Prognosis of the long-term ageing behaviour of drinking water abstraction wells using machine-learning approaches

Forecasting tool for strategic planning and maintenance of drinking water wells

H.Schwarzmüller*[1] S. Schimmelpfennig[2], A. Sperlich[2] and M. Riechel[1],

[1] Kompetenzzentrum Wasser Berlin gemeinnuützige GmbH, Cicerostr. 24, D-10709 Berlin, Germany
(E-mail: *mathias.riechel@kompetenz-wasser.de*; *hella.schwarzmueller@kompetenz-wasser.de)*
[2] Berliner Wasserbetriebe AöR, Neue Jüdenstr. 1, D-10179 Berlin
(E-mail: *sebastian.schimmelpfennig@BWB.de*; *alexander.sperlich@bwb.de*)
* Corresponding author

Abstract
The Berliner Wasserbetriebe (BWB) are operating more than 650 vertical filter wells supplying the drinking water for the city's nearly 3.7 Mio. inhabitants from groundwater resources within the city limits. In order to keep performance and water quality as high as possible, these wells require regular monitoring and maintenance. The main reason for inefficient well performance is so-called well ageing caused by deposit formation due to multiply correlated biological, chemical and physical clogging processes in and around the well that decrease the yield for a given drawdown. In order to better understand the key drivers for well ageing and to project the loss of capacity for a given time ahead, machine learning (ML) approaches were applied to selected data from routine well monitoring. The statistical programming language R was used for automated data processing, feature selection and assessment of the importance of the selected variables, and finally for model training and prediction of future loss of well capacity. Four variables were identified to be highly significant predictor variables. Multivariate linear regression, logistic regression, decision tree, random forest and gradient boosting were applied, the latter performing best with a sensitivity of 94% and precision of 88%. The approach is now transferred into a well condition index to be included in a well management and reporting tool box developed in the frame of the H2020 project digital-water.city.

Keywords
drinking water wells, machine learning, rehabilitation efficiency, well ageing, well maintenance

INTRODUCTION

Drinking water production from groundwater is done with horizontal or vertical filter wells. The lifetime of such drinking water wells typically ranges between 20 and 50 years. Statistical evaluation of well data from drinking water wells in Berlin, Germany, showed for example an average well age of 34 years (Schwarzmüller et al., 2010). The capacity of wells, that is the yield for a given drawdown, however, decreases with time of operation. This effect is called well ageing and is due to the formation of deposits of biochemical origin (e.g. iron oxides formed by iron bacteria; Figure 1) or particulate matter (e.g. clogging with silt or sand).

Figure 1: left: clogged well, ochre deposits inside the screen; right: clogged pump, ochre deposits at the pump intake, both ©BWB

Maintenance, such as cleaning the pump and filter screen as well as gravel pack, prolongs the functioning by removing these deposits. To identify and prioritize maintenance needs, well condition is monitored during operation in regular intervals or on demand and comprises parameters such as flow rate, water levels, water quality, power consumption, etc. These data are stored in a well management database together with static information such as well design and construction, geological information, and analytical data from raw water samples. Goal was to combine automated data processing of routine monitoring data and well characteristics, site properties and operational data with machine-learning (ML) approaches to identify well ageing and decreasing well capacity and prioritize maintenance or reconstruction needs. The application was developed as one of 15 digital solutions implemented in the H2020 project digital-water.city (DWC) aiming at leveraging the potential of data for boosting water management in cities.

METHODOLOGY

The ML approach considered 6.308 data sets of a total of 994 currently operated and abandoned wells operated by the Berliner Wasserbetriebe (BWB) since the 1950s. Data were obtained from a db2 database with SQL scripts and transferred to csv files. The statistical programming language R (R Core Team, 2021) was used to define the core algorithms to (i) pre-process the well data turning them into a data structure providing the explanatory variables to the ML model, (ii) assess the importance of the variables and select model features, and (iii) train the ML model and predict future loss of well capacity based on selected well characteristics. 36 features (26 numeric, 10 categorical) were initially tested describing well characteristics (e.g. well age, construction material, screen diameter), site properties (well location, aquifer coverage, groundwater level variation), operational data (abstraction volumes, flow rates), past maintenance events (number of rehabilitation events, time since last rehabilitation) and raw water quality (e.g. total iron, dissolved oxygen, total phosphate).

Target variable was the prediction of a numeric value for the specific capacity (that is the quotient of flow rate and drawdown) relative to the capacity at the time of initial operation (Qs_rel; Figure 2).

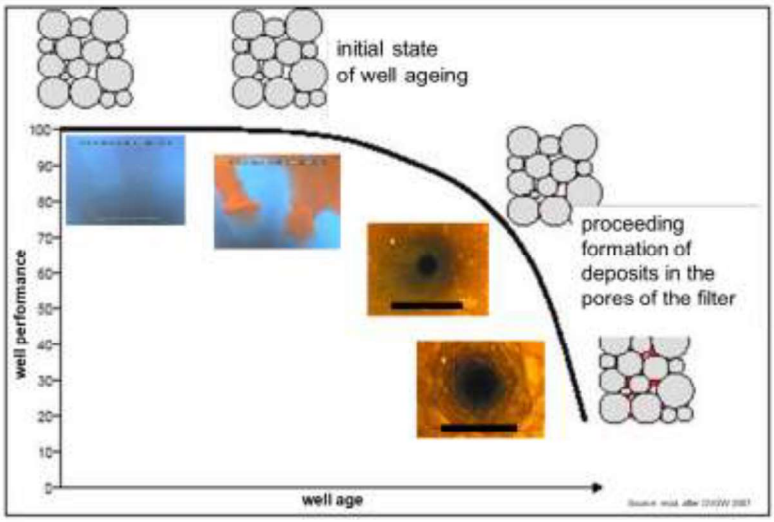

Figure 2: Typical Qs-curve (ageing curve) with decreasing specific capacity with time in operation (modified after DVGW, 2007)

RESULTS AND DISCUSSIONS

Based on the intercorrelation and variable importance tests, six of 36 variables were discarded as highly intercorrelated, another four were not contributing to the model accuracy. Of the remaining features, top five predictor variables were extracted from the random forest resulting in (i) well age, (ii) time since last rehabilitation, (iii) location, (iv) number of previous well rehabilitation events and (v) coefficient of variance in daily abstraction volume (Figure 3 left). 'Location' was further discarded in order to make the solution transferable to other well settings.

Training the model and applying it to the test data yielded a root mean square error (RMSE) of 15% and a coefficient of determination of $r^2 = 0.78$. The classification accuracy for specific capacity values < 80% was 94% (recall), with 12% wrong warnings (1 – precision) and 20% false positives (1 – specificity). The ML approach was thus rated well- applicable to forecast well ageing based on well and site-specific data (Figure 3 right). The trained model was subsequently used to predict the ageing curves for each single well of the test data set. Here too, the model showed a good fit to pumping test data from before and after past maintenance.

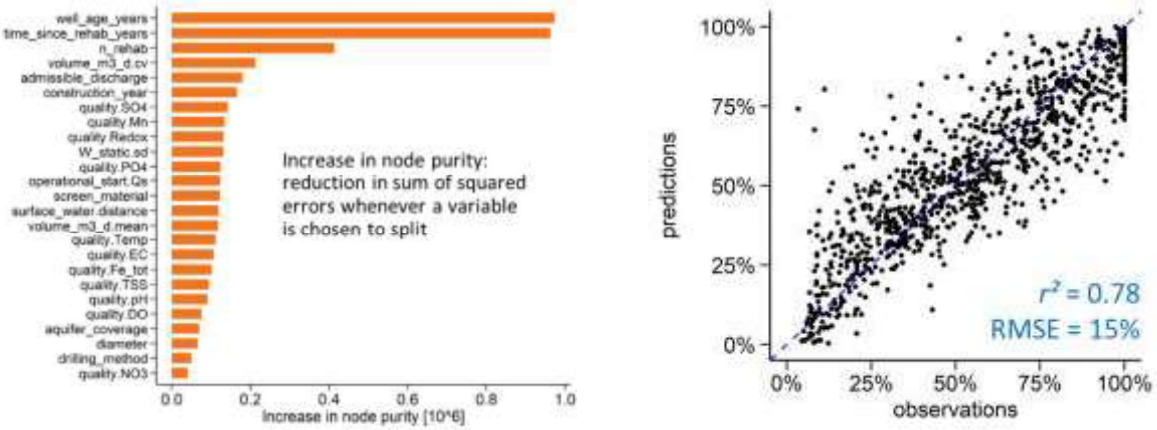

Figure 3: left: Results of random forest variable importance ranking; right: Model performance of gradient boosting approach with more than 78% of the variance in the observations explained by the model

CONCLUSIONS

Extended data collection and processing in combination with ML approaches provides data-driven analysis and identification of key variables related to the specific capacity development. One of the main advantages of the ML approach is that many variables could be included in the analyses and subsequently be narrowed down to a set of key variables, while no direct correlation was observed for single variables in previous research.

Gradient boosting showed a highly satisfying sensitivity and accuracy in the prediction and can assist well operators in planning well rehabilitations and renewals. Refinement of the solution within the DWC project will include further analyses such as clustering the ageing curves to narrow down preferred site conditions and factors that accelerate well aging. Data availability and remaining data gaps and/or pre-aggregated data showed to be a barrier and remain as crucial steps in proactive well maintenance.

ACKNOWLEDGEMENT

digital-water.city is a research project supported by the European Commission under the Horizon 2020 Framework Programme under the Gran Agreement No. 820954. The authors would further like to thank the colleagues from the Berliner Wasserbetriebe for their valuable contributions to the development of this digital solution.

REFERENCES

Cramér, H. (1946) *Mathematical Methods of Statistics. Princeton*: Princeton University Press, page 282 (Chapter 21. The two-dimensional case). ISBN 0-691-08004-6

Daniel, Wayne W. (1990) *Spearman rank correlation coefficient. Applied Nonparametric Statistics* (2nd ed.). Boston: PWS-Kent. pp. 358–365. ISBN 978-0-534-91976-4.

DVGW (2007) *DVGW-Arbeitsblatt W130* Brunnenregenerierung. 38p. DVGW. Bonn.

Friedman, J. H. (2001) *Greedy function approximation: a gradient boosting machine. Annals* of Statistics, 1189–1232.

Pearson, K. (1900) *On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50(5), 157–175.

R Core Team (2021) R: *A language and environment for statistical computing. R* Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwarzmüller, H., Orlikowski, D. and Grützmacher, G. (2010) *Survey on the implementation of DVGW - Drinking water well monitoring guidelines W125 in practice.* Bluefacts - International Journal of Water Management 2010, 40-46.