# Interval Forecast of Water Quality Parameters

**Orlando Ohashi** [1] and **Luís Torgo** [1] and **Rita P. Ribeiro** [1]

**Abstract.**

The current quality control methodology adopted by the water distribution service provider in the metropolitan region of Porto - Portugal, is based on simple heuristics and empirical knowledge. Based on the domain complexity and data volume, this application is a perfect candidate to apply data mining process. In this paper, we propose a new methodology to predict the range of normality for the values of different water quality parameters. These intervals of normality are of key importance to decide on costly inspection activities. Our experimental evaluation confirms that our proposal achieves good results on the task of forecasting the normal distribution of values for the following 30 days. The proposed method can be applied to other domains with similar network monitoring objectives.

## 1 Introduction

Given the strong socio-economical impact of the potable water in the human life it is important to have a strict control of the water quality. Moreover, controlling and/or correcting water quality problems has high costs for water distribution companies. Considering the nature and the high dimension of the data that is typically collected by automatic sensors, this application has been addressed using data mining techniques (e.g. [1]).

In this work we use data provided by the AdDP (*Águas do Douro e Paiva*) company, which is responsible for the water distribution network of the metropolitan region of Porto, the second largest city of Portugal. The AdDP company has an obligation to ensure the quality of the distributed water for the population. To achieve this goal the company continuously monitors several water quality parameters throughout the distribution network, verifying if any parameter overcomes the legal limits. For each water quality parameter these limits represent the acceptable range of values that guarantee the quality of the distributed water and are set by the government. Disrespecting these limits leads to severe fines and may put the public health at risk.

With the goal of timely detecting deviations from the normal behavior of each parameter, the AdDP company carries out regular monitoring activities. Figure 1 shows the time plot for the parameter pH. The values of this parameter show a strong seasonal influence. Based on the dynamic and seasonal behavior present on the parameters, it is important for the company to have a dynamic definition of normality that reflects the current state of the network. The current control methodology adopted by the company is based on a fixed definition of normality. This methodology has no flexibility (e.g. with respect to seasonal effects) and may lead to a higher number of false alarms increasing the costs for the company. In this domain a false alarm is the wrong identification of abnormal behavior on a parameter. In other worlds, is to generate a false alert indicating that a parameter presents an unusual behavior.

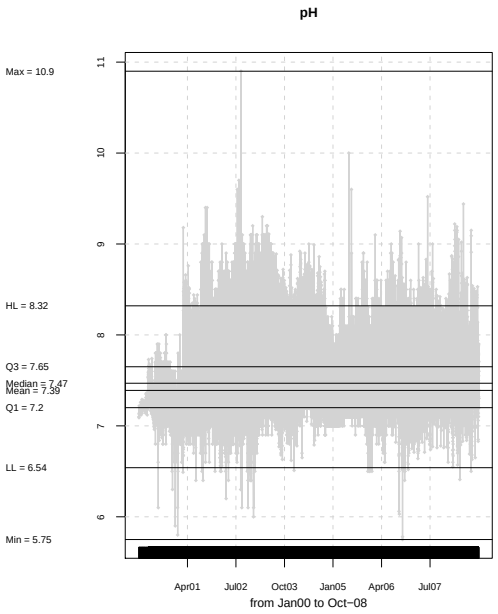


**Figure 1.** Global behavior of the parameter pH.

In this paper we are interested in dynamically defining the normal behavior of each water quality parameter. However, there is no general accepted rule that precisely defines what is the normal or abnormal behavior of a time series. We adopt a definition of normality based on the observed statistical distribution of the variable. We use the box-plot rule that is based on the quantiles of the variable to determine what are the most common values [11]. This rule is based on an assumption of near-normality of the variable under study, which is an acceptable assumption for water quality parameters. The rule is based on the 1st ($Q_1$) and 3rd ($Q_3$) quantiles, that are the values below which there are 25% and 75% of the data, respectively. This quantiles identify the central 50% of the data values. Our normality interval is $[LL, HL]$, where $LL = Q_1 - 1.5 \times (Q_3 - Q_1)$ and $HL = Q_3 + 1.5 \times (Q_3 - Q_1)]$. Any value outside this interval is considered an anomaly/outlier.

In terms of predictive models our goal is to accurately forecast the quantiles $Q_1$ and $Q_3$ for the following 30 days, using the current state of the distribution network, for each water quality parameter monitored by the company. The standard time series forecasting techniques focus on single point forecasts, typically the expected mean value of the target variable, that does not contain information about

[1] LIAAD INESC Porto L.A. / Faculty of Sciences-University of Porto. Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal. Emails: {ohashi, ltorgo, rpribeiro}@liaad.up.pt

the dispersion of the observations around the predicted value [12]. In our work we are interested in forecasting an interval of normality for a future period of time. We obtain this by forecasting the quantiles of the future distribution, more specifically the quantiles $Q_1$ and $Q_3$. We apply several standard machine learning models to achieve this task, namely: Regression Trees, Support Vector Machines, Random Forests and also a Quantile Regression method.

In Section 2 we describe the problem of interval forecasting and related approaches. Section 3 presents the models and the evaluation metrics used to compare them. Section 4 presents the results of the experimental comparison we have carried out with our data. The final section provides a summary and conclusions of this work.

## 2 Interval Forecasting

The majority of the research in time series forecasting is focused on single point prediction. Predicting the next value of a variable is of key importance for many applications. For example, a production manager wants to know what will be the future sales in order to plan the production. The sales manager wants to forecast the sales to optimize the stock policy of the company. Like these small examples, many other similar domains make decisions based on the forecasts of a single future value. However, there are different application domains where the prediction of a single point in the future is less important or insufficient to give enough information to take the appropriate decisions. For these domains having a model that forecasts the expected interval of values, gives more details that allows more informed decisions to be taken. For these domains it is not enough to have a prediction of the expected value of a variable. In these domains we want predictions of the dispersion of the target variable [12]. For example, in customer wallet estimation, the potential spending by costumers gives more information than the expected estimation [14]. Inventory control systems typically require the frequent updating of forecasts for many different products. Interval forecasts provide more information about the safe stock levels [16]. For wind velocity/power prediction, it is important to predict not only the next wind speed value, but the possible range of wind speeds in the near future so that the system operators can ensure that supply and demand are balanced [18]. A similar problem is faced with electricity forecasts, where it is more important to predict the interval of demand than a single value [17].

In this paper we are interested in the forecast of the quantiles that can be used to represent the normal behavior interval of each parameter. The task will be to obtain forecasts for these quantiles for the next 30 days, using the current information of the distribution network. With the prediction of the quantiles $Q_1$ and $Q_3$ we establish the interval of the normal behavior based on the box-plot rule. These predicted intervals of normality will be crucial in the company monitoring activities, and will be used to anticipate problems and take preventive actions.

Koenker and Bassett [9] first introduced quantile regression in 1978. Since then, their method has being attracting attention of the research community, as an alternative to the majority of the research methods that focus on the estimation of the mean. Several quantile regression approaches have been developed. Examples include an implementation based on Random Forests [12], an implementation using Support Vector Machines [8], using the MM algorithm [7] and an implementation using Neural Networks [19]. In this paper we used the implementation based on Random Forests with more details given in Section 3.

## 3 Building and Evaluating Interval Forecast Models

In this work we applied several models with the aim of accurately forecasting the quantiles $Q_1$ and $Q_3$ of four water quality parameters: Aluminum, Iron, pH and Turbidity. In this paper we did not use information about the geographical location of the measured values. Instead, we looked at the network globally using the mean daily value of each parameter that was calculated by aggregating different values measured across the water distribution network on the same day. For each parameter the goal is to accurately forecast the interval $I_i(x)$ for a parameter $x$ defined by the quartiles on the following 30 days (c.f. Equation 1)

$$I_i(x) = [\hat{Q}_1(x)_{i\cdots i+30}, \hat{Q}_3(x)_{i\cdots i+30}] \quad (1)$$

where $\hat{Q}_1(x)_{i\cdots i+30}$ and $\hat{Q}_3(x)_{i\cdots i+30}$ are the 1st and 3rd quantiles predicted for the next 30 days.

Given that our data set consists of a set of time series (one for each parameter), we have selected as experimental methodology a Monte Carlo simulation. We randomly selected 10 dates from the period for which we have data. Using these dates we have selected a training window using the values observed in the previous 365 days. The respective test window is formed with the values for the following 90 days. This process is repeated for the 10 randomly selected dates and all models are trained and tested using these same windows. Figures 2, 3, 4 and 5, show the randomly selected intervals for each parameter: Aluminum, Iron, pH and Turbidity, respectively. In these figures it is also possible to confirm the strong seasonal effects of the data. For each of the 10 intervals, different models are used to obtain predictions for the respective 90 days testing period (last part of the rectangles in the Figures). Using the first 365 days a first model is obtained for each of the techniques that will be compared. This model is used to obtain the first predictions. With a periodicity of 10 days the models are re-constructed adding the new known data from these 10 days. We consider both growing and sliding window strategies for this model re-construction. The former adds the new data to the existing training set, while the latter simply slides the training set to maintain it always at the same 365 days size.

### 3.1 Models

We have tried eight different models in our problem. However, only seven are reported in this paper as we were not able to obtain good results with the neural network variants we have tried. For the sake of reproducibility all experiments were carried out using the R [15] environment. We split the models in three main classes. In the first group are the models based on simple heuristics and empirical knowledge:

**AdDP** a fixed value of the distribution of each parameter that was calculated using the previous distribution of the parameters and empirical knowledge;

**Previous Distribution** a simple heuristic that applies the quantiles estimated using the most recent training data;

**Previous Season Distribution** a method similar to the previous one, but which uses the past data from the same season as the one we are trying to forecast.

The second group contains classical machine learning models. For these techniques we have obtained an individual predictive model for each quantile.

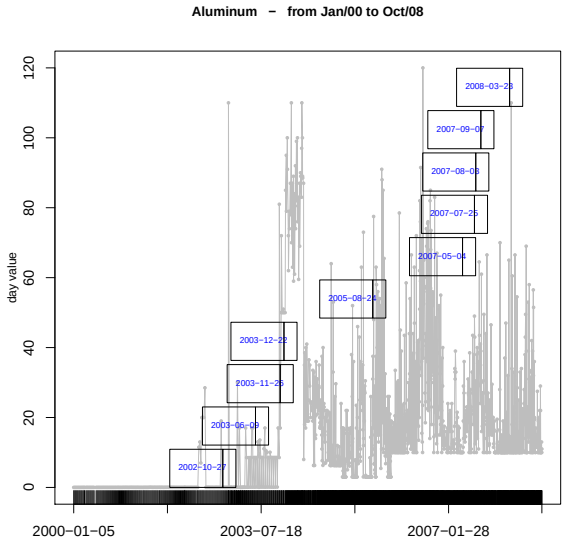


**Figure 2.** Monte Carlo simulation for Aluminum.

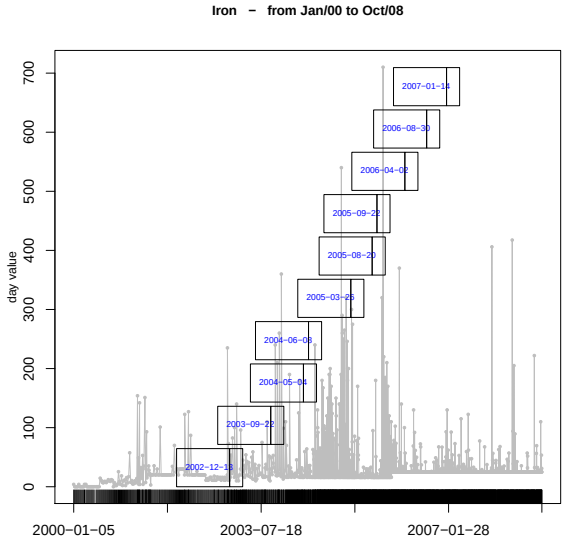


**Figure 3.** Monte Carlo simulation for Iron.

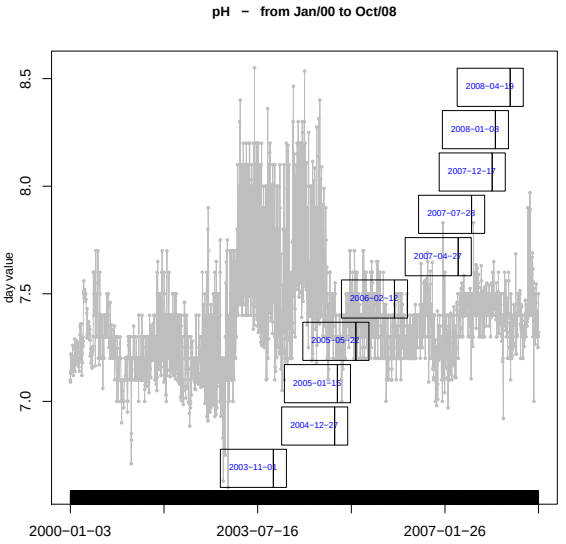


**Figure 4.** Monte Carlo simulation for pH.

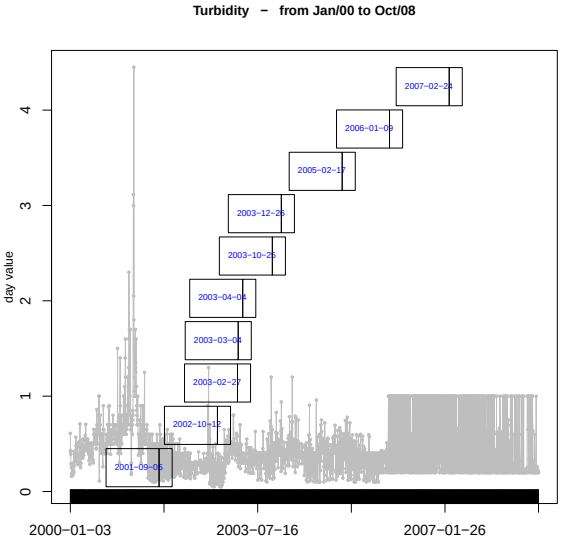


**Figure 5.** Monte Carlo simulation for Turbidity.

**Regression Trees [2]** a regression tree model developed for each quantile, $Q_1$ and $Q_3$;

**Neural Networks [6]** a neural network for each quantile (the results of this model were too bad, so they were omitted from our comparative tables);

**Support Vector Machines [4]** an SVM model to predict each quantile;

**Random Forest [3]** one random forest to predict each quantile.

The third group has just one model, based on quantile regression techniques.

**Quantile Regression Random Forest [13]** a random forest variant designed to optimize the prediction of quantiles.

The models in the first group are rather simple and can be regarded as a kind of baseline level of performance that the other approaches should clearly overcome. These models required no pre-processing steps. The remaining models are multiple regression models, while our data is a time series. We have used the classical time-delay embedding strategy to obtain a data set for these models. This strategy defines as target variable a future value of the series and uses the most recent past values as predictors to be used by the models. However, our application has some particularities as we are interested in predicting a quantile of the values in the next 30 days. We have calculated these quantiles and use them as the target variables of our models. As predictors we have used the values on the previous 10 days $(W_i, ..., W_{i-9})$ and also several statistics that describe the recent dynamics of the time series. Namely, we have used the mean $(\overline{W}_{i-30\cdots i})$, standard deviation $(\sigma_{i-30\cdots i})$, median $(\widetilde{W}_{i-30\cdots i})$, $Q_1$ $(Q_{1_{i-30\cdots i}})$ and $Q_3$ $(Q_{3_{i-30\cdots i}})$ measured in the previous 30 days. In summary, our prediction problem can be described as trying to obtain an approximation of the unknown regression function described in Equation 2,

$$(Q_{1_{i\cdots i+30}}, Q_{3_{i\cdots i+30}}) = f(W_i, W_{i-1}, ..., W_{i-9}, \overline{W}_{i-30\cdots i}, \sigma_{i-30\cdots i}, \widetilde{W}_{i-30\cdots i}, Q_{1_{i-30\cdots i}}, Q_{3_{i-30\cdots i}}) \quad (2)$$

With respect to quantile regression random forests we have used a different pre-processing as these models do not need that the target variable is a quantile to be able to predict these statistics. In this context, we have used the average target variable value in the next 30 days as the target of these models, as shown in Equation 3,

$$\overline{W}_{i\cdots i+30} = f(W_i, W_{i-1}, ..., W_{i-9}, \overline{W}_{i-30\cdots i}, \sigma_{i-30\cdots i}, \widetilde{W}_{i-30\cdots i}, Q_{1_{i-30\cdots i}}, Q_{3_{i-30\cdots i}}) \quad (3)$$

## 3.2 Evaluation

There is an extensive literature on evaluation metrics for single value prediction models. Usually, the preferred measure is the mean squared error (MSE). Our prediction problem is different as we have mentioned before. We selected two error measures that are more appropriated for this class of applications. The first is the Mean Absolute Deviation (MAD) for each quantile. See Equation 4, where $\alpha$ is the respective quantile,

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |\hat{Q}_\alpha - Q_\alpha| \quad (4)$$

When we forecast a value for a certain quantile, this is equivalent to say that we predict that a certain percentage of values will be below this prediction, as that is the definition of a quantile. The second error measure is the mean difference between the predicted and observed percentages, named Mean Quantile Error (MQE),

$$MQE = \frac{1}{n}\sum_{i=1}^{n} |\beta - \delta| \quad (5)$$

where $\beta$ is the observed percentage of data that is in the predict quantile, and $\delta$ is the correct percentage of the quantile. In our case we have two values of $\delta$. The value .25 for the quantile $Q_1$ and .75 for the quantile $Q_3$.

# 4 Comparison of Interval Forecast Models

This section describes the comparison of the models we have described previously, in the task of predicting the quantiles of the selected water quality parameters. The results we present were estimated using the Monte Carlo experimental method described previously. For each parameter we present a table of results with results of the following models,

- addp → model currently used by the AdDP company;
- dist → model that uses the previous distribution information;
- dist.seaso → model that uses the previous seasonal distribution information;
- rpart [20] → Regression Tree model;
- svm [5] → Support Vector Machine model;
- rf [10] → Random Forest model;
- qrrf [13] → Quantile Regression Random Forest model.

In the tables the model IDs have the terminations .s and .g appended. The first corresponds to a sliding window application of the model while the second is for the growing window approach.

For the parameter Aluminum the random forest ('rf') was the best model with both windowing strategies and error measures, as we can see in the Table 1. The second best model was the regression tree. The baseline models 'addp' and 'dist' achieved very poor results.

For the parameter Iron we did not observe a single model that was significant better than the others. Considering MAD statistic, the best model was the 'svm', followed by the random forest. However, considering the MQE error measure the best result was achieved by the 'addp' model followed by the 'dist', random forest and 'svm', respectively.

The best model for the parameter pH was the random forest in the both measures. However, regression trees achieved similar results considering the MAD error measure. Considering the MQE measure the random forest was significantly better than the reference models.

For the parameter Turbidity the best models in terms of MAD were the random forests and SVMs. With respect to MQE regression trees and SVMs appeared as the best, followed by random forests. These models were significantly better than the baseline models.

| | MAD | MAD.Q1 | MAD.Q3 | MQE | MQE.Q1 | MQE.Q3 |
|---|---|---|---|---|---|---|
| addp.s | 63.84 | 32.3 | 31.54 | 0.68 | 0.44 | 0.24 |
| addp.g | 63.84 | 32.3 | 31.54 | 0.68 | 0.44 | 0.24 |
| dist.s | 31.45 | 13.7 | 17.75 | 0.35 | 0.17 | 0.18 |
| dist.g | 31.03 | 13.58 | 17.45 | 0.35 | 0.17 | 0.18 |
| dist.seaso.s | 36.17 | 13.88 | 22.28 | 0.36 | 0.17 | 0.19 |
| dist.seaso.g | 40.21 | 14.13 | 26.09 | 0.38 | 0.18 | 0.2 |
| rpart.s | 23.91 | 10.62 | 13.3 | 0.34 | **0.15** | 0.18 |
| rpart.g | 24.03 | 10.78 | 13.25 | 0.34 | 0.16 | 0.18 |
| svm.s | 28 | 11.99 | 16.01 | 0.38 | 0.18 | 0.2 |
| svm.g | 28.93 | 12.21 | 16.73 | 0.39 | 0.18 | 0.2 |
| rf.s | **21.82** | **9.62** | **12.2** | **0.31** | **0.15** | **0.16** |
| rf.g | **22.35** | **9.66** | **12.69** | **0.32** | **0.15** | **0.17** |
| qrrf.m.s | 34.94 | 15.49 | 19.45 | 0.49 | 0.27 | 0.22 |
| qrrf.m.g | 35.2 | 15.61 | 19.59 | 0.5 | 0.27 | 0.23 |

**Table 1.** Evaluation error for the parameter Aluminum

| | MAD | MAD.Q1 | MAD.Q3 | MQE | MQE.Q1 | MQE.Q3 |
|---|---|---|---|---|---|---|
| addp.s | 18.99 | 3.99 | 15 | **0.42** | 0.32 | **0.1** |
| addp.g | 18.99 | 3.99 | 15 | **0.42** | 0.32 | 0.1 |
| dist.s | 14.28 | 2.05 | 12.23 | 0.45 | 0.35 | **0.1** |
| dist.g | 14.5 | 2.56 | 11.94 | 0.43 | 0.34 | **0.09** |
| dist.seaso.s | 14.61 | 2.14 | 12.47 | 0.47 | 0.34 | 0.13 |
| dist.seaso.g | 15.52 | 2.38 | 13.14 | 0.47 | 0.32 | 0.15 |
| rpart.s | 13.27 | 0.76 | 12.51 | 0.53 | 0.39 | 0.14 |
| rpart.g | 12.79 | 0.82 | 11.98 | 0.5 | 0.37 | 0.13 |
| svm.s | **10.82** | 0.76 | **10.06** | 0.46 | **0.3** | 0.16 |
| svm.g | **10.84** | 0.83 | **10** | 0.46 | **0.3** | 0.16 |
| rf.s | 11.39 | **0.75** | 10.64 | 0.45 | 0.32 | 0.13 |
| rf.g | 11.27 | **0.72** | 10.55 | 0.43 | 0.31 | 0.12 |
| qrrf.m.s | 35.08 | 14.42 | 20.66 | 0.65 | 0.53 | 0.11 |
| qrrf.m.g | 35.54 | 14.73 | 20.81 | 0.65 | 0.54 | 0.11 |

**Table 2.** Evaluation error for the parameter Iron

| | MAD | MAD.Q1 | MAD.Q3 | MQE | MQE.Q1 | MQE.Q3 |
|---|---|---|---|---|---|---|
| addp.s | 0.27 | 0.1 | 0.16 | 0.28 | 0.14 | 0.14 |
| addp.g | 0.27 | 0.1 | 0.16 | 0.28 | **0.14** | 0.14 |
| dist.s | 0.21 | 0.08 | 0.14 | 0.28 | 0.19 | 0.08 |
| dist.g | 0.2 | **0.07** | 0.13 | 0.26 | 0.19 | 0.08 |
| dist.seaso.s | 0.2 | 0.08 | 0.12 | 0.22 | 0.14 | 0.08 |
| dist.seaso.g | 0.2 | 0.08 | 0.11 | 0.22 | 0.15 | 0.07 |
| rpart.s | **0.18** | **0.07** | 0.11 | 0.23 | 0.15 | 0.08 |
| rpart.g | **0.18** | **0.07** | 0.11 | 0.23 | 0.15 | 0.08 |
| svm.s | 0.19 | 0.08 | 0.11 | 0.22 | 0.15 | 0.08 |
| svm.g | 0.19 | **0.07** | 0.11 | 0.23 | 0.15 | 0.08 |
| rf.s | **0.18** | 0.08 | **0.1** | **0.2** | **0.13** | **0.07** |
| rf.g | **0.18** | 0.08 | **0.1** | **0.21** | 0.15 | **0.06** |
| qrrf.m.s | 0.31 | 0.21 | **0.1** | 0.46 | 0.35 | 0.11 |
| qrrf.m.g | 0.31 | 0.21 | **0.1** | 0.46 | 0.35 | 0.11 |

**Table 3.** Evaluation error for the parameter pH

| | MAD | MAD.Q1 | MAD.Q3 | MQE | MQE.Q1 | MQE.Q3 |
|---|---|---|---|---|---|---|
| addp.s | 0.88 | 0.41 | 0.47 | 0.73 | 0.52 | 0.21 |
| addp.g | 0.88 | 0.41 | 0.47 | 0.73 | 0.52 | 0.21 |
| dist.s | 0.19 | 0.06 | 0.13 | 0.25 | 0.13 | 0.12 |
| dist.g | 0.19 | 0.06 | 0.13 | 0.25 | 0.13 | 0.12 |
| dist.seaso.s | 0.14 | 0.06 | **0.08** | 0.28 | 0.16 | 0.12 |
| dist.seaso.g | 0.16 | 0.07 | **0.09** | 0.29 | 0.16 | 0.13 |
| rpart.s | 0.15 | **0.04** | 0.1 | **0.22** | **0.11** | 0.11 |
| rpart.g | 0.16 | **0.04** | 0.12 | **0.21** | **0.1** | 0.11 |
| svm.s | **0.13** | **0.04** | 0.09 | **0.22** | **0.11** | 0.11 |
| svm.g | **0.13** | **0.04** | **0.09** | 0.23 | 0.12 | **0.1** |
| rf.s | **0.13** | **0.04** | 0.09 | 0.23 | 0.12 | **0.1** |
| rf.g | **0.13** | **0.04** | **0.09** | 0.22 | 0.13 | **0.1** |
| qrrf.m.s | 0.33 | 0.17 | 0.15 | 0.42 | 0.29 | 0.13 |
| qrrf.m.g | 0.33 | 0.18 | 0.15 | 0.42 | 0.29 | 0.13 |

**Table 4.** Evaluation error for the parameter Turbidity

## 5 Conclusions and Future Work

This paper described a concrete practical application of data mining. The task has strong socio-economical impact in a large region of Portugal. Our aim was to obtain models that could be used in forecasting the "normal" expected behavior of a series of water quality parameters in a distribution network. This is a key step in the task of monitoring the values of these parameters to enable preventive actions to be taken before the water quality is at risk. This problem has a high economic importance for the water distribution company on top of the mentioned socio-economical impacts to the region in general.

The proposed methods for this task achieved very good results clearly surpassing the current method that the company uses, on most setups. It is our belief that these results will be integrated in the current water quality process of the company thus allowing it to provide a better service to the society with reduced costs to the company.

Generally, both Random Forests and Support Vector Machines achieved the best performances. In particular, for the parameter Aluminum the Random Forest was significantly better than all others. A surprising result were the scores obtained by the Quantile Regression Random Forests. This model was developed specifically for this type of forecasting tasks. Nevertheless, it was not able to overcome the performance of more standard regression models, although the latter were used with pre-processed data to favor the prediction of quantiles. We did not observe any significant difference in terms of the windowing strategies for applying the models. This provides evidence for the absence of clear changes of regime on these time series.

A possible extension of this work is on the optimization of the configuration parameters of the models. Further optimization of these parameter values could eventually improve the results we have obtained. Another extension we are planning to carry out has to do with the spatial dimension of the data. In this research we analyzed only the temporal dimension of the water quality parameters, but it should be interesting to check whether these conclusions vary across the network by exploring the spatial information of the collected data.

## Acknowledgements

This work was partially supported by the FCT project MORWAQ (PTDC/EIA/68489/2006) and by a PhD scholarship of the Portuguese government (SFRH/BD/61795/2009) to Orlando Ohashi.

## REFERENCES


[1] A. Ailamaki, C. Faloutos, P.S. Fischbeck, M.J. Small, and J. VanBriesen, 'An environmental sensor network to determine drinking water quality and security', *ACM SIGMOD Record*, **32**(4), 47–52, (2003).

[2] L. Breiman, *Classification and regression trees*, Chapman & Hall/CRC, 1984.

[3] L. Breiman, 'Random forests', *Machine learning*, **45**(1), 5–32, (2001).

[4] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*, Cambridge Univ Pr, 2000.

[5] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, , and Andreas Weingessel, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2009. R package version 1.5-22.

[6] S. Haykin, *Neural networks: a comprehensive foundation*, Prentice Hall, 2008.

[7] David R. Hunter and Kenneth Lange, 'Quantile Regression via an MM Algorithm', *Journal of Computational and Graphical Statistics*, **9**(1), 60, (March 2000).

[8] C. Hwang and J. Shim, 'A simple quantile regression via support vector machine', *Lecture Notes in Computer Science*, **3610**, 512, (2005).

[9] R. Koenker and G. Bassett Jr, 'Regression quantiles', *Econometrica*, **46**(1), 33–50, (1978).

[10] Andy Liaw and Matthew Wiener, 'Classification and regression by randomforest', *R News*, **2**(3), 18–22, (2002).

[11] R. McGill, J.W. Tukey, and W.A. Larsen, 'Variations of box plots', *American Statistician*, **32**(1), 12–16, (1978).

[12] N. Meinshausen, 'Quantile regression forests', *The Journal of Machine Learning Research*, **7**, 999, (2006).

[13] Nicolai Meinshausen, *quantregForest: Quantile Regression Forests*, 2007. R package version 0.2-2.

[14] Claudia Perlich, Saharon Rosset, Richard D. Lawrence, and Bianca Zadrozny, 'High-quantile modeling for customer wallet estimation and other applications', *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, 977, (2007).

[15] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[16] J Taylor, 'Forecasting daily supermarket sales using exponentially weighted quantile regression', *European Journal of Operational Research*, **178**(1), 154–167, (2007).

[17] J W Taylor, 'Density forecasting for the efficient balancing of the generation and consumption of electricity', *International Journal of Forecasting*, **vol**, 22pp707–724, (2006).

[18] J W Taylor, P E Mcsharry, and R Buizza, 'Wind Power Density Forecasting Using Ensemble Predictions and Time Series Models', *IEEE Transactions on Energy Conversion, forthcoming*, 1–8, (1992).

[19] James W. Taylor, 'A quantile regression neural network approach to estimating the conditional density of multiperiod returns', *Journal of Forecasting*, **19**(4), 299–311, (July 2000).

[20] Terry M Therneau and Beth Atkinson. R port by Brian Ripley., *rpart: Recursive Partitioning*, 2009. R package version 3.1-44.