Applications of Advanced Analytics and Machine Learning

-

Optimising Asset Performance in the Utility Sector

by Astrid Bach Krabbe, Data Scientist

APX10

Abstract

The concept of machine learning in the context of asset management have long been used extensively and it is, by now, an established fact, that employing methods from these fields results in significant savings (Bluefield Research, 2018).

In this white paper, we will demonstrate how our solution is a game changer for utilities using a selection of machine learning methods as well as other relevant concepts. We will describe, based on specific cases, how these methods can be applied in the Utility Sector in order to reap the benefits of these technologies.

Introduction

The world is facing great challenges from climate changes and increasing urbanisation, both having a massive impact on the utility infrastructure established decades ago. This forces utilities to further optimize investments in their infrastructure.

At the same time new technological opportunities arise continuously as new methods are developed, and the existing data availability is expanded. The realization of the value, associated with the application of new computational methods for better decision making, is a key motivation for organizations.

In APX10 our ambition is to help utilities make optimal decisions while balancing their three major priorities:

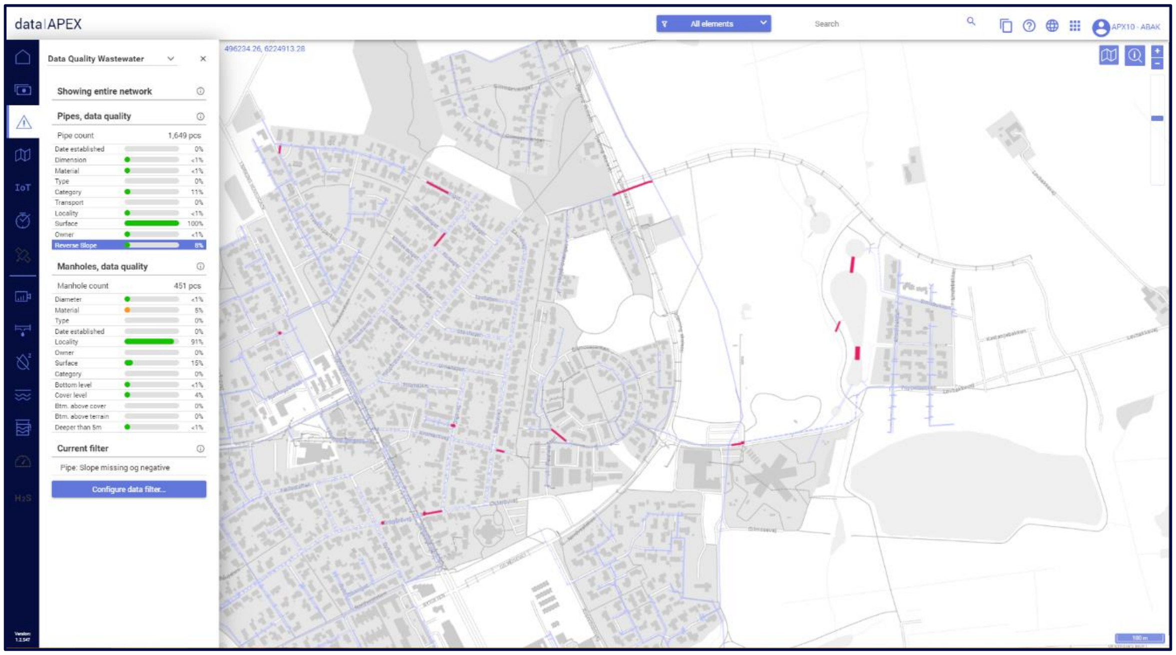
- Economy
- Consumers
- Climate and environment

Our efforts towards that goal is presented in our analysis platform data|APEX. Our solution offers an intuitive and data-driven approach to analytics and is identifying critical refurbishment potential that enables better decision making – ultimately optimising an often-regulated budget and improving customer service. We integrate statistics, machine learning and graph theory for predictive analyses and risk overview.

> Bluefield's forecasts indicate that advanced asset management solutions will save these utilities US$1.2 billion in annual CAPEX savings in 2018 and scale to US$7.3 billion in annual savings by 2027.
>
> Central to this shift toward more advanced asset management solutions is the increasing value in data collection, analytics, and visualization of network conditions and operations. Utilities are looking for ways to extend the service life of aging infrastructure assets and placing an increasing importance on data-based, predictive decisions, rather than being reactive.
>
> Bluefield Research (Bindler, 2018)

Available data

The cornerstone of our analyses and visualisations is the utility's own data on their physical assets (pipes, components, structures etc.). With a thorough understanding of the sector, we extract the data relevant to our analyses from the registration platform the utility uses, and we evaluate on the quality of these data. The result of this evaluation is presented in the *Data Quality Module* (shown in Figure 1), providing an approachable summary of the reliability of these data. The module enables the utility to visualize where data may, providing the opportunity to prioritize areas for improving the data quality.



Data Quality Wastewater

Showing entire network

Pipes, data quality

Pipe count	1,649 pcs
Date established	0%
Dimension	<1%
Material	<1%
Type	0%
Category	11%
Transport	0%
Locality	<1%
Surface	100%
Owner	<1%
Reverse Slope	8%

Manholes, data quality

Manhole count	451 pcs
Diameter	<1%
Material	5%
Type	0%
Date established	0%
Locality	91%
Owner	0%
Surface	15%
Category	0%
Bottom level	<1%
Cover level	4%
Btm. above cover	0%
Btm. above terrain	0%
Deeper than 5m	<1%

Current filter

Pipe: Slope missing og negative

Configure data filter...

Figure 1: *The Data Quality Module in data|APEX.*

We incorporate reports from CCTV inspections, leaks and bursts as part of the basis of our failure prediction modules for both sewer and water utilities. When leak data are combined with knowledge of different pipe features, such as age, material, location etc., it provides an excellent starting point for analysis of asset performance and life expectancy.

The use of both satellite data and aerial photographs provides dynamic, up to date knowledge of surface and terrain characteristics, such as surface coverage classification, soil wetness etc. When this is combined with weather data, a comprehensive representation of environmental and climatic impact can be achieved.

This representation, in turn, is supported by dynamic data from a multitude of IoT sources. These could be meters placed in the network to measure flow, temperature etc. More generally, it could be any type of sensor, which is connected to the internet, and thus is able to provide dynamic data on the features it measures.

Methods

In processing this abundance of data, we employ different statistical methods as well as experience-based risk matrices and machine learning algorithms. In this section, a selection of the methods used in the platform is described. The use of some of these methods will then be illustrated by the two cases below.

Random forest

Random Forest is a type of Machine Learning algorithm; a statistical model, which uses the concept of decision trees to calculate the possibility that an element with certain features falls into a certain category. As the name suggests, random forest uses an ensemble of decision trees to perform a classification of the elements into suitable classes. A class then consists of elements of similar attributes (or features) enabling the algorithm to make inferences for one element, from knowledge of the rest of the class.

The algorithm is first trained on a subset of the data, in order to create suitable classes. Then, the model is tested and validated using the remaining data. One use case for this algorithm, is the estimation of a pipe's condition. This is used in the module *Network Condition Assessment*, which takes a variety of pipe data as input, in order to estimate the target variable, the condition of the pipe.

Time series clustering

Clustering describes a range of methods that aims to group objects into sets (or clusters). More specifically, when data is in the form of a time series (that is, regularly sampled data points of one measurement type, consecutive in time) time series clustering is the process of dividing a larger group of time series into clusters of relatively uniform time series with similar characteristics. This means that, among other use cases, time series clustering is very well suited for detecting abnormal behaviour. In data|APEX we use this e.g. for estimation of dry weather flow variations in separate sewer system for estimating rainfall derived inflow and infiltration.

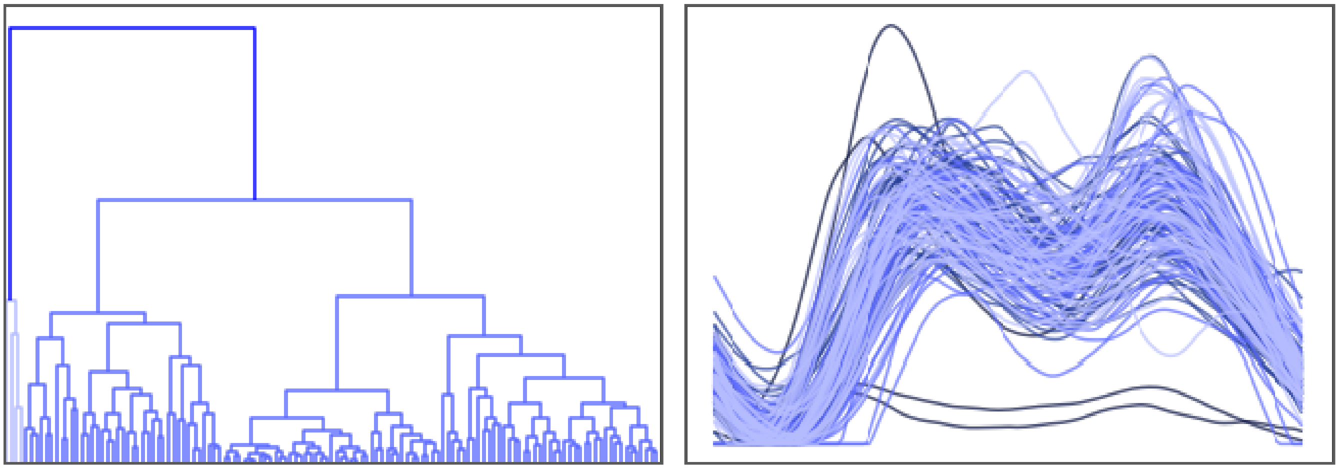


Figure 2: *An example of the results from using a clustering algorithm on a set of time series.*

An example demonstrating the use of time series clustering is shown in Figure 2. These time series represent pumped wastewater at a pump station on weekdays. On the right is a dendrogram visualising the arrangement of the clusters. On the left, the actual time series are shown, their colours designating which cluster they belong to.

Bayesian Statistics

Bayesian statistics are based on Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

stating that the probability of observing event A, given event B equals the probability of B occurring given A multiplied by the probability of the occurrence of A divided by the probability of the occurrence of B. Or, in other words, the posterior probability, can be obtained, given a set of observations, by updating the prior probability using these observations.

In Bayesian statistics, contrary to classical frequentist statistics, the probability is replaced by a probability distribution, specifying the degree of belief in all possible states, that the variable of interest may take. This view of the probability is inherently uncertain, and the posterior distribution provides an adequate description of this uncertainty. Furthermore, this makes possible the combination of a (possibly qualitative) preconception with actual observations of a given property.

Markov Chains and the Monte Carlo Method

The Monte Carlo method is a method for simulating the solution of a complex problem using random sampling from an appropriate probability distribution. Using Monte Carlo simulations, one may benefit from the increasing processing power of modern computers, to estimate a solution where problems are hard (or impossible) to solve analytically. Markov chains, on the other hand, are processes that satisfies the Markov property, meaning that each state depends only on the previous state and not on the entire history of the system. The Markov Chain Monte Carlo method thus is a combination of these two concepts, i.e. a Monte Carlo simulation in which each state is dependent only on the previous state. This can be used to e.g. estimate and update probability distributions in the context of Bayesian Statistics.

Case 1: Network Condition Assessment

Problem

Traditionally, network pipe replacement is prompted by one of the following: pipe age exceeding a given number of years; or network pipe failure (e.g. break, collapse or critical degree of deterioration). This approach has two disadvantages. Firstly, replacing pipes at a fixed age means replacing pipes that, in many cases, could have continued to function, with little or no loss of performance, for several years. Secondly, this approach is no guarantee against failures occurring before the end of the expected lifetime, resulting in great inconvenience to consumers and increased ad hoc repair costs.

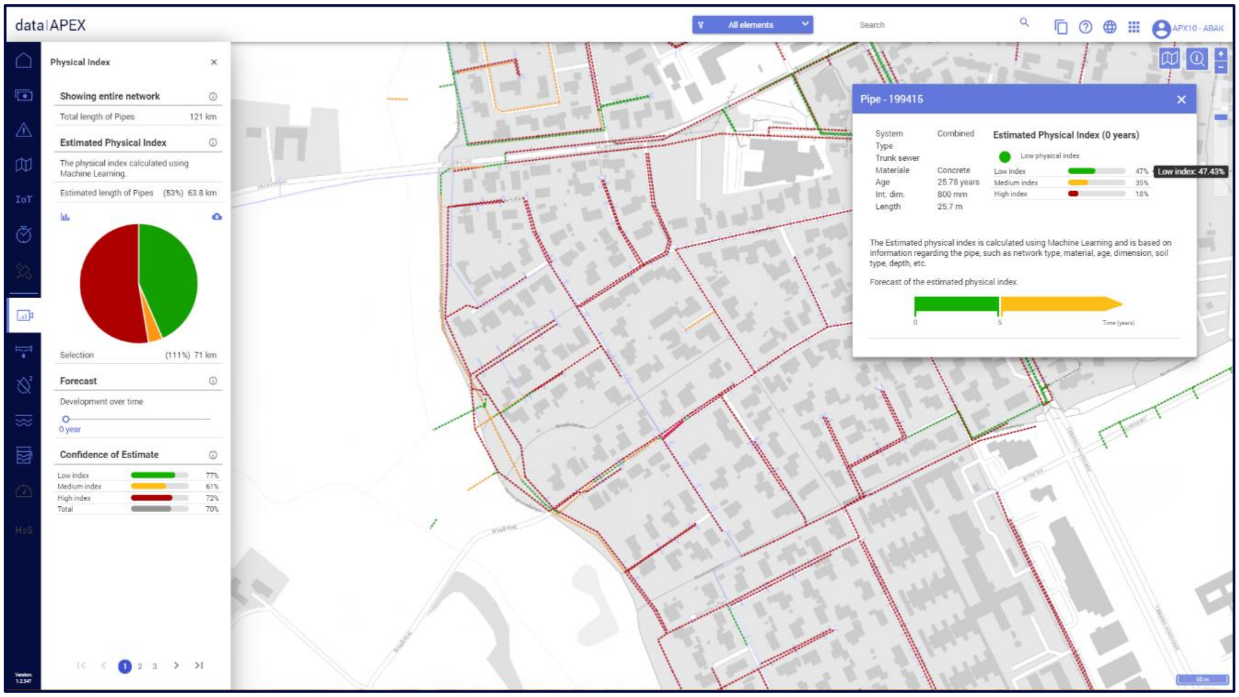
Another strategy for pipe replacement is to regularly inspect the pipes and replace the pipes that are categorised as being in poor condition. This method is effective, but inspections are expensive and only give the condition of a fraction of the network, at a single point in time. Inspections therefore must be repeated regularly to keep track of the network condition. Monitoring the entire network often enough to be able to prevent breaks isn't financially viable, so the Utility must choose between using a different method for estimating the network condition or employing an effective strategy for prioritising which pipes to inspect.

Solution

The *Network Condition Assessment Module* uses a random forest algorithm to determine the condition of uninspected pipes on a basis of historical pipe inspections in combination with pipe data.

As pipe age is a significant indicator of a pipe's condition, it is possible to adjust this variable in order to predict how the pipe's condition will be in the future. This provides an indication of the rate of decay of individual pipes and, combined with the estimate of pipe condition, equips the user/utility with a sound basis on which decisions on CCTV inspections can be made.

Due to the statistical nature of the random forest algorithm, the larger the dataset, the better the quality of the inference, at least up to a certain point. In the data|APEX platform, this is accommodated by combining data from multiple utilities into a large dataset, which is then used to train the model. This network effect greatly enhances the reliability of the model in comparison to a model trained on one utility's data alone. This network effect can enhance the reliability of the model in comparison to a model trained on a single utility's data.



Physical Index

Showing entire network

Total length of Pipes	121 km

Estimated Physical Index

The physical index calculated using Machine Learning.

Estimated length of Pipes	(53%) 63.8 km
Selection	(111%) 71 km

Forecast

Development over time

0 year

Confidence of Estimate

Low index	77%
Medium index	61%
High index	72%
Total	70%

Pipe - 199415

System	Combined
Type	
Trunk sewer	
Materiale	Concrete
Age	25.78 years
Int. dim.	800 mm
Length	25.7 m

Estimated Physical Index (0 years)

● Low physical index

Low index	47%
Medium index	35%
High index	18%

Low index: 47.43%

The Estimated physical index is calculated using Machine Learning and is based on information regarding the pipe, such as network type, material, age, dimension, soil type, depth, etc.

Forecast of the estimated physical index.

Figure 3: The Wastewater Condition Assessment Module in data|APEX*.*

An important consequence of the use of this algorithm to estimate the pipe's condition, is that the significance of the different features is not manually set based on assumptions and individual preference. Rather, it is determined by the algorithm, based on data alone.

In addition to pipe features, the algorithm relies on existing CCTV inspections for validation of the estimated condition. In the *Network Condition Assessment Module*, the results of previous CCTV inspections are shown in combination with the estimates, in order to give the most reliable prediction of the pipe's condition.

Case 2: Risk Analysis

Problem

When applying Asset Performance Management in an impact-based approach, we not only consider the probability of asset failure, but the consequences of a possible failure as well. An impact-based perspective offers a far better basis for risk analysis and decisions on refurbishment, as the risk of major incidents is a concern usually on a par with economic considerations.

> **Asset Performance Management (APM)** encompasses the capabilities of data capture, integration, visualisation and analytics tied together for the explicit purpose of improving the reliability and availability of physical assets. APM includes the concepts of condition monitoring, predictive forecasting and reliability-centred maintenance (RCM).
>
> *Gartner, Definition of Asset Performance Management* (Gartner, 2019)

Solution

The *Risk Analysis* uses experience-based likelihood and consequence matrices to evaluate the burst risk associated with each pipe. The likelihood and the consequence matrices are combined, in order to determine the overall risk associated with a possible leak or break. The likelihood is calculated from knowledge of a variety of pipe features, such as e.g. pipe age or pipe material. The consequence of a leak or a break is mainly determined by its effect on consumers. This effect can be estimated from features such as e.g. pipe dimension and location, from knowledge of the network e.g. number and type of consumers connected to a given pipe and from information on high risk areas such as highway or railway crossings.

Assessing Likelihood of Failure using Bayesian Statistics and Markov Chain Monte Carlo

As inspections of water pipes are rarely performed among Danish utilities, an approach, such as the one described in Case 1, is infeasible, lacking the substantial amounts of condition data from the inspections.

Instead, we have employed an approach based on Survival Analysis[1] and Bayesian Statistics. In the Bayesian framework, one considers inferences as beliefs rather than truth. These beliefs can then be updated as observations are made. This enables us to use the accumulated know-how of experienced professionals in the utilities to establish a plausible prior as an initial guess of the Likelihood of Failure (LoF) of water pipes. As pipe failures occur in the different utilities, data on these failures are recurrently being fed to the model. The ongoing inflow of observations of breaks continually improves the model, providing ever better estimates of the probability of failure.

In order to estimate the posterior distribution, we use a Markov Chain Monte Carlo simulation in order to draw samples from the posterior distribution, as described in the methods section.

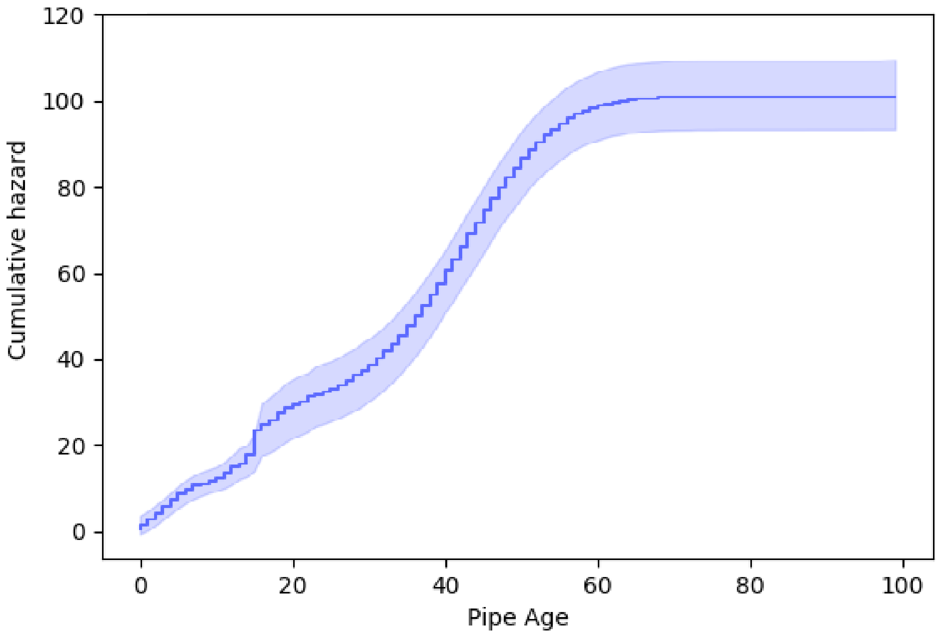


***Figure 4:** Likelihood of Failure as a function of pipe age for a PVC pipe with dimension 0 -100 mm.*

An example of a hazard function for a PVC pipe obtained using this method is shown in Figure 4. The blue curve in Figure 4, shows the mean value of the cumulative hazard, whereas the light blue area

[1] Using Cox's proportional hazards model. See (Ibrahim, Chen, & Sinha, 2001) and (Cox, 1972)

shows the 95% HPD interval[2]. From this figure, it is seen that the accumulated hazard is most likely to exceed 50% at a pipe age of around 40 years. This is consistent with the prior, but it is likely that this prior estimate was set a bit low[3]. If that is the case, then this estimate may well change, as more observations are added to the model.

This result reflects both the initial belief of the experienced technician as well as actual observations of pipe failures. The inherent uncertainty of the estimate is explicitly given by the 95% HPD interval, which gives the user an accurate account of the credibility of the method.

Assessing Consequence of Failure using specialized comprehension of the pipe network

In APX10 the pipe network is the central element in all our analyses. As such, the criticality of a pipe failure is quite naturally derived from knowledge of the network, demography and consumers that may be affected. This is done by combining network data from the Utility's own database with IoT sensors, satellite data, aerial photographs and other public data sources. These data sources are then analysed using routing algorithms and clustering. Relevant features are extracted from aerial or satellite imagery using remote sensing and image recognition.

Estimating criticality

Likelihood of Failure is combined with the corresponding Consequence of Failure to produce a criticality estimate. This is done using matrices that specifies how these two concepts should be related.

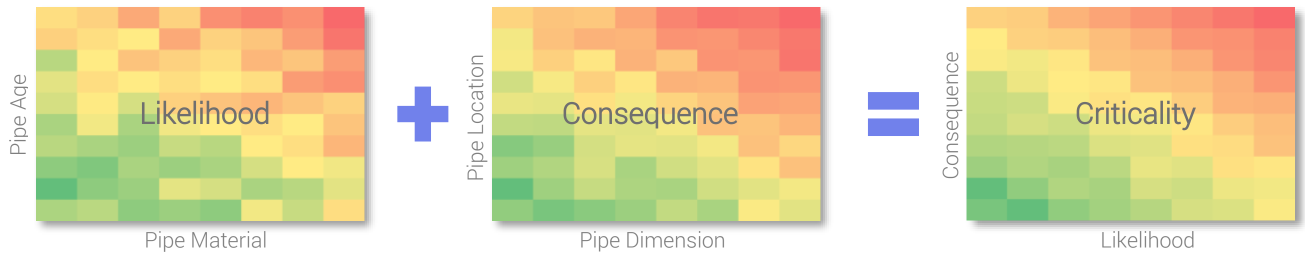


***Figure 5:** Visualisation of the criticality matrix obtained by combing the possibility and consequence matrices. This example uses two dimensional matrices for simplicity, but*

[2] The credible interval or the HPD (Highest Posterior Density) denotes the interval of values that are most probable. For instance, a 95% HPD interval signifies that the value of the estimate has a 95% probability of being in this interval. Thus, the credible interval in Bayesian statistics corresponds to the confidence interval of frequentist statistics.

[3] The matrices used for the prior were originally used to locate pipes at risk. As, in that context, it is safer to underestimate survival times than to overestimate them, these prior curves may have been somewhat skewed towards shorter survival times.

This approach is further expanded, so that the criticality may include all aspects that are deemed relevant to this criticality estimate. An example of this is nuisance and costs related to repairs or refurbishment. These considerations are included in the criticality estimate in order to give an exhaustive estimate of the implications of choosing whether to replace a pipe or not.

Conclusion

Addressing the demands of customers as well as requirements from governments or other regulatory instances, data|APEX assist utilities by enabling the transformation from a reactive to a predictive paradigm for managing their assets. This shift is enabled by the emergence of new technologies as well as the data that has accumulated since the dawn of computers in the second half of the twentieth century.

The result is a platform that provides the utility with the opportunity to evaluate and reduce risks relating to their assets. While the solution helps utilities optimizing budgets and reducing costs, the reduced risk of failure of critical infrastructure improves customer relations and helps utilities offer reliable services.

References

Bindler, E. (2018, May 8). *What can Advanced Asset Management Mean for Municipal Water?* Retrieved from bluefieldresearch.com: www.bluefieldresearch.com/can-advanced-asset-management-mean-municipal-water

Bluefield Research. (2018). *Advanced Asset Management for Municipal Water: Global Trends & Strategies, 2018 – 2027.* Bluefield Research.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*.

Gartner. (2019). *Asset Performance Management (apm)*. Retrieved from gartner.com: www.gartner.com/en/information-technology/glossary/asset-performance-management-apm

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian Survival Analysis.* Springer.

APX10

Phone: +45 87 38 61 66
Email: info@apx10.com

APX10 A/S
Jens Juuls Vej 16
DK-8260
Viby J
Denmark

Published November 2019
2nd edition