Identifying biases in deterioration models using synthetic sewer data

A. Scheidegger and M. Maurer

ABSTRACT

The assessment and validation of sewer deterioration models is difficult because reliable data are missing. This makes it hard to find the most suitable model for a particular application. A network condition simulator (NetCoS) is used to generate synthetic sewer data for defined test scenarios. Thereby, the deterioration and replacement of pipes, the expansion of the sewer network, and classification errors are considered. Based on such synthetic data, deterioration models are calibrated and their results compared with the predefined scenario. While this approach is not capable of proving that a model performs correctly on a real application, it highlights the strengths and weaknesses of a model. The influence of condition classification errors and the age of the sewer system is investigated for two deterioration models. The results show, that classification errors can introduce substantial biases in the parameter estimation of the Markov model while in comparison the applied cohort model is fairly robust. Young sewer systems with fewer pipes in bad condition states on the other hand, have a very strong influence on the parameter uncertainties of the cohort model while the Markov model proved to be less sensitive.

Key words | cohort model, Markov model, NetCoS, semi-Markov chain, sewer deterioration, synthetic data

INTRODUCTION

Substantial parts of many urban drainage systems are reaching their end-of-life. To assess the rehabilitation needs in the years to come, reliable estimations of the deterioration of the buried infrastructure are required (Davies et al. 2001; Wirahadikusumah et al. 2001; Maurer 2009).

To estimate the deterioration behavior of sewers from inspection data various statistical deterioration models have been proposed in the literature (for an overview see Kleiner & Rajani 2001; Tran 2007; Ana & Bauwens 2010). However, current investigations show that such models are hardly ever validated (Ana & Bauwens 2010). Statistical deterioration models always contain uncertainties in the parameters. Furthermore, all models make (implicit) assumptions about the deterioration behavior of the sewers and the inspection data. For the available data these model assumptions are hardly completely applicable and therefore, additional uncertainties and biases may be introduced. Such uncertainties are examined best by using simulated sewer data. Simulated data have the advantage that the ‘truth’ is known and can be compared with the estimation of a deterioration model. With this approach weaknesses of the deterioration model can be identified and it allows us to quantify the influence of input uncertainties.

To simulate synthetic data sets of sewer systems we developed the Network Condition Simulator (NetCoS) (Scheidegger et al. 2011). NetCoS generates complete and error-free data sets based on a predefined scenario. A scenario consists of the parameters of the deterioration, expansion and replacement processes. To investigate a deterioration model, first data are simulated with NetCoS, then the deterioration model is fitted to this data. The results of the deterioration model can then be compared with the predefined scenario of the data generation. By repeating these steps, parameter uncertainties of the deterioration models can be quantified. If scenarios for data generation are defined that violate some assumptions of the deterioration model, potentially resulting biases and uncertainties in the result of the deterioration model can be unveiled. An example is the investigation of the influence of misclassified condition states (CS) in the data on the deterioration model parameters.
In this paper we used the data generator NetCoS to test two common deterioration models, the Markov model and cohort model, on data deficits. First we investigated the influence of errors in condition state classification on the parameter estimation of models. Secondly, we examined how the overall age of a sewer system influences the deterioration models. The results show that the two deterioration models behave distinctly differently. The remainder of this paper is structured as follows. First, the data generator NetCoS and the formulation of the classification error are described briefly, followed by a demonstration of NetCoS with the two examples. Finally, the strength and limitations of our approach are discussed.

NETCOS: DATA GENERATOR

The intention of this section is to give a short description of the internals of NetCoS. More details are given in Scheidegger et al. (2011). The size and condition of a sewer system is the result of four main processes: (i) the physical deterioration of the pipes, (ii) the replacement of pipes in the past, (iii) the expansion of the network, and (iv) the rehabilitation of pipes. Based on the first three processes NetCoS generates randomly hypothetical but complete and ‘error free’ data sets of a sewer system. The data contain the condition of every pipe for all ages. The condition of a pipe is represented by condition states on an ordinal scale from 1 to 5, where 1 is the best condition and 5 the worst (corresponding with WRc 1986). Those data are the input for a simple error model which simulates the possibility of assigning the wrong CS. For the sake of simplicity, rehabilitation of pipes is ignored in this paper.

The deterioration behavior for the data generation is defined by a set of survival functions. A survival function describes the probability that a pipe has a certain CS or better at a given age. These probabilities are converted into the transition probabilities of a semi-Markov chain, which is used to sample the CS or its change of every pipe section in each time step. Due to the semi-Markov approach any survival functions can be used. The replacement of pipes is simulated with a CS depending on annual replacement probability. Even pipes in a good CS can be replaced, considering the fact that pipes are also replaced for hydraulic reasons or due to synergetic effects in larger construction projects. If a pipe is replaced, a new pipe is ‘built’ instead. Additionally, at every time step new pipes (typically in the best CS) can be added to represent the expansion of the sewer system. For this paper we added an inspection model that assigns an error into the condition rating. An ‘error matrix’ is defined, which describes the probability of the inspection outcome given the ‘true’ CS.

Deterioration

The time $T_{CS\theta}$ a pipe section spends in CS $k$ is modeled as a random variable. Consequently, $T_{CS\theta}$, the total time a section has spent in CS 1 to CS $j$, is a random variable, too:

$$T_{CS\theta} = \sum_{k=1}^{j} T_{CS\theta}, \quad j = 1, \ldots, 4$$

The probability that a pipe section of age $\theta$ has CS $j$ or better is expressed by the survival function $S_j(\theta)$, which is related to the cumulative distribution function $F_j(\theta)$ of $T_{CS\theta}$:

$$P(T_{CS\theta} > \theta) = 1 - F_j(\theta) = S_j(\theta)$$

To describe the transition probability among the five CSs used in this paper, four survival functions $S_j(\theta)$, $j = 1, \ldots, 4$ are required. The survival functions can be chosen from any distribution as long they satisfy the condition:

$$S_j(\theta) > S_{j-1}(\theta), \quad j = 2, \ldots, 4$$

for all $\theta > 0$. For these survival functions it is assumed that no replacement of pipe sections occurs. Therefore, the four survival functions represent the distribution of the physical lifespan. The difference $S_j(\theta) - S_{j-1}(\theta)$ is equal to the probability that a pipe with age $\theta$ is in CS $j$.

To predict the condition of a single pipe section at age $\theta_2$ given the CS at age $\theta_1$, a semi-Markov chain approach is used. Unlike homogeneous Markov chains, the transition probabilities of semi-Markov chains depend on the time the process has already spent in a specific state. This allows the modeling of a stochastic deterioration process that follows the predefined survival functions. The probability for a transition from CS $j$ to CS $(j + 1)$ in a small time interval $\Delta$ is derived from the survival functions following the description of Kleiner (2001):

$$P_{ij+1}(\theta) = \begin{cases} \frac{f_1(\theta) \cdot \Delta}{S_1(\theta)}, & j = 1 \\ \frac{f_j(\theta) \cdot \Delta}{S_j(\theta) - S_{j-1}(\theta)}, & j \in \{2, 3, 4\} \end{cases}$$

where $f_j(\theta)$ is the probability density function (pdf) of $F_j(\theta) = 1 - S_j(\theta)$. This allows us to set up a (semi-)Markov model with a time depending transition matrix to calculate...
the state probabilities for a certain age given a previous state. At every time step a CS is sampled for every pipe using these probabilities.

**Replacement**

Pipes in all condition classes have a chance of replacement. Examples for condition independent replacements are hydraulic capacity problems or coordinated construction efforts with other utilities. Additionally, pipes in a poor CS are more likely to be replaced. The model emulates this by assigning a replacement probability to every CS. The probabilities that a sewer line of a particular CS is replaced within the next year are expressed with the vector

$$p_{rep} = (p_{rep_{CS1}}, p_{rep_{CS2}}, p_{rep_{CS3}}, p_{rep_{CS4}}, p_{rep_{CS5}})$$

At the end of year $t$, new pipe sections for all replaced pipe sections are built. The replacement probability of sections in poor condition is generally higher compared with sections in good condition. For the examples in this paper, replacement was ignored by setting all replacement probabilities to zero.

**Network expansion**

Typically, a sewer system is not built at once but is likely to have grown over time. In NetCoS, this is taken into account by an annual expansion of the sewer system by a length $L_{exp}(t)$. At the end of year $t$, new pipe sections with a total length $L_{exp}(t)$ are added to the inventory.

**Classification error**

Classification errors by inspections are modeled with a probability matrix $J$. The element $J(i,j)$ denotes the probability that the inspection results in CS $j$ given the actual CS is $i$. The inspected CSs are sampled using these probabilities. If $J$ is symmetric, under- and overestimations have the same probability.

**APPLICATION OF NETCOS**

NetCoS can be used to examine the influence of incomplete data, violation of model assumptions, statistical uncertainties, and errors in the data on a deterioration model. Figure 1 shows the typical workflow: (1) the parameters of a test scenario (survival functions, growth rate, replacement strategy, matrix of classification error) are defined; (2) a complete data set is generated, whereof a realistic subset is used to fit a deterioration model; (3) the estimations of the deterioration model are compared with the predefined survival functions. NetCoS is stochastic and therefore, repeating data generation and model fit allows the identification of confidence intervals of the parameters and predictions to represent the statistical uncertainty.

In the following section, first two deterioration models are described briefly, a Markov model and the cohort model. Then, using NetCoS, the influence of classification errors and the age of the sewer system on the estimated survival functions are investigated for both deterioration models. Note however, that these examples illustrate the application of NetCoS and are not intended as a comprehensive comparison of the two models.

**Deterioration models**

**Cohort model**

To apply the cohort survival model proposed by Baur & Herz (2002), the data on pipe sections with a similar expected deterioration behavior are grouped together in so-called ‘cohorts’. For each of these groups the model is fitted separately. A special survival function for the network infrastructure suggested by Herz (1995) is used:

$$S_j(\theta|\beta_j) = \frac{\beta_{j,1} + 1}{\beta_{j,1} + e^{\theta j \beta_{j,2} - \beta_{j,3}}}$$

All elements of the parameter vector $\beta_j = (\beta_{j,1}, \beta_{j,2}, \beta_{j,3})$ have to be positive as well as $\beta_{j,2} < 1$ and $\beta_{j,3} \leq \theta$. The aging factor $\beta_{j,1}$ affects the speed of the aging at the beginning, $\beta_{j,2}$
determines the deterioration rate and $\beta_{i,j}$ is the resistance time before any deterioration occurs.

The parameters are estimated by minimizing the absolute distances between the survival functions $S_j(t|\theta|\beta_j)$ and the percentage of pipe sections which have at least a CS $j$:

$$
\sum_{t_built=1}^{t} \text{abs}\left( S_j(t - t_{built}|\beta_j) - \frac{L_{CS:j,active}(t_{built})}{L_{tot}(t_{built})} \right)
$$

where $t$ is the current year, $t_0$ the construction year of the oldest pipe section and $\beta_j$ the parameter vector of the survival function $S_j$. $L_{CS:j,active}(t_{built})$ is the length of all pipe sections built in year $t_{built}$ which are still in use and have CS $j$ or better. The length of all pipe sections (replaced or not) built at $t_{built}$ is $L_{tot}(t_{built})$. The model cannot incorporate data of successive inspections.

### Markov model

Dirksen & Clemens (2008) model the sewer conditions of a Dutch town using three condition classes and a discrete, homogeneous Markov chain. As a consequence the probability of changing a condition class depends only on the current state, but not on the time already spent in this class. Adapting their method to five condition classes’ results in the following transition probability matrix $B$:

$$
B = 
\begin{pmatrix}
1 - \sum_{k=2}^{5} b_{1,k} & b_{1,2} & b_{1,3} & b_{1,4} & b_{1,5} \\
0 & 1 - \sum_{k=3}^{5} b_{2,k} & b_{2,3} & b_{2,4} & b_{2,5} \\
0 & 0 & 1 - \sum_{k=4}^{5} b_{3,k} & b_{3,4} & b_{3,5} \\
0 & 0 & 0 & 1 - b_{4,5} & b_{4,5} \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

The probabilities $b_{i,k}$ that a pipe section changes from CS $j$ to CS $k$, $j < k$, are estimated by maximizing the likelihood function, which requires data from two consecutive inspections of the same pipe sections. $L_{CS:j-CS:k}(n)$ is the total length of all pipe sections which have changed from CS $j$ at the first inspection to CS $k$ at the second inspection after $n$ years. The logarithm of the likelihood function is:

$$
\log L(B) = \sum_{n=1}^{5} \sum_{j=1}^{5} \sum_{k=1}^{5} \log((B)^{n})L_{CS:j-CS:k}(n)
$$

For this Markov model, no survival functions are directly estimated. However, they can be easily derived as the predicted state probabilities after $\theta$ years of a pipe section with CS 1 at age 0.

### Example 1: Influence of condition classification errors

Sewer pipes are mostly inspected visually, based on CCTV images. Dirksen et al. (2011) describe the assessments of sewers in three steps which are all susceptible to errors: first the recognition of defects followed by their description and finally the interpretation of the inspection reports and assigning a CS. The uncertainty introduced thereby can be considerable (Müller 2006). We are using NetCoS to examine the sensitivity of the deterioration models regarding classification errors. Based on data of 213 pipes inspected twice within three years (from Hüben 2002), the error matrix $J_H$ was derived:

$$
J_H =
\begin{pmatrix}
0.54 & 0.18 & 0.08 & 0.10 & 0.10 \\
0.33 & 0.29 & 0.09 & 0.13 & 0.16 \\
0.22 & 0.13 & 0.22 & 0.13 & 0.30 \\
0.04 & 0.18 & 0.21 & 0.29 & 0.28 \\
0 & 0.02 & 0.20 & 0.07 & 0.71
\end{pmatrix}
$$

The element $J(i,j)$ denotes the probability that the classification results in CS $j$ given that the actual CS is $i$. However, for the sake of simplicity a simpler and symmetric probability matrix $J_S$ for the classification error was applied for the following examples. For symmetric probability matrices under- and overestimation of the CS have the same probability and therefore help to illustrate our case more clearly. The error matrix $J_S$ leads to qualitatively the same results as $J_H$ but generates smaller biases in our examples (results not shown).

With NetCoS data for hundred-year-old sewer systems have been generated without replacement and an annual extension of 20 pipes (each 50 m length). Starting with no sewer, the final systems consist of 2,000 pipes with a total length of 100 km. This leads to a uniform age distribution where all condition states are well represented. The survival functions for the data generation with NetCoS were chosen so that an ideal parameter estimation could be achieved if the data are free of errors. That means the survival function of the fitted deterioration model would fit the given survival function of NetCoS perfectly (results not shown). This choice isolates the influence of the classification error. In order to achieve that, Herz and Erlang survival functions were used for the cohort model and the Markov model, respectively (parameters are listed in Table 1). For the
Markov model an inspection frequency of 10 years was assumed.

The influence of the condition classification error on the cohort model is shown in Figure 2. Thereto, 150 data sets with classification errors were generated. The survival functions are estimated flatter (i.e. a higher variance of the life-span) if the data contain classification errors. The survival function $S_1$ between CSs 1 and 2 is slightly underestimated. This is because some pipes in CS 1 have been misclassified in a worse CS while a misclassification into a better CS was not possible as no better CSs exist. The opposite happens for $S_4$ between CS 4 and 5; pipes in the worst CS can only be misclassified with a better CS, hence the life-span is overestimated. The effect is more pronounced compared with $S_1$ as additional uncertainty stems from the extrapolation of $S_4$ (only data for a hundred years are available). This also explains why the error in the ‘lower’ part of the curve is larger.

The Markov model requires data of two consecutive inspections. For calibration, the length of all pipes changing $\Delta CS = CS(t_2) - CS(t_1)$ in between is used (see Equation (9)). In our case this is done by sampling the synthetic data set at two time steps. A symmetric classification error means that there is a probability that a pipe is incorrectly improving its CS. The model cannot deal with these cases where $\Delta CS < 0$. Instead of simply removing data of such pipes, somewhat better results can be obtained by assuming that the CS did not change and that the CS of the first inspection is correct.

Figure 3 depicts the survival functions estimated by the Markov model on 150 simulated data sets containing classification errors. All survival functions clearly underestimate the life-span. As for pipes with spuriously improving CS's no deterioration is assumed, hence, they cannot correct for pipes with too large observed deterioration. This error adds up for each survival function so that $S_4$ has the largest deviation.

**Example II: Influence of the overall age of a sewer system on prediction accuracy**

The identifiability of the parameters of a deterioration model depends on a sufficient number of pipes in all condition states. Intuitively it is clear that condition data of a very young sewer system cannot provide good information of the transition of CS 4 to CS 5 as pipes in poor CS are rare. To quantify this effect, data of sewer systems with different ages (30, 40, … 100 years) has been generated with NetCoS (see Table 1 for parameters). Analogous to example 1, the simulation starts with 0 pipes. The annual expansion has been chosen so that all final systems consist of 2,000 pipes (100 km) and 20,000 (1,000 km), respectively. No pipe replacement was considered for the data generation.

The variability of the results due to poorly identifiable parameters is characterized by the 80%-interquantile range (the difference between the 10%- and the 90%-quantile) of the median age for the transition of CS 4 to CS 5 in Figure 4. The interquantile ranges are based on 200 repetitions, slightly more than example 1 to guarantee reproducible results. The parametric survival functions can be calculated for every age based on the estimated parameters. However, such extrapolations for high ages are highly uncertain if the parameters are estimated on data of a young sewer system. The results show clearly that the cohort model is much

### Table 1 | Parameters of the survival functions used for NetCoS

<table>
<thead>
<tr>
<th></th>
<th>Herz distribution</th>
<th>Erlang distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$\beta_1 = (0.7, 0.05, 0)$</td>
<td>$\lambda = 4.591 \cdot 10^{-2}$, $n = 1$</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$\beta_2 = (5, 0.05, 0)$</td>
<td>$\lambda = 4.591 \cdot 10^{-2}$, $n = 2$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$\beta_3 = (18, 0.05, 0)$</td>
<td>$\lambda = 4.591 \cdot 10^{-2}$, $n = 3$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$\beta_4 = (50, 0.05, 0)$</td>
<td>$\lambda = 4.591 \cdot 10^{-2}$, $n = 4$</td>
</tr>
</tbody>
</table>

**Figure 2** | Influence of the condition classification error on mean of the survival functions estimated by the cohort model (solid lines) and the corresponding 80% interquantile ranges (dashed lines). The predefined survival functions used for data generations are shown as the borders of the shaded areas.

**Figure 3** | Influence of the condition classification error on the mean of the survival functions estimated by the Markov model (solid lines) and the corresponding 80% interquantile ranges (dashed lines). The predefined survival functions used for data generations are shown as the borders of the shaded areas.
more sensitive to the age of the sewer system than the Markov model.

DISCUSSION

The Network Condition Simulator (NetCoS) was used to: (i) examine the influence of classification errors, and (ii) the overall age of the sewer system on two different types of deterioration models.

Influence of condition classification errors

The results show that the cohort model is less affected by classification errors than the Markov model (Figures 2 and 3). The Markov model is sensitive to the number of CS changes over the inspection period. Errors in the condition rating alter artificially the number of CS changes. Without classification error, 6% of the pipes deteriorate within the examined 10 years into the worst CS. The classification error in the example above increases this number to 13%. The Markov model cannot ‘handle’ improving CS and therefore this information has to be discarded and introduces a systematic bias. False improvements of the CS are ignored while false degradation of the CS leads to an overestimation of the deterioration probability. Other data screening strategies, such as reversing the assumption (the second CS is correct) or deleting these data, lead to estimation biases that are similar or even more pronounced, respectively (data not shown).

The cohort model also shows a systematic bias, although substantially less strong. The bias is especially visible, where the error is not symmetrical; e.g. a pipe in the worst CS can only be assessed better but not worse. Considering that the probability of assigning the ‘correct’ CS is between 0.5 and 0.75, the cohort model is surprisingly robust against this type of data error.

For the sake of simplicity the presented examples use a symmetric error matrix. This enables us to show and compare the effects of CS rating errors more easily. Using the error matrix $J_H$ (Equation (10)) derived from data results in larger biases for both deterioration models (data not shown).

Overall age of the sewer system

The second example illustrates the influence of the amount of data available in each CS. Younger and smaller sewer systems have less data for worse CS and therefore increase the uncertainty of the parameter estimation. The results (Figure 4) show this behavior quite clearly for the last CS change (from CS4 to the worst CS5). The results also show that the two tested models have a different sensitivity for small data availabilities. The uncertainty of the cohort model decreases clearly as the fraction of pipes in CS 5 increases or if the analyzed data set is large. This result implies for instance, that survival functions estimated with the cohort Model with data of a sewer system built in the 1970s will have considerable uncertainties unless a very large data set is available. The Markov model is less sensitive on the age of the system because all parameters are estimated jointly, whereas the survival functions of the cohort model are estimated independently.

CONCLUSIONS

The intention of these examples was to illustrate how the data generator NetCoS can be applied to quantify uncertainties or detect biases of deterioration models. Both examined deterioration models showed a quite different behavior. While the CS classification error had a more drastic effect on the Markov model, the cohort model showed a strong sensitivity for younger or smaller networks. However, these are not the only aspects that should be considered for model selection. For example, further investigations demonstrated that the cohort model shows a very strong bias if replaced pipes are deleted from the database and only active sewer lines are considered, while the Markov model is fairly robust against this kind of data deficit (Scheidegger et al. 2011). This highlights the importance of model
selection for a particular application and the usefulness of testing quantitatively the behavior of deterioration models.

For our examples we generated data without consideration of any explanatory variables. Random processes and continuous deterioration processes can be described with time depended survival curves as long the conditions are stationary over time. The incorporation of explanatory variables would be an interesting extension of NetCoS and is discussed in Scheidegger et al. (2011).

It is important to emphasize that the presented approach is not capable of proving that a model performs correctly on a real application. It can only help to detect and quantify weaknesses of the deterioration model in question. However, this can help to eliminate less suitable models for the available data set – e.g. due to missing historic data. Despite the fact that the measured performance of a deterioration model depends on the – somewhat subjectively chosen – scenarios used for the data generation, relative comparisons for model selection are possible. If a model does not perform well on data generated by NetCoS, it is likely that it will perform even worse on real data.

We believe NetCoS is a helpful tool to assess and select deterioration models. Detected weaknesses and limitations of deterioration models may guide further development.

ACKNOWLEDGEMENTS

We thank Thomas Hug for his valuable comments and Karsten Müller for providing data about the classification error.

REFERENCES

Ana, E. V. & Bauwens, W. 2010 Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods. Urban Water 7 (1), 47.


Maurer, M. 2009 Specific net present value: an improved method for assessing modularisation costs in water services with growing demand. Water Research 45 (8), 2121–2130.

Müller, K. 2006 Strategien zur Zustandserfassung von Kanalisationen (Strategies to capture the condition state of sewers). Aachner Schriften zur Stadtentwässerung 7.


First received 23 February 2012; accepted in revised form 29 June 2012