Extreme-Value Analysis of Corrosion Data

Marcin Glegola

July 16, 2007

Supervisors:
Prof. Dr. Ir. Jan M. van Noortwijk
MSc Ir. Sebastian Kuniewski
Dr. Marco Giannitrapani

TUDelft

Outline

Motivation

Objective of the Thesis

Methods used

Data declustering

Examples of application

Framework for modelling extremes of corrosion

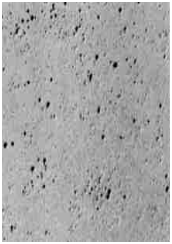
Conclusions and recommendations

Motivation

- in the oil industry, hundreds of kilometres of pipes and other equipment can be affected by corrosion
- it is extreme defect depth/wall loss that influences the system reliability ⇒ extreme-value methods are sensible for application
- only part of the system can be subjected to inspection ⇒ results extrapolation is needed

Motivation

- defects/wall losses caused by corrosion are likely to be locally dependent $\Rightarrow$ independence assumption questionable

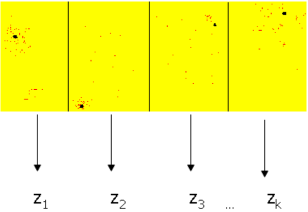


Objective of the Thesis

- present statistical methods to model extremes of corrosion data, taking into account local defect dependence
- spatial extrapolation of the results
- examples of application
- framework/guideline for modelling extreme-values of corrosion

Methods used

The Generalised Extreme-Value (GEV) distribution (block maxima data)

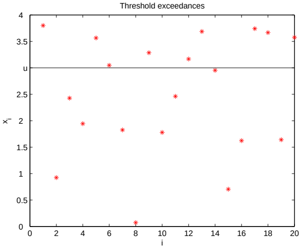


$$G(z) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, & \xi = 0, \end{cases}$$

Methods used

The Generalised-Pareto (GP) distribution (excess over threshold data)

$$Y_i = X_i - u, \text{ for } X > u, \ i = 1, \ldots, n_u$$

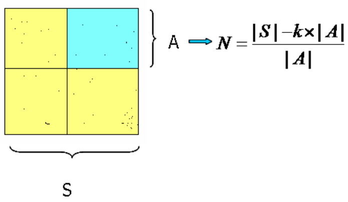


$$H(y) = \begin{cases} 1 - \left[1 + \frac{\xi y}{\bar{\sigma}}\right]_+^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\bar{\sigma}}\right), & \xi = 0 \end{cases}$$

Methods used (extrapolation-GEV)

- return-level method

$$G(z_p) = 1 - p \Leftrightarrow Pr\{M > z_p\} = p = \frac{1}{N}$$

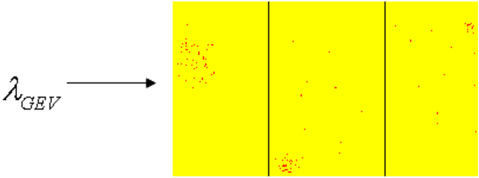


- implied distribution of the maximum corresponding to the not inspected area

$$Pr\{X_N \leq z\} = G_N(z) = G(z)^N$$

Methods used (extrapolation-GP)

based on Poisson frequency of threshold exceedances (Poisson-GP model)



- return-level method

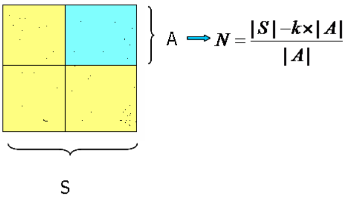
$$H(y_p) = 1 - p \Leftrightarrow Pr\{Y > y_p\} = p = \frac{1}{N_E}$$

$N_E = \lambda_{GEV} \times (|S| - k \times |A|)$ - expected number of exceedances on the not inspected area

Methods used (extrapolation-GP)

- implied distribution of the maximum corresponding to the not inspected area

$$Pr\{X_N \leq x\} = \exp\left\{-\lambda_{GEV}\left(1+\xi\frac{x-u}{\bar{\sigma}}\right)_+^{-1/\xi}\right\}^N$$



Methods used-summary

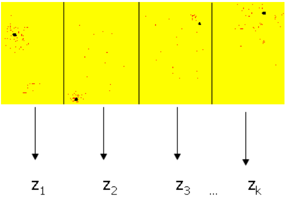
- two methods for statistical inference about extreme-values of corrosion
  - the GEV distribution (block maxima)
  - the GP distribution (excess over threshold data)
- two methods for spatial results extrapolation
  - return-level
  - distribution of the maximum corresponding to the not inspected area
- the GEV and GP distributions are closely related and theoretically, should give the same results

Modelling extremes of dependent data

- for the stationary data characterised by the limited extend of long-range dependence at extreme levels, the extreme-value methods can be still applied
- in corrosion application it is reasonable to assume that pit depths are locally dependent $\Rightarrow$ extreme-value methods are applicable

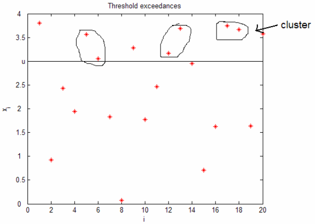
Modelling extremes of dependent data with the **GEV** distribution

- assuming local data dependence, block maxima of stationary data (for sufficiently large block sizes) can be considered as approximately independent
- the GEV distribution is used in its standard form



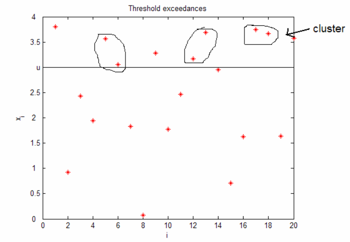
Modelling extremes of dependent data with the **GP** distribution

- neighbouring exceedances may be dependent, therefore the change of practise is needed
- one of the most widely adopted method is **data declustering** - filtering out dependent observations such that remaining exceedances can be considered as approximately independent



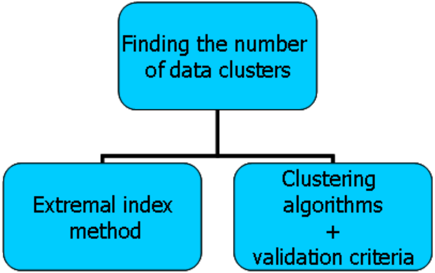
Modelling extremes of dependent data with the **GP** distribution-approach

- define clusters of exceedances
- identify maximum excess within each cluster
- assuming that cluster maxima are independent fit the GP distribution



Data declustering

Estimation of the number of data clusters



Data declustering

Extremal index method:

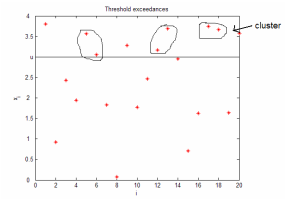
- the extent of short-range dependence of extreme events is captured by the parameter θ, called **extremal index**

$$\theta = \frac{1}{\text{limiting mean cluster size}}$$

- extremal index measures the degree of clustering of the process at extreme levels

Data declustering

Extremal index method



$$\theta = \frac{1}{\text{limiting mean cluster size}} \Rightarrow N_c = \theta \times N_e$$

where N_c - number of clusters, N_e - number of exceedances above threshold u

θ is estimated by the intervals estimator

Data declustering

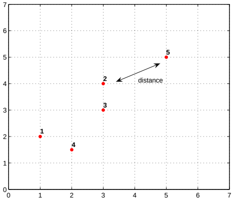
Finding number of clusters using clustering algorithm

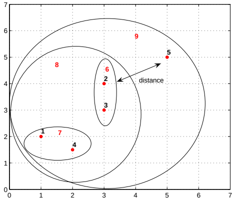
- define a validity criteria for the number N_c of found clusters
- run the clustering algorithm for a range of N_c
- as proper number of clusters choose the one for which the validity criteria are optimised

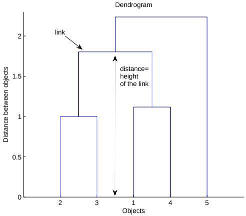
Agglomerative hierarchical clustering algorithm

- starts with single points as clusters
- at each step the two closest clusters are merged
- stops when only one cluster remains

Agglomerative hierarchical clustering algorithm







Data declustering

Validation criteria, that we used, aim at identifying clusters that are compact and well isolated:

- silhouette plot - maximum value indicates optimum
- Davies-Bouldin index - minimum indicates optimum

Data declustering-summary

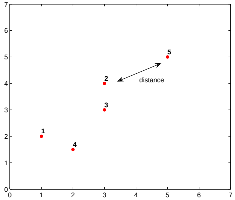
- two methods to estimate the number of data clusters
  - the extremal index method
  - clustering algorithm + validation criteria (Davies-Bouldin index, silhouette plot)
- prior to data declustering perform clustering tendency test

Simulated corroded surface

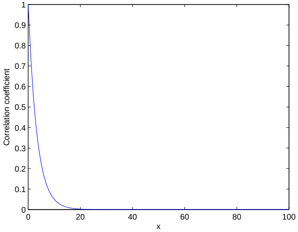
- application of the gamma-process model
- dependence in terms of the product moment correlation

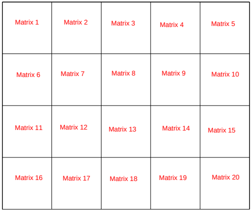
$$\rho(X_k, X_l) = \exp\left\{ -d \left(\sum_{i=1}^{2} |dist_i|^p \right)^{q/p} \right\}$$

$X_k = (x_k, y_k),\ X_l = (x_l, y_l),$
$dist_1 = |x_k - x_l|,\ dist_2 = |y_k - y_l|,$
$d = 0.3,\ p = 2,\ q = 1$

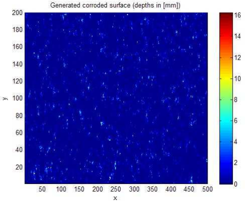


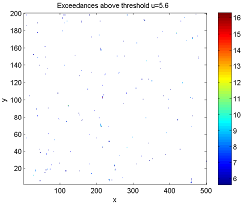
Simulated corroded surface

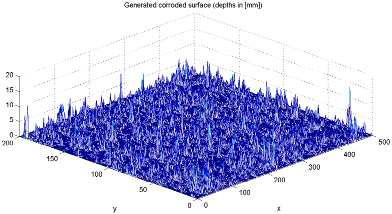




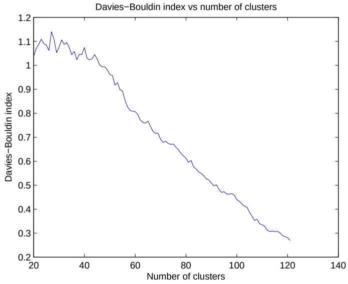
Simulated corroded surface

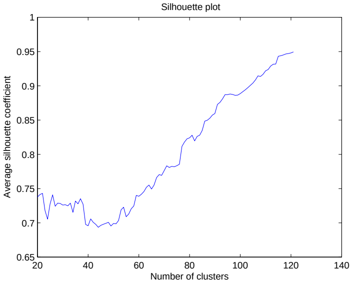






Simulated corroded surface - clustering algorithm





Simulated corroded surface - results

$\hat{\theta}$	n_c	$n_{c_{alg}}$
0.544	107	121

Table: The estimate of extremal index and determined number of clusters

$Number\ of\ clusters$	$AD^2_{up}\ p-v.$	$KS\ p-v.$
107	0.376	0.204
121	0.549	0.493

Table: Goodness-of-fit test results for different number of clusters

$\hat{\xi}$	$\hat{\hat{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
0.065	1.341	0.647	0.488

Table: GP fit to excess of dependent data

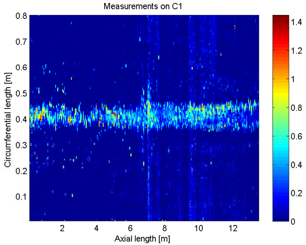
$\hat{\xi}$	$\hat{\sigma}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.0137	1.6839	0.555	0.493

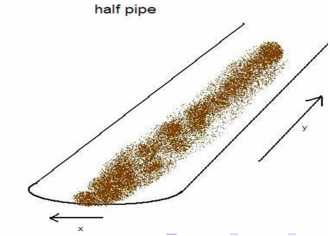
Table: GP fit to excess of declustered data

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.007	1.627	5.637	0.621	0.899

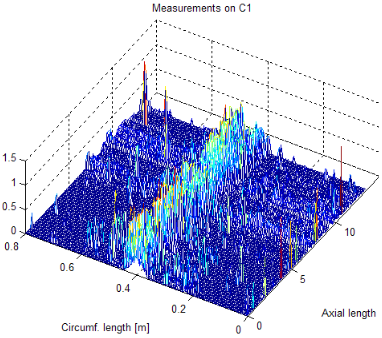
Table: GEV fit to block maxima data

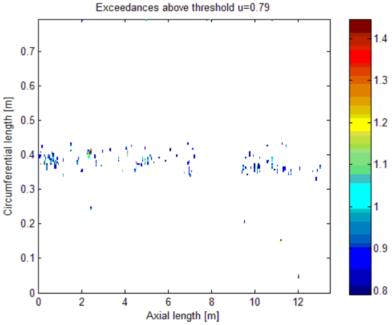
Real data example



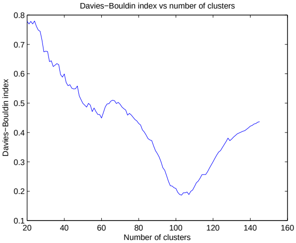


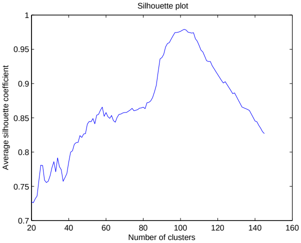
Real data example





Real data example





$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.082	0.182	0.757	0.713	0.986

Table: GEV fit to block maxima data

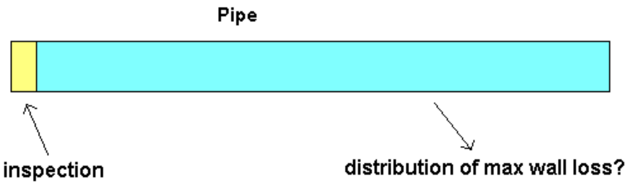
$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.008	0.133	0.616	0.074

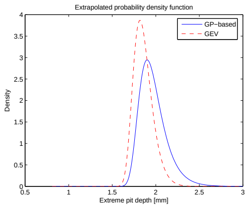
Table: GP fit to excess of dependent data

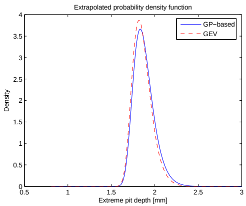
$\hat{\xi}$	$\hat{\tilde{\sigma}}$	AD^2_{up} $p-v.$	KS $p-v.$
-0.069	0.172	0.717	0.654

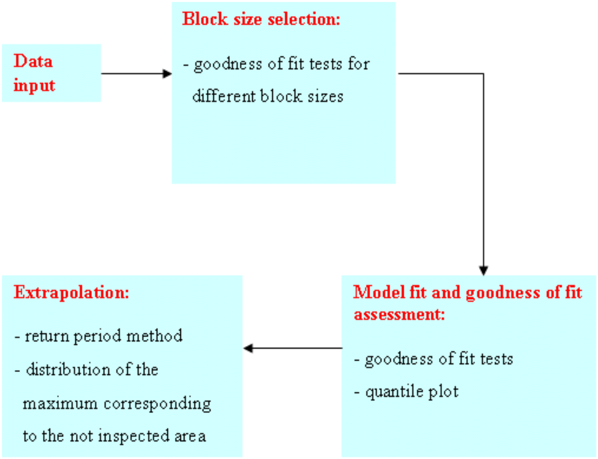
Table: GP fit to excess of declustered data

TUDelft

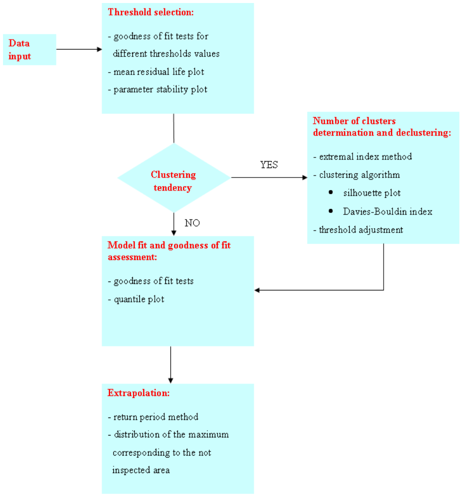

Pipe
inspection
distribution of max wall loss?


Extrapolated probability density function
GP−based
GEV
Density
Extreme pit depth [mm]


Extrapolated probability density function
GP−based
GEV
Density
Extreme pit depth [mm]



Framework for modelling extremes of corrosion with the GEV distribution



Framework for modelling extremes of corrosion with the GP distribution

Conclusions and recommendations

- the two applied distributions are closely related and lead to the consistent inference about extreme-values of corrosion
- data declustering improves the results given by the GP distribution
- the performance of other clustering algorithms could be checked
- in order to take into account corrosion nonstationarity due to space-varying environmental conditions, covariate-dependent extreme-value models with trends could be considered

THANK YOU FOR ATTENTION

Questions???