aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes

Karina Gibert [a,*] and Dante Conti [b,c]

[a] *Knowledge Engineering and Machine Learning group, Department Statistics and Operation Research, Universitat Politècnica de Catalunya–BarcelonaTech, Barcelona, Spain*
E-mail: karina.gibert@upc.edu
[b] *Department of Operations Research, Universidad de Los Andes, Mérida, Venezuela*
[c] *Instituto de Matemática, Estadistica e Computacao Cientifica, Departamento de Matemática Aplicada, U. Estadual de Campinas, Sao Paulo, Brasil*

Abstract. Clustering techniques find homogeneous and distinguishable prototypes. Careful interpretation of these prototypes is crucial to assist the experts to better organize this know-how and to really improve their decision-making processes. The Traffic Lights Panel was introduced in 2009 as a postprocessing tool to provide understanding of clustering prototypes. In this work, *annotated Traffic Lights Panel (aTLP)* is presented as an enrichment of the TLP to manage the intrinsic uncertainty related with prototypes themselves. The aTLP handles uncertainty through a quantification of the prototypes' purity based on the variation coefficients (VC) and an associated color-based uncertainty model, with two dimensions – tone and saturation – representing nominal trend and purity of the prototype. An application to a waste-water treatment plant in Slovenia, in a discrete and continuous approach, suggests that aTLP seems a useful and friendly tool able to reduce the gap between data mining and effective decision support, towards informed-decisions.

Keywords: Annotated Traffic Lights Panel, clustering, decision-making, KDD, prototypes interpretation, Traffic Lights Panel, uncertainty, waste-water treatment plant

1. Introduction and related work

It is well known that Knowledge Discovery in Databases (KDD) approach provides a good framework to analyse complex phenomena [17]. The core of KDD addresses to get novel and valid knowledge which can improve the *corpus doctrinae* of the target domain [4]. Fayyad's proposal marked the beginning of a new paradigm where both prior and posterior analysis becomes important: "*Most previous work on KDD has focused on [. . .] Data Mining (DM) step. However, the other steps are of considerable importance for the successful application of KDD in practice*" [4]. Indeed, posterior analysis or postprocessing tasks are crucial to complete the interpretation of data mining results and establish proper bridges between data-driven models and real decision making activities [3,13]. This seems to be in the core of current interests for the scientific community. Very recently, Alex "Sandy" Pentland, the head of MediaLab Entrepreneurship Programm MIT declared the need "*to be general literacy about data interpretation*" in his keynote of Campus Party Europa September 4th 2013 [20], where he stressed the importance of big data and the urgency to overcome the current lack of properly formed data scientists, as other scientists also claim [21].

Water Management in general and waste-water treatment in particular are among those phenomena that can benefit from KDD, as classical modelling approaches perform poorly due to the complexity of the related phenomena. The KDD approach might help to better understand waste-water treatment plants (WWTPs) processes, which is crucial to protect the environment [7]. An efficient WWTP should guarantee the effluent water quality (as well as fitting to legal requirements and government policies), in order to re-

*Corresponding author. E-mail: karina.gibert@upc.edu.

store the natural environmental balance which is disturbed by industry wastes, domestic waste-waters, etc. The process used to achieve this goal is really complex and delicate; on the one hand, because of the intrinsic features of waste-water treatment, whose correct behaviour depends on a number of numerical and non-numerical factors; on the other hand, because of the bad environmental consequences of an incorrect management of the plant [6]. When the plant is not operating under normal conditions, decisions have to be taken to modify some parameters of the waste-water treatment process in order to re-establish the normality as soon as possible. The management of these processes highly relies on the expertise of the decision-maker, even if he is currently well supported by monitoring data.

Learning a reduced set of typical situations that can be found in a WWTP might help to make better decisions and can contribute to standardisation of treatment protocols. Clustering techniques are, in fact, one of the most frequently used KDD tasks in real applications [14,19] and are useful to identify such a set of typical situations. Traditionally, the clustering results are expressed as a partition of the set of elements to be clustered. So, several groups of objects are listed as final result. The analyst is the responsible to identify the particularities of every group to assist the expert discovering the underlying clustering criteria that will allow a semantic interpretation of the resulting classes. This interpretation process is the key to obtain effective, valid and useful knowledge, typical situations happening in WWTP that might help to better organize the background domain knowledge and, as a consequence, to improve the associated decision-making processes. The expertise of both the end-user and the data miner are required for this purpose (Fig. 1). It is known that unless end-users/experts understand and trust data mining results, they are reluctant to use them in their daily decision-making [5,15,16,18]. Research oriented towards improving the interpretation processes in data mining contexts will contribute to guarantee the impact of knowledge extracted from data in the target domain. Although few authors pay attention to that topics, post-processing are among the tools that can play this role, by reducing the gap between data mining and effective decision support [2,3,13].

Traffic Lights Panel (TLP) is a symbolic postprocessing of the clustering results [11] proved extremely useful and well-accepted by domain experts in several real applications [8–10]. TLP exploits the association between the traffic light colours and the general trend of the variables in every class to help the expert to understand the clusters and to support the conceptualization of the discovered classes. Going further, authors have been working on the automatic construction of the TLP [8,9]. However, being TLP a symbolic representation of the prototypical patterns characterizing the classes, uncertainty propagation to final decisions is intrinsically involved, since the prototypes describe central class trends, by disregarding individual deviations from the main patterns. Indeed, the main principle governing the original TLP construction was to identify 3 qualitative labels over the variables representing the central trend of every variable inside every class and associate them to the traffic lights colors (red, green and yellow) accordingly. This construction is abstracting the contents of the class into a single indicator related to the central trend of the class, which might be too reductionists, by ignoring whereas the class is more or less homogeneously distributed around its central trend.

The *annotated TLP* (*aTLP*) was introduced in [8] in order to mitigate this effect. The aTLP enriches the original TLP with the uncertainty associated to the

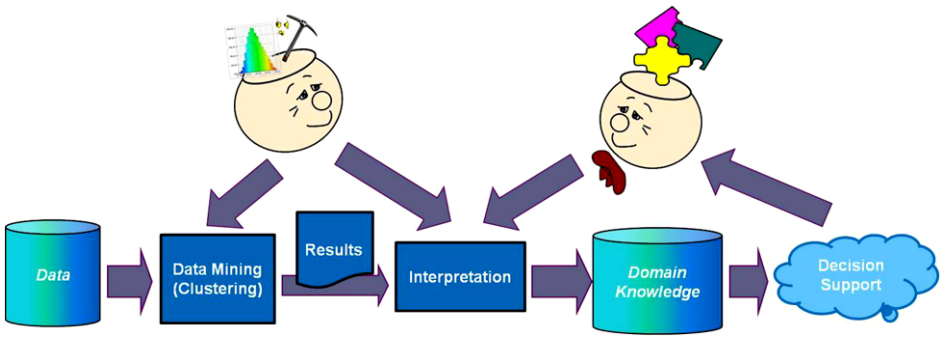

Fig. 1. Interpretation support tools bridging the gap between data mining and decision support. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

class prototypes. The variability within classes is represented by means of a degradation of the basic TLP colours, guided by a quantification of the heterogeneity of the class. In our first approach variation coefficients (VC) is used as a quantifier of heterogeneity. In the original formulation presented in [8], a cut-off over the VC is used to parameterize the aTLP and three levels of uncertainty are represented by using three levels of darkening the basic colors. This parameterization in three levels of darkening is useful, as the human eye cannot distinguish small variations on the light of a color, but seems somehow artificial and might be avoided. So, in this work a continuous approach for the uncertainty color-based modelling used to enrich the TLP is provided and a refinement of the original aTLP into a non-parametric tool is presented. Both TLP and aTLP either in the parametric or continuous form appear as useful and friendly tools to help the experts to understand the discovered profiles.

In this paper, an application to understanding of typical situations in a waste water treatment plant in Slovenia is presented. The paper has the following structure: Section 2 describes the application domain and previous work; Section 3 addresses the intrinsic ambiguity of TLP related with variability within classes; then in Section 4 the annotated TLP (aTLP) is presented as a tool to manage uncertainty depending on two parameters k_1, k_2 that the user has to choose according to the cost associated to wrong decisions in the target domain. In Section 5 the non-parametric version of the aTLP is introduced. The results regarding the Slovenian WWTP are presented and discussed along the paper after introducing the corresponding concepts. The paper ends with conclusions and future works in Section 6 where the impact, advantages and drawbacks of TLP and both parametric and non-parametric versions of aTLP are analyzed.

2. Application domain and previous work

This research regards a waste-water treatment plant (WWTP) placed in Slovenia, the Domzale-Kamnik waste-water treatment plant which is one of the largest Slovenian plants in operation (200,000 PE), treating municipal and industrial waste-water from four municipalities. It is placed near Ljubljana, the capital of the country and the receiving body is river Kamniska Bistrica. The waste-water treatment has become crucial and leader for the Slovenian environment protection. In Domzale-Kamnik, there is a second line where pilot methodologies are tested. In this case, the dataset is composed by a sample of 365 observations coming from the pilot plant during the period June 2005–May 2006, when a new technology (moving bed biofilm removal MBBR) was being tested in the process of upgrading the waste water treatment to also include nitrogen removal (see Fig. 2). The records represent the daily averages of 16 variables considered relevant by the experts (see Table 1). Those measurements are computed as the daily averages of the data recorded every hour (24 observations per day) by the monitoring system of the plant.

In [9] this dataset has been clustered and 4 classes were identified and validated by the experts. The association between the values of the variables and the traffic lights colours is made by taking into account the meaning of the variables themselves, and according to that, the experts provided the *polarity semantics table* (*PST*) where direct or reverse sense of the association is assigned depending on the semantics of the variable. In this particular case of study, a latent concept related with water quality or with the good operation of the plant permits to associate red colours to highest concentration of pollutants in water or non-efficient oper-

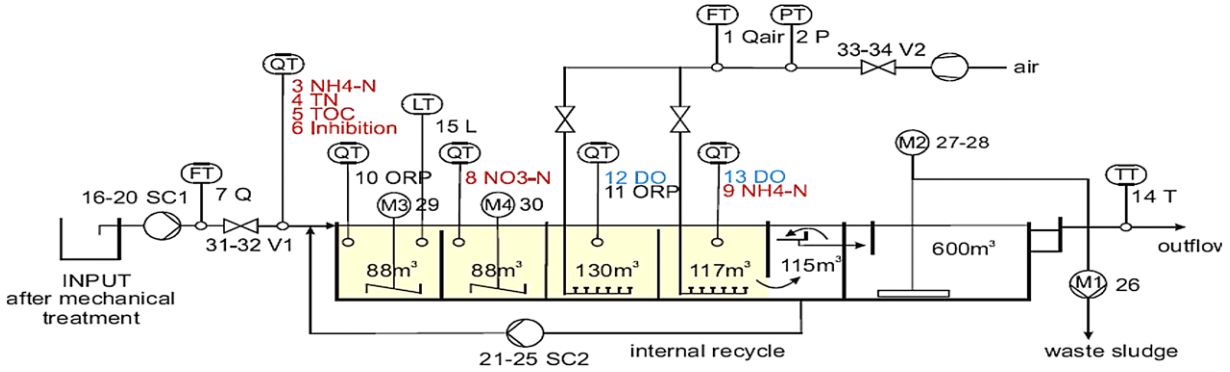

Fig. 2. Architecture of the Domzale-Kamnik pilot plant. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

Table 1
List of variables measured in the Domzale-Kamnik pilot plant

Phase of the process	Variable	Unit	Description
Influent	NH_4-influent	mg/l	Ammonia concentration at the influent of the pilot plant
	Q-influent	m^3/h	Wastewater influent flow rate
	FRI-DOTOK-20S	Hz	Frequency of the influent flow rate meter
	TN-influent	mg/l	Concentration of the total nitrogen at the influent of the pilot plant
	TOC-influent	mg/l	Total organic carbon concentration at the influent of the pilot plant
	Nitritox-influent	mg/l	Measurement of the inhibition at the influent of the pilot plant
2nd anoxic reactor	h-waste-water	m	Height of the waste-water in the reactor
1st aerobic reactor	O_2-1 aerobic	mg/l	Dissolved oxygen concentration in the 1st aerobic reactor
	Valve-air	%	Openness of the air valve in percentage
	Q-air	m^3/h	Total air flow that is dosed in both aerobic reactor
2nd aerobic reactor	NH_4-2 aerobic	mg/l	Ammonia concentration in the second aerobic reactor
	O_2-2 aerobic	mg/l	Dissolved oxygen concentration in the 2nd aerobic reactor
Effluent	TN-effluent	mg/l	Concentration of the total nitrogen at the effluent (outflow) of the pilot plant
	Temp-waste-water	°C	Temperature of the waste-water
	TOC-effluent	mg/l	Total organic carbon concentration at the effluent of the pilot plant
Other	Freq-rec	Hz	Frequency of the internal recycle flow rate meter (internal recycle flow rate)

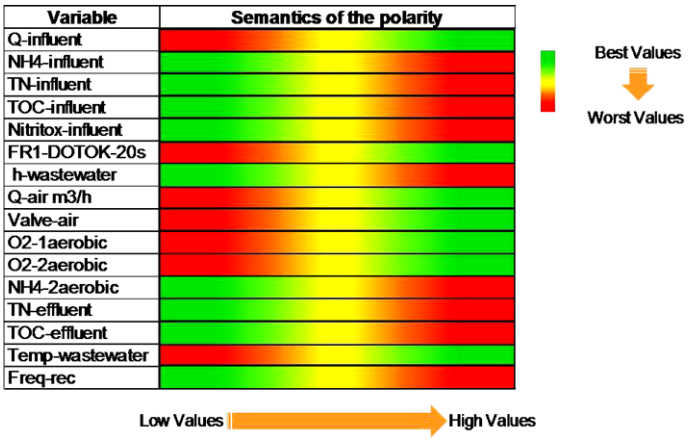

Fig. 3. Polarity semantics table for the set of variables considered in the WWTP pilot plant. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

ation of the plant, depending on the variable. Figure 3 shows the assignment for the case study.

An automatic procedure to build the TLP based on conditional Medians of variables with respect to classes was used and the information transferred to the system through the PST was integrated in the coloring process [9]. The result is the TLP shown in Fig. 4, where the quality of the water typically represented in every class can be suddenly understood.

In previous works, it was realized that the TLP constitutes a symbolic abstraction of the classes providing information closer to the expert knowledge and making easier the process of recognizing the concepts represented by the profiles. The conceptualization process performed by the experts upon the TLP from Fig. 4 provided the following description of the classes:

C353: Represents the plant operation under the high load. Influent nitrogen concentrations are high and influent flow rate is quite high as well. Even though the oxygen concentration in the aerobics reactors is high, aeration in 1st aerobic tank is higher than other classes;

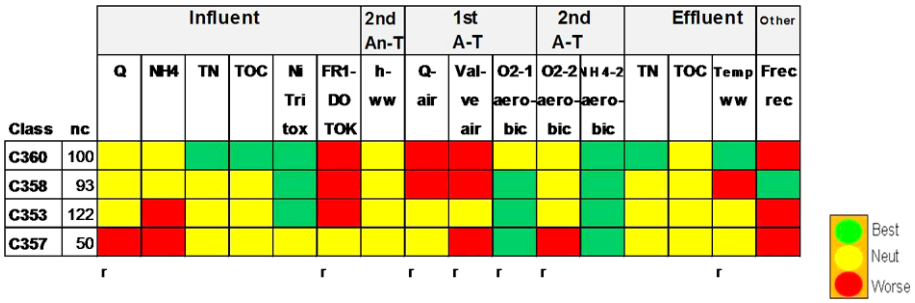

Fig. 4. Complete TLP for the whole set of variables considered in the WWTP pilot plant. (*Note*: for black and white versions, darker cells correspond to red, lighter to yellow and intermediate to green). (The colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

still the effluent nitrogen concentrations remain medium. It means that when the plant is overloaded non-low effluent concentrations at the effluent of the plant can be expected.

C357: Represents the situation of high load when the influent flow rate is low, that is, when the hydraulic retention time of the plant is high. In this case, middle effluent nitrogen concentration is obtained even if oxygen concentration in the 2nd aerobic tank is the lowest and internal recycling rate is higher than other classes. This means that when the influent flow rate to the plant is low the effluent concentrations of the plant could be obtained at the low level if the oxygen concentration in the aerobic tanks is high.

C358: Explains the situation when the waste-water temperature is low and internal recycling rate is low. In this case nitrogen removal efficiency of the plant is still not enough. This is happens because microorganisms in the reactors do not work so intensively in cold conditions and therefore non-low concentrations of the nitrogen at the effluent of the plant can be expected.

C360: Shows the situation when the waste-water temperature is high. In warmer conditions the microorganisms in the plant work faster, so the effluent nitrogen concentrations can be low even when the oxygen concentrations in the 1st aerobic tank is not so high as in other classes.

3. The intrinsic uncertainty associated to TLP

As said before, TLP was introduced in [11] and displays the qualitative dominant levels of a set of variables through a set of classes. TLP is based on the identification of qualitative levels on the variables. When 3 levels are used, it is very interesting to assign the colours of a traffic light to those levels, according to some latent concept (red for the negative value, yellow for the medium or neutral value and green for the positive value), being the negative values the higher or lower values of the variable depending on the variable's semantics expressed in the polarity semantics table. It is important to keep red and green as non-verbal codes for positive and negative values to connect with the expert's implicit codes for interpreting. Thus, in the presented application, red is associated to lower water quality or worse plant operation and in Fig. 4, class C360 is composed by those days where the concentration of nitrogen at the influence tends to be, *on average*, lower than in other classes.

TLP assigns colours to the cells taking only into account central trends presented in the classes by disregarding individual deviations from main patterns. The TLP is in fact a symbolic abstraction of the class panel graph (CPG) a previous tool introduced in [12] which displays the conditional distributions of the variables against classes. The abstraction provided through the TLP is much closer to the interpretation codes of a non-technical expert, like a physician, biologist, chemical engineer or psychologist than CPGs and permits a better class-conceptualization process. However, as TLP ignores variability within classes, it inherently introduces some uncertainty into the interpretation that is propagated to the later decision-making process. To better illustrate this limitation, Fig. 5 shows the CPG corresponding to the target application.

The ambiguity is associated to the meaning of the yellow colour. As an example, observe that in Fig. 4, yellow colour has been assigned to variable Temp-ww for both classes C353 and C357. Figure 5 shows

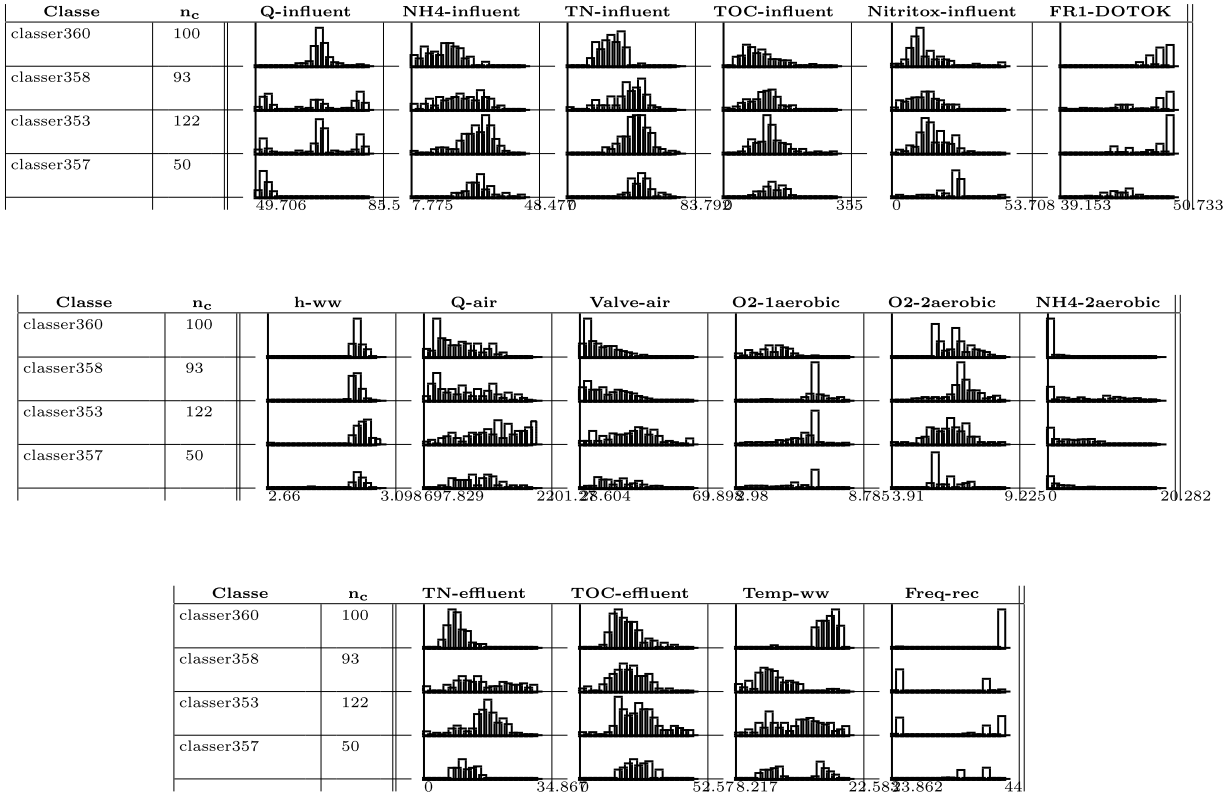

Fig. 5. Class panel graph for the whole set of variables in WWTP pilot plant.

the CPG where the local distribution of the variables within classes can be inspected. Getting the details provided by the CPG, it can be seen that C353 contains values distributed all along the range of the variable, so, for this case, the corresponding yellow cell in the TLP indicates that one can find all kind of temperatures in C353. For C357 situation is slightly different, a clear bimodal distribution is found there, giving an intermediate class average, located in a place where no real observations are found. The corresponding yellow cell should mean either low or high temperatures, but never intermediate. Another ambiguity appears in Q-air for classes C353 and C357: while C353 provides a yellow cell in the TLP indicating no clear trend in the class, whereas C357 provides a yellow cell really indicating a group of days where quite homogeneous intermediate values appear. The yellow cell of Nitritox-influent for class C357 is representing an even more homogeneous distribution around intermediate values. This phenomenon is somehow observable also for green and red cells. Looking at the TLP, as it was originally conceived in [11], one can learn that TN-Influent has smaller values in C360 than in other classes; the same situation happens for variable O2-aerobic. However, the uncertainty associated to both variables is radically different, as overlapping of local class-distributions is very small in O2-aerobic, whereas this is not the case for TN-influent. As a consequence, decisions related to low TN-influent in class C360 will involve higher risks than those associated with low O2-aerobic.

This arises the limitation that the original conception of TLP did not give enough information to disambiguate the risk associated to the colours interpretation.

4. The annotated TLP (aTLP): Uncertainty management

The annotated TLP (aTLP) was developed by the authors [8] as an enrichment of the original TLP that added to the representation sufficient information about the uncertainty associated to the class prototype, so that the previously related ambiguities disappeared. First of all a quantification of the homogeneity of the local distributions inside the classes is required. The variation coefficient conditioned to the classes is taken as a measure of uncertainty, because it is more robust than standard deviations or variances

and it behaves adimensionally, even if it is not upper-bounded. Given a numerical variable X_k, and a partition $\mathcal{P} = \{C_1, \ldots, C_\xi\}$ of the $\mathcal{I}$ set of target individuals, and being $s_{X_k|C}$ the standard deviation of X_k locally to class C and $\bar{X}_k|C$ the mean of X_k taken inside class C, the conditioned variation coefficient is:

$$\mathrm{VC}_k|C = \frac{s_{X_k|C}}{\bar{X}_k|C}. \quad (1)$$

VC is expected to perform better than direct variability measures like classical standard deviation (s_k) since VC is a normalized coefficient, adimensional, with a common interpretation of its values for all the variables. This permits to establish a single gradation of variability levels for all variables (from low to high variability). The aTLP is based on darkening basic colours cells according to VC levels. Darker colours will be associated to less pure classes (with higher VC, and darker colours). The degree of darkening can be determined upon the VC associated to the cell.

In the original formulation of aTLP, three levels of darkening where considered, by associating pure colours to low VC, slightly darkened colours to medium degrees of VC and strong darkened colours to high levels of VC, as it can be seen in Fig. 6. The first row of the figure shows the original pure colors used in the classical TLP, whereas the colors in second and third row where determined by finding darkened colors that were clearly distinguishable by human eyes

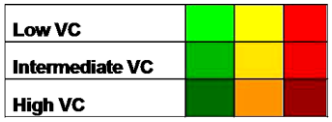

Fig. 6. Basic rank for gradation in the aTLP (in relation to VC levels). (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

without loosing the character of the three colors considered. This association implicitly required the definition of a cut-off over the VC values in three intervals that associated to low, middle and high variation levels. For this cut-off, two cutpoints (k_1, k_2) on the VC where required. The association worked such that left hand side VC interval ($\mathrm{VC} \leqslant k_1$) associates with basic TLP colour; central VC interval ($\mathrm{VC} \in (k_1, k_2)$) associates with light darkening and right hand side VC interval ($\mathrm{VC} \geqslant k_2$) associates with hard darkening [1].

The key concept is that the darker a cell is, the higher the heterogeneity of class individuals in that variable, and the more uncertainty about the standard behaviour of that variable in that class is propagated. In general, a darker aTLP can be associated with higher uncertainty related to the final profiles derived. This tells the experts that the induced profiles can involve high individual variability and decision-making should be corrected accordingly. In other words, strong decisions should be justified on the basic colour cells (pure classes) whereas major caution is required when decisions involve darker cells. For the target application, the variation coefficients of all cells are computed and displayed in Table 2.

Rules for determining the darkening of the TLP cells depend on the parameters k_1, k_2. Figure 7 contains the resulting aTLP for three different parameterizations, that go to a more conservative scenario towards the bottom of the figure. The aTLP$(0.3, 0.9)$, in the top figure, displays basic original colours in all cells with $\mathrm{VC} < 0.30$, hard darkening in all cells with $\mathrm{VC} > 0.90$, and slight darkening in between. Comparing the aTLP$(0.3, 0.9)$ with the original TLP displayed in Fig. 4 one can observe firstly that red cells remain as pure colours, indicating that low variability is associated to worse performance of the plant in all the variables. NH4-aerobic is the only variable displaying darker cells, those with higher degrees

Table 2
Variation coefficients (VC) associated to TLP shown in Fig. 4

Class	mc	Influent						Aerobic Tanks 1 and 2 – Anoxic Tank 2						Effluent			Other
		Q	NH_4	TN	TOC	Nitrite-tox	FR1 DO-TOK	h-ww	Q-air	Valve air	Q_2-1 aerobic	Q_2-2 aerobic	NH_4 aerobic	TN	TOC	Temp ww	Freq-rec
C360	100	0.04	0.31	0.24	0.43	0.54	0.02	0.01	0.24	0.15	0.17	0.11	2.48	0.28	0.30	0.06	0.04
C358	93	0.18	0.31	0.28	0.33	0.46	0.06	0.00	0.29	0.17	0.08	0.10	0.95	0.45	0.34	0.17	0.25
C353	122	0.14	0.19	0.19	0.31	0.34	0.04	0.01	0.23	0.19	0.14	0.12	0.86	0.28	0.34	0.22	0.24
C357	50	0.03	0.14	0.13	0.20	0.25	0.02	0.01	0.17	0.15	0.18	0.09	1.33	0.22	0.20	0.22	0.06
Global VC		0.16	0.31	0.28	0.35	0.46	0.05	0.01	0.29	0.21	0.19	0.12	1.41	0.43	0.33	0.23	0.21

of variability, indicating that it is better not to use this variable for decision-making. Caution is required on decisions made upon influent-NH4; influent-TOC, influent-Nitritox and Effluent TOC as a majority of cells on those variables have non-neglectable degrees of uncertainty. Influent Q, Influent-TN, influent FR1-DOTOK, all variables from 1st Aerobic Tank, O2-aerobic in the 2nd Aerobic Tank, temperature on the effluent and Freq-Rec are the variables that show reliable information and permit decisions with small risk.

However, according to the cost of wrong decisions, one could decide to be more or less conservative for the aTLP by controlling the cut points on VC levels. Thus, a more conservative aTLP would be obtained whereas the cut points are decreased. Thus, the central TLP in Fig. 4 is the aTLP$(0.1, 0.5)$, and depicts a scenario where the risk of decisions is moderate for variables with VC > 0.10 and high when VC > 0.5. Here, intermediate uncertainty levels dominate. Only two variables, FR1-DOTOK and h-ww, remain with pure colours. Q-influent and temperature in C360, 2nd Aerobic Tank-oxygen for C358 and Q and oxygen in 2nd aerobic tank for C357 keep sufficient purity to support non-risky decisions. Remaining variables increased their colour gradation going to light darkening colours. However, in this scenario, hard darkening is still scarce; indicating that high risk of decisions is still limited.

At the bottom of Fig. 7, an even more restrictive situation is provided, with the aTLP$(0.1, 0.3)$, which would represent a situation in which decisions cannot be based upon variables registering VC > 0.3 in some class, and obviously, this aTLP looks much more darker than the others. In this scenario, hard darkening is much present, indicating that no decisions should be taken regarding variables NH4, TOC or Nitritox at influent, NH4-2-aerobic, TOC-effluent and moderate risk will be associated to decisions regarding all other variables, except for 4 isolated cells in the aTLP.

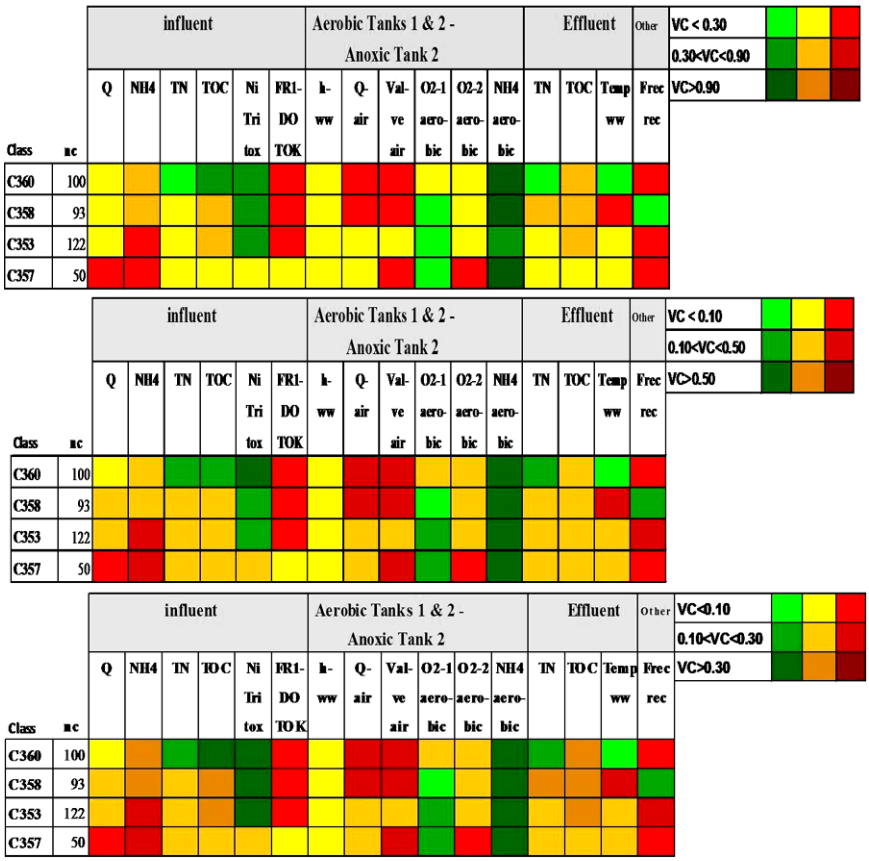

Fig. 7. Annotated TLP (aTLP) for different risk scenarios: (*top*) aTLP$(0.3, 0.9)$; (*center*) aTLP$(0.1, 0.5)$; (*bottom*) aTLP$(0.1, 0.3)$ (*Note*: colours gradation should be interpreted in the same way as in TLP, Fig. 3.) (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

Whereas the experts would consider that VC > 0.30 implies a non-assumable cost for wrong decisions, the aTLP$(0.1, 0.3)$ should be used as a reference. The aTLP provides a tool to visualize the risk of decisions depending on two parameters k_1, k_2; values of parameters should be chosen according to the assumable risk of wrong decisions in every particular application.

5. A continuous non-parametric approach for aTLP

However, determining k_1, k_2 is introducing some kind of arbitrariness into the aTLP that can be overcame by finding a model providing a continuous gradation of the colors on the basis of the variability indicator that moves the three levels of darkening proposed in previous section to a continuous darkening scale.

In fact, the darkening is performed by *de-saturating* the basic colours used for low, intermediate and high values of the variables. One can say that in the original TLP a color-based model is using the *tone* to refer to the *nominal* value of the variables in each class. In the aTLP, a second dimension is included, the *saturation* of the tone is used to represent the variability, in a clear mapping with the sufficient statistics required to describe a probability distribution (mean and variance). In our proposal, mean, or central trends are associated to tones, whereas variances (or heterogeneities) are associated to de-saturation of the tones. Originally, a discrete model is proposed in both dimensions, with three possible values for tones and also for saturations. Here, a generalization of the model is proposed in such a way that the saturation moves to a continuous space, whereas the tone remains in the discrete space with three possible values.

The idea is to associate a continuous function that given the value of the VC, provides the level of desaturation of the corresponding color, also in a continuous form. To this purpose, the RGB color model will be used to describe the colours in such a way that they can be computed automatically. The RGB model describes all colors as a 3-component vector, decomposing each color in the quantities of Red, Green and Blue required to form it. Each color can contribute with a quantity going from 0 to 255, which represents the saturation of the corresponding colour. In our formalization of the original basic TLP, two basic colours are used (Red, with RGB specification $(255, 0, 0)$ and Green, with RGB specification $(0, 255, 0)$) and the yellow is a composed color, also with maximum saturation, with RGB specification $(255, 255, 0)$. In the parametric aTLP, the darker levels of red and green are obtained by diminishing the saturation for some quantity. However, the gradation of yellows, that was adjusted visually, do not correspond to these principle and changes tone at the same time as diminishes saturation. Also, the desaturation provided between basic red and light dark red is much lower than the one occurring between the light dark red and the hard dark red. This happens in accordance with the fact that the human eye do not work linearly and can distinguish better close brightening colours than close dark colours.

Thus, in the formulation of the continuous saturation component, a function $S : \Re \rightarrow \Re$ is searched, such that:

- Increases inversely w.r.t. VC,
- behaves non-linearly w.r.t. VC.

The RGB definition of the 9 colours used in Fig. 6 is in Table 3.

For red and green colors, a single non-null component appears in the colour definition. Basic tones are associated with null VC, darker colours with VC $= 1$ and mid colours with VC $= 0.5$. The values for red and green colors (Table 3) are used to fit a quadratic function S, giving the desaturation degree associated to a certain VC:

$$S(x) = 80 + 125(1 - x) + 50(1 - x)^2. \qquad (2)$$

To find function S a quadratic regression has been estimated by using the non-null values of color definition as the response variable and associating the three tonalities to three critic values of the CV used as explanatory variable (0 for basic color, 0.5 for light darkening and 1 for hard darkening).

Thus, given a cell in the TLP, where the row represents class $C \in \mathcal{I}$ and the column is associated to the

Table 3
RGB definition of colours used in parametric aTLP

Color	R	G	B
Pure red	255	0	0
Mid red	200	0	0
Dark red	100	0	0
Pure yellow	255	255	0
Mid yellow	240	188	0
Dark yellow	200	95	0
Pure green	0	255	0
Mid greed	0	150	0
Dark green	0	90	0

variable X_k, the cell colour is denoted as h_{Ck}, and it is expressed under the RGB model (a 3-dimensional vector with red, green and blue components) in the following way:

- $h_{Ck}(x) = (S(\mathrm{VC}_{X_k}|C), 0, 0)$ for all the cells with negative values of the variables (corresponding to red colour).
- $h_{Ck}(x) = (0, S(\mathrm{VC}_{X_k}|C), 0)$ for all the cells with positive values of the variables (corresponding to red colour).

The corresponding gradation described by this functions is shown in the first and second scales of Fig. 9, and represents a vertical top down walk over the saturation scale of the Pantone based on pure red and green colors respectively (see Fig 8).

Accordingly, and taking into account that the basic yellow colour used in the original TLP contains the same quantity of red and green, one could think of defining the gradation of yellow colours as

$$h_{Ck}(x) = \big(S(\mathrm{VC}_{X_k}|C), S(\mathrm{VC}_{X_k}|C), 0\big). \qquad (3)$$

However, this provides the gradation shown in Fig. 9 right, which tends too much to the green for CV greater than 0.3 and do not distinguishes well enough from the green scale itself, specially when thinking of big aTLPs where cells seem small pixels. In fact, the colours used in Fig. 4 for the yellow scale do not maintain the same proportions of red and green. Following the same principle, a second function has been estimated for the quantity of the red component in yellow gradation and the following function was found:

$$S'(x) = 180 + 180(1-x) - 143(1-x)^2 + 38(1-x)^3. \qquad (4)$$

Thus, for the cells with intermediate or neutral values in the TLP, where yellow tone must be used, the following definition of the colour is used:

$$h_{Ck}(x) = \big(S'(\mathrm{VC}_{X_k}|C), S(CV_{X_k}|C), 0\big). \qquad (5)$$

In this case, two non-null components are used, one to define the basic tone, and the other defining saturation. See Fig. 8 to note the non-linear walks of those components over the Pantone.

As said before, Fig. 9 provides the gradation of the three colours in the interval [0, 1]. The models are chosen such that for $x = 0.5$ intermediate degrees of darkening are visually observed (what means about a 30% of desaturation of the basic tone, instead of 50%). Even when the VC is not upper bounded, the model

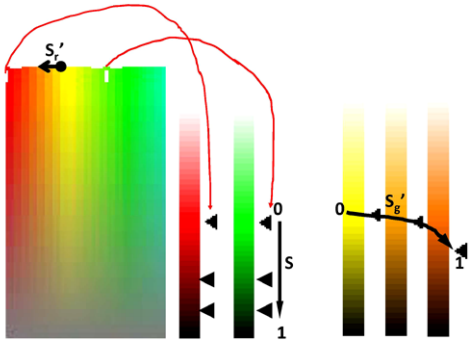

Fig. 8. Walk of the gradation of defined colours over the Pantone. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

RED Scale

x	R	G	B	Color
0,00	255	0	0	
0,05	244	0	0	
0,10	233	0	0	
0,15	222	0	0	
0,20	212	0	0	
0,25	202	0	0	
0,30	192	0	0	
0,35	182	0	0	
0,40	173	0	0	
0,45	164	0	0	
0,50	155	0	0	
0,55	146	0	0	
0,60	138	0	0	
0,65	130	0	0	
0,70	122	0	0	
0,75	114	0	0	
0,80	107	0	0	
0,85	100	0	0	
0,90	93	0	0	
0,95	86	0	0	
1,00	80	0	0	

GREEN Scale

x	R	G	B	Color
0,00	0	255	0	
0,05	0	244	0	
0,10	0	233	0	
0,15	0	222	0	
0,20	0	212	0	
0,25	0	202	0	
0,30	0	192	0	
0,35	0	182	0	
0,40	0	173	0	
0,45	0	164	0	
0,50	0	155	0	
0,55	0	146	0	
0,60	0	138	0	
0,65	0	130	0	
0,70	0	122	0	
0,75	0	114	0	
0,80	0	107	0	
0,85	0	100	0	
0,90	0	93	0	
0,95	0	86	0	
1,00	0	80	0	

Proposed YELLOW Scale

x	R	G	B	Color
0,00	255	255	0	
0,05	255	244	0	
0,10	254	233	0	
0,15	253	222	0	
0,20	252	212	0	
0,25	251	202	0	
0,30	249	192	0	
0,35	247	182	0	
0,40	245	173	0	
0,45	242	164	0	
0,50	239	155	0	
0,55	236	146	0	
0,60	232	138	0	
0,65	227	130	0	
0,70	222	122	0	
0,75	217	114	0	
0,80	211	107	0	
0,85	204	100	0	
0,90	197	93	0	
0,95	189	86	0	
1,00	180	80	0	

YELLOW Scale

x	R	G	B	Color
0,00	255	255	0	
0,05	244	244	0	
0,10	233	233	0	
0,15	222	222	0	
0,20	212	212	0	
0,25	202	202	0	
0,30	192	192	0	
0,35	182	182	0	
0,40	173	173	0	
0,45	164	164	0	
0,50	155	155	0	
0,55	146	146	0	
0,60	138	138	0	
0,65	130	130	0	
0,70	122	122	0	
0,75	114	114	0	
0,80	107	107	0	
0,85	100	100	0	
0,90	93	93	0	
0,95	86	86	0	
1,00	80	80	0	

Fig. 9. Gradation of colours. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

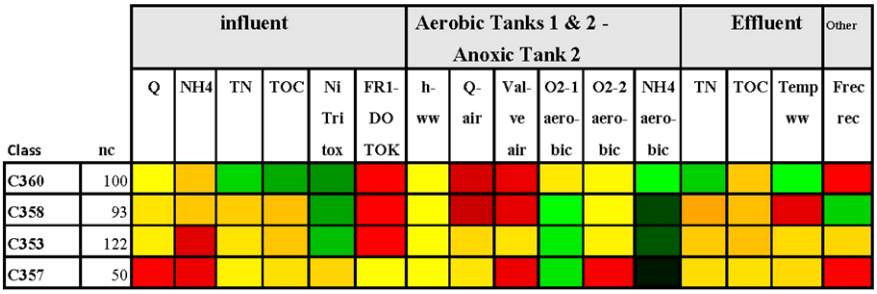

Fig. 10. aTLP non-parametric. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140611.)

stays valid, as for VC greater than one, the values of S and S' will approach to 0, providing darker and darker colours. In the limit, for 0 values, the black colour will be displayed, which will represent the highest variability and no availability of that variable for the decision-making.

Using the proposed model, the enriched non-parametric aTLP for the case study is shown in Fig. 10. From the figure, intermediate levels of heterogeneity are observed in general, except for NH4-aerobic that shows darker colours in most variables. In this non-parametric approach, the responsibility of assuming or not the cost of wrong decisions relies again on the expert. He must have in mind some threshold of darkening that becomes non-assumable according to his domain and operate accordingly.

6. Discussion

On the basis of the TLP presented for the first time in [11], the *annotated TLP* (*aTLP*) is introduced as an enriched tool to interpret classes obtained in a clustering process. While the TLP represent prototypes, giving information of the central pattern associated to each profile, the aTLP [8] takes into account the uncertainty associated to prototypes, which is directly related with the risk of involved decisions. In fact, the aTLP proposes a color-based model to represent prototypes, which includes two dimensions for the coloring:

- Tone: which is associated with central trend of the variable inside the class, in a simplified representation where three qualitative levels (low, intermediate or high values) are considered.
- Saturation: which is associated with the *purity* or homogeneity of the cell, and is based on the degree of variability of the variable around the central trend.

Those two dimensions are, indeed, visualizing the two sufficient characteristics, from a statistical point of view, to describe a distribution, and become, in turn, sufficient to understand the model and to support informed-decisions associated to prototypes. In the presented proposal, medians conditioned to classes are used to quantify central trends, whereas variation coefficients are used to quantify dispersion in a more robust representation than the classical one using means and standard deviations.

As in the original TLP, the central trend inside every cell is represented in a discrete space with three qualitative levels, associated to the basic traffic lights colours (red, green and yellow). In previous works [8,9], automatic methods are proposed to automatically determine the tone of every cell. The direct or revers assignment between colours and values of the variables includes the semantics associated to the polarity of the variable itself, and expressed by the experts in the *polarity semantics table*, a concept introduced in [9] to allow the expert an easy transfer to the system of that characteristic.

Pure colors are associated with no variability (or low), whereas aTLP introduces darkening to indicate increasing in heterogeneity, and, in consequence, lower reliability of associated decisions.

One of the main contributions of the paper is formalizing a model for the darkening of the cells. Darkening is implemented through desaturation of basic colors and two different models are provided in the paper:

- On the one hand a discrete modelling provides three levels of darkening (pure colour, light darkening and hard darkening), associated with three ranges of VC. Those three ranges are defined on the basis of two parameters $k_1, k_2 \in [0,1]^2$ inducing a cut-off over the VC range. Darker aTLPs are associated with higher uncertainty and less reliable decision support, as reliable decisions are supported by pure colour' cells. The values of k_1,

k_2 are provided by the expert according to the cost of wrong decisions. The lower the values of the parameters, the darker the aTLP and the more conservative the interpretation of the profiles.

- On the other hand, a continuous modelling provides real functions to determine the levels of desaturation on the basis of the values of VC. Here, a non-parametric approach appears, in the sense that the darkening do not depend on cut-offs induced over the VC range, but on their single values. The RGB colour-model has been used to find the desaturation function S, and colours of cells can be automatically determined on the basis of VC, provided that the medians indicates the basic tone to be graduated with S values. Fitting techniques have been used to find a quadratic function for S, according to the non-linear perception of the human eye. A correction had to be introduced for the yellow scale by introducing S', a cubic function moving the yellow scale from a basic yellow to a more brown color, perceived as a darker one by human eyes, but avoiding confusing effects with intermediate colours of the green scale. As in the previous formulation, darker aTLPs indicates less reliable decisions.

In a quick insight, the aTLP shows to the decision-maker which prototypes in which groups involve less variability and, in consequence, are more reliable. This permits to evaluate the risks of associated decisions. For both models, darkness is interpreted as low reliability of associated decisions, or higher risk of wrong decisions. However, there is a fundamental difference between the parametric aTLP and the non-parametric version. In the former, the colour of a cell depends, not properly on the value of their corresponding VC, but on the associated cut-offs, whereas a single colour is associated to a certain value of a VC in the non-parametric approach. Thus, the non-parametric approach provides an objective modelling that associates a single aTLP to a certain TLP and the expert must decide which level of darkness can be acceptable in his domain to identify which subset of cells can be good supports for his decisions. On the other hand, the parametric version of the aTLP, permits to introduce in the visualization the expert criteria themselves. More or less conservative aTLPs can be obtained by determining the cut points on the VC to be used for colour-gradation, thus enabling to adjust the uncertainty representation to the real costs of wrong decisions in the target application. When the cost of wrong decisions is clear, and experts can formulate consensual values for k_1, k_2, the aTLP(k_1, k_2) permits to standardize decisions over a set of experts, by homogenization of decision rules. In the parametric version of the aTLP, darker cells represent non-assumable risks, whereas pure cells represent reliability, and intermediate colours represent assumable risk in a rigid scheme.

In the paper, three different scenarios of the parametric version of the aTLP have been presented for the Slovenian WWTP, from a more permissive situation (risk levels 0.30–0.90) where quite high degrees of variability within classes are considered not critical for decisions, to a more conservative situation (risk levels 0.10–0.30) where small degrees of variability (0.3) are associated to non-assumable risk. Thus, the aTLP offers a friendly paradigm for bringing together the representation of typical situations in a system, with the risk evaluation. For the parametric version, the cost of wrong decisions is given by the experts and VC cut points determined accordingly, whereas for the non-parametric one, the assumable darkness is managed implicitly by the expert.

7. Conclusions and future work

Interpretation oriented tools are required to bridge the gap between raw data mining results and effective decision-making. Post-processing tools might contribute to this topic. Although understandability of data mining patters was highlighted from the seminal paper of KDD [4], unfortunately, not much works are found in this direction yet [3,13], although some of the most relevant members of the scientific community still stress the importance of developing these kind of tools linked to the urgent need of data scientists [20,21]. This paper proposes two different approaches to post-processing tools in the particular field of clustering that provide understanding of discovered profiles, and, as a consequence, better support to further decision-making processes.

The aTLP proposes a two-dimensional color-based model to represent prototypes, based on tone and saturation to respectively represent central trend and variability, thus enriching the original conceptualization of TLP with the uncertainty associated to prototypes.

Our conclusion, after discussing with several experts is that, both models are perceived as friendly tools, contributing to understand the profiling, and both are useful for different scenarios depending on the application goals.

In conclusions, when the aTLP is analyzed as a global picture of the domain, human eye is greatly powerful to catch the aTLP dominance and simple interpretation rules provide high information about the target phenomenon:

- Dominant green tone: profiles associated with benign situation. The meaning of benign depending on the context. In the particular application presented here, benignity is associated with correct plant operation and higher water quality.
- Dominant yellow tone: Neutral profiles, associated either to intermediate values or lack of trend depending on the darkening degree.
- Dominant red tone: profiles associated with malign situation. In the case study, lower water quality or abnormal plat operation.

For the parametric aTLP version, the lower the k_1, k_2 values the more conservative the aTLP; also:

- Dominance of pure colours: non-risky profiles, associated with well supported decisions.
- Dominance of light darkening cells: moderate risk for associated decisions.
- Dominance of hard darkening cells: non-assumable risks.

For the non-parametric aTLP version

- Dominance of pure, or highly saturated colours: non-uncertain profiles, associated with well supported decisions.
- Dominance of intermediate colours: moderate levels of uncertainty associated to classes, ill-supported decisions, the darker the colour, the higher the uncertainty, the user must decide which level of darkening represents non-assumable costs of wrong decisions, according to his background knowledge.
- Dominance of hard darkening cells: high levels of uncertainty, associated with too risky decisions.

Experts can quickly identify which variables have too high variations to make decisions based on them, which of them concentrates malignities and require more attention without requiring any technical skills to understand neither the clustering techniques nor the formal properties of the clusters. Both parametric and non-parametric versions of aTLP were used in a real application over the Slovenian WWTP and were perceived by the expert as a friendly and comprehensible tool to the decision-makers, which can contribute to reduce the gap between data mining and effective decision support. The aTLP also contributes to integrate the expert in the KDD process itself.

New criteria to improve automatic identification of the reference tone for each aTLP cell are in progress with the aim of standardising a general rule to find TLPs. Also, a general survey is in progress to build TLPs and both parametric and non-parametric versions of aTLP over different real datasets, to verify the concordance between the decisions suggested by those tools and reliability perceptions of the domain experts. In the long term, the possibility to generalize the tone model to a continuous space will also be explored.

Acknowledgements

We would like to thank the water expert Dr. Darko Vrecko from Jozeph Stephan Institute (Ljubljana, Slovenia) and the Domzale-Kamnik WWTP for providing data.

References

[1] D. Conti and K. Gibert, The use of the Traffic Lights Panel as a goodness-of-clustering indicator. An application to Financial Assets, in: *Procs CCIA 2013, Artificial Intelligence Research and Development*, Frontiers in Artificial Intelligence and Applications, Vol. 248, 2012, pp. 19–28.

[2] G. Corral, E. Armengol, A. Fornells and E. Golobardes, Explanations of unsupervised learning clustering applied to data security, *Neurocomputing* **72**(13–15) (2009), 2754–2762.

[3] P. Cortez and M.J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, *Information Sciences* **225** (2013), 1–17.

[4] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From data mining to knowledge discovery: an overview, in: *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[5] P. Feldman, P. Nadash and M. Gursen, Improving communication between researchers and policy makers in long-term care: or, researches are from Mars, Policy Makers are from Venus, *The Gerontologist* **41**(3) (2001), 312–321.

[6] M. Fiter, D. Güell, J. Comas, J. Colprim, M. Poch and I. Rodríguez-Roda, Energy saving in a wastewater treatment process: an application of fuzzy logic control, *Environmental Technology* **26**(11) (2005), 1263–1270.

[7] K. Gibert, G. Rodríguez-Silva and I. Rodríguez-Roda, Knowledge Discovery with Clustering based on rules by States: A water treatment application, *Environmental Modelling and Software.* **25** (2010), 712–723.

[8] K. Gibert, D. Conti and M. Sànchez-Marre, Decreasing uncertainty when interpreting profiles through traffic lights panel, in: *Advances in Computational Intelligence, Procs IPMU 2012, Part II*, Communications in Computer and Information Science, CCIS, Vol. 298, 2012, pp. 137–148, DOI:10.1007/978-3-642-31715-6.

[9] K. Gibert, D. Conti and D. Vrecko, Assisting the end-user in the interpretation of profiles for decision support. An application to wastewater treatment plants, *Environmental Engineering and Management Journal* **11**(5) (2012), 931–944.

[10] K. Gibert, A. García-Rudolph, L. Curcoll, D. Soler, L. Pla and J.M. Tormos, Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states, *Studies in Health Technology and Informatics* **150** (2009), 579–583.

[11] K. Gibert, A. Garcia-Rudolph, A. Garcia-Molina, T. Roig-Rovira, M. Bernabeu and J.M. Tormos, Response to TBI-neurorehabilitation through an AI & Stats hybrid KDD methodology, *Medical Archives* **62**(3) (2008), 132–135.

[12] K. Gibert, R. Nonell, J.M. Velarde and M.M. Colillas, Knowledge discovery with clustering: impact of metrics and reporting phase by using KLASS, *Neural Network World* **15**(4) (2005), 319–326.

[13] K. Gibert, G. Rodríguez-Silva and R. Annicchiarico, Post-processing: bridging the gap between modelling and effective decision-support. The profile assessment grid in human behaviour, *Mathematical and Computer Modelling* **57**(7,8) (2013), 1633–1639.

[14] K. Gibert, J. Spate, M. Sànchez-Marrè, I.N. Athanasiadis and J. Comas, Data mining for environmental systems, in: *Environmental Modelling, Software and Decision Support*, A.J. Jakeman, A.A. Voinov, A.E. Rizzoli and S.H. Chen, eds, Developments in Integrated Environmental Assessment, Vol. 3, Elsevier, 2008, pp. 205–228.

[15] R. Gunther, Business models: a discovery driven approach, *Long Range Planning* **43** (2010), 247–261.

[16] M. Hammond, in: *The Fact Gap: the Disconnect Between Data and Decisions. A Study of Executives in USA and Europe*, Business Research Services, 2004.

[17] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd edn, Morgan Kaufmann Publishers, 2011.

[18] S. Hutchins, in: *Principles for Intelligent Decision Aiding*, Kluwer Academic Publishers, 1996.

[19] Kdnuggets, November 2011, www.kdnuggetscom/polls.

[20] D. Palmer, Not enough data scientists, MIT experts tells computing, *Computing* (2014), 2292485, available at: http://www.computing.co.uk/ctg/news/2292485.

[21] S. Shah, Analysis: It takes skills to explore big data, *Computing* (2013), 2237857.