb tres atributs (5) 57
Taula 18 : NaiveBayes amb tres atributs (6) 57
Taula 19 : J48 amb dos atributs (1) 57
Taula 20 : J48 amb dos atributs (2) 57
Taula 21 : J48 amb dos atributs (3) 57
Taula 22 : J48 amb dos atributs (4) 58
Taula 23 : SimpleKMeans (1) 58
Taula 24 : SimpleKMeans (2) 58
Taula 25 : SimpleKMeans (3) 58
Taula 26 : SimpleKMeans (4) 59
Taula 27 : SimpleKMeans (5) 59
Taula 28 : SimpleKMeans (6) 59
Taula 29 : MakeDensityBasedClustered (1) 59
Taula 30 : MakeDensityBasedClustered (2) 60
Taula 31 : MakeDensityBasedClustered (3) 60
Taula 32 : MakeDensityBasedClustered (4) 60
Taula 33 : MakeDensityBasedClustered (5) 60
Taula 34 : MakeDensityBasedClustered (6) 61
Taula 35 : Simbologia dels contaminants 72
Taula 36 : Unitats de mesura 72
Taula 37 : Weka amb detall (1) 77
Taula 38 : Weka amb detall (2) 78
Taula 39 : Weka amb detall (3) 79

1. Introducció

Cada persona respira diàriament de mitjana entre 14 i 18 kg d'aire, mentre que tan sols consumeix de 1,5 a 2 kg d'aigua d'una forma o altra i no més de 0,7 kg de matèria sòlida seca com a aliment. Només pot viure uns minuts sense aire, mentre que pot mantenir-se viva durant dies sense beure aigua i durant setmanes sense ingerir aliments.

És per aquesta raó, perquè l'aire és un mitjà imprescindible per al desenvolupament de la vida al planeta, que el manteniment de la seva qualitat és de suma importància.

Quan es parla de qualitat de l'aire s'ha de tenir en compte que la seva contaminació és un procés que s'inicia a partir de les emissions a l'aire des dels diferents focus emissors de contaminants a l'atmosfera i que aquesta ha estat el tipus de contaminació més ignorada, ja que :

- Fins ben entrat el segle XX només se'n consideraven els fums.
- L'augment de contaminació i l'aparició de nous contaminants es dóna amb l'increment dels derivats del petroli, especialment la gasolina.
- Fins la dècada dels 60 la contaminació es contemplava com a problema local, proper a les fonts de contaminació.

L'atmosfera és un medi fluid amb una dinàmica que fa que la dispersió i el transport dels contaminants siguin difícils d'estudiar i de preveure. Així la relació entre la quantitat de contaminants emesos a l'aire i la presència d'aquests a l'aire en un moment i en un lloc determinat no és una relació directa ni proporcional ni senzilla de conèixer ja que l'atmosfera és un sistema complex amb un comportament caòtic, d'aquí les dificultats en l'evolució dels conceptes de contaminació, dels mecanismes de transmissió, i els acords en les solucions.

1.1. Context i justificació del Treball :

La contaminació atmosfèrica és un dels riscos ambientals més freqüents en totes les poblacions del planeta, afecta a tota la gent, des del seu naixement fins a la seva mort. Tot i les lleis i regulacions sobre les emissions contaminants, aquestes, no han parat de créixer en els darrers anys, a causa de l'augment en la quantitat de vehicles que circulen i del nombre d'indústries que s'estableixen i es desenvolupen dins o a prop d'aquestes poblacions.

La definició legal de contaminant segons el B.O.E. 290 (3-12-1976) és tota aquella partícula sòlida, líquida o gasosa continguda en l'atmosfera, que no forma part de la composició normal de l'aire o que està present en quantitat anormal.

Un contaminant és el nom que rep tota substància aliena a la composició de l'atmosfera que passa a ella i roman durant un cert temps. També s'inclou dins d'aquesta categoria totes aquelles substàncies que conformen l'atmosfera però que es presenten en concentracions superiors a les naturals.

Per tant direm que l'atmosfera està contaminada quan s'hi emeten, de forma natural o per acció de l'home, substàncies alienes a la seva composició normal.

Es coneix per **emissió** l'alliberament de substàncies a l'atmosfera a partir d'un punt concret.

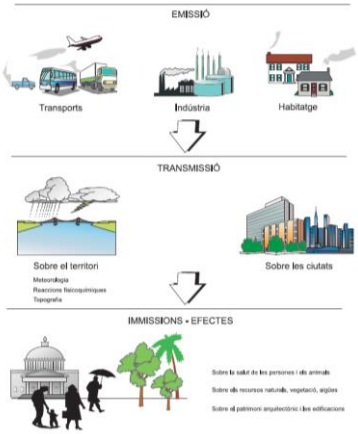
Per contra, **immissió** és el nivell de substàncies contaminants que podem mesurar en un punt concret de l'espai, independentment de la font d'on prové.

Conèixer el nivell d'immissió dels diferents contaminants és conèixer el nivell de qualitat de l'aire.

Els efectes de la contaminació de l'aire poden ser crònics o aguts. Els primers són aquells associats a rebre l'impacte de nivells d'immissió moderats durant llargs períodes de temps. Els segons són aquells produïts en rebre l'impacte de nivells d'immissió molt alts durant curts períodes de temps, és a dir, com a conseqüència d'episodis de contaminació.

La valoració dels efectes de cada contaminant és un dels criteris fonamentals a l'hora d'establir les normes de qualitat o els nivells permesos.

En la il·lustració següent es mostren totes les fases del procés de la contaminació atmosfèrica, fonts d'emissió, transmissió sobre el territori i les ciutats i els seus efectes sobre la salut de les persones i els animals, sobre el recursos naturals, vegetació i aigües, i sobre el patrimoni arquitectònic i les edificacions:



Il·lustració 1 : El procés de la contaminació atmosfèrica.
Font : La Qualitat de l'aire a Catalunya 2000-2001, Direcció General de Qualitat Ambiental, Generalitat de Catalunya

Cal tenir en compte també que la naturalesa i l'estructura de les fonts contaminants és decisiva pel que fa als efectes que puguin produir posteriorment, és a dir, que els mateixos contaminants, emesos d'una manera o d'una altra, poden tenir efectes molt diferents.

Hi ha multitud de substàncies contaminants i de fonts de contaminació. Pel que fa a les substàncies encara que les veurem més endavant amb profunditat cal esmentar aquelles que destaquen per la seva importància quantitativa o qualitativa o pels seus efectes. Així podem parlar sobretot de compostos de carboni, compostos de sofre, compostos de nitrogen, ozó i oxidants, hidrocarburs i partícules, també de metalls, compostos halogenats, contaminants biològics, compostos orgànics i altres. Els mateixos contaminants, emesos d'una manera o altra, poden tenir efectes molt diferents.

Els contaminants atmosfèrics més importants, segons la seva composició química, són :

- Partícules : Segons la mida :
  - Sedimentables (> 30 µm).
  - Partícules en suspensió (< 30 µm).
  - Partícules respirables (< 10 µm).
  - Fums (< 1 µm).
- Compostos de sofre : SO_2, H_2S, H_2SO_4 mercaptans, sulfurs.
- Compostos de nitrogen : NO, NO_2, NOx, NH_3.
- Compostos de carboni : CO, CO_2, CH_4, HCT.
- Halògens i compostos halogenats : Cl_2, HCl, HF, CFC.
- Oxidants fotoquímics : O_3, peròxids, aldehids.

D'altra banda, els contaminants atmosfèrics també es poden classificar segons la seva procedència de la manera següent :

- Contaminants Primaris : Procedents directament de fonts d'emissió fixes o mòbils, que es poden trobar amb la mateixa forma química en els focus emissors (per exemple : SO2, H2S, NO, NH3, CO,CO2, HCl, HF, PST ...).
- Contaminants Secundaris : Originats en l'atmosfera mateixa, com a conseqüència de transformacions de contaminants primaris; és a dir, no es poden trobar amb la mateixa forma química en els focus emissors (per exemple : O3, SO3, H2SO4, NO2, HNO3 ...).

L'activitat humana, especialment l'activitat associada a l'emissió de contaminants, comporta efectes preocupants sobre el medi ambient :

- Efectes sobre la salut.
- Efectes sobre la seguretat viària.
- Efectes sobre els materials.
- Efectes econòmics.
- Efectes sobre el clima.

La contaminació de l'aire produeix una notable deterioració de l'atmosfera, sobretot de la capa fronterera o capa d'ozó.

La contaminació atmosfèrica està causada principalment per les emissions incontrolades de fums que genera l'activitat industrial i, sobretot als nuclis urbans, per l'ús de certs mitjans de transport i de determinades calefaccions.

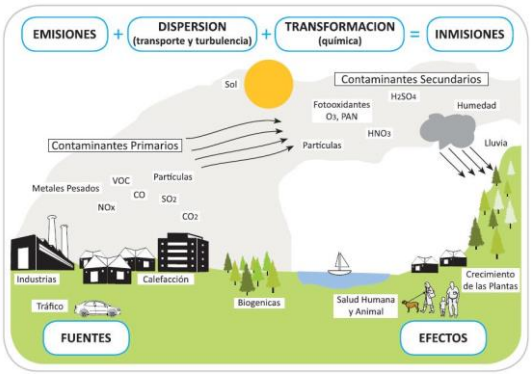
Una vegada a l'atmosfera, els contaminants són transportats pel vent o la pluja o es transformen en altres substàncies. Com a resultat d'aquests canvis, cada punt de l'atmosfera té un nivell d'immissió, és a dir, un nivell de contaminació diferent.

Els nivells d'immissió mesuren la qualitat de l'aire i, per tant, determinen l'efecte d'un contaminant sobre la salut i el medi ambient.

Pel que fa a les fonts de contaminació les podem diferenciar de manera senzilla com:

- Fonts naturals, són aquelles en les que l'home no hi té intervenció i no és capaç d'evitar-ne la contaminació.
- Fonts naturals accelerades per l'activitat humana.
- Fonts antropogèniques : És sobre aquestes fonts que es pot intervenir de forma més eficaç per tal d'evitar o minimitzar la contaminació atmosfèrica.

En la il·lustració següent es mostra visualment un esquema de la contaminació :



Il·lustració 2 : Esquema general de la contaminació.
Font : Grupo de Estudios en Sostenibilitat Urbana y Regional (SUR) Universidad de los Andes

En la taula següent es mostren les característiques i les fonts principals dels contaminants més importants que es mesuren actualment :

SO2 (diòxid de sofre)	
Característiques	**Fonts emissores antropogèniques**
• Gas incolor i d'olor forta i sufocant. • En una atmosfera humida es transforma en àcid sulfúric i causa la deposició àcida. • A partir de concentracions >0.1 ppm es produeix una important reducció de la visibilitat.	• Refineries de petroli. • Transport: Vehicles de gasoil. • Centrals tèrmiques. • Combustió de carburants : Carbó, fueloil. • Cimenteres. • Incineració de residus.
NO (òxid nitrós)	
Característiques	**Fonts emissores antropogèniques**
• Gas incolor, poc soluble en l'aigua i difícilment liquable. • És lleugerament més dens que l'aire i soluble en sulfur de carboni. • Es combina ràpidament amb l'oxigen de l'aire per formar NO2, fins i tot a temperatura ambient.	• Transport. • Automobilisme. • Elaboració de productes químics.
NO2 (diòxid de nitrogen)	
Característiques	**Fonts emissores antropogèniques**
• Gas de color amarronat i d'olor irritant. • Tòxic a altes concentracions. • Intervé en la formació de la boira fotoquímica.	• Transport. • Centrals tèrmiques. • Combustió de carburants : Gas natural, carbó, fueloil, líquids i sòlids. • Incineradores. • Cremacions agrícoles. • Cimenteres. • Fàbriques de vidre. • Refineries.
O3 (ozó)	
Característiques	**Fonts emissores antropogèniques**
• Gas incolor i d'olor agradable. • Molt oxidant i irritant.	• És un contaminant secundari, és a dir, no és emès per cap focus. • D'origen fotoquímic, és a dir, es forma per l'acció de la llum solar i en presència d'òxids de nitrogen i hidrocarburs.
H2S (sulfur d'hidrogen)	
Característiques	**Fonts emissores antropogèniques**
• Gas incolor i amb forta olor (olor a ous podrits). • Límit olfactible molt baix (a partir de 2 ppb). • Tòxic a altes concentracions i a exposicions curtes de temps.	• Fabricació de pasta de paper. • Refineries. • Indústria de curtits i colorants. • Depuradores d'aigües residuals i clavegueram.
CO (monòxid de carboni)	
Característiques	**Fonts emissores antropogèniques**
• Gas inodor i incolor. • Tòxic a altes concentracions i a exposicions curtes de temps. • Gran indicador del trànsit.	• Transport : Principalment vehicles de gasolina. • Centrals tèrmiques. • Combustió de carburants : Gas natural, líquids i sòlids • Incineració de residus. • Cremacions agrícoles.
Benzè	
Característiques	**Fonts emissores antropogèniques**
Es un hidrocarbur aromàtic, de fórmula molecular C6H6. Es un líquid incolor a temperatura ambient. Es un compost orgànic volàtil.	• Transit i altres font mòbils. • Producció d'energia. • Calefacció domèstica.
Cl2 (clor)	
Característiques	**Fonts emissores antropogèniques**
• Gas de color groc-verdós i d'olor sufocant. • Tòxic a altes concentracions.	• Petroquímiques. • Indústria química.

Taula 1 : Característiques i fonts principals dels contaminants (part 1)

HCl (clorur d'hidrogen)

Característiques	Fonts emissores antropogèniques
• Gas incolor d'olor intensa i irritant.	• Petroquímiques. • Indústria química. • Processos de neteja i decapat de metalls. • Incineradores.

Hidrocarburs Aromàtics Policíclics

Característiques	Fonts emissores antropogèniques
Constitueixen un grup de compostos que es caracteritzen per tenir dos o mes anells aromàtics formats íntegrament per carboni i hidrogen. Les propietats fisicoquímiques varien considerablement i depenen de cada compost en concret. Algun són semivolàtis la qual cosa fa que es distribueixin entre l'aire, l'aigua i el sòl seguint fenòmens de deposició i revolatilització.	• Industrials: processos de fosa, producció d'alumini primari i conservació de la fusta. • Domèstiques: associada a combustibles fòssils (fusta i carbó). • Trànsit. • Agricultura.

Material particulat

Característiques	Fonts emissores antropogèniques
• Matèria en suspensió a l'aire. • PST: partícules de diàmetre <30µm. • PM10:partícules de diàmetre <10µm. • FN:(fums negres) partícules de diàmetre <1µm.	• Centrals tèrmiques. • Processos de foneria. • Processos de molturació. • Incineradors. • Plantes asfàltiques. • Fàbriques de vidre. • Fàbriques de ceràmica. • Combustió de carburants : Carbó, fueloil, gas natural, fusta. • Transport : Principalment vehicles de gas-oil. • Cimenteres i mineries. • Extracció d'àrids.

Metalls (Ni, Cd, As, Pb)

Característiques	Fonts emissores antropogèniques
Es determina el Níquel (Ni), Cadmi (Cd) Arsènic (As) i Plom (Pb) presents en la fracció PM10 del material particulat.	
1- Plom: és un element d'alta densitat, flexible i mal·leable. Presenta un gran nombre de aplicacions industrials, tant en la seva forma elemental com els seus compostos i aliatges.	1 - Plom • Mineria. • Fosa, producció, ús, reciclatge i eliminació de productes amb plom. • Crema de fusta i combustibles fòssils.
2- Níquel: és un element molt abundant en el nucli de la Terra però menys abundant en l'escorça terrestre. Presenta un gran nombre de aplicacions industrials.	2- Níquel • Combustió. • Operacions metal·lúrgiques a alta temperatura. • Operacions de producció de níquel primari.
3- Cadmi: és un element poc abundant en l'escorça terrestre. Es presenta associat principalment amb zinc, plom i coure i es produeix principalment com a subproducte de la indústria del zinc.	3- Cadmi • Combustió • Processos de producció de zinc, coure i plom. • Restes d'incineració. • Producció de ferro i acer.
4- Arsènic: presenta una gran varietat de compostos inorgànics i orgànics.	4- Arsènic • Transit i altres font mòbils. • Producció d'energia. • Calefacció domèstica.

Taula 2 : Característiques i fonts principals dels contaminants (part 2)

Tant aquesta taula com la següent en que es mostra el resum de les principals fonts emissores contaminants a l'aire i els contaminants més significatius que emeten, vénen avalades per la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica de Catalunya (XVPCA), la Direcció General de Qualitat Ambiental del Departament de Medi Ambient de la Generalitat de Catalunya, que es pot trobar en el següent enllaç :

http://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/avaluacio_qualitat_aire_catalunya_altres/Informes/

Principals fonts emissores antropogèniques	**SO2**	**NO2**	**CO**	**H2S**	**COV's**	**HCl**	**Cl2**	**PST**	**Pb**	**Altres metalls pesants**
Centrals tèrmiques	x	x	x					x		
Cimenteres	x	x	x					x		
Cremacions agrícoles			x					x		
Depuradores d'aigües residuals				x	x					
Extracció d'àrids i mineria								x		
Fàbriques de ceràmica			x					x	x	
Fàbriques de vidre	x	x	x					x		x
Fabricació de pintures					x					
Fabricació de pasta de paper				x				x		
Foneries								x	x	x
Incineradores	x	x	x			x		x		x
Indústria de curtits				x	x					
Indústria química				x	x	x	x			
Indústria que utilitza dissolvents					x					
Plantes asfàltiques								x		
Processos de combustió: • gas natural		x	x					x		
• combustibles líquids i sòlids	x	x	x					x		
Processos de molturació								x		
Refineries	x	x	x	x	x			x		
Transport: • gasolina		x	x		x			x		
• gasoil	x	x			x			x		
• GLP		x	x		x			x		
• GN		x	x					x		
• Biodièsel		x	x		x			x		
• Bioetanol		x	x					x		

Taula 3 : Principals fonts emissores antropogèniques

Cal tenir en compte, que això no exclou l'existència d'altres fonts emissores i emissions de més contaminants de forma menys important.

Així doncs, es fa important conèixer l'evolució temporal que han patit els diferents contaminants, així com establir les diferents relacions que pugui haver-hi entre ells.

1.1.1. Motivació i referències a treballs anteriors

La motivació d'aquest treball és poder combinar els coneixements estadístics adquirits amb els meus estudis anteriors de matemàtiques i els coneixements de mineria de dades de l'especialitat de computació d'aquest grau d'enginyeria informàtica per extreure informació i poder establir les bases per la millora del medi ambient.

En els següents enllaços s'observa l'interès creixent en l'aplicació de la mineria de dades a qüestions mediambientals.

a) U-Air: When Urban Air Quality Inference Meets Big Data
http://131.107.65.14/pubs/193973/U-Air-KDD-camera-ready.pdf

En aquest article es fa una anàlisi de la qualitat ambiental de l'aire en les ciutats, basada en les dades proporcionades per les diferents estacions de control i mesurament instal·lades a la ciutat de Pequin.

Per a la realització d'aquest estudi s'han inferit dades en temps real i recol·lectades posteriorment. Aquestes dades no solament es refereixen a dades de contaminants pròpiament dits, sinó de tots aquells factors que afectin a aquests contaminants com són el nivell de trànsit a la ciutat, la meteorologia... .

El fet d'escollir aquest article en el treball és a causa de la utilització d'informació facilitada per estacions de mesura i control semblants a les utilitzades en el present treball.

Així mateix s'empren dos tipus de classificadors en les dades, un primer en xarxes neuronals artificials per a relacionar dades espacials, i un altre basat en una cadena lineal amb camp aleatori condicional (CRF), que implica les característiques relacionades temporalment.

Tot i que en el treball actual no s'utilitzin aquests tipus de classificadors l'article aporta informació del tractament i font de dades en un estudi de tractament de dades en ambient urbà.

Cal tenir en compte que una de les localitzacions escollides en el present TFG és un ambient urbà per a poder-ne determinar o observar algun tipus de tendència en els contaminants.

b) Environmental Monitoring Using Big Data
http://axibase.com/environmental-monitoring-using-big-data/

El present article presenta un estudi mitjançant anàlisi i tractament de dades emprant la tecnologia Big Data per a conèixer l'estat de la contaminació en una localitat i establir-ne una previsió per als dies següents.

Està pensat per a viatgers que han d'anar a grans ciutats amb alts nivells de pol·lució i veure'n la situació actual i quina n'és la tendència.

Per a fer aquesta anàlisi els autors utilitzen el programa estadístic R, amb el seu paquet per a anàlisi de Big Data.

La utilitat d'aquest article en el TFG és que explica el procediment, així com les tècniques emprades en un estudi d'aquest tipus. Es podria dir que pot ésser emprat com a protocol d'estudi per a contaminants.

Les dades són extretes de diferents organismes públics que disposen d'estacions de mesura de contaminants, igual com es fa en aquest TFG.

Esmentar el cas de la referència provinent de Beijing (apartat a) que és una de les ciutats més contaminades del món.

Es per això que es fa necessari establir les bases d'un estudi exhaustiu de totes les dades recollides pels sensors ubicats en una zona d'un territori, podent ser extrapolable a qualsevol país amb varietat d'entorns mediambientals i de focus contaminant.

1.2. Objectius del Treball

L'objectiu del treball és usar les dades públiques dels contaminants en l'aire obtingudes per la Xarxa de Vigilància i Previsió de la Qualitat de l'Aire (XVPCA), a partir d'equips automàtics i manuals (web de la Generalitat de Catalunya http://dtes.gencat.cat/icqa/), per tractar-les amb programari lliure de mineria de dades i d'aquesta manera estudiar l'evolució d'aquests contaminants, així com, les diferents relacions que pugui haver entre ells.

Més concretament, l'objectiu és emprar el màxim possible d'algorismes de mineria de dades, per tractar les dades preparades amb antelació i establir les bases per poder agrupar objectes semblants, classificar objectes, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants a Catalunya.

Finalment, es tracta també d'esbrinar si els contaminants estan a prop o superen el llindar establert per les agències de protecció mediambiental i si és així, parlar breument de les conseqüències que poden portar tant sanitàriament com sòcio-econòmicament sense confrontar les dades dels contaminants de la Generalitat de Catalunya amb les d'altres fonts com el número d'ingressos hospitalaris en una zona o període donat, doncs en els últims estudis epidemiològics que s'han fet, la Generalitat de Catalunya relaciona la morbiditat i/o mortalitat amb el temps d'exposició, més que amb els nivells d'exposició com es pot veire en aquest enllaç :

http://www.diba.cat/es/web/salutpublica/qualitat-aire-exterior dient textualment :

"... **que no hi ha nivells segurs de contaminació atmosfèrica, que l'exposició crònica a contaminants pot ser més perjudicial per a la salut que l'exposició aguda i que millorar la qualitat de l'aire en una determinada zona es tradueix de seguida en una reducció de les taxes de mortalitat i morbiditat**".

Il·lustració 3 : Influència del medi ambient en la salut humana.
Font : Revista Medi Ambient. Tecnologia i Cultura. Núm. 31

1.2.1. Objectius generals

- Objectius de caire acadèmic :
  - Millora en l'aprenentatge de la preparació de les dades.
  - Millora en l'aprenentatge de la tria d'algorismes de mineria de dades.
  - Millora en l'aprenentatge del programari necessari per dur a terme el tractament de les dades obtingudes.
  - Millora en la interpretació dels resultats obtinguts en el tractament de les dades.

- Objectius de caire medi ambiental :
  - Evolució dels contaminants en els darrers temps (període 2005-2015).
  - Establir les bases per poder agrupar objectes semblants, classificar objectes, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants a Catalunya.

- Objectius sòcio-econòmics :
  - Implicació dels contaminants en els aspectes sanitaris.
  - Implicació dels contaminants en els aspectes socials.
  - Implicació dels contaminants en els aspectes econòmics.

1.2.2. Objectius específics

- Agafar les dades de l'històric de la web de la Generalitat de Catalunya dels contaminants entre el 2005 i el 2015 de 17 estacions meteorològiques representatives entre les 76 existents, de totes les zones del territori català.
- Preparar les dades per poder extreure resultats de :
  - Quatre contaminants principals :
    - CO (Monòxid de carboni).
    - H2S (Sulfur d'hidrogen).
    - NO2 (Diòxid de nitrogen).
    - SO2 (Diòxid de sofre).
  - Dos contaminants secundaris però molt presents en les dades :
    - NO (Òxid nitrós).
    - O3 (Ozó).
- Preparar les dades per poder extreure resultats segons:
  - Dia.
  - Mes.
  - Any.
  - Població.
  - Comarca.
  - Zona.
  - Contaminant.
- Millorar en la utilització del programari lliure Weka.
- Provar un gran nombre d'algorismes de tots els models per veure la resposta de les dades i establir les bases per poder agrupar, classificar, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants.
- Evolució temporal i realització d'inferència per als diferents contaminants seleccionats.
- Seguiment i predicció dels nivells de contaminants.
- Establir les bases per poder agrupar objectes semblants, classificar objectes, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants.

1.3. Enfocament i mètode seguit

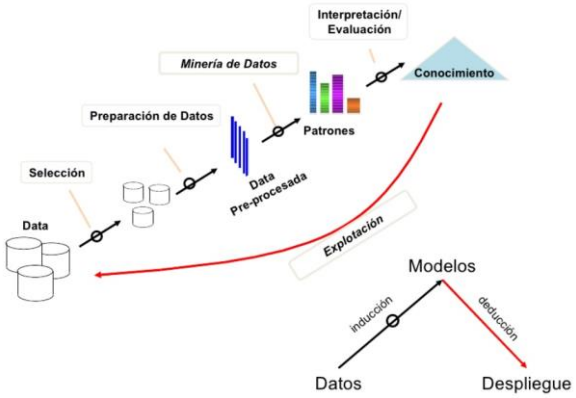
1.3.1. Introducció

En aquest apartat primerament es presenten vàries definicions del que s'entén per mineria de dades, tot seguit, es poden veure en una imatge les fases d'un projecte de mineria de dades, posteriorment comentades de forma teòrica i al final de cadascuna d'elles es fa una petita valoració del què es farà en aquest punt del projecte.

La mineria de dades es pot definir, entre d'altres, com :

- L'extracció no trivial d'informació implícita, desconeguda prèviament, i potencialment útil des de les dades.
- El procés d'extracció i refinament de coneixement útil des de grans bases de dades.
- El procés d'extracció d'informació prèviament desconeguda, vàlida i processable des de grans bases de dades per després ser utilitzada en la presa de decisions.
- L'exploració i anàlisi, a través de mitjans automàtics i semiautomàtics, de grans quantitats de dades amb la finalitat de descobrir patrons i regles significatius.
- El procés de plantejament de diferents consultes i extracció d'informació útil, patrons i tendències prèviament desconegudes des de grans quantitats de dades.
- El procés de descobrir models en les dades.

En la il·lustració següent, es mostren les fases d'un projecte de mineria de dades que són : la selecció de les dades, la preparació de les dades, la mineria de dades i l'anàlisi / interpretació / avaluació de resultats i l'assimilació del coneixement extret. Aquests passos es realitzen en l'ordre en què apareixen, essent el procés altament iteratiu, establint-se retroalimentació entre els mateixos.

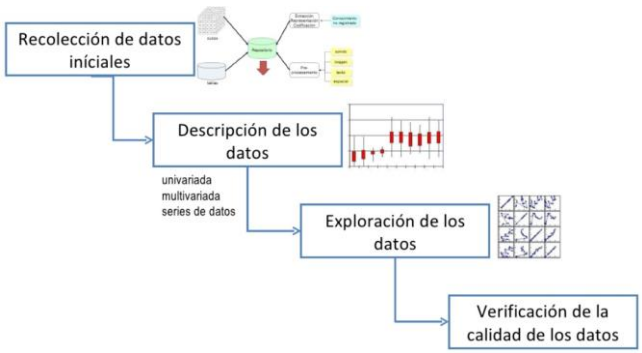


Il·lustració 4 : El procés de la mineria de dades.
Font : Mg. Samuel Oporto Díaz http://www.wiphala.net/oporto

1.3.2. Selecció i exploració de les dades

En aquesta fase, s'extreuen les dades necessàries que són la matèria primera del projecte de mineria de dades.

En la següent il·lustració es mostren totes les etapes de la fase de selecció i exploració de les dades que són: la recol·lecció de dades, la descripció de les dades, l'exploració de les dades i la verificació de la qualitat de les dades.



Il·lustració 5 : Enteniment de les dades.
Font : Mg. Samuel Oporto Díaz http://www.wiphala.net/oporto

En aquesta etapa es determinen les fonts de dades i el tipus d'informació a emprar. És l'etapa on les dades rellevants per l'anàlisi són extretes des de la o les fonts de dades.

- Recol·lecció de dades : Seleccionar i agrupar les dades necessàries pel projecte de mineria de dades extraient-les de les diferents fonts.
  - Les fonts poden ser variades :
    - Bases de dades generades en l'organització.
    - Històrics de bases de dades.
    - Bases de dades transaccionals.
    - Data Warehousing.
- Descripció de les dades : Descriure les dades recol·lectades de les diferents fonts per saber què es té, en què es compta per començar el projecte de mineria de dades.
- Exploració i verificació de les dades : Fer un anàlisi de les dades recol·lectades i descrites anteriorment, verificant-ne la validitat i qüestionant-se :
  - Si les dades estan completes.
  - Si són correctes o contenen errors.
  - Si hi ha errors, si són molt comuns.
  - Si hi ha valors omesos en les dades, on passa, com estan representats... .

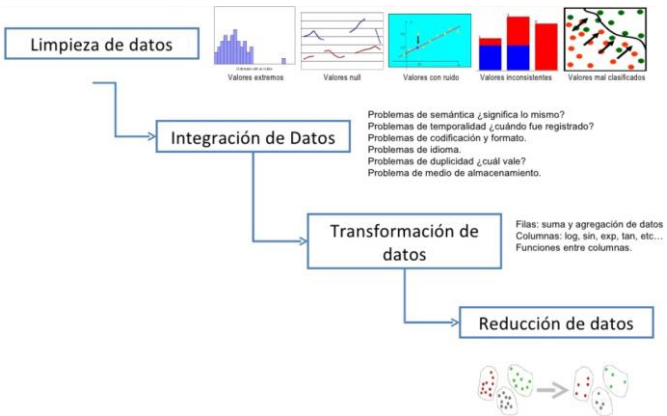
Les dades d'aquest projecte, provindran de les dades públiques de les mesures dels contaminants en l'aire obtingudes per la Xarxa de Vigilància i Previsió de la Qualitat de l'Aire (XVPCA), a partir d'equips automàtics i manuals (web de la Generalitat de Catalunya http://dtes.gencat.cat/icqa/), aquesta, conté la possibilitat de descarregar arxius tipus Excel, de com a màxim un any de durada, amb l'històric de les mesures automàtiques (mitjanes) diàries d'alguns dels contaminants següents : *CO (mg/m³), H2S (µg/m³), NO (µg/m³), NO2 (µg/m³), O3 (µg/m³), PM10 (µg/m³), PST (µg/m³) i SO2 (µg/m³)*, en cadascun d'aquests arxius, poden no estar continguts tots els contaminants esmentats.

1.3.3. Preparació de les dades (neteja i transformació)

És la fase on es preparen les dades seleccionades, estudiant la seva qualitat i determinant les operacions de neteja i/o transformació que es poden realitzar, per poder aplicar els mètodes o eines que permetin construir el model o models desitjats.

En finalitzar aquesta fase cal assegurar que les dades tenen la qualitat suficient, que són les necessàries i que tenen la forma adient.

En aquesta il·lustració es mostren totes les etapes de la fase de preparació de les dades (neteja i transformació) que són : la neteja de dades, la integració de les dades, la transformació de les dades i la reducció de la dades.

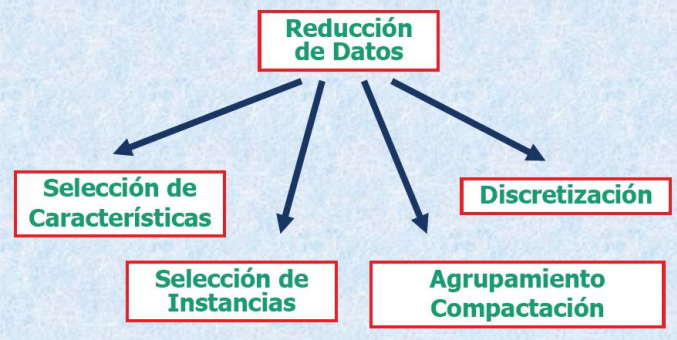


Il·lustració 6 : Preparació de les dades.
Font : Mg. Samuel Oporto Díaz http://www.wiphala.net/oporto

Les tècniques emprades per desenvolupar aquesta fase són :

- Neteja de dades : Eliminar dades errònies o redundants.
- Integració de les dades : Homogeneïtzar totes les dades.
- Transformació de les dades : Modificar les dades a una forma més adient perquè puguin ser emprades pels diferents models.
- Reducció de les dades : Reducció de la dimensionalitat, si s'escau, de les dades per obtenir els mateixos resultats.

En certes ocasions es fa necessari disminuir la mida del conjunt emmagatzemat objectiu de l'estudi, les tècniques més habituals en la reducció de les dades es mostren en la següent il·lustració :



Il·lustració 7 : Reducció de les dades.
Font : J.L. Cubero, F. Berzal, F. Herrera Dpto. Ciencias de la Computación e I.A. Universidad de Granada

Les dades d'aquest projecte (emeses en fitxers de tipus Excel), només contenen, bàsicament, la data en què es realitza la mesura i les mitjanes diàries dels contaminants, això comporta una tasca important en la preparació d'aquestes, per poder extreure informació entre el 2005 i el 2015 relacionada amb els contaminants.

En aquest cas es crea un fitxer Excel que les contingui totes i s'afegeixen uns quants camps per poder extreure resultats per dia, mes, any, població, comarca, zona, a més de dissenyar una estratègia en el tractament de dades inexistents, ja que en segons quines estacions i en segons quins contaminants hi ha una manca parcial o absoluta d'informació.

1.3.4. Mineria de dades

En aquesta fase, es fa un tractament automatitzat de les dades seleccionades amb una combinació apropiada d'algorismes.

Una tècnica o model (Models d'agregació, Models d'associació, Regles de classificació, Arbres de decisió, Xarxes neuronals, Regressió estadística, Xarxes bayesianes, Regles d'associació) constitueix l'enfocament conceptual per extreure la informació de les dades, i, en general és implementada per diversos algorismes.

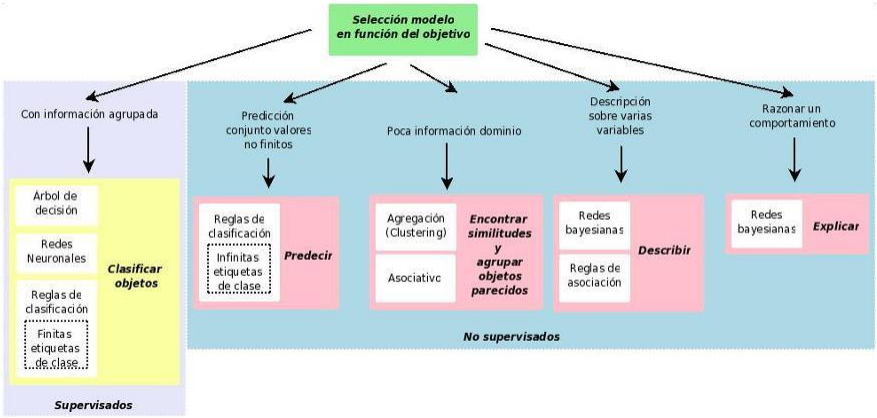
Les tècniques de Mineria de Dades es classifiquen en dos grans categories:

- Supervisades o Predictives.
- No supervisades o Descriptives.

Les prediccions es fan servir per preveure el comportament futur d'algun tipus d'entitat mentre que, una descripció pot ajudar a la seva comprensió.

De fet, els models predictius (supervisats) poden ser descriptius (fins on siguin comprensibles per persones) i els models descriptius (no supervisats) poden emprar-se per realitzar prediccions.

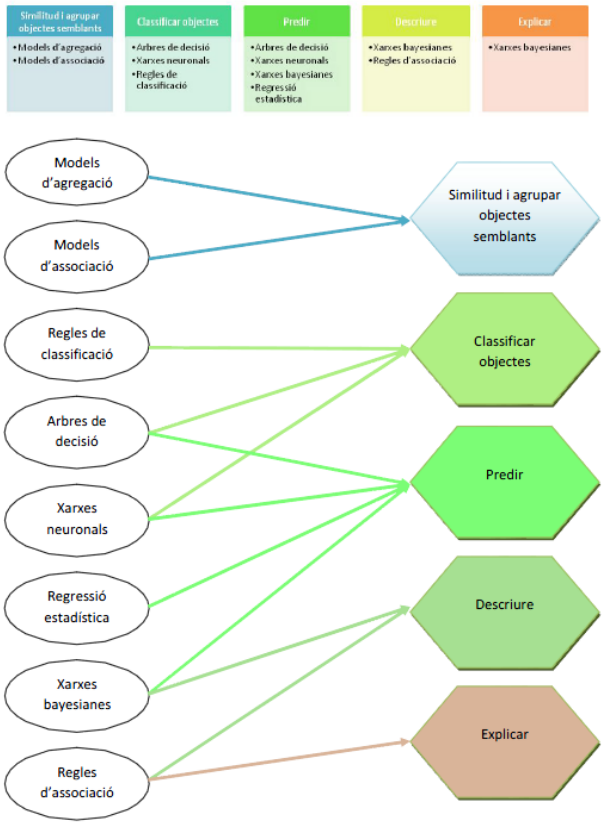
En la il·lustració següent es mostra un resum esquematitzat de com es pot decidir el tipus de tècnica o model de mineria de dades a emprar, tenint en compte les necessitats del projecte (agrupar, classificar, predir, descriure i/o explicar) :



Il·lustració 8 :Tècniques i objectius.
Font : 2009 Gutiérrez Covarrubias, Manuel Ramón PEC 1

D'aquesta manera, hi ha algorismes i tècniques que poden servir per a diferents propòsits.

La següent il·lustració mostra un resum dels models més habituals que es poden emprar per poder agrupar, classificar, predir, descriure i/o explicar de manera profitosa.



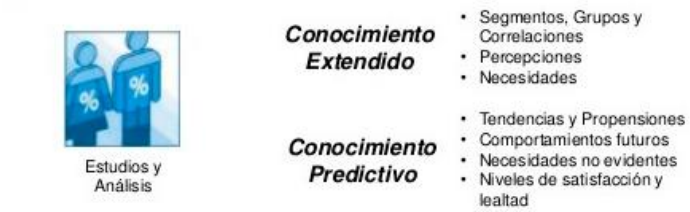
Il·lustració 9 : Models i objectius.
Font : 2013 Antoni Corral Herrerías TFC Àrea Mineria de dades.

En aquest projecte es vol extreure el màxim de coneixement possible, per tant la intenció es provar el màxim d'algorismes de tots els models per veure la resposta de les dades i establir les bases per poder agrupar, classificar, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants.

1.3.5. Anàlisi de resultats i assimilació del coneixement

Arriba el moment d'interpretar els resultats obtinguts en l'etapa anterior, generalment amb l'ajuda d'una tècnica de visualització, es realitza la seva exposició, i s'aplica el coneixement extret en la mesura de lo possible.

Els tipus de coneixements que hi ha es detallen en la següent il·lustració :

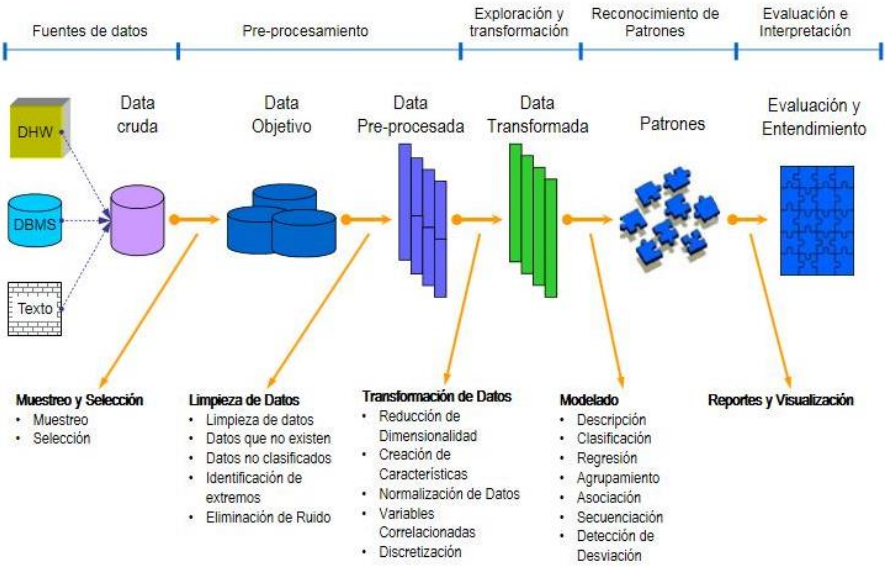


Il·lustració 10 : Conocimiento extendido y predictivo.
Font : PASS (Predictive Analytical Software and Solutions) http://pass.mx

Al concloure aquesta fase s'espera haver extret el màxim de coneixement de les proves realitzades i poder establir una base sòlida per poder agrupar, classificar, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants.

Tot i que els passos anteriors es realitzen en l'ordre en què apareixen, el procés és altament iteratiu, establint retroalimentació entre els mateixos. A més, no totes les vies requereixen el mateix esforç, generalment l'etapa de pre-processament és la més costosa ja que representa aproximadament el 60% de l'esforç total, mentre que l'etapa de mineria només representa el 10%.

En aquesta il·lustració es resumeixen totes les fase del projecte, detallant en cadascuna d'elles els aspectes més rellevants en el seu desenvolupament :



Il·lustració 11 : El procés de la mineria de dades.
Font : Mg. Samuel Oporto Díaz http://www.wiphala.net/oporto

1.4. Planificació del treball

La planificació del treball, ve donada per les dates de lliurament de les PACs (de la 1 a la 3) i del lliurament final marcades per la UOC dins l'assignatura 05.629 TFG - Intel·ligència artificial (TFG : Treball de Final de Grau).

1.4.1. Fites

En la **primera PAC**, amb data de lliurament el **9 de març del 2016** es presenta una planificació durant un període de temps específic (el semestre que dura l'assignatura TFG) que identifica els problemes que s'han de solucionar i proposa maneres de solucionar-los, servint com a document guia per dur a terme les activitats durant aquest període de temps.

Entre les **PACs 2 i 3** amb dates **6 d'abril del 2016** i **4 de maig del 2016** respectivament, se seleccionen les dades, es preparen i es transformen per poder assegurar que aquestes tenen la qualitat suficient, que són les necessàries i que tenen la forma adient.

En el **lliurement final** amb data **1 de juny del 2016**, primer es decideix quins algorismes actuaran sobre aquestes en funció de si es vol agrupar, classificar, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants.

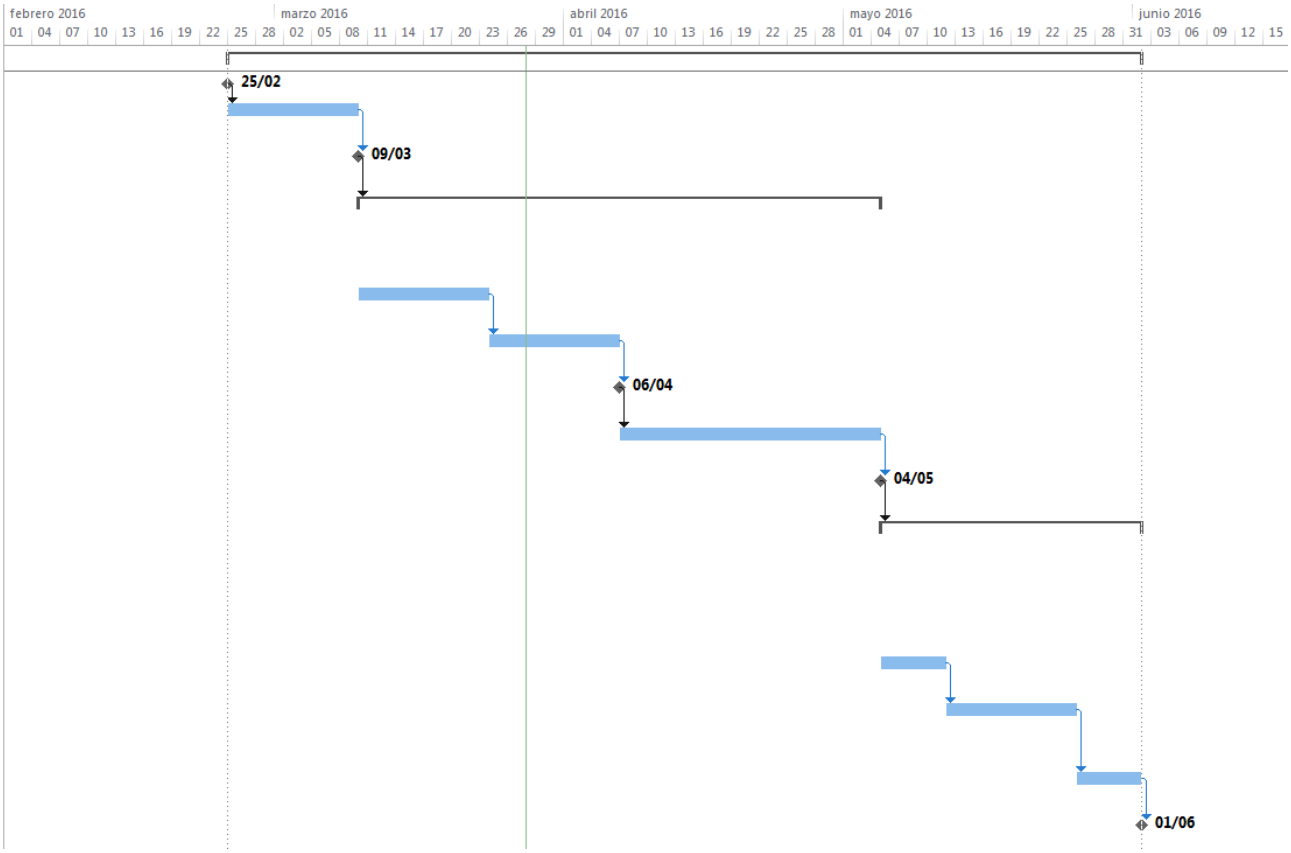
Finalment es fa un tractament automatitzat de les dades seleccionades amb una combinació apropiada d'algorismes i es presenta la interpretació dels resultats obtinguts amb el tractament d'aquests tant a nivell medi ambiental com sòcio-econòmic, a més de les conclusions i línies de futures investigacions.

1.4.2. Cronograma

DIAGRAMA DE GANTT

	Modo de	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1		**Projecte TFG**	**70 días**	**jue 25/02/16**	**mié 01/06/16**	
2		Inici	0 días	jue 25/02/16	jue 25/02/16	
3		Fase 0 : Pla de Treball	10 días	jue 25/02/16	mié 09/03/16	2
4		Lliurar Pla de Treball	0 días	mié 09/03/16	mié 09/03/16	3
5		**Fase 1 : Selecció, Preparació i Transformació de les Dades**	**40 días**	**jue 10/03/16**	**mié 04/05/16**	**4**
6		Selecció de les Dades	10 días	jue 10/03/16	mié 23/03/16	
7		Preparació de les Dades	10 días	jue 24/03/16	mié 06/04/16	6
8		Lliurar Arxiu Excel i Actualitzacions de Productes Anteriors	0 días	mié 06/04/16	mié 06/04/16	7
9		Transformació de les Dades	20 días	jue 07/04/16	mié 04/05/16	8
10		Lliurar Arxiu ARF i Actualitzacions de Productes Anteriors	0 días	mié 04/05/16	mié 04/05/16	9
11		**Fase 2 : Mineria de Dades, Interpretació dels Resultats, Conclusions i Línies Futures**	**20 días**	**jue 05/05/16**	**mié 01/06/16**	**10**
12		Mineria de Dades	5 días	jue 05/05/16	mié 11/05/16	
13		Interpretació dels Reultats i Conclusions	10 días	jue 12/05/16	mié 25/05/16	12
14		Línies Futures d'Investigació	5 días	jue 26/05/16	mié 01/06/16	13
15		Fi : Lliurament de la Memòria amb Tots els Productes Actualitzats	0 días	mié 01/06/16	mié 01/06/16	14

1.4.3. Diagrama de Gantt



1.5. Breu sumari de productes obtinguts

Els productes obtinguts, igual que la planificació del treball, venen donats per les dates de lliurament de les PACs (de la 1 a la 3) i del lliurament final marcades per la UOC dins l'assignatura 05.629 TFG - Intel·ligència artificial (TFG : Treball de Final de Grau).

El **primer producte** obtingut és el pla de treball (primer capítol de la memòria final) en el qual es detallen :

- El context i la justificació del treball.
- Els objectius del treball.
- L'enfocament i el mètode seguit en el treball.
- La planificació del treball.
- Un breu sumari dels productes obtinguts.
- Una breu descripció dels altres capítols de la memòria que es lliura al final del TFG.

Juntament amb aquest document, també es lliura un arxiu fet amb Microsoft Project de la planificació temporal de tot el projecte.

El **segon producte** obtingut és un arxiu fet amb Microsoft Excel que conté totes les dades recol·lectades per la seva preparació/transformació, juntament amb aquest lliurable, s'adjunten els capítol 2 de la memòria final en els quals s'explica amb detall quines dades s'han seleccionat i quins passos s'han dut a terme per preparar-les.

El **tercer producte** obtingut és un arxiu amb format **.arff** que es pot llegir amb el Bloc de Notes en el qual estan totes les dades netejades, homogeneïtzades, transformades i reduïdes, si s'escau, és a dir, totalment preparades per tractar-les amb el programari de mineria de dades Weka.

Juntament amb aquest arxiu es lliura el capítol 3 de la memòria final en el que s'expliquen detalladament les accions dutes a terme per transformar les dades i deixar-les llestes per aplicar-hi els algorismes de mineria de dades que s'escolliran.

El **quart i darrer lliurable** és la memòria final en la que s'expliquen els algorismes de mineria de dades aplicats, s'interpreten els resultats obtinguts amb aquests, s'extreuen les conclusions finals del projecte, s'expliquen les possibles línies futures d'investigació i es resumeix tot el procés del treball realitzat.

Il·lustració 12 : Entrega de producte.
Font : https://ciitlaliiflores.wordpress.com/2013/10/03/analisis-de-la-empresa-privalia/

Cal tenir amb compte que a partir del segon lliurable, tots els canvis, que s'hagin realitzat tan en els lliurables anteriors, com en els capítols ja redactats de la memòria, s'actualitzaran, si s'escau, pels resultats obtinguts en la fase en curs del projecte.

1.6. Breu descripció dels altres capítols de la memòria

En els **capítols 2 i 3** s'explica detalladament tot el procés de selecció i preparació de les dades, és a dir :

- Quines dades es tenen al començament.
- Quines d'aquestes i perquè se seleccionen.
- Es tracten totes de la mateixa forma o separadament.
- Quines accions es prenen davant la manca total o parcial de dades.
- Quines accions es realitzen per poder primer crear un fitxer de tipus Excel.
- Quines accions es realitzen per traslladar totes aquestes dades a un fitxer en format **.arff** que és el més adient per emprar en el programari lliure de mineria de dades Weka.

En el **capítol 4** de la memòria, s'expliquen detalladament els algorismes de mineria de dades emprats amb el programari lliure Weka i el perquè d'aquestes tries.

En els **capítols 5 i 6** de la memòria, s'expliquen les interpretacions fetes dels resultats obtinguts amb els algorismes de mineria de dades i les conclusions i possibles línies futures extretes de totes elles.

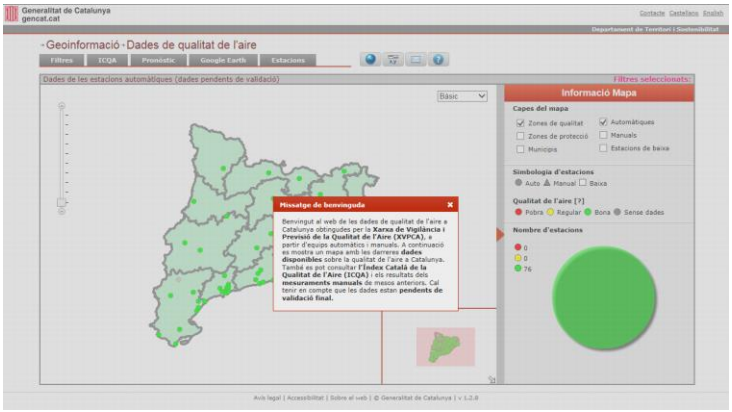
2. Selecció i exploració de les dades

La mineria de dades s'ha de considerar com una fase del descobriment de coneixement a partir de dades. Trobar les dades que calen és més fàcil de dir que de fer.

2.1. Recol·lecció de dades

La font d'on s'han obtingut les dades és pública i aquestes provenen de les mesures dels contaminants en l'aire obtingudes per la Xarxa de Vigilància i Previsió de la Qualitat de l'Aire (XVPCA), a partir d'equips automàtics i manuals (web de la Generalitat de Catalunya http://dtes.gencat.cat/icqa/), aquesta, conté la possibilitat de descarregar arxius tipus Excel, de com a màxim un any de durada, amb l'històric de les mesures automàtiques (mitjanes) diàries d'alguns dels contaminants següents : *CO (mg/m³), H2S (µg/m³), NO (µg/m³), NO2 (µg/m³), O3 (µg/m³), PM10 (µg/m³), PST (µg/m³) i SO2 (µg/m³)*, en cadascun d'aquests arxius, poden no estar continguts tots els contaminants esmentats.

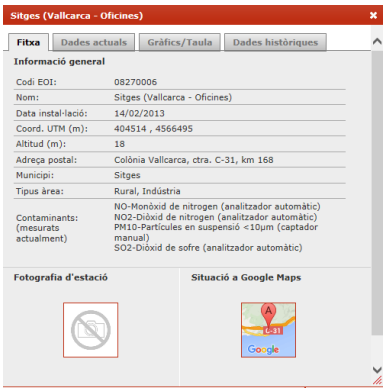
En la següent captura de pantalla es pot veure una imatge inicial de la web en qüestió:



Il·lustració 13 : Captura de pantalla de la web de la qualitat de l'aire de la Generalitat de Catalunya.
Font : 2015 http://dtes.gencat.cat/icqa/

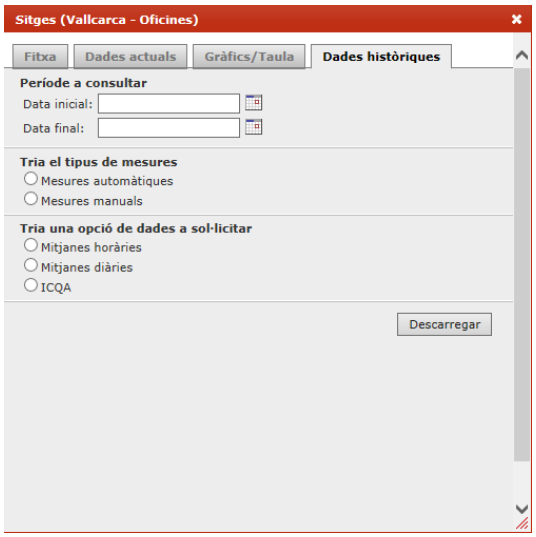
Cada punt verd representa una de les 76 estacions meteorològiques repartides per totes les zones del territori català.

A l'accedir a un d'aquests punts apareix una finestra amb la fitxa tècnica de cada estació, que conté la Informació general, les Dades actuals, Gràfics/Taules i Dades històriques, com es mostra en la següent captura de pantalla :



Il·lustració 14 : Captura de pantalla d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya.
Font : 2015 http://dtes.gencat.cat/icqa/

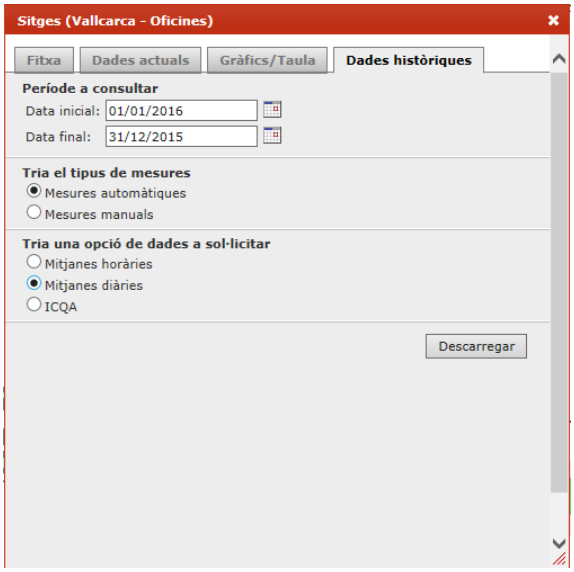
Accedint a la pestanya Dades històriques de la finestra anterior s'està en disposició de poder descarregar l'històric de les mesures automàtiques (mitjanes) diàries d'alguns d'aquests contaminants :*CO (mg/m³), H2S (µg/m³), NO (µg/m³), NO2 (µg/m³), O3 (µg/m³), PM10 (µg/m³), PST (µg/m³) o SO2 (µg/m³)*, com es mostra en la següent captura de pantalla.



Il·lustració 15 : Captura de pantalla de la pestanya Dades històriques d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya.
Font : 2015 http://dtes.gencat.cat/icqa/

El màxim que permet extreure la web en un arxiu de tipus Excel són les dades en un període d'un any, així doncs, per seleccionar les dades del període 2005-2015 d'una estació en concret, s'han descarregat onze arxius de tipus Excel, un per any desitjat.

En la següent captura de pantalla es mostra aquesta situació en l'estació de Sitges (Vallcarca -oficines) com exemple :



Il·lustració 16 : Captura de pantalla de la pestanya Dades històriques amb les opcions i les dates escollides d'una estació de la web de la qualitat de l'aire de la Generalitat de Catalunya.
Font : 2015 http://dtes.gencat.cat/icqa/

D'acord amb la legislació vigent (Directiva 2008/50/CE, Reial Decret 102/2011, etc.), Catalunya està dividida en 15 zones de qualitat de l'aire (ZQA) segons les emissions i les condicions de dispersió, tal i com es mostra en la següent captura de pantalla:

ZQA	Nom de les zones
1	Àrea de Barcelona
2	Vallès - Baix Llobregat
3	Penedès - Garraf
4	Camp de Tarragona
5	Catalunya Central
6	Plana de Vic
7	Maresme
8	Comarques de Girona
9	Empordà
10	Alt Llobregat
11	Pirineu Oriental
12	Pirineu Occidental
13	Prepirineu
14	Terres de Ponent
15	Terres de l'Ebre

Il·lustració 17 : Captura de pantalla del mapa de les zones de qualitat de l'aire de (XVPCA).
Font : 2015 http://www.navas.cat/media/download/3291

En aquest treball, s'ha incorporat un altre criteri de definició de zones a més de l'oficial, redefinint el territori en 6 zones representatives de diferents àmbits: Ciutat Poblada, Costa, Delta de l'Ebre, Interior, Pirineu-Prepirineu i Industrial, aquestes zones, han estat les que han marcat la tria d'estacions i són ampliables i/o canviables si s'incorporen més estacions d'entre les 76 existents.

A més poden donar un altre enfoc a l'estudi dels contaminants i possibilitar la comparativa de resultats d'aquests amb els de les zones oficials.

Les estacions d'estudi escollides segons les zones redefinides i incorporant la zona oficial han estat:

Zona Redefinida	Comarca	Població	Estació	ZQA	Zona Oficial
Ciutat Poblada	Vallès Occidental	Sabadell	Gran Via	2	Vallès - Baix Llobregat
Ciutat Poblada	Vallès Occidental	Terrassa	Pare Alegre	2	Vallès - Baix Llobregat
Costa	Baix Llobregat	Gavà	Parc del Mil·leni	1	Àrea de Barcelona
Costa	Garraf	Cubelles	Poliesportiu	3	Penedès - Garraf
Costa	Garraf	Sitges	Vallcarca - Oficines	3	Penedès - Garraf
Costa	Garraf	Vilanova	Plaça de les Danses de Vilanova	3	Penedès - Garraf
Delta de l'Ebre	Montsià	Alcanar	Llar de Jubilats	15	Terres de l'Ebre
Delta de l'Ebre	Montsià	Amposta	Sant Domènec - Itàlia	15	Terres de l'Ebre
Delta de l'Ebre	Montsià	La Sénia	Repetidor	15	Terres de l'Ebre
Industrial	Baix Camp	Reus	El Tallapedra	4	Camp de Tarragona
Industrial	Tarragonès	Tarragona	Sant Salvador	4	Camp de Tarragona
Interior	Osona	Manlleu	Hospital Comarcal	6	Plana de Vic
Interior	Osona	Tona	Zona Esportiva	6	Plana de Vic
Interior	Osona	Vic	Estadi	6	Plana de Vic
Pirineu-Prepirineu	Cerdanya	Bellver de Cerdanya	CEIP Mare de Déu de Talló	11	Pirineu Oriental
Pirineu-Prepirineu	Pallars Sobirà	Sort	Escola de Caiac	12	Pirineu Occidental
Pirineu-Prepirineu	Ripollès	Pardines	Ajuntament	11	Pirineu Oriental

Taula 4 : Estacions d'estudi escollides segons les zones redefinides incorporant la zona oficial

Les estacions de la ciutat de Barcelona han quedat excloses d'aquest projecte, ja que, Barcelona per si sola, considerada com una gran aglomeració urbana i pel nombre d'estacions que té, seria objecte d'un projecte similar exclusivament per a ella. Barcelona es considerada zona de protecció especial de l'ambient atmosfèric (ZPE) amb l'aprovació el maig de 2006 del Decret 226/2006, zona on s'havia registrat problemes de contaminació.

Un cop seleccionades les disset estacions, es descarreguen onze arxius (un per cada any del període 2005-2015 de les mesures automàtiques (mitjanes) diàries) per cada estació seleccionada.

2.2. Descripció de les dades

Els arxius de Excel descarregats presenten en general el següent format, primer de tot una columna sense nom en la qual hi ha les dates del període de temps seleccionat, tenint en compte que no pot ser superior a un any, seguidament poden aparèixer unes quantes columnes cadascuna amb el nom d'un dels contaminants esmentats anteriorment que contenen les mesures automàtiques (mitjanes) diàries del contaminant que porta per nom la columna, per finalitzar amb una columna de nom Dades pendents de validació que està sempre totalment buida.

Les diferents opcions trobades són :

- Hi ha arxius que només contenen les dates de l'any seleccionat i una columna anomenada Dades pendents de validació en la qual no hi ha res com es mostra en aquesta captura de pantalla.

	A	B
1		Dades pendents de validació
2	01/01/2005	
3	02/01/2005	
4	03/01/2005	
5	04/01/2005	
6	05/01/2005	

Il·lustració 18 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

- Hi ha arxius que només contenen les dates de l'any seleccionat, una o varies columnes que tenen per capçalera el nom i les unitats de mesura d'un contaminant i que contenen les seves mesures automàtiques (mitjanes) diàries i finalment la columna anomenada Dades pendents de validació en la qual no hi ha res com es mostra en aquestes captures de pantalla.

	A	B	C
1		O3 (µg/m³)	Dades pendents de validació
2	01/01/2005	22	
3	02/01/2005	38	
4	03/01/2005	41	
5	04/01/2005	20	
6	05/01/2005	18	

Il·lustració 19 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D
1		NO (µg/m³)	NO2 (µg/m³)	Dades pendents de validació
2	01/01/2013	1	10	
3	02/01/2013	1	8	
4	03/01/2013	4	26	
5	04/01/2013	4	29	
6	05/01/2013	3	22	
7	06/01/2013	1	13	

Il·lustració 20 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D	E
1		SO2 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	Dades pendents de validació
2	01/01/2014	2	1	8	
3	02/01/2014	2	3	21	
4	03/01/2014	2	3	25	
5	04/01/2014	2	1	9	
6	05/01/2014	2	1	6	

Il·lustració 21 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D	E	F
1		NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	PM10 (µg/m³)	Dades pendents de validació
2	01/01/2014	6	21	31	15	
3	02/01/2014	20	26	15	29	
4	03/01/2014	30	34	8	42	
5	04/01/2014	9	18	36	14	
6	05/01/2014	7	16	59	10	

Il·lustració 22 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D	E	F	G
1		SO2 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	CO (mg/m³)	Dades pendents de validació
2	01/01/2010	2	6	13	70	0,4	
3	02/01/2010	3	34	44	23	0,4	
4	03/01/2010	2	33	54	15	0,3	
5	04/01/2010	3	69	64	6	0,7	
6	05/01/2010	3	75	61	9	1	

Il·lustració 23 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D	E	F	G	H
1		SO2 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	H2S (µg/m³)	CO (mg/m³)	Dades pendents de validació
2	01/01/2006	5	1	4	59	1,1	0,2	
3	02/01/2006	2	1	8	52	1,1	0,2	
4	03/01/2006	1	1	10	43	1,2	0,2	
5	04/01/2006	1	8	24	30	1,3	0,3	
6	05/01/2006	3	16	33	12	1,2	0,5	

Il·lustració 24 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

	A	B	C	D	E	F	G	H	I
1		SO2 (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	PST (µg/m³)	O3 (µg/m³)	H2S (µg/m³)	CO (mg/m³)	Dades pendents de validació
2	01/01/2005	6	18	20	15	37	1,2	0,3	
3	02/01/2005	8	15	28	21	30	1,1	0,2	
4	03/01/2005	2	55	26	38	34	1,4	0,5	
5	04/01/2005	23	94	40	60	13	1,1	0,6	
6	05/01/2005	17	74	42	65	17	1,1	0,6	

Il·lustració 25 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

2.3. Exploració de les dades

Després d'haver recol·lectat i descrit les dades, aquestes s'han d'explorar i verificar, amb un primer cop d'ull als arxius descarregats, s'observa que respecte a les dades trobades en les cel·les de les columnes que tenen per capçalera el nom i les unitats de mesura d'un contaminant (*CO, H2S, NO, NO2, O3, PM10, PST o SO2)*, es poden classificar en :

- Cel·les en blanc.
- Cel·les amb les paraules **Sense dades**.
- Cel·les amb valors numèrics.

La següent captura de pantalla mostra les diferents possibilitat esmentades :

1	Data	CO (mg/m³)	H2S (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	SO2 (µg/m³)
2	01/01/2007	0,6		49	43	13	1
3	02/01/2007	0,6		114	65	8	3
4	03/01/2007	0,7		97	79	12	2
5	04/01/2007	0,5		80	68	12	3
6	05/01/2007	0,6		77	67	9	2
7	06/01/2007	0,6		51	50	14	3
8	07/01/2007	0,5		68	68	7	4
9	08/01/2007	0,7		85	67	13	3
10	09/01/2007	0,7		119	85	9	7
11	10/01/2007	0,9		124	86	7	7
12	11/01/2007	0,8		124	80	8	7
13	12/01/2007	0,3		39	37	4	4
14	13/01/2007	Sense dades		Sense dades	Sense dades	Sense dades	Sense dades
15	14/01/2007	Sense dades		Sense dades	Sense dades	Sense dades	Sense dades
16	15/01/2007	1,1		145	95	1	7
17	16/01/2007	1,1		135	90	2	6

Il·lustració 26 : Captura de pantalla de la capçalera amb unes poques files d'un històric d'una estació. Font : XVPCA.

2.4. Verificació de les dades

Respecte la garantia de les dades cal assenyalar que el Departament de Territori i Sostenibilitat és l'òrgan responsable de l'avaluació de la qualitat de l'aire a Catalunya, i que la principal eina per realitzar aquesta tasca és la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica (XVPCA) que integra els diferents punts de mesura distribuïts al territori.

Fins a l'entrada en vigor del nou *Reial decret 102/2011, de 28 de gener, relatiu a la millora de la qualitat de l'aire, el marc normatiu relatiu per a l'avaluació i la gestió de la qualitat de l'aire ambient* era definit pel *Reial decret 1073/2002, de 18 d'octubre (sobre avaluació i gestió de la qualitat de l'aire ambient en relació amb el diòxid de sofre, el diòxid de nitrogen, òxids de nitrogen, partícules, plom, benzè i monòxid de carboni)*, el *Reial decret 812/2007, de 22 de juny (sobre avaluació i gestió de la qualitat de l'aire ambient en relació amb l'arsènic, el cadmi, el mercuri, el níquel i els hidrocarburs aromàtics policíclics)*, i la *Directiva 2008/50/CE, de 21 de maig (relativa a la qualitat de l'aire ambient i a una atmosfera més neta a Europa)*. Aquest Reial decret 102/2011 desenvolupa els aspectes relacionats amb la qualitat de l'aire de la *Llei 34/2007, del 15 de novembre, de qualitat de l'aire i protecció de l'atmosfera i transposa la nova directiva europea*, tot i integrant tots els reials decrets anteriorment aprovats. Per tant, la legislació de referència per a l'avaluació de la qualitat de l'aire és la Llei 34/2007 i el Reial decret 102/2011.

D'acord amb aquest marc, l'avaluació de la qualitat de l'aire es fa per zones. Per aquest motiu es va classificar el territori en 15 zones de qualitat de l'aire (ZQA) i aquestes zones estan definides per tal que la seva superfície presenti unes característiques similars respecte a la qualitat de l'aire considerant elements com: l'orografia, la climatologia, la densitat de població, el volum d'emissions industrials i de transport.

3. Preparació de les dades

En la mineria de dades, el més normal és que les dades necessàries per a dur a terme un projecte de mineria de dades, un cop seleccionades, hagin de ser preparades i, si s'escau, modificades/transformades per a poder-les-hi aplicar el mètode de construcció del model escollit per a la tasca (agrupar, classificar, predir, descriure i/o explicar).

3.1. Estadístiques i gràfics de les dades

Per poder realitzar aquestes estadístiques s'han agrupat totes les dades en un arxiu de tipus Excel que conté 68290 files (68289 files de dades i una de capçalera), aquest arxiu permet poder extreure a més de les estadístiques, gràfics, valors màxims i mínims de les dades que permetran preparar-les, i si s'escau transformar-les, per poder crear l'arxiu definitiu tipus .arff, que utilitzarà el programari lliure de mineria de dades Weka.

En aquest arxiu de tipus Excel s'ha optat per primer de tot, per afegir el màxim d'informació possible, per després, si és necessari anar-la reduint mentre s'avança en les fases de la preparació de dades descrites en el punt 1.3.3 de la memòria.

Per tant la capçalera d'aquest arxiu conté els següents camps :

- Data : Data de la mesura automàtica (mitjana) diària.
- Dia : Número de dia del mes de la mesura automàtica (mitjana) diària.
- Mes : Nom del mes de la mesura automàtica (mitjana) diària.
- Any : Any de la mesura automàtica (mitjana) diària.
- CO (mg/m³) : Valor de la mesura automàtica (mitjana) diària de CO en les unitats especificades.
- H2S (µg/m³) : Valor de la mesura automàtica (mitjana) diària de H2S en les unitats especificades.
- NO (µg/m³) : Valor de la mesura automàtica (mitjana) diària de NO en les unitats especificades.
- NO2 (µg/m³) : Valor de la mesura automàtica (mitjana) diària de NO2 en les unitats especificades.
- O3 (µg/m³) : Valor de la mesura automàtica (mitjana) diària de O3 en les unitats especificades.
- SO2 (µg/m³) : Valor de la mesura automàtica (mitjana) diària de SO2 en les unitats especificades.
- Població : Nom de la població on està ubicada l'estació que ha fet la mesura automàtica (mitjana) diària.
- Estació : Nom de l'estació que ha fet la mesura automàtica (mitjana) diària.
- Comarca : Nom de la comarca on està ubicada l'estació que ha fet la mesura automàtica (mitjana) diària.
- ZQA. : Numero de la zona oficial on està ubicada l'estació que ha fet la mesura automàtica (mitjana) diària.
- Zona Oficial : Nom de la zona oficial on està ubicada l'estació que ha fet la mesura automàtica (mitjana) diària.
- Zona Redefinida : Nom de la zona redefinida en aquest estudi on està ubicada l'estació que ha fet la mesura automàtica (mitjana) diària.

Aquesta captura de pantalla de la capçalera de l'arxiu i quatre files d'aquest, mostra l'esmentat :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Data	Dia	Mes	Any	CO (mg/m³)	H2S (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	SO2 (µg/m³)	Població	Estació	Comarca	ZQA	Zona Oficial	Zona Redefinida
2	01/01/2005	1	Gener	2005	0,7		22	21	15	9	Sabadell	Gran Via	Vallès Occidental	2	Vallès - Baix Llobregat	Ciutat Poblada
3	02/01/2005	2	Gener	2005	0,5		19	18	14	7	Sabadell	Gran Via	Vallès Occidental	2	Vallès - Baix Llobregat	Ciutat Poblada
4	03/01/2005	3	Gener	2005	0,7		48	24	12	7	Sabadell	Gran Via	Vallès Occidental	2	Vallès - Baix Llobregat	Ciutat Poblada
5	04/01/2005	4	Gener	2005	0,9		97	20	9	7	Sabadell	Gran Via	Vallès Occidental	2	Vallès - Baix Llobregat	Ciutat Poblada

Il·lustració 27 : Captura de pantalla de la capçalera amb unes poques files de l'arxiu Excel creat.
Font : 2016 Document Excel Dades lliurat.

Mitjançant la creació de taules dinàmiques i fórmules, a continuació es mostra amb una captura de pantalla el resum del tipus de cel·les (en blanc, amb la paraula “Sense dades” o amb valors numèrics) que té cadascun dels contaminants, així com, l’existència o no d’aquests en les mesures de cadascuna de les disset poblacions escollides.

Aquest resum especifica per cada contaminant el número total de cel·les en el període 2005-2015, el número total de cel·les en blanc i amb dades i d’aquest darrer, el número total de cel·les amb valors numèrics, així com el número total de cel·les amb dades numèriques que aporta cada població, a més dels totals generals de cel·les en blanc, amb dades i amb dades numèriques.

Resum Dades

Poblacions	CO (mg/m³)	H2S (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	SO2 (µg/m³)	Total Dades Població
Alcanar	No	No	Si	Si	No	No	2108
Amposta	No	No	Si	Si	Si	No	8165
Bellver de Cerdanya	No	No	Si	Si	Si	Si	11990
Cubelles	No	No	Si	Si	No	Si	11499
Gavà	Si	No	Si	Si	Si	Si	11557
La Sénia	No	No	No	No	Si	No	3953
Manlleu	No	No	Si	Si	Si	Si	14325
Pardines	No	No	No	No	Si	No	3903
Reus	Si	Si	Si	Si	Si	Si	21714
Sabadell	Si	No	Si	Si	Si	Si	16159
Sitges	No	No	Si	Si	No	Si	1947
Sort	No	No	No	No	Si	No	2686
Tarragona	Si	Si	Si	Si	No	Si	17682
Terrassa	Si	No	Si	Si	Si	Si	16821
Tona	No	No	Si	Si	Si	No	6910
Vic	No	No	No	No	Si	No	3975
Vilanova	Si	No	Si	Si	Si	Si	19427
Total Cel·les	68289	68289	68289	68289	68289	68289	Totals Generals
Total Cel·les en Blanc	48204	60255	25928	25928	19355	38344	218014
Total Dades Contaminant	20085	8034	42361	42361	48934	29945	191720
Total Cel·les Numèriques	17964	7738	38787	38725	44975	26632	174821
Total Cel·les Sense Dades	2121	296	3574	3636	3959	3313	16899

Il·lustració 28 : Captura de pantalla taula resum general de l’arxiu Excel creat.
Font : 2016 Document Excel Dades lliurat.

Mitjançant la creació de taules dinàmiques i fórmules, a continuació es mostra amb sis captures de pantalla el resum del tipus de cel·les (en blanc, amb la paraula “Sense dades” o amb valors numèrics) que té cadascun dels contaminants per població.

CO (mg/m³)

Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar			4017	0	4017
Amposta			4017	0	4017
Bellver de Cerdanya			4017	0	4017
Cubelles			4017	0	4017
Gavà	2922	2280	1095	642	4017
La Sénia			4017	0	4017
Manlleu			4017	0	4017
Pardines			4017	0	4017
Reus	4017	3952	0	65	4017
Sabadell	2556	2131	1461	425	4017
Sitges			4017	0	4017
Sort			4017	0	4017
Tarragona	2556	2076	1461	480	4017
Terrassa	4017	3619	0	398	4017
Tona			4017	0	4017
Vic			4017	0	4017
Vilanova	4017	3906	0	111	4017

Il·lustració 29

H2S (µg/m³)

Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar			4017	0	4017
Amposta			4017	0	4017
Bellver de Cerdanya			4017	0	4017
Cubelles			4017	0	4017
Gavà			4017	0	4017
La Sénia			4017	0	4017
Manlleu			4017	0	4017
Pardines			4017	0	4017
Reus	4017	3858	0	159	4017
Sabadell			4017	0	4017
Sitges			4017	0	4017
Sort			4017	0	4017
Tarragona	4017	3880	0	137	4017
Terrassa			4017	0	4017
Tona			4017	0	4017
Vic			4017	0	4017
Vilanova			4017	0	4017

Il·lustració 30

NO (µg/m³)					
Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar	1461	1054	2556	407	4017
Amposta	2191	2112	1826	79	4017
Bellver de Cerdanya	4017	3882	0	135	4017
Cubelles	4017	3784	0	233	4017
Gavà	2922	2334	1095	588	4017
La Sénia			4017	0	4017
Manlleu	4017	3563	0	454	4017
Pardines			4017	0	4017
Reus	4017	3841	0	176	4017
Sabadell	4017	3964	0	53	4017
Sitges	1095	624	2922	471	4017
Sort			4017	0	4017
Tarragona	4017	3890	0	127	4017
Terrassa	4017	3559	0	458	4017
Tona	2556	2297	1461	259	4017
Vic			4017	0	4017
Vilanova	4017	3883	0	134	4017

Il·lustració 31

NO2 (µg/m³)					
Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar	1461	1054	2556	407	4017
Amposta	2191	2112	1826	79	4017
Bellver de Cerdanya	4017	3825	0	192	4017
Cubelles	4017	3789	0	228	4017
Gavà	2922	2336	1095	586	4017
La Sénia			4017	0	4017
Manlleu	4017	3554	0	463	4017
Pardines			4017	0	4017
Reus	4017	3847	0	170	4017
Sabadell	4017	3964	0	53	4017
Sitges	1095	620	2922	475	4017
Sort			4017	0	4017
Tarragona	4017	3893	0	124	4017
Terrassa	4017	3559	0	458	4017
Tona	2556	2298	1461	258	4017
Vic			4017	0	4017
Vilanova	4017	3874	0	143	4017

Il·lustració 32

O3 (µg/m³)					
Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar			4017	0	4017
Amposta	4017	3941	0	76	4017
Bellver de Cerdanya	4017	3972	0	45	4017
Cubelles			4017	0	4017
Gavà	2922	2358	1095	564	4017
La Sénia	4017	3953	0	64	4017
Manlleu	4017	3765	0	252	4017
Pardines	4017	3903	0	114	4017
Reus	4017	3933	0	84	4017
Sabadell	4017	3955	0	62	4017
Sitges			4017	0	4017
Sort	3286	2686	731	600	4017
Tarragona			4017	0	4017
Terrassa	4017	3656	0	1702	4017
Tona	2556	2315	1461	241	4017
Vic	4017	3975	0	42	4017
Vilanova	4017	3904	0	113	4017

Il·lustració 33

SO2 (µg/m³)					
Poblacions	Cel·les Escrites	Cel·les Numèriques	Cel·les En Blanc	Cel·les Sense dades	Total Cel·les 2005-2015
Alcanar			4017	0	4017
Amposta			4017	0	4017
Bellver de Cerdanya	365	311	3652	54	4017
Cubelles	4017	3926	0	91	4017
Gavà	2922	2249	1095	673	4017
La Sénia			4017	0	4017
Manlleu	4017	3443	0	574	4017
Pardines			4017	0	4017
Reus	2922	2283	1095	639	4017
Sabadell	2556	2145	1461	411	4017
Sitges	1095	703	2922	392	4017
Sort			4017	0	4017
Tarragona	4017	3943	0	74	4017
Terrassa	4017	3769	0	248	4017
Tona			4017	0	4017
Vic			4017	0	4017
Vilanova	4017	3860	0	157	4017

Il·lustració 34

Il·lustracions 29, 30, 31, 32, 33 i 34 : Captures de pantalla taules específiques resum cel·les per contaminant i població de l'arxiu Excel creat.
Font : 2016 Document Excel Dades lliurat.

Mitjançant la creació de taules dinàmiques, a continuació es mostra amb dues captures de pantalla el resum dels valors màxim i mínim de cada contaminant per població i el màxim i el mínim general de cadascun d'ells que mostrarà el rang de valors que prenen les dades.

Poblacions	Máx. de CO (mg/m³)	Máx. de H2S (µg/m³)	Máx. de NO (µg/m³)	Máx. de NO2 (µg/m³)	Máx. de O3 (µg/m³)	Máx. de SO2 (µg/m³)
Alcanar			999	999		
Amposta			41	48	125	
Bellver de Cerdanya			30	43	129	9
Cubelles			54	81		14
Gavà	2,2		117	89	113	39
La Sénia					139	
Manlleu			168	100	143	154
Pardines					144	
Reus	1,6	6,3	171	83	113	101
Sabadell	2,7		283	173	110	43
Sitges			28	33		4
Sort					124	
Tarragona	1	14,6	56	98		80
Terrassa	2,9		248	119	108	34
Tona			37	48	144	
Vic					142	
Vilanova	1,1		71	79	111	41
Total general	**2,9**	**14,6**	**999**	**999**	**144**	**154**

Il·lustració 35

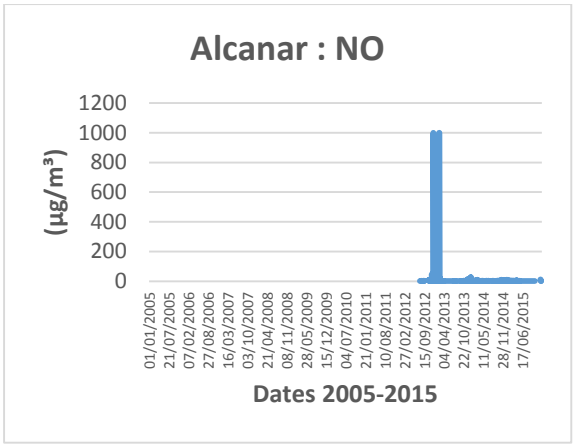
Poblacions	Mín. de CO (mg/m³)	Mín. de H2S (µg/m³)	Mín. de NO (µg/m³)	Mín. de NO2 (µg/m³)	Mín. de O3 (µg/m³)	Mín. de SO2 (µg/m³)
Alcanar			1	1		
Amposta			1	1	4	
Bellver de Cerdanya			1	1	2	1
Cubelles			0	1		0
Gavà	0,2		0	1	0	0
La Sénia					30	
Manlleu			1	1	1	1
Pardines					30	
Reus	0,2	1	1	1	2	1
Sabadell	0,2		2	5	1	1
Sitges			1	1		1
Sort					4	
Tarragona	0,2	1	1	1		1
Terrassa	0,2		1	3	1	1
Tona			1	1	3	
Vic					1	
Vilanova	0,2		1	1	1	1
Total general	**0,2**	**1**	**0**	**1**	**0**	**0**

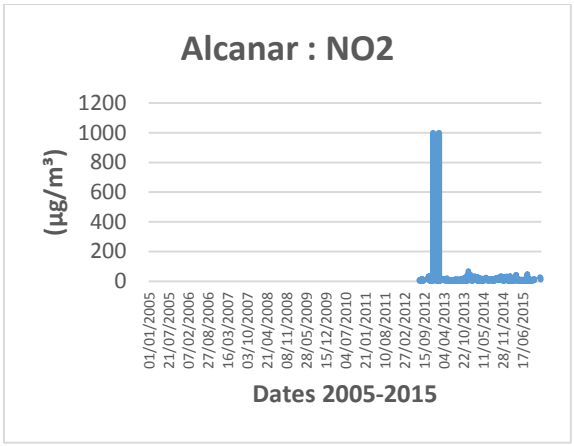
Il·lustració 36

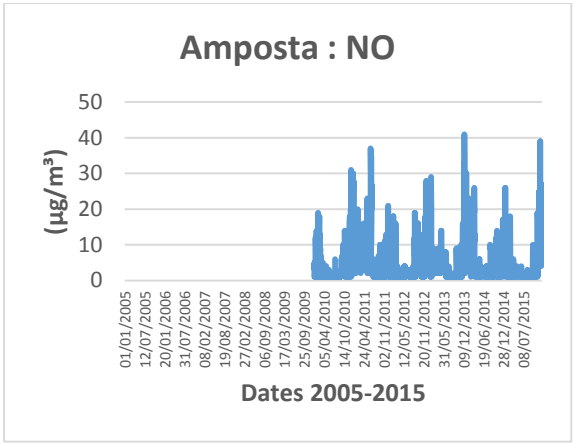
Il·lustracions 35 i 36 : Captures de pantalla taules específiques per contaminant i població de valors màxims i mínims de l'arxiu Excel creat.
Font : 2016 Document Excel Dades lliurat.

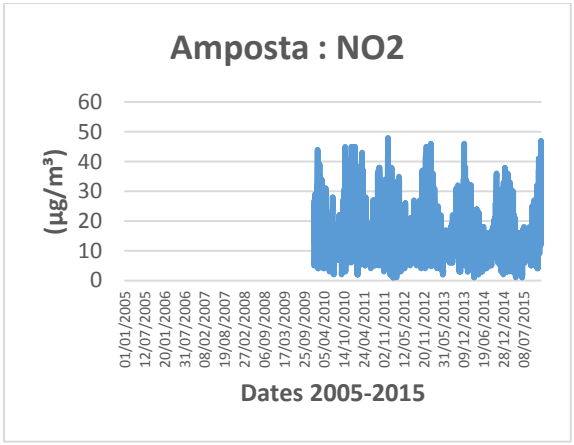
Mitjançant una plantilla personalitzada per la creació de gràfics, s'ha creat un de cada contaminant per població durant el període 2005-2015, on es pot veure l'evolució de les mesures automàtiques (mitjanes) diàries d'aquests i si hi ha algun patró (per exemple de tipus lineal o sinusoïdal) que servirà per tenir més clara la opció a escollir en front la problemàtica dels buits i la manca parcial o absoluta d'informació en segons quines cel·les.

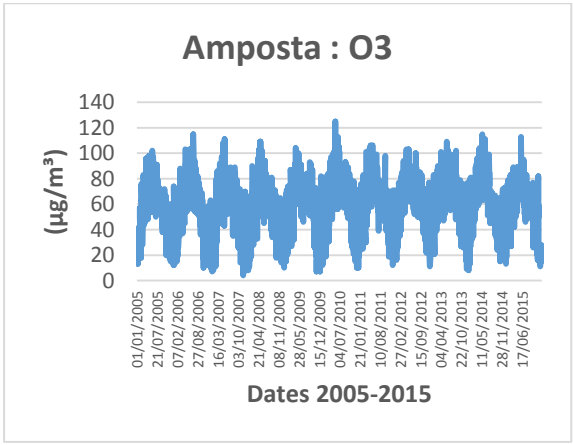
Els següents gràfics tenen a l'eix horitzontal les dates del període 2005-2015 i a l'eix vertical les mesures automàtiques (mitjanes) diàries de cada contaminant en cada població. També queden reflectits els períodes de manca parcial o absoluta d'informació.

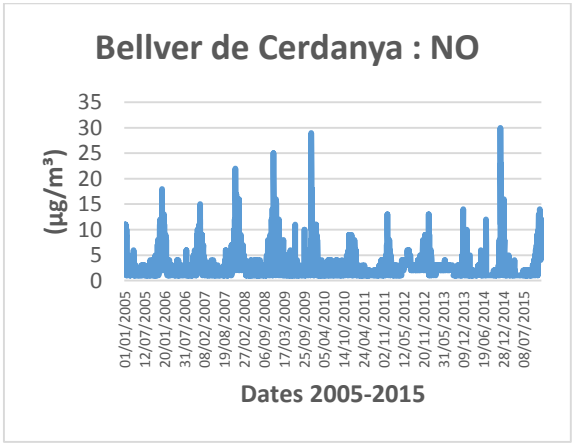


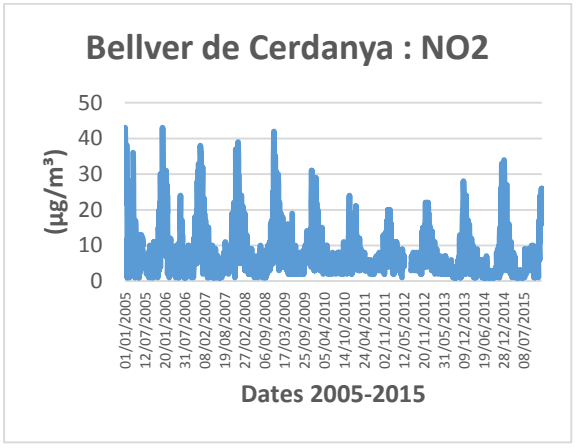


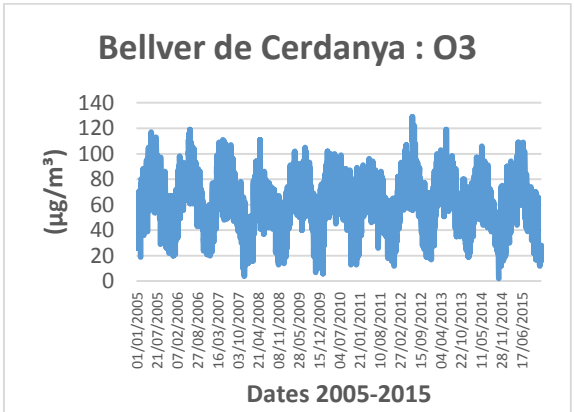


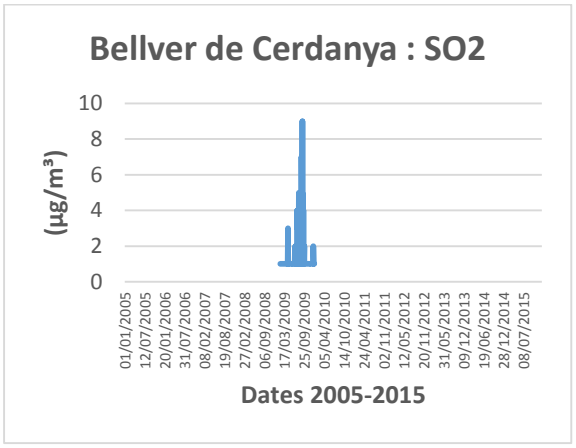


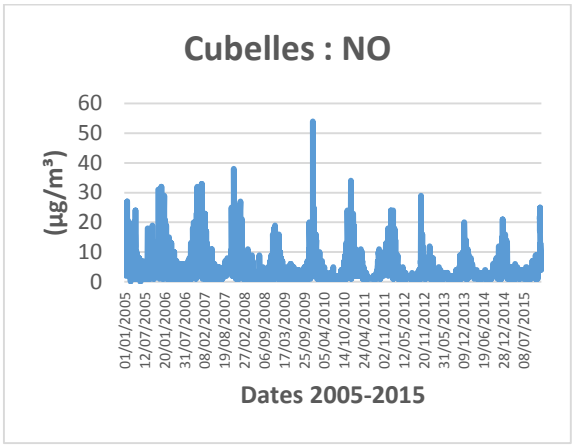


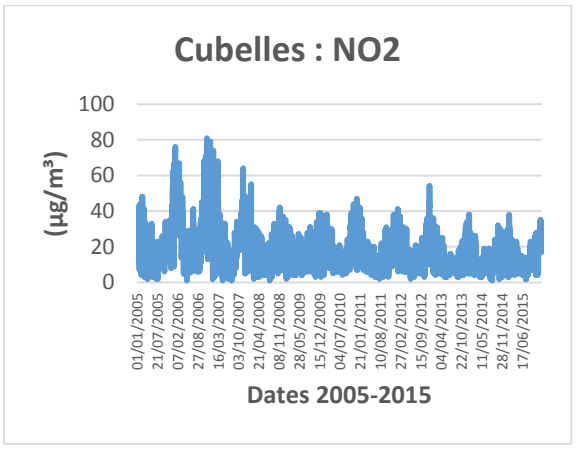


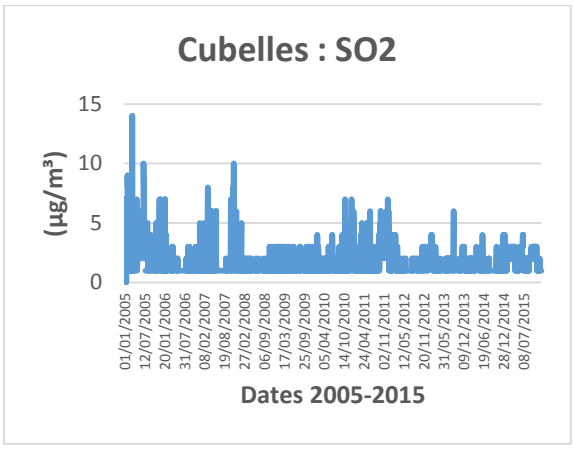

Bellver de Cerdanya : NO2
(µg/m³)
Dates 2005-2015

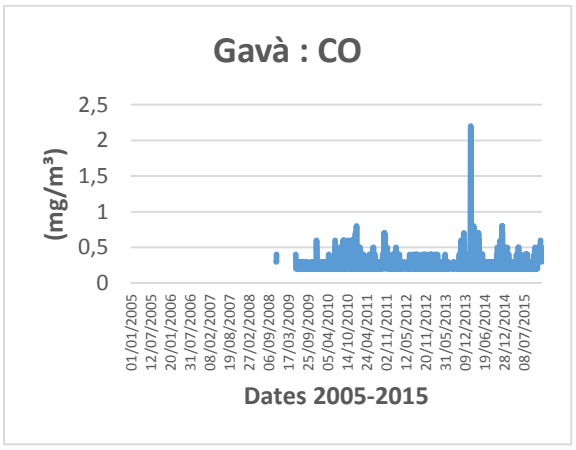

Bellver de Cerdanya : O3
(µg/m³)
Dates 2005-2015

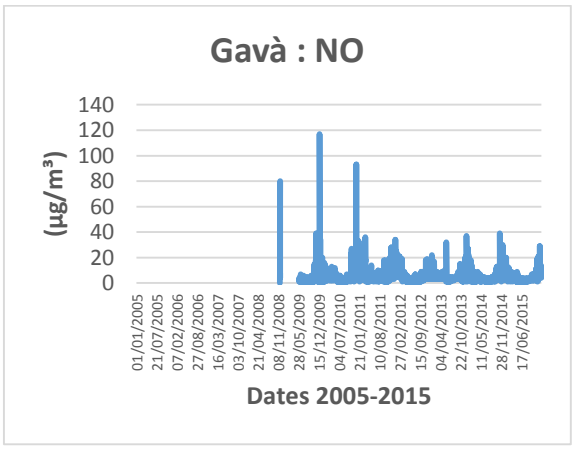

Bellver de Cerdanya : SO2
(µg/m³)
Dates 2005-2015

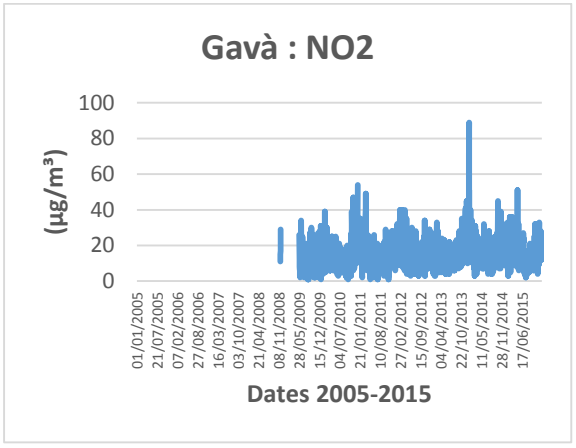

Cubelles : NO
(µg/m³)
Dates 2005-2015

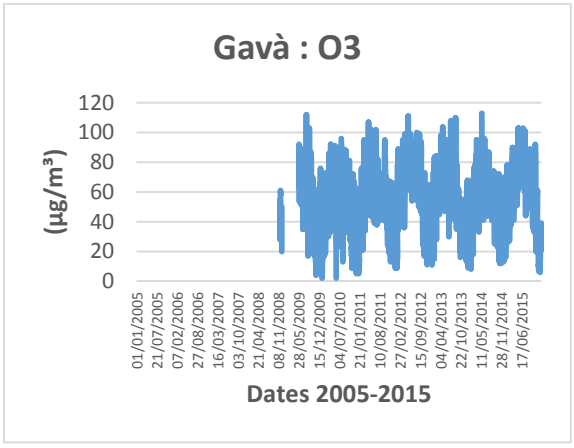

Cubelles : NO2
(µg/m³)
Dates 2005-2015

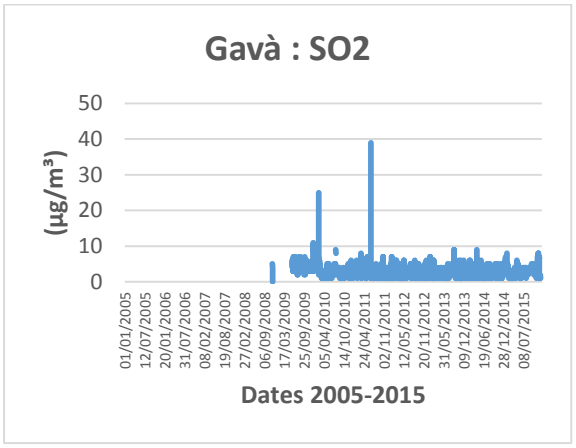

Cubelles : SO2
(µg/m³)
Dates 2005-2015

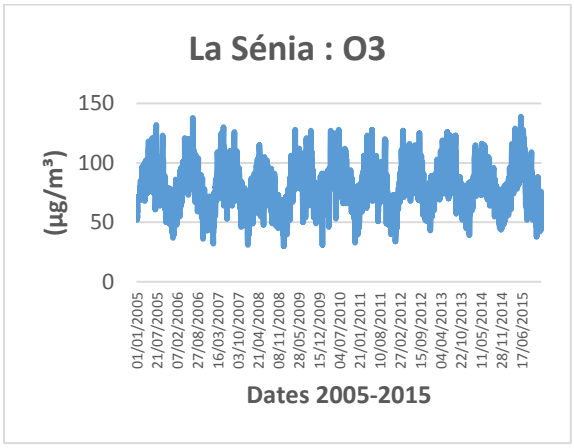

Gavà : CO
(mg/m³)
Dates 2005-2015

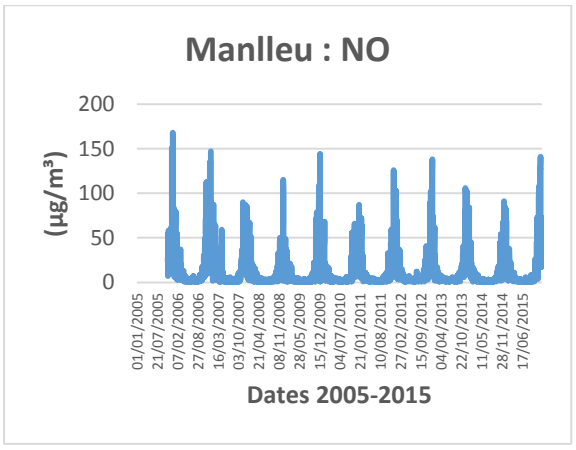

Gavà : NO
(µg/m³)
Dates 2005-2015

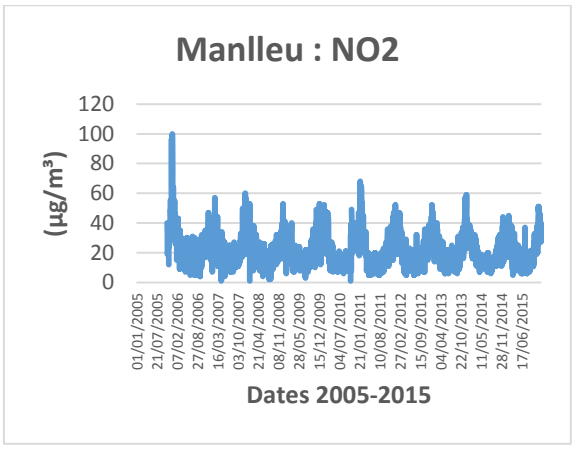

Gavà : NO2
(µg/m³)
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

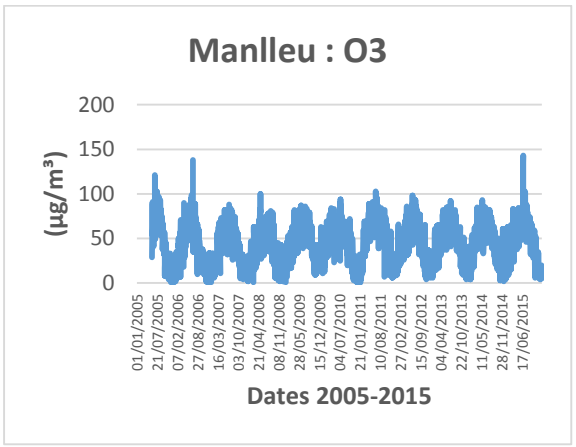

Gavà : O3
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

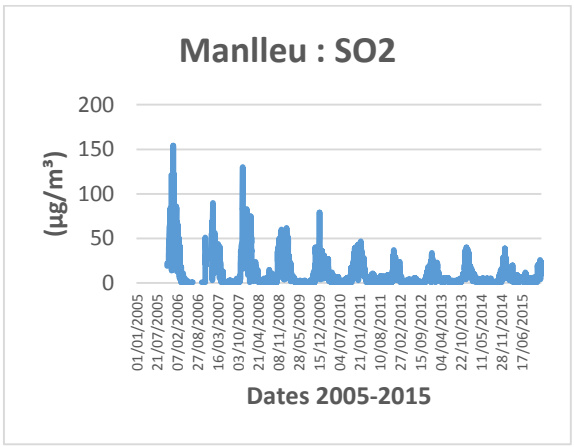

Gavà : SO2
(µg/m³)
50
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

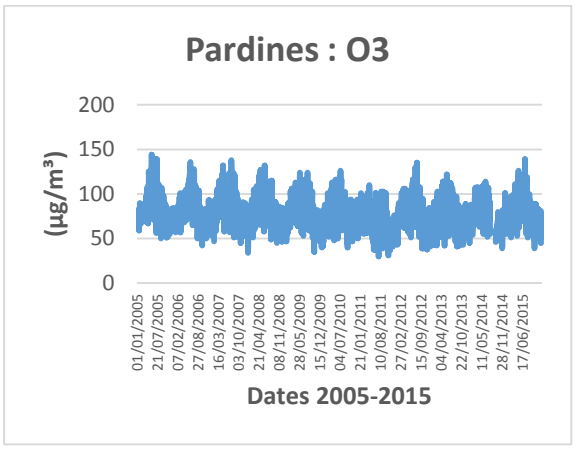

La Sénia : O3
(µg/m³)
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

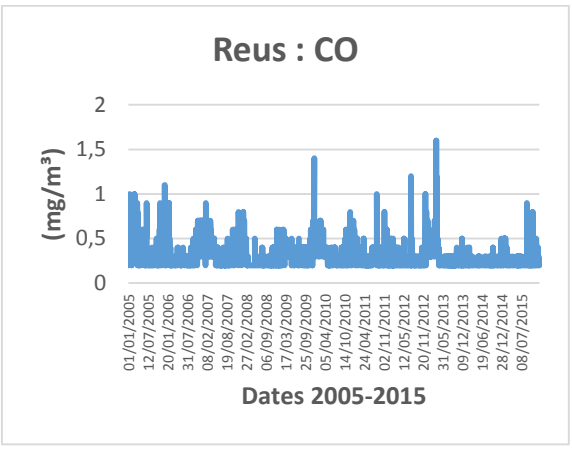

Manlleu : NO
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

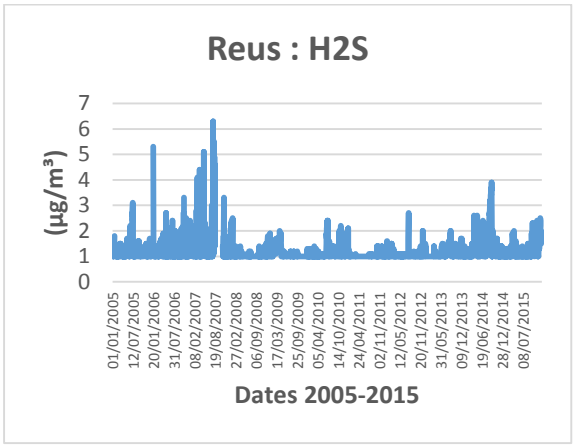

Manlleu : NO2
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

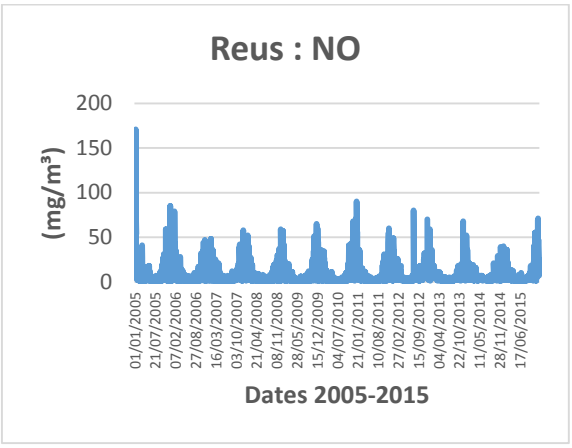

Manlleu : O3
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

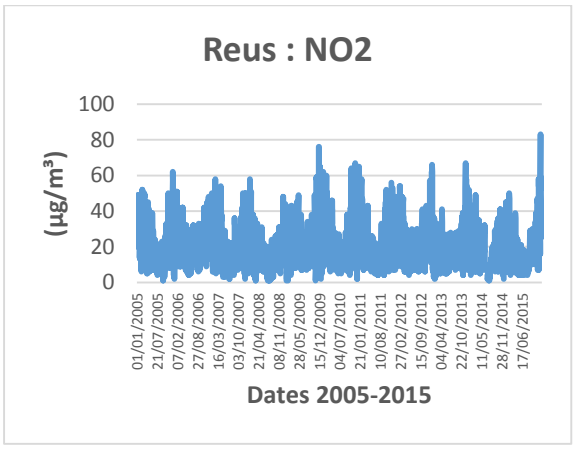

Manlleu : SO2
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

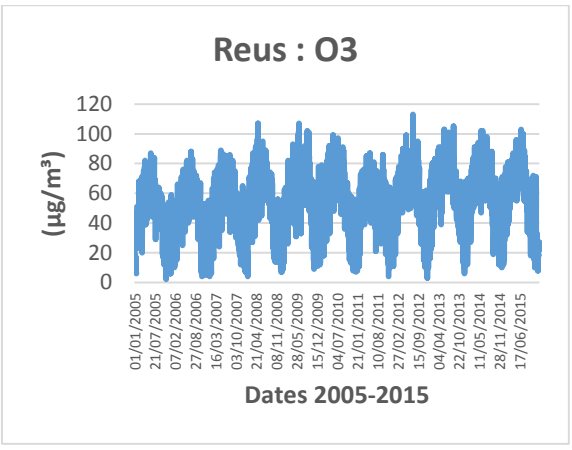

Pardines : O3
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

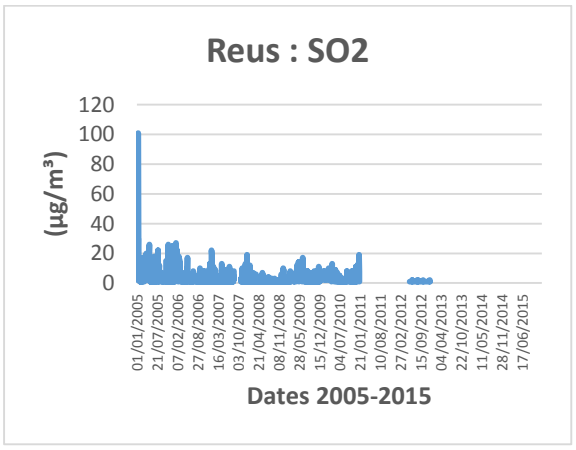

Reus : CO
(mg/m³)
2
1,5
1
0,5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

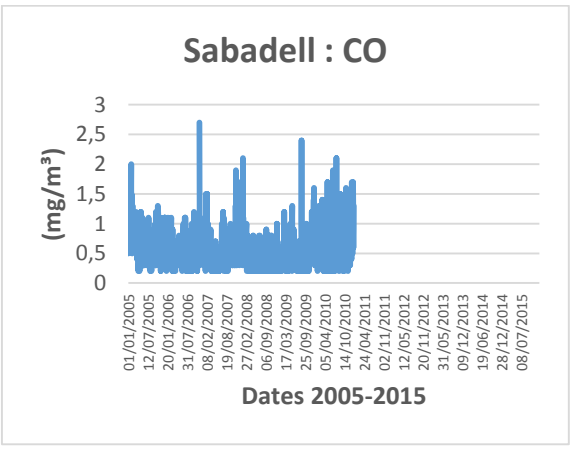

Reus : H2S
(µg/m³)
7
6
5
4
3
2
1
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

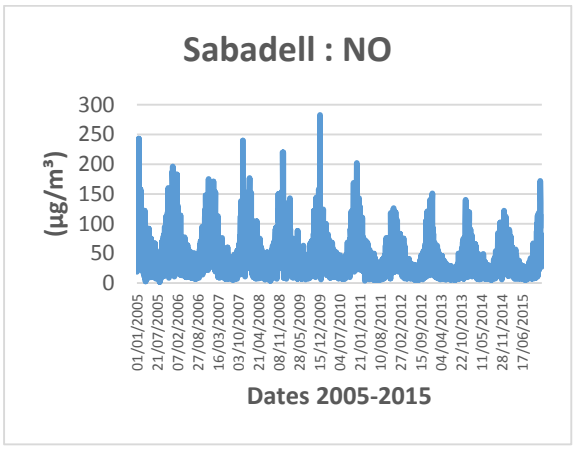

Reus : NO
(mg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

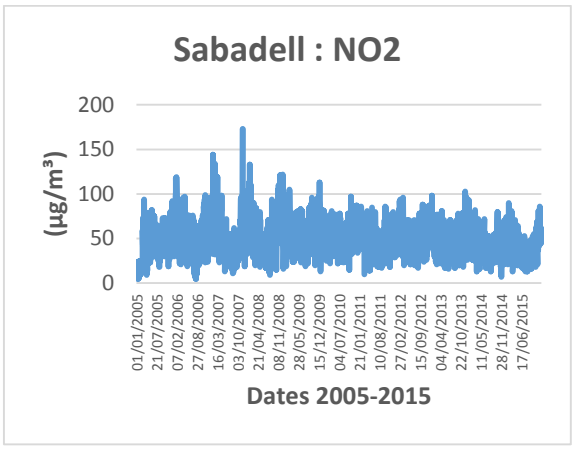

Reus : NO2
(µg/m³)
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

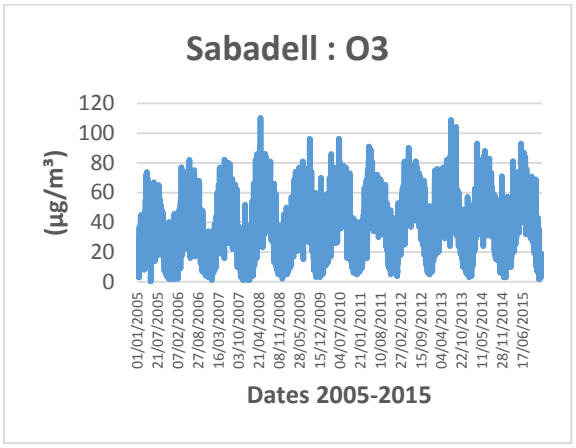

Reus : O3
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

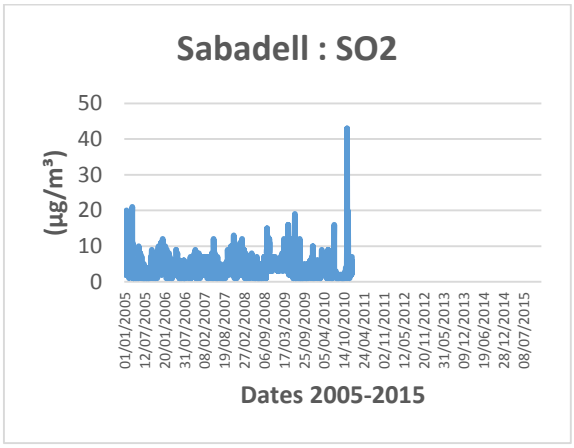

Reus : SO2
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

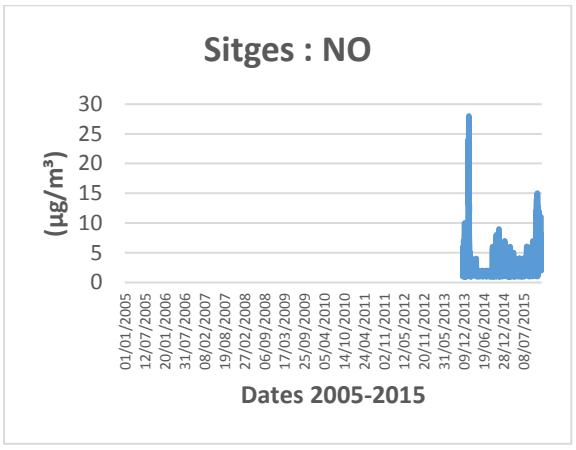

Sabadell : CO
(mg/m³)
3
2,5
2
1,5
1
0,5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

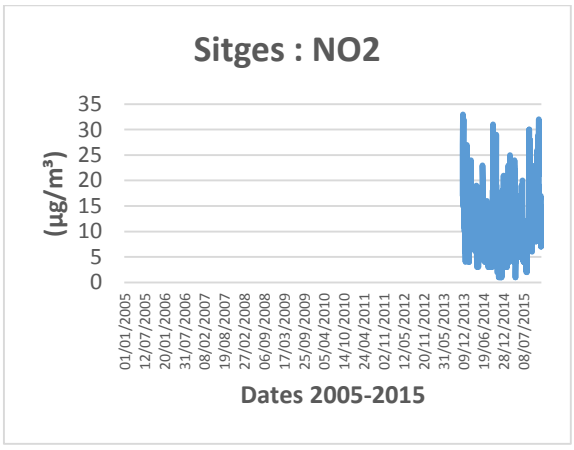

Sabadell : NO
(µg/m³)
300
250
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

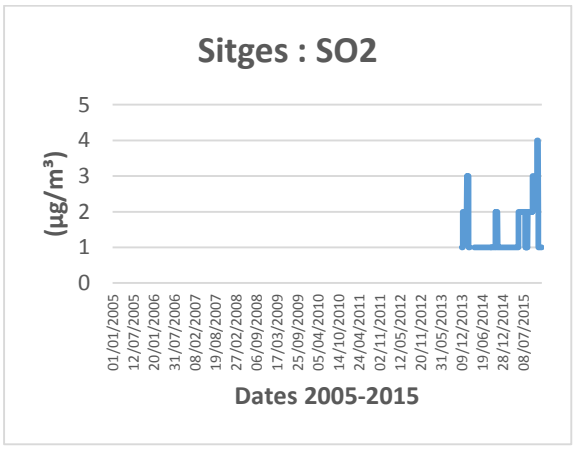

Sabadell : NO2
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

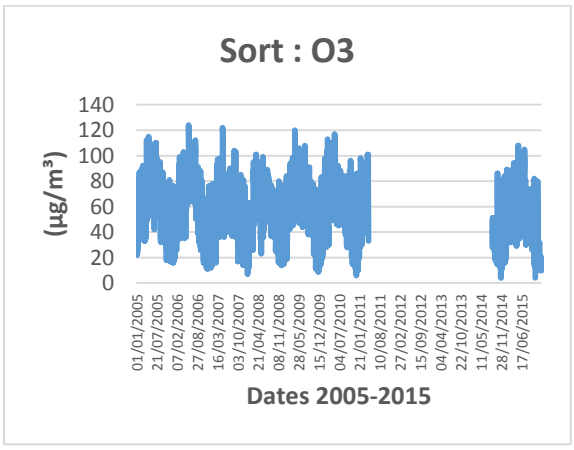

Sabadell : O3
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

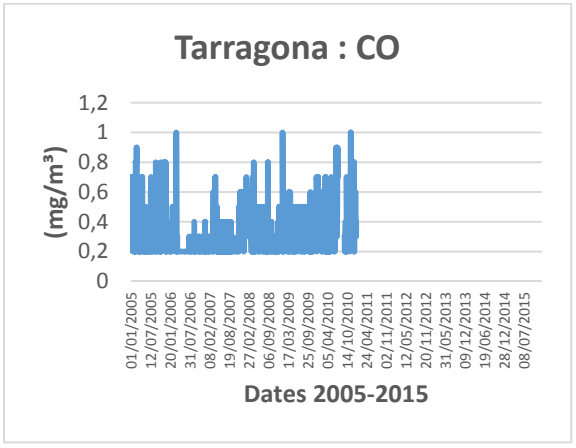

Sabadell : SO2
(µg/m³)
50
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

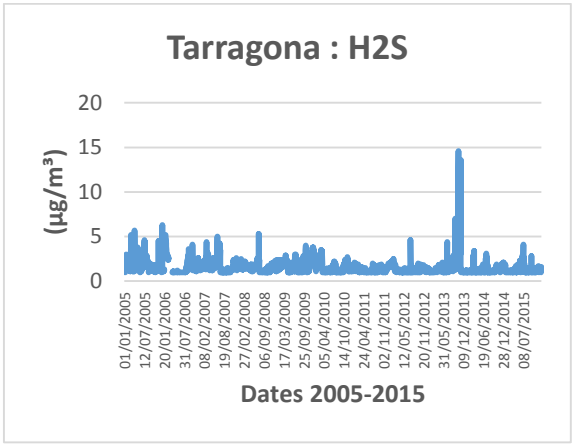

Sitges : NO
(µg/m³)
30
25
20
15
10
5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

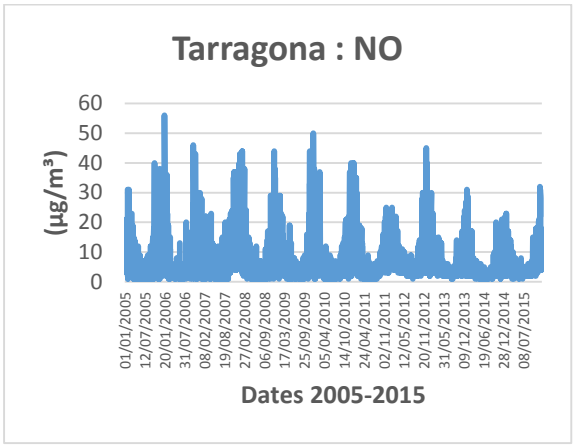

Sitges : NO2
(µg/m³)
35
30
25
20
15
10
5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

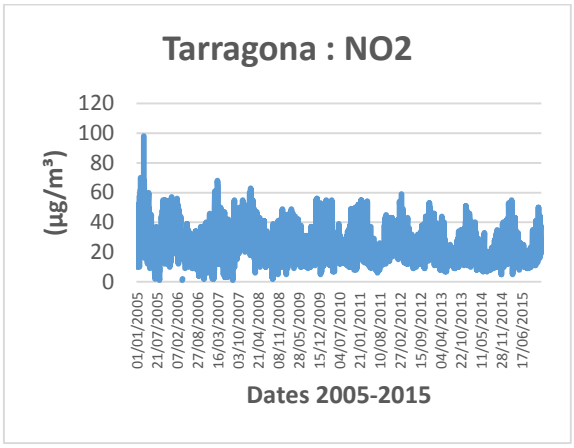

Sitges : SO2
(µg/m³)
5
4
3
2
1
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

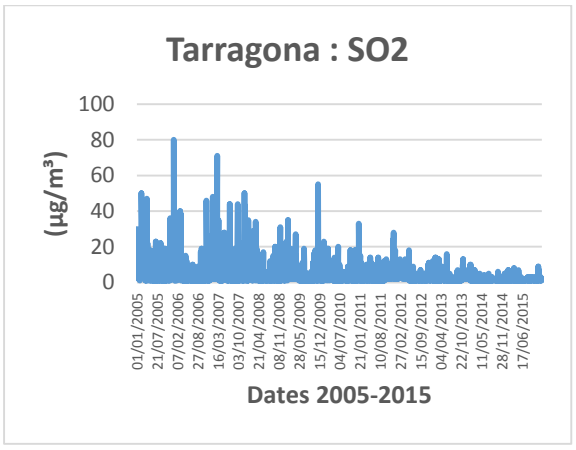

Sort : O3
(µg/m³)
140
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

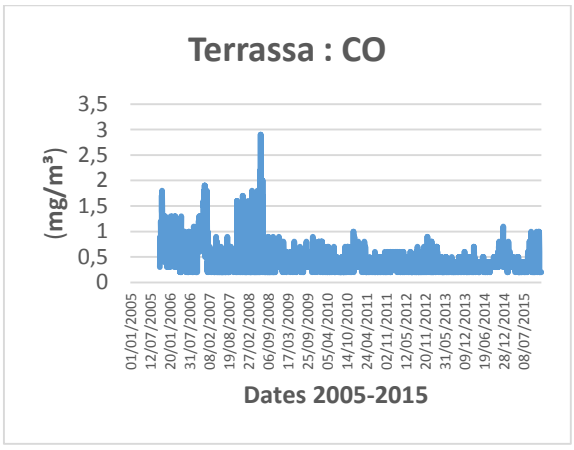

Tarragona : CO
(mg/m³)
1,2
1
0,8
0,6
0,4
0,2
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

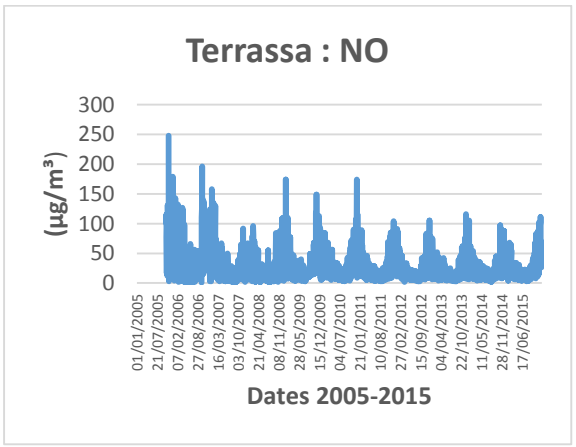

Tarragona : H2S
(µg/m³)
20
15
10
5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

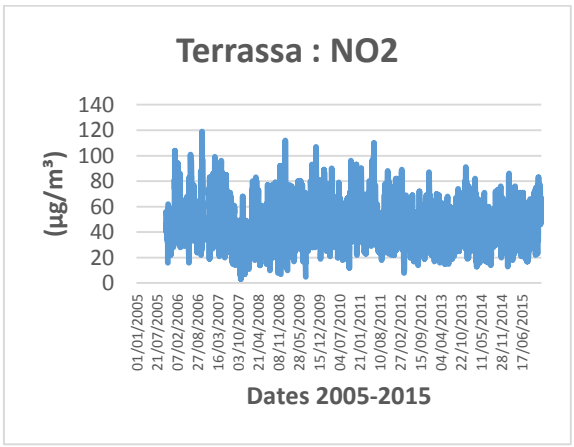

Tarragona : NO
(µg/m³)
60
50
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

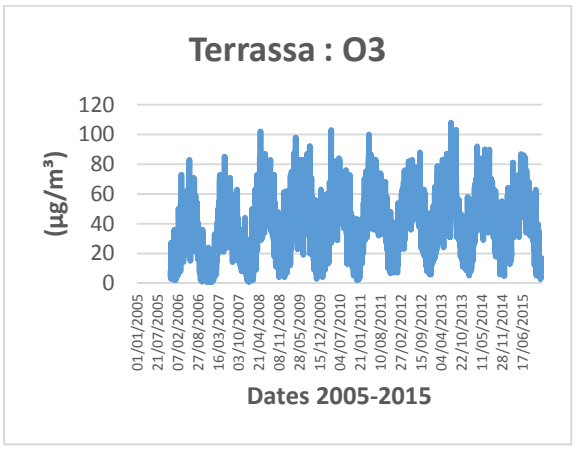

Tarragona : NO2
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

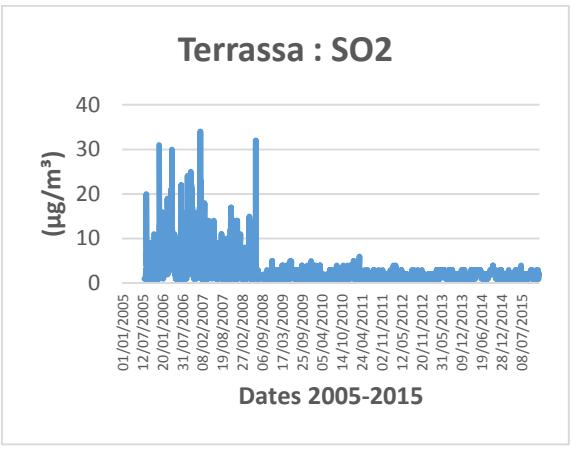

Tarragona : SO2
(µg/m³)
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

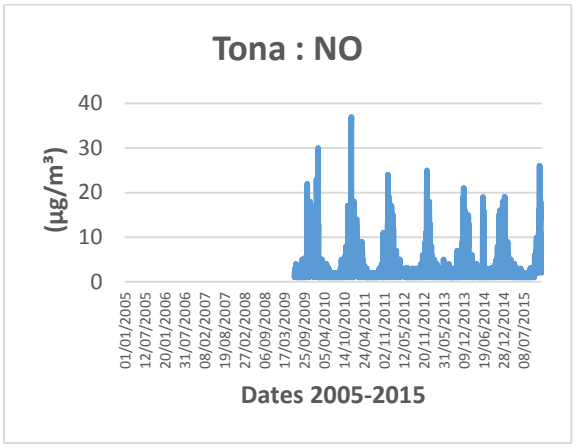

Terrassa : CO
(mg/m³)
3,5
3
2,5
2
1,5
1
0,5
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

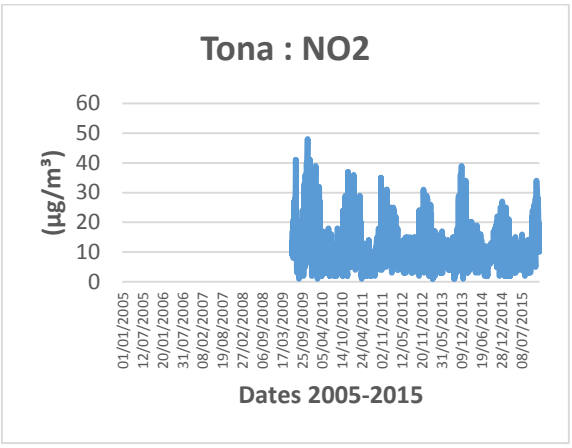

Terrassa : NO
(µg/m³)
300
250
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

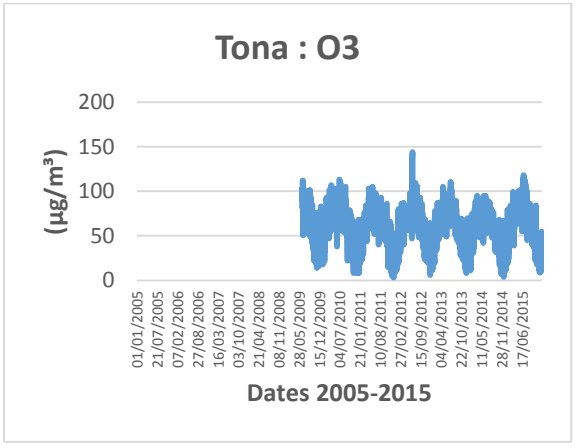

Terrassa : NO2
(µg/m³)
140
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

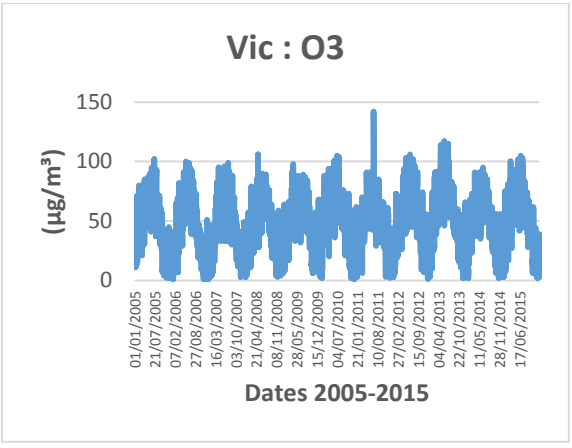

Terrassa : O3
(µg/m³)
120
100
80
60
40
20
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

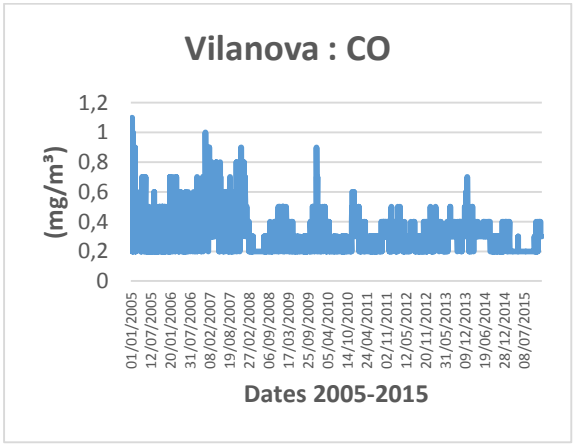

Terrassa : SO2
(µg/m³)
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

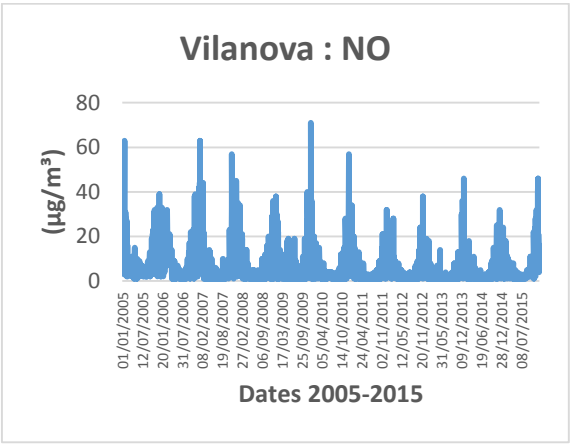

Tona : NO
(µg/m³)
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

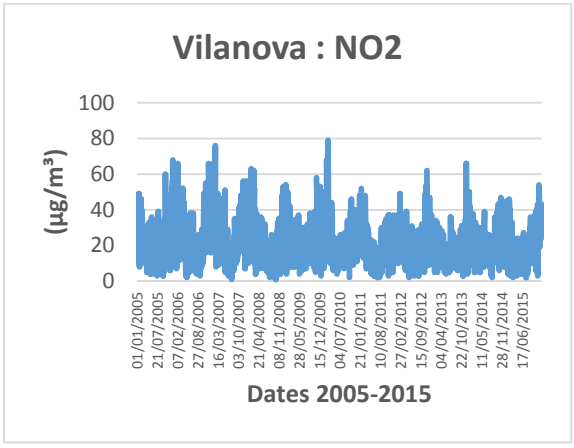

Tona : NO2
(µg/m³)
60
50
40
30
20
10
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015

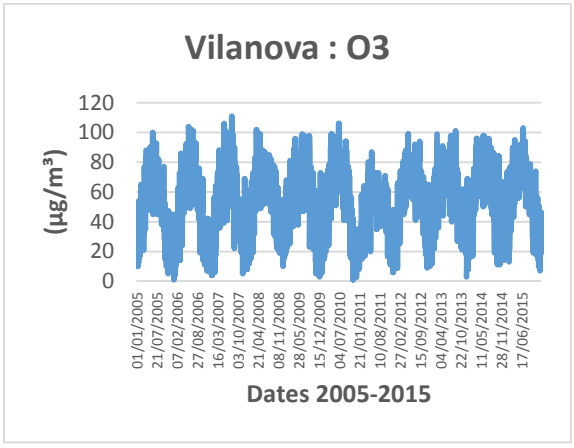

Tona : O3
(µg/m³)
200
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015

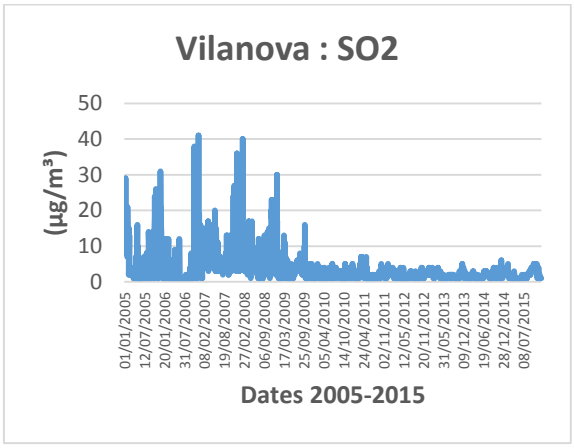

Vic : O3
(µg/m³)
150
100
50
0
01/01/2005
21/07/2005
07/02/2006
27/08/2006
16/03/2007
03/10/2007
21/04/2008
08/11/2008
28/05/2009
15/12/2009
04/07/2010
21/01/2011
10/08/2011
27/02/2012
15/09/2012
04/04/2013
22/10/2013
11/05/2014
28/11/2014
17/06/2015
Dates 2005-2015


Vilanova : CO
(mg/m³)
1,2
1
0,8
0,6
0,4
0,2
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015


Vilanova : NO
(µg/m³)
80
60
40
20
0
01/01/2005
12/07/2005
20/01/2006
31/07/2006
08/02/2007
19/08/2007
27/02/2008
06/09/2008
17/03/2009
25/09/2009
05/04/2010
14/10/2010
24/04/2011
02/11/2011
12/05/2012
20/11/2012
31/05/2013
09/12/2013
19/06/2014
28/12/2014
08/07/2015
Dates 2005-2015







Després de revisar totes les dades, s'ha detectat que les cel·les de les columnes de la capçalera de l'arxiu Excel que tenen per nom el contaminant i la seva unitat de mesura, són les úniques que necessiten tractament, ja que les estadístiques, els valors màxims i mínims i els gràfics detallats, així ho demostren.

Aquestes columnes són les úniques que presenten cel·les amb valors textuals concretament amb les paraules "Sense dades" en lloc de cel·les amb valors numèrics o cel·les buides, també són les úniques que presenten valors impossibles com 999 o valors redundants de 0.

3.2. Neteja de dades

Una part de la preparació de les dades és la neteja d'aquestes processant-les per tal de, completar possibles dades incompletes, eliminar dades redundants o dades incorrectes o inconsistents, arreglar o unificar criteris per errors de transcripció, actualitzar dades envellides o dades procedents de diverses fonts representant un mateix concepte i dades esbiaixades.

S'han netejat les següents cel·les de les columnes de la capçalera de l'arxiu Excel que tenen per nom el contaminant i la seva unitat de mesura, convertint-les en cel·les buides per procedir a la seva posterior transformació si s'escau :

- Les cel·les que contenen les paraules "Sense dades", ja que s'han considerat cel·les amb dades redundants.
- Algunes cel·les que contenen valors erronis de 999, pertanyents a la localitat d'Alcanar (columna de capçalera Població), concretament dels anys 2012 i 2013 (columna de capçalera Any), en les mesures dels contaminants NO i NO2.

- Algunes cel·les que contenen valors erronis/redundants de 0 pertanyents a les localitats de Cubelles i Gavà (columna de capçalera Població), concretament :
  - Any 2005 Cubelles, en les mesures dels contaminants NO i NO2.
  - Any 2008 Gavà, en les mesures dels contaminants NO i SO2.
  - Any 2010 Gavà, en la mesura dels contaminant O3.

3.3. Integració de les dades

Els diversos tipus d'atributs amb què es pot descriure un domini són :

- Numèrics : Prenen valors en els reals o enters o naturals (en general, en un conjunt numèric que pot prendre valors infinits).
- Lògics : Prenen els valors 'Cert' o 'Fals', normalment representats per 0 ('Fals') i 1 ('Cert').
- Categòrics o discrets: els que prenen valors en un conjunt finit, per exemple {1, 2, 3} o {'Baix', 'Mitjà', 'Alt'}.
- Ordenats : Tenen una relació d'ordre entre si. Per exemple, en principi els colors no tenen cap ordre establert; per tant, si un atribut pren els valors {'Vermell', 'Verd', 'Blau'}, no necessàriament hem de considerar que és un atribut amb valors ordenats.

Es poden emprar també altres criteris com les dimensions del rang de valors i l'escala de mesura emprada.

Respecte al rang de valors de la variable, es distingeixen els tipus de variables següents :

- Les variables contínues : Tenen un conjunt de valors infinit no numerable. Cas típic : els nombres reals.
- Les variables discretes : Prenen valors en un conjunt finit o, com a molt, infinit numerable. Exemple: els enters.
- Les variables binàries : Variables discretes que només poden prendre dos valors. Cas típic : la representació dels valors lògics 'Cert' i 'Fals' mitjançant els nombres 1 i 0.

El primer pas que s'ha fet per homogeneïtzar les dades ha estat reunir-les en un sol arxiu.

Posteriorment, amb la valoració de les estadístiques, valors màxims i mínims i els gràfics de l'apartat 3.1, s'ha decidit que:

- El rang de valors numèrics dels contaminants serà :

Contaminants	CO (mg/m³)	H2S (µg/m³)	NO (µg/m³)	NO2 (µg/m³)	O3 (µg/m³)	SO2 (µg/m³)
Valors Mínims	0,1	1	1	1	1	1
Valors Màxims	3	15	300	200	150	175

Taula 5 : Rang de valors numèrics dels contaminants

- Respecte als valors mínims :
  - En el CO, 0,1, una dècima menys del mínim trobat.
  - En la resta de contaminants, 1.
- Respecte als valors màxims :
  - En el CO, 3, una dècima més del màxim trobat.
  - En el H2S, 15, quatre punts més del màxim trobat degut a la seva fluctuació.
  - En el NO, 300, disset punts més del màxim trobat degut a la seva fluctuació.
  - En el NO2, 200, vint-i-set punts més del màxim trobat degut a la seva fluctuació.
  - En el O3, 150, sis punts més del màxim trobat degut a la seva fluctuació.
  - En el SO2, 175, vint-i-un punts més del màxim trobat degut a la seva fluctuació.

- Els mesos de les cel·les de la columna de capçalera Mes de l'arxiu Excel, són noms, és a dir, una variable categòrica, això s'ha fet així, ja que molts algorismes empren les distàncies per establir criteris i per exemple, allunyarien molt el mes 1 del mes 12, cosa que no ha de ser així, ja que gener i desembre estan a tocar.

3.4. Transformació de les dades

La transformació de valors es defineix bàsicament com les tècniques per a canviar els valors sense perdre informació i fer que puguin ser tractats pel mètode que interessi.

No sempre les dades estan en la forma més adequada per a poder aplicar els mètodes que calen per a la tasca que s'ha de portar a terme i el model que es vol obtenir. En general ens trobarem que haurem d'efectuar alguna d'aquestes transformacions :

- Dades numèriques a categòriques.
- Dades categòriques a numèriques.
- Simplificació de valors.
- Agrupació de valors continus.
- Normalització de dades : Posar els valors numèrics en un interval determinat.
- Afegir-hi una etiqueta que digui a quina classe pertany un registre.
- Expansió d'un atribut pel fet que el valor d'un atribut pot prendre valors en un conjunt limitat de categories.
- Derivació de dades : Es poden emprar els atributs de les dades existents per a derivar atributs nous (i generar, de fet, un conjunt nou de dades) que siguin més útils per a la mena d'estudi de mineria de dades que es porti a terme.
- Fusió de dades o enriquiment : Pot interessar afegir dades procedents d'altres relacions o, fins i tot, altres bases de dades aportades des d'altres fonts.

No s'aplicarà cap tipus de normalització ni cap anàlisi de distàncies, ja que no és bo normalitzar mesures de contaminants i no té cap sentit fer un anàlisi de distàncies, ja que la distància s'usa per agrupar contaminants segons una jerarquia i en aquest projecte bàsicament es vol estudiar l'evolució temporal.

Després de revisar les estadístiques, valors màxims i mínims i gràfics i veure que la majoria d'aquests darrers tenen forma sinusoïdal, s'ha decidit :

- Primer de tot s'aplicaran les transformacions exposades en els apartats anteriors (3.2 Neteja de dades i 3.3 Integració de dades).
- La problemàtica de les cel·les buides se soluciona de la següent forma :
  - Si no hi ha cap valor numèric del contaminant en tot el període 2005-2015, es deixaran totes les cel·les en blanc.
  - Si no hi cap valor numèric del contaminant en tot un any, es substituiran totes les cel·les d'aquest any, per la mitjana de tots el altres anys que sí contenen valors numèrics del contaminant.
  - Si no hi cap valor numèric del contaminant en tot un mes, es substituiran totes les cel·les d'aquest mes, per la mitjana de tots el altres mesos del mateix nom que sí contenen valors numèrics del contaminant.
  - Si no hi cap valor numèric del contaminant en tot un dia, es substituirà per la mitjana de tots el altres dies del mateix número i mes que sí contenen valors numèrics del contaminant.
  - Aquestes mesures s'aplicaran per ordre, és a dir primer anys després mesos i després dies.
  - Casos especials en la problemàtica dels buits :
    - Si no hi ha cap valor numèric en tots els mesos del mateix nom durant el període 2005-2015, però hi han dades numèriques en els altres mesos d'aquest període, es substituirà per la mitjana anual del contaminant.

- Si no hi ha cap valor numèric en tots els dies del mateix número i mes durant el període 2005-2015, però hi ha dades numèriques en els altres dies del mes d'aquest període, es substituirà per la mitjana mensual del dia del contaminant.

S'ha optat per aquesta solució (omplir el màxim de cel·les possibles amb criteri) ja que el programari lliure Weka té una pestanya anomenada Preprocess, des de la que es pot accedir a diferents filtres que permeten preparar les dades carregades al programa i aquests es poden fer servir, si s'escau, al llarg del treball.

3.5. Reducció de les dades

Conèixer les principals tècniques de preparació de dades i adonar-se de la seva conveniència per a cada tipus diferent de tasca de construcció de models és fonamental. Una d'elles és la reducció de la dimensionalitat mitjançant la selecció d'atributs.

- Reducció de dimensionalitat : Eliminar casos dins el conjunt de dades original, o bé eliminar atributs, o totes dues coses al mateix temps per a obtenir models de la mateixa qualitat amb menys esforç computacional.

Els mètodes de reducció de dimensionalitat cerquen justament treballar amb menys dades i obtenir els mateixos resultats.

Des de bon principi (les seves dades ja ni s'han adjuntat en l'arxiu de tipus Excel creat) s'han descartat els contaminants PM10 i PST per centrar-se en els altres sis (*CO, H2S, NO, NO2, O3, i SO2)*, això ja és un primer pas en la reducció de les dades.

Es podrien reduir les dades eliminant per exemple la zona redefinida, però no es farà ja que pot aportar un punt de vista diferent de l'oficial i pot permetre la comparació de resultats.

3.6. Construcció de l'arxiu .arff

Adonar-se que, tot i la potència del suport de maquinari i programari que poden tenir els mètodes de mineria de dades, la seva complexitat requereix molt sovint introduir simplificacions per a mantenir el cost computacional dins uns nivells adequats sense comprometre la qualitat final dels models obtinguts, és clau alhora de prepara les dades.

Un cop seleccionades totes les dades, aquestes s'han de preparar perquè se les pugui aplicar els mètodes o eines que construiran el model volgut. Aquesta fase, encara que sembli senzilla, conjuntament amb la de selecció de dades, s'emporta el 80% de l'esforç en els projectes de mineria de dades de nova implantació.

L'objectiu principal de la preparació de dades consisteix a organitzar-les de manera que puguin ser processades pels programes de construcció de models que s'hagin escollit i, al mateix temps, assegurar que les dades estan de manera que s'obtingui el millor model que es pot extreure del conjunt de dades.

Els objectius del projecte de mineria de dades determinen el tipus de model de coneixement a extreure. També determinen que una dada o una relació entre dades sigui significativa o no en relació amb els objectius marcats.

En aquest projecte, per construir un arxiu .arff que abasti el màxim de possibilitats s'ha procedit de la següent forma :

- Crear un arxiu Excel de nom Dades Preparades copia de l'arxiu Excel Dades, sense els fulls Estadístiques i Gràfics.
- Ordenar les dades del full Dades, per Població i Data.
- Reemplaçar les cel·les del full Dades, amb les paraules "Sense dades" per cel·les en blanc.
- Reemplaçar les cel·les del full Dades, amb valor 999 per cel·les en blanc.
- Reemplaçar les cel·les del full Dades, amb valor 0 per cel·les en blanc.
- Crear un nou full de nom Mitjanes amb tres taules dinàmiques que permeten fer :
  - Les mitjanes anuals de cada contaminant per població i període 2005-2015.
  - Les mitjanes mensuals de cada contaminant per població i període 2005-2015.
  - Les mitjanes de cada dia de cada mes per població i període 2005-2015.
- Crear 17 fulls, un per cada població amb el nom d'aquesta, que contenen les dades extretes de les taules dinàmiques creades en els full Mitjanes.
- Substituir tots els espais en blanc que s'han de substituir, en el full Dades, per les dades dels fulls que tenen per nom, el nom de les poblacions.
- Homogeneïtzar el format de les dades del full Dades a 2 decimals.
- Crear un arxiu Word de nom Dades Preparades amb les dades del full Dades de l'arxiu Excel Dades Preparades.
- Reemplaçar totes les comes decimals dels valors dels contaminants per punts.
- Canviar totes les tabulacions de l'arxiu Word Dades preparades, per comes, ja que, els camps en un arxiu .arff han d'estar separats per comes.
- Obrir l'arxiu Dades_Preparades.txt amb el Notepad++, reemplaçar tots el espais buits entre paraules per guions baixos.
- Reemplaçar també la manca de valors, és a dir quan apareix",," per ,?, perquè el Weka entengui que no hi ha valor.

Els arxius .arff (Attribute-Relation File Format), són el tipus de format reconegut per Weka, el fitxer .arff està definit per tres parts, la primera és la capçalera, en la qual es defineix el nom de la relació; la segona consta de la declaració d'atributs, on apareixen els atributs que compondran l'arxiu juntament amb el seu tipus de dada; finalment es troba la part de dades, on es declaren les dades que componen la relació, separant per comes els atributs i són salts de línia les relacions.

Per poder emprar dades provinents d'altres formats, cal transformar les fonts de dades en format arff, podent convertir fitxers en text contenint un registre per línia i amb els atributs separats amb comes (format .csv) a fitxers arff.

- Crear un arxiu .txt de nom Dades_Preparades que contindrà tots les dades de l'arxiu Word de nom Dades Preparades en la secció @DATA i en la secció @RELATION contaminants, contindrà :

  - @ATTRIBUTE **Data DATE "dd/MM/yyyy"**
  - @ATTRIBUTE **Dia {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}**
  - @ATTRIBUTE **Mes** {**Gener,Febrer,Març,Abril,Maig,Juny,Juliol,Agost,Setembre,Octubre,Novembre,Decembre}**
  - @ATTRIBUTE **Any {2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015}**
  - @ATTRIBUTE **CO REAL**
  - @ATTRIBUTE **H2S REAL**
  - @ATTRIBUTE **NO REAL**
  - @ATTRIBUTE **NO2 REAL**
  - @ATTRIBUTE **O3 REAL**
  - @ATTRIBUTE **SO2 REAL**
  - @ATTRIBUTE **Poblacio STRING**
  - @ATTRIBUTE **Estacio STRING**

- @ATTRIBUTE **Comarca STRING**
- @ATTRIBUTE **ZQA {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}**
- @ATTRIBUTE **Zona_Oficial STRING**
- @ATTRIBUTE **Zona_Redefinida STRING**

- Finalment, canviar-li l'extensió a l'arxiu Dades_Preparades.txt per .arff, és a dir, Dades_Preparades.arff.

En la següent captura de pantalla es mostra l'aspecte general d'aquest arxiu :

```
@RELATION contaminants


@ATTRIBUTE Data DATE "dd/MM/yyyy"
@ATTRIBUTE Dia {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}
@ATTRIBUTE Mes {Gener,Febrer,Març,Abril,Maig,Juny,Juliol,Agost,Setembre,Octubre,Novembre,Desembre}
@ATTRIBUTE Any {2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015}
@ATTRIBUTE CO REAL
@ATTRIBUTE H2S REAL
@ATTRIBUTE NO REAL
@ATTRIBUTE NO2 REAL
@ATTRIBUTE O3 REAL
@ATTRIBUTE SO2 REAL
@ATTRIBUTE Poblacio STRING
@ATTRIBUTE Estacio STRING
@ATTRIBUTE Comarca STRING
@ATTRIBUTE ZQA {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@ATTRIBUTE Zona_Oficial STRING
@ATTRIBUTE Zona_Redefinida STRING


@DATA


01/01/2005,1,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
02/01/2005,2,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
03/01/2005,3,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
04/01/2005,4,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
05/01/2005,5,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
06/01/2005,6,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
07/01/2005,7,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
08/01/2005,8,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
09/01/2005,9,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
10/01/2005,10,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
11/01/2005,11,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
12/01/2005,12,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
13/01/2005,13,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
14/01/2005,14,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
15/01/2005,15,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
16/01/2005,16,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
17/01/2005,17,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
18/01/2005,18,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
19/01/2005,19,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
20/01/2005,20,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
21/01/2005,21,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
22/01/2005,22,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
23/01/2005,23,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
24/01/2005,24,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
25/01/2005,25,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
26/01/2005,26,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
27/01/2005,27,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
28/01/2005,28,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_de_l'Ebre,Delta_de_l'Ebre
```

Il·lustració 37 : Captura de pantalla arxiu .arff.
Font : 2016 Document Dades_Preparades.arff. Capçalera i Dades, lliurat.

Tota la tasca realitzada en els apartats d'aquest capítol es tradueix en un arxiu en format .arff que intenta abastar el màxim d'algorismes possibles amb el mínim de cel·les buides.

Aquesta tasca ha permès emplenar moltes cel·les amb els criteris descrits anteriorment per assolir la màxima productivitat possible dels algorismes amb el programari lliure Weka.

4. Mineria de dades

4.1. El programari lliure Weka 3.8.0

Hi ha molts programes per poder treballar algorismes de mineria de dades. Alguns d'ells són privatius i d'altres són de codi lliure. D'aquests últims hi ha diferents estudis de quins són els millors o més emprats.

De la font publicada el 16 de abril del 2016 en "EL RINCÓN DE JMACOE" : http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/, indiquen que els cinc millors són :

- **Orange** : http://orange.biolab.si/
- **RapidMiner** : http://rapid-i.com/content/view/181/190/
- **Weka** : http://www.cs.waikato.ac.nz/~ml/weka/
- **JHepWork** : http://jwork.org/jhepwork/
- **Knime** : http://www.knime.org/

Weka (Waikato Environment for Knowledge Analysis), en català «entorn per l'anàlisi del coneixement de la Universitat de Waikato», és un programari que ha estat desenvolupat a la universitat de Waikato (Nova Zelanda) sota llicència GPL la qual cosa ha impulsat que sigui una de les suites més emprades en l'àrea en els últims anys. El seu nom prové d'un ocell (Gallirallus australis) famós per la seva curiositat.

El programari consta d'una col·lecció d'algorismes d'aprenentatge automàtic per a tasques de mineria de dades. Els algorismes, bé es poden aplicar directament a un conjunt de dades o cridades des del seu propi codi Java. Weka conté eines per a les dades pre-processament, classificació, regressió, clustering, regles d'associació, i la visualització.

En aquest projecte s'ha escollit Weka, ja que és un dels millors programaris lliures apresos i practicats en l'assignatura 05.584 Mineria de Dades de la UOC.

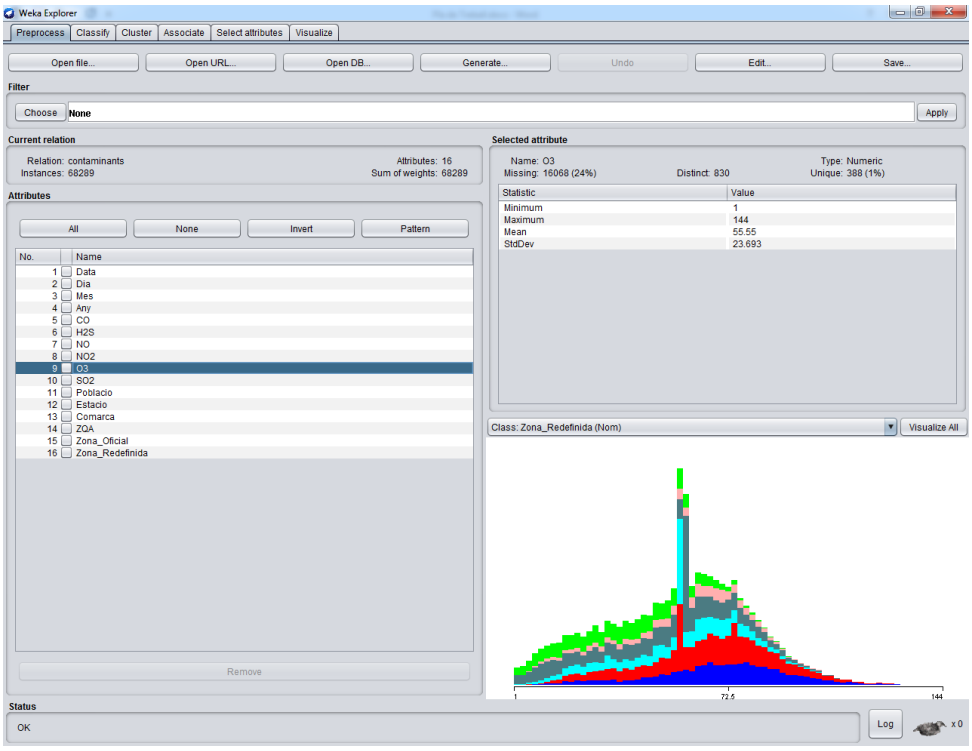
La versió de Weka 3.8.0 defineix 5 entorns de treball (Applications), que es detallen a continuació i es visualitzen en la il·lustració següent que acompanya el text :

- **Explorer** : Entorn visual, que ofereix una interfície gràfica per a l'ús dels paquets, és amb el que es treballarà en aquest projecte.
- **Experimenter** : Entorn centrat en l'automatització de tasques, s'empra bàsicament per facilitar la realització d'experiments a gran escala.
- **KnowledgeFlow** : Permet generar projectes de mineria de dades mitjançant la generació de fluxos d'informació.
- **Workbench** : Aquesta és una interfície gràfica unificada que combina els altres tres (i qualsevol plugin que l'usuari hagi instal·lat) en una sola aplicació. El banc de treball és altament configurable, el que permet a l'usuari especificar quines aplicacions i plugins apareixeran juntament amb els ajustos relatius als mateixos.
- **Simple CLI** : Entorn de consola, per invocar directament amb Java els paquets de Weka.

Il·lustració 38 : Captura de pantalla entorns Weka.
Font : 2016 Programari instal·lat al PC.

Existeixen 6 sub-entorns d'execució dins de l'opció Explorer del Weka 3.8.0, es poden observar aquestes opcions en les pestanyes de la part superior de la il·lustració següent :

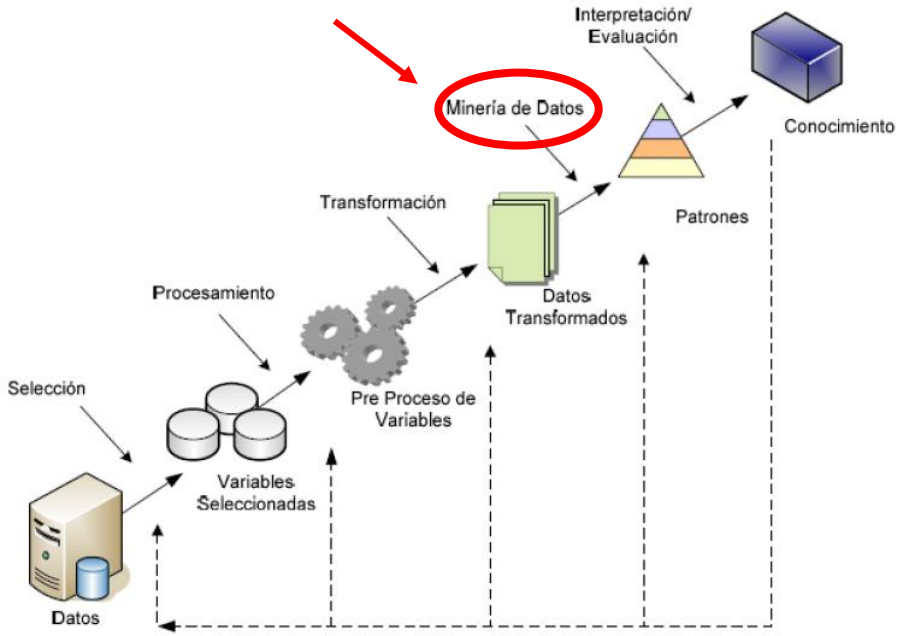
- **Preprocess**: Inclou les eines i filtres per carregar i manipular les dades, permet realitzar manipulacions sobre les dades aplicant filtres. Es poden aplicar en dos nivells: atributs i instàncies. A més les operacions de filtrat poden aplicar-se en cascada, de manera que l'entrada de cada filtre és la sortida d'haver aplicat l'anterior filtre.
- **Classify** : Permet accedir a les tècniques de classificació i regressió.
- **Cluster** : Integra diversos mètodes d'agrupament.
- **Associate** : Inclou unes poques tècniques de regles d'associació.
- **Select attributes** : Permet aplicar diverses tècniques per a la reducció del nombre d'atributs.
- **Visualize** : Permet estudiar el comportament de les dades mitjançant tècniques de visualització, representant gràfiques 2D que relacionen parells d'atributs.



Il·lustració 39 : Captura de pantalla sub-entorns Weka.
Font : 2016 Programari instal·lat al PC.

4.2. Millora de l'arxiu .arff

La mineria de dades, fase en la que ens trobem del projecte, és una de les darreres etapes del procés d'extracció de coneixement tal i com es recorda en la següent il·lustració :



Il·lustració 40 : Etapes del procés d'extracció de coneixement.
Font : FAYYAD, Usama, PIATETSKY, Gregory, Smyth, Padharic.

Un cop descrit el programari que s'ha emprat, a l'obrir l'arxiu Dades_Preparades.arff al Weka i fent una ràpida exploració, s'ha presentat un problema important, al no poder emprar la majoria d'algorismes que proposa el programa, la raó fonamental ha estat que uns quants dels atributs eren de tipus **STRING**.

Per tant, el primer que s'ha fet abans de seleccionar els algorismes que s'aplicaran ha estat solucionar aquest contratemps, per fer-ho s'ha creat un segon arxiu .arff anomenat Dades_Preparades2, en el qual s'han fet els següents canvis respecte l'anterior :

- Els cinc atributs que eren definits com **STRING** (**Poblacio, Estacio**, **Comarca, Zona_Oficial** i **Zona_Redefinida**) han estat canviats a categòrics, per poder realitzar aquests canvis s'ha hagut de:
  - Canviar els espais per un guió baix (exemple: Camp de Tarragona per Camp_de_Tarragona).
  - Suprimir l'apòstrof de la paraula "de l'Ebre" modificant-la per "del_Ebre".

Tot això ha modificat la secció @RELATION (capçalera) i la secció @DATA (dades de l'arxiu) tal i com es mostra en les següents captures de pantalla del nou arxiu Dades_Preparades2.arff :

@ATTRIBUTE Poblacio {Alcanar,Amposta,Bellver_de_Cerdanya,Cubelles,Gavà,La_Sénia,Manlleu,Pardines,Reus,Sabadell,Sitges,Sort,Tarragona,Terrassa,Tona,Vic,Vilanova}

@ATTRIBUTE Estacio {Llar_de_Jubilats,Sant_Domènec_-_Itàlia,CEIP_Mare_de_Déu_de_Talló,Poliesportiu,Parc_del_Mil·leni,Repetidor,Hospital_Comarcal,Ajuntament,El_Tallapedra,Gran_Via,Vallcarca_-_Oficines,Escola_de_Caiac,Sant_Salvador,Pare_Alegre,Zona_Esportiva,Estadi,Plaça_de_les_Danses_de_Vilanova}

@ATTRIBUTE Comarca {Montsià,Cerdanya,Garraf,Baix_Llobregat,Osona,Ripollès,Baix_Camp,Vallès_Occidental,Pallars_Sobirà,Tarragonès}

@ATTRIBUTE Zona_Oficial {Àrea_de_Barcelona,Vallès_-_Baix_Llobregat,Penedès_-_Garraf,Camp_de_Tarragona,Catalunya_Central,Plana_de_Vic,Maresme,Comarques_de_Girona,Empordà,Alt_Llobregat,Pirineu_Oriental,Pirineu_Occidental,Prepirineu,Terres_de_Ponent,Terres_del_Ebre}

@ATTRIBUTE Zona_Redefinida {Delta_del_Ebre,Pirineu-Prepirineu,Costa,Interior,Industrial,Ciutat_Poblada}

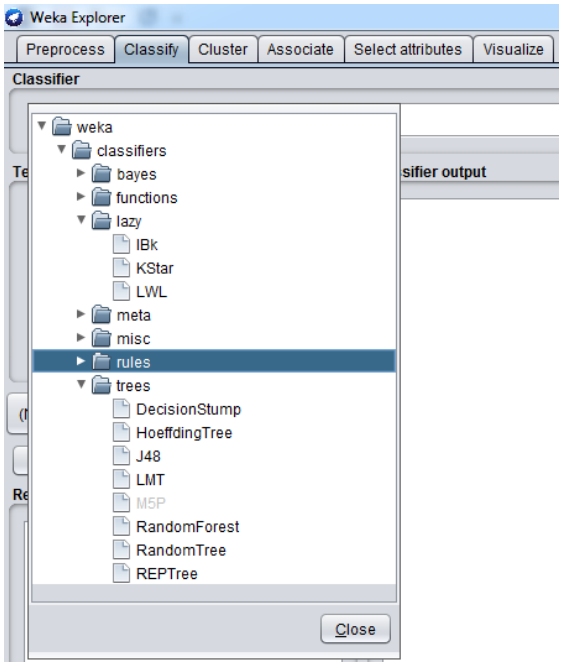
Il·lustració 41 : Captura de pantalla segon arxiu .arff.
Font : 2016 Document Dades_Preparades2.arff. Canvis en la Capçalera, secció @RELATION, lliurat.

```
@DATA

01/01/2005,1,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
02/01/2005,2,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
03/01/2005,3,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
04/01/2005,4,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
05/01/2005,5,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
06/01/2005,6,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
07/01/2005,7,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
08/01/2005,8,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
09/01/2005,9,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
10/01/2005,10,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
11/01/2005,11,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
12/01/2005,12,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
13/01/2005,13,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
14/01/2005,14,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
15/01/2005,15,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
16/01/2005,16,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
17/01/2005,17,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
18/01/2005,18,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
19/01/2005,19,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
20/01/2005,20,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
21/01/2005,21,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
22/01/2005,22,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
23/01/2005,23,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
24/01/2005,24,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
25/01/2005,25,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
26/01/2005,26,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
27/01/2005,27,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
28/01/2005,28,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar_de_Jubilats,Montsià,15,Terres_del_Ebre,Delta_del_Ebre
29/01/2005,29,Gener,2005,?,?,2.22,12.11,?,?,Alcanar,Llar de Jubilats,Montsià,15,Terres del Ebre,Delta del Ebre
```

Il·lustració 42 : Captura de pantalla segon arxiu .arff.
Font : 2016 Document Dades_Preparades2.arff. Canvis en les dades, secció @DATA, lliurat.

Les modificacions descrites han permès accedir a la majoria d'algorismes del Weka, tal i com es mostra en la següent captura de pantalla :



Il·lustració 43 : Captura de pantalla Weka amb la majoria d'algorismes actius.
Font : 2016 Programari instal·lat al PC.

Hem arribat al punt d'escollir els algorismes que s'aplicaran a l'arxiu Dades_Preparades2.arff creat, per agrupar objectes semblants, classificar objectes, predir, descriure i/o explicar de manera profitosa la problemàtica dels contaminants a Catalunya, treballant amb els sub-entorns Classify, Cluster i Associate.

4.3. Sub-entorns Explorer de Weka

Weka presenta varis sub-entorns per realitzar la mineria de dades, els més importants són :

4.3.1. Sub-entorn Cluster

Els algorismes d'agrupament cerquen grups d'instàncies amb característiques similars, segons un criteri de comparació entre valors d'atributs de les instàncies definits en els algorismes.

És la classificació d'objectes similars en diferents grups o el que és el mateix fer la partició de les dades en diferents subconjunts (clústers). Els criteris per fer l'assignació a un clúster o altre s'estableix a partir de mesures de distància en l'espai d'observacions o que volen reflectir la proximitat de les distribucions de probabilitat conjunta dels atributs que hi ha en les observacions realitzades.

Hi ha diversos modes de realitzar l'avaluació :

- **Use training set** : Permet l'avaluació del classificador sobre el mateix conjunt que es construeix el model predictiu per determinar l'error, que en aquest cas es denomina "error de resustitució".
- **Supplied test set** : Permet avaluar sobre un conjunt independent. Permet carregar un conjunt nou de dades. Sobre cada dada es pot fer una predicció de classe per comptar els errors.
- **Percentage Split** : Permet dividir les dades en dos grups, d'acord amb el percentatge indicat (%). El valor indicat és el percentatge d'instàncies per construir el model, que tot seguit és avaluat sobre les que s'han deixat a part.
- **Classes to clusters evaluation** : Permet comparar com de bé els grups seleccionats s'ajusten a una classe assignada prèviament en les dades, seleccionant una classe verdadera de caràcter nominal i la classe majoritària de cada grup, així com una matriu de confusió que mostra la quantitat d'errors que hi haurà si s'empren els grups en lloc de la classe vertadera.

4.3.2. Sub-entorn Associate

Mitjançant algorismes d'associació es pot fer la cerca automàtica de regles que relacionen conjunts d'atributs entre si. Són algorismes no supervisats, ja que no hi ha relacions conegudes a priori amb què contrastar la validesa dels resultats, sinó que s'avalua si aquestes regles són estadísticament significatives.

Les regles d'associació cerquen trobar concurrències prou significatives entre grups de variables. L'únic requeriment que imposen és que s'indiqui el "nivell de suport" que es vol que tinguin a partir de les dades, la proporció de les dades que es vol cobrir amb aquesta regla. Llavors cal trobar grups de variables i combinacions de valors que arribin a tenir aquest grau de suport.

El principal algorisme implementat en Weka és l'algorisme "Apriori", el qual només busca regles entre atributs simbòlics, per la qual cosa tots els atributs numèrics haurien de ser discretitzats prèviament, això és fàcil de fer emprant el sub-entorn Preprocess.

4.3.3. Sub-entorn Classify

El problema de la classificació és el més freqüent en la pràctica. Una vegada aplicats els algorismes no supervisats d'agrupament i associació s'aplica la classificació com un refinament en l'anàlisi. D'aquesta manera, es construeix un model que permet predir la categoria de les instàncies en funció d'una sèrie d'atributs d'entrada. La classe es converteix en la variable objectiu a predir.

Els arbres de decisió donen una estructura tal que a cada node se li fa una pregunta sobre un atribut determinat: el valor que pren indica que cal seguir la branca corresponent a l'atribut. Els nodes finals corresponen a un conjunt d'exemples que pertanyen a la mateixa classe. Alguns arbres permeten mètodes de poda que eliminen la generació de sub-arbres que compliquen la seva comprensió.

Les xarxes neuronals també són bons models classificatoris i predictius. Tenen certes analogies amb la manera en què estan connectades les neurones cerebrals i s'organitzen en forma de molts nodes de procés connectats que donen una o més sortides. Les diverses capes de nodes estan connectades entre si amb més o menys força a través d'uns factors o pesos que indiquen la importància de les sortides produïdes per cada node. En conjunt, el que fan és aprendre a ajustar els valors d'aquests pesos per a ser tan predictives com sigui possible.

Les regles de classificació imposen una sèrie de condicions sobre els valors que prenen els atributs d'entrada per tal d'indicar a quina classe poden pertànyer.

El resultat d'aplicar l'algorisme de classificació s'efectua comparant la classe predita amb la classe real de les instàncies.

Les Xarxes Bayesianes són models gràfics que representen la relació probabilística entre certes variables d'interès.

Hi ha diversos modes de realitzar l'avaluació :

- **Use training set** : Permet l'avaluació del classificador sobre el mateix conjunt que es construeix el model predictiu per determinar l'error, que en aquest cas es denomina "error de resustitució".
- **Supplied test set** : Permet avaluar sobre un conjunt independent. Permet carregar un conjunt nou de dades. Sobre cada dada es pot fer una predicció de classe per comptar els errors.
- **Cross-Validation** : Permet avaluar amb validació creuada. Es dividiran les instàncies en tantes carpetes com indica el paràmetre "Folds", i en cada avaluació es prenen les instàncies de cada carpeta com dades de test, i la resta com a dades d'entrenament per construir el model. Els errors calculats seran la mitjana de totes les execucions
- **Percentage Split** : Permet dividir les dades en dos grups, d'acord amb el percentatge indicat (%). El valor indicat és el percentatge d'instàncies per construir el model, que tot seguit és avaluat sobre les que s'han deixat a part.

A més, es poden fer servir, entre d'altres, les opcions addicionals següents :

- **Output model** : Permet visualitzar el model construït pel classificador.
- **Output per-class stats** : Permet obtenir estadístiques dels errors de classificació per cada un dels valors que pren l'atribut de classe.
- **Output entropy evaluation measures** : Permet generar mesures d'avaluació d'entropia.
- **Output confusion matrix** : Permet veure la sortida de la matriu de confusió, on els elements a la diagonal principal són els elements que ha encertat el classificador i la resta són els errors.
- **Store predictions for visualization** : Permet analitzar els errors de classificació.

4.4. Algorismes escollits

Un algorisme de Mineria de dades és un conjunt de càlculs i regles que permet crear un model de mineria de dades a partir de les dades. Aquests algorismes analitzen les dades d'entrada, a la recerca de patrons, trobant totes les connexions possibles que pugui haver-hi en tota la informació.

Per a la qual s'ha de tenir els paràmetres d'ingrés que s'analitzaran en el conjunt de dades per obtenir com a resultat patrons de comportament en base als atributs analitzats. Entre les tècniques de mineria de dades mes emprades estan els algorismes de predicció, classificació, clustering i associació.

4.4.1. Algorismes de predicció

Són algorismes que mitjançant tècniques i operacions matemàtiques donen com a resultat un estimat de la veritat a curt termini. Aquests algorismes pretenen predir una o més variables contínues d'un conjunt de dades basant-se en altres atributs o patrons del mateix conjunt.

La veracitat en les prediccions depèn del coneixement i habilitat de l'usuari, a més del conjunt de paràmetres emprats per a la predicció que aconsegueixi la millor simulació. Diversos algorismes avaluen diferents conjunts de valors i en base als resultats de les simulacions es van millorant aquests valors amb simulacions posteriors.

S'ha de considerar la utilització d'un escenari bo en un instant de temps per predir què és el que passarà en els instants de temps posterior, considerant algun criteri de selecció i la combinació dels valors dels paràmetres que permeti convergir cap a combinacions de valors que donin bones simulacions.

Regressions : A continuació es presenta una introducció intuïtiva de les idees de regressió lineal, múltiple, i no lineal, així com la generalització als models lineals.

- **Regressió lineal**

Es poden resoldre molts problemes per mitjà de la regressió lineal, i pot aconseguir-se encara més aplicant les transformacions a les variables perquè un problema no lineal pugui convertir-se a un de lineal.

La regressió lineal és la forma més simple de regressió, ja que en ella es modelen les dades emprant una línia recta. Es caracteritza, per tant, per la utilització de dues variables, una aleatòria, i (anomenada variable resposta), que és funció lineal d'una altra variable aleatòria, x (anomenada variable predictora), formant-se l'equació y = ax + b. En aquesta equació la variació de y s'assumeix que és constant, i a i b són els coeficients de regressió que especifiquen la intersecció amb l'eix d'ordenades, i el pendent de la recta, respectivament.

Aquests coeficients es calculen emprant el mètode dels mínims quadrats que minimitza l'error entre les dades reals i l'estimació de la línia.

També cal saber com de bona és la recta de regressió construïda. Per a això, s'empra el coeficient de determinació que és una mesura de l'ajust de la mostra.

Coeficient de Determinació : $R^2 = S^2_{xy}/S^2_x S^2_y$

El valor de R^2 ha d'estar entre 0 i 1. Si s'apropa a 0 la recta de regressió no té un bon ajust, mentre que si s'acosta a 1 l'ajust és "perfecte".

- **Regressió Lineal Múltiple**

La regressió lineal múltiple és una extensió de regressió lineal que involucra més d'una variable predictora, i permet que la variable resposta i sigui plantejada com una funció lineal d'un vector multidimensional. El model de regressió múltiple per n variables predictores seria com el que es mostra en l'equació $y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$

Per trobar els coeficients b_i es planteja el model en termes de matrius, com es mostra en la següent il·lustració:

$$Z = \begin{pmatrix} z_{11} & \cdots & z_{1n} \\ z_{21} & \cdots & z_{1n} \\ \vdots & & \vdots \\ z_{m1} & \cdots & z_{mn} \end{pmatrix}; \; Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}; \; B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Il·lustració 44 : Regressió lineal múltiple.
Font : 1984. Breiman, L.; Friedman, H.; Olshen, R.; Stone, C. Classification and Regression Trees. Belmont: Wadsworth.

A la matriu Z, les files representen els m exemples disponibles per calcular la regressió, i les columnes dels n atributs que formaran part de la regressió. D'aquesta manera, z_{ij} serà el valor que pren en l'exemple i l'atribut j. El vector Y està format pels valors de la variable dependent per a cada un dels exemples, i el vector B és el que es desitja calcular, ja que es correspon amb els paràmetres desconeguts necessaris per construir la regressió lineal múltiple. Representant amb X^T la matriu transposada de X i amb X^{-1} la inversa de la matriu X, es calcularà el vector B mitjançant l'equació :

$$B = (Z^T Z)^{-1} Z^T Y$$

Per determinar si la recta de regressió lineal múltiple està ben ajustada, s'empra el mateix concepte que en el cas de la regressió lineal simple: el coeficient de determinació.

$$R^2 = 1 - \frac{(Y - ZB)^T (Y - ZB)^T}{\sum_{i=1}^{m} (y_i - \bar{y})^2}$$

Il·lustració 45 : Regressió lineal múltiple, coeficient de determinació.
Font : 1984. Breiman, L.; Friedman, H.; Olshen, R.; Stone, C. Classification and Regression Trees. Belmont: Wadsworth.

Igual que en el cas de la regressió simple, el valor de R^2 ha d'estar entre 0 i 1, sent 1 l'indicador d'ajust "perfecte".

- **Regressió No Lineal**

En moltes ocasions les dades no mostren una dependència lineal. Això és el que passa si, per exemple, la variable resposta depèn de les variables independents segons una funció polinòmica, donant lloc a una regressió polinòmica que es pot plantejar afegint les condicions polinòmiques al model lineal bàsic.

D'aquesta forma i aplicant certes transformacions a les variables, es pot convertir el model no lineal en un lineal que es pot resoldre pel mètode de mínims quadrats.

No obstant això, alguns models són especialment no lineals com, per exemple, la suma de termes exponencials i no poden convertir-se a un model lineal. Per a aquests casos, pot ser possible obtenir les estimacions del mínim quadrat a través de càlculs extensos en formules més complexes.

Els models lineals generalitzats representen el fonament teòric en què la regressió lineal es pot aplicar per modelar les categories de les variables dependents. En els models lineals generalitzats, la variació de la variable y és una funció del valor mitjà de y, diferent a la regressió lineal on la variació de y és constant. Els tipus comuns de models lineals generalitzats inclouen regressió logística i regressió del Poisson. La regressió logística modela la probabilitat d'algun esdeveniment que ocorre com una funció lineal d'un conjunt de variables independents. Freqüentment les dades exhibeixen una distribució de Poisson i es modelen normalment emprant la regressió de Poisson.

Els algorismes a emprar en aquest apartat són : **SimpleLinearRegression**, **LinearRegression**, **M5P** i **M5Rules**.

4.4.2. Algorismes de classificació

Els algorismes de classificació permeten l'ordenament i disposició d'un conjunt de dades en diferents categories basades en un criteri de similitud a partir d'un patró comú entre les dades. Aquestes tècniques permeten classificar un conjunt d'objectes en unitats més petites que faciliten la seva administració, compressió i interpretació.

Els algoritmes de classificació són del tipus supervisat la qual cosa permet obtenir un model o regla general per a la classificació i així ajudar a tractar casos futurs, on el sistema sigui capaç d'aprendre del que té per poder generalitzar i tractar el que no té.

Classificadors bayesians : Els classificadors Bayesians són classificadors estadístics, que poden predir tant les probabilitats del nombre de membres de classe, com la probabilitat que una mostra donada pertanyi a una classe particular.

La classificació Bayesiana es basa en el teorema de Bayes, i els classificadors Bayesians han demostrat una alta exactitud i velocitat quan s'han aplicat a grans bases de dades. Diferents estudis comparant els algorismes de classificació han determinat que un classificador bayesià senzill conegut com el classificador "naive bayesià" és comparable en rendiment a un arbre de decisió i a classificadors de xarxes de neurones.

A continuació s'expliquen els fonaments del classificador naive Bayesià.

- **Naive Bayes**

El que normalment es vol saber en aprenentatge és quina és la millor hipòtesi (més probable) donades les dades. Si es denota P(D) com la probabilitat a priori de les dades, P(D|h) la probabilitat de les dades donada una hipòtesi, el que es vol estimar és: P(HD), la probabilitat posterior de h donades les dades. Això es pot estimar amb el teorema de Bayes, tal i com es mostra en la següent il·lustració :

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Il·lustració 46 : Teorema de Bayes.
Font : Xarxes bayesianes. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724

Per estimar la hipòtesi més probable (MAP, màximum a posteriori hipòtesi) es busca el major P(h|D), tal i com es mostra en la següent il·lustració :

$$\begin{aligned} h_{MAP} &= argmax_{h \in H}\left(P(h \mid D)\right) \\ &= argmax_{h \in H}\left(\frac{P(D \mid h)P(h)}{P(D)}\right) \\ &= argmax_{h \in H}\left(P(D \mid h)P(h)\right) \end{aligned}$$

Il·lustració 47 : MAP, màxim a posteriori hipòtesi.
Font : Xarxes bayesianes. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724

Ja que P(D) és una constant independent de h. Si s'assumeix que totes les hipòtesis són igualment probables, aleshores resulta la hipòtesi de màxima versemblança (ML, màximum likelihood).

$$h_{ML} = argmax_{h \in H}\left(P(D \mid h)\right)$$

Il·lustració 48 : ML, màximum likelihood.
Font : Xarxes bayesianes. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724

El classificador naive (ingenu) bayesià s'empra quan es vol classificar un exemple descrit per un conjunt d'atributs (a_i) en un conjunt finit de classes (v_j).

Els classificadors naive Bayesians assumeixen que l'efecte d'un valor de l'atribut en una classe donada és independent dels valors dels altres atributs. Aquesta suposició es diu "independència condicional de classe". Aquesta simplifica els càlculs involucrats i, en aquest sentit, és considerat "ingenu" (naive). Aquesta assumpció és una simplificació de la realitat.

$$P(v_j \mid a_1,\ldots,a_n) = P(v_j) \times \prod_i P(a_i \mid v_j)$$

Il·lustració 49 : Simplificació.
Font : Xarxes bayesianes. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724

Malgrat el nom del classificador i de la simplificació realitzada, el naive Bayesià funciona molt bé, sobretot quan es filtra el conjunt d'atributs seleccionat per eliminar redundància, amb el que s'elimina també dependència entre dades.

Regles de Classificació :Les tècniques d'inducció de Regles van sorgir fa més de dues dècades i permeten la generació i contrast d'arbres de decisió, o regles i patrons a partir de les dades d'entrada. La informació d'entrada serà un conjunt de casos on s'ha associat una classificació o avaluació a un conjunt de variables o atributs. Amb aquesta informació aquestes tècniques obtenen l'arbre de decisió o conjunt de regles que suporten l'avaluació o classificació.

En els casos en què la informació d'entrada posseeix algun tipus de "soroll" o defecte (insuficients atributs o dades, atributs irrellevants o errors o omissions en les dades) aquestes tècniques poden habilitar mètodes estadístics de tipus probabilístic per generar arbres de decisió retallats o podats. També en aquests casos es poden identificar els atributs irrellevants, la manca d'atributs discriminants o detectar "gaps" o buits de coneixement. Aquesta tècnica sol portar associada una alta interacció amb l'analista de manera que aquest pugui intervenir en cada pas de la construcció de les regles, bé per acceptar-les, bé per modificar-les

La inducció de regles es pot aconseguir fonamentalment mitjançant dos camins: Generant un arbre de decisió i extraient d'ell les regles o bé mitjançant una estratègia de covering, que consistent en tenir en compte cada vegada una classe i cercar les regles necessàries per cobrir (cover) tots els exemples d'aquesta classe; quan s'obté una regla s'eliminen tots els exemples que cobreix i es continuen cercant més regles fins que no hi hagi més exemples de la classe.

- **Algorisme C4.5 (Weka : J48)**

Aquest permet la predicció i classificació basada en la teoria de la informació de dades. És un arbre multinivell que per al seu càlcul realitza la comparació dels valors d'informació abans i després de cada un dels possibles candidats.

Permet treballar amb valors continus per als atributs, separant els possibles resultats en dues branques i escollir un rang de mesura apropiada.

Aquest algorismes és una evolució de ID3 que permet :

- Ocupació del concepte raó de guany (GR, Gain Ratio).
- Construir arbres de decisió quan alguns dels exemples presenten valors desconeguts per a alguns dels atributs.
- Treballar amb atributs que presentin valors continus.
- La poda dels arbres de decisió.
- Obtenció de Regles de Classificació.

Els algorismes a emprar en aquest apartat són : **NaiveBayes** i **J48**.

4.4.3. Algorismes de Clustering (classificació)

Els de classificació o d'aprenentatge no supervisat empren tècniques iteratives sobre les dades d'entrada per agrupar elements d'un conjunt de dades amb similars característiques basant-se en atributs que es coneguin, els mateixos que no disposen d'un conjunt d'entrenament per tant no posseeixen coneixement a priori. Aquests algoritmes no posseeixen atributs que diferenciïn la classe a la qual pertany cadascuna de les instàncies d'entrada pel fet que no posseeixen informació inicial que validi la pertinença a un determinat clúster.

En els algoritmes de clustering no es tria el camp de predicció per generar classes d'agrupació. El que s'ha de definir és el nombre possible de classes que s'obtindran com a resultat després del processament.

L'objectiu de l'agrupament és classificar un conjunt d'objectes en grups, de manera que els objectes dins d'un grup posseeixin un alt grau de semblança, mentre que els pertanyents a grups diferents siguin poc semblants entre si.

Els algoritmes de clustering fan servir criteris de comparació de similitud o divergència entre les dades analitzades que anomenen distància entre dades.

- **k veïns més propers**

El mètode dels k-veïns més propers està considerat com un bon representant d'aquest tipus d'aprenentatge, i és de gran senzillesa conceptual. Se sol denominar mètode perquè és l'esquelet d'un algorisme que admet l'intercanvi de la funció de proximitat donant lloc a múltiples variants.

La funció de proximitat pot decidir la classificació d'un nou exemple atenent a la classificació de l'exemple o de la majoria dels k exemples més propers. Admet també funcions de proximitat que considerin el pes o cost dels atributs que intervenen, el que permet, entre altres coses, eliminar els atributs irrellevants.

Una funció de proximitat clàssica entre dues instàncies x_i i x_j, si suposem que un exemple ve representat per una n-tupla de la forma $(a_1(x), a_2(x), ..., a_n(x))$ en la qual $a_r(x)$ és el valor de la instància per a l'atribut a_r, és la distància euclidiana.

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

Il·lustració 50 : Distància Euclidiana.
Font : https://es.wikipedia.org/wiki/Distancia_euclidiana

Atès que l'algorisme k-NN permet que els atributs dels exemplars siguin simbòlics i numèrics, així com que hi hagi atributs sense valor (missing values) l'algorisme per al càlcul de la distància entre exemplars es complica lleugerament.

A més dels diferents tipus d'atributs cal tenir en compte també, en el cas dels atributs numèrics, els rangs en els quals es mouen els seus valors. Per evitar que atributs amb valors molt alts tinguin molt més pes que atributs amb valors baixos, es normalitzaran aquests valors, tal i com es mostra en la següent il·lustració :

$$\frac{x_{il} - min_l}{Max_l - min_l}$$

Il·lustració 51 : Distància Euclidiana Normalitzada.
Font : Agregació. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165722

En aquesta imatge x_{if} és el valor i de l'atribut f, sent min_f el mínim valor de l'atribut f i Max_f el màxim.

D'altra banda, l'algorisme permet donar més preferència a aquelles instàncies més properes al que es desitja classificar. En aquest cas, en lloc d'emprar directament la distància entre instàncies, s'emprarà :

$$\frac{1}{1 + d(x_i, x_j)}$$

Il·lustració 52 : Distància entre instàncies properes.
Font : Agregació. Ramon Sangüesa i Solé http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165722

- **Algorisme MakeDensityBasedClustered**

MakeDensityBasedClusterered és un meta- clúster que envolta un algorisme d'agrupament per fer-lo tornar una distribució de probabilitat i densitat. Per a cada grup d'atributs s'ajusta a una distribució discreta o una distribució normal simètrica (dels que la desviació estàndard mínima és un paràmetre).

Els algorismes a emprar en aquest apartat són : **SimpleKMeans** i **MakeDensityBasedClustered**.

4.4.4. Algorismes d'associació

Aquests algoritmes permeten descobrir regles d'associació que ocorren en comú entre elements o objectes que pertanyen a un conjunt de dades. Per a la qual es consideren totes les possibles combinacions d'atribut-valor de totes les dades emmagatzemades en el conjunt.

La problemàtica que es pretén resoldre amb aquests problemes és : Donat un conjunt de registres, trobar regles que prediuen l'ocurrència d'un ítem, basant-se en les ocurrències d'altres ítems en el registre.

Els problemes d'associació es defineixen típicament com :

- Sigui I = i_1, i_2, ..., i_n, un conjunt d'n atributs binaris anomenats ítems.
- Sigui D = t_1, t_2, ..., t_n, un conjunt de transaccions emmagatzemades en una base de dades.

Cada transacció té un identificador únic i conté un subconjunt d'ítems de I. Una regla es defineix com una implicació de la forma X=> Y, on : $X, Y \subseteq I$ y $X \cap Y = \phi$.

Els conjunts d'ítems X i Y es denominen respectivament antecedent i conseqüent de la regla respectivament.

El principal algorisme implementat en Weka és l'algorisme "Apriori", el qual només busca regles entre atributs simbòlics, per la qual cosa tots els atributs numèrics han de ser discretitzats prèviament.

Els algorismes a emprar en aquest apartat són : **Apriori** i **FilteredAssociator**.

4.5. Mineria

4.5.1. Significat dels paràmetres en els resultats

En primer lloc s'explica quin significat té cadascun d'aquests paràmetres amb la seva respectiva fórmula, això servirà per interpretar els resultats obtinguts en aplicar els algorismes escollits i detallats en l'apartat anterior amb el programari lliure Weka. Per dur a terme aquesta tasca, s'han emprat les referències bibliogràfiques següents : [Bat04], [Cha02], [Gar79], [Han11], [He09], [He11], [Her04], [Hul09], [Hya76], [Jap02], [Lue09] i [Wit11].

- **Kappa statistic**

El coeficient Kappa mostra la concordança entre les dades de prova i la classificació feta pel model, és a dir, quan el resultat del coeficient de Kappa és 1, vol dir que totes les instàncies són classificades correctament, és allà on es diu que té màxima concordança, per contra, quan el valor és igual a zero, la concordança es deu a l'atzar.

- **Mean absolute error**

L'error absolut mitjà és la diferència entre els valors previstos en les instàncies de prova i els valors reals sobre el nombre total. Si és un valor baix, fa que el model classifiqui bé.

$$\frac{|p_1-a_1|+\cdots+|p_n-a_n|}{n}$$

Il·lustració 53 : Mean Absolute Error.
Font : [Wit11]

- **Root mean squared error**

És l'arrel del quadrat de l'error absolut mitjà, mesura la magnitud mitjana de l'error. Si és un valor baix, fa que el model classifiqui bé.

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}}$$

Il·lustració 54 : Root Mean Squared Error.
Font : [Wit11]

- **Relative absolute error**

És la diferència entre els valors previstos en les instàncies de prova i els valors reals sobre la diferència entre els valors reals i el valor mitjà de les dades d'entrenament.

$$\frac{|p_1-a_1|+\cdots+|p_n-a_n|}{|a_1-\bar{a}|+\cdots+|a_n-\bar{a}|}$$

Il·lustració 55 : Relative Absolute Error.
Font : [Wit11]

- **Root relative squared error**

És l'arrel de l'error absolut relatiu al quadrat.

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}}$$

Il·lustració 56 : Root Relative Squared Error.
Font : [Wit11]

- **Detall de precisió per classe**

A continuació es poden observar els detalls de la precisió per classe, els quals són TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area i PRC Area.

- TP Rate : Mostra els veritables positius.
- FP Rate : Mostra els falsos positius.
- Precision : S'obté a partir de les mesures anteriors, la qual cosa indica el percentatge d'encert del model després de fer les classificacions en cada classe, és a dir, mesura el nombre de termes correctament reconeguts respecte al total de termes predits.

$$Precision = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}$$

Il·lustració 57 : Precision.
Font : [Her04]

- Recall (cobertura) : Mesura la proporció d'instàncies correctament reconegudes, respecte al total de termes reals.

$$Recall = \frac{verdaderos\ positivos}{Verdaderos\ positivos + falsos\ negativos}$$

Il·lustració 58 : Recall.
Font : [Her04]

- F-Measure : Caracteritza amb un únic valor la bondat d'un algorisme, mentre més propera sigui a 1, més gran serà la fiabilitat del model en la classe.

$$F - Measure = \frac{2 \times recall \times precision}{recall + precision}$$

Il·lustració 59 : F-Measure.
Font : [Wit11]

- MCC : El coeficient de correlació de Matthews, s'empra com una mesura de la qualitat de les classificacions de dues classes. Retorna un valor entre -1 (cap relació entre predicció i observació) i 1 (predicció perfecta).

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Il·lustració 60 : MCC.
Font : [He09]

- ROC (Receiver Operating Characteristic) Àrea : És l'àrea sota la corba entre els veritables positius (eix Y) i els falsos positius (eix X), mentre més proper sigui a 1 el test és vist com excel·lent.

- PRC Area : Les corbes PRC (Precision-Recall Curve) poden oferir més informació sobre la valoració de l'acompliment en el cas de conjunts de dades altament esbiaixats, és per això que en molts treballs actuals s'usen aquest tipus de corbes per fer avaluacions d'acompliment i comparacions. Aquestes corbes es defineixen fent la gràfica de la taxa de precisió contra la taxa de Recall. Les corbes PR tenen una estreta relació amb les corbes ROC: una corba domina en l'espai ROC si i només si, domina en l'espai PR (La corba PRC dominant, es troba a la part superior dreta de l'espai PR).

- **Matriu de confusió**

A la matriu de confusió es poden observar les instàncies classificades correctament, es reconeixen perquè estan a la diagonal, aquests són els encerts i la resta de valors són els errors, és a dir, mostra quantes instàncies són predites a cada un dels valors possibles de cada classe, la matriu és de tipus n x n.

4.5.2. Presentació dels resultats obtinguts

En aquest apartat, es descriuran els resultats que es mostren en les taules, tenint en compte que en el títol de la taula, es descriuen els atributs i l'algorisme que se'ls ha aplicat. Això permetrà prendre la decisió sobre quin mètode ha obtingut millors resultats i quin o quins models implantar.

- **SimpleLinearRegression amb les dades dels sis contaminants :**

Select the attribute to use as the class	Correlation Coefficient	Mean absolute error	Prediction
CO	0.5993	0.0951	Linear regression on NO : 0 * NO + 0.29
H2S	0.2708	0.3367	Linear regression on O3 : -0 * O3 + 1.31
NO	0.7762	5.8649	Linear regression on NO2 : 0.85 * NO2 - 8.04
NO2	0.7761	7.1281	Linear regression on NO : 0.71 * NO + 14.1
O3	0.5285	15.6651	Linear regression on NO : -0.63 * NO + 59.01
SO2	0.3109	2.5102	Linear regression on O3 : -0.09 * O3 + 8.22

Taula 6 : SimpleLinearRegression

- **LinearRegression amb les dades dels sis contaminants :**

Select the attribute to use as the class	Correlation Coefficient	Mean absolute error	Prediction
CO	0.6167	0.0932	**CO =** 0.0033 * NO + 0.0018 * NO2 + 0.0006 * O3 + 0.0032 * SO2 + 0.279
H2S	0.2432	0.3438	**H2S =** 0.2338 * CO - 0.0016 * NO + 0.0077 * NO2 + 0.0013 * O3 + 0.0226 * SO2 + 0.9677
NO	0.8271	5.7157	**NO =** 30.4477 * CO - 1.7832 * H2S + 0.6341 * NO2 - 0.1638 * O3 + 0.35 * SO2 - 5.0479
NO2	0.783	7.0475	**NO2 =** 2.64 * CO + 1.9732 * H2S + 0.6491 * NO - 0.1022 * O3 - 0.1227 * SO2 + 16.7108
O3	0.4778	16.2631	**O3 =** - 1.6509 * CO - 3.2985 * H2S - 0.3543 * NO - 0.3329 * NO2 - 0.5707 * SO2 + 74.443
SO2	0.3384	2.5365	**SO2 =** - 4.8305 * CO + 1.4618 * H2S + 0.0778 * NO - 0.0214 * NO2 - 0.0648 * O3 + 5.9115

Taula 7 : LinearRegression

S'observa en els dos mètodes de regressió, que en varis dels contaminants, s'obtenen uns coeficients de correlació propers a 1 i les seves mitjanes d'error absolut són força baixes, el contaminant H2S obté el pitjor resultat, ja que es tenen poques dades (només les de dues poblacions, Tarragona i Reus).

- **M5P amb les dades dels sis contaminants :**

Select the attribute to use as the class	Correlation Coefficient	Mean absolute error	Number of Rules
CO	0.6612	0.0828	43
H2S	0.3731	0.3144	6
NO	0.9179	3.0131	102
NO2	0.9012	4.3073	118
O3	0.6408	13.5115	67
SO2	0.7068	1.7192	109

Taula 8 : M5P

- **M5Rules amb les dades dels sis contaminants :**

Select the attribute to use as the class	Correlation Coefficient	Mean absolute error	Number of Rules
CO	0.6612	0.0822	29
H2S	0.3711	0.3146	3
NO	0.9139	3.0776	43
NO2	0.9002	4.3228	52
O3	0.6408	13.4996	45
SO2	0.696	1.753	40

Taula 9 : M5Rules

S'observa que amb els dos mètodes M5P i M5Rules, en tots els contaminants s'obtenen uns coeficients de correlació propers a 1 i les seves mitjanes d'error absolut també són força baixes, excepte el contaminant H2S, que com s'ha dit anteriorment es tenen poques dades.

- **NaiveBayes amb dos atributs :**

Contaminant	Select the attribute to use as the class	Kappa statistic
CO	Comarca	0.1346
H2S	Comarca	0.1003
NO	Comarca	0.0835
NO2	Comarca	0.1712
O3	Comarca	0.0659
SO2	Comarca	0.2387

Taula 10 : NaiveBayes amb dos atributs (1)

Contaminant	Select the attribute to use as the class	Kappa statistic
CO	Zona_Redefinida	0.0827
H2S	Zona_Redefinida	0.1688
NO	Zona_Redefinida	-0.0493
NO2	Zona_Redefinida	0.1135
O3	Zona_Redefinida	0.1119
SO2	Zona_Redefinida	-0.0593

Taula 11 : NaiveBayes amb dos atributs (2)

Contaminant	Select the attribute to use as the class	Kappa statistic
CO	Zona_Oficial	0.0794
H2S	Zona_Oficial	0.1499
NO	Zona_Oficial	0.0278
NO2	Zona_Oficial	0.131
O3	Zona_Oficial	0.1157
SO2	Zona_Oficial	-0.0246

Taula 12 : NaiveBayes amb dos atributs (3)

Els coeficients Kappa statistic del mètode NaiveBayes amb dos atributs són molt baixos la qual cosa ens diu que la concordança es deu a l'atzar.

- **NaiveBayes amb tres atributs :**

Contaminant/Comarca	Select the attribute to use as the class	Kappa statistic
CO	Zona_Redefinida	0.999
H2S	Zona_Redefinida	1
NO	Zona_Redefinida	0.9937
NO2	Zona_Redefinida	0.9977
O3	Zona_Redefinida	1
SO2	Zona_Redefinida	0.9955

Taula 13 : NaiveBayes amb tres atributs (1)

Contaminant/Zona_Redefinida	Select the attribute to use as the class	Kappa statistic
CO	Comarca	0.7427
H2S	Comarca	0.7411
NO	Comarca	0.7449
NO2	Comarca	0.7587
O3	Comarca	0.7653
SO2	Comarca	0.75

Taula 14 : NaiveBayes amb tres atributs (2)

Contaminant/Comarca	Select the attribute to use as the class	Kappa statistic
CO	Zona_Oficial	0.9989
H2S	Zona_Oficial	1
NO	Zona_Oficial	0.9939
NO2	Zona_Oficial	0.998
O3	Zona_Oficial	1
SO2	Zona_Oficial	0.9957

Taula 15 : NaiveBayes amb tres atributs (3)

Contaminant/Zona_Oficial	Select the attribute to use as the class	Kappa statistic
CO	Comarca	0.8792
H2S	Comarca	0.878
NO	Comarca	0.8814
NO2	Comarca	0.8953
O3	Comarca	0.8956
SO2	Comarca	0.8868

Taula 16 : NaiveBayes amb tres atributs (4)

Contaminant/Zona_Oficial	Select the attribute to use as the class	Kappa statistic
CO	Zona_Redefinida	0.999
H2S	Zona_Redefinida	1
NO	Zona_Redefinida	0.9938
NO2	Zona_Redefinida	0.9978
O3	Zona_Redefinida	1
SO2	Zona_Redefinida	0.9956

Taula 17 : NaiveBayes amb tres atributs (5)

Contaminant/ Zona_Redefinida	Select the attribute to use as the class	Kappa statistic
CO	Zona_Oficial	0.8606
H2S	Zona_Oficial	0.8612
NO	Zona_Oficial	0.8556
NO2	Zona_Oficial	0.8594
O3	Zona_Oficial	0.8615
SO2	Zona_Oficial	0.857

Taula 18 : NaiveBayes amb tres atributs (6)

Els coeficients Kappa statistic del mètode NaiveBayes amb tres atributs són molt propers a 1 en la majoria dels casos la qual cosa ens diu que pràcticament totes les instàncies són classificades correctament.

- **J48 amb dos atributs :**

Contaminant	Select the attribute to use as the class	Kappa statistic	Number of Leaves	Size of the tree
CO	Zona_Redefinida	0.1119	8	15
H2S	Zona_Redefinida	0	1	1
NO	Zona_Redefinida	0.258	91	181
NO2	Zona_Redefinida	0.2952	122	243
O3	Zona_Redefinida	0.1531	63	125
SO2	Zona_Redefinida	0.1732	68	135

Taula 19 : J48 amb dos atributs (1)

Contaminant	Select the attribute to use as the class	Kappa statistic	Number of Leaves	Size of the tree
CO	Comarca	0.1711	10	19
H2S	Comarca	-0	1	1
NO	Comarca	0.3064	141	281
NO2	Comarca	0.3198	184	367
O3	Comarca	0.1682	128	255
SO2	Comarca	0.3267	82	163

Taula 20 : J48 amb dos atributs (2)

Contaminant	Select the attribute to use as the class	Kappa statistic	Number of Leaves	Size of the tree
CO	Zona_Oficial	0.1755	9	17
H2S	Zona_Oficial	-0	1	1
NO	Zona_Oficial	0.3076	136	271
NO2	Zona_Oficial	0.3421	154	307
O3	Zona_Oficial	0.153	151	301
SO2	Zona_Oficial	0.249	72	143

Taula 21 : J48 amb dos atributs (3)

Contaminant	Select the attribute to use as the class	Kappa statistic	Number of Leaves	Size of the tree
CO	Població	0.16	14	27
H2S	Població	0.0646	7	13
NO	Població	0.308	201	401
NO2	Població	0.2921	234	467
O3	Població	0.1403	247	493
SO2	Població	0.2567	92	183

Taula 22 : J48 amb dos atributs (4)

Amb l'algorisme J48 amb dos atributs passa exactament el mateix que amb el mètode NaiveBayes amb dos atributs, com a curiositat s'observa el valor Kappa statistic 0 en el contaminant H2S. Afegir també, que amb tres atributs ja millora molt més com passa amb el NaiveBayes.

- **SimpleKMeans amb les dades dels sis contaminants i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (52191.0)	Clúster 1 (16098.0)		Clustered Instances
CO	0.3663	0.3482	0.4248		Clúster 0 52191 (76%)
H2S	1.3511	1.3522	1.3475		
NO	9.8984	5.4422	24.3457		
NO2	21.131	16.8389	35.0462		Clúster 1 16098 (24%)
O3	55.5503	64.2084	27.48		
SO2	3.3745	2.8438	5.0951		

Taula 23 : SimpleKMeans (1)

- **SimpleKMeans amb les dades dels sis contaminants més l'atribut Mes i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (47479.0)	Clúster 1 (20810.0)		Clustered Instances
Mes	Gener	Agost	Novembre		
CO	0.3663	0.3479	0.4082		Clúster 0 47479 (70%)
H2S	1.3511	1.3505	1.3526		
NO	9.8984	5.2777	20.4407		
NO2	21.131	16.4645	31.7777		Clúster 1 20810 (30%)
O3	55.5503	65.1772	33.5861		
SO2	3.3745	2.8154	4.6502		

Taula 24 : SimpleKMeans (2)

- **SimpleKMeans amb les dades dels sis contaminants més l'atribut Any i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (46054.0)	Clúster 1 (22235.0)		Clustered Instances
Any	2008	2007	2005		
CO	0.3663	0.3498	0.4005		Clúster 0 46054 (67%)
H2S	1.3511	1.3489	1.3558		
NO	9.8984	5.804	18.3789		
NO2	21.131	16.7671	30.1697		Clúster 1 22235 (33%)
O3	55.5503	63.9828	38.0847		
SO2	3.3745	2.9416	4.2712		

Taula 25 : SimpleKMeans (3)

- **SimpleKMeans amb les dades dels sis contaminants més l'atribut Comarca i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (39855.0)	Clúster 1 (28434.0)	Clustered Instances
CO	0.3663	0.3423	0.4	
H2S	1.3511	1.3437	1.3615	Clúster 0 39855 (58%)
NO	9.8984	5.8276	15.6043	
NO2	21.131	15.8153	28.5818	
O3	55.5503	60.138	49.12	Clúster 1 28434 (42%)
SO2	3.3745	3.367	3.385	
Comarca	Montsià	Osona	Montsià	

Taula 26 : SimpleKMeans (4)

- **SimpleKMeans amb les dades dels sis contaminants més l'atribut Zona_Oficial i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (39855.0)	Clúster 1 (28434.0)	Clustered Instances
CO	0.3663	0.3423	0.4	
H2S	1.3511	1.3437	1.3615	Clúster 0 39855 (58%)
NO	9.8984	5.8276	15.6043	
NO2	21.131	15.8153	28.5818	
O3	55.5503	60.138	49.12	Clúster 1 28434 (42%)
SO2	3.3745	3.367	3.385	
Zona_Oficial	Penedès_-_Garraf	Plana_de_Vic	Terres_del_Ebre	

Taula 27 : SimpleKMeans (5)

- **SimpleKMeans amb les dades dels sis contaminants més l'atribut Zona_Redefinida i dos clústers:**

Attribute	Full Data (68289.0)	Clúster 0 (37668.0)	Clúster 1 (30621.0)	Clustered Instances
CO	0.3663	0.3397	0.399	
H2S	1.3511	1.3432	1.3609	Clúster 0 39855 (58%)
NO	9.8984	5.0735	15.8337	
NO2	21.131	16.1593	27.2468	
O3	55.5503	62.8484	46.5727	Clúster 1 28434 (42%)
SO2	3.3745	2.6366	4.2823	
Zona_Redefinida	Costa	Costa	Delta_del_Ebre	

Taula 28 : SimpleKMeans (6)

- **MakeDensityBasedClustered amb les dades dels sis contaminants i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (52191.0)	Clúster 1 (16098.0)	Prior probability	Clustered Instances
CO	0.3663	0.3482	0.4248		
H2S	1.3511	1.3522	1.3475	Clúster 0 0.7643	Clúster 0 56718 (83%)
NO	9.8984	5.4422	24.3457		
NO2	21.131	16.8389	35.0462		
O3	55.5503	64.2084	27.48	Clúster 1 0.2357	Clúster 1 11571 (17%)
SO2	3.3745	2.8438	5.0951		

Taula 29 : MakeDensityBasedClustered (1)

- **MakeDensityBasedClustered amb les dades dels sis contaminants més l'atribut Mes i dos clústers:**

Attribute	Full Data (68289.0)	Clúster 0 (47479.0)	Clúster 1 (20810.0)	Prior probability	Clustered Instances
Mes	Gener	Agost	Novembre		
CO	0.3663	0.3479	0.4082	Clúster 0 0.6953	Clúster 0 51476 (75%)
H2S	1.3511	1.3505	1.3526		
NO	9.8984	5.2777	20.4407		
NO2	21.131	16.4645	31.7777	Clúster 1 0.3047	Clúster 1 16813 (25%)
O3	55.5503	65.1772	33.5861		
SO2	3.3745	2.8154	4.6502		

Taula 30 : MakeDensityBasedClustered (2)

- **MakeDensityBasedClustered amb les dades dels sis contaminants més l'atribut Any i dos clústers:**

Attribute	Full Data (68289.0)	Clúster 0 (46054.0)	Clúster 1 (22235.0)	Prior probability	Clustered Instances
Any	2008	2007	2005		
CO	0.3663	0.3498	0.4005	Clúster 0 0.6744	Clúster 0 52317 (77%)
H2S	1.3511	1.3489	1.3558		
NO	9.8984	5.804	18.3789		
NO2	21.131	16.7671	30.1697	Clúster 1 0.3256	Clúster 1 15972 (23%)
O3	55.5503	63.9828	38.0847		
SO2	3.3745	2.9416	4.2712		

Taula 31 : MakeDensityBasedClustered (3)

- **MakeDensityBasedClustered amb les dades dels sis contaminants més l'atribut Comarca i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (39855.0)	Clúster 1 (28434.0)	Prior probability	Clustered Instances
CO	0.3663	0.3423	0.4		
H2S	1.3511	1.3437	1.3615	Clúster 0 0.5836	Clúster 0 44742 (66%)
NO	9.8984	5.8276	15.6043		
NO2	21.131	15.8153	28.5818		
O3	55.5503	60.138	49.12	Clúster 1 0.4164	Clúster 1 23547 (34%)
SO2	3.3745	3.367	3.385		
Comarca	Montsià	Osona	Montsià		

Taula 32 : MakeDensityBasedClustered (4)

- **MakeDensityBasedClustered amb les dades dels sis contaminants més l'atribut Zona_Oficial i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (39855.0)	Clúster 1 (28434.0)	Prior probability	Clustered Instances
CO	0.3663	0.3423	0.4		
H2S	1.3511	1.3437	1.3615	Clúster 0 0.5836	Clúster 0 44763 (66%)
NO	9.8984	5.8276	15.6043		
NO2	21.131	15.8153	28.5818		
O3	55.5503	60.138	49.12	Clúster 1 0.4164	Clúster 1 23526 (34%)
SO2	3.3745	3.367	3.385		
Zona_Oficial	Penedès_-_Garraf	Plana_de_Vic	Terres_del_Ebre		

Taula 33 : MakeDensityBasedClustered (5)

- **MakeDensityBasedClustered amb les dades dels sis contaminants més l'atribut Zona_Redefinida i dos clústers :**

Attribute	Full Data (68289.0)	Clúster 0 (37668.0)	Clúster 1 (30621.0)		Prior probability	Clustered Instances
CO	0.3663	0.3397	0.399			
H2S	1.3511	1.3432	1.3609		Clúster 0 0.5516	Clúster 0 44029 (64%)
NO	9.8984	5.0735	15.8337			
NO2	21.131	16.1593	27.2468			
O3	55.5503	62.8484	46.5727		Clúster 1 0.4484	Clúster 1 24260 (36%)
SO2	3.3745	2.6366	4.2823			
Zona_Redefinida	Costa	Costa	Delta_del_Ebre			

Taula 34 : MakeDensityBasedClustered (6)

Tan l'algorisme SimpleKMeans com l'algorisme MakeDensityBasedClustered mostren els dos grups d'exemples més similars (Clústers 0 i 1), i els seus centroides, a més en la columna Full Data es mostren les mitjanes per atributs numèrics i valor més repetit per atribut simbòlic (si n'hi ha).

- **Apriori de CO discretitzat amb Comarca i Zona Oficial :**

```
Best rules found:

 1. Zona_Oficial=Terres_del_Ebre 12051 ==> Comarca=Montsià 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 2. Comarca=Montsià 12051 ==> Zona_Oficial=Terres_del_Ebre 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 3. Zona_Oficial=Penedès_-_Garraf 12051 ==> Comarca=Garraf 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 4. Comarca=Garraf 12051 ==> Zona_Oficial=Penedès_-_Garraf 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 5. Zona_Oficial=Plana_de_Vic 12051 ==> Comarca=Osona 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 6. Comarca=Osona 12051 ==> Zona_Oficial=Plana_de_Vic 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 7. Zona_Oficial=Vallès_-_Baix_Llobregat 8034 ==> Comarca=Vallès_Occidental 8034    <conf:(1)> lift:(8.5) lev:(0.1) [7088] conv:(7088.82)
 8. Comarca=Vallès_Occidental 8034 ==> Zona_Oficial=Vallès_-_Baix_Llobregat 8034    <conf:(1)> lift:(8.5) lev:(0.1) [7088] conv:(7088.82)
```

Il·lustració 61 : Captura de pantalla algorisme Apriori.
Font : 2016. Programari lliure Weka.

En aquesta captura de pantalla es poden llegir regles òbvies (totes tenen un nivell de confiança de 1, es compleix al 100%) com per exemple que la comarca del Montsià pertany a la Zona Oficial de les Terres de l'Ebre (Regla 2) o que la comarca del Garraf pertany a la Zona Oficial del Penedès-Garraf (Regla 4).

- **FilteredAssociator sense discretitzar CO amb Comarca i Zona Oficial :**

```
=== Run information ===

Scheme:       weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1-4,6-12,14,16
Instances:    68289
Attributes:   3
              CO
              Comarca
              Zona_Oficial
```

Il·lustració 62 : Captura de pantalla algorisme FilteredAssociator sense discretitzar.
Font : 2016. Programari lliure Weka.

En aquesta captura de pantalla es pot observar que no apareix cap regla, tot i que sigui obvia com les de la captura anterior. Això succeeix amb cadascun dels contaminants combinats amb un o varis dels atributs Dia, Mes, Any, Estació, Comarca, Població, ZQA, Zona Oficial o Zona Redefinida.

- **FilteredAssociator sense discretitzar NO2 amb tots els atributs menys la data i la resta de contaminants :**

```
=== Run information ===

Scheme:       weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1,5-7,9-10
Instances:    68289
Attributes:   10
              Dia
              Mes
              Any
              NO2
              Poblacio
              Estacio
              Comarca
              ZQA
              Zona_Oficial
              Zona_Redefinida
```

Il·lustració 63 : Captura de pantalla algorisme FilteredAssociator sense discretitzar amb 10 atributs.
Font : 2016. Programari lliure Weka.

En aquesta captura de pantalla es pot observar que no apareix cap regla, fins i tot combinant nou atributs amb un contaminant sense discretitzar no és capaç de trobar res, ni obvi ni interessant.

- **FilteredAssociator de CO discretitzat amb Comarca i Zona Oficial :**

```
Best rules found:

 1. Comarca=Montsià 12051 ==> CO='(-inf-0.21]' 12051    <conf:(1)> lift:(1.37) lev:(0.05) [3241] conv:(3241.76)
 2. Comarca=Osona 12051 ==> CO='(-inf-0.21]' 12051    <conf:(1)> lift:(1.37) lev:(0.05) [3241] conv:(3241.76)
 3. Zona_Oficial=Plana_de_Vic 12051 ==> CO='(-inf-0.21]' 12051    <conf:(1)> lift:(1.37) lev:(0.05) [3241] conv:(3241.76)
 4. Zona_Oficial=Terres_del_Ebre 12051 ==> CO='(-inf-0.21]' 12051    <conf:(1)> lift:(1.37) lev:(0.05) [3241] conv:(3241.76)
 5. Zona_Oficial=Terres_del_Ebre 12051 ==> Comarca=Montsià 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 6. Comarca=Montsià 12051 ==> Zona_Oficial=Terres_del_Ebre 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 7. Zona_Oficial=Penedès_-_Garraf 12051 ==> Comarca=Garraf 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 8. Comarca=Garraf 12051 ==> Zona_Oficial=Penedès_-_Garraf 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 9. Zona_Oficial=Plana_de_Vic 12051 ==> Comarca=Osona 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
10. Comarca=Osona 12051 ==> Zona_Oficial=Plana_de_Vic 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
```

Il·lustració 64 : Captura de pantalla algorisme FilteredAssociator amb CO discretitzat.
Font : 2016. Programari lliure Weka.

En aquesta captura de pantalla com ha passat amb la captura de pantalla de l'algorisme Apriori, es poden llegir regles òbvies (totes tenen un nivell de confiança de 1, es compleix al 100%) com per exemple que la comarca del Montsià pertany a la Zona Oficial de les Terres de l'Ebre (Regla 6) o que la comarca del Garraf pertany a la Zona Oficial del Penedès-Garraf (Regla 8).

5. Conclusions

5.1. Conclusions dels algorismes

- **Algorismes de predicció**

El resultat de la regressió lineal proporciona una funció lineal que permet calcular i predir la variable de sortida en funció de com afecten en major o menor mesura les variables d'entrada.

Mitjançant l'algorisme M5P s'evidencia consistència entre les dades de CO extrapolades (mitjançant predicció) i les dades de CO mesurades. El mateix succeeix amb la resta de contaminants, excepte H2S a causa del baix nombre de mesures fetes. Aquest baix nombre impedeix obtenir un coeficient de correlació que garanteixi la consistència entre les dades mesurades i les obtingudes per predicció.

Pel que fa a l'algorisme M5Rules, al tractar-se d'una derivació de l'algorisme M5P, s'obtenen els mateixos resultats que l'anterior, observant-ne el baix coeficient de correlació en el H2S.

Els algorismes són més precisos com més variables tenen per calcular la predicció, però és interessant perdre una mica de precisió per aconseguir un estalvi de temps i de càrrega en aquests processos.

Si es vol prioritzar la precisió els més precisos son el M5P i el M5Rules, en canvi si el que es vol és velocitat, l'algorisme de regressió lineal, com s'ha pogut veure en els resultats mostrats en l'apartat anterior, és amb molt el més senzill d'implementar i el més ràpid en la seva execució. En aquest cas, s'hauria d'avaluar l'impacte d'aquest error, i si no és prou significatiu, seria l'algorisme guanyador.

- **Algorismes de classificació**

Amb l'aplicació de l'algorisme de Naive Bayes s'observa com al aplicar-se amb dos atributs, el valor de kappa statistic és baix, cosa que determina poca relació entre les dades mesurades i el model. En canvi, al augmentar a tres els atributs, el valor de Kappa statistic augmenta considerablement, apropant-se al valor unitari, cosa que indica una molt bona relació entre dades i model. Això vol dir, que per a poder aproximar un model acceptable, són necessaris tres atributs com a mínim.

En el J48 amb tres atributs passa exactament el mateix que amb dos, és a dir el Kappa statistic és molt baix i la majoria dels arbres són enormes, difícils de llegir i pràcticament no aporten informació rellevant.

- **Algorismes de Clustering (classificació)**

Les informacions aportades per l'algorisme SimpleKMeans i per l'algorisme MakeDensityBasedClustered són la mitjana per a cada contaminant del total de mostres i la determinació de dos clústers, amb els seus centroides corresponents, així com la probabilitat associada a cadascun d'ells. A més l'algorisme MakeDensityBasedClustered incorpora mesures de la distribució normal i de la desviació estàndard mínima.

La interpretació dels resultats indica petites variacions entre la mitjana de cadascun dels contaminants sobre el total de mostres i el clúster 0. Les variacions més significatives es donen en el NO; pel que fa a la resta de contaminants es mantenen en unes variacions relativament acceptables. Ara bé, pel que fa al clúster 1, les variacions s'accentuen en el NO, NO2, O3 i SO2.

Els òxids de nitrogen estan associats amb el trànsit de vehicles, per tant, els dos clústers poden relacionar-se amb èpoques i/o zones depenent de les variacions en la quantitat de vehicles.

L'ozó és un contaminant secundari que prové de reaccions de la resta de contaminants afavorides per la llum solar, en el nostre país aquest contaminant és abundant a causa de la forta radiació solar que tenim, essent un gas que tendeix a concentrar-se en capes baixes de l'atmosfera en moments d'estabilitat atmosfèrica a causa del seu pes molecular. Així doncs, mitjançant aquest algorisme es poden observar dos clústers amb valors d'ozó molt diferenciats que poden ésser atribuïbles a moments d'estabilitat atmosfèrica (màxim valor) i valor mínim en èpoques amb moviment meteorològic. En qualsevol cas, el valor mitjà és més proper al valor elevat del clúster 0, ja que com s'ha comentat abans, l'ozó apareix en zones amb alta radiació solar, com és el cas de Catalunya.

La resta de contaminants es mouen en ambdós clústers amb valors força propers, establint-se una relació amb períodes de més immissió.

- **Algorismes d'associació**

La utilització dels algorismes Apriori i FilteredAssociator implica una discretització de les dades en el cas del primer. Un cop aplicats ambdós filtres es pot observar que el nivell de confiança en tots els casos és 1, cosa que ens indica que els resultats són totalment certs. Evidentment, vistos els resultats, les comarques coincideixen plenament amb la zona oficial; és a dir, dur a una trivialització dels resultats.

5.2. Conclusions generals

En aquest treball s'ha presentat Weka com una alternativa de programari de mineria de dades, s'ha demostrat que és una eina lliure i molt interessant a l'hora d'aplicar diverses tècniques de mineria de dades, observant la gran utilitat que té la mineria de dades en aplicar-la a un cas real.

S'ha experimentat el senzill que és mitjançant Weka l'anàlisi i estudi d'aquestes dades, i la seva posterior interpretació. S'ha decidit emprar totes les possibilitats que ofereix aquesta eina per demostrar que es pot fer un estudi molt complet.

El preprocessat, la classificació, l'agrupament, l'associació i la visualització previs de les dades d'entrada permeten obtenir, amb més facilitat, millors resultats.

S'ha vist també la gran diversitat d'algoritmes inclosos en el programari lliure Weka que es poden emprar segons es vulgui obtenir uns o altres objectius.

Tot això fa que Weka sigui una eina principal en les cada vegada més importants tecnologies basades en el processament d'informació en els diferents àmbits de la societat.

Per tant, es dedueix que la mineria de dades es una tecnologia innovadora, que ofereix una sèrie de beneficis: d'una banda, resulta un bon punt de trobada entre els investigadors i les persones de negocis; de l'altra, pot estalviar grans quantitats de diners a una empresa i obre noves oportunitats de negocis. A més, no hi ha dubte que treballar amb aquesta tecnologia implica tenir cura d'un gran nombre de detalls pel fet que el producte final involucra "presa de decisions".

Després de fer forces proves amb diferents models per observar el tipus d'informació que donava cada model de mineria de dades, es conclou entre d'altres que :.

L'evolució temporal dels contaminants presenta, en la majoria dels casos, una alta estacionalitat, ja que aquests estan íntimament relacionats amb les condicions meteorològiques o les condicions de tràfic i/o processos industrials. Cal esmentar que mentre les concentracions d'ozó venen determinades per les condicions d'estabilitat atmosfèrica com s'ha comentat anteriorment, els contaminants que depenen del tràfic com els òxids de nitrogen, augmenten els seus valors fluctuant i tenint pics en èpoques de màxim consum energètic.

En el cas del SO2, es nota un decreixement des del 2008 a la zona de Tarragona. Aquest fet és atribuïble a la crisi econòmica, ja que les immissions per part de la indústria petroquímica de la ciutat han disminuït al minvar la demanda de petroli per als processos industrials. La resta de contaminants segueixen una certa estacionalitat o bé es mantenen constants amb variacions a causa de les condicions meteorològiques.

Pel que fa a la inferència, s'ha realitzat una regressió lineal en la que és possible obtenir el valor d'un contaminant en funció de la resta. S'ha fet la regressió de cada contaminant respecte dels altres, com s'observa en les taules 6 i 7.

La principal diferència observada en l'aplicació dels algorismes entre zona oficial i redefinida és que la comarca recollida en zona oficial i redefinida pot no ser la mateixa. Per exemple, en l'aplicació de SimpleKMeans entre zona oficial i clústers, dóna el Penedès-Garraf com a zona que "compleix" els valors mitjans, i al avaluar el mateix algorisme amb zona redefinida, adjudica a la zona "costa" els valors mitjans. Es veu com en aquest cas, ambdós atributs coincideixen.

No passa el mateix amb el clúster 0, on mentre adjudica Plana de Vic a aquest clúster, en la zona redefinida la zona trobada és Costa. Aquestes variacions poden ser degudes a la diferència de criteri entre ambdues zones, de manera que aquest clúster pot ser adjudicat a les dues a la vegada.

5.3. Recomanacions

Si es pretén aplicar mineria de dades, cal tenir unes dades ben definides i seguir cada pas del procés d'extracció del coneixement és fonamental per aconseguir un encert alt en aplicar la tècnica més eficaç.

Per a realitzar els passos del procés d'extracció del coneixement s'ha de començar en la selecció de dades, obtenint unes dades, després depurar-les de manera que quedin només les instàncies necessàries; posteriorment s'han de preprocessar, això és perquè només s'analitzin les rellevants, per finalment aplicar els algoritmes continguts en Weka, per així poder agrupar objectes semblants, classificar objectes, predir, descriure i/o explicar de manera profitosa.

És realment important a l'hora de definir la tècnica de mineria de dades conèixer la més adequada, per això és recomanable realitzar l'anàlisi de les dades pels algorismes de les tècniques disponibles en Weka i escollir la que millor percentatge d'encert tingui respecte a les altres.

Concretament en el seguiment de la qualitat de l'aire s'ha millorat en els darrers 30 anys, però si un es fixa en el marc normatiu, ja que les millores ambientals sempre van lligades a la regulació, se'n dona compte que la llei catalana de contaminació és del 1983 i no s'hauria de regular la qualitat de l'aire, en aquests temps de canvis constants, amb una llei tan obsoleta.

6. Línies futures

6.1. Sobre el programari Weka

Sobre el programari lliure Weka es plantegen dues línies futures d'investigació, ja que en aquest treball no hi ha prou temps per explorar a fons tot el programari i extreure'n totes les possibilitats, aquestes són :

Els Filtres :

En el sub-entorn Preprocess, Weka permet aplicar una gran diversitat de filtres sobre les dades, permetent realitzar transformacions sobre ells de tot tipus. En prémer el botó Choose dins el requadre Filter es desplega un arbre en el qual seleccionar els filtres a escollir. Entre d'altres hi ha :

- **Attribute** : Els filtres agrupats en aquesta categoria són aplicats a atributs.
  - Add : Afegeix un atribut més. Com a paràmetres s'ha de proporcionar la posició que ocuparà aquest nou atribut (aquest cop començant des de l'1), el nom de l'atribut i els possibles valors d'aquest atribut separats entre comes. Si no s'especifiquen, se sobreentén que l'atribut és numèric.
  - AddExpression : Aquest filtre és molt útil ja que permet afegir al final un atribut que sigui el valor d'una funció. Cal especificar la fórmula que descriu aquest atribut, on es pot calcular aquest atribut a partir dels valors d'un altre o altres, referint-nos als altres atributs per "a" seguit del número de l'atribut (començant per 1). Un altre argument d'aquest filtre és el nom del nou atribut.
  - AddNoise : Afegeix soroll a un determinat atribut que ha de ser nominal. Es pot especificar el percentatge de soroll, la llavor per generar-lo, i si es vol que en introduir el soroll compti o no amb els atributs que manquen.
  - ClusterMembership : Filtre que donat un conjunt d'atributs i l'atribut que defineix la classe dels mateixos, torna la probabilitat de cadascun dels atributs d'estar classificats en una classe o una altra.
  - Copy : Realitza una còpia d'un conjunt d'atributs en les dades. Aquest filtre és útil en conjunció amb altres, ja que hi ha certs filtres (la majoria) que destrueixen les dades originals.
  - Discretize : Discretitza un conjunt de valors numèrics en rangs de dades. Com a paràmetres pren els índexs dels atributs discretitzar (attribute índexs) i el número de particions en què es vol dividir les dades.
  - FistOrder : Aquest filtre realitza una transformació de les dades obtenint la diferència de parells consecutius de dades, suposant una dada inicial addicional de valor 0 per aconseguir que la cardinalitat del grup de dades resultant sigui la mateixa que la de les dades origen.
  - MakeIndicator : Crea un nou conjunt de dades reemplaçant un atribut nominal per un booleà (Assignarà "1" si en una instància es troba l'atribut nominal seleccionat i "0" en cas contrari).
  - MergeTwoValues : Fusiona dos atributs nominals en un de sol. Pren com a arguments la posició de l'argument resultat i la dels arguments font.
  - NominalToBinary : Transforma els valors nominals d'un atribut en un vector les coordenades del qual són binàries.
  - Normalize : Normalitza totes les dades de manera que el rang de les dades passi a ser [0,1].
  - NumericToBinary : Converteix dades en format numèric a binari. Si el valor d'una dada és 0 o desconegut, el valor en binari resultant serà el 0.
  - NumericTransformFiltre : Similar a AddExpression però molt més potent. Permet aplicar un mètode Java sobre un conjunt d'atributs donant-li el nom d'una classe i un mètode.
  - Obfuscate : Ofusca totes les cadenes de text de les dades. Aquest filtre és molt útil si es desitja compartir una base de dades però no es vol compartir informació privada.
  - PKIDiscretize : Discretitza atributs numèrics (igual que Discretize), però el nombre d'intervals és igual a l'arrel quadrada del nombre de valors definits.

  - RandomProjection : Redueix la dimensionalitat de les dades (útil quan el conjunt de dades és molt gran) projectant-la en un subespai de menor dimensionalitat emprant per a això una matriu aleatòria. Tot i reduir la dimensionalitat de les dades resultants es procura conservar l'estructura i propietats fonamentals de les mateixes.
  - Remove : Esborra un conjunt d'atributs del fitxer de dades.
  - RemoveType : Elimina el conjunt d'atributs d'un tipus determinat.
  - RemoveUseless : Elimina atributs que oscil·len menys que un nivell de variació. És útil per eliminar atributs constants o amb un rang molt petit.
  - ReplaceMissingValues : Reemplaça tots els valors indefinits per la moda en el cas que sigui un atribut nominal o la mitjana aritmètica si és un atribut numèric.
  - Standardize : Estandarditza les dades numèriques de la mostra perquè tinguin de mitjana 0 i la unitat de variància.
  - StringToNominal : Converteix un atribut de tipus cadena a un tipus nominal.
  - StringToWordVector : Converteix els atributs de tipus String en un conjunt d'atributs representant l'ocurrència de les paraules del text.
  - SwapValues : Intercanvia els valors de dos atributs nominals.
  - TimeSeriesDelta : Filtre que assumeix que les instàncies formen part d'una sèrie temporal i reemplaça els valors dels atributs de manera que cada valor d'una instància és reemplaçat amb la diferència entre el valor actual i el valor pronosticat per a aquesta instància.

- **Instance** : Els filtres són aplicats a instàncies concretes senceres.
  - NonSparseToSparse : Converteix una mostra de mode complert a mode abreujat.
  - Randomize : Modifica l'ordre de les instàncies de forma aleatòria.
  - RemoveFolds : Permet eliminar un conjunt de dades. Aquest filtre està pensat per eliminar una partició en una validació creuada.
  - RemoveMisclassified : Donat un mètode de classificació l'aplica sobre la mostra i elimina aquelles instàncies mal classificades.
  - RemovePercentage : Suprimeix un percentatge de mostres.
  - RemoveRange : Elimina un rang d'instàncies.
  - RemoveWithValues : Elimina les instàncies d'acord a una determinada restricció.
  - Resample : Obté un subconjunt del conjunt inicial de forma aleatòria.
  - SparseToNonSparse : Converteix una mostra de mode abreujat a mode complet. És l'operació complementària a NonSparseToSparse.

I la Selecció d'atributs :

Un altre possible treball a realitzar seria automatitzar una recerca d'atributs més apropiada per explicar un atribut objectiu, en un sentit de classificació supervisada.

Així, es podria explorar què subconjunts d'atributs són els que millor poden classificar la classe de la instància.

Aquesta selecció supervisada tindria dos components :

- Mètode d'avaluació (Attribute Evaluator) : Funció que determina la qualitat del conjunt d'atributs per discriminar la classe.
- Mètode de cerca (Search Method) : Forma de realitzar la recerca de conjunts. Si es vol realitzar una avaluació exhaustiva de tots els subconjunts, apareix un problema combinatori inabordable quant creix el nombre d'atributs. Per això apareixen aquestes estratègies que permeten fer la recerca d'una forma més eficient.

Respecte al component d'avaluació es poden distingir dos tipus :

- Els que directament empren un classificador específic per mesurar la qualitat del subconjunt d'atributs a través de la taxa d'error del classificador. S'anomenen mètodes Wrapper, ja que "envolten" al classificador per explorar la millor selecció d'atributs que optimitza les seves prestacions. Necessiten un procés complet d'entrenament i avaluació en cada cas de recerca, per això són molt costosos.

- Els que no empren aquest classificador específic. Dins d'aquest tipus està el mètode CfsSubsetEval, que calcula la correlació de la classe amb cada atribut, i elimina atributs que tenen una correlació molt alta com a atributs redundants.

Respecte al component de cerca dir que :

- Un d'aquests mètodes, que es caracteritza per la seva rapidesa, és el ForwardSelection. Es tracta d'un mètode de recerca subòptima en escalada. El procediment és el següent: es tria primer el millor atribut, després afegeix el següent atribut que més aporta i continua així fins arribar a la situació en què afegir un nou atribut empitjora la situació.
- Un altre d'aquest tipus de mètodes seria el BestSearch, que permet cercar interaccions entre atributs més complexes. El seu procediment és anar analitzant el que millora i empitjora un grup d'atributs en afegir elements, amb la possibilitat de fer retrocessos per explorar amb més detall.
- Un altre d'aquests mètodes que es podria emprar és el ExhaustiveSearch, que enumera totes les possibilitats i les avalua per seleccionar la millor.

S'haurà, en la configuració del problema, d'escollir l'atribut que s'empra per a la selecció supervisada, i determinar si l'avaluació es realitzaria amb totes les instàncies disponibles, o mitjançant la validació creuada.

Una bona solució seria escollir els algorismes més eficients d'avaluació i recerca, CsfSubsetEval i ForwardSelection. D'aquesta manera es podrien estudiar els atributs que millor expliquen altres clústers, arribant a relacions trivials o a altres no conegudes.

6.2. Sobre Medi ambient

L'aire contaminat que sura en la superfície de la terra és arrossegat pel vent i la pluja cap a altres zones. Els núvols i les altes temperatures també contribueixen a que la contaminació es dispersi i arribi a grans distàncies, allunyades del punt d'origen.

A l'hivern, la contaminació atmosfèrica es produeix per estancament de l'aire. Aquest fenomen ocorre quan els contaminants, procedents de la combustió, com el SO2 i altres partícules en suspensió, s'acumulen a l'atmosfera, essent una de les causes del clima extrem.

A l'estiu, la contaminació de l'aire afecta més en els dies calorosos i assolellats. Durant aquests dies es produeixen reaccions fotoquímiques de gasos com l'òxid de nitrogen i els hidrocarburs. Ells contribueixen a la formació d'un contaminant molt perjudicial per a salut com és l'ozó i d'altres substàncies tòxiques.

El vent, la humitat, la inversió i les precipitacions tenen un paper important en l'augment o disminució de la contaminació. El vent generalment afavoreix la difusió dels contaminants ja que desplaça les masses d'aire en funció de la pressió i la temperatura. L'efecte que pot causar el vent depèn dels accidents del terreny o fins i tot de la configuració dels edificis a les zones urbanitzades.

Al contrari del vent, la humitat juga un paper negatiu en l'evolució dels contaminants ja que afavoreix l'acumulació de fums i pols. D'altra banda, el vapor d'aigua pot reaccionar amb certs anions augmentant l'agressivitat dels mateixos, per exemple el tri-òxid de sofre en presència de vapor d'aigua es transforma en àcid sulfúric, el mateix passa amb els clorurs i els fluorurs per donar àcid clorhídric i fluorhídric respectivament.

Així doncs, una línia futura és fer predicció de sèries temporals relacionades amb la contaminació atmosfèrica per a realitzar actuacions efectives tendents a reduir els nivells de contaminació o adoptar a temps sistemes de prevenció i alerta.

Caldrà buscar relacions entre les concentracions dels contaminants O3 i PM10 (No estudiat per manca de dades en aquest treball) juntament amb les variables atmosfèriques com : precipitació, velocitat dels vents, radiació solar i temperatura mitjançant tècniques de mineria de dades.

6.3. Salut

Els principals efectes de la contaminació atmosfèrica sobre la salut van des d'alteracions de la funció pulmonar, problemes cardíacs i altres símptomes i molèsties fins a un augment del nombre de defuncions, d'ingressos hospitalaris i de visites a urgències, especialment per causes respiratòries i cardiovasculars. L'efecte de la contaminació atmosfèrica manté una gradació tant en la gravetat de les seves conseqüències com a la població de risc afectada. Així, a mesura que els efectes són menys greus, el percentatge de població afectada és més gran.

Il·lustració 65 : Els diferents efectes de la contaminació atmosfèrica sobre la salut.
Font : 2009. Tenías y Ballester.

L'augment ràpid i quantiós de vehicles automotors que empren combustible de mala qualitat i tenen motors antiquat i l'augment de la generació d'electricitat a força de carbó i altres combustibles contaminants, han augmentat el risc sanitari per aire contaminat a la població.

Cal destacar que els efectes de l'exposició crònica superen en magnitud als efectes aguts deguts a exposicions per un curt termini de temps.

- Efectes a curt termini : Irritació d'ulls, nas i gola, infeccions respiratòries, atacs d'asma, canvis en el bombeig del cor.
- Efectes a llarg termini : Desenvolupament pulmonar en nens molt lent, malalties respiratòries cròniques, malalties del cor, càncer de pulmó.

Son molts els estudis que han posat de manifest la importància de la qualitat de l'aire en la salut de la població i han permès identificar els principals mecanismes d'acció pels quals l'exposició a la contaminació atmosfèrica causa danys a la salut.

A finals dels anys 70 i durant la dècada següent, la majoria d'experts pensaven que, amb els nivells que es registraven en la majoria de ciutats dels països més desenvolupats, la contaminació atmosfèrica no representava un perill important per a la salut.

Avui dia, uns 30 anys després, les principals agències encarregades de la protecció de la salut i del medi ambient (OMS, Agència Europea del Medi Ambient i EPA), reconeixen que la inhalació de contaminants, especialment de partícules fines, representa un augment de risc de defunció prematura. Aquest canvi tan important, va començar amb l'anàlisi dels efectes aguts o a curt termini, dels increments de la contaminació atmosfèrica. Amb el temps, i els resultats d'estudis posteriors, se sap que els efectes deguts a l'exposició crònica (efectes a llarg termini), poden ser considerablement més importants en termes de reducció de l'esperança de vida i morbiditat crònica.

Per tot això una de les línies futures a estudiar és el creuament de dades dels contaminants amb el numero d'ingressos hospitalaris i de visites a urgències, especialment per causes respiratòries i cardiovasculars, per establir, en cas de que existeixi, una relació entre ambdós factors.

7. Glossari

Arff : Un arxiu arff (Atribut-Relation File Format) és un arxiu de text ASCII que descriu una llista d'instàncies que comparteixen un conjunt d'atributs, van ser desenvolupats pel Machine Learning Project en el Departament de Ciències de la Computació de la Universitat de Waikato per al seu ús amb el programari d'aprenentatge automàtic Weka.

C4.5 : És un algorisme extensió de l'algorisme ID3 emprat per generar un arbre de decisió desenvolupats ambdós per Ross Quinlan. Els arbres de decisió generats per C4.5 poden ser usats per classificar, i per aquesta raó, C4.5 està gairebé sempre referit com un classificador estadístic.

Coeficient de Correlació de Pearson : És una mesura de la relació lineal entre dues variables aleatòries quantitatives. A diferència de la covariància, la correlació és independent de l'escala de mesura de les variables. De manera menys formal, es pot definir el coeficient de correlació de Pearson com un índex que pot emprar-se per mesurar el grau de relació de dues variables sempre que ambdues siguin quantitatives.

CSV : Els arxius CSV (de l'anglès Comma-Separated Values) són un tipus de document en format obert senzill per representar dades en forma de taula, en què les columnes es separen per comes (o punt i coma on la coma és el separador decimal : Argentina, Brasil ...) i les files per salts de línia. Els camps que continguin una coma, un salt de línia o una cometa doble han de ser tancats entre cometes dobles. El format CSV és molt senzill i no indica un joc de caràcters concret, ni com van situats els bytes, ni el format per al salt de línia. Aquests punts s'han d'indicar-moltes vegades en obrir l'arxiu.

CRF : Un camp aleatori condicional (Conditional Random Field o CRF en espanyol) és un model estocàstic emprat habitualment per etiquetar i segmentar seqüències de dades o extreure informació de documents.

DEI : Directiva sobre les emissions industrials.

EPA : Agència de Protecció Ambiental dels EUA.

GR (Gain Ratio) : En el context dels arbres de decisió, el terme s'empra a vegades com sinònim d'informació mútua, que és el valor esperat de la divergència de Kullback-Leibler d'una distribució de probabilitat condicional. El criteri de maximitzar el guany té com a biaix l'elecció d'atributs amb molts valors. Això és degut a que com més fina sigui la participació produïda pels valors de l'atribut, normalment, la incertesa o entropia en cada nou node serà menor, i per tant també serà menor la mitjana de l'entropia a aquest nivell.

ID3 : Algorisme emprat dins de l'àmbit de la intel·ligència artificial. El seu ús s'engloba en la recerca d'hipòtesis o regles en ell, donat un conjunt d'exemples. El conjunt d'exemples haurà d'estar conformat per una sèrie de tuples de valors, cadascun d'ells denominats atributs, en el qual un d'ells, (l'atribut a classificar) és l'objectiu i és de tipus binari. ID3 realitza aquesta tasca mitjançant la construcció d'un arbre de decisió, els elements del qual són :

- Nodes : Els quals contindran atributs.
- Arcs : Els quals contenen valors possibles del node pare.
- Fulles : Nodes que classifiquen l'exemple com a positiu o negatiu.

L'elecció del millor atribut s'estableix mitjançant l'entropia. Escollint aquell que proporcioni un millor guany d'informació.

IPPC : Directiva europea de prevenció i control integrats de la contaminació.

J48 : És una implementació de codi obert en llenguatge de programació Java de l'algorisme C4.5 en l'eina Weka de mineria de dades.

Java : És un llenguatge de programació de propòsit general, concurrent, orientat a objectes que va ser dissenyat específicament per tenir tan poques dependències d'implementació com fos possible. La seva intenció és permetre que els desenvolupadors d'aplicacions escriguin el programa una vegada i ho s'executin en qualsevol dispositiu.

MAP (Màximum a posteriori hipòtesi) : És un mètode que es pot emprar per estimar un nombre de paràmetres desconeguts, com ara els paràmetres d'una densitat de probabilitat, vinculats a una mostra donada.

ML (maximum likelihood) : Hipòtesi de màxima versemblança és un mètode habitual per ajustar un model i trobar els seus paràmetres. En general, per a un conjunt fix de dades i un model estadístic subjacent, el mètode de màxima versemblança selecciona el conjunt de valors dels paràmetres del model que maximitza la funció de probabilitat.

Notepad++ : És un editor de text de codi font lliure amb suport per a diversos llenguatges de programació i de suport natiu a Microsoft Windows. S'assembla al Bloc de notes pel que fa al fet que pot editar text sense format i de forma simple. No obstant això, inclou opcions més avançades que poden ser útils per a usuaris avançats com a desenvolupadors i programadors. Es distribueix sota els termes de la Llicència pública general de GNU.

OMS : Organització Mundial de la Salut

PAC : Prova d'Avaluació Continuada.

PAMQA : Pla d'actuació per a la millora de la qualitat de l'aire a les zones de protecció especial de l'ambient atmosfèric.

P(D) : És la notació emprada per definir la probabilitat a priori de les dades.

P(D|h) : És la notació emprada per definir la probabilitat de les dades donada una hipòtesi h.

P(HD) : És la notació emprada per definir la probabilitat posterior de h donades les dades.

R2 (Coeficient de Determinació) : És un estadístic emprat en el context d'un model estadístic el principal propòsit és predir futurs resultats o provar una hipòtesi. El coeficient determina la qualitat del model per replicar els resultats, i la proporció de variació dels resultats que pot explicar-se pel model. En el cas de la regressió lineal simple és el quadrat del coeficient de correlació de Pearson.

TFG : Treball Final de Grau.

UOC : Universitat Oberta de Catalunya.

WEKA (Waikato Environment for Knowledge Analysis) : En català «entorn per l'anàlisi del coneixement de la Universitat de Waikato», és un programari de mineria de dades que ha estat desenvolupat a la universitat de Waikato (Nova Zelanda) sota llicència GPL.

X^{-1} : Matriu inversa de la matriu X.

X^{T} : Matriu transposada (s'han canviat files per columnes) de la matriu X.

XVPCA : Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica.

Z_{ij} : Element de la matriu Z, concretament, es troba en la intersecció de fila i amb la columna j.

ZPE : Zona de Protecció Especial de l'ambient atmosfèric.

ZQA (Zona de Qualitat de l'Aire) : Porció del territori amb una qualitat de l'aire similar en tots els seus punts, tant des del punt de vista de les condicions de dispersió com de les emissions de contaminants a l'atmosfera.

Simbologia dels contaminants :

As	Arsènic
BaP	Benzo(a)pirè
C6H6	Benzè
Cd	Cadmi
Cl2	Clor
CO	Monòxid de Carboni
H2S	Sulfur d'Hidrogen
HAP	Hidrocarburs Aromàtics Policíclics
HCl	Clorur d'Hidrogen
Ni	Níquel
NO2	Diòxid de Nitrogen
NOx	Òxids de Nitrogen
O3	Ozó
Pb	Plom
PM10	Partícules en suspensió de diàmetre inferior a 10 micròmetres
PM2.5	Partícules en suspensió de diàmetre inferior a 2.5 micròmetres
SO2	Diòxid de Sofre

Taula 35 : Simbologia dels contaminants

Unitats de mesura :

ng/m³	Nanograms de contaminant per metre cúbic d'aire [1ng = 10^{-9}g]
µg/m³	Micrograms de contaminant per metre cúbic d'aire [1µg = 10^{-6}g]
mg/m³	Mil·ligrams de contaminant per metre cúbic d'aire [1mg = 10^{-3}g]
ppm o µmol/mol	Parts per milió en volum (cm^3/m^3) [1 ppm = (pes molecular/V) mg/m^3]. Equival a 1µmol/mol
ppb o nmol/mol	Parts per bilió americà en volum (mm^3/m^3) [1 ppb = (pes molecular/V) $\mu g/m^3$]. Equival a 1nmol/mol

(on V és el volum a una pressió i temperatura determinades. Segons la legislació actualment vigent cal prendre 101.3kPa i 293K)

Taula 36 : Unitats de mesura

8. Bibliografia / Webgrafia

8.1. Bibliografia

[Bat04] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6(1), 20-29 (Jun 2004).

[Cha02] Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16, 321-357 (2002).

[Gar79] Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1979).

[Han11] Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edn. (2011).

[He09] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. on Knowl. And Data Eng. 21(9), 1263{1284 (Sep 2009).

[He11] He, H.: Self-Adaptive Systems for Machine Intelligence. Wiley (2011).

[Her04] HERNANDEZ ORALLO, José, RAMIREZ QUINTANA, María José, FERRI RAMIREZ, César. Introducción a la minería de datos.Madrid.2004.

[Hul09] Hulse, J.V., Khoshgoftaar, T.: Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering 68(12), 1513-1542 (2009).

[Hya76] Hyafil, L., Rivest, R.L. : Constructing optimal binary decision trees is np-complete. Inf. Process. Lett. 5(1), 15{17 (1976).

[Jap02] Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429 (2002).

[Lue09] Luengo, J., Fernandez, A., Herrera, F., Herrera, F.: Addressing data-complexity for imbalanced datasets: A preliminary study on the use of preprocessing for c4.5. In: Intelligent Systems Design and Applications, 2009. ISDA '09. Ninth International Conference on. pp. 523-528 (2009).

[Wit11] WITTEN, Ian, FRANK Eibe, HALL, Mark. DATA MINING. 2011.

8.2. Webgrafia

Materials Mineria de dades UOC

- Preparació de dades. Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165719
- Classificació: Arbres de decisió. Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165720
- Classificació: xarxes neuronals. Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165721
- Agregació (clustering). Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165722
- Regles d'associació. Luis Carlos Molina Félix, Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165723
- Xarxes bayesianes. Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165724
- Avaluació de models. Luis Carlos Molina Félix, Ramon Sangüesa i Solé
  http://cv.uoc.edu/autors/MostraPDFMaterialAction.do?id=165725
- Aprenentatge. Vicenç Torra i Reventós, David Masip i Rodó
  http://cvapp.uoc.edu/autors/MostraPDFMaterialAction.do?id=200707

Medi ambient :

- Contaminació atmosfèrica
  http://aransa.upc.es/tmma/atmos-unitat/atmos-unitat.html
- Departament de Medi Ambient: ICQA
  http://www.gencat.es/mediamb/aire/cicqa.htm
- Departament de Medi Ambient: XVPCA
  http://www.gencat.es/mediamb/aire/cqaire.htm
- Departament de Territori i Sostenibilitat. Secretaria de Medi Ambient i Sostenibilitat
  http://mediambient.gencat.cat/ca/05_ambits_dactuacio/
  http://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/plans_de_millora

- Pla interdepartamental de salut pública (PINSAP)
  http://salutpublica.gencat.cat/ca/
- Autoritat del Transport Metropolità (ATM) - Pla director de mobilitat
  http://www.atm.cat/web/ca/PDM.php
- Ecovia't
  http://www.ecoviat.com/ca-es/inici.aspx

Normativa :

- Avaluació de la qualitat de l'aire
  http://www.idescat.cat/cat/idescat/biblioteca/docs/pec/paae2011/gi08222010.pdf
- El medi ambient i la salut. Qualitat de l'aire, contaminació química, soroll i radiacions. Anàlisi de legislació i experiències de bones pràctiques de millora del medi i la salut. Propostes per a Catalunya.
  http://www.diba.cat/c/document_library/get_file?uuid=77df4f4a-4d4d-4135-a32f-c044b9611831&groupId=7294824
- El Pla d'actuació per a la millora de la qualitat de l'aire a les zones de protecció especial de l'ambient atmosfèric
  http://mediambient.gencat.cat/web/.content/home/ambits_dactuacio/atmosfera/qualitat_de_laire/oficina_tecnica_de_plans_de_millora/enllacos/Resum1PAMQAMaster6_5-de-juny.pdf
- La qualitat de l'aire a Catalunya : Informes anuals
  http://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/avaluacio_qualitat_aire_catalunya_altres/Informes/

Weka

- Manual de Weka. Diego García Morate
  http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/weka.pdf
- Técnicas de análisis de datos. Aplicaciones prácticas usando Microsoft Excel y Weka. José, Manuel Molina López y Jesús García Herrero
  http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf
- Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de Weka (Waikato Environment for Knowledge Analysis). Paula Andrea, Vizcaíno Garzón.
  http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf
- WEKA, University of Waikato. [En línia] Disponible en Internet :
  http://weka.wikispaces.com/Can+I+use+WEKA+for+time+series+analysis%3F
- WEKA. University of Waikato. [En línia] Disponible en Internet :
  http://weka.wikispaces.com/Making+predictions
- CORZO, Cynthia Lorena. Aplicación de algoritmos de clasificación supervisada usando Weka. [En línea] Disponible en Internet :
  http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

9. Annexos

9.1. Aspectes legals : Dates d'interès

1983. Llei 22/1983, de protecció de l'ambient atmosfèric

Amb aquesta Llei, a Catalunya es regula per primer cop la contaminació atmosfèrica. La seva aprovació va comportar la creació de la Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica a Catalunya (XVPCA), que és l'eina de què es disposa per avaluar la qualitat de l'aire.

La Llei de protecció de l'ambient atmosfèric preveu les actuacions que s'han de dur a terme quan se superen els nivells establerts de qualitat de l'aire i estableix les declaracions de zones de protecció especial i d'atenció especial.

1996-2015. Les directives europees

Amb la incorporació a la Unió Europea, l'any 2000, Espanya se sotmet a les regulacions ambientals comunitàries, que estaven molt més avançades que la regulació pròpia de l'Estat espanyol. Entre els anys 90 i principis del 2000, la Comissió Europea havia aprovat la *Directiva marc 96/62/CE, de 27 de setembre, sobre avaluació i gestió de la qualitat de l'aire ambient*, i les 4 directives "filles" que establien valors límit per a diferents contaminants i altres directives relacionades amb l'intercanvi d'informació i dades.

L'any 2008 es va aprovar la *Directiva 2008/50/CE, de 21 de maig, relativa a la qualitat de l'aire ambient i a una atmosfera més neta a Europa*, que refon els aspectes més importants de les anteriors directives, esdevé el nou marc regulador de la qualitat de l'aire a Europa i preveu la possibilitat de prorrogar el compliment dels valors límit de qualitat de l'aire en tres anys per a les partícules en suspensió de diàmetre inferior a 10 micres (PM10) i en cinc anys per al diòxid de nitrogen (NO2) i el benzè. Els plans de millora de la qualitat de l'aire associats a aquestes sol·licituds han de complir certes condicions que han de ser avaluades per la Comissió Europea.

També en el període 2000-2010 la Comissió va actualitzar altres directives relacionades amb la qualitat de l'aire, com les grans instal·lacions de combustió, d'incineració i de coincineració de residus, que limiten la transferència de la contaminació a l'atmosfera.

Al desembre de 2013 la Comissió Europea va aprovar el programa Aire net per a Europa (A Clean Air Programme for Europe) que és la nova estratègia europea per a la millora de la qualitat de l'aire que estableix el compliment dels valors límit actuals de qualitat de l'aire com a màxim per al 2020 i estableix nous objectius per al 2030. Alhora es va presentar la proposta de revisió de la Directiva de sostres nacionals d'emissió que fixa valors més estrictes per a 6 contaminants crítics per a cadascun dels estats de la Unió o la Proposta de directiva per a les instal·lacions de combustió de mitjana potència.

2011. Reial decret 102/2011, relatiu a la millora de la qualitat de l'aire

És el referent de la normativa bàsica estatal en matèria de qualitat de l'aire, que transposa la *Directiva 2008/50/CE, relativa a la qualitat de l'aire ambient i a una atmosfera més neta a Europa*, i desplega parcialment la *Llei estatal 34/2007, de qualitat de l'aire i protecció de l'ambient atmosfèric*.

Aquest Reial decret defineix els objectius de qualitat de l'aire per a diversos contaminants, i regula l'avaluació, el manteniment i la millora de la qualitat de l'aire. També estableix criteris i mètodes comuns d'avaluació de les concentracions de les substàncies regulades, determina com s'ha d'efectuar la informació a la població i a la Comissió Europea sobre les concentracions dels contaminants i el compliment dels objectius de qualitat de l'aire i dels plans de millora, amb la finalitat d'evitar, prevenir i reduir els efectes nocius de la contaminació sobre la salut humana i el medi ambient.

2007-2014. El primer Pla d'actuació per a la millora de la qualitat de l'aire a les zones de protecció especial de l'ambient atmosfèric, 2007-2010

El maig de 2006 el Govern de la Generalitat, atenent l'evolució de les dades de qualitat de l'aire, va declarar quaranta municipis de la conurbació de Barcelona zones de protecció especial de l'ambient atmosfèric (ZPE) per al contaminant partícules de diàmetre inferior a 10 micres (PM10); 16 d'aquests municipis també van ser-ne declarats per al contaminant diòxid de nitrogen (NO2). El juliol de 2012 el Govern va homogeneïtzar totes les zones per als dos contaminants, NO2 i PM10.

El juliol de 2007, el Govern de la Generalitat va aprovar el Pla d'actuació per a la millora de la qualitat de l'aire 2007-2010, que estableix 73 mesures que s'havien d'adoptar per millorar la qualitat de l'aire als municipis declarats zones de protecció especial. Aquest Pla, que acabava la vigència el desembre de 2010, va ser prorrogat pel Govern fins que no s'aprovés un nou pla que el substituís.

La implantació de les mesures del Pla 2007-2010, de la seva pròrroga i de les noves mesures que s'han anat implementant han suposat una millora significativa en l'evolució dels nivells de partícules en suspensió (PM10), detectable a partir de 2008, però la tendència de millora de la mitjana anual dels nivells de diòxid de nitrogen (NO2) s'ha demostrat molt més lenta. Tant en un cas com en l'altre, la millora registrada de la qualitat de l'aire ha estat influïda positivament per la disminució de l'activitat i la mobilitat que ha comportat la greu crisi econòmica, però també per la implantació de la *Directiva europea de prevenció i control integrats de la contaminació (IPPC)* i de la *Directiva de grans instal·lacions de combustió*, que el novembre de l'any 2010 han estat refoses en la nova *Directiva sobre les emissions industrials (DEI)*, i per les directives "Euro", que regulen els nivells d'emissió de contaminants permesos per als vehicles de motor a la Unió Europea.

2014-2020. El Pla d'actuació per a la millora de la qualitat de l'aire a les zones de protecció especial de l'ambient atmosfèric (PAMQA), amb horitzó 2020

El 23 de setembre de 2014 s'aprova mitjançant l'Acord de Govern 127/2014, el Pla d'actuació per a la millora de la qualitat de l'aire a les zones de protecció especial de l'ambient atmosfèric (PAMQA); les mesures d'aquest Pla abasten àmbits clau com el trànsit rodat, la indústria, el port i l'aeroport, la sensibilització ciutadana o la fiscalitat, entre d'altres, i estableixen responsabilitats, indicadors i calendaris d'execució que impliquen diverses fases d'aplicació, de manera que els seus objectius s'han d'acomplir completament l'any 2020, d'acord amb les orientacions del nou programa europeu Aire net per a Europa.

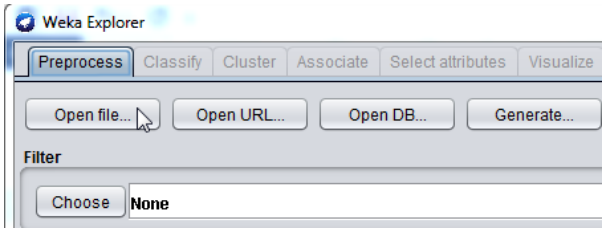
En aplicació del Pla, el Parlament de Catalunya va aprovar la *Llei 12/2014, del 10 d'octubre, de l'impost sobre l'emissió d'òxids de nitrogen a l'atmosfera produïda per l'aviació comercial, de l'impost sobre l'emissió de gasos i partícules a l'atmosfera produïda per la indústria i de l'impost sobre la producció d'energia elèctrica d'origen nuclear*. Els impostos creats per aquesta Llei tenen, d'una banda, un caràcter extra-fiscal per orientar el comportament dels agents econòmics afectats, però són també un instrument de política econòmica que generaran ingressos addicionals per finançar, parcialment, despeses i inversions públiques per a la millora de la qualitat de l'aire.

La mateixa Llei modifica la *Llei de 1983, de protecció de l'ambient atmosfèric*, de manera que modifica, adapta a les necessitats actuals i posa en funcionament, per primera vegada, el Fons per a la protecció de l'ambient atmosfèric, que es destina a finançar les despeses i les inversions públiques en matèria de protecció de l'ambient atmosfèric i de millora de la qualitat acústica i, en general, a les polítiques de prevenció i millora de la qualitat atmosfèrica

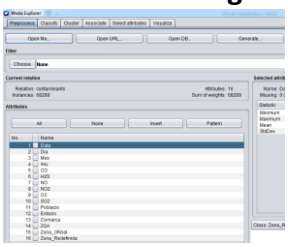
9.2. Passos detallats de la mineria de dades amb Weka

Entrar al Weka

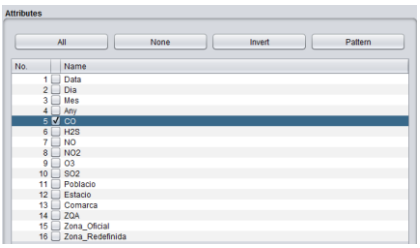
Obrir l'arxiu arff



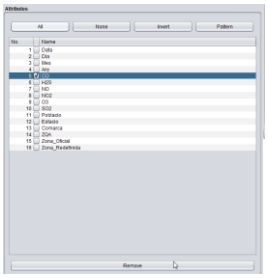
Arxiu .arff carregat



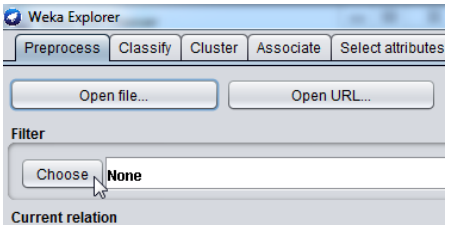
Seleccionar atributs



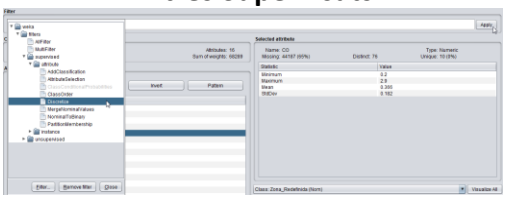
Eliminar atributs



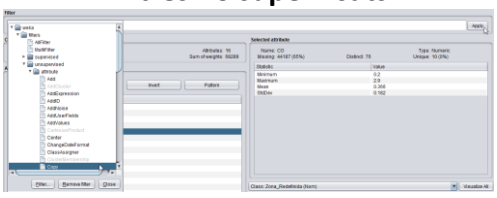
Filtres



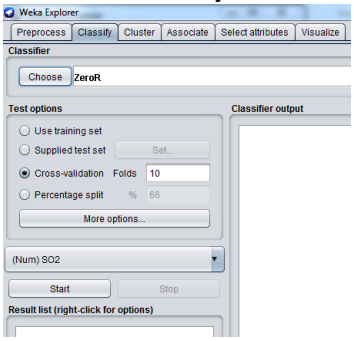
Filtres Supervisats



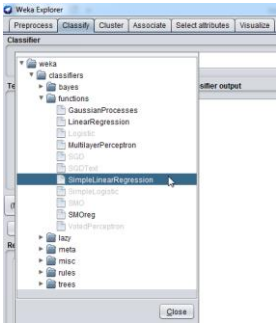
Filtres no supervisats



Classify

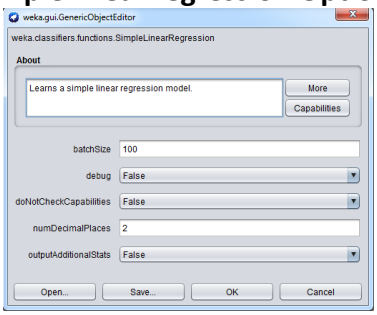


SimpleLinearRegression

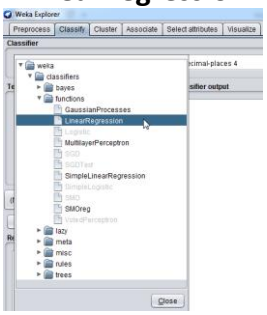


Taula 37 : Weka amb detall (1)

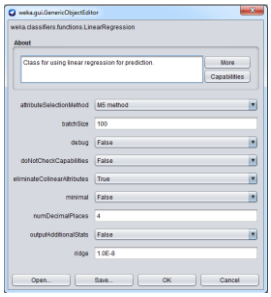
SimpleLinearRegression Options



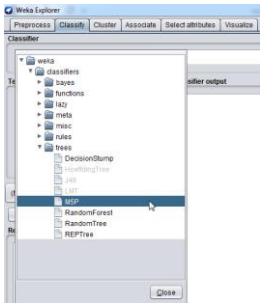
LinearRegression



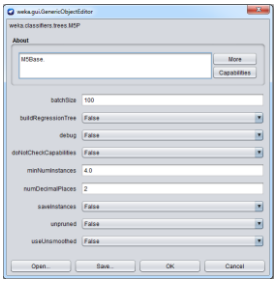
LinearRegression Options



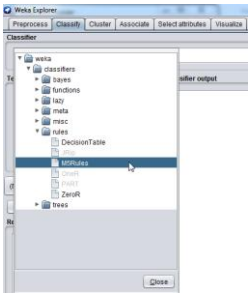
M5P



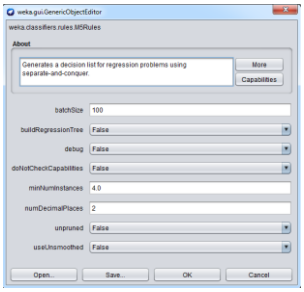
M5P Options



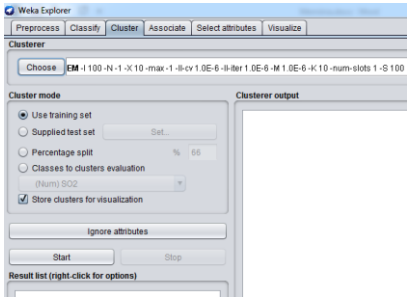
M5Rules



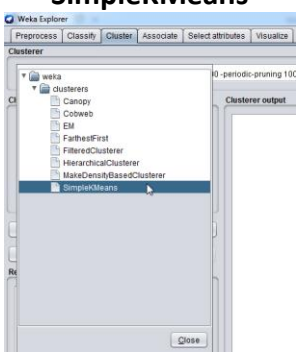
M5Rules Options



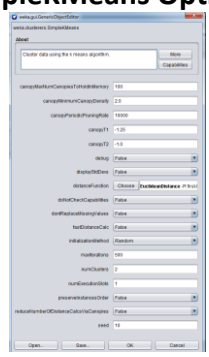
Cluster



SimpleKMeans

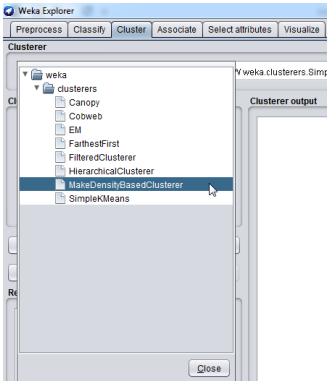


SimpleKMeans Options

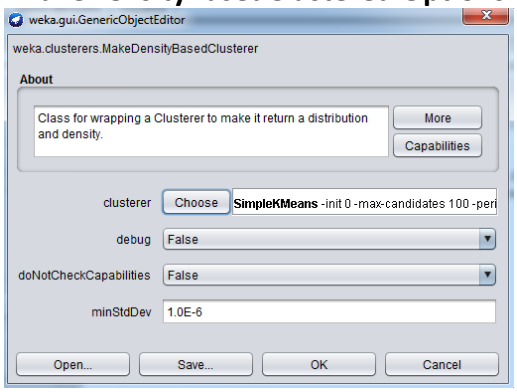


Taula 38 : Weka amb detall (2)

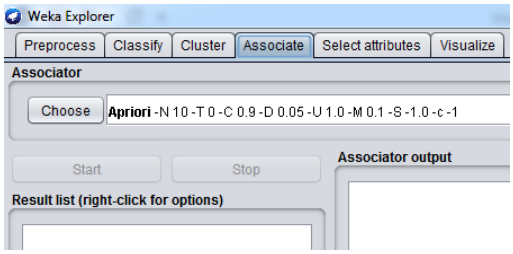
MakeDensityBasedClustered



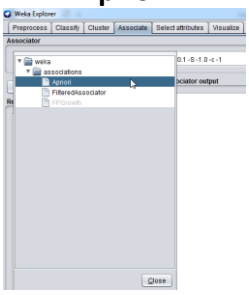
MakeDensityBasedClustered Options



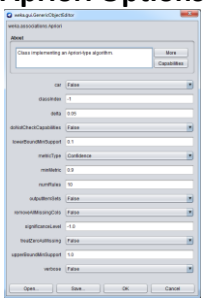
Associate



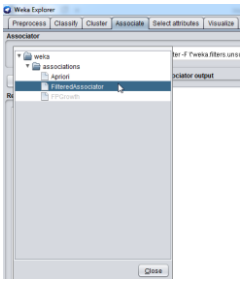
Apriori



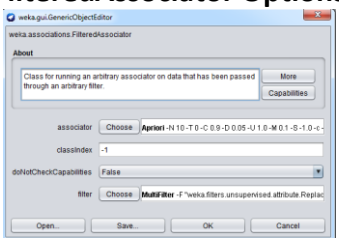
Apriori Options



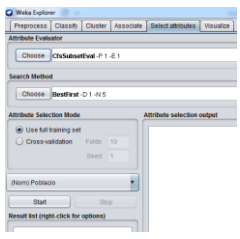
FilteredAssociator



FilteredAssociator Options

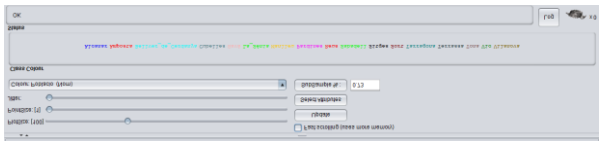


Select Attributes



Visualize (1)

Visualize (2)



Taula 39 : Weka amb detall (3)

9.2.1. Captures Rellevants

- **LinearRegression Ozó per Mes :**

```
Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1-2,4-8,10-16
Instances:    68289
Attributes:   2
              Mes
              O3
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

O3 =

      2.1404 * Mes=Gener,Novembre,Octubre,Febrer,Setembre,Març,Agost,Abril,Maig,Juliol,Juny +
      1.8022 * Mes=Novembre,Octubre,Febrer,Setembre,Març,Agost,Abril,Maig,Juliol,Juny +
      8.363  * Mes=Octubre,Febrer,Setembre,Març,Agost,Abril,Maig,Juliol,Juny +
      2.3852 * Mes=Febrer,Setembre,Març,Agost,Abril,Maig,Juliol,Juny +
     11.1514 * Mes=Setembre,Març,Agost,Abril,Maig,Juliol,Juny +
      1.8026 * Mes=Març,Agost,Abril,Maig,Juliol,Juny +
      5.295  * Mes=Agost,Abril,Maig,Juliol,Juny +
      2.2747 * Mes=Abril,Maig,Juliol,Juny +
      1.9577 * Mes=Maig,Juliol,Juny +
      1.0348 * Mes=Juliol,Juny +
     33.17

Time taken to build model: 0.88 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6043
Mean absolute error                     15.2586
Root mean squared error                 18.8775
Relative absolute error                 80.6143 %
Root relative squared error             79.6753 %
Total Number of Instances           52221
Ignored Class Unknown Instances             16068
```

Il·lustració 66 : Captura de pantalla LinearRegression Ozó per Mes
Font : 2016, Programari lliure Weka.

- **Apriori de NO2 discretitzat amb comarca i zona redefinida :**

```
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1-7,9-12,14-15-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances:    68289
Attributes:   3
              NO2
              Comarca
              Zona_Redefinida
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (6829 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 4

Best rules found:

 1. Zona_Redefinida=Delta_del_Ebre 12051 ==> Comarca=Montsià 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 2. Comarca=Montsià 12051 ==> Zona_Redefinida=Delta_del_Ebre 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 3. Comarca=Garraf 12051 ==> Zona_Redefinida=Costa 12051    <conf:(1)> lift:(4.25) lev:(0.13) [9215] conv:(9215.47)
 4. Zona_Redefinida=Interior 12051 ==> Comarca=Osona 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 5. Comarca=Osona 12051 ==> Zona_Redefinida=Interior 12051    <conf:(1)> lift:(5.67) lev:(0.15) [9924] conv:(9924.35)
 6. Zona_Redefinida=Ciutat_Poblada 8034 ==> Comarca=Vallès_Occidental 8034    <conf:(1)> lift:(8.5) lev:(0.1) [7088] conv:(7088.82)
 7. Comarca=Vallès_Occidental 8034 ==> Zona_Redefinida=Ciutat_Poblada 8034    <conf:(1)> lift:(8.5) lev:(0.1) [7088] conv:(7088.82)
```

Il·lustració 67 : Captura de pantalla Apriori de NO2 discretitzat amb comarca i zona redefinida
Font : 2016, Programari lliure Weka.

- **MakeDensityBasedClusterer amb les dades dels sis contaminants :**

```
Scheme:       weka.clusterers.MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1-4,11-16
Instances:    68289
Attributes:   6
              CO
              H2S
              NO
              NO2
              O3
              SO2
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 1249.570086574939

Initial starting points (random):

Cluster 0: 0.366298,1.351123,3.13,11.44,59.26,3.374517
Cluster 1: 0.366298,1.351123,4.27,15.95,21,3.374517

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute   Full Data          0          1
            (68289.0) (52191.0) (16098.0)
=========================================
CO             0.3663    0.3482    0.4248
H2S            1.3511    1.3522    1.3475
NO             9.8984    5.4422   24.3457
NO2           21.131    16.8389   35.0462
O3            55.5503   64.2084     27.48
SO2            3.3745    2.8438    5.0951

Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.7643

Attribute: CO
Normal Distribution. Mean = 0.3482 StdDev = 0.0685
Attribute: H2S
Normal Distribution. Mean = 1.3522 StdDev = 0.2209
Attribute: NO
Normal Distribution. Mean = 5.4422 StdDev = 4.4801
Attribute: NO2
Normal Distribution. Mean = 16.8389 StdDev = 8.5324
Attribute: O3
Normal Distribution. Mean = 64.2084 StdDev = 13.9713
Attribute: SO2
Normal Distribution. Mean = 2.8438 StdDev = 2.0726

Cluster: 1 Prior probability: 0.2357

Attribute: CO
Normal Distribution. Mean = 0.4248 StdDev = 0.1725
Attribute: H2S
Normal Distribution. Mean = 1.3475 StdDev = 0.115
Attribute: NO
Normal Distribution. Mean = 24.3457 StdDev = 25.6236
Attribute: NO2
Normal Distribution. Mean = 35.0462 StdDev = 18.5025
Attribute: O3
Normal Distribution. Mean = 27.48 StdDev = 12.5354
Attribute: SO2
Normal Distribution. Mean = 5.0951 StdDev = 8.071


Time taken to build model (full training data) : 3.14 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      56718 ( 83%)
1      11571 ( 17%)


Log likelihood: -12.25871
```

Il·lustració 68 : Captura de pantalla MakeDensityBasedClusterer amb les dades dels sis contaminants
Font : 2016, Programari lliure Weka.

- **NaiveBayes amb les dades dels 6 contaminants comarca i zona redefinida :**

```
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     contaminants-weka.filters.unsupervised.attribute.Remove-R1-4,11-12,14-15
Instances:    68289
Attributes:   8
              CO
              H2S
              NO
              NO2
              O3
              SO2
              Comarca
              Zona_Redefinida
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier
```

	Class					
Attribute	Delta_del_Ebre (0.18)	Pirineu-Prepirineu (0.18)	Costa (0.24)	Interior (0.18)	Industrial (0.12)	Ciutat_Poblada (0.12)
CO						
mean	0	0	0.2901	0	0.3192	0.4933
std. dev.	0.006	0.006	0.0977	0.006	0.1124	0.2279
weight sum	0	0	8034	0	8034	8034
precision	0.036	0.036	0.036	0.036	0.036	0.036
H2S						
mean	0	0	0	0	1.3574	0
std. dev.	0.0186	0.0186	0.0186	0.0186	0.5836	0.0186
weight sum	0	0	0	0	8034	0
precision	0.1115	0.1115	0.1115	0.1115	0.1115	0.1115
NO						
mean	3.1872	2.6123	4.3238	7.8562	7.4566	35.7114
std. dev.	3.0796	2.4158	4.8394	14.841	9.1235	28.6945
weight sum	8034	4017	16068	8034	8034	8034
precision	0.4237	0.4237	0.4237	0.4237	0.4237	0.4237
NO2						
mean	14.1085	7.4499	16.4446	16.3013	22.3593	48.138
std. dev.	5.8161	6.2406	9.1538	9.4164	11.1063	17.0012
weight sum	8034	4017	16068	8034	8034	8034
precision	0.2219	0.2219	0.2219	0.2219	0.2219	0.2219
O3						
mean	69.6917	64.5699	54.2663	49.565	54.9364	38.4736
std. dev.	22.0043	21.4867	19.5095	23.3875	20.5146	19.7858
weight sum	8034	12051	8034	12051	4017	8034
precision	0.1725	0.1725	0.1725	0.1725	0.1725	0.1725
SO2						
mean	0	1.3571	2.2401	9.9288	3.5106	3.1256
std. dev.	0.0764	0.2626	2.0818	14.476	4.5872	2.498
weight sum	0	4017	16068	4017	8034	8034
precision	0.4583	0.4583	0.4583	0.4583	0.4583	0.4583
Comarca						
Montsià	12052.0	1.0	1.0	1.0	1.0	1.0
Cerdanya	1.0	4018.0	1.0	1.0	1.0	1.0
Garraf	1.0	1.0	12052.0	1.0	1.0	1.0
Baix_Llobregat	1.0	1.0	4018.0	1.0	1.0	1.0
Osona	1.0	1.0	1.0	12052.0	1.0	1.0
Ripollès	1.0	4018.0	1.0	1.0	1.0	1.0
Baix_Camp	1.0	1.0	1.0	1.0	4018.0	1.0
Vallès_Occidental	1.0	1.0	1.0	1.0	1.0	8035.0
Pallars_Sobirà	1.0	4018.0	1.0	1.0	1.0	1.0
Tarragonès	1.0	1.0	1.0	1.0	4018.0	1.0
[total]	12061.0	12061.0	16078.0	12061.0	8044.0	8044.0

```
Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       67706               99.1463 %
Incorrectly Classified Instances       583                0.8537 %
Kappa statistic                          0.9896
Mean absolute error                      0.0035
Root mean squared error                  0.0496
Relative absolute error                  1.2776 %
Root relative squared error             13.3879 %
Total Number of Instances            68289

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,990    0,000    1,000      0,990    0,995      0,994    0,999     0,998     Delta_del_Ebre
               0,995    0,000    1,000      0,995    0,997      0,997    0,999     0,998     Pirineu-Prepirineu
               0,977    0,000    0,999      0,977    0,988      0,984    0,992     0,992     Costa
               0,998    0,000    0,998      0,998    0,998      0,998    1,000     1,000     Interior
               1,000    0,004    0,974      1,000    0,987      0,985    1,000     0,997     Industrial
               0,999    0,006    0,960      0,999    0,979      0,976    0,999     0,991     Ciutat_Poblada
Weighted Avg.  0,991    0,001    0,992      0,991    0,991      0,990    0,997     0,996

=== Confusion Matrix ===

     a     b     c     d     e     f   <-- classified as
 11930     0     0    22    50    49 |     a = Delta_del_Ebre
     0 11988    16     0    31    16 |     b = Pirineu-Prepirineu
     0     0 15703     0   120   245 |     c = Costa
     0     0     0 12032     0    19 |     d = Interior
     0     0     0     0  8031     3 |     e = Industrial
     0     0     0     0    12  8022 |     f = Ciutat_Poblada
```

Il·lustració 69 : Captura de pantalla NaiveBayes amb les dades dels 6 contaminants comarca i zona redefinida
Font : 2016, Programari lliure Weka.