© *Acta hydrotechnica* **20/33 (2002)**, Ljubljana
ISSN 0352-3551

UDK / *UDC*: 004.8:628.32
Izvirni znanstveni prispevek – *Original scientific paper*

Prejeto / *Received*: 17.7.2002
Sprejeto / *Accepted*: 28.12.2002

UPORABA ODLOČITVENIH DREVES PRI MODELIRANJU ČISTILNE NAPRAVE ZA ODPADNO VODO
THE USE OF DECISION TREES IN THE MODELLING OF A WASTEWATER TREATMENT PLANT

Nataša ATANASOVA, Boris KOMPARE

Čistilne naprave (ČN) za odpadno vodo so dinamični in kompleksni sistemi, katerih vodenje lahko izboljšamo z različnimi pristopi k modeliranju in napovedovanju delovanja ČN. V nalogi smo poskušali zgraditi uporabne modele za napoved delovanja čistilne naprave z orodji strojnega učenja, točneje z odločitvenimi drevesi. Podatkovna baza, iz katere smo modele gradili, je sestavljena iz enodnevnih povprečnih merjenih podatkov na ČN. Poleg kvantitativnih podatkov je baza sestavljena iz številnih kvalitativnih ocen, kakor tudi iz obsežne mikrobiološke analize. Dosedanja obdelava podatkov je obsegala klasifikacijo podatkov z Linneo+ postopkom. Temu smo dodali izgradnjo preprostih, a dovolj natančnih modelov, ki predvidevajo funkcionalno stanje ČN na podlagi merjenih (kvantitativnih) vhodnih podatkov. Za izgradnjo modelov smo uporabili programski paket WEKA, ki ima vgrajeno večino popularnih algoritmov strojnega učenja.
Ključne besede: *odpadna voda, čistilna naprava, modeliranje, strojno učenje, odločitvena drevesa*

Wastewater treatment plants (WWTP) are dynamic and complex systems, the management of which can be improved by different approaches to modeling and predicting their operation. Machine learning tools (decision trees) were used to build useful prediction models for wastewater treatment plant operation. The data base used for building the models is composed of measured quantitative as well as qualitative data on the WWTP. We were also provided with a microbiological analysis. The data are presented as a one-day situation of the plant operation. So far, classification of the data was made using the Linneo+ methodology. We extended the knowledge gained by classification by analyzing the classified data and constructing useful models that predict WWTP operation from inflow data. The WEKA program package, which includes most of the popular machine learning algorithms, was used for constructing the models.
Key words: *wastewater, wastewater treatment plant, modeling, machine learning, decision trees*

1. UVOD

Biološko čiščenje odpadne vode predstavlja vrsto procesov, s katerimi se odstranjujejo nezaželene snovi iz vode – organsko onesnaženje, dušik, fosfor. Procese izvajajo različne vrste mikroorganizmov, ki za svoj metabolizem (razvoj) uporabljajo onesnaženje v vodi. Za uspešen potek čiščenja mora biti zagotovljena ustrezna koncentracija mikro-organizmov v bioreaktorju ter ustrezni pogoji, v katerih lahko delujejo.

1. INTRODUCTION

Biological treatment of wastewater is composed of numerous complex processes by which organic matter, nitrogen and phosphorus are removed from water. The processes are carried out by many different microorganisms, using the compounds to be removed for their metabolism. Adequate concentration and conditions for growth of microorganisms must be achieved in the bioreactor for successful cleaning of wastewater.

Izredna kompleksnost in občutljivost procesov čiščenja vode otežuje optimalno dimenzioniranje in vodenje čistilne naprave. Modeliranje postaja nepogrešljivo orodje za simulacijo in vodenje čistilne naprave. V svetu se večinoma uporabljajo matematični modeli. Toda dejavnike, kot so spremenljiva sestava odpadne vode, ki doteka na čistilno napravo, številne vrste mikroorganizmov, ki sodelujejo pri procesu čiščenja, ter spremenljivi pogoji delovanja čistilne naprave, je tako rekoč nemogoče zajeti v matematične enačbe brez večjih poenostavitev in številnih predpostavk. Tako imamo ali preveč sofisticirane modele, s številnimi parametri, ali pa preproste modele, ki pa imajo omejeno uporabnost.

Po drugi strani so tu orodja strojnega učenja, s katerimi lahko dobimo uporabne preproste konceptualne modele (Kompare, 1995) na podlagi merjenih podatkov. Ti modeli imajo svojo največjo uporabnost pri odkrivanju novega znanja med podatki ter simulaciji (napovedi) novih situacij. Pri strojnem učenju moramo ločiti med pridobivanjem novega znanja iz podatkov in izdelavo uporabnih modelov. V prvem primeru nam algoritem sam ponuja soodvisnosti med podatki in pravzaprav ne potrebujemo znanja s področja uporabe (pridobivamo novo znanje). Pri izdelavi uporabnih modelov pa moramo vedeti, kaj želimo dobiti. Skrbno moramo izbirati atribute, iz katerih bo model zgrajen, ter paziti, da bo še dovolj natančno napovedoval.

V literaturi smo zasledili navedbe uporabe nekaterih tehnik strojnega učenja na merskih podatkih na ČN za odpadno vodo. Dosedanja analiza in obdelava teh podatkov je potekala v smislu klasifikacije podatkov (Sanchez et al., 1997; Comas et al., 1999) in odkrivanja novega znanja in soodvisnosti med podatki z različnimi metodami umetne inteligence (Roda et al., 1998; Comas et al., 2001, Belanche et. al., 1999). Klasifikacija podatkov je bila narejena s programom Linneo+, kateri so avtorji potem kot področni strokovnjaki dopisali funkcionalna stanja, tj. opis stanja, v katerem deluje ČN. Zgrajen je bil model, ki povezuje tako vhodne kot izhodne spremenljivke s posamezno klasifikacijo.

Due to the complexity and sensibility of the cleaning processes it is difficult to maintain optimal operational conditions in a WWTP. Therefore modelling is becoming a very useful and commonly used tool for simulation and control of a WWTP. Usually mathematical models are used for modelling of a WWTP. Since we are dealing with a dynamic and complex process, the equations in every mathematical model are more or less simplified, meaning that they contain more or less parameters which need to be estimated. The complexity of the mathematical models can vary. Thus, we have either too sophisticated models with many parameters, difficult to be estimated, or else simple models with limited use.

On the other hand, machine learning tools offer simple conceptual models (Kompare, 1995) built from a data set. This approach is most suitable for discovering new knowledge among measured data or predicting situations in a specific domain. A distinction should be made between knowledge acquisition and constructing prediction models. In the first case, the learning algorithm discovers connections and dependencies among measured data and we do not really need specific knowledge of the domain. We are gaining new knowledge. If we want to construct useful prediction models then we need to carry out the measurements in a proper way and to make a choice of relevant data in the domain, so that the models are built from suitable attributes and predict the situations with satisfactory accuracy.

The WWTP data set used in this paper was found in the literature. So far analysis of the data set includes classification of data (Sanchez et al., 1997; Comas et al. 1999), and knowledge discovery with different machine learning tools (Roda et al., 1998; Comas et al., 2001; Belanche et. al., 1999). Classification of data, by which several clusters were obtained, was made with Linneo+. Domain experts then characterised and identified each cluster with a different state or situation on the WWTP. New knowledge was discovered by building a decision tree, which is composed of quantitative and qualitative data. The model predicts situations obtained on the WWTP

Model odkriva zakonitosti, kot je na primer povezava pojava napihnjenega blata z ustreznimi mikroorganizmi, vendar pa tak način ne omogoča napovedovanja delovanja ČN.

Glavni namen naše naloge je z uporabo drugih (naprednih) tehnik strojnega učenja izluščiti iz obstoječih podatkov še več znanja, kot je bilo to storjeno do sedaj. Cilj naloge je torej napovedati predhodno določena stanja ČN (npr. plavajoče blato) na podlagi meritev na vtoku ČN. Tak model bi omogočal upravljalcem hitrejše odkrivanje težav pri delovanju ČN in lažje odločanje pri vodenju ČN.

Poudarjamo, da so podatki pridobljeni preko osebne komunikacije z avtorji prej navedenih člankov, v obliki, kot so jo uporabljali oni. Vse ostale obdelave podatkov so navedene v tem članku.

2. ČISTILNA NAPRAVA

Glavni elementi biološke čistilne naprave z aktivnim blatom so: primarni usedalnik, biološki reaktor in naknadni ali sekundarni usedalnik. Surova odpadna voda doteka v primarni usedalnik 1, kjer poteka mehansko čiščenje tj. odstranjevanje usedljivih organskih in anorganskih primesi. Nato gre voda v biološki reaktor, kjer mikroorganizmi razgrajujejo preostalo (neusedljivo) organsko onesnaženje in ga uporabljajo za svojo rast. Tako nastalo mešanico vode in biološkega blata vodimo nato iz reaktorja v sekundarni usedalnik, kjer se biološko blato prične usedati, preostala vodna masa pa odteka naprej v recipient. Usedlo biološko blato večinoma recirkuliramo nazaj v reaktor in s tem vzdržujemo zadostno koncentracijo biološkega blata za učinkovit potek čiščenja v reaktorju. V primeru, da je dotok surove vode večji, kot ga prenese čistilna naprava, se višek odpadne vode razbremenjuje po razbremenilnem kanalu preko primarnega usedalnika (SP3) v recipient. Ta del odpadne vode ne vpliva na učinek čiščenja biološkega dela ČN. Obremenitev obravnavane ČN je značilna za turistična mesta, saj je pozimi (30 000 populacijskih enot) bistveno manj obremenjena kot poleti (150 000 PE). Tehnološka shema obravnavane ČN je prikazana na sliki 1.

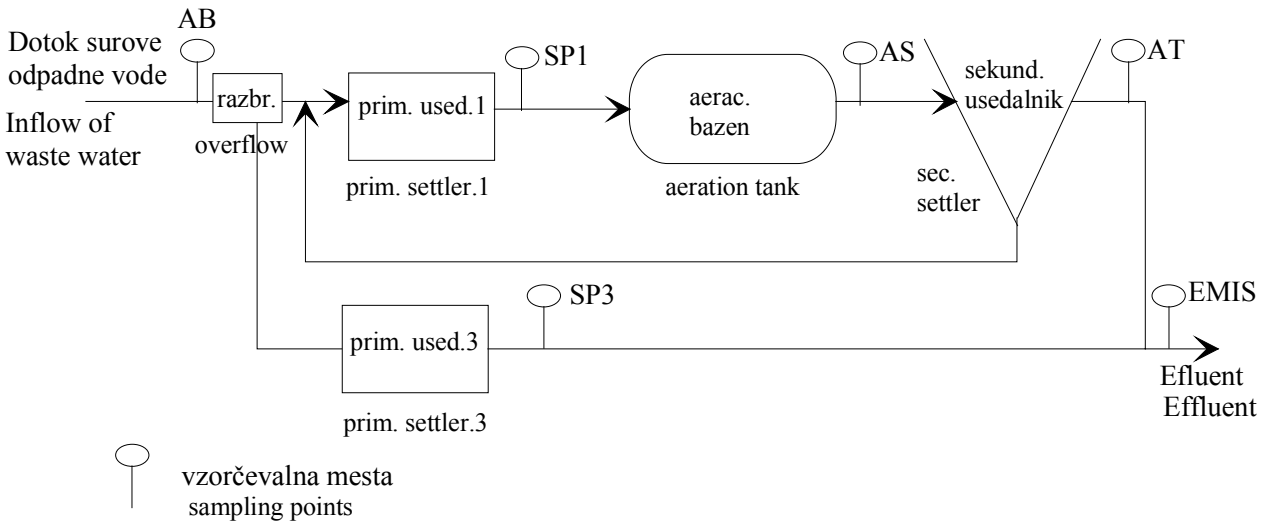
using all data (quantitative, qualitative and estimations). However such a model cannot be used for predictions, i.e. for managing of the WWTP.

The main goal of this paper is, by using different (advanced) ML techniques, to gain some more knowledge from given data, than has been gained so far. Thus, we wanted to predict the situations on the WWTP (i.e. bulking sludge) from measured data on the WWTP inflow. Such models can be of help in detecting the conditions when operational problems on the WWTP occur and in managing the WWTP.

At this point we wish to stress that the data were obtained by personal communications with the previously cited authors, and in the form as they were using them. All additional processing of the data is described in this article.

2. WWTP

An activated sludge WWTP contains three main units: primary settler, aeration tank and secondary settler or clarifier. The inflow of raw wastewater is in primary settler 1, where mechanical treatment is taking place. Water is then transported to the aeration tank, where the microorganisms carry out the biological treatment. The produced mixture of water and activated sludge is finally sent to the secondary settler, where the flocks of microorganisms settle down and clean water is discharged to the recipient. The settled sludge is recycled back to the aeration tank to maintain a satisfactory concentration of activated sludge for a successful treatment process in the reactor. If the inflow to the WWTP exceeds a certain value then the bypass is activated. Excess water is led to primary settler 3 and then discharged to the recipient without any biological treatment. This part of the wastewater does not affect biological treatment of the WWTP. The load of the WWTP under study is typical for tourist resorts. It is much lower during the winter (30 000 PE) and then in summer (150 000 PE). The configuration of the WWTP is presented in Figure 1.



Slika 1. Tehnološka shema čistilne naprave (ČN).
Figure 1. Scheme of the water line of WWTP.

3. OPIS PODATKOVNE BAZE

3.1 MERJENI PODATKI (ATRIBUTI)

Visoka stopnja avtomatizacije ČN omogoča zapisovanje in shranjevanje večine podatkov, ki so značilni za delovanje ČN. Merjeni so kvantitativni (preglednica 1) in kvalitativni (preglednica 2) podatki. Kvantitativni podatki vključujejo pretoke, analize odpadne vode in biološkega blata na različnih vzorčevalnih mestih (AB, SP1, AS, AT, SP3), medtem ko kvalitativni podatki vključujejo ocene samega delovanja čistilne naprave ter obsežno mikroskopsko analizo, iz katere so razvidne vrste mikroorganizmov, ki sodelujejo v procesu.

Podatki v bazi so podani za en dan oz. en zapis v podatkovni bazi prikazuje enodnevno situacijo delovanja čistilne naprave. Uporabljena podatkovna baza je sestavljena iz 243 situacij (zapisov) in 63 atributov (merjenih kvantitativnih in kvalitativnih podatkov na vzorčevalnih mestih).

3. DATA BASE

3.1 MEASURED DATA

The WWTP has a high level of automation that collects all the data measured on the plant. Thus, we are dealing with a large amount of historical data typical for a WWTP operation. The data set is composed of quantitative data (Table 1), such as flow rate, wastewater and sludge analysis at different sampling points (AB, SP1, AS, AT, SP3) and qualitative data (Table 2), including estimations of WWTP operation and sludge quality as well as large microscopic analysis revealing microorganism species involved in the treatment process.

The data are presented as average values during one day, i.e. one record in the data set represents a one-day situation of the WWTP operation. There are 243 situations and 63 attributes (measured quantitative and qualitative data at different sampling points, listed in Table 1 and Table 2) in the database.

Preglednica 1. Kvantitativno merjeni podatki.
Table 1. Quantitative data.

Oznaka atributa Attribute	Opis *Description*	Enota *Unit*	Vzorčevalno mesto *Sampling point*
Q	pretok *flow rate*	m3/dan m3/day	AB, SP1, SP3
DQO	kemijska potreba po kisiku *chemical oxygen demand*	mg/l	AB, SP1, SP3, AT, EMIS
DBO5	biokemijska potreba po kisiku *biochemical oxygen demand*	mg/l	AB, SP1, SP3, AT, EMIS
SST	suspendirani delci *total suspended solids*	mg/l	AB, SP1, SP3, AT, EMIS
NKT	amonijak in organski dušik *total kjeldhal nitrogen*	mg/l	AB, AT
N-NH4	amonijak *ammonium*	mg/l	AB, AT
NO2	nitrit *nitrite*	mg/l	AT
NO3	nitrat *nitrate*	mg/l	AT
PH	PH	$\log[H^+]$	AB, AT
COND	prevodnost *conductivity*	µ Siemens/cm	AB, AT
TERB	motnost (NTU) *turbidity*	NTU	AB, AT
CL	slanost *chlorine*	mg/l	AB, AT
V30	volumen blata po 30min usedanju *sludge volume after 30min settling*	ml	AT
LM-B	masa vseh delcev biološkega blata *mixed liquor suspended solids*	g/l	AT
MLVSS	masa organskega dela biološkega blata *mixed liquor volatile suspended solids*	g/l	AT
IVF	indeks blata *sludge volume index*	ml/g	AT
SST-R	recirkulacija blata *recirculation*	m^3/dan m^3/day	AT
CM	obremenitev blata *food to microorganism ratio*	DBO5/(MLSS*dan) DBO5/(MLSS*day)	AT
TRC	starost blata *sludge retention time*	dan *day*	AT

Preglednica 2. Kvalitativno ocenjeni podatki.
Table 2. Qualitative data.

Atribut *Attribute*	Opis *Description*
ESC-B	pene v aeracijskem bazenu *presence of foam at the aeration tank*
FLOC	prisotnost flokul v sekundarnem usedalniku *presence of activated sludge flocs at the secondary settler*
ASP-AT	ocena kvaliteta efluenta qua*lity of the final treated water*
ASP-FLOC	ocena flokul *microscopic appearance of the activated sludge floc*
ZOO	Zooglea
NOC	Nocardia
A21N/THIO	021N/Thiothrix
A41	A41
MICO	Microthrix parvicella
FILAM-DOMI	dominantna vrsta nitastih bakterij *dominant filamentous bacteria*
NFILAM	število vrst nitastih bakterij *number of different filamentous bacteria*
ASPID	Aspidisca
EUPL	Euplotes
VORT	Vorticella
EPIST	Epystylis
OPER	Opercularia
CIL-CAR	Carnivorous ciliates
GPROTO-DOMI	Dominantna vrsta praživali *dominant protozoa-group*
BIODIV-MIC	št. različnih mikroorganizmov *biodiversity of the microfauna*
G-FLAG	bičkarji > 20µm *flagellates* > 20µm
P-FLAG	bičkarji < 20µm *flagellates* < 20µm
AMEB	Amoebae
AMEB-TECA	Testate Amoebae
ROTIFERS	Rotifers

3.2 RAZREDI PODATKOV

Vsaka situacija v obravnavani bazi merjenih podatkov ima določen razred. Razredi so bili določeni s programom Linneo+ (Sanchez et al., 1997; Comas et al., 1999), ki uporablja koncept razdalje med objekti (zapisi v podatkovni bazi). Klasifikacija je potekala v dveh korakih: (1) Določanje skupin podobnih primerov (zapisov v bazi) na podlagi razdalje in (2) Ovrednotenje dobljenih skupin s strani strokovnakov. Dobljene razrede so analizirali,

3.2 CLASSIFICATION OF DATA

Linneo+ has been used for classification of the data (Sanchez et al., 1997; Comas et al., 1999). It is an unsupervised learning method that determines clusters of the data. The methodology implemented can be considered as a two-step task: (1) Clustering, which determines useful subsets of data using the conventional concept of distance between objects, and (2) Validation of the subsets by experts. The obtained subsets are accepted or

modificirali in interpretirali strokovnjaki. Vsaki skupini je bilo pripisano funkcionalno stanje, v katerem deluje ČN. Dobljeni razredi so prikazani v preglednici 3.

rejected by experts and identified with a different state or situation on the WWTP operation. The clusters are depicted in Table 3.

Preglednica 3. Razredi podatkov.
Table 3. Classes of data.

Opis razreda *Class description*	Oznaka *Mark*	Št. primerov *No. of examples*
Nornalno delovanje ČN pozimi *Normal WWTP-operation in winter days*	c1	81
Normalno delovanje ČN poleti *Normal WWTP-operation in summer days*	c2	55
Optimalno delovanje ČN poleti Summer *days with optimal WWTP-operation*	c11	24
Deževje *Rainy days*	c3	3
Nevihte *Storm days*	c4	3
Podobremenitev (obdobje po dežju) *Underloading (final period of a storm)*	c5	12
Nitrifikacija *Nitrification*	c7	2
Denitrifikacija v sekundarnem usedalniku (dviganje blata) *Denitrification in the secondary settler (rising)*	c13	7
Deflokulacija (motnost efluenta zaradi majhnih bičkarjev) *Deflocculation (effluent turbidity due to small-flagellates)*	c8	5
Plavajoče blato zaradi Thiotrix-a (vpliv na efluent) *Bulking sludge due to Thiotrix (affecting the effluent)*	c9	3
Začetek in konec plavajočega blata zaradi Thiotrix-a *Beginning and end of a strong bulking-sludge episode due to Thiotrix*	c14	2
Peneče blato, ki ga povzroča Mcrothrix (normalna diverziteta mikrofavne) *Foaming sludge due to Microthrix (with normal microfauna biodiversity)*	c10	17
Peneče blato, ki ga povzroča Mcrothrix (nizka diverziteta mikrofavne in dominairajo bičkarji < 20 μm) *Foaming sludge due to Microthrix (with very low microfauna biodiversity and dominant < 20 μm flagellates)*	c18	1
Peneče blato, ki ga povzroča Mcrothrix (in plavajoče blato zaradi Zooglee) *Foaming sludge due to Microthrix (and viscous bulking due to Zooglea)*	c19	6
Peneče blato, ki ga povzroča Nocardia, ki vpliva na proces *Foaming sludge due to Nocardia, affecting the process*	c15	4
Peneče blato, ki ga povzroča Nocardia, ki vpliva na proces (odplavljanje delcev) *Severe foaming sludge due to Nocardia, affecting the process (loss of solids)*	c16	5
Peneče blato, ki ga povzroča Nocardia in zastopanje bičkarjev < 20 μm *Foaming sludge due to Nocardia with abundant < 20 μm flagellates*	c17	8
Sprememba konfiguracije ČN iz zimskega v letno obdobje *Winter-summer Plant configuration change*	c20	3
Preobremenitev *Overloading*	c6	1
Klorni šok *Chlorine shock*	c12	1

4. UPORABLJENE METODE

Odločitvena drevesa so ena izmed učnih shem (opisov) danega pojma ali koncepta, ki jo generira učni algoritem iz danih primerov tega pojma. Primere podajamo v preglednici tako, da je en primer (ena vrstica v preglednici) sestavljen iz atributov t.j. neodvisnih spremenljivk in razreda primera t.j. odvisnih spremenljivk. Razred predstavlja pojem, ki se ga algoritem nauči napovedovati. Za ponazoritev si poglejmo preprost primer. Koncept, ki se ga želimo naučiti, je uporaba *prevoznega sredstva*. Vprašamo se, kdaj (v kakšnih vremenskih pogojih) uporabiti kolo in kdaj avto. V preglednici 4 je podano 9 primerov, iz katerih se bo algoritem učil koncepta. Primere sestavljajo naslednji opisi (atributi): sončno, temperatura in dež. Zadnja kolona v preglednici (prevozno sredstvo) predstavlja razred primera.

4. METHODS

Decision trees are one of the learning schemes of a given concept, generated by a learning algorithm from given instances of that concept. The instances are presented in a table, each row being an instance and the column being a description i.e. an attribute of the instance or an independent variable. One of the columns contains the values of the instances class (dependant variable), which is the concept to be learned by the algorithm. The following example will illustrate the methodology. Let the *means of transportation* be the concept to be learned, i.e. when (under what circumstances) to use a bike and when to use a car as a means of transportation. There are nine examples, depicted in Table 4, that will be used to learn the concept. The examples have the following descriptors (attributes): sunny, temperature and rain. The last column (means of transportation) contains the values of each examples class.

Preglednica 4. Preprost primer učenja koncepta.
Table 4. Simple example of concept learning.

Primer *Example*	Sonce *Sunny*	Temperatura *Temperature*	Dež *Rainy*	prevozno sredstvo *means of transportation*
1	Da *Yes*	Visoka *High*	Ne *No*	Kolo *Bike*
2	Da *Yes*	Srednja *Med*	Ne *No*	Kolo *Bike*
3	Da *Yes*	Nizka *Low*	Ne *No*	Avto *Car*
4	Ne *No*	Visoka *High*	Da *Yes*	Avto *Car*
5	Ne *No*	Srednja *Med*	Da *Yes*	Avto *Car*
6	Ne *No*	Nizka *Low*	Da *Yes*	Avto *Car*
7	Ne *No*	Visoka *High*	Ne *No*	Kolo *Bike*
8	Ne *No*	Srednja *Med*	Ne *No*	Kolo *Bike*
9	Ne *No*	Nizka *Low*	Ne *No*	Avto *Car*

V obravnavani podatkovni bazi je primer podan za posamezen dan delovanja ČN, atributi so merjene in ocenjene količine na vzorčevalnih mestih ČN, razred pa je opis delovanja ČN, ki je odvisen od meritev na ČN, t.j. podanih atributov za vsak primer.

Najpogosteje uporabljen algoritem za generacijo odločitvenih dreves je Top-Down Induction of Decision Trees (Quinlan, 1986; Bratko et.al., 1999), ki se glasi:

1. Podani niz primerov (S) se razdeli na podnize (S_i) glede na izbrani najboljši atribut tako, da imajo vsi primeri v podnizu isto vrednost izbranega atributa. V zgornjem primeru atribut *dež* razdeli celotni niz na dva podniza: prvi niz so tisti primeri, ki imajo vrednost tega atributa DA, drugi pa tisti, ki imajo vrednost atributa NE. Vsak podniz predstavlja vozlišče v drevesu. Najboljši atribut se izbere avtomatsko z računanjem entropije in pridobitvijo informacije.
2. Vsak podniz predstavlja vozlišče drevesa. Če so vsi primeri v podnizu S_i istega razreda, potem se ustvari list drevesa s tem razredom, sicer pa se postopek ponovi za $S = S_i$.

Če algoritem apliciramo na podani primer, dobimo odločitveno drevo na sliki 2. Drevo ima dve vozlišči (*dež* in *temperatura*) in štiri liste (kvadrati). Na vrhu drevesa je atribut *dež*. Če je vrednost enaka DA, uporabljamo avto, sicer pa se vprašamo, kakšna je *temperatura*. Če je VISOKA oz. SREDNJA, se peljemo s kolesom, če je NIZKA, pa z avtom. Glede na podane primere, se je algoritem naučil, da *sonce* ne igra nobene vloge pri uporabi prevoznega sredstva in ga zato ni vstavil v odločitveno drevo.

Algoritem deluje za diskretne vrednosti atributov. Če imamo opravka z zveznimi vrednostmi, lahko uvedemo diskretizacijo zveznih atributov. Postavimo mejne vrednosti intervalov atributa, ki jih nato obravnavamo kot diskretne vrednosti. Recimo, da ima atribut temperatura namesto nominalnih realne vrednosti (preglednica 5). Potem bi odločitveno drevo izgledalo, kot kaže slika 3.

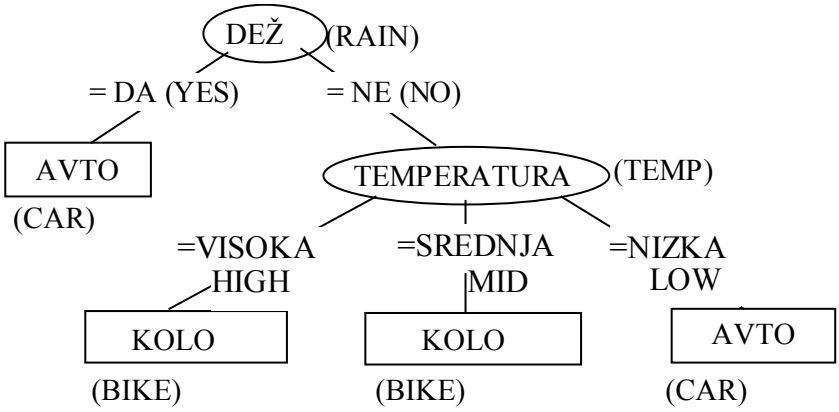
In the presented database an example is given for one day of WWTP operation, the attributes are measured and the estimated data on the sampling points and class provides a description of a WWTP operation for each day.

A commonly used algorithm for induction of decision trees is Top-Down Induction of Decision Trees (Quinlan, 1986; Bratko et.al., 1999):

1. A given set of examples (S) is divided into subsets (S_i) according to the best attribute, so that all examples in the subset have an equal value of the chosen attribute. For instance attribute *rain* in the presented example would divide the given set of nine examples into two subsets: the first one having all examples with the value of attribute *rain* YES, and second one containing all examples that have the value NO standing for the attribute *rain*. The best attribute is chosen by calculating the entropy and information gained.
2. Each subset represents a node in the tree and if all examples in a subset have same class, then a leaf is created, otherwise the procedure is repeated for $S = S_i$.

When applied on our example the algorithm constructs the decision tree shown in Figure 2. The tree has two nodes (*rain* and *temperature*) and four leaves (ends of the branches). The attribute *rain* is on top of the tree. If the value of *rain* is YES, then we use a car, or else we ask about the *temperature*. If it is HIGH or MED then we ride a bike and if LOW we use a car. According to the given examples the algorithm has learned that the attribute *sunny* is not important for the concept and therefore can not be found in the tree.

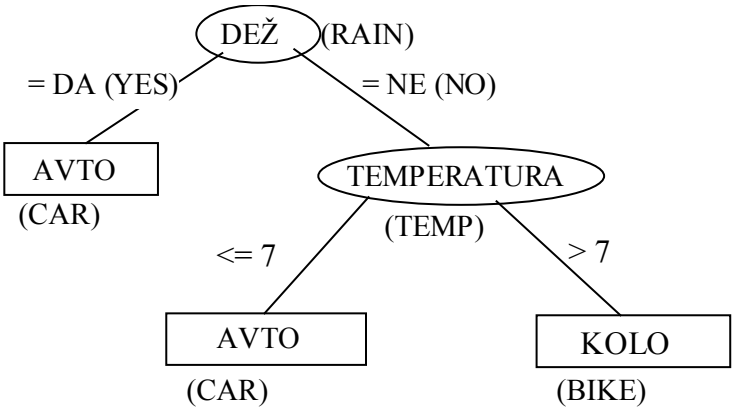
This algorithm works for discrete attributes. If we are dealing with continuous attributes, then binarization of the numeric attributes can be applied i.e. thresholding their ranges into pairs of subintervals, that are treated as symbols. Suppose that attribute temperature has real (continuous) values instead of nominal (Table 5). Then the decision tree would look as shown in Figure 3.



Slika 2. Odločitveno drevo preprostega primera.
Figure 2. Decision tree of a simple example.

Preglednica 5. Primer z zveznim atributom.
Table 5. Example with continuous attribute.

Primer *Example*	Sonce *Sunny*	Temperatura *Temperature*	Dež *Rainy*	Prevozno sredstvo *Transportation means*
1	Da *Yes*	25	Ne *No*	Kolo *Bike*
2	Da *Yes*	15	Ne *No*	Kolo *Bike*
3	Da *Yes*	5	Ne *No*	Avto *Car*
4	Ne *No*	22	Da *Yes*	Avto *Car*
5	Ne *No*	10	Da *Yes*	Avto *Car*
6	Ne *No*	7	Da *Yes*	Avto *Car*
7	Ne *No*	30	Ne *No*	Kolo *Bike*
8	Ne *No*	14	Ne *No*	Kolo *Bike*
9	Ne *No*	2	Ne *No*	Avto *Car*



Slika 3. Odločitveno drevo, zgrajeno iz nominalnih in numeričnih atributov.
Figure 3. Decision tree induced from nominal and numeric attributes.

5. IZVEDENE ANALIZE (EKSPERIMENTI)

Ločiti moramo med pridobivanjem novega znanja in izdelavo uporabnih modelov za napoved delovanja ČN. V prvem primeru lahko izberemo poljubno spremenljivko za odvisno, t.j. za razred primera, ki jo napovedujemo na podlagi vseh podanih atributov. Z učnim algoritmom odkrivamo njihovo povezanost oz. odvisnost med sabo. V tem primeru se lahko zgodi, da odkrivamo trivialne soodvisnosti.

Pri izdelavi uporabnih modelov pa mora biti razred, ki ga napovedujemo, odvisna spremenljivka, atributi, iz katerih je model zgrajen, pa striktno neodvisne spremenljivke, da ne bi dobili trivialnih povezav. V našem primeru želimo kvalitativne (ocenjene) podatke napovedati iz merjenih (kvantitativnih), ki so na voljo.

5.1 SELEKCIJA POMEMBNEJŠIH ATRIBUTOV

Modeli, ki jih želimo zgraditi, naj bi dali odgovore na vprašanja, kot so: Kaj pomeni določen dotok na ČN? Kaj se zgodi, če se obremenitev poveča ali pomanjša? Kako vplivajo karakteristike biološkega blata na proces čiščenja? itd. Za konstrukcijo takih modelov je treba najprej izbrati ustrezne atribute, iz katerih bo model zgrajen. To so predvsem kvantitativno merjeni podatki in so

5. EXPERIMENTS

A distinction should be made between knowledge discovery from data and building prediction models on WWTP operation. In the first case we can pick any attribute to be a dependent variable, i.e. the class which is predicted from all given attributes. The learning algorithm discovers connections and dependencies between data, where also trivial dependencies can be found.

If we want to construct useful prediction models then the class which is predicted has to be a dependent variable, while the attributes from which the model is built need to be independent variables. In our case, the models should predict qualitative data from the measured (quantitative) data that are available.

5.1 SELECTION OF RELEVANT ATTRIBUTES

Models to be constructed should give answers to questions like: What is the influence of a specific inflow on WWTP operation? What happens if the load of WWTP increases or decreases? What is the influence of sludge characteristics on the treatment process? and so on. The first step in building such models is selection of relevant attributes (Table 6), from which the model is built. These are mainly quantitative data, measured

upravljalcem na voljo oz. jih lahko predvidijo za določena obdobja, npr. dotok na ČN poleti ali pozimi. Izbrani atributi so prikazani v preglednici 6.

5.2 REDUKCIJA OBSTOJEČIH RAZREDOV

Z razvrščanjem podatkov v skupine, upoštevajoč vseh 63 atributov, je bilo v izvirni podatkovni bazi določeno 20 razredov (Sanchez et al., 1997; Comas et al., 1999). Glede na številčnost primerov, ki spadajo v nek razred, vidimo, da od dvajsetih izstopa le nekaj razredov: c1 (81 primerov), c2 (55 primerov), c11 (24 primerov), c10 (17 primerov). Če pa želimo da se algoritem nauči nekega razreda (pojma), mora imeti dovolj primerov (opisov) tega razreda, na podlagi katerih se uči. Odločitveno drevo, ki napovedujejo vse razrede, ne glede na številčnost primerov, odkriva soodvisnosti med atributi in razredi, vendar pa je precej veliko in nepregledno (Comas et al., 2001). Redukcijo smo naredili tako, da smo združili vse razrede, ki opisujejo probleme z blatom iz različnih razlogov (c9, c10, c14 do c19) v en razred cb (preglednica 7).

on regular bases and usually available to the manager. These data can also be predicted in certain periods of the year, e.g. inflow to the WWTP in summer or winter.

5.2 REDUCTION OF CLASSES

There were 20 clusters determined using the linneo+ classification (Sanchez et al., 1997; Comas et al., 1999). But according to the number of instances belonging to a specific cluster only a few clusters can be depicted: c1 (81 examples), c2 (55 examples), c11 (24 examples), c10 (17 examples). These clusters can be easily learned by the algorithm, since they have enough descriptions. If we want to predict all clusters (also those which have only one example) then we expect the decision tree to be big and not understandable (Comas et al., 2001). Therefore in our analysis we reduced the number of clusters by joining all clusters that describe sludge problems for different reasons (c9, c10, c14 to c19) into one cluster cb (Table 7). Now we have 46 examples belonging to this class and we expect the algorithm to learn to predict this class properly.

Preglednica 6. Izbira pomembnejših atributov.
Table 6. Choice of relevant attributes.

Vzorčevalno mesto *Sampling point*	Atributi *Attribute*
AB	Q, DQO, DBO5, SST, NKT, N-NH4, PH, COND, TERB, CL
SP1	Q, DQO, DBO5, SST
SP3	Q, DQO, DBO5, SST
PODATKI O BIOLOŠKEM BLATU SLUDGE DATA	V30, LM-B, IVF, MLVSS, SST-R, CM, TRC

Preglednica 7. Reducirani razredi (z 20 na 13).
Table 7. Reduction of classes (from 20 to 13).

Opis razreda *Class description*	Oznaka *Class*	Št. Primerov *Number of examples*
Nornalno delovanje ČN pozimi *Normal WWTP-operation in winter days*	c1	81
Normalno delovanje ČN poleti N*ormal WWTP-operation in summer days*	c2	55
Optimalno delovanje ČN poleti *Summer days with optimal WWTP-operation*	c11	24
Deževje *Rainy days*	c3	3
Nevihte *Storm days*	c4	3
Podobremenitev (obdobje po dežju) *Underloading (final period of a storm)*	c5	12
Nitrifikacija *Nitrification*	c7	2
Denitrifikacija v sekundarnem usedalniku (dviganje blata) *Denitrification in the secondary settler (rising)*	c13	7
Problemi z biološkim blatom *Sludge problems*	cb	46
Sprememba konfiguracije ČN iz zimskega v letno obdobje W*inter-summer plant configuration change*	c20	3
Preobremenitev *Overloading*	c6	1
klorni šok *Chlorine shock*	c12	1

5.3 UPOŠTEVANJE ČASOVNEGA SOSLEDJA PRI MERJENIH PODATKIH

Glede na navedbe o načinu vzorčevanja sklepamo, da so podatki v bazi podani kot povprečen enodnevni podatek. Prave informacije nimamo, vendar predpostavljamo, da so podatki v bazi podani zaporedno s korakom en dan oz. z enodnevnim zamikom. V izvirni podatkovni bazi je klasifikacija v razrede opravljena glede na podane atribute za isti dan. To pomeni, da vtočni parametri vplivajo na kakovost biološkega blata še isti dan, kar pa ne drži povsem. Če se kakšen dan pojavi problem plavajočega blata, je to posledica vtočnih parametrov merjenih pred enim ali dvema dnevoma. Zato smo zapise v podatkovni bazi spremenili tako, da vsak zapis vsebuje vrednosti posameznega atributa, ki so se pojavile pred enim oz. dvema dnevoma (preglednica 8). Oznaka ATT-n pomeni vrednost atributa ATT, ki se je pojavila pred n dnevi. Npr. DQO-1 pomeni vrednost atributa DQO pred enim dnevom.

5.3 INCORPORATING TIME DELAY ATTRIBUTES

Data in the database are presented as one-day average values of the measured attributes, and one record represents a one-day situation on the WWTP. We assume that the records are given with a one-day step and without too many missing days. Classes (situations) of the records are determined with respect to the attributes given for the same day, meaning that sludge quality is a result of the inflow from the same day. Based on the assumption that, in reality, problems with bulking sludge can not appear on the same day as they are presented in the data set, we incorporated a one- and two-day delay for sludge problems. Thus, we added new attributes which describe the obtained classes as a result of one or two days before inflow. The records are changed so that each record has information on values of the specific attribute that occurred one and two days before. In Table 8 are given the new attributes where the sign ATT–n denotes the value of attribute ATT that occurred n days ago, e.g. DQO-1 denotes the value of attribute DQO one day ago.

Preglednica 8. Upoštevanje časovnega zamika pri atributih.
Table 8. Incorporated time delay among measured attributes.

Vzorčevalno mesto *Sampling point*	Atributi *Attributes*
AB	Q, DQO, SST, COND, Q-1, DQO-1, SST-1, COND-1, Q-2, DQO-2, SST-2, COND-2
SP1	DQO, SST, DQO-1, SST-1, DQO-2, SST-2
SP3	Q, Q-1, Q-2
AT	DQO, SST, NH4, NO3, DQO-1, SST-1, NH4-1, NO3-1, DQO-2, SST-2, NH4-2, NO3-2
PODATKI O BIOLOŠKEM BLATU SLUDGE DATA	LM-B, IVF, SST-R, CM, TRC
OCENE ESTIMATIONS	ESC-B, FLOC, ASP-AT, ESC-B-1, FLOC-1, ASP-AT-1, ESC-B-2, FLOC-2, ASP-AT-2

6. REZULTATI

Za konstrukcijo modelov smo uporabili programski paket WEKA, ki vsebuje večino popularnih algoritmov strojnega učenja. Za indukcijo odločitvenih dreves je bil razvit algoritem J48.

Podatkovno bazo razdelimo na učno množico, na podlagi katere se model zgradi, ter testno množico, na kateri program ovrednoti natančnost modela. Napovedani razredi so v listih drevesa, kjer se vidi tudi natančnost napovedovanja posameznih razredov. Prva številka pomeni število primerov, uvrščenih v ta razred, druga pa število nepravilno uvrščenih primerov. Zgradili smo tri modele, ki imajo svojo uporabnost pri vodenju ČN, saj napovedujejo njeno delovanje na podlagi kvantitativnih atributov, merjenih na vhodu ČN.

6.1 M1 – NAPOVED VSEH RAZREDOV NA PODLAGI VHODNIH ATRIBUTOV

Pričujoči model naj bi napovedoval vseh 20 razredov. Vendar pa imajo nekateri razredi premalo primerov (opisov), da bi jih lahko napovedovali z razumljivo velikim drevesom. Zato smo z rezanjem zgradili manjše in razumljivejše drevo z natančnostjo 71.2 odstotka na testni množici podatkov, ki pa napoveduje le razrede, ki imajo več primerov: c5, c1, c10, c2 in c11 (slika 4). Model je zgrajen iz petih atributov Q-SP3, DQO-SP1, Q-AB, Q-SP1 in LM-B. Na vrhu drevesa se

6. RESULTS

We used the WEKA package to build the models. WEKA incorporates most of the popular machine learning algorithms that can be applied to a dataset and analyses its output to extract information about the data. Algorithm J48 was developed for induction of decision trees.

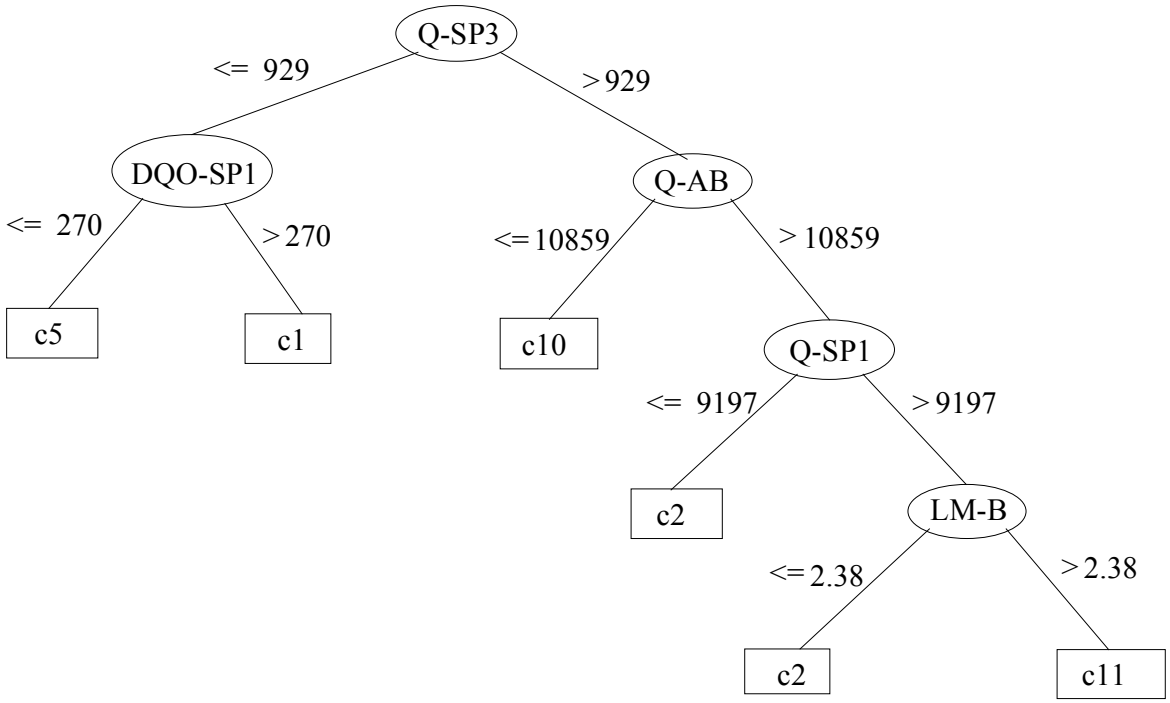
The data set is divided into a learning set used to build the model (scheme), and test data used to determine the model accuracy. In the leaves of the tree are the predicted classes. There are two numbers in each leaf. The first means the number of the examples going to the leaf and the second is for misclassified examples. We built three models, which can be of help in managing the WWTP, since they predict the WWTP operation according to the inflow data.

6.1 M1 – PREDICTING ALL CLASSES FROM INFLOW ATTRIBUTES

Model M1 is built to predict all 20 classes. Due to the small number of examples describing some of the classes, the tree that predicts all classes is not understandable. Therefore we used the pruning option and built an understandable and smaller tree with 71.2 % accuracy on test split, predicting only the classes that have more descriptions: c5, c1, c10, c2 and c11 (Figure 4). The model is built of five attributes: Q-SP3, DQO-SP1, Q-AB, Q-SP1 and LM-B. The model put the attribute Q-SP3, which describes the flow rate in the

nahaja atribut Q-SP3, ki opisuje pretok po razbremenilnem kanalu. Če je manjši od 929, model napove razred c1 (dobro delovanje pozimi) in razred c2 (dobro delovanje poleti) v veji, kjer je Q-SP3 < 929. Torej na podlagi vrednosti atributa Q-SP3 lahko sklepamo, ali gre za zimski (zunaj turistične sezone) oz. letni čas (v turistični sezoni) delovanja ČN.

bypass and on top of the tree. According to this attribute value we can decide if the WWTP is operating in winter or in summer (tourist season). We can expect class c1 (normal operation in winter) if Q-SP3 < 929 according to the model. Similarly we can find the class c2 (normal operation in summer) in the branch where Q-SP3 > 929.



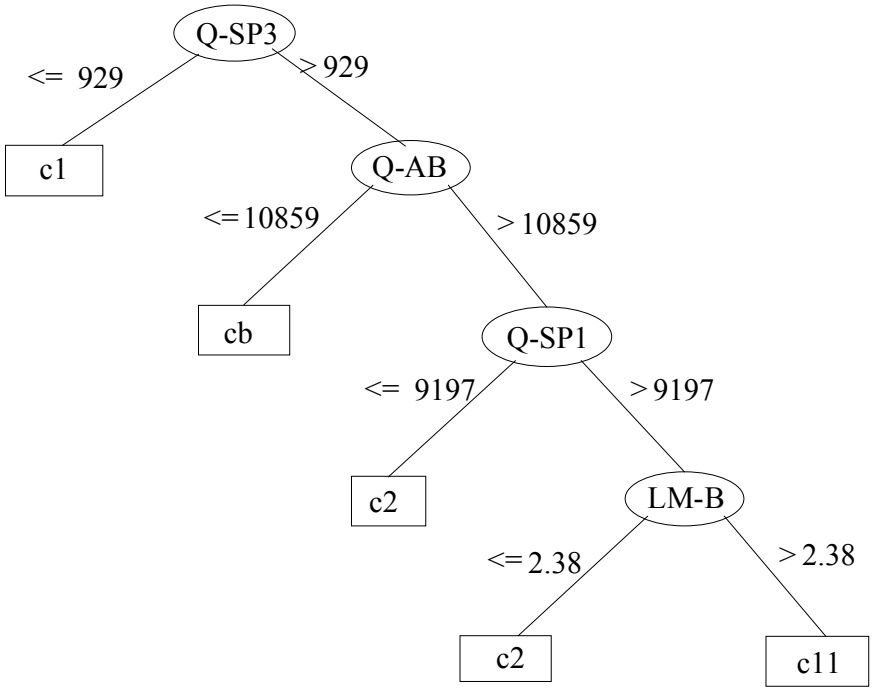
Slika 4. Odločitveno drevo modela M1.
Figure 4. Decision tree of model M1.

6.2 M2 – NAPOVED REDUCIRANIH RAZREDOV

Iz istega razloga kot pri modelu M1 smo tudi pri izgradnji modela M2 uporabili možnost rezanja. Orodje je avtomatsko našlo štiri najpomembnejše atribute, iz katerih je nato zgrajen model (Q-SP3, Q-AB, Q-SP1 in LM-B). Od podanih razredov iz preglednice 7 pa napoveduje naslednje: c1, cb, c2 in c11 (slika 5). Natančnost, ki jo dosega, znaša 73.9 odstotka na testni množici podatkov. Torej, dosega večjo natančnost od modela M1. Na vrhu drevesa se nahaja atribut Q-SP3, podobno kot prejšnji primer.

6.2 M2 – PREDICTION OF REDUCED CLASSES

For the same reason as with model M1 we used the pruning option to build model M2. The algorithm has found four important attributes of which the model is built (Q-SP3, Q-AB, Q-SP1 and LM-B). From the classes given to the model (Table 7) only c1, cb, c2 and c11 are predicted (Figure 5) with accuracy 73.9 % on test split. Thus, it is a little more accurate then M1. Similarly to M1 attribute, Q-SP3 is on top of the tree.



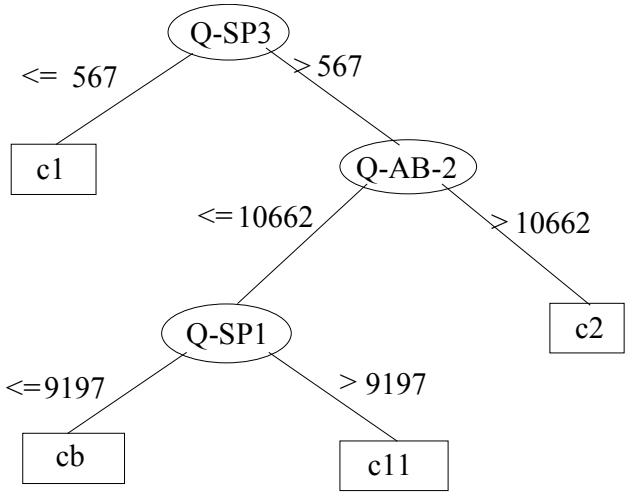
Slika 5. Odločitveno drevo modela M2.
Figure 5. Decision tree of model M2.

6.3 M3 - UPOŠTEVANJE ČASOVNEGA ZAMIKA

Model je zgrajen iz atributov, ki opisujejo časovno zakasnitev (preglednica 8) in napoveduje reducirane razrede z natančnostjo kar 75.3 odstotka. V modelu (slika 6) nastopajo le trije atributi (Q-SP3, Q-AB-2, Q-SP1) in štirje razredi (c1, cb, c11, c2).

6.3 M3 – INCORPORATING TIME DELAY

The model is predicting the following classes: c1, cb, c11, and c2, using the attributes that incorporate time delay (Table 8). Accuracy achieved is 75.3 % on test split. Only three attributes are needed for this prediction: Q-SP3, Q-AB-2 and Q-SP1 (Figure 6).



Slika 6. Odločitveno drevo modela M3.
Figure 6. Decision tree of model M3.

7. DISKUSIJA

Temeljni pogoj za konstrukcijo uporabnih in natančnih modelov na podlagi merjenih podatkov je dovolj velika, kakovostna in ustrezno izdelana podatkovna baza. Obravnavana ČN ima možnost zapisovanja velikega števila podatkov, vendar pa ima kar nekaj težav:

1. Velikost podatkovne baze je relativno majhna (243 zapisov) glede na število atributov (63) in število razredov (20).
2. Način merjenja atributov. V tem primeru je ena situacija delovanja ČN prirejena na enodnevni interval, kar pomeni, da se vpliva vtočnih parametrov na kakovost iztoka in kskovost blata ne vidi v celoti. To težavo smo poskušali rešiti z upoštevanjem ustreznega časovnega zamika med atributi, ki se vidi pri modelu M3. Model dosega večjo natančnost in prispeva k večji informiranosti.
3. Vprašljiva je tudi konsistentnost kvalitativnih podatkov. Atribut FLOC opisuje stanje flokul, vendar zavzema le problematične vrednosti, medtem ko opisa dobrega delovanja ne zasledimo. Domnevamo, da bi to lahko bile manjkajoče vrednosti tega atributa. Atribut torej nima pozitivnih vrednosti in ga zato ne moremo napovedovati, t.j. algoritem bi se učil pojma na nepopolnih primerih. Vprašljiv je tudi atribut ASP-FLOC, ki podaja oceno flokul b.b. Plavajoče blato (fangflotant) ima enkrat oceno 1, drugič pa oceno 2 ali več (preglednica 9). Nekonsistentnost se pojavlja tudi v povezavi z indeksom blata (IFV). Med podatki zasledimo za isti dan delovanja ČN normalno vrednost volumskega indeksa blata IFV (100 do 150) in hkrati plavajoče blato. Zato smo napovedovanje kvalitativnih atributov opustili.
4. Razredi, ki naj bi jih napovedovali, so preveč natančno določeni oz. jih je bilo preveč. Če bi želeli napovedovati vse razrede, bi bilo odločitveno drevo temu ustrezno veliko in nerazumljivo, hkrati pa preveč specifično (preveč prilagojeno učni množici). Zato smo z izbiro rezanja zgradili

7. DISCUSSION

Preparation of data is one of the most important, difficult and time-consuming steps in the knowledge acquisition process. Good quality of data is often the key to success in making predictions. The analysed data set had several problems:

1. The data set was quite small (243 records) in relation to the number of measured attributes (63) and classes (20).
2. Data preparation and measurements. Each situation on the WWTP consists of attributes average values during the day, which does not reveal the complete influence of the inflow data to outflow characteristics and on sludge quality. We tried to solve this problem by incorporating a time delay among the attributes. Model M3 has highest accuracy and contributes to better information
3. Consistency of qualitative data. Attribute FLOC describes the quality of flocks, but it only has negative values and no positive values i.e. *good operation*. There are a lot of missing values, so we assume that positive values are some of the missing ones. Thus the attribute has no positive values and therefore cannot be predicted since the algorithm would learn the concept on incomplete examples (descriptions). The value of this attribute should be in correlation with attribute ASP-FLOC, which is an estimation (mark from 1 to 4) of the flocs. But the same value of FLOC (i.e. fangflotant) corresponds ones to mark 1 and ones to mark 2 or more (Table 9). Attribute IFV (sludge volume index) is also inconsistent. The value of 100 to 150 corresponds to bulking sludge. For all these reasons we omitted prediction of the qualitative attributes.
4. According to the number of instances belonging to a specific cluster, the clusters were determined too accurately. If we want to predict all clusters (also those which have only one example) then we expect the decision tree to be big and not

manjše, bolj posplošene, hkrati pa razumljive in dovolj natančne modele, ki pa napovedujejo le tiste razrede, v katere sodi večje število primerov.

understandable. So in our analysis we used the pruning option and built smaller and understandable trees, that predict only those clusters with more examples.

Preglednica 9. Prikaz neustreznosti podatkov v bazi.
Table 9. Presentation of data inconsistency.

IVF	SSVLM-B	SST-R	CM	TRC	ESC-B	FLOC	ASP-AT	ASP-FLOC	WWTP
194	1.49	4.5	0.31	4.81	2	fangflotant	1	2	c8
186	1.07	3.21	0.41	8.23	1	?	2	?	c1
469	1.68	5.94	0.24	7.58	3	desfloculacio	2	2	c1
374	1.82	4.05	0.37	7.87	2	desnitrificacio	2	1	c1
?	?	?	?	?	2	?	2	?	c1
?	?	?	?	?	1	fangflotant	2	3	c1

Kljub naštetim težavam smo zgradili nekaj uporabnih modelov, iz katerih lahko razberemo kar nekaj zakonitosti in informacij pri delovanju obravnavane ČN. Najbolj očitno je, da se težave pojavljajo v turistični sezoni, ko imamo večji dotok na ČN in se zato aktivira razbremenilni kanal. Vsi modeli (M1, M2 in M3) postavljajo na vrhu drevesa atribut Q-SP3 (pretok v razbremenilnem kanalu). Dokler je manjši od določene vrednosti (926 m^3/dan glede na modela M1 in M2 oz. 526 m^3/dan po modelu M3), napovedujejo modeli razred c1 (dobro delovanje pozimi). V tej veji drevesa ne zasledimo težav z biomaso. Edino model M1 napoveduje podobremenitev (razred c5) v zimskem času in sicer če je DQO-SP1 (koncentracija KPK) < 270 mg/l (glej sliko 4).

Težave z blatom zasledimo v veji dreves, kjer je Q-SP3 večji od zgoraj omenjenih vrednosti, oz v času turistične sezone. Turistična sezona torej predstavlja prevelik obremenitveni šok (hidravlični), ki ga naprava rešuje z razbremenjevanjem (preko SP3). Istočasno je koncentracija organskega onesnaženja na ČN manjša zaradi večje porabe vode na prebivalca v poletnem času. To povzroča, da se nitaste bakterije v flokulah hitreje razvijejo (ker jim je hrana lažje dostopna kot ostalim) oz. napihnjeno blato.

Običajno so te posledice razvidne nekaj dni po nizko obremenjenem (biokemijsko) vtoku na ČN. Model, ki upošteva časovno zakasnitev, je zelo preprost in razumljiv,

In spite of all these difficulties we constructed three useful models that discover certain knowledge and information about WWTP operation. The most obvious information revealed from the models is that problems appear during the tourist season, when the inflow increases and bypass is activated. All models (M1, M2 and M3) put the attribute Q-SP3 (flow rate in the bypass) on the top of the tree. We can expect class c1 (normal operation in winter) until Q-SP3 is less than certain value (929 m^3/day according to models M1 and M2 or 526 according to model M3). In this branch of the trees there are no sludge problems. Only model M1 predicts class c5 (underloading) in winter time if DQO-SP1 (COD concentration) < 270 mg/l (see Figure 4).

Sludge problems can be found in the branches where Q-SP3 is greater than 929 or 526, i.e. in summer (tourist season.). It is clear that in summer (tourist season) the plant is hydraulically overloaded, which is solved by activating the overflow bypass. At the same time organic matter concentration (microorganisms food) decreases due to greater water consumption per capita, i.e. the wastewater is more diluted. This causes bulking sludge, since the filamentous bacteria in flocks develop faster (better food access).

These consequences are usually visible after a few days of underloaded (biochemically) inflow to the plant. This is

dosega pa kar veliko natančnost. Za razliko od modelov M1 in M2, ki napovedujta težave z blatom na podlagi vtoka na ČN istega dne, model M3 napoveduje problem plavajočega blata v turistični sezoni (pri Q-SP3 > 567 m^3/dan) na podlagi pretoka skozi AB pred dvema dnevoma. Če je dotok na čistilno napravo izpred dveh dni (Q-AB-2) manjši od 10 662 m^3/dan ter iztok iz primarnega usedalnika 1 manjši od 9 197 m^3/dan, potem lahko pričakujemo težave z blatom (slika 6).

8. ZAKLJUČKI

Iz podane podatkovne baze, merjene na ČN za odpadno vodo, smo poskušali z odločitvenimi drevesi predvideti delovanje ČN. Kljub pomankljivostim v podatkovni bazi, smo s selekcijo relevantnih atributov (domenski ekspert) in redukcijo razredov uspeli zgraditi modele, ki so uporabni pri vodenju ČN, še zlasti pa pri odkrivanju težav, ki se pojavljajo na ČN.

Uvedba časovne zakasnitve se je izkazala kot dobra rešitev, saj je model zelo preprost, dosega pa največjo natančnost in prispeva k večji informiranosti.

Način merjenja in klasificiranja podatkov je bistvenega pomena za konstrukcijo uporabnih odločitvenih dreves. Za povečanje natančnosti modelov bi bilo smiselno ustrezno povečati bazo podatkov, ali pa prestrukturirati zapise v podatkovni bazi tako, da bo en zapis predstavljal dejansko situacijo delovanja ČN. Podatki na iztoku iz ČN bi morali biti rezultat vtočnih podatkov in ne povprečne dnevne vrednosti. Potrebna bi bila tudi klasificikacija podatkov z drugo, fleksibilnejšo metodo.

ZAHVALA

Del raziskave je nastal v okviru seminarja "Analiza ekoloških podatkov z orodji umetne inteligence", katerega nosilec je bil dr. Sašo Džeroski. Sredstva za podiplomsko raziskovalno usposabljanje prve avtorice je omogočilo Ministrstvo za šolstvo, znanost in šport Republike Slovenije po pogodbi št. S2-792-020/19443/99.

confirmed by model M3, which incorporates time delay attributes. The model is very simple and understandable and achieves quite high accuracy. M3 predicts sludge problems during the tourist season, i.e. when Q-SP3 > 567 m^3/dan according to the flow rate through AB, two days before. If the inflow to the WWTP two days ago was less than 10 662 m^3/dan and the flow rate through point SP1 is less then 9 197 m^3/dan, then we can expect sludge problems (see Figure 6).

8. CONCLUSIONS

We used the given data set measured on WWTP to predict WWTP operation by constructing decision trees. By selecting relevant attributes and reducing the classes we built three models that can be used in managing the plant and revealing the problems that occur in WWTP operation.

Incorporating time delay attributes appeared to be a good solution. Model M3 is very simple, has the highest accuracy and extracts more information.

Data preparation and classification is a crucial step in constructing useful decision trees. To increase the accuracy of the constructed models it would be necessary to enlarge the database or reorganise the data, so that one record would represent the actual situation of the treatment process. Outflow data should be a result of the inflow data and not average values during the day. Also the classification of data should be done by some other method.

ACKNOWLEDGEMENTS

Part of this research was elaborated within the seminar "Analysis of Environmental Data with Machine Learning Methods", conducted by Dr. Sašo Džeroski. The postgraduate study of the first author was subsidised by the Ministry of Education, Science and Sport of the Republic of Slovenia (contract No. S2-792-020/19443/99).

VIRI - REFERENCES

Belanche, L.A., Valdes, J.J., Comas, J., Roda, I.R., Poch, M. (1999). Towards a Model of Input-Output Behaviour of Wastewater Treatment Plants Using Soft Computing Techniques. *Environmental Modelling and Software* **14**, 409–419.

Bratko, I., Džeroski, S., Kompare, B., Urbančič, T. (2000). *Analysis of environmental data with machine learning methods. I., II*. Ljubljana: Jožef Stefan Institute, Center for Knowledge Transfer in Information Technologies.

Comas, J., Roda, I.R., Ceccaroni, L., Sánchez-Marré, M. (1999). *Semi-automatic learning with quantitative and qualitative features*, CAEPIA.

Comas, J., Dzeroski, S., Gibert, K., Roda, I.R., Sánchez-Marré, M. (2001). Knowledge discovery by means of inductive methods in wastewater treatment data. *AI Communications* **14**, 45-62.

Kompare, B. (1995). *The use of artificial intelligence in ecological modelling.* Ph.D. Thesis. Ljubljana: FGG; Kopenhagen: Royal Danish School of Pharmacy.

Quinlan, J.R. (1986). Induction of Decision Trees, *Machine Learning* **1**, 81–106.

Roda, I.R., Poch, M., Lafuente, J., Sánchez-Marré, M., Gimeno, J.M., Cortés, U. (1998). An Integrated Intelligent System to Improve Wastewater Treatment Plant Operation. *Int. Workshop on Decision and Control in Waste Bio-Processing (WASTE DECISION 98)*. Montpellier-Narbonne, France.

Sanchez, M., Cortes, U., Bejar, J., DeGracia, J., Lafuente, J., Poch, M. (1997). Concept Formation in WWTP by Means of Classification Techniques: A Compared Study. *Applied Intelligence* **7**(2): 147–166.

Witten, I. H., Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

Naslova avtorjev – Authors' Addresses

mag. Nataša Atanasova
Univerza v Ljubljani – University of Ljubljana
Fakulteta za gradbeništvo in geodezijo – Faculty of Civil and Geodetic Engineering
Inštitut za zdravstveno hidrotehniko – Institute of Sanitary Engineering
Jamova 2, SI–1000 Ljubljana, Slovenia
E–mail: natanaso@fgg.uni-lj.si

izr. prof. dr. Boris Kompare
Univerza v Ljubljani – University of Ljubljana
Fakulteta za gradbeništvo in geodezijo – Faculty of Civil and Geodetic Engineering
Inštitut za zdravstveno hidrotehniko – Institute of Sanitary Engineering
Jamova 2, SI–1000 Ljubljana, Slovenia
E–mail: bkompare@fgg.uni-lj.si