© *Acta hydrotechnica* **24/40 (2006)**, Ljubljana
ISSN 1581-0267

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

Prejeto/*Received*: 27. 7. 2007
Sprejeto/*Accepted*: 19. 8. 2008

# DVA PRISTOPA K MODELIRANJU ČISTILNE NAPRAVE ZA ODPADNO VODO
# TWO APPROACHES TO WASTEWATER TREATMENT PLANT MODELLING

**Boris KOMPARE, Meta LEVSTEK, Nataša ATANASOVA**

*Osnovni cilj vsakega modela je služiti namenu, za katerega je bil zgrajen. Modele čistilnih naprav za odpadno vodo (ČN) gradimo za boljše razumevanje procesov v sistemu, za odkrivanje novega znanja o sistemu ali pa za napoved delovanja ČN. Zato je uporaba različnih pristopov in metod modeliranja zelo zaželena, saj na ta način najlaže pridemo do modela, ki bi ustrezal našim potrebam. V tem prispevku uporabljamo dva pristopa k modeliranju pilotne čistilne naprave, postavljene znotraj centralne ČN Domžale-Kamnik. Prvi je t. i. konceptualno modeliranje, kjer gradimo modele izhajajoč iz osnovnih teoretičnih spoznanj o domeni. Tu smo uporabili poznani model ASM1, s katerim smo simulirali dogajanje v čistilni napravi. Drugi pristop se ukvarja z učenjem modelov iz podatkov. Uporabili smo metodo strojnega učenja za gradnjo regresijskih dreves. Uporabo metode so omogočile on-line meritve dušikovih spojin, iz katerih smo zgradili model za napoved iztočne koncentracije amonija. Vse dobljene modele smo analizirali z vidika ekspertovih pričakovanj in potreb.*
**Ključne besede:** *čistilna naprava za odpadno vodo, aktivno blato, modeliranje, konceptualni modeli, strojno učenje, regresijska drevesa.*

*The main goal of any model is to serve the purpose for which it was built. Models of wastewater treatment plants (WWTP) are built to better explain how processes work in the system, to reveal some new knowledge, or, simply, to predict their future behaviour. Therefore it should be strongly encouraged to use a variety of approaches and methods to find a model that will best satisfy our needs. In this paper we are using two basic approaches to model a pilot WWTP within the central WWTP Domžale-Kamnik (Slovenia). The first includes conceptual modelling, where models are constructed based on theoretical knowledge about the domain. For this approach we apply the well-known conceptual model ASM1 to simulate the processes in the WWTP. The second approach is concerned with learning from data. In this case we use a machine learning method for inducing regression trees. The on-line measurements of nitrogen compounds enabled the ML tool to build a model for prediction of ammonia outflow concentration. The models obtained are discussed in the light of the field expert's expectations and needs.*
**Key words**: *wastewater treatment plant, activated sludge, modelling, conceptual models, machine learning, regression trees*.

## 1. UVOD

Naraščajoče onesnaževanje okolja nalaga vedno strožje predpise za čiščenje odpadne vode. Zato je treba optimizirati vodenje in delovanje čistilnih naprav (ČN). Zaradi kompleksnosti samega procesa čiščenja je včasih težko zagotoviti optimalno delovanje ČN in s tem tudi iztok iz ČN, ki bi po kakovosti ustrezal vsem predpisom.

## 1. INTRODUCTION

The increasing pollution in the environment has reflected in severe laws and regulations regarding wastewater treatment. To meet these requirements, proper management of WWTPs has become a necessity. Due to the complexity of the cleaning process it is sometimes difficult to ensure the optimal WWTP operation and good quality outflow from the WWTP.

Modeliranje je eno izmed orodij, ki ga lahko uporabimo za optimiziranje delovanja ČN. Večinoma se uporabljajo konceptualni matematični modeli, ki izhajajo iz osnovnih fizikalnih, kemijskih in bioloških zakonitosti. Med najširše uporabljane sodi Activated Sludge Model No. 1 (ASM1), ki ga je razvila skupina IAWPRC (1986), ter njegove kompleksnejše različice ASM2, ASM2d in ASM3, ki jih je razvila ista skupina (kasneje preimenovana v IAWQ in nazadnje v IWA). Podrobno razlago modelov najdemo v literaturi (Henze *et al.*, 1986; 1995; 2000), kakor tudi obširni pregled literature obstoječih modelov (Gernaey *et al.*, 2004). Ti modeli so razumljivi in transparentni, kar je poglavitni vzrok za njihovo popularnost. Vendar pa je domensko znanje s področja delovanja ČN lahko zelo kompleksno, kar posledično povečuje kompleksnost matematičnih modelov. Medtem ko obstajajo uspešne aplikacije kompleksnejših modelov (npr. Iacopozzi *et al.*, 2005), številni primeri potrjujejo problem neizraznosti preveč kompleksnih modelov. Ta se kaže predvsem v oteženem umerjanju številnih parametrov, ki jih taki modeli vsebujejo. Čeprav se na prvi pogled zdi, da kompleksnejši modeli (s številnimi parametri) bolje (natančneje) opisujejo realnost, po drugi strani predstavljajo resen numerični problem v fazi umerjanja modela (Petersen *et al.*, 2002). Umerjanje zahteva veliko meritev in tudi veliko dragih eksperimentov. Checcy *et al.* (2005) predlagajo metodo za testiranje zanesljivosti določenih vrednosti parametrov. Zato je izbira ustreznega modela pravzaprav kompromis med (pre)kompleksnimi modeli s številnimi parametri in (pre)enostavnimi modeli z zelo omejeno rabo.

Druga možnost je ‘pozabiti’ na teoretično domensko znanje in se učiti delovanja sistema samo iz merjenih podatkov. Orodja umetne inteligence (AI), med katere prištevamo tudi strojno učenje (ML), omogočajo konstrukcijo enostavnih modelov, ki potrjujejo domensko znanje in prav tako odkrivajo novega, kar je bilo že dokazano na mnogih okoljskih domenah (Kompare, 1995).

Na področju ČN so bila uporabljena številna orodja AI, kot so npr. ekspertni

Therefore modelling is becoming a commonly used tool for optimizing the WWTP operation. Mathematical models constructed from the basic physical, chemical and biological principles are mostly used. One of the most popular models is the Activated Sludge Model No. 1 (ASM1), developed by IAWPRC (1986), and its more complex versions ASM2, ASM2d, and ASM3, developed by the same group (later renamed to IAWQ, and finally to IWA). A detailed explanation of these models can be found in literature (Henze *et al.*, 1986; 1995; 2000), as well as a substantial overview of the existing models (Gernaey *et al.*, 2004). The popularity of these models is a result of their transparency and clearness to the domain experts. However, the domain specific knowledge can be very complex, which results in increasing complexity of the mathematical models. While successful applications of complex models can be found in the literature (e.g. Iacopozzi *et al.*, 2005) in many cases scientists encounter problems due to too complex models, which are reflected in unsuccessful estimation of their numerous parameters. Although complex models (with many parameters) may seem to give a more complete (correct) presentation of the reality, they, on the other hand, represent a serious numerical problem in the calibration phase (Petersen *et al.*, 2002). This process requires good quality data and many costly experiments. To this end Checcy *et al.* (2005) suggest a method for testing the reliability of parameter estimation. Thus, in order to find the suitable model we need to balance between too sophisticated models with many parameters, and too simple mathematical models with very limited use.

The other option is to ‘forget’ the theoretical domain knowledge and to learn the system from the measured data. Artificial intelligence (AI) tools, which include machine learning (ML), enable a construction of simple models that confirm the background knowledge and also discover new knowledge, which was confirmed in many environmental domains (Kompare, 1995).

Different approaches to modelling and control of WWTP operation have already been applied, such as expert systems (Baeza *et al.*,

sistemi (Baeza *et al.*, 1999; Roda *et al.*, 1999), pristop zasnovan na bazi znanja (Comas *et al.*, 2003), nevronske mreže (Belanche *et al.*, 1999), hibridni pristopi (Sanchez-Marre *et al.*, 1996, Grieu *et al.*, 2005) in strojno učenje (Comas *et al.*, 2001; Atanasova & Kompare, 2002).

V tem prispevku uporabljamo oba osnovna principa k modeliranju. Osnovni namen je oceniti primernost obeh pristopov glede na podatke, ki so na razpolago, in v končni fazi zgraditi uporaben model z eno izmed metod. To analizo sta nam omogočila dva podatkovna niza iste čistilne naprave, ki se razlikujeta po kakovosti in količini podatkov.

# 2. ČISTILNA NAPRAVA

## 2.1 POSTOPEK Z AKTIVNIM BLATOM

Čiščenje odpadne vode lahko poteka po številnih postopkih (Henze *et al.*, 1997, Tschobanouglos & Burton, 1991), od katerih je za komunalne vode najširše uporabljan postopek z aktivnim blatom. Tehnologija je zasnovana na bioloških procesih, s katerimi se iz odpadne vode odstranjujejo organska snov, dušik in fosfor. Čiščenje vršijo različne vrste mikroorganizmov (biomasa) s svojim metabolizmom.

Glavna elementa biološke čistilne naprave z aktivnim blatom sta biološki reaktor in naknadni ali sekundarni usedalnik (slika 1). Odpadna voda doteka v biološki reaktor, kjer mikroorganizmi razgrajujejo onesnaženje s tem, da ga uporabljajo za svojo rast. Med tem ko rastejo, se mikroorganizmi združujejo v kosme (aktivna biomasa ali blato), ki so sestavljeni iz različnih vrst mikroorganizmov. Tako nastalo mešanico vode in biološkega blata (BB) vodimo nato iz reaktorja v usedalnik, kjer se biološko blato začne usedati, očiščena voda pa odteka običajno v površinsko vodo oziroma recipient. Usedlo biološko blato večinoma recirkuliramo nazaj v reaktor in s tem vzdržujemo zadostno koncentracijo biološkega blata za učinkovit potek čiščenja v reaktorju. Odvišno biološko blato odvajamo in ga obdelujemo ločeno.

Čiščenje različnih sestavin odpadne vode zahteva vzdrževanje ustreznih pogojev. Odstranjevanje organskega onesnaženja lahko

1999; Roda *et al.*, 1999), knowledge-based approach (Comas *et al.*, 2003), neural networks (Belanche *et al.*, 1999), hybrid approaches (Sanchez-Marre *et al.*, 1996, Grieu *et al.*, 2005) and machine learning methods (Comas *et al.*, 2001; Atanasova & Kompare, 2002).

In this paper we use both, conceptual and ML modelling. The aim is to estimate and point the appropriateness of both modelling approaches regarding the data set available and finally to build a useful model. Such research was conducted using two data sets from the same WWTP that differ in their quality and quantity.

# 2. WWTP

## 2.1 ACTIVATED SLUDGE PROCESS

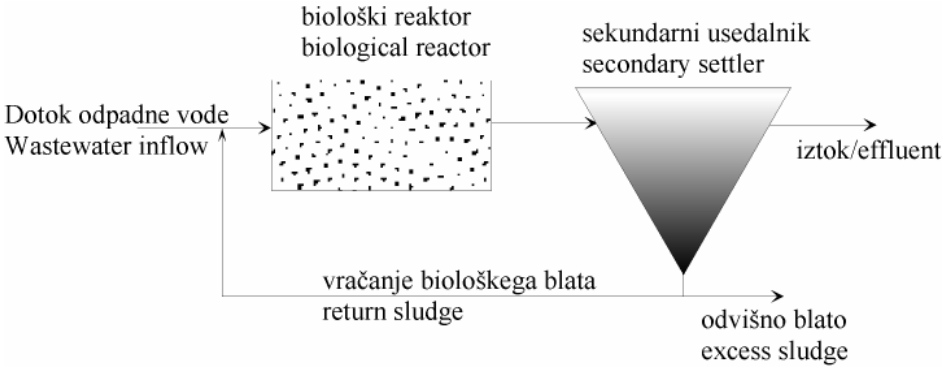
There is a variety of technologies (Henze *et al.*, 1997, Tschobanouglos & Burton, 1991) applied to a WWTP, among which activated sludge is most widely used for municipal wastewater treatment. This technology is based on biological processes by which organic matter, nitrogen and phosphorous are removed from water. The processes are carried out by many different microorganisms, using the compounds for their metabolism.

Activated sludge WWTP (see Figure 1) contains two main units: a bio-reactor and a secondary settler or clarifier. Wastewater enters the reactor, where microorganisms carry out the biological treatment, by utilizing the compounds to be removed from the water for their growth. While growing, the microorganisms form flocks, i.e. activated sludge, composed of different microorganisms species. The produced mixture of water and activated sludge (microorganisms) is then sent to the clarifier, where the flocks of microorganisms settle down, thus clean water is discharged to the recipient. Most of the settled sludge is returned to the biological reactor in order to maintain the sufficient concentration of microorganisms for the cleaning process. Excess sludge is removed and handled separately.

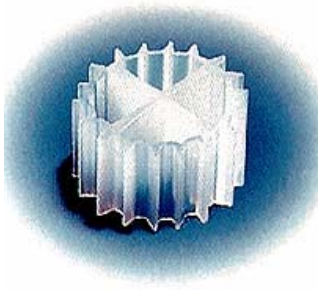
Suitable conditions must be maintained in the reactor for adequate treatment. For

poteka bodisi v oksičnih ali anoksičnih pogojih, medtem ko odstranjevanje dušika poteka v dveh korakih: (1) nitrifikacija, tj. pretvorba amonija v nitrat v oksičnih pogojih, in (2) denitrifikacija, tj. pretvorba nitrata v atmosferski dušik pod anoksičnimi pogoji. V tem primeru je biološki reaktor sestavljen iz dveh tipov reaktorjev, in sicer najprej anoksičnih (ali zaporedja anoksičnih) in nato oksičnih (ali zaporedja oksičnih) reaktorjev (glej sliko 2).

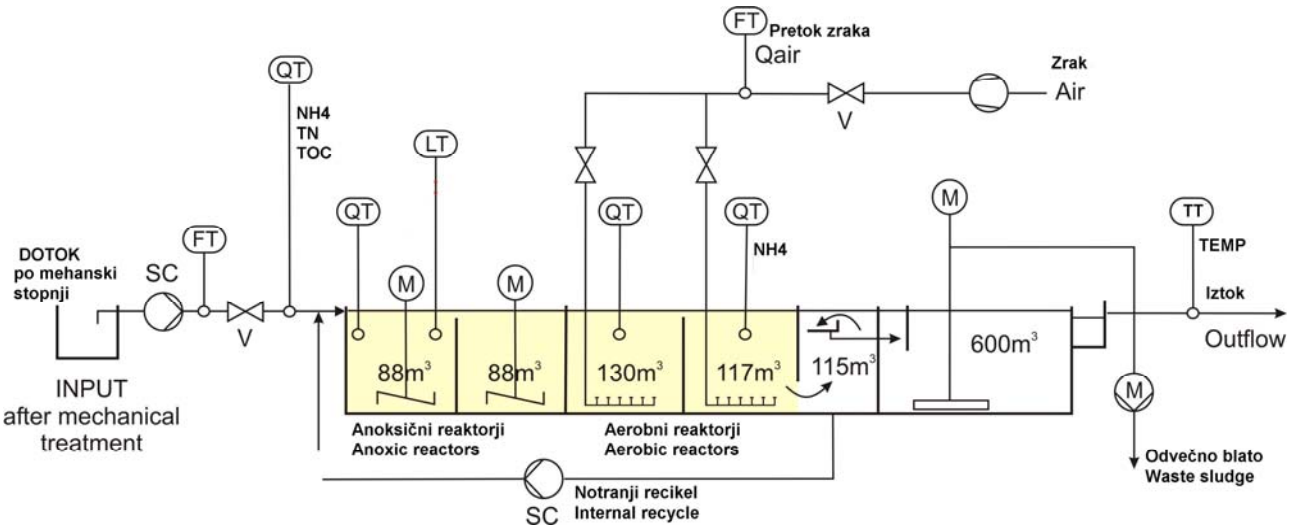
removal of organic matter oxic conditions must be ensured. Nitrogen removal is a two-step process: (1) nitrification, i.e. conversion of ammonia to nitrate under oxic conditions, and (2) denitrification, i.e. conversion of nitrate to gas nitrogen under anoxic conditions. In this case the biological reactor is composed of two types of reactors, anoxic (or series of anoxic) and aerobic (or series of aerobic) reactors (Figure 2).




Slika 1. Shema procesa z aktivnim blatom.
*Figure 1. Scheme of the activated sludge process.*



Slika 3. Plavajoči nosilec biomase.
*Figure 3. Free-floating biomass carrier.*




Slika 2. Tehnološka shema pilotne ČN.
*Figure 2. Technological scheme of the pilot wastewater treatment plant.*

## 2.2 PILOTNA ČN

Pilotna naprava skupne prostornine 500 $m^3$ in hidravličnega zadrževalnega časa 6 ur je bila postavljena v aeracijskem bazenu obstoječe čistilne naprave Domžale-Kamnik (slika 2). Pilotna naprava je bila sestavljena iz

## 2.2 CASE STUDY – PILOT WWTP

A pilot plant with a total volume of 500 $m^3$ and hydraulic retention time of 6 hours was put in the aeration reactor of the existing WWTP Domžale-Kamnik (Slovenia), Figure 2. The plant is composed of four sequential

štirih zaporedno vezanih bioreaktorjev, od tega dveh za denitrifikacijo (anoksična) in dveh za nitrifikacijo (oksična). V vseh bazenih se nahajajo plavajoči nosilci biomase (slika 3), ki zasedejo 60 vol %.

Tehnološka shema obravnavane čistilne naprave je prikazana na sliki 2. Voda priteka najprej v anoksična reaktorja, kjer poteka denitrifikacija in z njo delno odstranjevanje organske snovi. Nato gre voda v prezračevana (oksična) reaktorja, kjer potekata odstranjevanje organske snovi (s privzemom v biomaso in dihanjem biomase) ter nitrifikacija. Proizvedeni nitrat v procesu nitrifikacije se iz reaktorja recirkulira v anoksični reaktor, kjer se v procesu denitrifikacije pretvori v atmosferski dušik. V čistilni proces sta vključeni dve glavni skupini mikroorganizmov: (1) heterotrofna biomasa, ki vrši odstranjevanje organske snovi in denitrifikacijo, in (2) avtotrofna biomasa, ki izvaja nitrifikacijo.

Kot smo navedli zgoraj, čistilna naprava uporablja tehnologijo s plavajočimi nosilci biomase. Razlika med to tehnologijo in tehnologijo z aktivnim blatom je v tem, da je pri nosilcih aktivno blato pritrjeno in raste na prosto plavajočih nosilnih elementih. Ti so premešani v reaktorju. Gre za kombinacijo tehnologij s pritrjenim blatom (biofilm) in prosto plavajočim. Vračanje blata tu načeloma ni potrebno, saj večina blata ostane na nosilcih v reaktorju.

## 2.3 PODATKOVNA BAZA

Podatkovna baza ČN vsebuje dva niza podatkov. V prvem nizu je merjena večina količin, ki določajo delovanje čistilne naprave. Merjeni so sledeči atributi: Q (pretok), COD (KPK), $BOD_5$ ($BPK_5$), TKN (skupni Kjeldahlov dušik), $NH_4$ (amonijev dušik), $NO_x$ (nitratni dušik), DBOD (raztopljen BOD), DCOD (raztopljen COD), DTKN (raztopljen TKN), T (temperatura), pH. Podatki so merjeni dvakrat dnevno za obdobje petih dni. Ta niz podatkov je primeren za postavitev konceptualnega modela, saj vsebuje skoraj vse količine, ki jih lahko modeliramo. Vendar pa je obdobje merjenja zelo kratko, kar pomeni, da je premalo meritev za uporabo metod strojnega učenja.

Drugi niz podatkov so *on-line* meritve

bio-reactors. Two of them are anoxic (for denitrification) and two are oxic (for nitrification). Biomass carriers were introduced to all reactors (Figure 2). They take 60% of the volume.

The technological scheme of the WWTP under study is shown in Figure 2. Wastewater goes first to the anoxic tanks, where denitrification and removal of organic matter under anoxic conditions is performed. Water is then transported to the aerobic (oxic) tanks, where oxidation of organic matter and nitrification take place. The produced nitrate in the nitrification process is sent back to the anoxic tank, by internal recycle to be removed in the denitrification process. Two main groups of micro-organisms are involved in the treatment process: (1) heterotrophic biomass, performing organic matter removal and denitrification, and (2) autotrophic biomass, performing nitrification.

As stated above, this plant is using a technology with suspended carriers. The difference between this technology and the technology with activated sludge is that here the activated sludge is fixed and grows on free-floating carrier elements, which are mixed in the reactor. It is a combination of biofilm and activated sludge processes. No sludge recycle is needed.

## 2.3 DATA BASE

There are two series of data measured on the WWTP. The first comprises most of the quantities that determine a WWTP operation. The following is measured: Q (flow rate), COD, $BOD_5$, TKN (total Kjeldahl nitrogen), $NH_4$ (ammonium nitrogen), $NO_x$ (nitrate), DBOD (dissolved BOD), DCOD (dissolved COD), DTKN (dissolved TKN), T (temperature), pH. The meausurements took place twice a day for a period of five days. This data set is suitable for the conceptual modelling purpose. It contains nearly all quantities that can be modelled. However the measuring period is rather short, which means there are not enough measurements (of the quantities) for machine learning methods.

The second series of data are on-line measured data of nitrogen compounds, total

dušikovih spojin, skupnega organskega ogljika in temperature. Količine so merjene vsakih 15 minut za obdobje štirih mesecev (od začetka septembra do konca decembra). Merjeno je sledeče: skupni dušik (TN-IN) in skupni organski ogljik (TOC-IN) na vtoku ter amonijev dušik (NH4-OUT) in temperatura (TEMP) na iztoku. Očitno je, da ta niz vsebuje premalo merjenih količin za postavitev konceptualnega modela. So pa zato te količine merjene dovolj pogosto, da je niz primeren za obdelavo s strojnim učenjem.

## 3. KONCEPTUALNO MODELIRANJE ČN

Konceptualne modele ČN gradimo na podlagi osnovnih fizikalnih, bioloških in kemijskih principov. V tej raziskavi uporabljamo poznani model za ČN ASM1, ki simulira odstranjevanje organskega onesnaženja in dušika, tj. nitrifikacijo in denitrifikacijo. Sestavlja ga sistem navadnih diferencialnih enačb, ki predstavljajo masne bilance snovi, ki jih modeliramo, oziroma spremenljivk stanja ali odvisnih spremenljivk. Z reševanjem enačb modela dobimo vrednosti spremenljivk stanja (koncentracije snovi) v odvisnosti od časa. Gre za razmeroma kompleksen model, ki vključuje večje število odvisnih spremenljivk in parametrov (Henze *et al.*, 1986). V nadaljevanju podajamo odvisne spremenljivke, ki jih upošteva model ASM1.

### *Organska snov*

Skupno organsko onesnaženje, izraženo kot KPK (kemijska potreba po kisiku), na vtoku sestavljata biološko razgradljivi in nerazgradljivi delež, od katerih je vsak delno raztopljen in delno netopen (slika 4). Nerazgradljivi raztopljeni del ($S_I$) zapušča sistem v enaki koncentraciji kot ob vstopu, medtem ko se nerazgradljivi netopni del ($X_I$) adsorbira na biomasi in zapusti sistem s pretokom odvišnega blata. Razgradljivi topni delež ($S_S$) porabljajo mikroorganizmi (heterotrofna biomasa) za rast. Mikroorganizmi prav tako porabljajo razgradljivi netopni del ($X_S$), vendar se mora ta najprej pretvoriti v topno obliko ($S_S$) v procesu hidrolize.

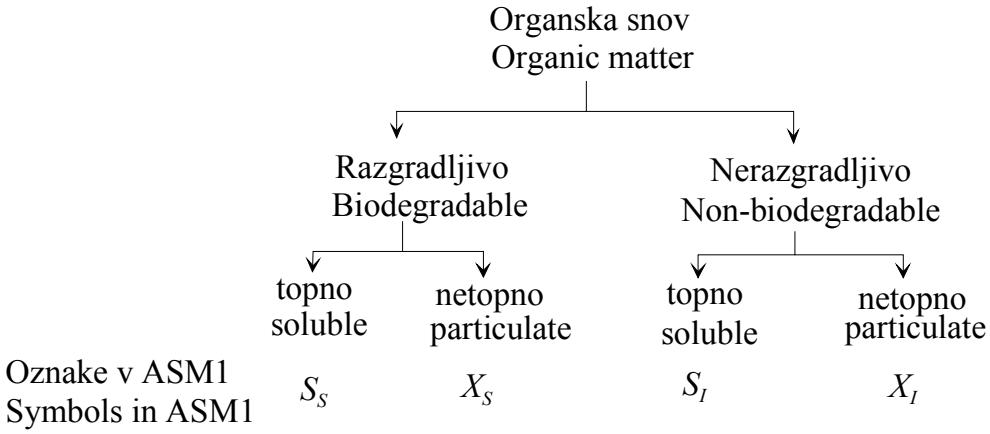
organic carbon and temperature. Data are taken each 15 minutes for a four-month period (September to December). Total nitrogen (TN-IN) and total organic carbon (TOC-IN) are measured at the plant inflow, while ammonium nitrogen (NH4-OUT) and temperature (TEMP) at the outflow. It is obvious that this data set is missing quite some quantities in order to be used for conceptual modelling. Yet, the quantities here are measured frequently enough and thus, suitable for modelling with ML tools.

## 3. CONCEPTUAL MODELLING OF WWTP

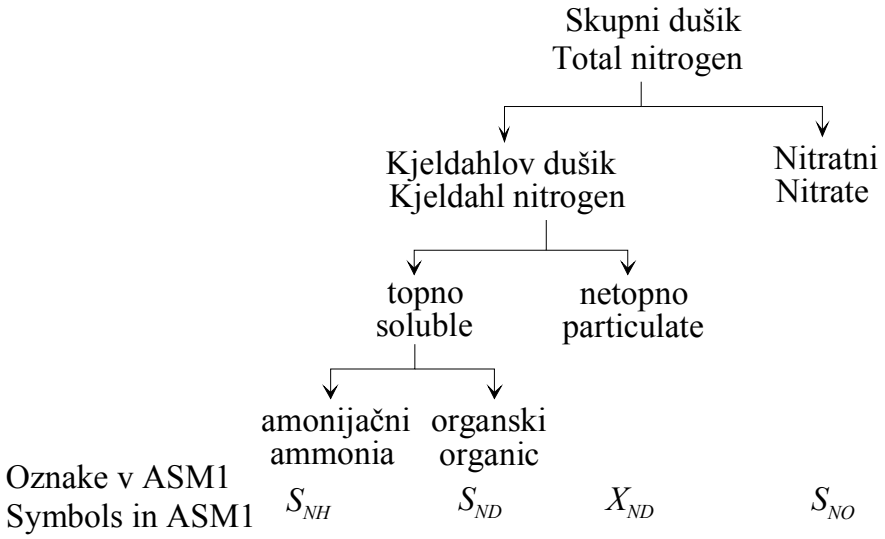
Conceptual models of WWTP are built from basic physical, biological and chemical principles. In our study, we use a well-known model ASM1, which simulates the removal of organic matter, nitrification and denitrification. It is composed of a set of ordinary differential equations, representing mass balances for each substance of interest, i.e. state variable. By solving these model equations we calculate the values of each state variable (concentrations of the substances) as a function of time. It is a fairly complex model, including a large amount of state variables and parameters (Henze *et al.*, 1986). The ASM1 model incorporates the following state variables.

### *Organic matter*

Total organic matter expressed as COD (chemical oxygen demand) on the inflow is composed of biodegradable and non-biodegradable parts, each of which have dissolved and particulate parts (see Figure 4). Non-biodegradable dissolved organic matter ($S_I$) leaves the system in the same concentration as it reaches it, while the non-biodegradable particulate matter ($X_I$) is adsorbed on biomass and leaves the system with the waste flow rate. Biodegradable dissolved matter ($S_S$) is directly available to the heterotrophic biomass. Biodegradable particulate matter ($X_S$) is also consumed by microorganisms, but first it has to be converted in easily degradable substrate ($S_S$) in the hydrolysis process.




Slika 4. Komponente organske snovi (spremenljivke stanja) v modelu ASM1.
*Figure 4. Organic components (state variables) in the ASM1 model.*




Slika 5. Komponente dušika (spremenljivke stanja) v modelu ASM1.
*Figure 5. Nitrogen components (state variables) in the ASM1 model.*

### *Dušik*

Skupni dušik v odpadni vodi je sestavljen iz skupnega Kjeldahlovega dušika in nitrata ($S_{NO}$), glej sliko 5. Kjeldahlov dušik pa sestavljajo netopni biorazgradljivi organski dušik ($X_{ND}$) in topni Kjeldahlov dušik, ki ga sestavljata amonijak ($S_{NH}$) in (topni) biorazgradljivi organski dušik ($S_{ND}$).

### *Biomasa*

Biomasa je v modelu upoštevana preko treh komponent – heterotrofne biomase ($X_{B,H}$), avtotrofne biomase ($X_{B,A}$) in komponente, ki nastaja zaradi odmiranja biomase ($X_P$).

### *Nitrogen*

The total nitrogen in wastewater is composed of total Kjeldahl nitrogen and nitrates (SNO), see Figure 5. The components of total Kjeldahl nitrogen are particulate biodegradable organic nitrogen (XND) and dissolved Kjeldahl nitrogen, composed of ammonia (SNH) and dissolved biodegradable organic nitrogen (SND).

### *Biomass*

Biomass is presented through three components – heterotrophic biomass ($X_{B,H}$), autotrophic biomass ($X_{B,A}$) and a component that arises from biomass decay ($X_P$).

***Alkalnost***

Skupna alkalnost ($S_{ALK}$) je vsota negativno nabitih ionov in daje informacijo o spremembi pH-vrednosti.

Kinetični procesi v ČN delujejo na opisane spremenljivke stanja, tj. vplivajo na spremembo njihove koncentracije s časom. Model ASM1 upošteva naslednje procese:

- rast heterotrofnih organizmov (v oksičnih in anoksičnih pogojih)
- aerobna rast avtotrofnih organizmov
- odmiranje heterotrofov
- odmiranje avtotrofov
- amonifikacija topnega organskega dušika
- hidroliza netopne organske snovi
- hidroliza organskega dušika

Matematični model sestavlja sistem navadnih diferencialnih enačb, po ena za vsako spremenljivko stanja. Natančnejši opis enačb je podan v poglavju 5.1.

## 4. MODELIRANJE S STROJNIM UČENJEM

Za razliko od konceptualnih modelov, ki so osnovani na teoretičnem znanju o domeni, so modeli strojnega učenja zgrajeni na podlagi učenja sistema iz meritev. Osnovna naloga strojnega učenja je torej, naučiti se neki koncept iz podanih meritev, ki ta koncept opisujejo. Celotni učni postopek strojnega učenja je sestavljen iz koncepta, primerov (meritve), učnega algoritma in učne sheme oziroma modela (slika 6). Opisom koncepta pravimo *primeri*. Primere podajamo običajno v tabeli tako, da je en primer (ena vrstica v tabeli) sestavljen iz atributov, tj. neodvisnih spremenljivk, in razreda primera (koncept), tj. odvisnih spremenljivk. Učni algoritem nato iz primerov in področnega znanja generira učno shemo, ki predstavlja model oziroma predstavitev naučenega. Učna shema ali model je lahko odločitveno drevo, regresijsko drevo, klasifikacijska pravila, odločitvene tabele in podobno.

V tej raziskavi smo uporabili učni algoritem, ki generira regresijska drevesa.

***Alkalinity***

Total alkalinity ($S_{ALK}$) is a sum of negative ion charges and gives information about pH changes.

The state variables described are involved in kinetic processes, by which their concentration is changed. The following processes are modelled in ASM1:

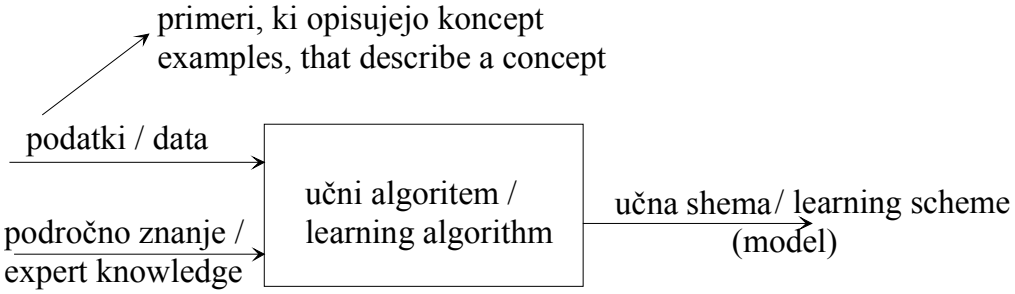
- microbial growth of heterotrophs (aerobic and anoxic)
- aerobic growth of autotrophs
- decay of heterotrophs
- decay of autotrophs
- ammonification of soluble organic nitrogen
- hydrolysis of entrapped organics
- hydrolysis of entrapped organic nitrogen

The mathematical model is represented by a set of ordinary differential equations, one for each state variable. A more detailed description on model equations is given in section 5.1.

## 4. MODELLING WITH MACHINE LEARNING TOOLS

Unlike the conceptual models, which are based on theoretical knowledge about the domain, ML models are constructed by learning the system from measured data. Thus, the main task of ML is to learn a concept from given measurements, describing that concept. The machine learning procedure consists of the concept, examples (measurements), learning algorithm and learning scheme or model (Figure 6). The instances can be presented in a table, each row being an instance, composed of attributes of the instance or independent variables and a class of instance or dependant variables, which is the concept to be learned by the algorithm. The learning algorithm then, from the examples and some background knowledge, generates the learning scheme, or a model, which is a presentation of what has been learned. The learning scheme or model can be a decision tree, regression tree, classification rules, decision tables, and so on.

In this paper we are using a learning algorithm that generates regression trees.




Slika 6. Postopek strojnega učenja.
*Figure 6. Machine learning procedure.*

### 4.1 REGRESIJSKA DREVESA

Linearna regresija je statistična metoda, katere cilj je izraziti odvisno spremenljivko kot linearno kombinacijo neodvisnih spremenljivk iz podanih meritev (primerov). Vsak primer določajo vrednosti atributov (neodvisnih spremenljivk) in razreda primera, tj. odvisne spremenljivke. Če je $x$ razred in $a_i$ atributi primera, potem je:

$$x = a_1*w_1 + a_2*w_2 + \ldots + a_n*w_n = \Sigma a_i*w_i,$$

kjer so $w_i$ uteži, izračunane iz podatkov oziroma učnega niza. Z izračunanim vektorjem uteži lahko določimo razred (odvisno spremenljivko) za vsak primer v učnem nizu.

Medtem ko enostavna linearna regresija določi en (linearni) model odvisne spremenljivke za celotni niz podatkov, drevesno strukturirana regresija najprej smiselno razdeli niz na podnize ter vsakemu izmed njih določi linearni model. V tem smislu lahko drevesna linearna regresija mnogo bolje opiše nelinearno obnašanje odvisne spremenljivke. Regresijsko drevo je zgrajeno iz vozlišč in vej. Veje povezujejo vozlišča in liste drevesa, ki predstavljajo končna vozlišča. V listih se napoveduje odvisna spremenljivka. Tu imamo lahko eno vrednost spremenljivke – takim drevesom rečemo enostavna regresijska drevesa – ali pa linearno enačbo, takim pa rečemo modelna regresijska drevesa. Za naše eksperimente smo uporabili učni algoritem M5 (Quinlan, 1992).

Ko je model (regresijsko drevo) zgrajen iz učnega niza podatkov, je treba ovrednotiti njegovo uporabnost oziroma natančnost. To

### 4.1 REGRESSION TREES

Linear regression is a method, which aims to express the dependent variable as a linear combination of the independent variables from the given measurements (examples). Each example is given by the values of the attributes (independent variables) and by the class (dependent variable). If $x$ is the class and $a_i$ are the attributes of an example then:

$$x = a_1*w_1 + a_2*w_2 + \ldots + a_n*w_n = \Sigma a_i*w_i$$

where $w_i$ are weights, which are learned (calculated) from the training set. Using the calculated vector of weights we can determine the class for every example in the data set.

While the simple linear regression calculates one equation (one weight vector) for the entire data set, piecewise or tree-structured regression divides the data set to several subsets on which uniform class value or linear equation can be applied. In this manner the piecewise linear parts can much better cover the non-linear behaviour of the dependent variable. The regression tree consists of nodes (branching points) and branches. The branches connect the nodes and the leaves of the tree, which are terminal nodes where the dependant variable is predicted. If we have a single value for the class prediction then we are talking about simple regression trees, while if linear equation is used for prediction in the leaf we are speaking of (regression) model trees. For our experiments we use the M5 algorithm (Quinlan, 1992).

After the tree is constructed from the training (learning) set of data, it is necessary to assess the model quality, i.e. the accuracy of prediction. This can be done by providing an

lahko storimo z uporabo modela na testnem nizu podatkov, torej na podatkih, ki jih učni algoritem ne pozna. Druga izbira je t. i. navzkrižna validacija. Po tej metodi se celotni niz razdeli na izbrano število podnizov (*n*). V vsaki iteraciji se en podniz izloči za testiranje, medtem ko se preostali niz (*n-1* podnizov) uporabi za učenje. Po *n*-ti iteraciji se določi model, ki se najbolje prilega vsem podnizom. Na ta način metoda izboljšuje delovanje modela na novih podatkih.

Napaka modela se vrednoti po nekaj statističnih merah, kot so: koren srednje napake kvadratov, srednja absolutna napaka, koren relativne napake kvadratov, relativna absolutna napaka in koeficient korelacije.

# 5. EKSPERIMENTI

## 5.1 KONCEPTUALNI MODEL

Model ASM1 simulira dogajanje v čistilni napravi z aktivnim biološkim blatom, tj. mikroorganizmi se prosto gibljejo in formirajo flokule. Obravnavana čistilna naprava ima nekoliko drugačno tehnologijo, tj. organizmi rastejo na majhnih nosilcih, ti pa so premešani v reaktorju. Za dosledno simulacijo takega sistema je treba poleg procesov, formuliranih v modelu ASM1, upoštevati še procese, značilne za pritrjeno biomaso, tj. upoštevati stratifikacijo sistema in prehajanje snovi skozi sloje z difuzijo. Vendar pa je z upoštevanjem določenih poenostavitev (spomnimo se, da je vsak še tako kompleksen model le poenostavitev naravnih procesov) mogoče modelirati ČN kot sistem z razpršeno aktivno biomaso. Nenazadnje je biomasa, čeprav pritrjena na nosilcih, premešana v reaktorju skupaj z nosilci, ki jih lahko upoštevamo kot flokule mikroorganizmov.

Za modeliranje takega sistema smo model ASM1 še dodatno poenostavili tako, da simuliramo le osnovne procese: aerobno rast heterotrofnih organizmov, anoksično rast heterotrofnih organizmov, rast avtotrofnih organizmov, odmiranje heterotrofov in odmiranje avtotrofov, ter sledeče spremenljivke stanja: raztopljeni delež organskega onesnaženja (merjeni $BPK_5$), amonijev dušik, nitrat, heterotrofne in avtotrofne organizme.

additional testing set of data. Another option is to employ cross-validation. The given data set is partitioned on a chosen number of folds (*n*). In turn, each fold is used for testing, while the remainder (*n-1* folds) is used for training. Finally the model that fits best to all folds is chosen. This method prevents the tree from over-fitting the learning data and improves the quality (performance) of the tree on unseen data.

The size of the error between the actual and the predicted values is calculated by several measures to evaluate the model accuracy: root mean-squared error, mean absolute error, root relative squared error, relative absolute error, and correlation coefficient.

# 5. EXPERIMENTS

## 5.1 SETTING THE CONCEPTUAL MODEL

The ASM1 model simulates processes in an activated sludge WWTP, i.e. free floating microorganisms. The technology in our WWTP differs from the typical activated sludge in a sense that microorganisms are fixed on small carriers and mixed in the reactor. To model such system it is necessary to incorporate the biofilm processes, i.e., to consider the stratification of the system and passing of the substances through the layers by diffusion, besides the processes formulated in the ASM1 model. However, with some modifications (recall that even complex models represent a simplification of natural processes) it is possible to model the system as an activated sludge system. After all, the biomass (although fixed on the carriers) is mixed in the reactor together with the carriers and we can consider the carriers as flocks of microorganisms.

To model such system we simplified the ASM1 model by taking only the basic processes into account: aerobic growth of heterotrophs, anoxic growth of heterotrophs, growth of autotrophs, decay of heterotrophs and decay of autotrophs. The following state variables were modelled: Dissolved substrate (measured $BOD_5$), ammonium nitrogen, nitrate, heterotrophic and autotrophic organisms.

Preglednica 1. Kinetika procesov in stehiometrični parametri v poenostavljenem modelu ASM1.
*Table 1. Process kinetics and stoichiometric parameters in the simplified ASM1 model.*

| | $S_S$ | $X_{B,H}$ | $X_{B,A}$ | $S_{NO}$ | $S_{NH}$ | Kinetics [$ML^{-3}T^{-1}$] |
|---|---|---|---|---|---|---|
| Aerobna rast heterotrofov / *Aerobic growth of heterotrophs* | $-\frac{1}{Y_H}$ | 1 | | | $-i_{XB}$ | $\mu_H \cdot \left(\frac{S_S}{K_S + S_S}\right) \cdot \left(\frac{S_0}{K_{O,H} + S_0}\right) \cdot f\_NH4 \times x_{B,H}$ |
| Anoksična rast heterotrofov / *Anoxic growth of heterotrophs* | $-\frac{1}{Y_H}$ | 1 | | $-\frac{1-Y_H}{2{,}86 \cdot Y_H}$ | $-i_{XB}$ | $\mu_H \cdot \left(\frac{S_S}{K_S + S_S}\right) \cdot \left(\frac{K_{O,H}}{K_{O,H} + S_O}\right) \cdot \left(\frac{S_{NO}}{K_{NO} + S_{NO}}\right) \cdot f\_NH4 \cdot \eta_g \cdot x_{B,H}$ |
| Aerobna rast avtotrofov / *Aerobic growth of autotrophs* | | | 1 | $\frac{1}{Y_A}$ | $-i_{XB} - \frac{1}{Y_A}$ | $\mu_A \cdot \left(\frac{S_{NH}}{K_{NH} + S_{NH}}\right) \cdot \left(\frac{S_0}{K_{O,A} + S_0}\right) \cdot X_{B,A}$ |
| Odmiranje heterotrofov / *Decay of heterotrophs* | | - 1 | | | | $b_H \cdot X_{B,H}$ |
| Odmiranje avtotrofov / *Decay of autotrophs* | | | - 1 | | | $b_A \cdot X_{B,A}$ |
| **Parametri modela / *Parameters*** | **Stehiometrijski / *stoichiometric*:**<br>$Y_H$ prirast heterotrofne biomase na enoto porabljenega substrata / *Heterotrophic yield*<br>$Y_A$ prirast avtotrofne biomase na enoto porabljenega substrata (amonij) / *Autotrophic yield*<br>$i_{XB}$ delež dušika v biomasi / *nitrogen fraction in biomass*<br><br>**Kinetični / *kinetic*:**<br>$\mu_H$ maksimalna rast heterotrofov v oksičnih pogojih / *max. growth rate (het)*<br>$\mu_A$ maksimalna rast avtrotrofov / *max. growth rate (aut)*<br>$b_H$ hitrost odmiranja heterotrofov / *decay rate (het)*<br>$b_A$ hitrost odmiranja avtotrofov / *decay rate (aut)*<br>$K_S$ polsaturacijski koeficient za substrat / *half-saturation constant for substrate*<br>$K_{O,H}$ polsaturacijski koeficient za kisik (het) / *half-saturation constant for oxygen (het)*<br>$K_{O,A}$ polsaturacijski koeficient za kisik (avt) / *half-saturation constant for oxygen (aut)*<br>$K_{NO}$ polsaturacijski koeficient za nitrat / *half-saturation constant for nitrate*<br>$K_{NH}$ polsaturacijski koeficient za amonij / *half-saturation constant for ammonia*<br>$f\_NH4$ funkcija vpliva amonija na biomaso / *ammonia influence function* | | | | | |

Poenostavljeni model ASM1 prikazuje preglednica 1. V prvi vrstici so zapisane modelirane spremenljivke stanja, v prvem stolpcu pa procesi, ki na njih delujejo. Kinetične enačbe procesov so navedene v zadnjem stolpcu. Enačba posamezne spremenljivke stanja je podana z naslednjim izrazom:

The simplified ASM1 model is shown in Table 1. The first row contains the modelled state variables and the first column the processes of their conversions. Kinetic equations of the processes are listed in the last column. The equation for each state variable is given by the following expression:

$$r_i = \sum_j \upsilon_{i,j} \cdot \rho_j \tag{1}$$

kjer $r_i$ pomeni hitrost reakcije za spremenljivko $i$, $\upsilon_{ij}$ je stehiometrijski koeficient v $j$-tem procesu spremenljivke i in $\rho_j$ kinetika $j$-tega procesa. Na primer, formulacija hitrosti reakcije spremenljivke $S_s$ (koncentracija topnega organskega onesnaženja) v procesu aerobne rasti heterotrofov se glasi:

where $r_i$ is reaction rate of the variable $i$, $\upsilon_{ij}$ is the stoichiometric coefficient in the $j$-th process of the variable $i$ and $\rho_j$ the kinetic of the $j$-th process. For instance, the rate of reaction for soluble substrate $S_s$ in the process of aerobic growth of heterotrophic organisms would be:

$$\frac{dS_S}{dt} = -\frac{1}{Y_H} \cdot \mu_H \cdot \frac{S_S}{K_S + S_S} \cdot \frac{S_O}{K_O + S_O} \cdot f\_NH4 \cdot X_{B,H} \tag{2}$$

Sistem diferencialnih enačb smo reševali s programom za identifikacijo in simulacijo vodnih ekosistemov, AQUASIM (Reichert, 1998).

Parametre modela (stehiometrijske in kinetične) smo določili z izbiro intervala vrednosti vsakega parametra in z začetno vrednostjo. Nato smo z optimizacijsko funkcijo, ki je vgrajena v program AQUASIM, umerili vrednosti parametrov na podane meritve. Podrobnejša razlaga programa AQUASIM je podana v Reichert (1998).

A computer program for identification and simulation of aquatic systems AQUASIM Reichert (1998) was used to solve the equations of the model.

The parameters (stoichiometric and kinetic) were estimated by choosing the interval of possible values for each parameter and setting the initial value of the parameters. Then, we used the parameter estimation function (built in AQUASIM) to determine the optimal parameters' values, i.e. calibration of the model. For a more detailed description of the program see Reichert (1998).

## 5.2 MODELIRANJE Z REGRESIJSKIMI DREVESI

Regresijska drevesa smo uporabili za numerične napovedi merjenih koncentracij na iztoku iz ČN. Uporabljena podatkovna baza za učenje ima zelo malo merjenih atributov (poglavje 2.3). Zato je mogoče napovedovati le amonij ($NH_4$). Ker je v podatkovni bazi preveč zapisov za normalno procesiranje uporabljenega programa, smo 15-minutne zapise pretvorili v povprečne zapise, ki si sledijo z enournim korakom. Na ta način smo

## 5.2 MODELLING WITH REGRESSION TREES

The experiments with regression trees include numeric predictions of WWTP outflow concentrations. Since the data base comprises very few measured attributes (section 3.2) the only quantity that can be predicted here is the ammonia outflow concentration. In order to easier process the data we transformed the 15-minute time step records into hourly records by averaging the values. In this way we reduced the number of

zmanjšali število podatkov, ne da bi izgubili informacijo o časovni dinamiki sistema. Zapise v bazi smo uredili tako, da smo pri učenju modela zajeli vpliv hidravličnega zadrževalnega časa in upoštevali zgodovino. Vsak zapis v času t ima informacijo o vtoku pred tremi, šestimi in devetimi urami. Eksperiment smo nastavili tako, da iščemo model (regresijsko drevo) za napoved amonijevega dušika čez 6 (NH4-OUT6) ur, skladno s hidravličnim zadrževalnim časom. Uporabljeni atributi in razred, ki smo ga napovedovali, so podani v preglednici 2, kjer oznaka *–x* pomeni vrednost atributa pred *x* urami, *x* pa vrednost čez *x* ur.

records without losing the information about the system dynamics. Next, we rearranged the records in the data set, so that the hydraulic retention time was taken into account, i.e. the influence of the inflow data on the outflow data. Each record (at time t) carries the information about the inflow data before 3, 6, and 9 hours. We set the experiment so that we learn a model that predicts ammonium nitrogen for 6 (NH4-OUT6) hours in advance, accordingly with the hydraulic retention time. The attributes and the class are depicted in Table 2, where the tags *–x* denote values of the attribute from *x* hours ago and *x* value of *x* hours in advance.

# 6. REZULTATI

## 6.1 SIMULACIJA IZTOKA ČISTILNE NAPRAVE Z MODELOM ASM1

S poenostavljenim modelom ASM1 smo simulirali spremenljivke na iztoku iz obravnavane ČN, pri čemer smo enačbe reševali z orodjem AQUASIM. Rezultati petdnevne simulacije so prikazani na sliki 7 (simulacija koncentracije organskega onesnaženja, $BPK_5$), sliki 8 (simulacija koncentracije amonija) in sliki 9 (simulacija koncentracije nitrata). Vsaka slika prikazuje dva grafa: merjeno in simulirano vrednost spremenljivke na iztoku ČN.

## 6.2 REGRESIJSKO DREVO

Za izgradnjo regresijskih dreves smo uporabili programski paket WEKA, ki vključuje večino algoritmov strojnega učenja v eno okolje (Witten & Frank, 2000). Med njimi je tudi algoritem M5 (Quinlan, 1992) za gradnjo regresijskih dreves.

# 6. RESULTS

## 6.1 SIMULATIONS OF THE PLANT OUTFLOW QUALITY WITH ASM1

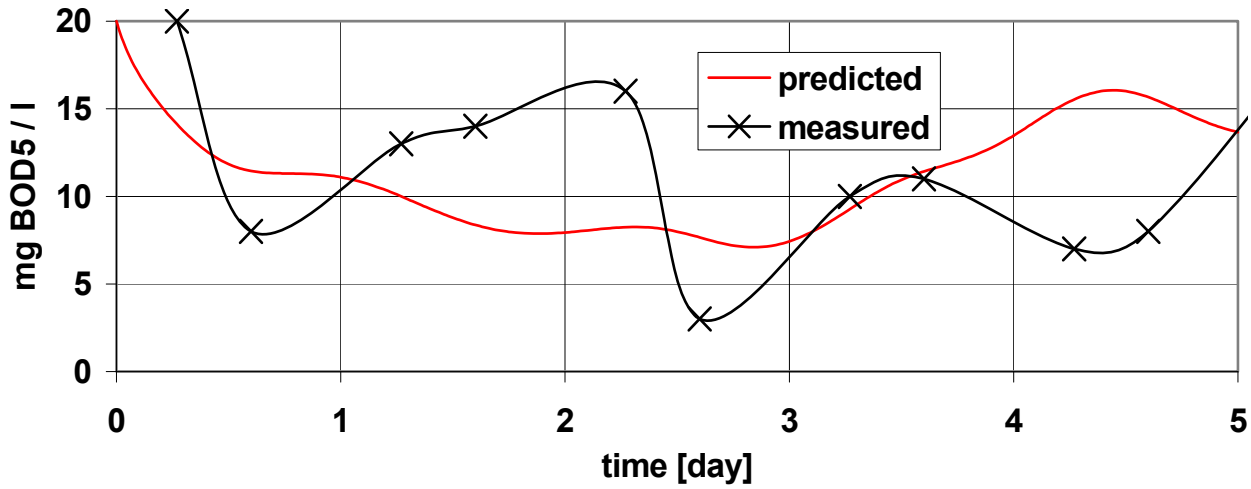
We performed the simulations of the plant outflow quality variables using the simplified ASM1 model. For solving the equations we used the AQUASIM computer program. The results of five-day simulations are shown on Figures 7 (simulation of BOD concentration), 8 (simulation of ammonia concentration) and 9 (simulation of nitrate concentration). Each figure shows two graphs: measured and simulated values of the variable on the plant outflow.

## 6.2 REGRESSION TREES

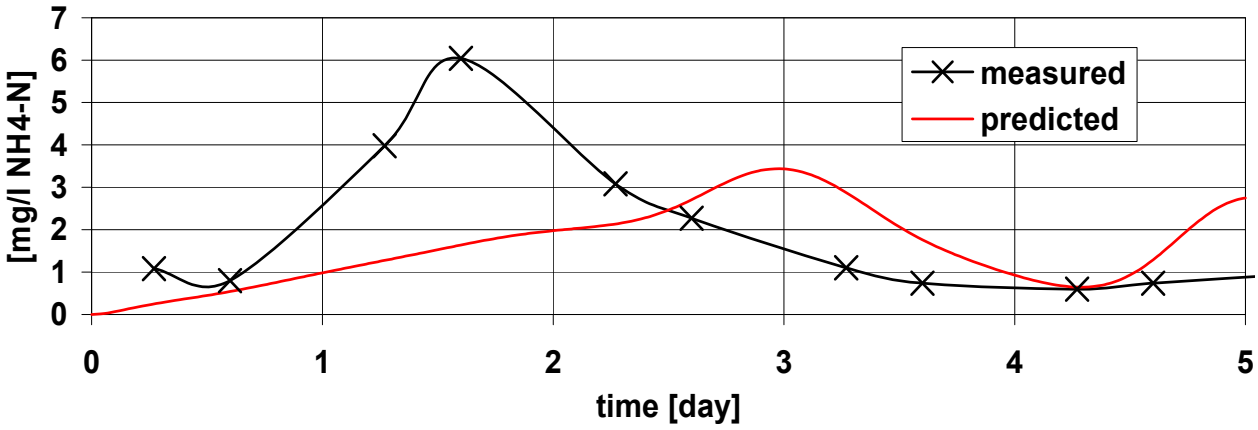
We used the WEKA package to build the models. WEKA incorporates most of the popular machine learning algorithms (Witten & Frank, 2000), including M5 (Quinlan, 1992) for building regression trees.

Preglednica 2. Atributi (merjene spremenljivke) za napoved amonija z regresijskim drevesom.
*Table 2. Attributes (measured variables) for ammonia prediction with the regression tree.*

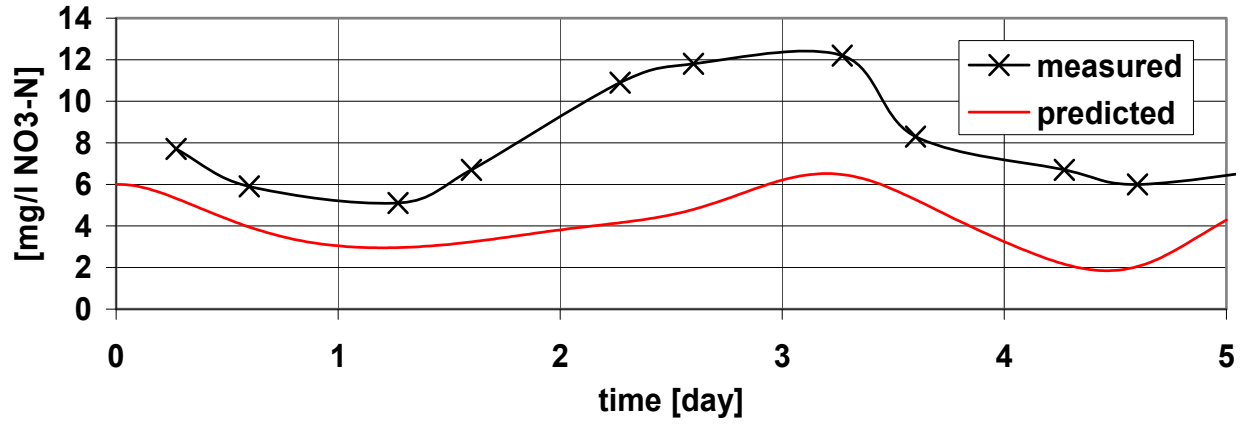
| Atributi (neodvisne spremenljivke)<br>*Attributes (independent variables)* | Razred (odvisna spremenljivka)<br>*Class (dependent variable)* |
|---|---|
| TEMP-3, TN-IN-3, TOC-IN-3, TEMP-6, TN-IN-6, TOC-IN-6, TEMP-9, TN-IN-9, TOC-IN-9 | NH4-OUT6 |




Slika 7. Petdnevna simulacija iztočne koncentracije $BPK_5$: merjena in simulirana vrednost.
*Figure 7. Five-day simulation of BOD5 concentration on the WWTP outflow: measured and simulated values.*




Slika 8. Petdnevna simulacija iztočne koncentracije amonija: merjena in simulirana vrednost.
*Figure 8. Five-day simulation of ammonium nitrogen concentration on the WWTP outflow: measured and simulated values.*



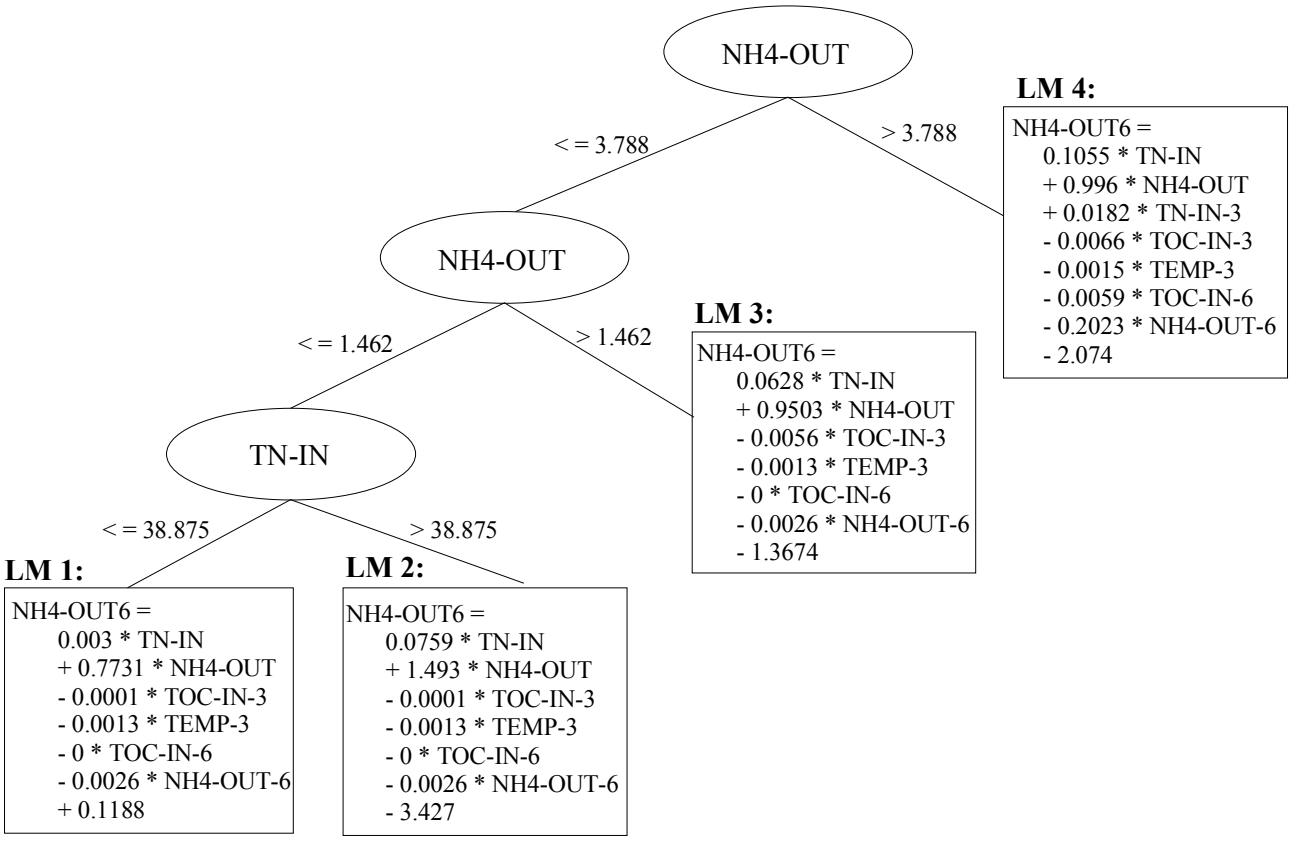

Slika 9. Petdnevna simulacija iztočne koncentracije nitratnega dušika: merjena in simulirana vrednost.
*Figure 9. Five-day simulation of nitrate nitrogen concentration on the WWTP outflow: measured and simulated values.*

Model, ki napoveduje iztočno koncentracijo amonija za 6 ur vnaprej, je prikazan na sliki 10. Model je sestavljen iz štirih linearnih enačb (LM1, LM2, LM3 in LM4) in treh vozlišč, v katerih nastopata dva atributa: NH4-OUT (trenutna koncentracija amonija na iztoku) in TN-IN (trenutna koncentracija totalnega dušika na vtoku).

Za napoved koncentracije amonija čez 6 ur je treba ustrezno izbrati linearno enačbo, glede na vrednosti atributov v vozliščih drevesa. Če je vrednost atributa NH4-OUT večja od 3,788, potem računamo po enačbi LM4. Če je ta vrednost med 3,788 in 1,462, uporabimo enačbo LM3, če pa je manjša od 1,462, preverimo še vrednost atributa TN-IN. Če je manjša od 38,875, velja enačba LM1, sicer pa LM2.

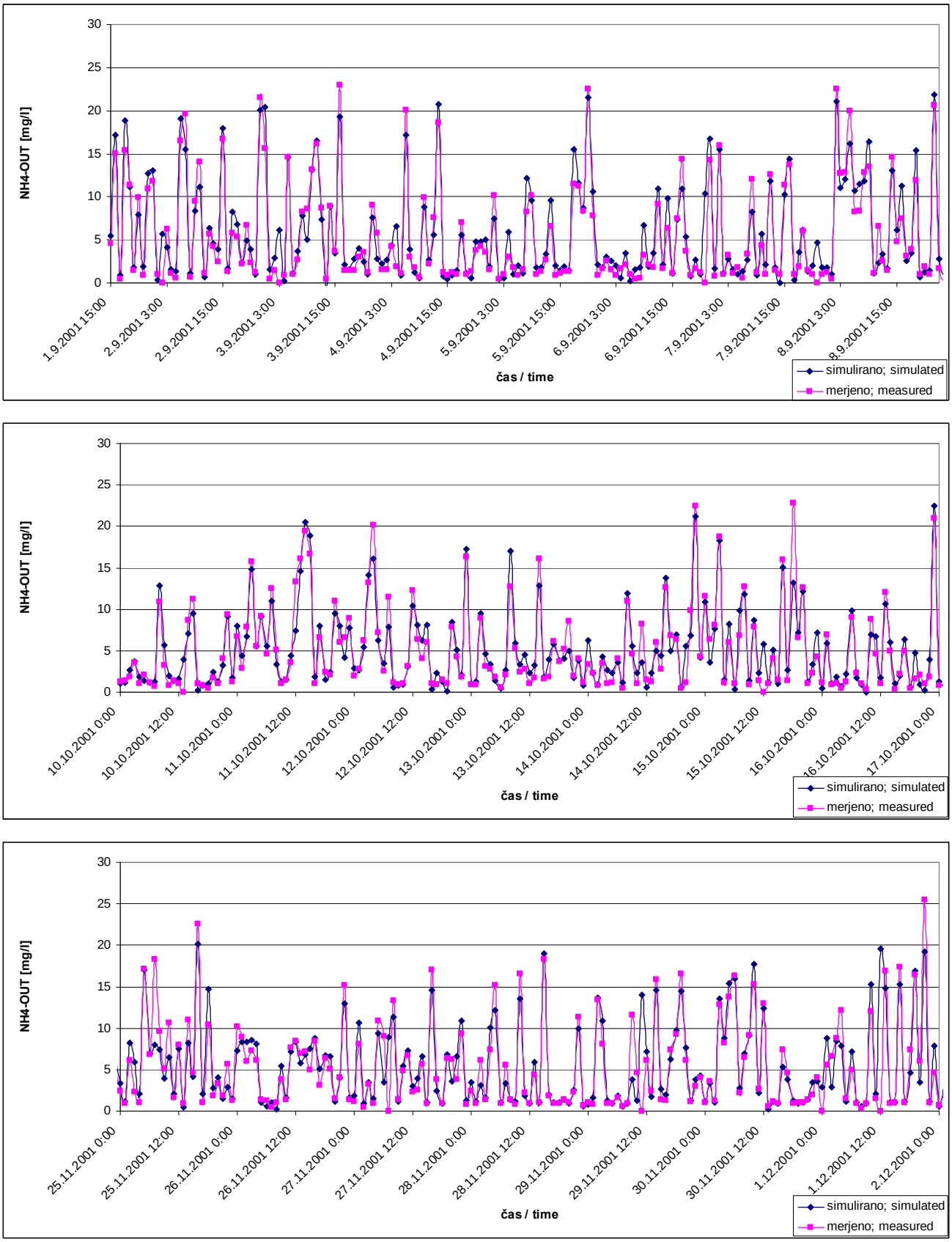
Figure 10 presents the model that predicts ammonia concentration in the effluent for six hours in advance. It is composed of four linear equations (LM1, LM2, LM3 in LM4) and three junctions, where two attributes can be found: NH4-OUT (present the concentration of effluent ammonia) and TN-IN (present the concentration of total nitrogen at the inflow).

In order to predict the ammonia concentration in 6 hours a suitable equation needs to be selected, regarding the attribute values in the junctions. If the value of NH4-OUT is larger than 3.788 we take the equation LM4. If this value is between 3.788 in 1.462, we use the equation LM3, and if it is less than 1.1462 then the TN-IN value needs to be checked. If it is less than 38.875 we take equation LM1, else we take LM2.




Slika 10. Model regresijskega drevesa, ki simulira koncentracijo amonija na iztoku čistilne naprave za šest ur vnaprej.

*Figure 10. Regression tree model for prediction of ammonia concentration at the WWTP outflow for six hours in advance.*




Slika 11. Simulacija modela, ki simulira koncentracijo amonija na iztoku čistilne naprave za šest ur vnaprej. Trije grafi prikazujejo različne dele simulacijskega obdobja.

*Figure 11. Performance of the model that predicts ammonia concentration at the WWTP outflow for six hours in advance. The three diagrams present different segments of the simulation period*

Model je bil validiran s postopkom navzkrižne validacije. Dosega visoko natančnost, saj je koeficient korelacije med merjenimi in napovedanimi vrednostmi 0,934. Slika 11 prikazuje simulirane vrednosti amonijevega dušika na iztoku v primerjavi z merjenimi. Zaradi preglednosti prikazujemo le del simuliranega obdobja, od 1. 9. 2001, 15:00, do 29. 12. 2001, 3:00, s korakom ene ure.

## 7. RAZPRAVA

Rezultati te raziskave jasno pokažejo, da je smiselno uporabljati različne pristope k modeliranju. Uporabili smo dva pristopa, tj. konceptualni pristop in modeliranje z uporabo strojnega učenja. Pri konceptualnem modeliranju smo aplicirali model ASM1 na sistem s nosilci. Tu se izkažejo pozitivne strani tovrstnega modeliranja. Brez težav smo model modificirali za modeliranje tehnologije z nosilci biomase, pri čemer smo privzeli določene poenostavitve, kot je recimo zanemarjanje procesov na samih nosilcih. Pri verifikaciji modela se ta smiselno odziva na spremembo parametrov in vhodnih spremenljivk, s čimer smo tudi upravičili uporabo ASM1 za naš sistem. Kljub temu je bilo umerjanje modela na meritve nezadovoljivo, kot je razvidno iz rezultatov (slike 7, 8 in 9). Nobena od izvedenih simulacij, tako koncentracije $BOD_5$, $NH_4$-N kot tudi $NO_3$-N, ne dosega zadovoljive natančnosti, za kar obstajata vsaj dva razloga. Prvi je ta, da kljub izvedbi ciljnih (in dragih) meritev, vseeno potrebujemo še dodatne meritve, in sicer pogostejše (saj se sistem spreminja tekom dneva, česar dva vzorca dnevno ne pokažeta) in za daljše obdobje. Drugi razlog pa bi lahko bil preenostaven model. Vendar pa se je treba zavedati, da bi kompleksnejši model, ki bi vključeval še procese na samih nosilcih, zahteval še več meritev.

Istočasno so na ČN potekale pogoste meritve (on-line) in za daljše obdobje, a žal le dušikovih spojin. Zaradi premalo merjenih količin so te meritve neuporabne za konceptualno modeliranje. Tu se izkaže uporabnost metod strojnega učenja. Z algoritmom M5 smo zelo uspešno zgradili regresijsko drevo (slika 10) za napoved iztočne koncentracije amonija za šest ur vnaprej. Model je zelo enostaven, razumljiv in

The model was evaluated by the cross-validation procedure. It has high accuracy of prediction, since the correlation coefficient between measured and predicted values is 0.934. Figure 11 presents the simulated values of ammonium nitrogen concentration together with the measured ones. For clarity reasons we only present three segments of the simulated period from 1/9/2001, 15:00, to 29/12/2001, 3:00, with one-hour step.

## 7. DISCUSSION

The results in this study reveal that it is very useful to tackle the modelling problem with different approaches. We applied two approaches to modelling of WWTP – conceptual modelling and modelling with ML tools. In conceptual modelling we adjusted the activated sludge model (ASM1) for modelling of the system with suspended carriers by taking some simplification into account, for example neglecting the biofilm processes. The model responds logically when conditions, parameters, and the forcing functions are changed. This suggests that the system with suspended carriers can be approximated with ASM1. However, none of the performed simulations (concentrations of $BOD_5$, $NH_4$-N and $NO_3$-N) achieves a satisfactory accuracy (see Figures 7, 8 and 9), which makes the model quite unreliable. There are at least two reasons for this. The first is unsuccessful calibration due to the measurements. In spite of the (costly) measurements that were performed to set the model, the sampling frequency was still not high enough (only twice a day) for model calibration. Also the sampling period was very short (one week). The second reason may be that the model was too simple. However, increasing the complexity of the model, like including biofilm equations, would require even more data for calibration.

At the same time on-line measurements of nitrogen compounds took place on the WWTP. Unfortunately these cannot be used for conceptual modelling because there are too many quantities missing. But machine learning appeared to be the right tool for this data set. By the application of the M5 algorithm we successfully built a regression tree model to

natančen.

Pri učenju iz podatkov je zelo pomembno, da podatkovni vzorec zajema dovolj različnih situacij, iz katerih se algoritem uči napovedovanja odvisne spremenljivke. Za pretvorbo dušikovih spojin je eden izmed pomembnejših faktorjev temperatura. Glede na to, da je obdobje naših meritev sep–dec, je zajeti razpon temperature od približno 20 do 8 °C. Torej so zabeleženi primeri delovanja ČN tako ob višji (ko je denitrifikacija običajno bolj učinkovita) kot tudi ob nizki temperaturi (ko je denitrifikacija običajno manj učinkovita). Prav tako je pomembno, da imamo merjene vse atribute, ki bi lahko vplivali na odvisno spremenljivko. Še zlasti je to pomembno, če želimo s strojnim učenjem odkrivati neke (očem in ekspertom) skrite vzorce v podatkih. Glede na to, da v naši bazi nastopa malo atributov, ne moremo pričakovati odkrivanja nekega novega znanja (saj se učimo iz podatkov!). Je pa model v skladu s pričakovanji domenskih ekspertov. Celotni niz podatkov se razdeli na smiselne podnize (glej razlago algoritma M5) glede na trenutne vrednosti atributov NH4-OUT in TN-IN. Za vsakega od teh podnizov je algoritem določil ustrezno linearno enačbo, v kateri nastopajo še ostali atributi, TOC in TEMP. Zgodovina je upoštevana do šest ur nazaj.

## 8. ZAKLJUČEK

V tej raziskavi smo uporabili dva pristopa k izgradnji matematičnega modela pilotne ČN, v odvisnosti od merjenih podatkov, ki so bili na razpolago. Meritve na ČN so obsegale dva podatkovna niza.

Prvi niz (merjena večina količin na ČN dvakrat dnevno), primeren za konceptualno modeliranje, smo uporabili za umerjanje poenostavljenega modela ASM1. Model je razumljiv ekspertom, a žal premalo natančen v napovedih. Za izboljšanje natančnosti modela potrebujemo bolj pogoste meritve in verjetno bolj kompleksno različico modela ASM1.

Drugi podatkovni niz obsega *on-line* meritve, a žal le dušikovih spojin. Uporabili smo ga za izgradnjo regresijskega drevesa za napoved iztočne koncentracije amonija za šest ur vnaprej. Model je enostaven, razumljiv in natančen v svojih napovedih. Žal večina

predict ammonia outflow concentration for 6 hours in advance.

When learning from data it is very important to have a data set that includes a wide variety of examples used by the algorithm to learn how to predict the concept. Temperature is one of the most important factors for nitrogen compounds conversions. In this case we capture the temperature range from 8 to 20 °C in the entire data set. Thus, we have examples of WWTP operation during high temperatures, when denitrification is more efficient, and during low temperatures with poor denitrification. It is also important that we measure all the attributes that might influence the dependant variable, especially if we are using ML for new knowledge discovery. Since very few attributes are measured in our data set it is not reasonable to expect the discovery of hidden patterns in the data. However, the model is in accordance with the expectations of the domain experts. The entire data set is divided in sub-sets with respect to the NH4-OUT and TN-OUT values. For these subsets the algorithm finds the suitable equation, composed of the rest of the attributes, i.e. TOC and TEMP. The history of six hours is taken into account.

## 8. CONCLUSIONS

In this paper we use two different approaches to build a mathematical model of a pilot WWTP. The use of each is dependant on the quality and quantity of the available data set. There were two data sets measured on the WWTP.

The first (measured most WWTP features twice a day), suitable for conceptual modelling, was used to calibrate a simplified version of the ASM1 model. The model is transparent and clear to the experts, but not accurate in its predictions. To improve the model accuracy we need more frequent measurements and probably a more complex version of ASM1.

The second data set, comprising on-line measurements of (only) nitrogen compounds was used to build a regression tree model for ammonia effluent prediction for six hours in

ključnih količin, ki so potrebne za opis delovanja ČN, ni merjena v tem podatkovnem nizu.

Nadaljnje delo je torej usmerjeno k razširitvi podatkovnega niza s temi količinami. S takim podatkovnim nizom bi lahko:

- dodatno izboljšali natančnost modela,
- zgradili modele za ostale spremenljivke, kot je totalni dušik ali KPK, in ne le za amonij, in
- odkrivali nove zveze in vzorce med podatki.

advance. The model is simple, understandable and accurate in predictions. Unfortunately, some crucial features for describing the WWTP operation are missing in the data set, used for building regression trees.

Thus future work is aimed at the collection of a more comprehensive data set in order to:

- improve the predictive power of the model,
- to build models for other features, such as total nitrogen, and
- (3) reveal new relations among the measured features.

## VIRI – REFERENCES

Atanasova, N., Kompare, B. (2002). Uporaba odločitvenih dreves pri modeliranju čistilne naprave za odpadno vodo = The use of decision trees in the modelling of a wastewater treatment plant, *Acta hydrotechnica* **20**(33), 351–370.

Baeza, J., Gabriel, D., Lafuente, J. (1999). An expert supervisory system for a pilot WWTP, *Environmental Modelling and Software* **14**, 383–390.

Belanche, L. I., 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.

Checci, N., Marsili-Libelli, S. (2005). Reliability of parameter estimation in respirometric models, *Water Research* **39**, 3686–3696.

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

Comas, J., Rodríguez-Roda, I. R., Sànches-Marré, M., Cortés, U., Freixó, A., Arráez, J. Poch, M. (2003). A knowledge-based approach to the defloculation problem: integrating on-line, off-line, and heuristic information, *Water Research* **37**, 2377–2387.

Gernaey, K., van Loosdrecht, C. M. M., Henze, M., Lind, M., Jørgensen, B. S. (2004). Activated sludge wastewater treatment plant modelling and simulation: state of the art, *Environmental Modelling and Software* **19**, 763–783.

Grieu, S., Traoré, A., Polit, M., Colprim, J. (2005). Prediction of parameters characterising the state of a pollution removal biologic process, *Engineering Applications of Artificial Intelligence* **18**, 559–573.

Henze, M., Harremoës, P., Jansen, J. l. C., Arvin, E. (1997). *Wastewater Treatment: Biological and Chemical Processes*, 2nd Ed, Springer-Verlag, Berlin, ISBN 3-540-62702-2.

Henze, M., Grady, C. P. L. Jr., Gujer, W., Marais, G. v. R., Matsuo, T. (1986). Activated Sludge Model No. 1. Scientific and Technical Reports No. 1. International association on water pollution research and control. IAWPRC task group on mathematical modelling for design and operation of biological wastewater treatment processes.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M. (2000). Activated Sludge Models ASM1, ASM2, ASM2d, and ASM3. Scientific and Technical Report No. 9. IWA task group on mathematical modelling for design and operation of biological wastewater treatment, ISBN 1-900222-24-8.

Henze, M., Mino, W., Gujer, W., Wentzel, M.C., Marais, G. v. R., and Matsuo, T. (1995). Activated Sludge Model No. 2. Scientific and Technical Report No. 3. IAWQ task group on mathematical modelling for design and operation of biological wastewater treatment processes,

ISBN 1-900222-00-0.

Iacopozzi, I., Innocenti, V., Marsili-Libelli, S., Giusti, E. (2007). A modified Activated Sludge Model No. 3 (ASM3) with two-step nitrification-denitrification, *Environmental Modelling and Software* **22**, 847–861.

Petersen, B., Gernaey, K., Henze, M., Vanrolleghem, A.P. (2002). Evaluation of an ASM1 model calibration procedure on a municipal-industrial wastewater treatment plant, *Journal of Hydroinformatics*, **04**, **1**, 15–37.

Quinlan, J. R. (1992). Learning with continuous classes. Proceedings AI'92 (Australian Conference on AI), Singapore, World Scientific, 343–348.

Reichert P. (1998). AQUASIM 2.0 – User Manual: Computer Program for the Identification and Simulation of Aquatic Systems, ISBN: 3-906484-16-5.

Roda, I. R., Comas, J., Sànches-Marré, M., Cortés, U., Lafuente, J., Poch, M. (1999). Expert system development for a real wastewater treatment plant. Chemical Industry and Environment III, Proceedings. Kraków, Poland, 653–660.

Sànches-Marré, M., Cortés, U., Lafuente, J., Roda, I. R., Poch, M. (1996). DAI-DEPUR: a distributed architecture for wastewater treatment plants supervision, *Artificial Intelligence in Engineering* **10**(3), 379–423.

Tschobanouglos, G., Burton, L. F. (1991). *Wastewater Engineering: Treatment Disposal and Reuse*, 3rd Ed., Metcalf & Eddy, Copyright McGraw–Hill, USA, ISBN 0-07-041690-7.

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

## Naslovi avtorjev – Authors' Addresses

*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

*mag. Meta Levstek*
Centralna čistilna naprava Domžale-Kamnik –
Central Waste Water Treatment Plant Domžale-Kamnik
Študljanska 91, SI-1230 Domžale, Slovenia
E-mail: levstek@ccn-domzale.si

*asist. dr. 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