Classificació Automàtica amb KLASS de les Dades de Procés d'una EDAR

Xavier Flores[1], Ignasi Rodriguez-Roda[1] i Karina Gibert [2].

[1]Laboratori d'Enginyeria Química i Ambiental, Universitat de Girona. Campus Montilivi s/n 17071Girona. Campus Montilivi s/n 17071 Girona.
[2]Knowledge Engineering and Machine Learning group & Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya. C/ Pau Gargallo, 5. Barcelona(*).

En aquest estudi es porta a terme una classificació automàtica de dades per tal de trobar episodis típics que es poden donar en dominis poc estructurats com són els processos que tenen lloc a una estació depuradora d'aigües residuals (EDAR) juntament amb el coneixement dels experts . A partir d'aquesta informació es podrà determinar quins son les situacions en que pot presentar la planta i a la vegada serà un punt de partida per elaborar un sistema de raonament basat en casos que juntament amb un sistema de raonament basat en regles que formarà part d'un sistema supervisor com a suport a la presa de decisions .

1. INTRODUCCIÓ I OBJECTIUS

El procés de tractament d'aigües residuals, com altres processos biotecnològics, genera una gran quantitat de dades. Les dades que provenen de les EDARs, com la majoria de les dades ambientals son ambigües i tenen particularitats que fan difícil la interpretació. L'Estadística és una ciència molta antiga que neix de la necessitat d'analitzar dades i ara en l'actualitat es pot definit com la ciència que s'ocupa de la recollida i anàlisi de dades amb l'objectiu d'extreure la informació que contenen i presentar-la de forma resumida i esquematitzada

Dins de les múltiples branques de l'Estadística existeixen les tècniques de clustering o classificació automàtica. Les tècniques de clustering es pot dir que neixen el 1963 quan Sokal i Sneath publiquen la obra *Numerical Taxonomy* com alternativa a les anàlisis factorials. El desenvolupament de la informàtica va ser crucial, per que aquest tipus de tècniques prenguessin cos, al incrementar la potència del càlcul i facilitar tasques que sense ells serien extremadament costoses

Així neixen nous algorismes que milloren les propostes originals, neixen els processos jeràrquics, els processos de partició, mètodes de classificació piramidal... Però en l'àmbit estadístic tots aquests mètodes comparteixen un substrat comú: la definició d'una mesura de distància entre objectes i un criteri de qualitat que permeti escollir la millor distància en un sentit algebraic

El 1983, Michalsky [Michalsky & Stepp 1983], presenta el clustering conceptual que basant-se en els clàssics algorismes de classificació ascendent jeràrquica, canvia de paradigma i es recolza em l'associació de conceptes a les classes i generalització successiva dels mateixos. En aquests casos el substrat és la lògica de Predicats de primer ordre i a partir d'aquí s'han desenvolupat diversos mètodes de classificació automàtica en el si de l'aprenentatge automàtic com AUTOCLASS [Cheeseman et al. 1988], CODWEB [Fisher 1987]

L'Objectiu d'aquest treball es aplicar tècniques de clustering, en un domini ampli i poc estructurat com en el camp de les aigües residuals, per adquirir així sota la supervisió dels experts bases de coneixement per la gestió d'EDARs

2. EL DOMINI

2.1. LES PLANTES DEPURADORES D'AIGÜES RESIDUALS

El camp d'aplicació és la línia d'aigües d'una EDAR Diverses operacions i processos unitaris són requerits per tal de tractar adequadament les aigües residuals. La combinació d'aquestes operacions i processos, tant físics, químics com biològics, conformen el diagrama de procés de cada estació depuradora. El procés global sempre segueix una seqüència lògica de tractament [Metclaf & Eddy 2003]:

- Una primera etapa, anomenada pretractament, que realitza un primer desbast groller dels sòlids més grans arrossegats per l'aigua residual que arriba del col·lector. La finalitat del pretractament és evitar possibles obturacions posteriors, així com eliminar l'efecte abrasiu d'aquests materials sobre mecanismes com les bombes i vàlvules que es troben al llarg del procés. Aquesta operació física es sol realitzar mitjançant una seqüència de reixes, amb diferent obertura i automatisme, però també existeix la possibilitat d'incloure un triturador que redueixi la mida de les partícules i n'eviti la seva separació. L'addició d'un desarenador a continuació permet separar les sorres més fines i els greixos o olis presents, aprofitant la major velocitat de sedimentació

(*) Aquesta recerca ha estat parcialment finançada pel projecte TIC'2000.

de les primeres, i la flotació dels segons, fet que s'afavoreix amb l'aportació d'un cabal controlat d'aire i el característic disseny del desarenador.

- Una segona etapa on l'aigua es deixa reposar unes hores a un tanc de sedimentació primària, per tal que decanti la matèria orgànica sedimentable, així com la resta de sorres o partícules inorgàniques que no han quedat retingudes al pretractament. Els sòlids sedimentats són enviats cap a una línia de tractament específic, la línia de fangs, essent habitual el seu pas previ per un garbell que separi les petites partícules inorgàniques contingudes. Quan la càrrega és força elevada, o el temps de retenció és insuficient, es pot complementar la decantació natural de la matèria en suspensió amb l'addició de coagulants químics que afavoreixin la seva floculació. Aquest tractament químic és gairebé obligat quan l'aigua contingui metalls o algun tòxic que pugui malmetre el funcionament de la posterior etapa biològica.
- Seguidament l'aigua arriba ja a l'etapa més important del procés anomenat tractament secundari. El fonament d'aquesta etapa no és altra que accelerar el procés biològic que es donaria a la natura, és a dir, la degradació, per part d'una població multiespecífica de microorganismes, de la matèria orgànica dissolta a l'aigua residual. Aquesta reacció té lloc a uns bioreactors (Figura 1), fortament airejats en cas que el procés sigui aerobi.

- Seguint el camí que recorre l'aigua dins l' EDAR, l'última de les etapes habituals és una nova decantació a uns sedimentadors secundaris (Figura1). L'objectiu és assolir una bona separació entre l'aigua tractada i la biomassa present. El sobrenedant, ben clarificat, sol ser abocat directament cap al llit receptor on segueix el seu cicle natural. La major part dels microorganismes separats al decantador és retornada cap el reactor biològic per tal de mantenir el nivell de depuració necessari, mentre que una petita fracció és apartada diàriament del sistema i enviada cap a la línia de fangs, per tal d'evitar un augment i envelliment excessiu de la biomassa present al sistema. Aquestes dues accions, clau per garantir el correcte desenvolupament del procés de fangs actius, són les anomenades recirculació (RASS) i purga(WASS)

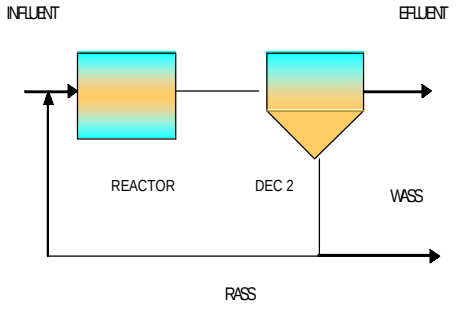


Figura 1. Esquema del procés de llots actius

2.2. COMPOSICIÓ DE L'AIGUA RESIDUAL URBANA

S'ha definit una sèrie de paràmetres que permeten quantificar i normalitzar el nivell de contaminació de les aigües residuals, establint-se una classificació principal [Metcalf & Eddy 2003] que els diferencia entre físics (temperatura, aigua, color i terbolesa) i químics (principalment sòlids, matèria orgànica, nutrients, pH, alcalinitat, duresa, clorurs i greixos).

2.2.1. PARÀMETRES FÍSICS

Temperatura: varia en funció de l'estació de l'any, però sol ser lleugerament superior a la de l'aigua corrent. Té efectes sobre l'activitat microbiana, la solubilitat dels gasos, i la viscositat

- Olor: l'aigua residual fresca es caracteritza per una olor lleugerament desagradable que denota presència d'olis i detergents. Quan envelleix es produeixen males olors com a resultat de la descomposició dels productes que conté l'aigua.
- Color: l'aigua residual presenta un color gris clar, però s'enfosqueix amb el pas del dies o en condicions de sèptiques. Qualsevol altra color que presenti és producte de la presència de determinats compostos (tints, sang, crom, residus làctics, etc..)
- Terbolesa: manca de transparència deguda a la presència d'una àmplia varietat de sòlids en suspensió presents a l'aigua residual

2.2.2. PARÀMETRES QUÍMICS

- Sòlids totals (ST): aquest paràmetre inclou tant la matèria en suspensió com dissolta.
  - ➢ Sòlids en suspensió totals (SST o MES): correspon a la fracció de sòlids, orgànics i inorgànics, que no estan dissolts. Aquests a l'hora es poden classificar com:
    - ♦ Fixes (SSF): compostos minerals o fracció no combustible dels SST (mg/l).
    - ♦ Volàtils (SSV): compostos orgànics o fracció combustible dels SST (mg/l).
  - ➢ Sedimentables: fracció de sòlids en suspensió, orgànics i inorgànics, que sedimenta en 1 hora en un con d'Imhoff. Representa aproximadament el fang que es pot eliminar al tanc de sedimentació (ml/l).
  - ➢ Dissolts: fracció de sòlids, orgànics i inorgànics, que no és filtrable. Inclou tots aquells sòlids inferiors a 1 mil.limicra (mμ).
    - ♦ Fixes: compostos minerals o fracció no combustible dels sòlids dissolts totals (mg/l).
    - ♦ Volàtils: compostos orgànics o fracció combustible dels sòlids dissolts totals (mg/l).

- Matèria orgànica : actualment són tres els mètodes que s'utilitzen per determinar el contingut orgànic en les aigües residuals.
  - ➢ Demanda bioquímica d'oxigen (DBO_5): representa la fracció orgànica biodegradable present a l'aigua residual, i és una mesura de l'oxigen dissolt requerit pels microorganismes per consumir aquesta matèria orgànica en 5 dies i a 20 ºC (mg/l).

- Demanda química d'oxigen (DQO): mesura de la fracció de matèria orgànica que és oxidada químicament en utilitzar un agent oxidant fort (dicromat de potassi). També es mesura per la quantitat estequiomètrica d'oxigen dissolt requerit (mg/l). Aquest paràmetre sol ser superior a la DBO_5 en ser més gran el número de compostos que poden oxidar-se químicament que per via biològica.
- Carbó orgànic total (COT): determina el contingut en matèria orgànica de l'aigua mitjançant la conversió d'aquest carbó a CO_2 a altes temperatures i en presència d'un catalitzador (mg/l). Aquesta tècnica és especialment aplicable quan les concentracions de matèria orgànica són petites.
- Nitrogen total (NT): el nitrogen es pot trobar en l'aigua formant part de compostos orgànics, com a nitrogen amoniacal (com ió amoni o amoníac depenen del pH) i com a nitrit i nitrat.
- Nitrogen orgànic (Norg): inclou el nitrogen lligat a les proteïnes, als aminoàcids i a la urea (mg/l).
- Nitrogen amoniacal (NH_4^+): primer producte de la descomposició del nitrogen orgànic (mg/l).
- Nitrogen Kjeldah l(TKN): paràmetre resultant de la suma dels dos anteriors, el nitrogen amoniacal i el nitrogen orgànic (mg/l).
- Nitrits i nitrats(NO_3^-): formes més oxidades del nitrogen (mg/l).

- Fòsfor total (P_T): el fòsfor es pot trobar formant part dels compostos orgànics i com a polifosfat i ortofosfat essent aquesta última espècia l'única directament assimilable pel metabolisme biològic.
- Orgànic (P_{org}): fracció del fòsfor que es troba lligat a la matèria orgànica (mg/l).
- Inorgànic (P): fracció inorgànica del fòsfor que existeix com a ortofosfats i polifosfats (mg/l).

- pH: indicatiu de la natura bàsica o àcida de l'aigua residual.

- Alcalinitat: deguda a la presència d'ions bicarbonat, carbonat i hidròxid a l'aigua residual, ofereix resistència als canvis de pH (mg $CaCO_3$/l).

- Duresa: deguda principalment als ions calci i magnesi dissolts a l'aigua (mg $CaCO_3$/l).

- Clorurs: proporcionen major conductivitat a l'aigua i n'augmenten la seva densitat (mg/l).

- Greixos: fracció de matèria orgànica soluble en hexà. Inclou greixos i olis d'origen animal i vegetal (mg/l).

2.3. LEGISLACIÓ

El Pla de Sanejament de la Generalitat de Catalunya, basant-se en la directiva del Consell 91/271/CEE (Taula 1) del Diari Oficial de las Comunitats Europees, insta a la depuració dels principals components de l'aigua residual urbana que s'aboqui a llera pública.

Taula 1. Quadre resum de la directiva 91/271 CEE

Paràmetre	Concentració	Percentatge de reducció
DBO_5 (mg O_2 /l)	25	70 - 90 %
DQO (mg O_2 /l)	125	75 %
SS (mg/l)	35	90 %
Fòsfor total (mg-P /l)	2 1 (> 100,000 h.-e.)	80 %
Nitrogen Total (mg-N/l)	15 10 (> 100,000 h.-e.)	70 - 80 %

S'estableix un límit de fins a 25 mg/l per la DBO_5, amb un percentatge mínim de reducció del 70-90 %, mentre que pels SS el límit queda fixat en 35 mg/l amb un percentatge mínim de reducció del 90%. En quant als abocaments a zona sensible (amb risc d'eutrofització), s'hi addiciona un límit pels nutrients establert en 1 mg/l (i 80% de reducció) pel fòsfor, i en 10 mg/l pel nitrogen total (amb 70-80 % mínim d'eliminació). Aquests límits són lleugerament més permissibles en cas que la depuradora tracti cabals petits (d'entre 10,000 i 100,000 habitants equivalents).

3. PLANTA OBJECTE D'ESTUDI

Els resultats obtinguts d'aquest estudi provenen línia d'aigües d'una planta, situada dins de la conca del riu Besòs, a prop de Barcelona. La planta rep un cabal diari aproximat de 27000 m^3/dia. El diagrama de flux de l'EDAR , es pot veure a la Figura 2 i consta de: pre-tractament i un tractament biològic de doble etapa

3.1. DESCRIPCIÓ DE LA PLANTA

El pre-tractament consta d'un sistema de desbast tant de fins com de gruixuts. Els primers tenen una llum de 10 mm mentre els altres de 6. La planta disposa d'un dessorrador-desgreixador per injecció d'aire. El tractament biològic esta format per una doble etapa. La primera consta de dos reactors aerobis tipus mescla complerta de 1590 m^3 de volum airejats per difusors de membrana. Aquesta etapa opera a alta càrrega (2 Kg DBO_5/Kg MLSS) i presenta una sonda d'oxigen a la sortida del reactor amb una consigna pre-establerta establint un llaç de control. La planta disposa de dos decantadors primaris de 2200 m^3 circulars. La segona etapa consta de tres reactors tipus flux pistó de 4790 m^3 cadascun, que al igual que en el cas anterior estan airejats per difusors de membrana. Presenta un càrrega inferior (0,26 kg DBO_5/Kg MLSS) a l'anterior etapa i amb un llaç de control per l'oxigen. Finalment l'aigua roman durant unes hores en tres decantadors secundaris de 2600 m^3 cadascun, on es separa la biomassa tractada del líquid clarificat.

Tant dels decantadors de la primera etapa com dels de la segona, en surt la purga i la recirculació. La purga en la segona etapa pot anar directament a espessidors o pot anar a primera etapa o es barreja amb el fang primari (ja que aquest presenta dificultats al decantar, millorant d'aquesta manera les seves condicions) i és enviat a espessidors

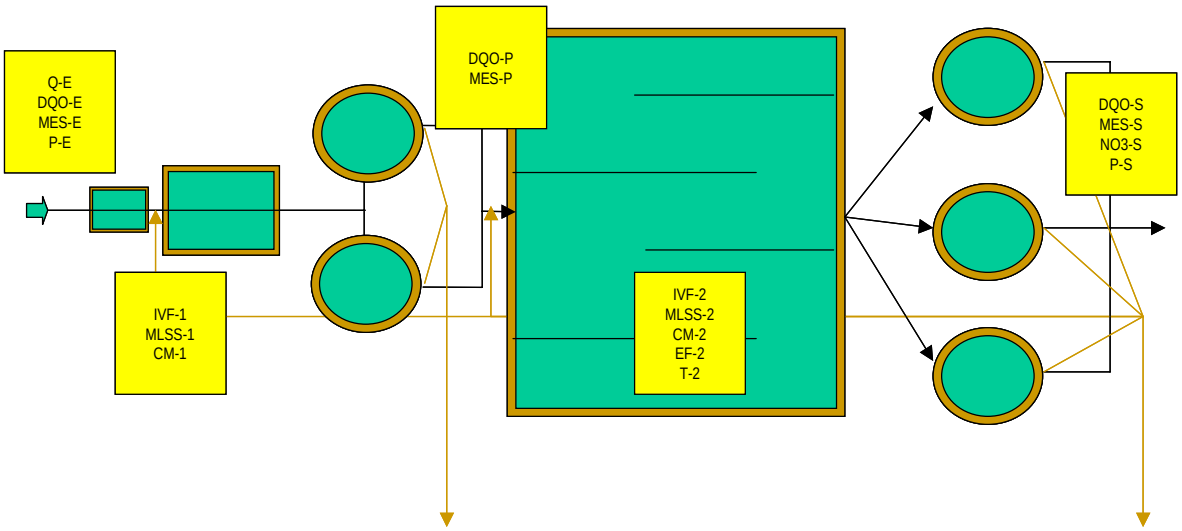


Figura 2. Diagrama de flux de la línia d'aigües de l'EDAR objecte d'estudi

3.2. LES DADES

Les dades estudiades en aquest treball (Taula 2)corresponen als dies d'operació de la planta, durant el període comprés entre el gener i el maig de 2002. Son un total de 149 dies i s'han considerat 18 variables corresponents a diferents mesures preses a diferents punts de la planta

Taula 2. Variables utilitzades en la classificicació

variable	Nom	unitats
Q	Cabal entrada	m3/dia
DQO-E	Demanda química d'oxigen	mg/l
DQO-P		mg/l
DQO-S		mg/l
MES-E	Sòlids en suspensió	mg/l
MES-P		mg/l
MES-S		mg/l
NO3-S	nitrats	mg/l
P-E	fosfats	mg/l
P-S		mg/l
MLSS-1	Sòlids en el reactor -1	mg/l
IVF-1	Índex volumètric al reactor 1	
CM-1	Càrrega màssica al reactor-1	Kg DBO/kg MLSS/
T-2	Temperatura al reactor-2	mg/l
MLSS-2	Sòlids en el reactor -2	mg/l
IVF-2	Índex volumètric al reactor 2	mg/l
EF-2	Edat del fang al reactor -2	mg/l
CM-2	Càrrega massica al reacitor-2	Kg DBO/kg MLSS/

4. ANÀLISI EXPLORATÒRIA DE DADES

Un cop definides les variables que es classificaran, es farà una descripció exhaustiva de cadascuna d'elles. L'anàlisi descriptiu, permet una primera aproximació, sobre la composició de la mostra

Aquest anàlisi es portarà a cap sobre els 149 dies i les 18 variables disponibles. Aquest anàlisi es pot veure en els sumaris estadístics anteriors i consta de màxims, mínims, mitjanes, desviació estàndard, coeficient de variació i els quartils. (Taula 3)

També es farà una representació gràfica de d'algunes de les variables al llarg del període d'estudi (Figures 3,4,5 i 6)

A partir dels valors dels sumaris estadístics i l'evolució temporal de les variables es poden fer un seguit d'apreciacions.

Taula 3 .Resultat de l'anàlisi exploratòria de dades

	Q	DQO-E	DQO-P	DQO-S	MES-E	MES-P	MES-S	NO3-S	P-E	P-S	MLSS-1	IVF-1	CM-1	T-2	MLSS-2	IVF-2	EF-2	CM-2
MIN	12480	264	126	43	92	44	5	0.3	4.7	0.4	200	6	0.7	14.9	935.0	67.0	2.7	0.1
MAX	44630	838	664	230	724	375	70	1	11	4.5	5700	177	22.8	23.5	3750.0	527.0	4.8	0.6
MEAN	30040	472	345	92	286	133	20	0.4	7.12	1.5	2484	51	5.8	18.7	2330.9	207.9	3.3	0.3
DESV	5987.8	142.5	119.8	25.6	104.3	47.1	9.2	0.2	1.8	0.9	1891.4	29.5	6.6	2.3	650.5	131.4	0.4	0.2
CV	0.20	0.30	0.35	0.28	0.36	0.35	0.46	0.44	0.25	0.63	0.76	0.58	1.14	0.12	0.28	0.63	0.11	0.58
Q1	25848	362	245	74	222	104	15	0.3	6	0.9	351	22	1.2	17.0	1762.5	93.2	3.0	0.1
Q3	34186	540	435	106	350	149	23	0.4	8.1	1.8	4185	68	10.5	20.0	2860.0	290.5	3.6	0.4
MED	29990	456	358	90	268	120	19	0.3	7	1.3	2905	56	2.0	18.5	2247.5	170.0	3.3	0.2

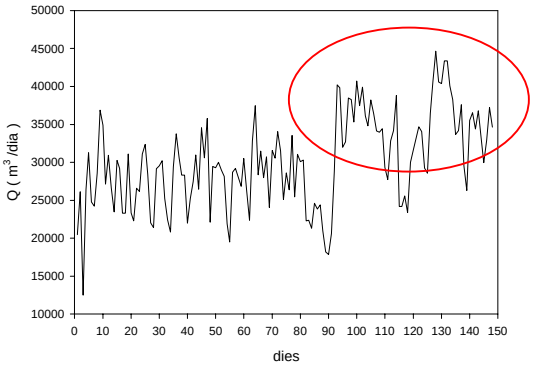


Figura 3. Representació del cabal en el temps

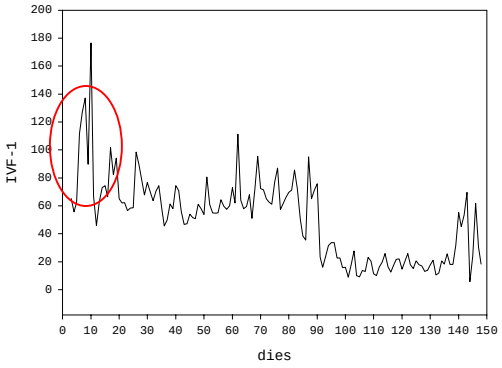


Figura 5. Representació de IVF-1 en el temps

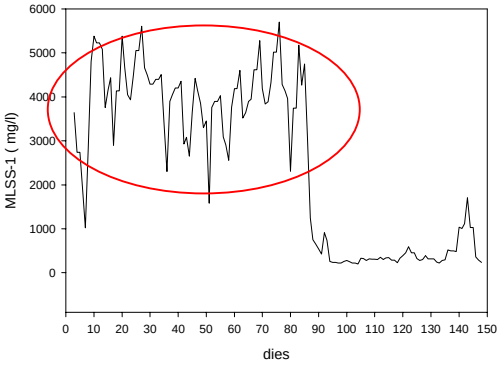


Figura 4. Representació de la MLSS-1 en el temps

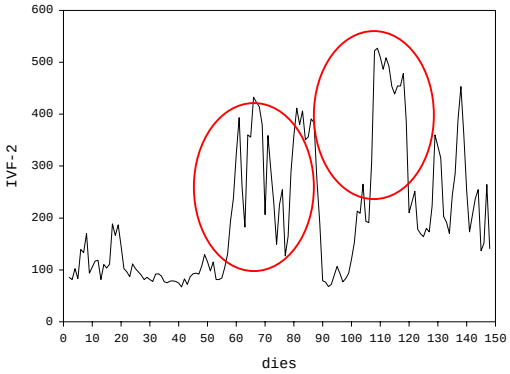


Figura 6. Representació de la IVF-2-1 en el temps

La variable Q-E, presenta una gran variabilitat, amb una pujada important dels valors al final del període d'estudi. Això es degut a que aquests dies corresponen al mes de maig, mes caracteritzat per intenses pluges, cosa que incrementa l'arribada d'aigua a l planta. Si es comparen els valors dels quartils amb els màxims es pot observar que les desviacions son degudes a puntes.

La variable MLSS-1 presenta dos períodes diferenciats, els primers 90 dies amb valors elevats i la resta del període on els valors son baixos. Això es degut a la flexibilitat operacional que presenta la planta on la primera etapa pot funcionar com a tal mantenint una població estable de microorganismes (valors elevats de MLSS-1) o com a pre-aeració, eliminant el cultiu microbiològic a la primera etapa (valors baixos de MLSS-1). Tot i això es pot observar que en els períodes on la primera etapa funcionava com a tal presentava elevats valors de IVF-1, cosa que li ocasionava problemes de separació de sòlids. Es per aquesta raó que es va obtar per treballar com a preaireació, eliminant el cultiu a primera etapa (valors baixos de MLSS-1) amb una consequent baixada de la IVF-1 i millorant així la decantació a primera etapa

La variació de la IVF-2 presenta al final del període una pujada de valor. Si s'observa l'evolució de la variable Q-E es pot observar que les dues incrementen a la vegada. Això es degut a que els periodes de pluja de ocasionen desequilibris entre els nutrients, cosa que afavoreix el desenvolupament de Zooglea, que dificulta la sedimentabilitat del fang

.

5. CLASSIFICACIÓ AUTOMÀTICA DE LES DADES

5.1. EL PROCÉS DE CLASSIFICACIÓ AUTOMÀTICA

La classificació automàtica és una part de la estadística que te com a objectiu agrupar objectes o individus per variables o característiques homogènies i diferenciades entre elles. Aquest tipus de tècniques són apropiats per a grans matrius de dades sense absència d'estructura a priori. Els resultats de les classificacions es poden obtenir per arbres jeràrquics o per particions

El procés de classificació que s'utilitzarà es del tipus jeràrquic. Les classificacions jeràrquiques es basen en particions successives de la matriu de dades E en classes cada cop més fines obtingudes per dicotomies (algorismes descendents) o menys fines obtingudes per reagrupaments (algorismes ascendents).

Els resultats es poden representar amb dendogrames o arbres de classificació. Com més properes estan els individus en el dendograma, més similars són (Figura 7)

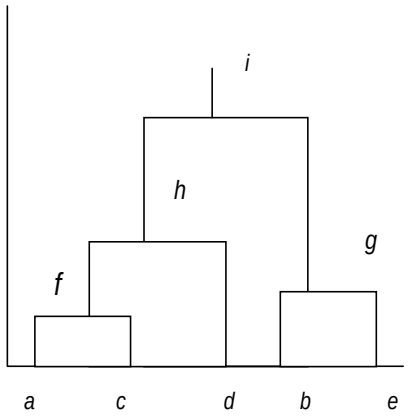


Figura 7. Representació d'un dendograma

El procés de classificació ascendent jeràrquic es basa en la mesura de la proximitat entre individus fusionant dos elements *a* i *b* en un nou element *h*. Llavors es torna a calcular la distància amb el nou

La distància utilitzada es l'Euclídea normalitzada per tractar-se de variables continues i pel gran varietat d'unitats existents. L'expressió d'aquesta distància

$$d_{ii'} = \left\{ \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{i'j}}{s_j} \right)^2 \right\}^{\frac{1}{2}}$$

Els algorismes corresponents a classificacions ascendents jeràrquiques (CAJ) es poden formular

Es defineix *E= E'*
Calcular la matriu de distàncies D (E,D)
Mentre Card (E')>1

1. *Trobar els dos elements més propers* (D,a,b)
2. *h =a agregat amb b*
3. *E'=E'-(a, b)+(h)*
4. *Actualitzar matriu de distàncies (E',D)*

Fimentre

El procés d'agregació contempla diverses possibilitats, en aquest informe s'ha utilitzat el criteri de Ward. Aquest es defineix
Siguin *x, y, z* tres objectes
Sigui h l'agrupació d' *x* i *y*
Sigui n_x, n_y , n_z i n_h els cardinals de *x, y,* z i *h*
Sigui g_x, g_y, $g_h$els centres de gravetat dels objectes *x, y* i *h*

$$D = \frac{\left\{(n_x + n_x)d(x,z) + (n_y + n_z)d(x,y) - n_z d(x,y)\right\}}{(n_x + n_y + n_z)}$$

Aquest tipus d'agrupació busca obtenir a cada pas hi hagi la mínima pèrdua d'inèrcia agregant els dos objectes que ho acompleixin. La variació de la inèrcia en un pas és:

$$\Delta I = d^2(g_x, g) + d^2(g_y, g) - d^2(g_h, g) =$$

$$= d^2(g_x, g_y) > 0$$

I d'aquí s'obté aïllant D(h ,z)

L'índex de nivell és la distància a la que es troben dos elements abans de ser agregats (Figura8)

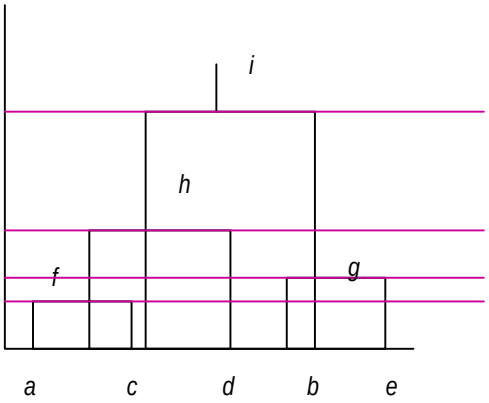


Figura8. Representació de l'índex de nivell entre els diferents talls del dendograma

Per obtenir les diferents classes, es talla l'arbre per diferents zones (Figura9)

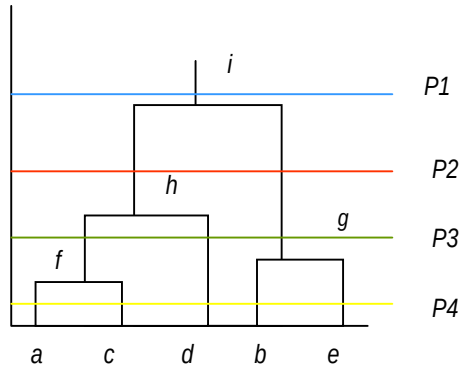


Figura9. Representació dels talls en el dendograma que donaran les classes

P4 és la partició més fina i conte tots els elements {(a),(b),(c),(d),(e)}
P1 és la partició més gruixuda: conté conjunt E {(a,b,c,d,e,)}
P2 conte {(a,c,d),(d,e))}

Les classes resultants seran més diferenciades com més gran sigui el salt en l'índex de nivell, per tant es tallarà allà on hi hagi salts més grans en l'índex de nivell.

5.2. SOFTWARE UTILITZAT

5.2.1. KLASS

KLASS és un sistema orientat a la classificació automàtica de dominis poc estructurats implementat amb LISP, del que s'ha utilitzat una versió pel sistema operatiu Solaris (sobre UNIX) de SUN. S'ha desenvolupat al Departament de EIO de la Universitat Politècnica de Catalunya, a partir de la tesi de K.Gibert. Per aquest treball s'ha treballat amb variables continues, de forma que els resultats obtinguts serien els mateixos que en paquets comercials (SPSS,SAS..)

5.2.1.1. ESTRUCTURA DELS FITXERS

<Fitxer.dat> És un fitxer que conte les dades en forma de matriu per files. Per a cada fila hi ha una llista de coordenades que defineixen cadascun dels valors de les diferents variables

<Fitxer.pro>. Conte la informació referida a les variables segons els quals s'ha descrit l'aigua de cadascun dels dies. Cada variable te el nom i el seu índex numèric associat, el tipus de variable (qualitativa o numèrica) i nombre d'individus.

<Fitxer.obj>. Fitxer que conte la caracterització dels objectes de la mostra. Al igual que en el cas anterior, hi ha un identificador dels objectes (el dia de les mesures), un índex associat i una llista de propietats que ho descriuen

5.2.2. EXCEL I SIGMAPLOT

Són dos programes de gran difusió en el mon universitari. S'han fet servir per la facilitat de manipulació a l'hora de fer la descriptiva de les classes proporcionades per KLASS tant de forma numèrica com gràfica i l'accessibilitat que s'hi tenia.

6. INTERPRETACIÓ DELS RESULTATS

Al finalitzar la classificació de les dades per KLASS, es pot triar quantes classes volem que ens ofereixi el programa. Normalment el programa ofereix la llista dels millors talls, en funció del salt en l'índex de nivell.

Per interpretar els resultats s'han fet servir diagrames de caixa múltiples(annex 1), per trobar *variables caracteritzadores,* variables que entre el primer i el tercer quartil prenen valors exclusius d'aquesta classe. Si no es troben variables caracteritzadores es pot treballar amb variables parcialment caracteritzadores, on és un fragment del diagrama de caixa, el que pren valors exclusius de la classe. Un pas molt important és la interpretació de resultats per part de l'expert, un cop les dades han estat classificades. Donada una classificació qualsevol, i sigui quin sigui el seu origen, l'objectiu és la identificació de les variables més rellevants en cadascuna de les classes formades, d'una forma àgil i poc costosa des del punta de vista computacional, de forma que es pugui establir el significat de les classes resultants.

Un altre mètode consisteix en fer una comparació de la mitjana del grup en relació a la mitjana total (annex3), Per veure si una variable x te importància es fa un contrast sobre la mitjana en la classe sota la hipòtesi que els individus han estat assignats a aquella classe de forma aleatòria. Per fer-ho es fa a partir del següent estadístic:

$$t(x) = \frac{x_j - x}{s_j(x)}$$

$$s^2{}_j(x) = \frac{n - n_j}{n - 1} \frac{s^2(x)}{n_j}$$

El coneixement del domini jugarà un paper molt important en la interpretació dels resultats. Gràcies a aquest coneixement es podran triar les variables més rellevants, per donar una categoria a les classes.

A partir del coneixement del domini, i les ajudes a la interpretació, s'han obtingut situacions que defineixen les classes proporcionades per KLASS. Se li ha donat el nom de variables indicadores, a les variables (tant caracteritzadores, com parcialment caracteritzadores) per determinar aquests episodis o situacions típiques

7. RESULTATS DE LES CLASSIFICACIONS

Un cop el programa ha classificat les dades, es decideix tallar el dendograma en un punt on s'obtenen 12 classes.

La primera gran diferenciació que es pot fer a partir dels valors promig de les dades, es si la primera etapa funciona com a primera etapa o com a pre-aeració. Si la primera etapa funciona com a tal presenta valors de MLSS-1 més elevats de 3000 mg/l Si funciona com a pre-aeració presenta valors de MLSS-1 més baixos de 3000 mg /l. Si la primera etapa funciona com a pre-aeració no es recircula licor mescla i per tant tots els microorganismes que es poguessin desenvolupar s'eliminen per rentat (veure Figura 19)

Les sis primeres classes corresponen al funcionament de la primera etapa com reactor biològic. Dins d'aquestes classes es diferencien 6 situacions:

- classe 1: situació de normalitat, aquesta es caracteritza per no tenir valors que difereixen molt dels promitjos i que ronden la normalitat

- classe 2: situació de normalitat però amb una alta eficiència d'eliminació a primera etapa. Hi ha valors de MLSS-1 elevats que asseguren una població abundant de microorganismes a primera etapa (veure Figures 12 i 21)

- classe 3: situació de bulking a primera etapa. Aquest episodi es caracteritza pel desenvolupament de microorganismes filamentosos que dificulten la sedimentabilitat del fang. Els indicadors que ens informen de la presència de bulking son valors de IVF-1 elevats (veure Figura 19)

- classe 4: sobrecàrrega orgànica, en aquesta situació hi ha una arribada d'una càrrega no biodegradable, que comporta una sortida de sòlids i DQO (veure annex 3)

- classes 5 i 6: aquestes classes corresponen a episodis de bulking. Les primeres per ser una transició al bulking amb valors de IVF-2 intermitjos i la segona per tractar-se d'episodis de bulking seriosos (veure Figura 22). Les sis següents classes corresponen al funcionament de la planta com pre-aeració. Dins d'aquestes classes es poden diferenciar 6 situacions:

- classes 7 i 12: situació de normalitat funcionant com a pre-aeració, la primera es diferencia de la segona ja que la primera no presenta casi sòlids a primera etapa, MLSS-1 molt baixos i la segona per estar amb un estat de transició entre pre-aeració i primera etapa amb valors de MLSS-1 rondant als 1000 (veure Figura 19)

- classe 8: La planta nitrifica. Això es degut al desenvolupament de microorganismes autrotròfics, encarregats de catalitzar les reaccions d'oxidació de l'amoni a nitrat. Aquesta classe presenta valors elevats de nitrat (veure Figura 17)

- classe 9: En aquesta ocasió la planta ha rebut una entrada molt gran de cabal, degut a les pluges que hi ha en l'època (abril i maig). Aquesta situació es caracteritza per tenir valors elevats de cabal. Les pluges també porten associats fenòmens de rentat dels microorganismes i desequilibris entre nutrients que portaran a situacions de bulking (veure Figures 10 i 21)

- classe 10 i 11: es tracta de fenòmens de bulking causat per zooglea, el bulking viscós normalment té lloc després de períodes de pluja degut als desequilibris entre nutrients que arriben a la planta i associat a situacions d'alta càrrega . L'altra classe es la transició del bulking a a la normalitat (Figura 22 i 23)

Un cop s'han obtingut les classificacions es pot representar la tendència temporal de les classes.(annex2)

8. CONCLUSIONS

A partir de l'informe que s'ha realitzat i després de treballar amb diferents tècniques de classificació automàtica, s'ha arribat a una sèrie de conclusions

Les bases de dades ambientals són complexes i estan poc estructurades. L'Ús de tècniques estadístiques clàssiques com l'anàlisi exploratòria de dades pot ajudar a augmentar el coneixement es te del domini

Les tècniques de cluster són un bon mètode, per explorar aquests tipus de bases de dades, ja que permeten obtenir coneixement del domini sobre el que es treballa a partir dels diferents valors que poden tenir les variables involucrades en el procés.

L'ús de diagrames de caixa múltiples i les proves d'hipòtesi per veure si una variable és rellevant en la classificació són eines molt útils a l'hora d'interpretar les classes, que et proporciona el programa, on es pot apreciar la importància de la variable dins de la classe i respecte la mostra

Les tècniques de cluster també son un bon mètode per determinar episodis típics o escenaris que poden donar-se en determinats processos, com en les plantes depuradores d'aigües residuals de la que no se'n te coneixement, a partir dels diferents valors que prenen les variables involucrades en el procés

Les tècniques de cluster poden ser un bon mètode poden ser de gran ajut a l'hora de implantar diverses tècniques d'intel·ligència artificial com els sistemes experts i els sistemes de raonament basats en casos. En sistemes experts per saber quins episodis es donen en el procés i en sistemes de raonament basats en casos, per tenir una base de casos inicial

A partir dels resultats obtinguts en aquestes classificacions s'ha elaborat la base inicial de casos que formarà part del sistema supervisor d'una planta depuradora real

Està en marxa un segon estudi amb les dades de procés de tot un any. En aquest estudi es partirà d'un nombre de variables superior i se'n aniran eliminant prenen com a referència el test de Kruskal-Wallis.

9. REFERÈNCIES

[Comas *et al.*, 2000a] Comas J., Dzeroski S., Gibert K., R.-Roda I. & Sànchez-Marrè M.Knowledg Discovery by means of Inductive Methods in Wastewater treatment Plant. AI Communications, 14(1), pp. 45-62. 2001.

[Comas et al., 2000b] Comas J., Colprim J., Baeza J. and Poch M., A Hybrid Supervisory System to support Wastewater, Treatment Plant Operation: Implementation and Validation. IWA conference on Instrumentation, Control and Automation Malmo.2001

[Cheeseman *et al.*, 1988] Cheeseman P., kelly J., Self M., Stutz J., Taylor W., Freeman D.. AUTOCLASS: a Bayesian Classification system. Procc. Of 5th Conference on machine learning (ICML-1988).pp.54-64. 1988

[Comas *et al.*, 1999] Comas J., R.-Roda I., Ceccaroni L. and Sànchez-Marrè M. "Semi-automatic learning with quantitative and qualitative features", CAEPIA-TTIA'99 5th Conference of the Spanish Association for Artificial Intelligence, vol. 1, pp. 17-25, Ana Mª García Serrano, Ramón Rizo, Serafín Moral, Francisco Toldeo editors. Murcia (España). 1999

[Gibert K *et al.*, 1998] Gibert K.,Cortés U. " clustering based on rules and knowledge discovery in ill-structured domains" Computacion y sistemas. 1998

[Gibert K *et al.*, 2000] Gibert K., Salvador A.. Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. Congreso español sobre tecnologías i lógica fuzzy. 2000.

[Gibert K *et al.*, 2000] Gibert K., R-Roda I. Identifying characteristics situations in wastewater treatment plants. 2nd ECAI Workshop on Binding Environmental sciences and artificial intelligence. Berlin 2000

[Lebart, 1985] lebart L., and Morineau, A. and Fenelon J.P., Tratamiento estadístico de datos. Marcombo, 1985

[Metcalf and Eddy, 2003] Metcalf & Eddy. Wastewater engineering treatment, disposal, reuse. 4rd ed. revised by George Tchobanoglous, Frankliln L. Burton, McGraw-Hill, NY. 2003.

[Michalski and Stepp, 1983] Michalski R.S., Stepp R.E., Automated construction of classifications: Conceptual clustering versus numerical taxonomy, IEEE trans. On PAMI (5): 395-410,1983

[R-Roda, 1998] R.-Roda I., *Desenvolupament d'un Protocol per l'Aplicació de Sistemes Basats en el Coneixement a la Gestió d'Estacions Depuradores d'Aigües Residuals Urbanes*, PhD Thesis, Universitat de Girona 1998

[R-Roda, 1998] R-Roda I., Comas J., Colprim J., Poch M.,SànchezMarrè M., Cortés U., Baeza, J & Lafuente J..A hybrid supervisory system to support wastewater treatment plant: operation: implementation and validation. Water Science & Technology, 45(4-5), pp. 289-297.2002.

[Sànchez-Marrè *et al.*, 1998] Sànchez-Marrè M., R.-Roda I., Comas J., Cortés U. and Poch M. "L'Eixample Distance: A New Similarity Measure for Case Retrieval", CCIA'98 1st Catalan Conference in Artificial Intelligence, 14-15, pp. 246-253 Tarragona (Catalunya). October 1998.

[Sànchez-Marrè *et al.*, 1997b] Sànchez, M., Cortés, U., Béjar, J., de Gracia, J., Lafuente, J. and Poch, M. Concept formation in WWTP by means of classification techniques: a compared study. *Applied Intelligence*, 7, 147, 1997

10. ANNEX 1. Diagrames de caixa múltiples per interpretar i taules amb valors per interpretar els resultats

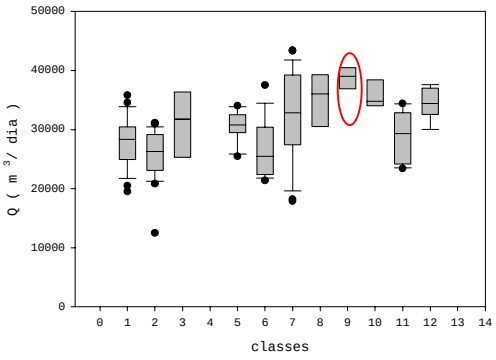


Figura 10 Box-plots múltiples per a Q-E

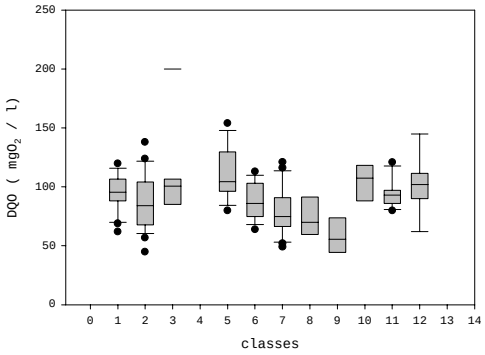


Figura 13. Box-plots múltiples per a DQO-S

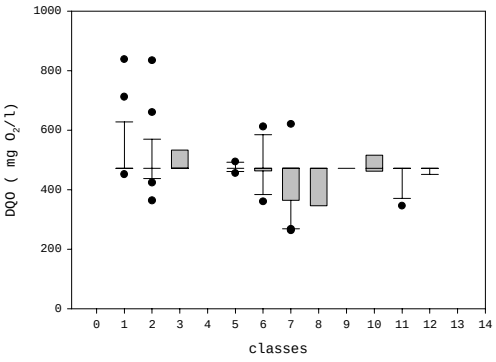


Figura 11. Box-plots múltiples per a DQO-E

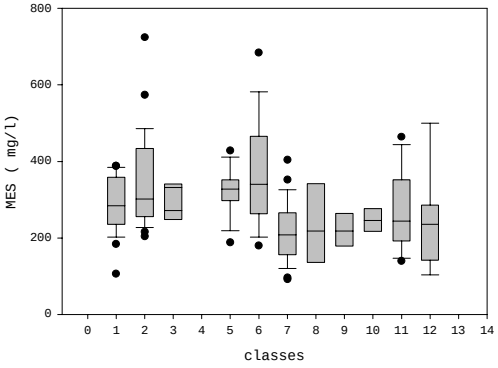


Figura 14. Box-plots múltiples per a MES-E

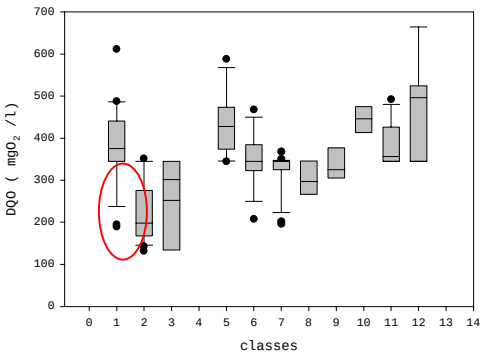


Figura 12. Box-plots múltiples per a DQO-P

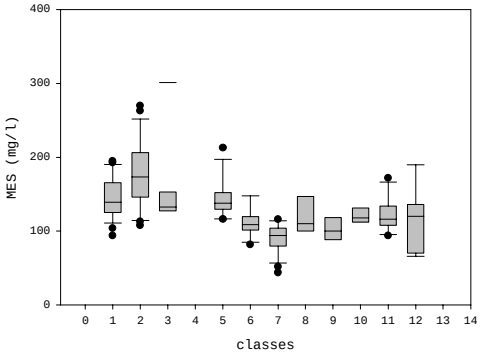


Figura 15. Box-plots múltiples per a MES-P

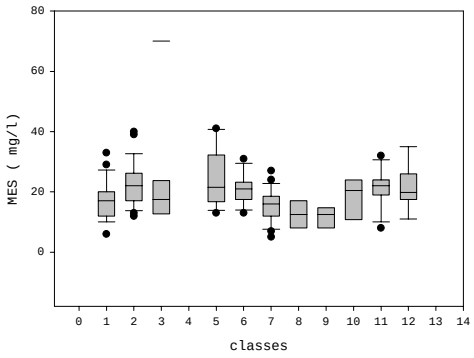


Figura 16. Box-plots múltiples per a MES-S

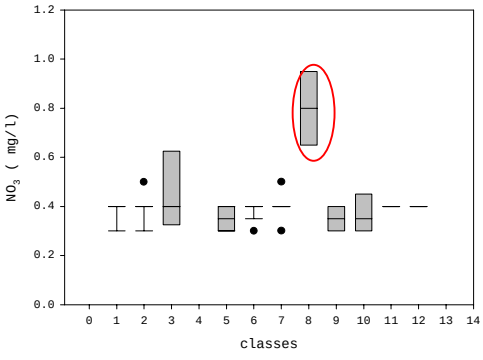


Figura 17. Box-plots múltiples per a NO_3-S

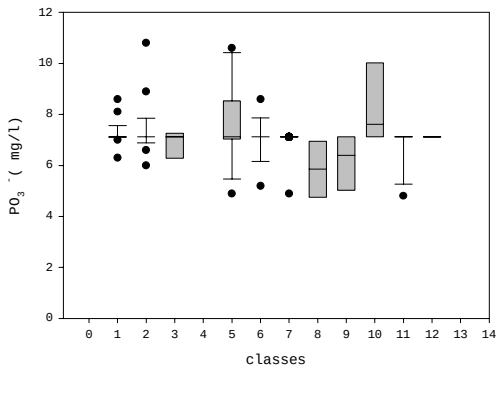


Figura 18. Box-plots múltiples per a PO_3-E

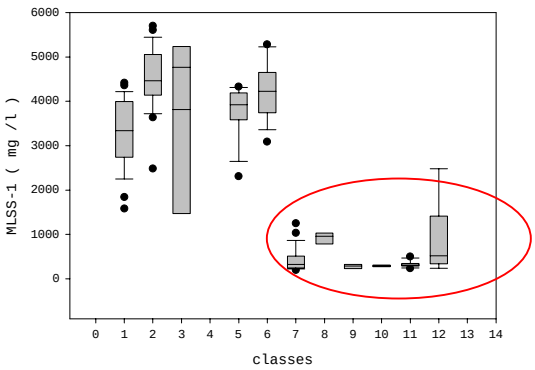


Figura 19. Box-plots múltiples per a MLSS-1

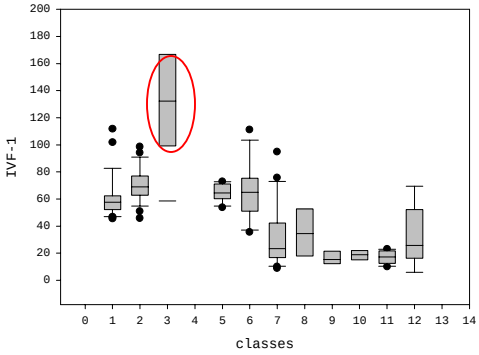


Figura 22. Box-plots múltiples per aIVF-1

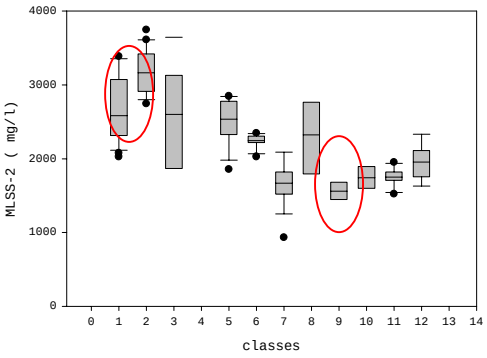


Figura 21. Box-plots múltiples per a MLSS-2

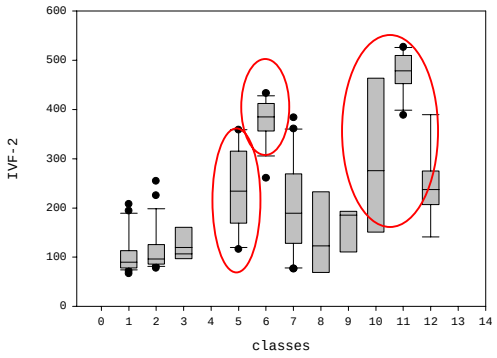


Figura 22. Box-plots múltiples per aIVF-2

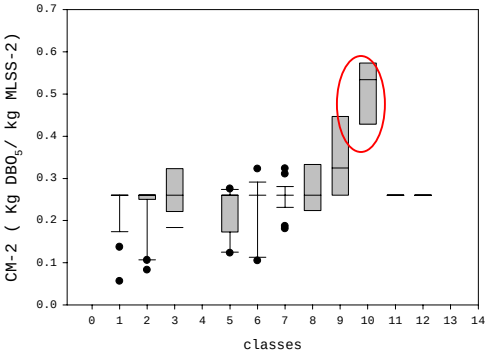
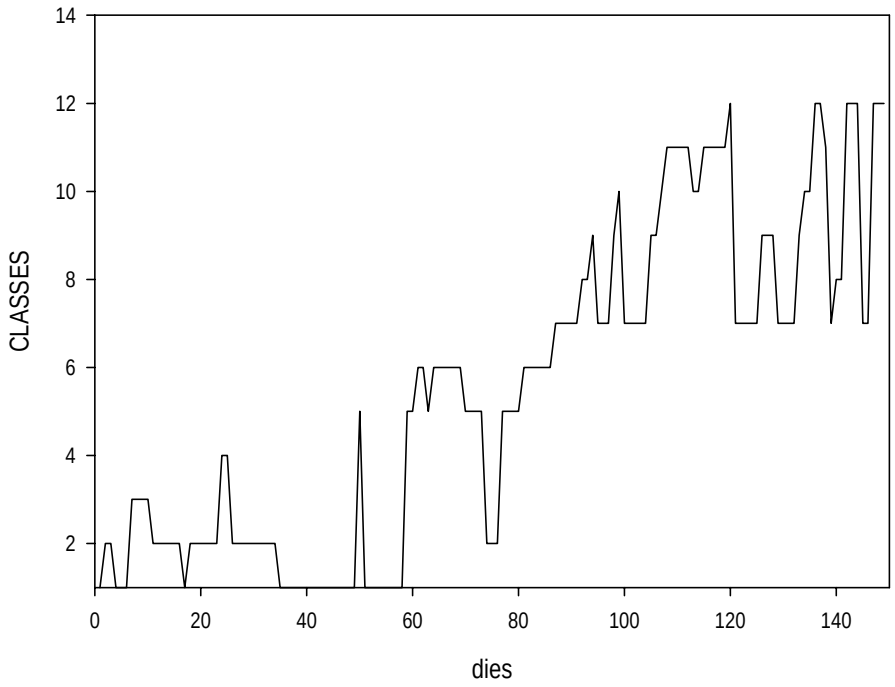


Figura 23. Box-plots múltiples per a CM-2

ANNEX 2. Evolució de les classes al llarg del temps



Annex 3. taules amb les mitjanes de les classes i valors del contrast de mitjanes

Taula 4. Valors de les mitjanes de les variables per a cadascuna de les classes

	Q	DQO-E(F)	DQO-P(F)	DQO-S	MES-E	MES-P	MES-S	NO3-S	P-E	P-S	MLSS-1	IVF-1	CM-1	T-2	MLSS-2	IVF-2	EF-2	CM-2
1.00	27703.00	502.79	381.59	96.00	289.50	143.89	17.29	0.38	7.19	1.39	3315.68	61.23	4.16	18.41	2691.46	103.57	3.55	0.24
2.00	25705.58	488.90	226.32	87.77	345.85	178.81	22.85	0.38	7.28	1.40	4529.50	70.84	4.17	17.47	3170.19	115.72	3.15	0.23
3.00	31156.75	492.23	243.92	97.50	287.00	137.80	18.00	0.45	6.89	2.93	3510.00	132.71	2.80	15.38	2532.50	125.70	2.95	0.27
4.00	31736.50	471.64	301.50	200.00	332.00	301.50	70.00	0.40	7.12	1.50	4770.00	58.48	5.76	16.50	3647.50	106.93	2.99	0.18
5.00	30608.42	473.56	433.74	111.83	324.33	144.25	24.58	0.35	7.59	1.68	3822.08	64.74	2.79	19.88	2511.67	240.65	3.08	0.23
6.00	26678.79	472.91	349.01	88.57	365.86	111.50	21.00	0.39	7.09	1.56	4254.64	66.46	4.41	18.44	2242.14	379.74	3.19	0.24
7.00	32309.29	430.72	324.82	79.11	216.16	89.76	15.48	0.40	7.03	1.48	434.37	32.50	5.83	18.65	1657.30	203.00	3.19	0.26
8.00	35271.00	429.73	303.00	73.75	232.00	119.00	12.50	0.80	5.85	1.65	925.00	34.97	3.65	20.75	2295.00	141.49	3.33	0.27
8.00	39319.25	451.19	346.75	59.38	221.00	102.75	12.00	0.39	6.13	1.18	290.56	17.14	14.93	17.50	1594.38	165.67	3.61	0.36
10.00	35743.00	481.15	446.67	107.00	246.00	122.00	18.33	0.38	8.36	1.50	292.50	18.84	17.70	20.83	1742.50	293.43	3.50	0.51
11.00	28649.82	460.07	387.47	94.27	263.27	123.09	21.27	0.40	6.91	1.45	326.05	17.17	7.08	19.97	1750.00	474.70	3.54	0.26
12.00	34455.15	469.46	454.81	101.29	243.06	111.84	21.51	0.40	7.12	1.50	920.14	31.85	5.76	22.50	1942.87	244.05	3.45	0.26

Taula 5. Valors del test per a cada variable i per les diferents classes

	Q	DQO-E(F)	DQO-P(F)	DQO-S	MES-E	MES-P	MES-S	NO3-S	P-E	P-S	MLSS-1	IVF-1	CM-1	T-2	MLSS-2	IVF-2	EF-2	CM-2
1.00	-2.30	2.52	2.15	0.97	0.21	1.43	-1.61	-1.15	0.46	-1.36	2.60	2.07	-2.31	-1.14	3.27	-4.69	4.20	-1.43
2.00	-4.08	1.34	-6.63	-0.89	3.24	5.61	1.87	-0.95	0.97	-1.21	6.11	3.83	-2.18	-3.87	7.27	-3.96	-2.77	-2.14
3.00	0.38	0.58	-2.04	0.45	0.02	0.23	-0.40	1.11	-0.54	6.25	1.11	5.69	-1.47	-3.64	0.63	-1.28	-2.18	0.15
4.00	0.40	0.00	-0.62	6.05	0.63	5.20	7.80	0.00	0.00	0.00	1.73	0.37	0.00	-1.71	2.89	-1.10	-1.40	-1.43
5.00	0.34	0.10	3.20	2.84	1.34	0.91	1.89	-1.98	1.87	1.37	2.57	1.71	-2.63	2.10	1.01	0.91	-2.53	-1.56
6.00	-2.21	0.07	0.16	-0.50	3.03	-1.79	0.52	-0.31	-0.14	0.49	3.71	2.09	-1.30	-0.69	-0.54	5.18	-1.47	-0.95
7.00	2.08	-3.10	-1.09	-2.73	-3.67	-5.08	-2.58	0.00	-0.55	-0.19	-5.98	-3.47	0.10	-0.35	-5.69	-0.21	-2.07	-0.27
8.00	1.78	-1.17	-0.85	-1.44	-1.05	-0.60	-1.61	8.90	-2.88	0.66	-1.68	-1.11	-1.05	2.12	-0.11	-1.03	0.10	0.25
8.00	4.52	-0.82	0.06	-3.70	-1.81	-1.88	-2.47	-0.40	-3.21	-2.04	-3.40	-3.37	6.52	-1.96	-3.30	-0.94	2.62	3.59
10.00	2.39	0.33	2.54	1.49	-0.96	-0.57	-0.40	-0.46	3.46	0.00	-2.92	-2.75	7.30	2.72	-2.27	1.64	1.45	7.70
11.00	-0.80	-0.55	1.47	0.33	-0.75	-0.71	0.56	0.00	-0.82	-0.41	-3.96	-3.99	1.11	2.18	-3.09	7.05	2.33	-0.11
12.00	2.29	-0.09	3.40	1.15	-1.27	-1.39	0.58	0.00	0.00	0.00	-2.58	-2.03	0.00	6.09	-1.85	0.86	1.30	-0.10