South Staffs Water

Network Strategy

Review of Mains Deterioration Model
v2.0
June 2013

Technical Report

Contents

1. Introduction 1
2. Current Model Evaluation 2
3. Future Requirements 3
3.1 Why SSW needs a deterioration model 3
3.2 Specific requirements for new model 4
3.3 Intended phases of implementation 5
3.3.1 Phase 1 5
3.3.2 Phase 2 5
3.3.3 Phase 3 5
4. Preliminary Data Analysis 6
4.1 Type of Pipe 6
4.2 Material of Pipe 7
4.3 Data Availability 9
4.3.1 Failure Data (the 'bursts envelope') 9
4.3.2 Data Quality 11
4.3.3 Predictors 11
4.3.3.1 Included predictors 11
4.3.3.2 Excluded predictors 11
4.3.3.3 Potential future inclusions 12
4.4 Conclusions – Preliminary analysis 12
5. Developing Analysis 13
5.1 Age Trends 13
5.2 The 'rehab effect' 15
5.3 'Dual Curves' approach 16
5.4 'State Transition' approach 18
5.5 Final model development 18
5.5.1 SPSS investigations 18
5.5.2 Scattermatrix of scale variables (investigation of cross correlation) 20
5.5.3 Grouping and testing 21
5.6 Conclusions - Developing analysis 21
6 Final Analysis 22
6.1 Regression analysis by material 22
6.1.1 HPPE 22
6.1.2 AC 27
6.1.3 DI 33
6.1.4 MDPE 35
6.1.5 Steel 41
6.1.6 'Other' materials 46

6.1.7 mPVC 47
6.1.8 PVC 49
6.1.9 CI 60
6.1.9.1 CI where year installed less than 1952: 78
6.1.9.2 CI where year installed greater than or equal to 1952: 85
6.2 Final models for all cohorts 90
7. Application of model curves 91
7.1 Flow Diagram 91
7.2 Scheme selection logic diagram 92
7.3 Stochastic effects 93
8. Validating outputs 96
8.1 Sense check – scale and rates 96
8.2 Backcast analysis 97
8.3 Sensitivity checks 98
8.3.1 Minimum scheme size 99
8.3.2 Maximum scheme size 99
8.3.3 Minimum allowable failure rate 100
8.3.4 Summary 100
8.4 Benchmarking against other companies 101
8.5 Expert review and critical appraisal 101
8.5.1 Mott MacDonald – Statistics Review 101
8.5.2 Seams Analytics – Approach Review 102
8.5.3 Birmingham University 103
8.5.4 Summary of comments 103
9. Communicating results and identifying improvements 104
9.1 Presentation of results 104
9.2 Future improvements 104
9.3 Action plan 104

Appendix 1: Critical review – Mott MacDonald
Appendix 2: Critical review – Seams Analytics

1. Introduction

South Staffs Water uses a deterioration model for forecasting the long term performance levels of pipes. The current model was developed by WRc around 7-8 years ago, and uses SSW's historic failure records to assess trends in bursts for different pipe cohorts. The trends for these cohorts are then used to forecast future failure rates, and to simulate the effect of applying differing lengths of asset renewal each year; the accuracy of the model is therefore critically important for SSW's future investment planning.

This document assesses the suitability of the current model, and describes the process undertaken to build the new deterioration model. This is intended to be a descriptive and technical document describing the entire analysis process, but is accompanied by a shorter summary document for wider distribution. SSW has followed the good practice guide published by the Water Statistics User Group (WSUG) to ensure that the approach to any statistical modelling is as robust as possible. This guide is available here:
http://www.statsusernet.org.uk/StatsUserNet/Communities/Resources/ViewDocument/?DocumentKey=58b5544a-01bd-4031-83da-41f091713e58

SSW recognises that developing skills in-house for the creation and maintenance of complex modelling is crucial in order to extract the best out of such techniques for effective future asset management; with this aim, the intention was for SSW to deliver this project using in-house effort and skills wherever possible. This has been supplemented with statistical guidance and assistance from Adam Grove at Mott MacDonald, both at the start of the project and for critical appraisal towards the end of the project. Another independent critical review has also been requested from Joe Roebuck at Seams Ltd Analytics. These reviews are attached as appendices for reference.

2. Current Model Evaluation

The current deterioration model divides SSW's 6,000km of network assets into approximately 260 individual cohorts – groups of assets that share certain characteristics. These cohorts are determined using material, diameter, era (date installed), soil corrositivity and soil fracture potential.

Current failure rates of each cohort are assessed and then linear deterioration rates are applied to each cohort to model future levels of performance. It is also possible to model the impact of removing different lengths of the worst performing assets each year, simulating the effect that proposed lengths of renewal intervention might have on the observed bursts. This approach is utilised to determine the appropriate scale of the renewal programme required to effectively manage the number of burst mains that occur on the SSW network. It is then assumed that by managing bursts, SSW are, by proxy, managing the service experienced by each customer.

The model has proven very useful for making strategic decisions, but it is not an effective operational tool for identifying assets for renewal. This is undertaken using cluster analysis and condition inference based on recent burst patterns. Schemes are then selected to tackle recently observed 'problem areas' – not to tackle modelled 'problem areas'.

A critical review of the existing model has been undertaken. The key points of this review are listed below:

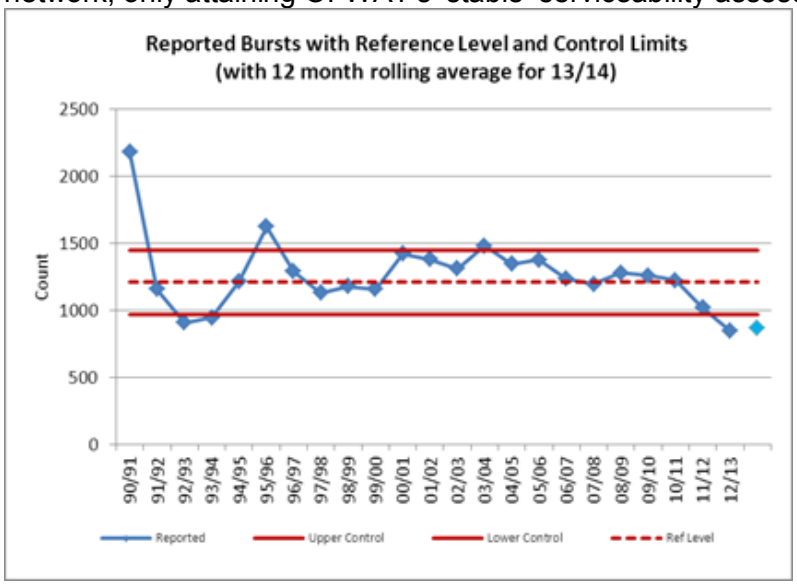
- The model is non-selective in it's application, assumes removing 50% of a cohort length will remove 50% of bursts;
- Previous analysis has suggested linear deterioration of asset performance, which seems unlikely;
- The excel based application of the model doesn't utilise any link between cohorts and specific assets, so it is impossible to use the deterioration model to highlight specific assets for renewal – this was previously criticised by both the regulator OFWAT and the Company's reporters at PR09;
- The model uses just four individual 'bins' of data, so the high-level grouping ensures a relatively good statistical fit;
- Previous analysis didn't over-complicate regression, because it was recognised that there was an insufficient period of failure data available. The regression was therefore based on general trends per failure year available, and not based on age of assets;
- Transparent workings and application – no black box;
- The model doesn't reflect increased chance of failure if pipe has already failed;
- The model has proven easy to outperform. This is because large cohorts are used for prediction. It's then assumed that identical benefit can be realised by replacing any component pipes within each cohort. In reality, burst rate disparity within groups is quite large, and by selecting the worst assets for replacement, a disproportionate (larger than predicted) number of bursts can be prevented;
- Limited factors included within regression analysis, based on material and SMD & min winter temp;
- Only 4 deterioration rates were applied to over 260 cohorts.

3. Future Requirements

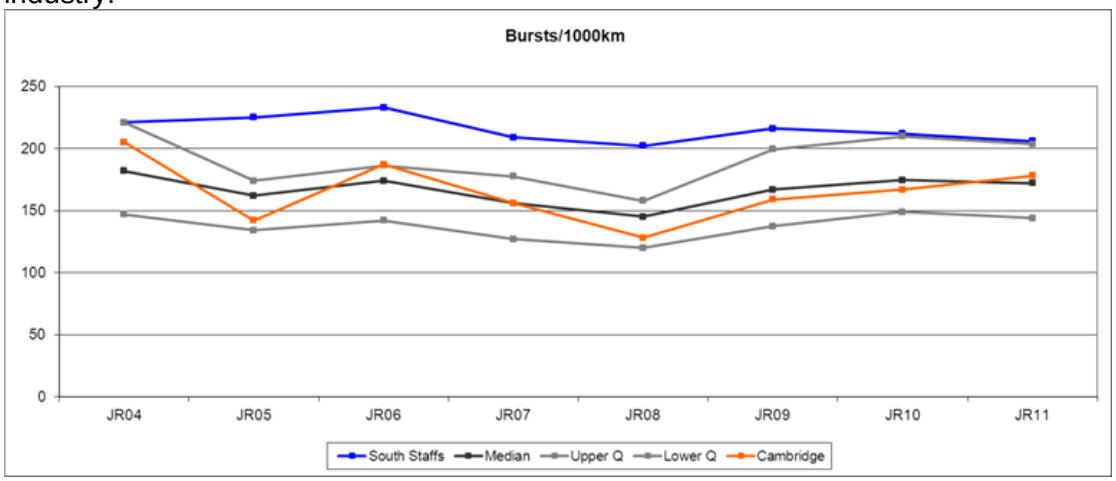
3.1 Why SSW needs a deterioration model

It is important to review exactly why South Staffs needs a deterioration model before discussing the specific requirements for the any new model.

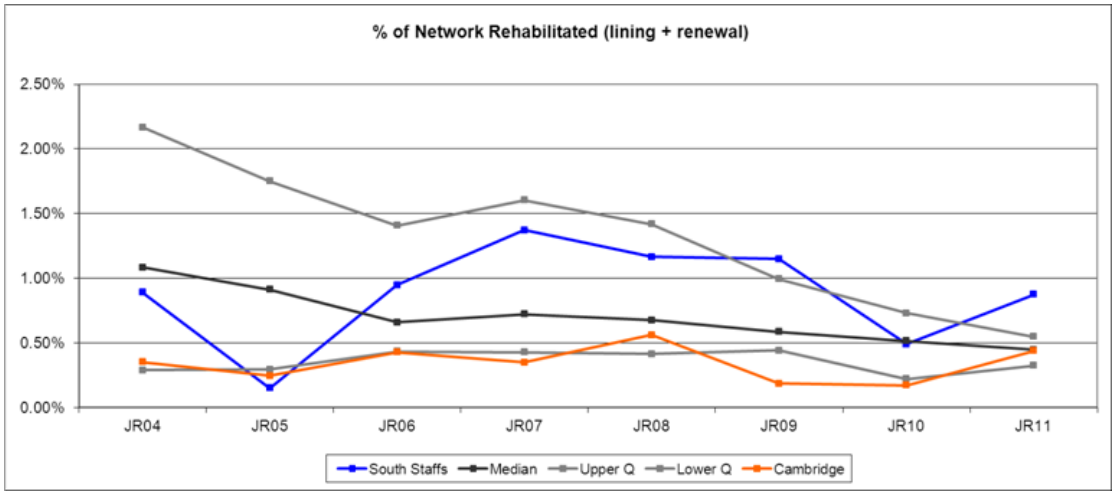
Historically, South Staffs have had a high number of burst mains relative to the size of it's network, only attaining OFWAT's 'stable' serviceability assessment in 2008.



The high level of bursts has begun to reduce during the last few years, but South Staffs still rank highly overall, with burst numbers consistently exceeding the upper quartile for the industry:



Current rehab levels are also in the upper quartile in an effort to tackle the problem:



It's therefore critically important for South Staffs to accurately forecast the future behaviour of the network in order to invest in suitable levels of asset renewal:

- How quickly will the network deteriorate with no proactive intervention?
- Is PVC an emerging problem, will it continue to deteriorate exponentially?
- How do we balance today's investment to control failure rates in 20 years?
- How long can we really expect our assets to last, is it reasonable to assume 150+ year lives for pipes?

3.2 Specific requirements for new model

The new deterioration model needs to deliver similar capability to the current model, for example, using the company's asset data and failure data to forecast failure trends up to 25 years from today. In addition to this functionality, the following improvements are required:

1. The model must **be able to simulate more detailed scenarios**, such as:
    a. Target end rate (varying rates of rehab to hit end rate at most efficient level)
    b. Fixed rates (specifying flat rate per year, or specific rates for each year, to forecast end rate)
    c. Limited budget (varying rates of rehab between min and max limits to hit end rate at most efficient level)
    d. Max failure rate (specifying max failure events and modelling for minimum levels of rehab to avoid breaching)
2. The model must be able to **execute numerous randomly seeded 'runs'**, evaluating 100's or 1000's of possible combinations to ascertain the most efficient strategy.
3. There must be a clear **link between the top-down deterioration model outputs and the bottom-up renewal selection**. The deterioration model will therefore need to reference and identify specific pipes and flag them for renewal. These flags can then be used to select which pipes are renewed, and when they are programmed.
4. The model must be **easy to refresh**. Failure and asset data is constantly changing and improving. It should be expected that the deterioration curves will need to be refreshed once per year, with the selection database being updated every time a batch of rehabilitation feasibility is prepared. There may also be a requirement to incorporate the pipe and failure data from Cambridge Water.
5. The model should **reflect known patterns** – i.e. pipes at higher pressure may fail more frequently, pipes that have burst once are more likely to fail again because of disruption caused by repair of first failure.
6. The model must be able to incorporate **stochastic variables** such as weather, i.e. applying a known distribution for weather events to the forecast burst rates to simulate high and low years and demonstrate sensitivity to **unusual weather trends and climate change**.

7. The application of the model will need to allow future inclusion of **'risk to customer' scores**, i.e. target renewals where there is a provable problem with serviceability, not just where there are bursts.

3.3 Intended phases of implementation

It should be noted that the list of requirements listed in section 3.2 is a significant change from the current approach. This could take a long time to design, develop and implement and is therefore likely to be approached in phases. The initial focus will be to review and revise the current deterioration rates and curves and then to develop a method of applying these curves whereby specific assets can be recommended for intervention (effectively joining our top-down approach to the bottom-up scheme selection).

Future phases are likely to address optimisation (selecting most efficient investment profiles), sensitivity (applying variance for weather, pressure changes etc), customer risk (determining intervention need by reviewing impact, not just burst frequencies) and more sophisticated uncertainty assessment.

3.3.1 Phase 1

- Build deterioration curves using currently available 'pipe level' data (such as age, material , diameter, soils etc)
- Build an application to apply these models iteratively to the existing asset data for forecasting future performance of the network
- Attempt to implement a module within the modelling application to automatically select schemes

3.3.2 Phase 2

- Review the inclusion of other predictors/consequences from hydraulic and DG2 models – specifically: modelled roughness coefficients and individual pipe criticality (isolation risks)
- Potential inclusion of alternative modelling techniques – GLM etc, allowing greater ability to assess model error
- Undertake more advanced analysis for weather effects and apply potential future impacts of climate change
- In addition to pipe-level prediction of burst numbers, implement pipe level consequence assessment for effects such as: interruptions, discolouration and traffic impact

3.3.3 Phase 3

- Full pipe level failure modelling, consequence modelling and intervention option modelling (including cost differences)
- Scenario results logging and presentation via qlikview (or equivalent) for improved usability
- Easier data refresh and potential use of SPSS modeller for advanced/semi-automated refresh of the deterioration curves

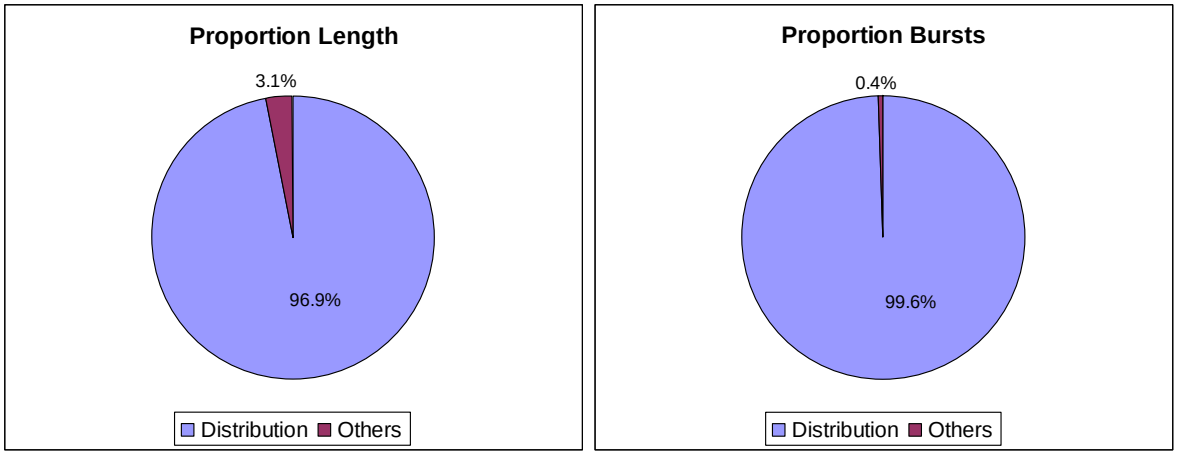
4. Preliminary Data Analysis

This section provides a high level summary of initial / preliminary analysis which has informed the approach to the detailed analysis:

- Initial graphs of asset data/burst data
- Quality and quantity of data
- Means and std devs etc, scatter plots of materials
- Look at where bulk of failures or bulk of assets are, where should effort be focussed
- Explanation of bursts envelope
- Investigate percentage of data in each cohort represented within envelope
- Look at percentages of assets in each material and roughly forecast how this is likely to change over 25 years
- Look at where bursts are occurring, which are the materials that require the most focus
- Analyse availability and reliability of data

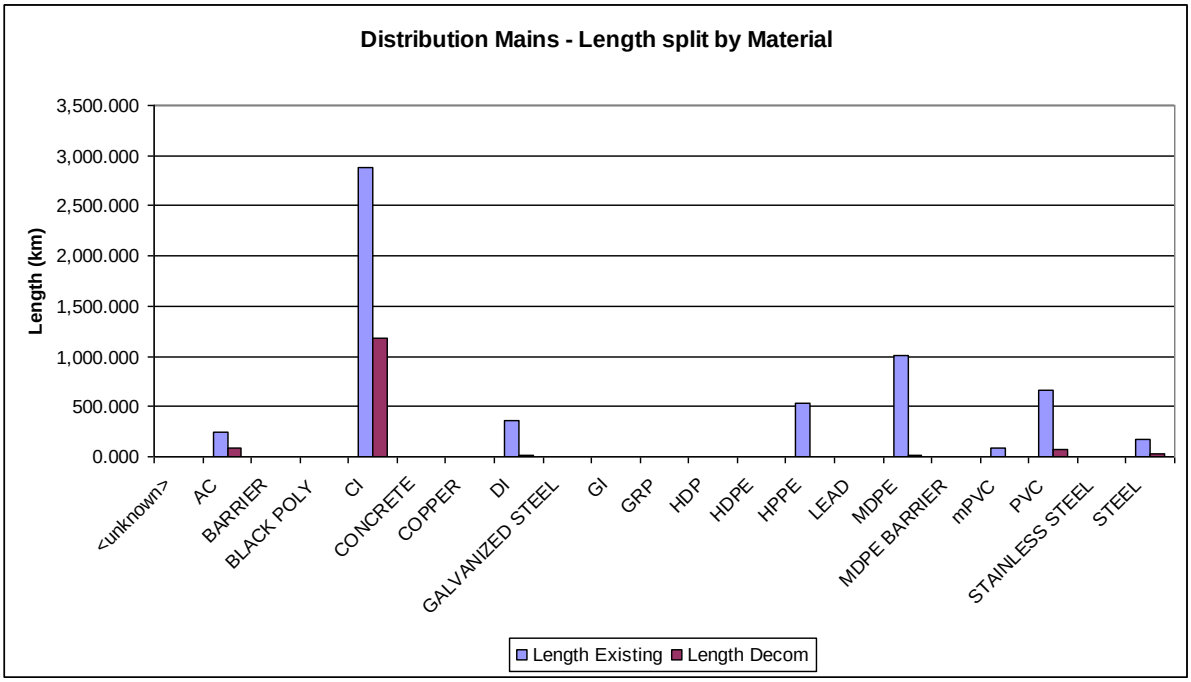
4.1 Type of Pipe

The following charts show that the network comprises almost 97% 'distribution' type assets, and that these assets account for an average of 99.6% of the annual burst events. It is therefore considered unnecessary to build a detailed model for the remaining types; these can be explained using simple mean burst rates. The current burst rate on 'other' types is just 5.5 bursts per year, so any attempt at complex modelling of these events would be significantly hindered by lack of data. There is also further justification for analysing distribution mains in isolation because other types of mains are more likely to have different pumping conditions (i.e. raw/blending mains may have more pressure spikes or fire mains may have long periods of standing water); these differing conditions are likely to provoke a different failure trend when compared against the consistently used distribution mains.



The distribution mains contain several materials, all of which behave and deteriorate differently. It's expected that most of these materials will require an individual model. The following charts explore the distribution of the asset and failure data between these materials to ascertain which materials justify further analysis.

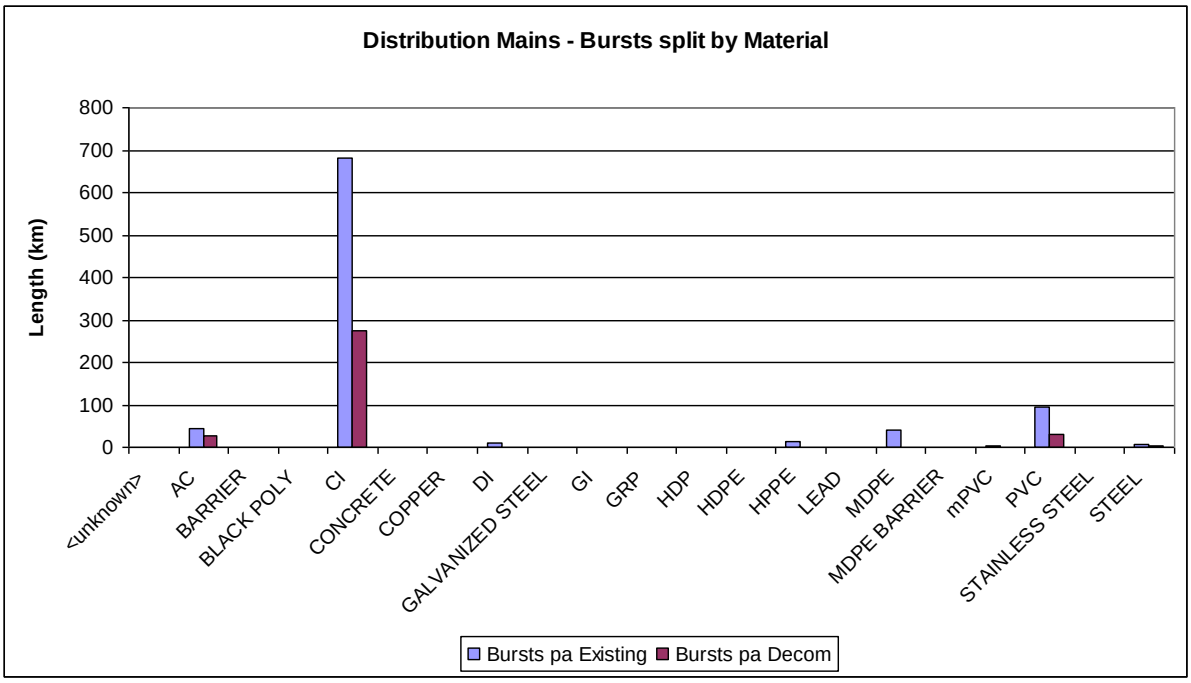
4.2 Material of Pipe



As can be seen in the above chart, there are just 8 of the 21 material types that seem to have sufficient length of assets to enable robust deterioration modelling. These are:

- Asbestos Cement/AC (4.43%)
- Cast Iron/CI (55.15%)
- Ductile Iron/DI (5.06%)
- High Performance Polyethylene/HPPE (7.24%)
- Medium Density Polyethylene/MDPE (13.92%)
- Modified Polyvinylchloride/mPVC (1.23%)
- Standard Polyvinylchloride/PVC (10.0%)
- Steel (2.81%)

This leaves a total of 0.16% of the total asset stock in the ‘other’ category.

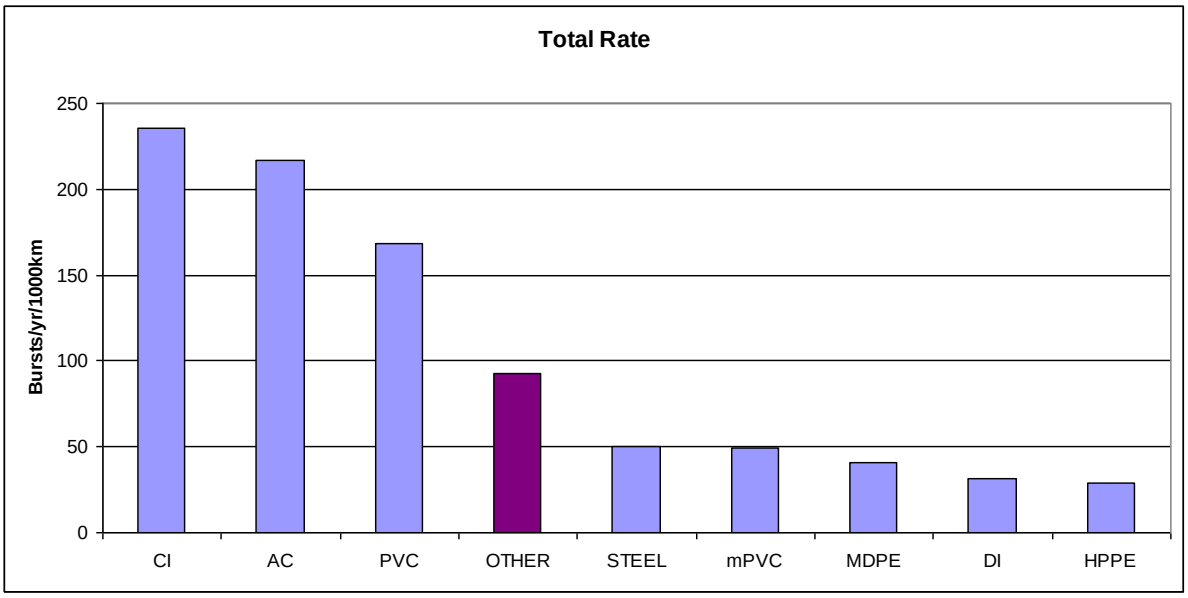


Similar to the previous length chart, this graph displays the average annual bursts observed against each material type. Again there are really only 4 (possibly 7) dominant materials out of the 21 present in the dataset. These are the same materials as the length categories, with the exception of mPVC which demonstrates such a low burst rate that it may prove difficult to model, with few observed events (average 4.5 per year) and a relatively small exposure of length. The materials considered appropriate for modelling are:

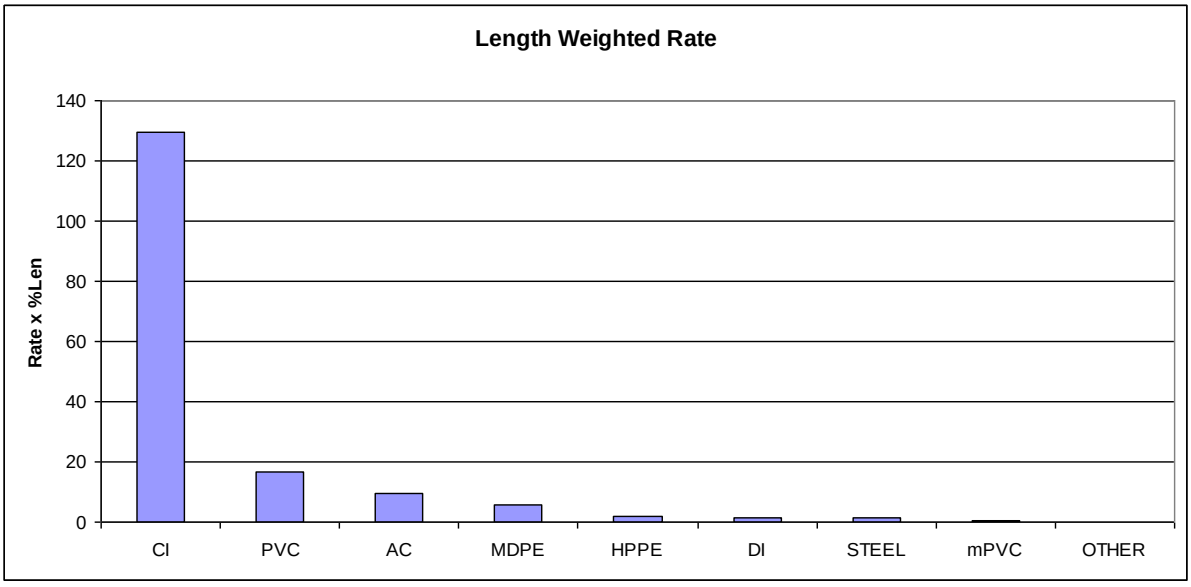
- Asbestos Cement/AC (5.73%)
- Cast Iron/CI (77.34%)
- Ductile Iron/DI (0.95%)
- High Performance Polyethylene/HPPE (1.25%)
- Medium Density Polyethylene/MDPE (3.4%)
- Standard Polyvinylchloride/PVC (10.05%)
- Steel (0.84%)

This would leave just 0.45% of the annual bursts in the ‘others category’ (est. average 5.6 bursts per year).

The following chart displays the rate of failure on each of the identified material groups:



This shows the expected high failure rates on Cast Iron, Asbestos Cement and PVC mains, suggesting that the deterioration analysis should focus more on these materials. There is a slight concern that the 'other' group has the fourth highest rate, although this should be considered alongside the previous discussion – there is very little asset stock in this group. The following graph displays the same data, but with the rate in each material group weighted by total asset length:



Again this suggests the same three materials make up the top 3, but importantly, shows that the impact of the 'other' group is negligible.

4.3 Data Availability

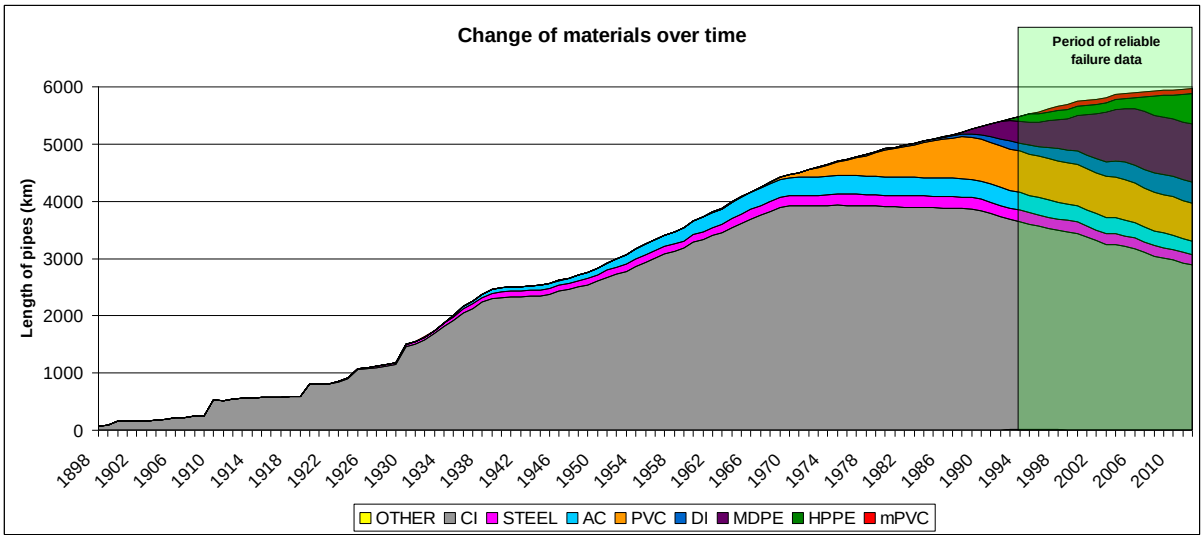
4.3.1 Failure Data (the 'bursts envelope')

SSW are using the past 16 years of failure data for building the deterioration model; although more data is available, this more recent period is more reliably validated and each failure record is tagged to individual assets. Each asset that has existed during the 16 year 'bursts

envelope' is then being logged (at each year) in the analysis database to enable the most effective analysis possible with the limited amount of data; for example:

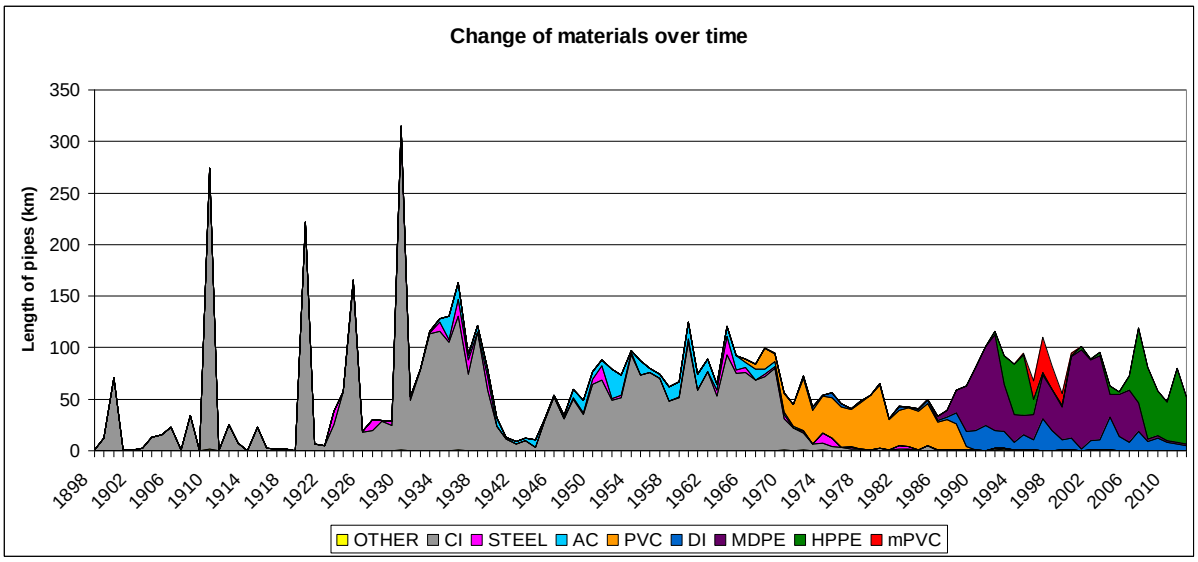
- A pipe may be recorded in 1996 at 40 years of age, with 0 bursts
- The pipe is then recorded again in 1997, now at 41 years, with 2 bursts
- The pipe is then recorded again in 1998, 42 years old, 0 bursts
- The pipe may then cease to be recorded again if it is decommissioned in 1999 at 43 years old

So each pipe could appear within the dataset up to 16 times, with varying numbers of bursts. The sum of bursts in each year and the sum of applicable length (exposure) can then be assessed to calculate the mean burst rates.



The above chart shows a timeline for SSW's assets over the last 120 years. The green rectangle on the right hand side of the chart shows that despite having almost 20,000 failure records, this only covers a short period of the timeline. It is also evident that the rectangle covers different parts of each material's lifespan, with cast iron assets already in decline before failures were ever recorded, whilst HPPE is almost fully covered, offering an opportunity to measure early life failures. This creates difficulties for modelling, especially when asset renewal has been taking place since the early 1970s so that the 16 year's worth of failure data has always been artificially moderated by ongoing renewal intervention; i.e. we have no way of knowing what today's failure rates would have been like if no intervention had ever taken place. It is therefore likely that any attempt at deterioration modelling may need to include a coefficient to measure the effect of the renewal efforts within each asset group.

The following chart is a variation of the chart above, but only shows the pipe lengths installed within each year (not accumulating the length over time). This shows, as expected, that particular materials dominate specific eras. It also shows that there are some anomalies in the GIS asset data for older pipes, whereby year installed may have been rounded up or down to the nearest 5 years, resulting in the spikes on the left hand side of the chart. These spikes are likely to cause problems during the analysis, because many of the assets will all reach the same age at the same time. This is not too problematic when modelling that specific age, because any modelling is likely to be length weighted; but it may affect the neighbouring ages where the asset length (and therefore failures) will be much lower than expected. In this case, a few additional bursts could dramatically increase the failure rate for that age, resulting in an outlier in the data (or vice versa with absence of bursts). This is difficult to repair ahead of modelling, but will need to be noted when analysing outputs.



4.3.2 Data Quality

SSW considers it's asset data to be of very good quality, with very few data gaps or anomalies. The company's GIS (geographical information system) has been in place since around 1990 and a comprehensive 3 year resurvey programme was undertaken during the implementation phase, this resurvey covered the entire network and delivered a reliable and trustworthy asset database.

The failure data is captured within the works management system. Each record is individually validated to ensure correct coding and then links to the appropriate asset ID are added. This gives a reliable link between the failure and the asset records; ongoing maintenance of these links is then automatically managed by the GIS, handling the frequent splitting and merging of asset IDs that typically occur with linear assets.

4.3.3 Predictors

Comprehensive analysis of the available data and experienced engineering judgement has informed the decisions on which predictor variables could potentially be included within the models.

4.3.3.1 Included predictors

The following predictors are currently being considered for inclusion within the deterioration analysis:

- Material;
- Diameter;
- Age;
- Date Installed and/or Era;
- Soil Corrosivity;
- Soil Fracture Potential

These have been chosen because they are well-populated and readily available at individual pipe level granularity.

4.3.3.2 Excluded predictors

The following predictors are considered incomplete or inaccurate and are not being considered:

- Internal protection/lining;
- External protection/coating

The attributes are not well populated due to a mix of manufacturer and in-situ linings applied to some pipe materials. Records of manufacturer linings could be estimated from the dates

installed, but the in-situ linings are more difficult to ascertain because of poor record keeping during extensive relining of mains during the 1970's and early 80's. It is therefore highly unlikely that these attributes would make reliable predictors within a model; it is also impossible to improve the data quality for these attributes; hence their early exclusion from the model building process.

4.3.3.3 Potential future inclusions

The following predictors are not being included at this stage, but are possible considerations for future development/improvements:

- Typical operating pressure;
- Modelled internal corrosion build-up/scale;

These attributes are readily available within our InfoWorks hydraulic models, but would take time to migrate into our pipe-level analysis datasets. From an engineering perspective, it is likely that they would prove to be significant for predicting pipe failure rates and future action plans will reflect this; but it is also recognised that pipes are long-life assets and the typical operating pressure now may be quite different to that experienced at earlier phases of each asset's life cycle (before the implementation of large scale district zoning and pressure management). However, the inclusion of these attributes within the initial modelling exercise is nevertheless prevented by time limits and impending business planning deadlines.

4.4 Conclusions – Preliminary analysis

The preliminary analysis has shown that:

- there are notable differences between the failure rates of different materials. This will be expanded upon in the next chapter;
- the asset data available is considered very accurate and already contains many of the predictor variables required for any predictive modelling;
- the failure data is effectively limited to 16 years (1996 – 2012). But it is also recognised that this dataset (fully linked to assets) is unusually large compared to other UK companies;
- there are other predictors, that make engineering sense, but cannot be included within this version of the analysis because of time constraints. These should be reconsidered in future, because they could make a valuable and significant contribution to the predictive models;
- there are some anomalous GIS asset ages, especially for earlier eras. These anomalies may affect regression outputs and will need to be considered when analysing for patterns;
- there are some materials where the number of bursts experienced (or the length of assets) is so low that reliable regression analysis is impossible. It is recognised up-front that the most appropriate way of accounting for failures on these materials will be to apply simple averages.

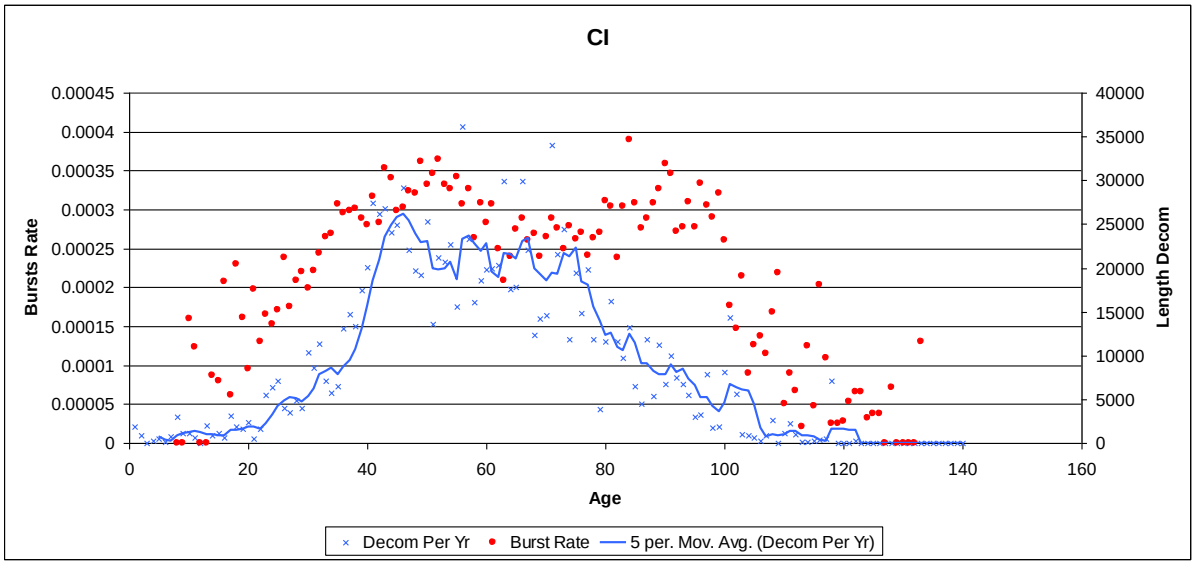
5. Developing Analysis

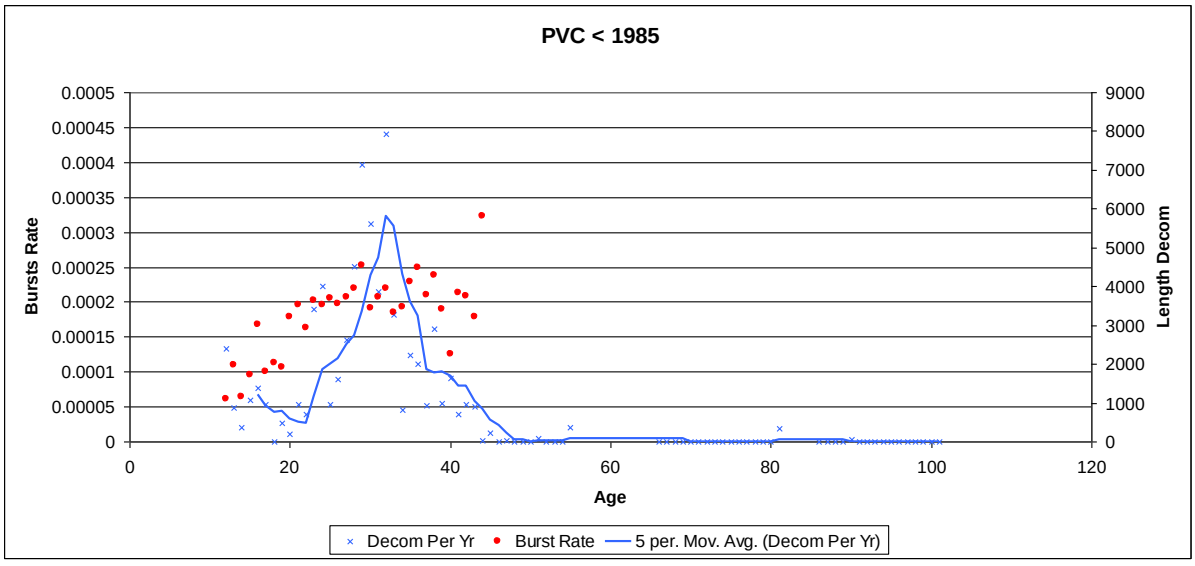
Many hours of analysis have contributed to the final model, and this section attempts to describe some of the work undertaken; describing the 'analysis journey' that has driven some of the decisions and direction for the final SPSS analysis detailed in the next section.

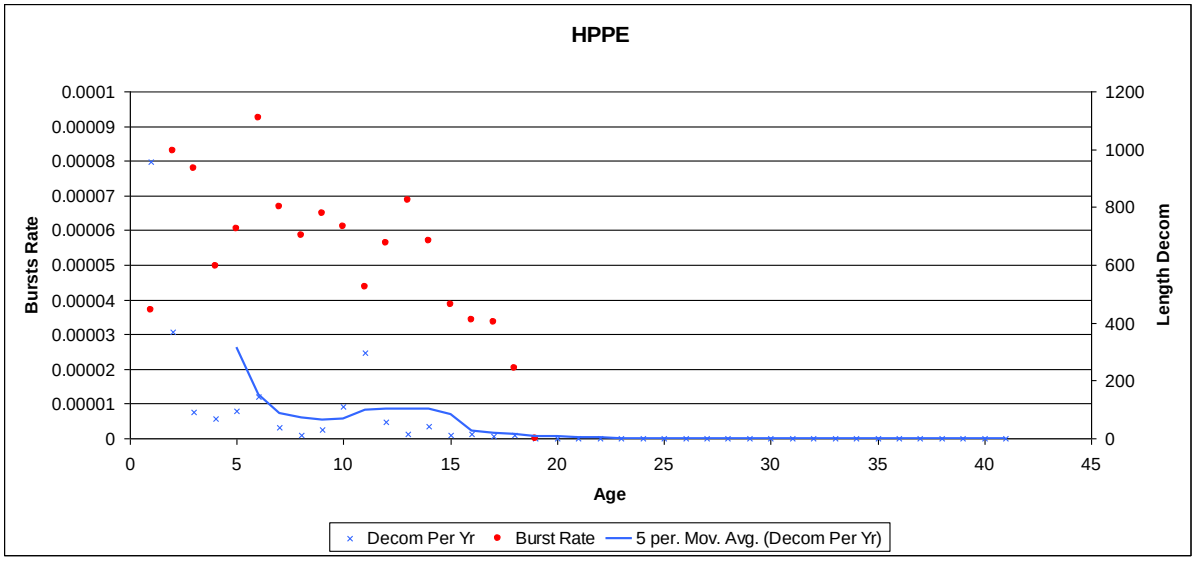
5.1 Age Trends

The intention was to construct a deterioration model at a pipe level, as set out in the original aims for the model review. This would rely on the ability to accurately model the average failure rates of each pipe at a particular point in it's life cycle. This differs from the previous deterioration model, which applied more generic deterioration rates to an entire cohort, based on observed failures for much larger groups of assets (of differing ages).

The following graphs show some examples of obvious age related relationships (red dots are mean failure rates and the blue line represents a rolling average of the renewal intervention applied at each age):







These patterns are confirmed by the following R output showing normal GLM for PVC bursts rate. Note that R has identified age as a highly significant predictor for burst rate: t = 5.4e-11

```
> fit_N<-glm(Rate  ~  Age ,  data = pvcAge, family = gaussian)
> fit_N

Call:  glm(formula = Rate ~ Age, family = gaussian, data = pvcAge)

Coefficients:
(Intercept)         Age
  3.454e-05    5.324e-06

Degrees of Freedom: 39 Total (i.e. Null);  38 Residual
Null Deviance:      2.215e-07
Residual Deviance: 7.043e-08    AIC: -686.8
> summary(fit_N)

Call:
glm(formula = Rate ~ Age, family = gaussian, data = pvcAge)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.082e-04  -2.349e-05   2.077e-06   2.285e-05   1.047e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.454e-05  1.759e-05   1.964   0.0569 .
Age         5.324e-06  5.897e-07   9.029  5.4e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.853533e-09)

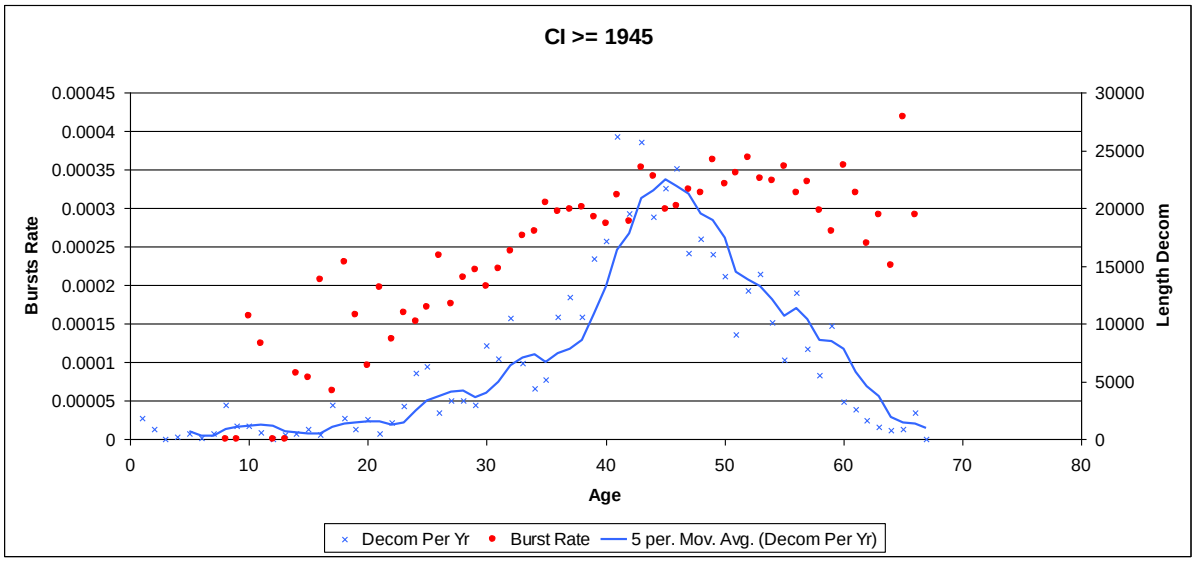
    Null deviance: 2.2153e-07  on 39  degrees of freedom
Residual deviance: 7.0434e-08  on 38  degrees of freedom
AIC: -686.78

Number of Fisher Scoring iterations: 2
```

However, there are some notable anomalies appearing on some of the graphs, such as the 'M' shaped trend for cast iron.

Age seems to be a significant predictor for most materials, but there are distinct eras within materials too. This is to be expected where manufacturing techniques and installation techniques have evolved during a material's installation period.

The following graph is for post 1945 cast iron only:

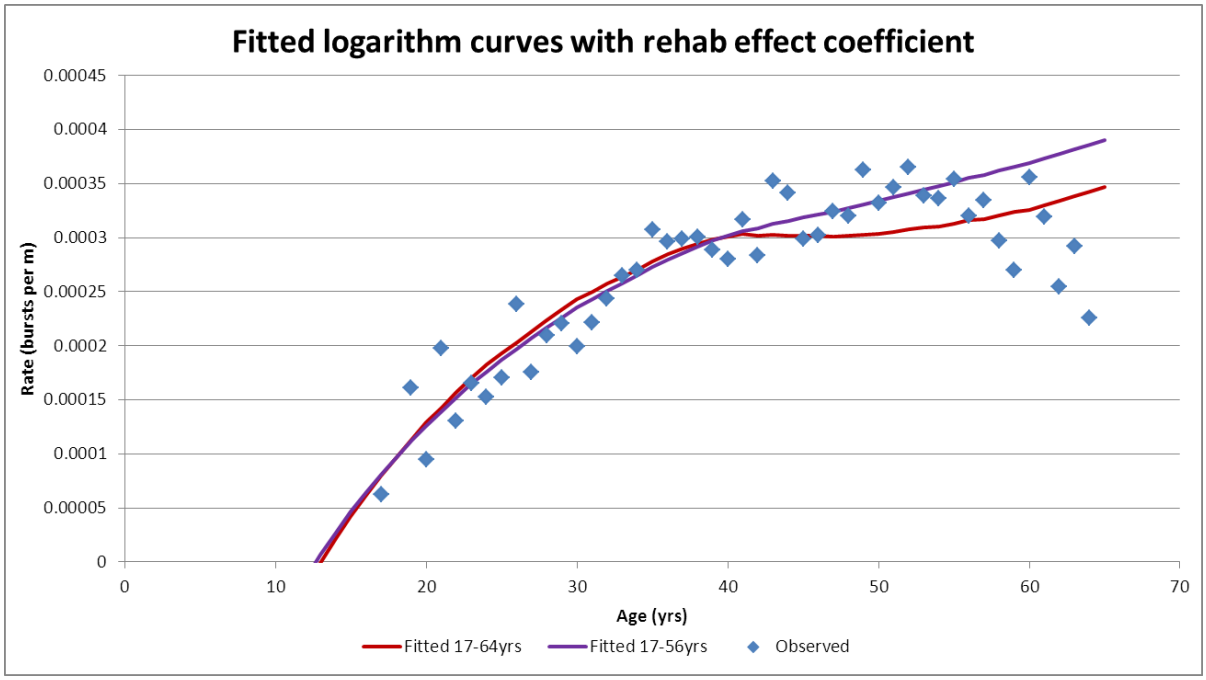


This graph shows a better trend with age, although it doesn't make sense that the burst rate increases with age, to a point, and then reverses and improves again as pipes get really old.

This effect seems to be a result of the renewals intervention already applied by the time the pipes reach the later ages. By removing the mains that fail the most, the overall average failure rate in each age band is significantly reduced, artificially exaggerating the flattening of the curve. This can't be a true deterioration pattern, with pipes beginning to fail less as they age, and ultimately ceasing to fail by the time the assets hit a certain age (i.e. approx 100yrs in above example).

5.2 The 'rehab effect'

Attempting to integrate a cumulative amount of rehab into the regression generated mixed results:



The intention is to balance the observed failures against an additional coefficient representing the cumulative renewal activity applied to the asset group up to each individual age, thereby countering the improvement effect delivered by the renewal activity to reveal the true base level of deterioration. This however, proves very difficult, because it relies on trying to fit a line to data that doesn't exist (i.e. the performance of the asset group if renewals had never been undertaken).

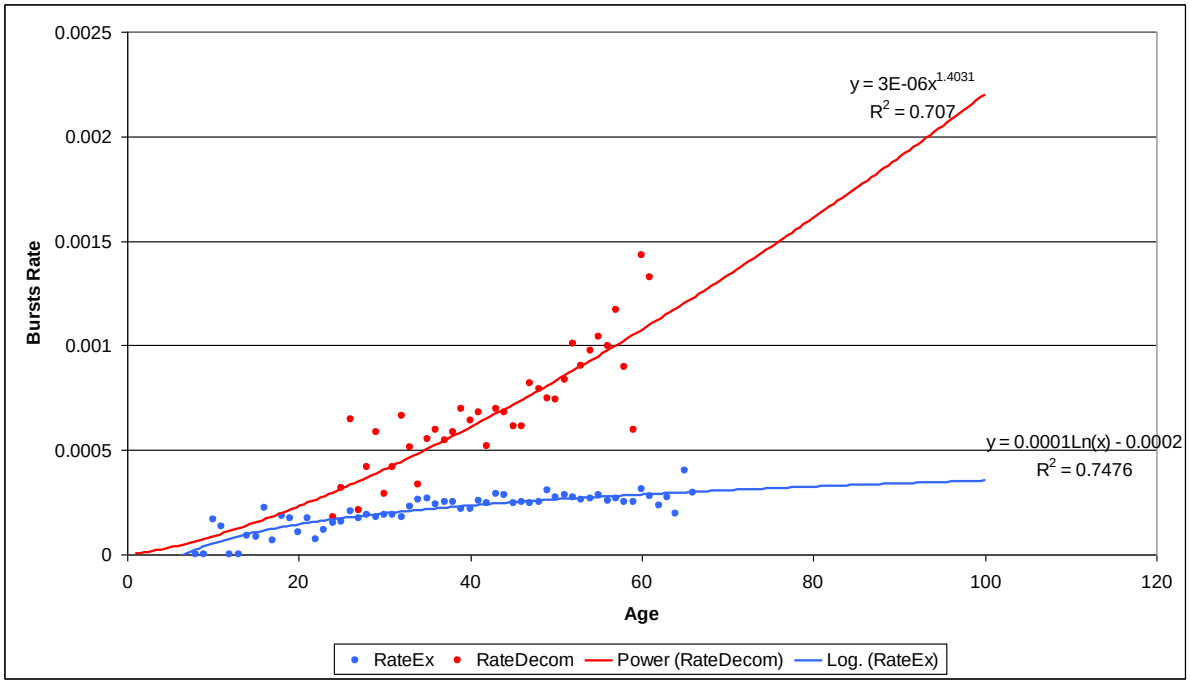
This approach is also of limited use for forecast models, because the user would need to predict the amount of renewal activity in each asset group for each year. The model then becomes autoregressive and reliant on the previous year's rehab amount (and more significantly the specific age bins where that rehab was undertaken), so the level of error increases quickly over time and reliability of predictions for 25 years from now would be extremely questionable.

5.3 'Dual Curves' approach

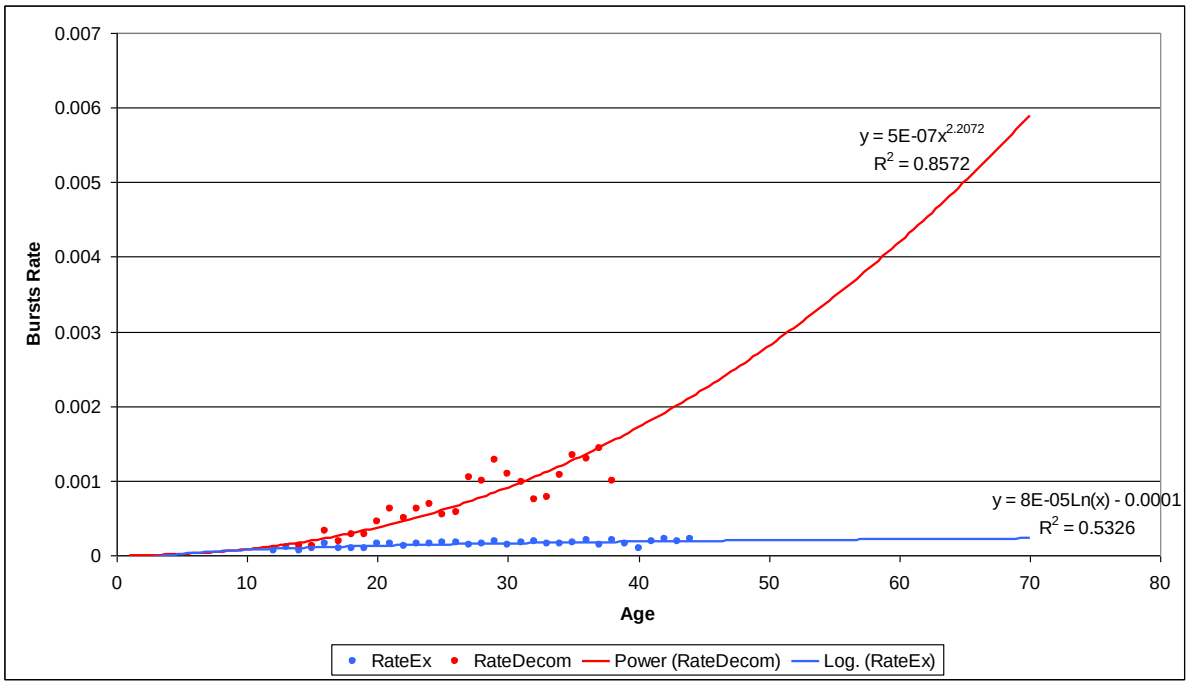
An alternative approach was then trialled, attempting to separate the models by analysing the assets still existing in a separate bin to the assets already decommissioned. This removes the need to take account of the 'rehab effect'.

The following charts show the observed trends for decommissioned assets in red and the existing assets in blue:

Examples for post 1945 cast iron:



And again for pre-1985 PVC:



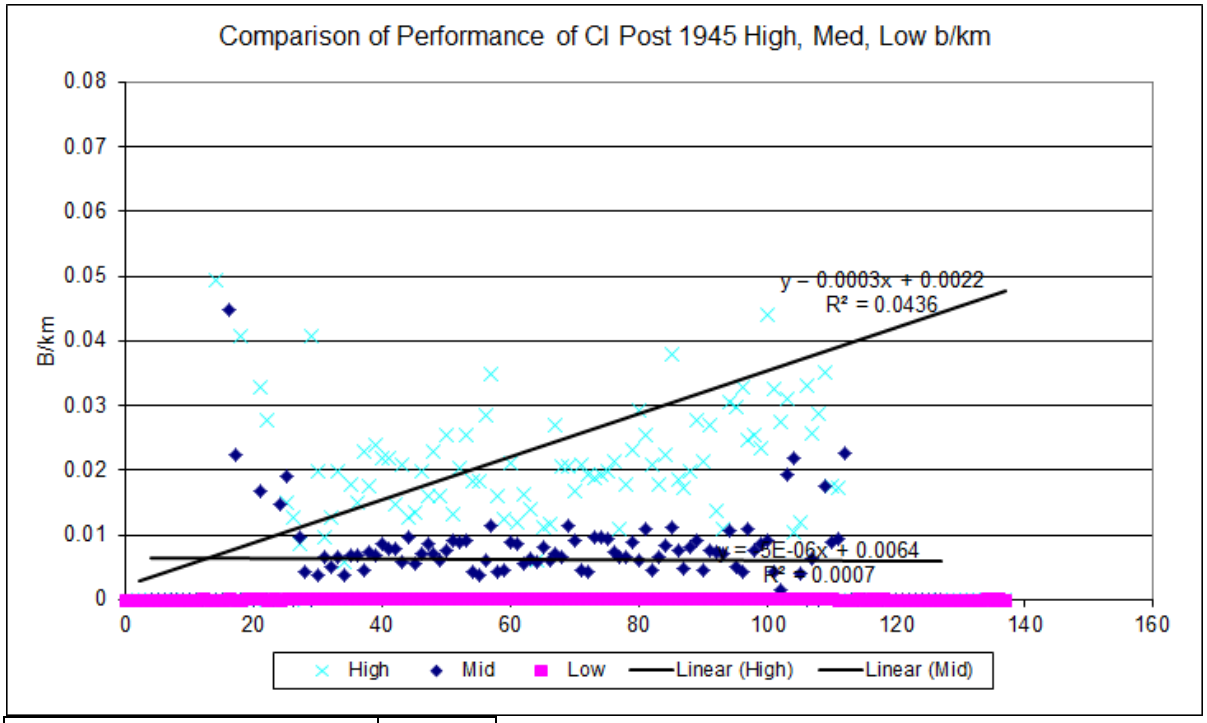
There are two problems with utilising the above method:

- Interpolating the two lines into a single useable model is difficult. The red lines (already decommissioned) represent the worst performing assets, historically these have made up a fairly small percentage of the asset stock; i.e. you could argue that the combined model should be influenced by 99% of the blue line and 1% of the red line (since only 1% of the asset group burst at or above the rate of the red line). But this percentage split won't remain constant over time, and it's expected that the disparity of burst rate at an individual asset level is likely to increase over time.
- The second problem is that by separating the existing and decommissioned pipes, the amount of data available for analysis in each bin is reduced, affecting statistical accuracy.

However, the fit statistics for the above lines are generally better than those achieved by trying to fit lines to combined failure data.

5.4 'State Transition' approach

There is another potential method, only explored at a basic level, analysing the distribution of poor performing pipes v's average performing pipes in each age. This would involve specifying different states (i.e. banded failure frequencies), and then modelling the transition of assets from one state into the next. This could be a better method of predicting when assets will move up the scale into the 'burstier' state.



Sum of SumOfSumOfS_SLEN	
Band	Total
HIGH	0.08%
LOW	99.64%
MID	0.28%
Grand Total	100.00%

As the R^2 value indicates, this method isn't robust in forecasting the performance, especially when considering that the asset length included in the high burst group accounts for less than 1% of the cohort.

There is also more scatter than expected in the 'High' and 'Mid' bands, although this may reduce if other predictors are included, but this could reduce the asset exposure even further, reducing the model stability.

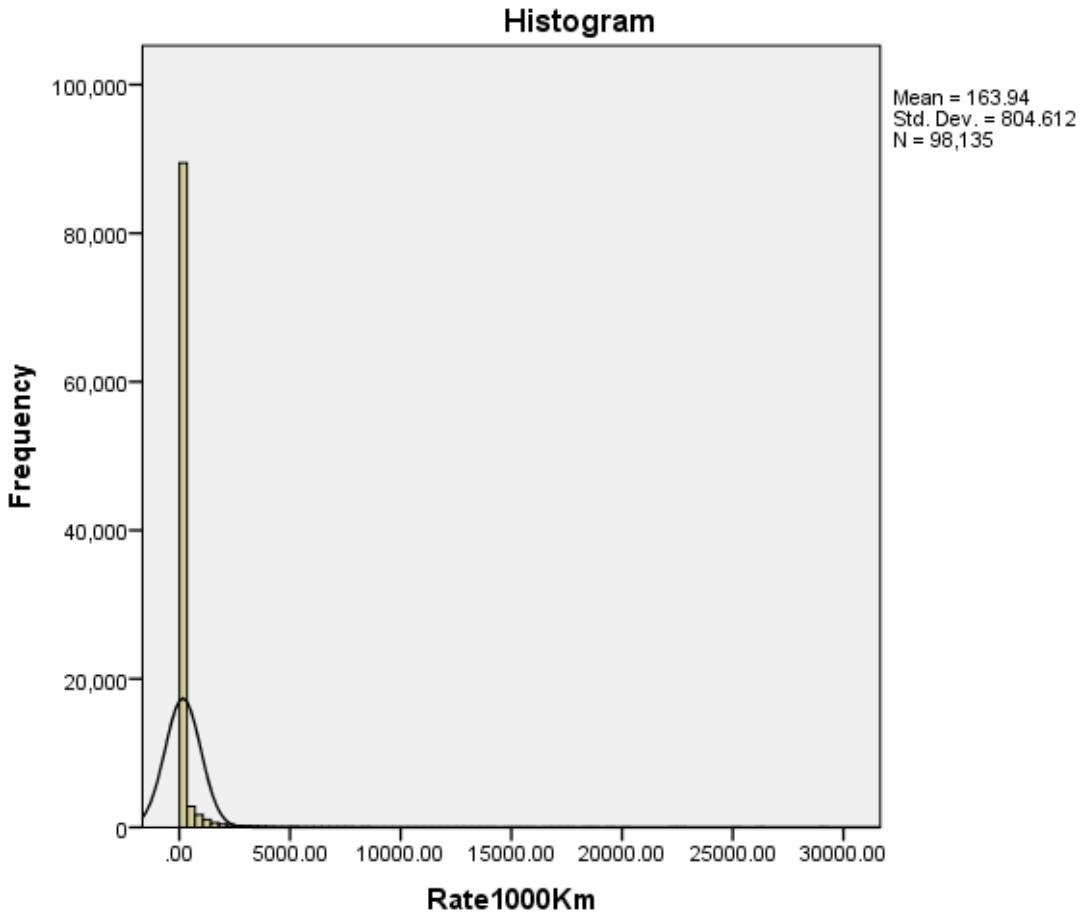
5.5 Final model development

5.5.1 SPSS investigations

The following couple of pages include some charts from SPSS. These are the results of some preliminary analysis looking at the normality of the whole dataset, and at some of the correlations between predictor variables.

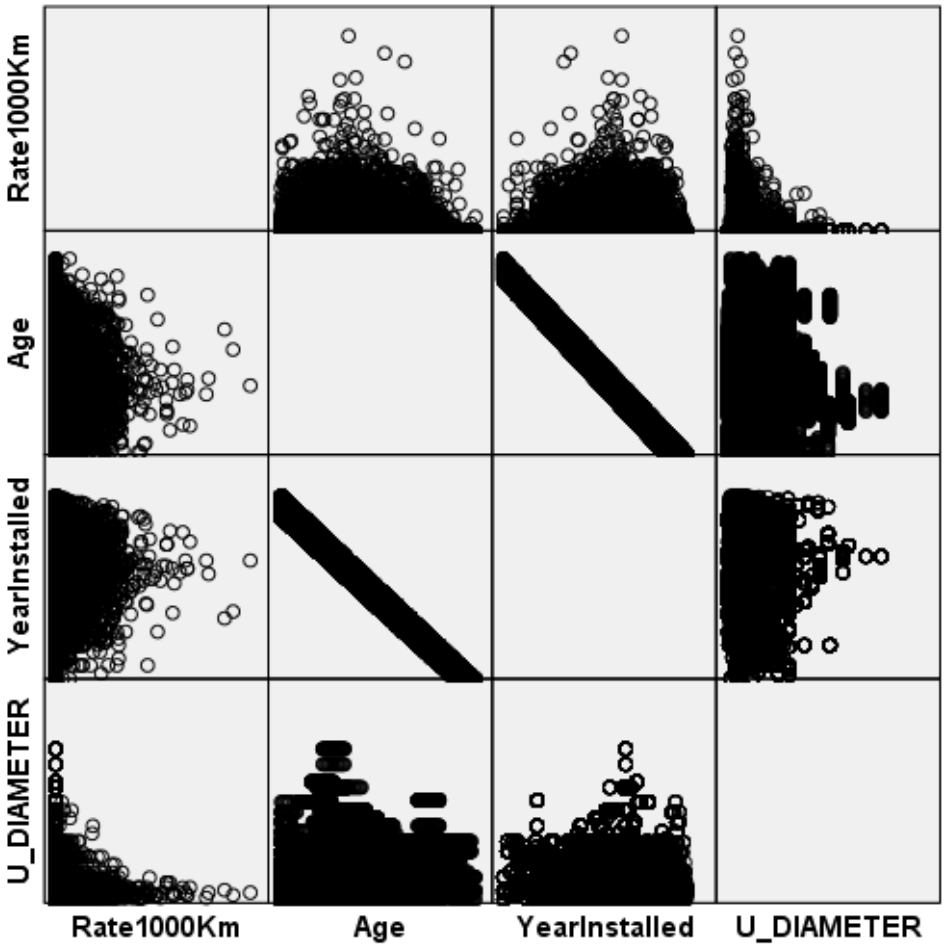
Frequency plot

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Rate1000Km	98135	.00	28981.20	163.9354	804.61168	647399.957	9.849	.008	151.265	.016
Valid N (listwise)	98135									



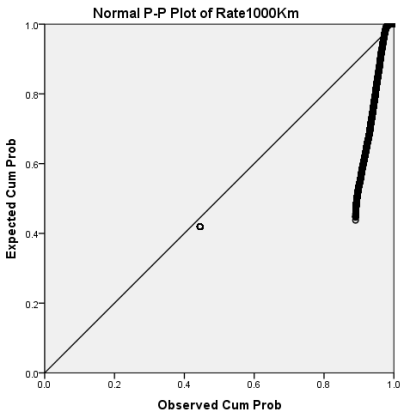
Findings: the data shows a lot of groups with a zero failure rate. Despite showing almost 100,000 samples, the observed population data isn't normally distributed. The 'heavy-tailed' population distribution (lots of zero rate) probably indicates the distribution of errors may also be non-normal and that efforts need to be taken to assess normality in the sub-sets of data utilised later in the analysis.

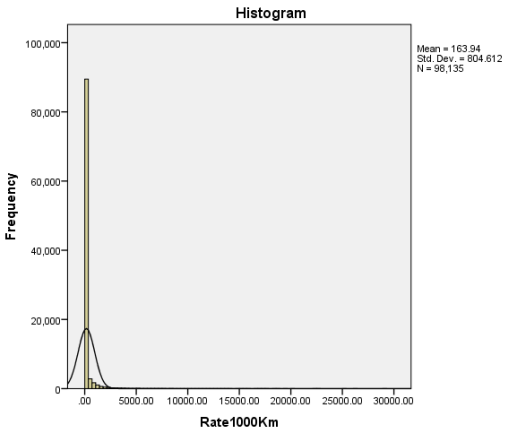
5.5.2 Scattermatrix of scale variables (investigation of cross correlation)



Findings: seems to indicate strong inverse relationship between diameter and failure rate. Indicates a pattern between age (or year installed) and failure rate, although this is messy due to all materials being analysed together. Shows perfect relationship, as expected, between age and year installed – need to take care with future analysis, because this co-linearity could be problematic for certain modelling techniques (i.e. need to use one or the other, but not both).

P-P Plots – Assessing Normality





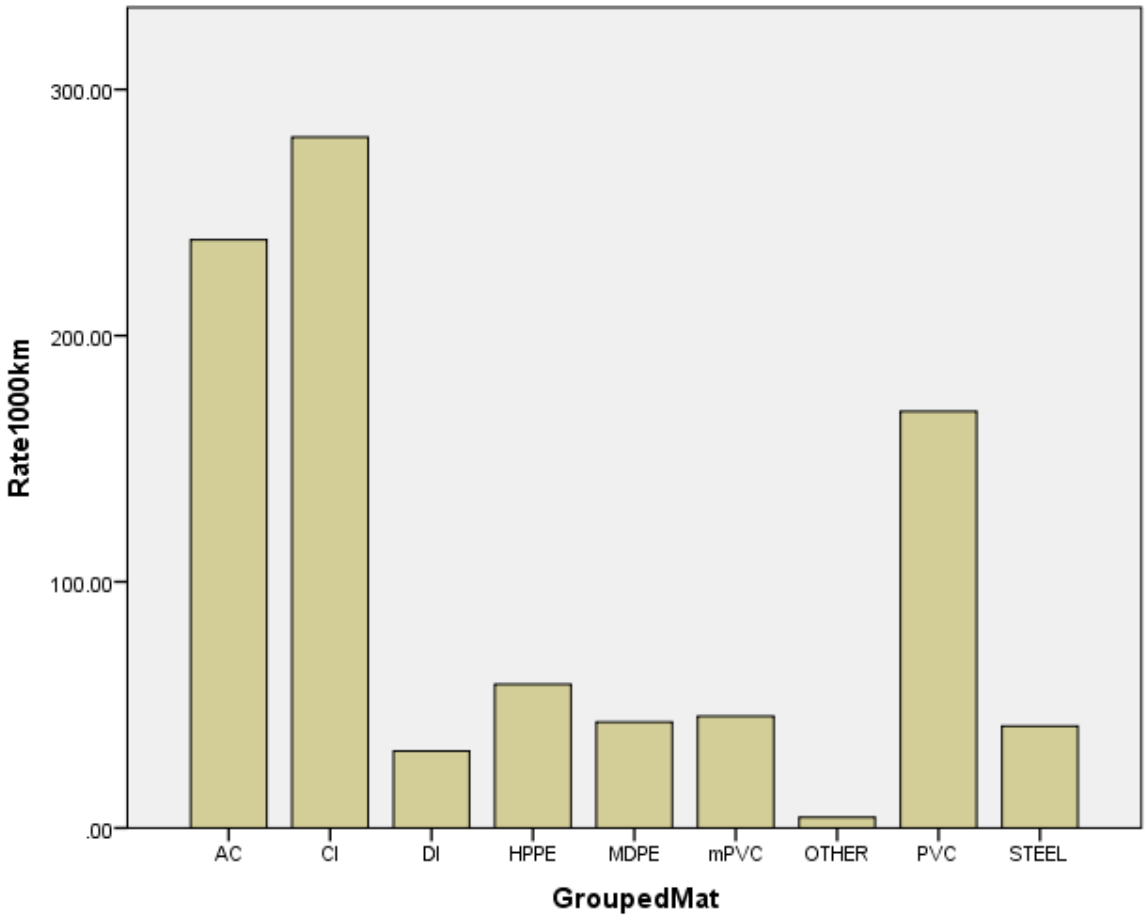
Findings: data is not normally distributed. Very positive skew (lots of low values), and very high positive kurtosis (pointy profile)

This is to be expected when analysing the full dataset, with a mode of zero, no negative bursts.

5.5.3 Grouping and testing

As expected, there are obvious issues with applying parametric statistical techniques to the raw string data. The following charts look at analysing smaller grouped sub-sets of the data, to see if this addresses normality issues etc:

Material Aggregation



As determined in the preliminary analysis stage, burst rates differ greatly between materials.

5.6 Conclusions - Developing analysis

6 Final Analysis

This section of the report gives a detailed breakdown of how the models have been generated and chosen for each of the material groups.
This analysis builds upon the work in section 5, utilising 'SPSS Statistics' to carry out regression, visual analysis and decision tree modelling.

Note; some of the charts do not include axis titles or units. Unless otherwise stated:

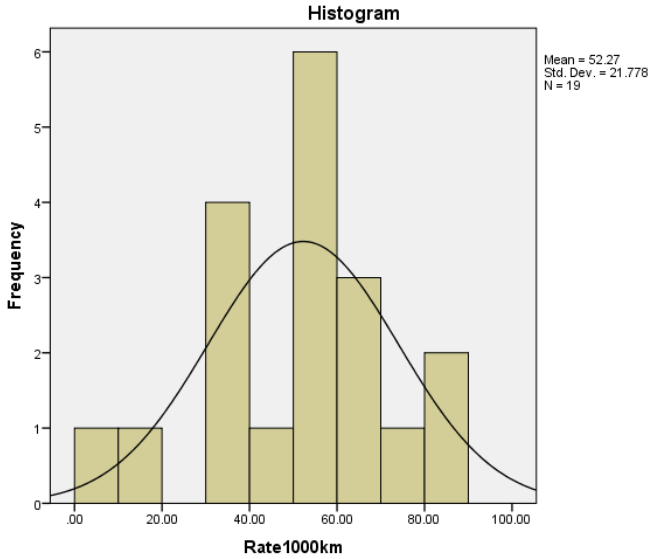
- Rate refers to the bursts rate, given in number of bursts per year per 1000km of mains length (b/1000km/yr)
- Diameters are generally quoted in mm, although some of the comments may refer to imperial measurements for older materials. Note that any SPSS analysis is always performed on metric equivalents – i.e. 4" = 101.8mm

6.1 Regression analysis by material

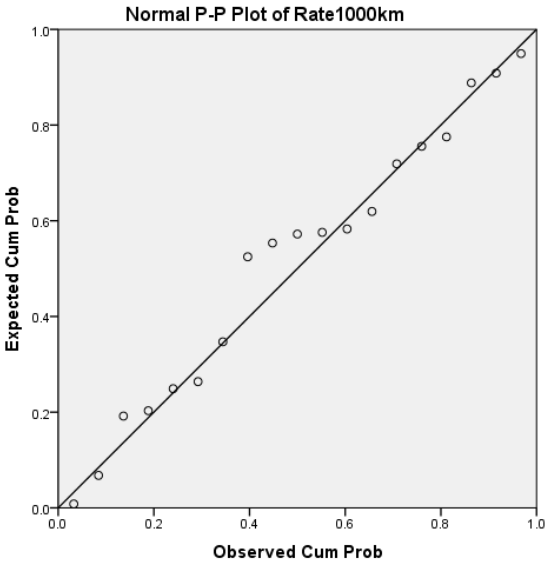
Note that this section isn't particularly easy to read, containing many charts and tables but only comments where they are deemed necessary to explain decisions.

6.1.1 HPPE

Removing outlying ages with <5km applicable length yields following rate frequency distribution:

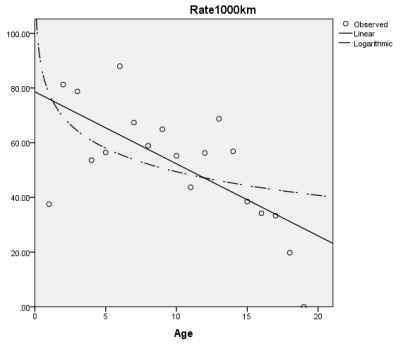


P-P plot also indicates reasonable normality:



Regression

HPPE



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

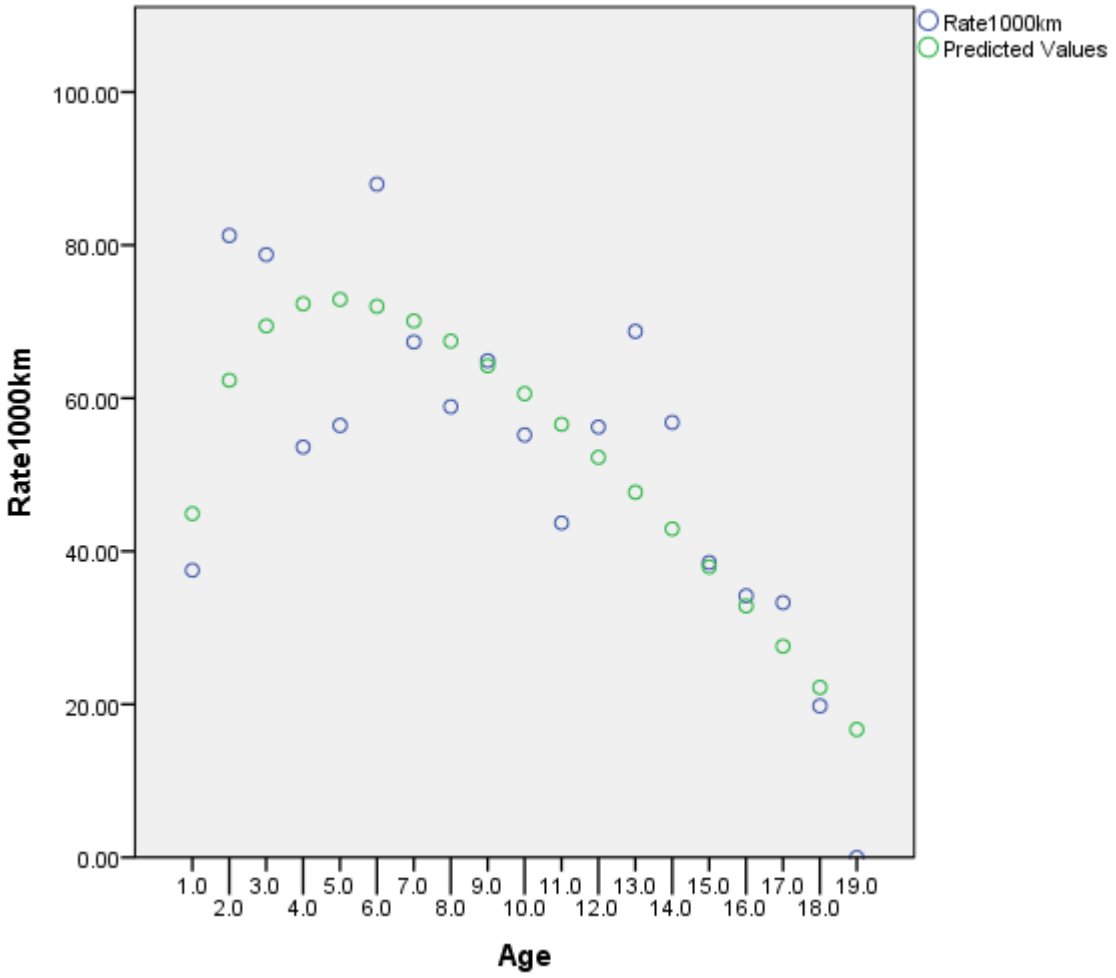
Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.463	14.651	1	17	.001	78.603	-2.633
Logarithmic	.213	4.606	1	17	.047	78.056	-12.452
Compound[a]	.	.	.	.	.	.	.
Power[a]	.	.	.	.	.	.	.
S[a]	.	.	.	.	.	.	.
Growth[a]	.	.	.	.	.	.	.
Exponential[a]	.	.	.	.	.	.	.
Logistic[a]	.	.	.	.	.	.	.

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Linear is effective for age, but would result in negative rate at future ages. Logarithmic is more appropriate in reality.

Non-linear modelling achieves better fit, but retains issues that negative burst rates are possible: (green shows model predictions, blue is observed)



ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	57831.419	4	14457.855
Residual	2622.650	15	174.843
Uncorrected Total	60454.069	19	
Corrected Total	8536.671	18	

Dependent variable: Rate1000km

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .693.

Based on y= a +(bx) + (c * ln(x))
Estimated parameters:
a: 52.347
b: -7.433
c: 35.863

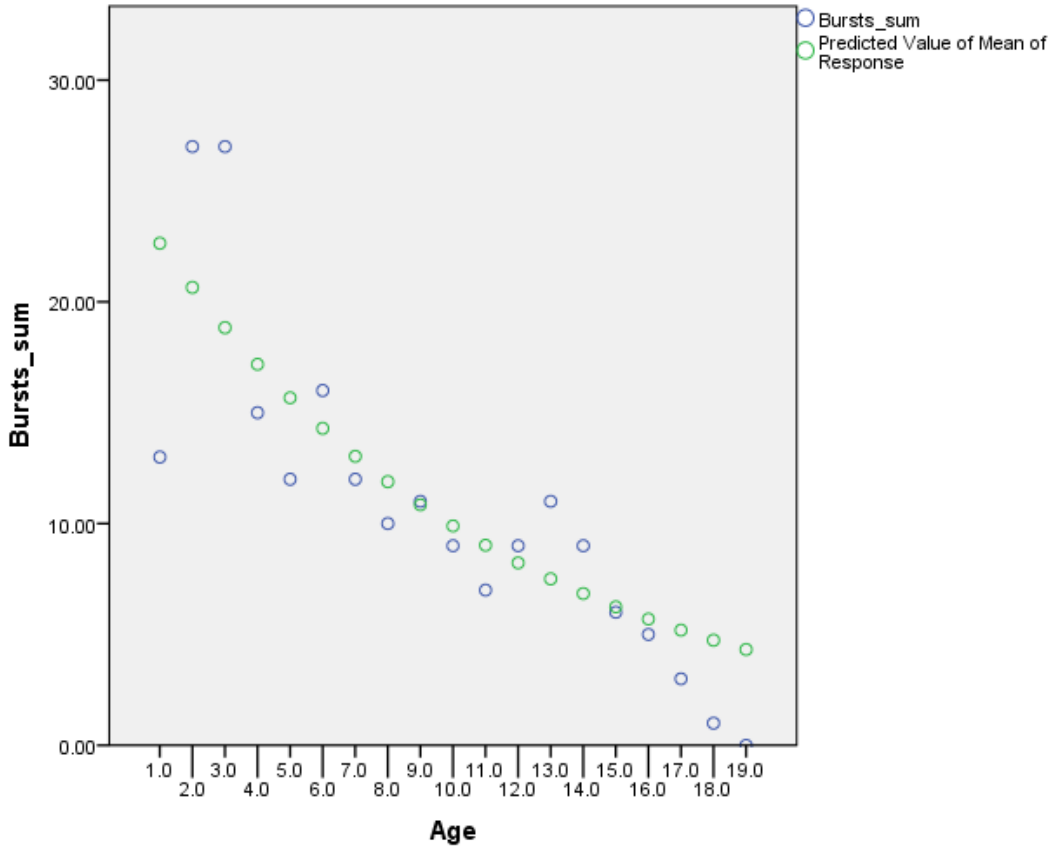
It still seems likely that any fitted curve is likely to be stretched along x-axis if regression is repeated in future (i.e. next amp)

Regression Tree

No significant variables identified when aggregating data bins for soil corrosivity, fracture potential and pipe diameter.

The regression tree didn't select any sub-categories even with reduced significance constraints. GLM also utilised, showing similar results with stepwise method using fracture potential first, although examination of means show that effect isn't significant enough to justify dividing data.

Curve selection for age predictor



Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	3.212	.0002	3.211	3.212	187633396.5	1	.000
Age	-.092	3.3339E-005	-.092	-.092	7615600.390	1	.000
(Scale)	1[a]						

Dependent Variable: Bursts_sum
Model: (Intercept), Age

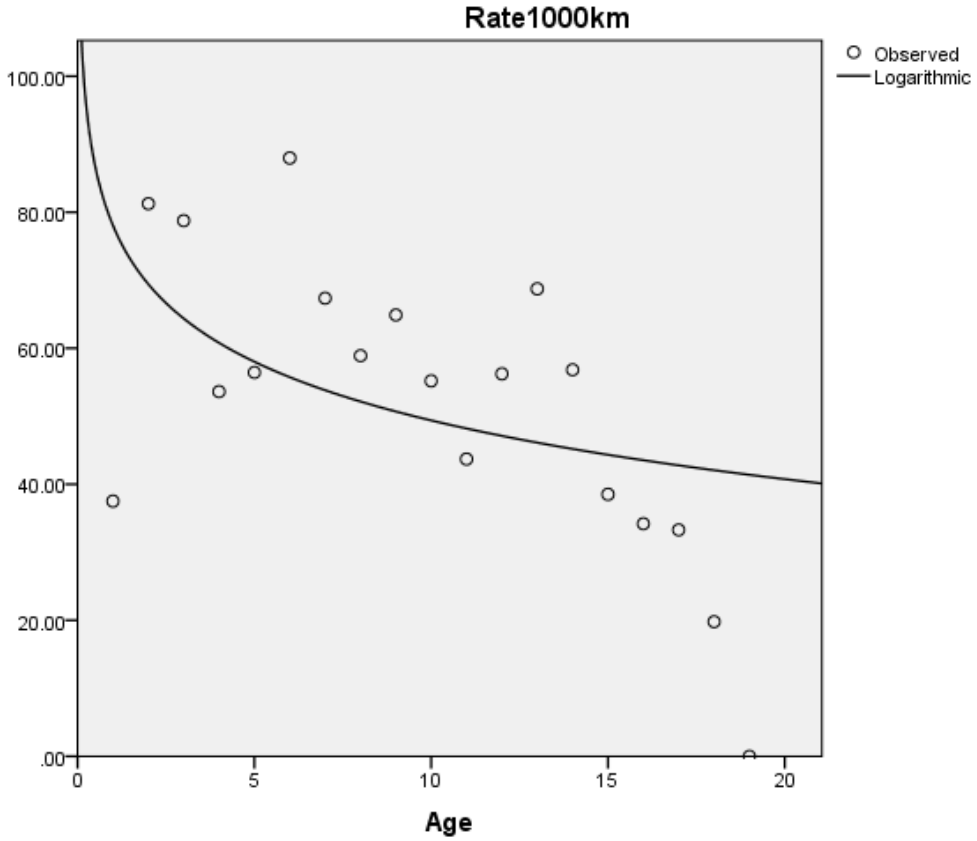
a. Fixed at the displayed value.

Above chart shows observed and predicted burst counts using poisson loglinear GLM. Deterioration model will use this curve for all HPPE up to 50yrs age. An alternative presumptive model will then be used to reflect probable age related deterioration of pipe material and joints. Model will presume simple and gradual linear deterioration until future industry research or observed bursts are able to make better prediction.

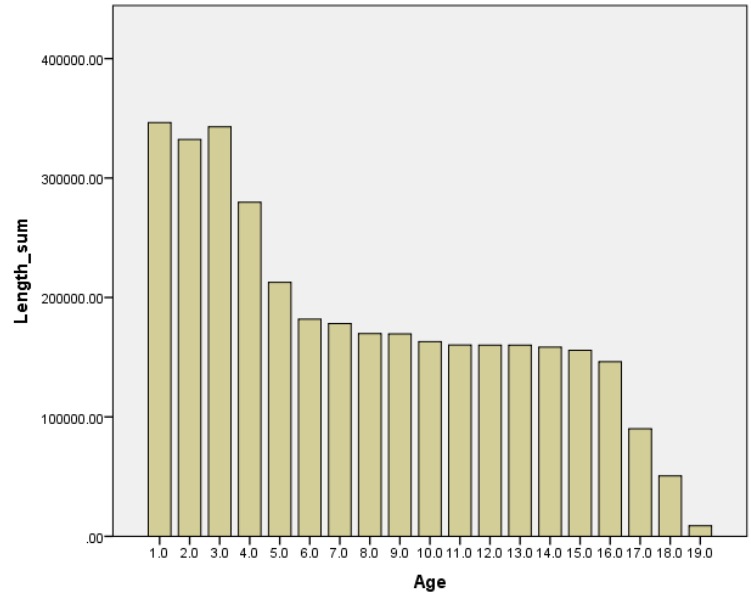
Reapply: count bursts = exp(intercept + coefficient * age)
Need to think about how this fits with proposed application of models, can this curve be effectively reapplied within an iterative process where lengths increase?

Answer to above is ‘not easily’. Despite apparently better fit, the model becomes difficult to apply over a 40 year time period because the lengths at each age will vary significantly. This is where ‘rate predicting’ models are easier, because they are already normalised against the pipe exposure and can be reapplied ‘as is’ to the forecast lengths.

Following graph shows ordinary logarithmic regression for rate:



Seems to overpredict at extremes, but these points have minimal exposure, less confidence.



Final model: <50 yrs

With low exposure outliers removed:

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	55439.672	2	27719.836
Residual	3514.224	14	251.016
Uncorrected Total	58953.896	16	
Corrected Total	3716.112	15	

Dependent variable: Rate1000km

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .054.

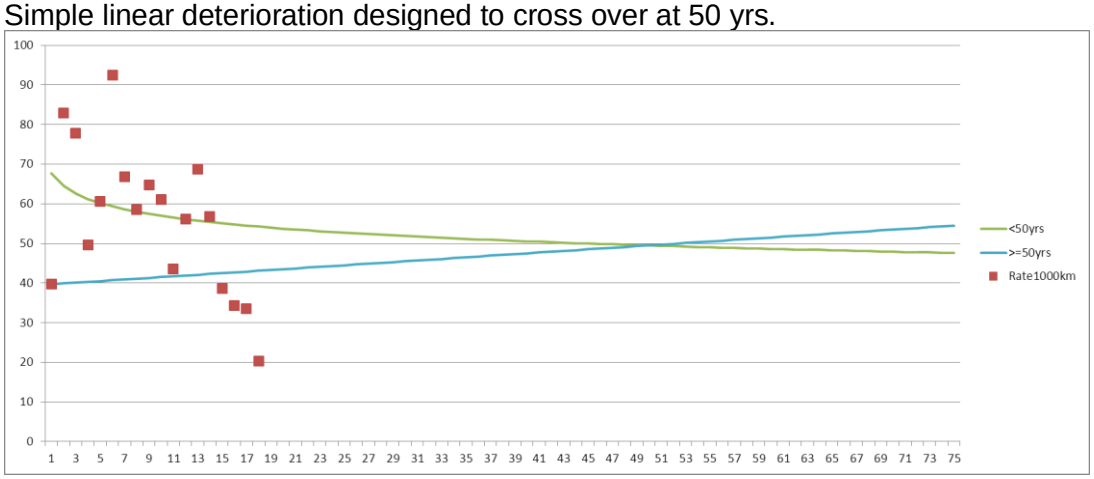
Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	-4.649	5.184	-15.767	6.469
b2	67.669	10.697	44.725	90.612

y (Rate1000km) = - 4.649 * log(Age) + 67.669

HPPE >= 50yrs

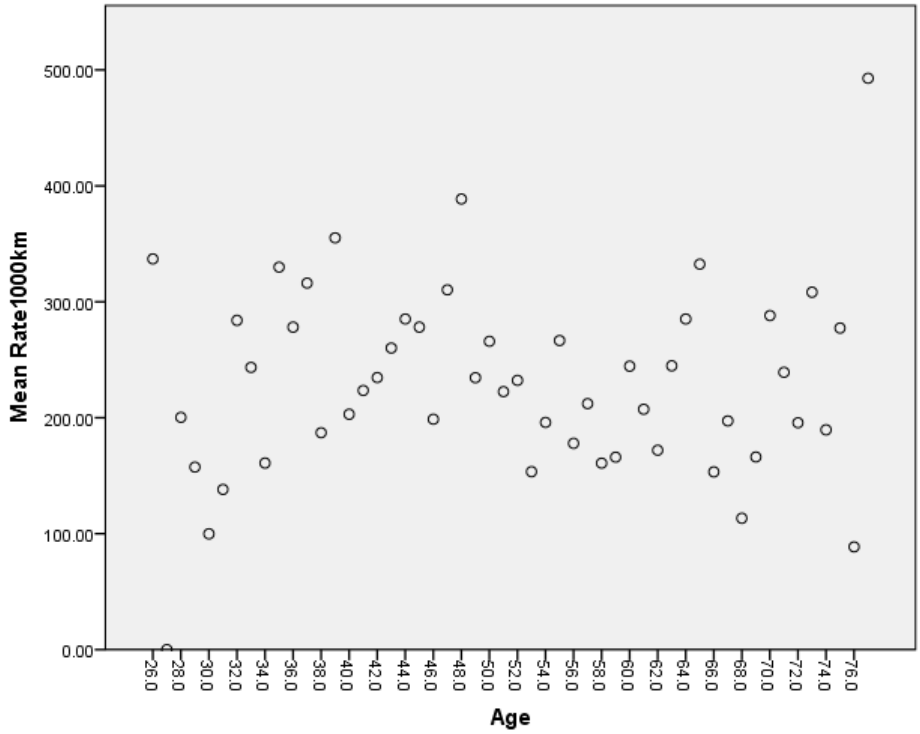
Simple linear deterioration designed to cross over at 50 yrs.



y = 0.2 * Age + 39.5

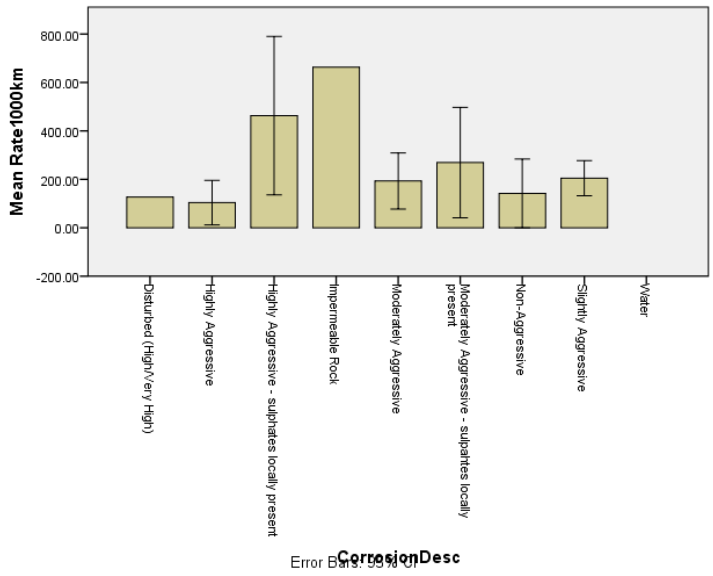
6.1.2 AC

Analysis by age shows little correlation/pattern:

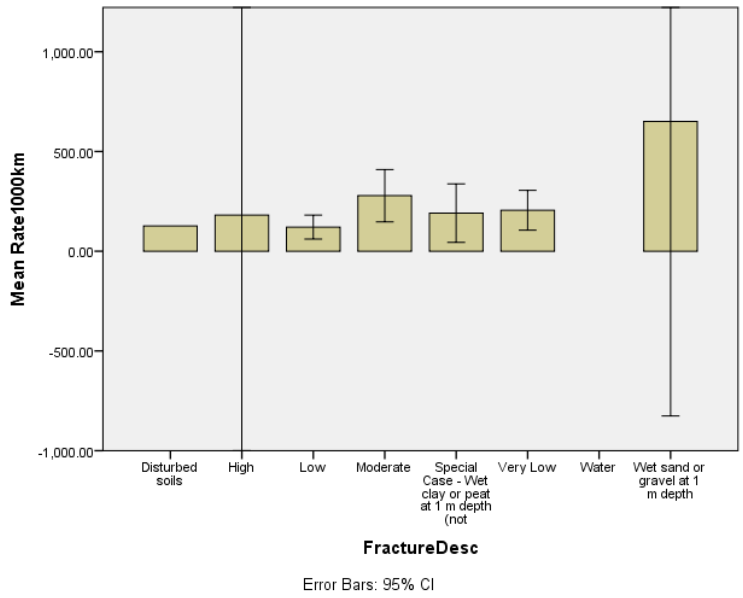


Either there is no age related deterioration for AC, or there are some sub-groups with no age correlation that are masking others.

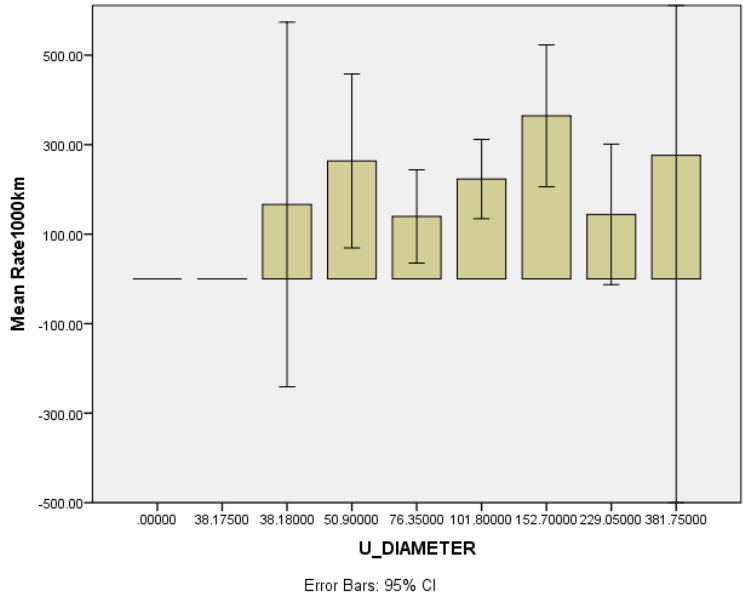
Analysis by other factors shows:



Some correlation by corrosion, although confidence intervals are large



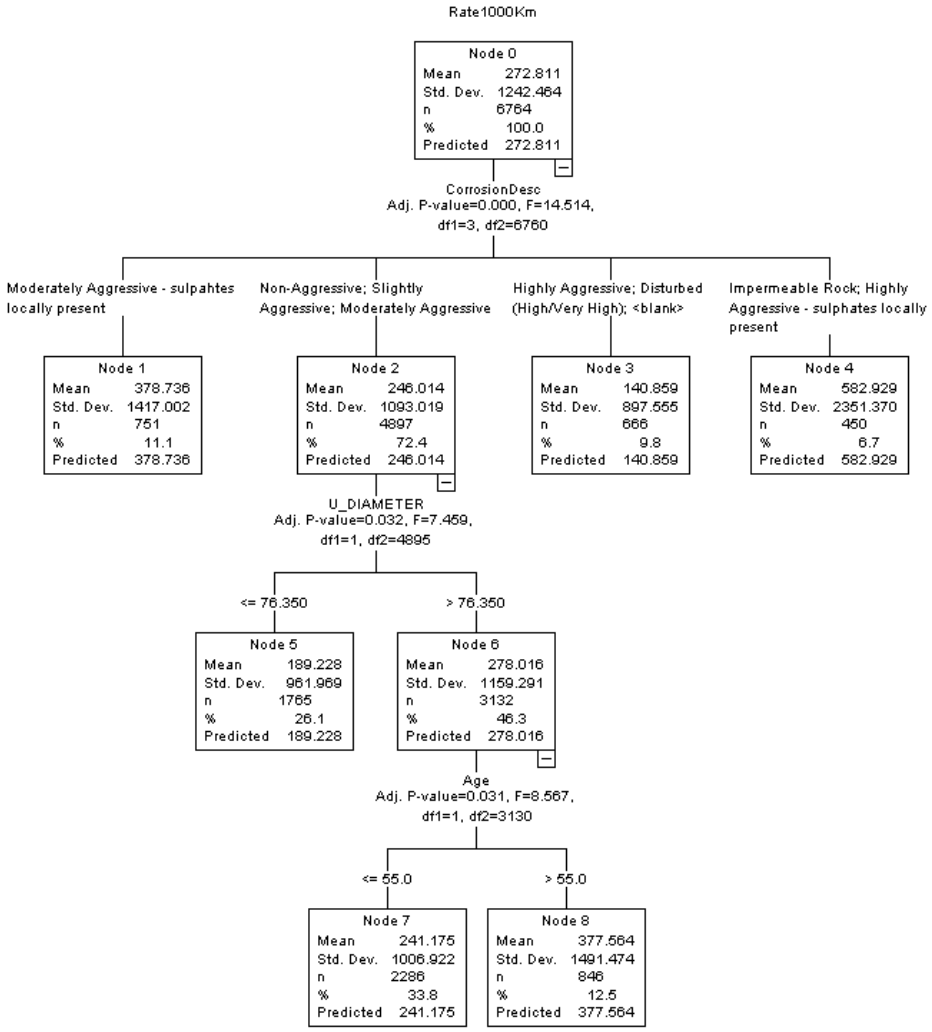
Less relationship with Fracture potential and highly variable around means.



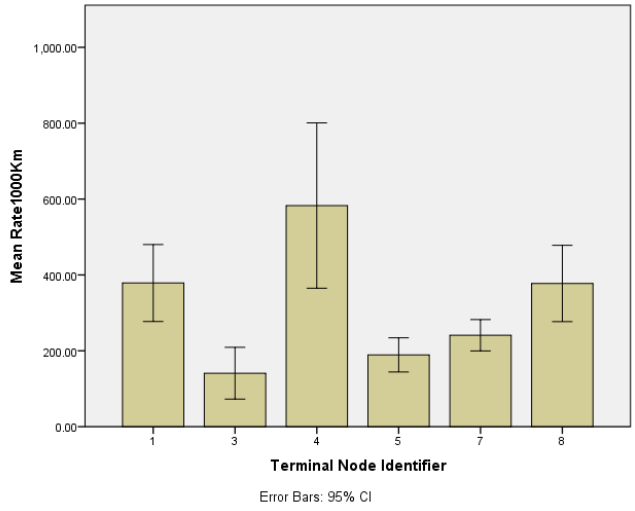
Reasonable correlation with diameter, but again, variable around means.
Interestingly, AC seems to show reverse of usual trend, with 6" worse than 4" worse than 3"…

Decision Tree for other factors

As with HPPE, this material shows no appropriate binning when age is ignored.
By retesting all predictors (age, diameter, corr & frac), then the following tree is generated:

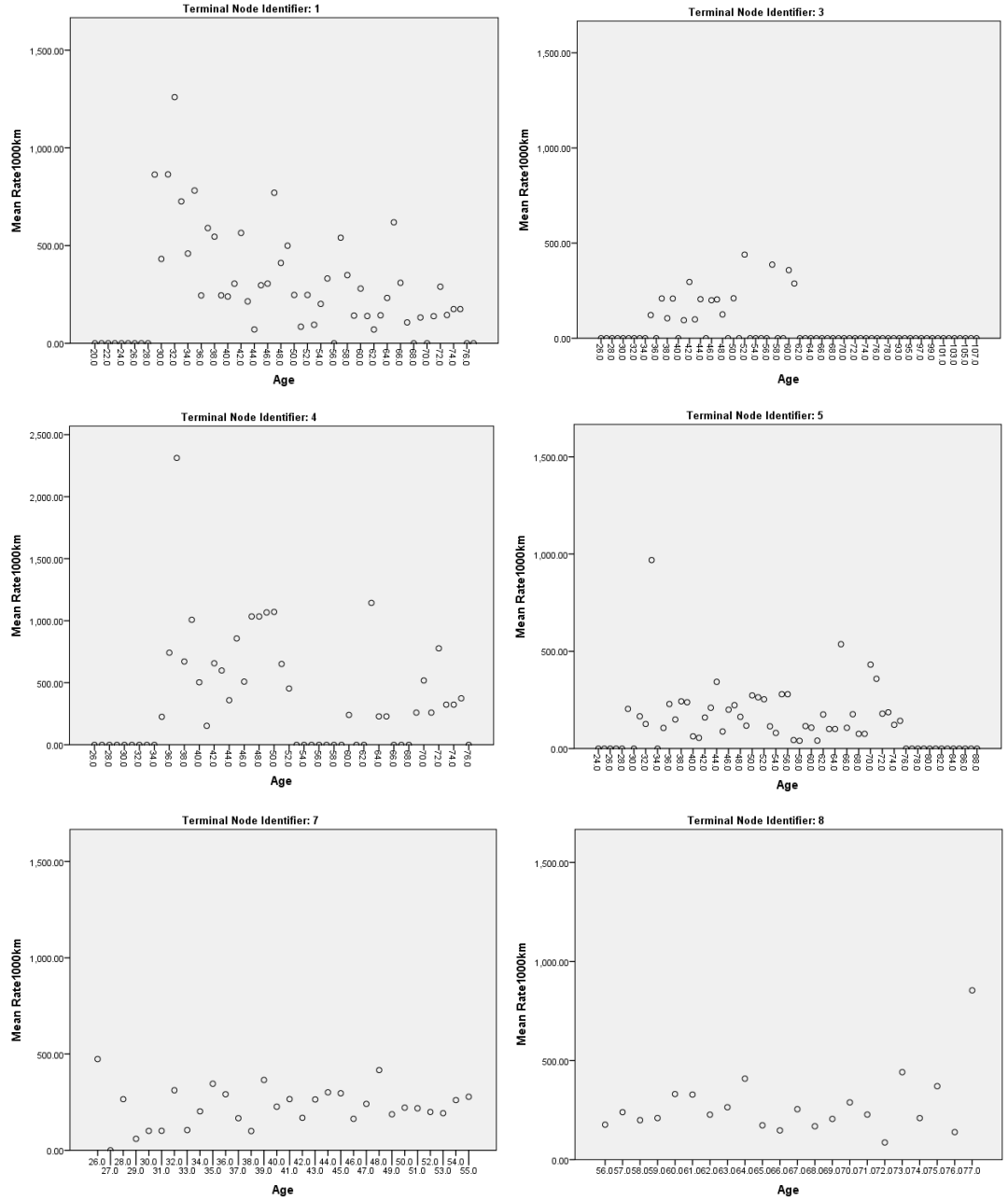


Analysing each terminal node then gives mean failure rates and confidence intervals as below (some nodes e.g 1, 4 & 8 are reliable combinations of significant predictors):



The grouped data does show reasonable differences in the means for each group, and reassuringly, the confidence intervals are notably narrower.

Best apparent method for modelling AC seems to be to take mean failure rate for each of these groups.



The only groups with sufficient length of assets and clear enough patterns to determine age related curve are nodes 7 & 8. These are far from perfect regression fits, but it makes engineering sense that there is some predictable age related deterioration in the more stable, less corrosive soils:

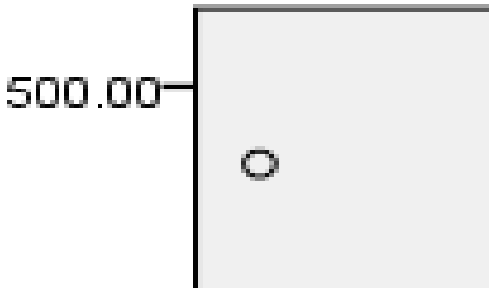
Node 7:

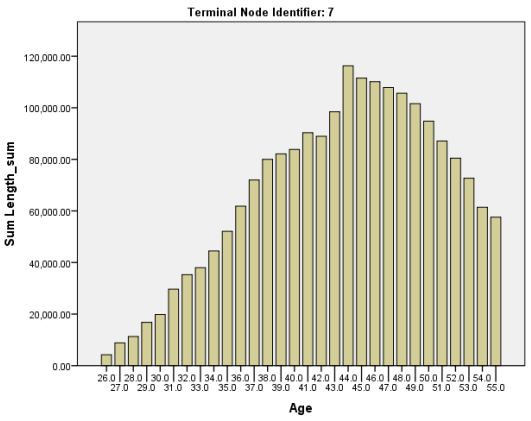
Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.034	.973	1	28	.332	138.090	2.178

The independent variable is Age.





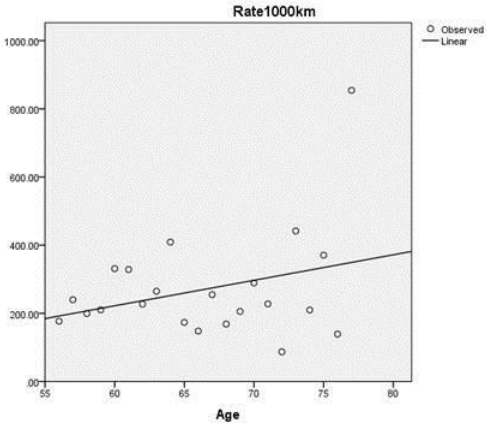
Node 8:

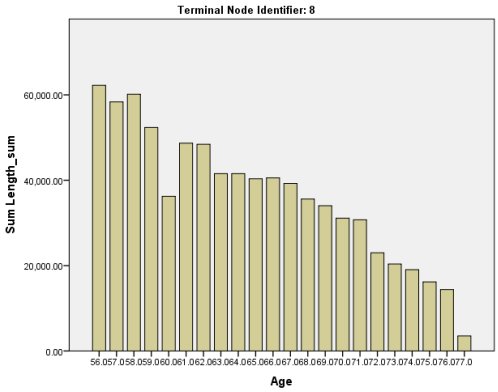
Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.095	2.096	1	20	.163	-227.004	7.485

The independent variable is Age.





Final 6 x AC models:

Node	Filter
1	(CorrosionDesc = "Moderately Aggressive - sulpahtes locally present")
3	(CorrosionDesc = "Highly Aggressive " OR CorrosionDesc = "Disturbed (High/Very High) " OR CorrosionDesc = " ")
4	(CorrosionDesc = "Impermeable Rock" OR CorrosionDesc = "Highly Aggressive - sulphates locally present ")
5	(CorrosionDesc != "Moderately Aggressive - sulpahtes locally present" AND CorrosionDesc != "Highly Aggressive" AND CorrosionDesc != "Disturbed (High/Very High) " AND CorrosionDesc != " " AND CorrosionDesc != "Impermeable Rock" AND CorrosionDesc != "Highly Aggressive - sulphates locally present") AND (U_DIAMETER NOT MISSING AND (U_DIAMETER <= "76.35000"))
7	(CorrosionDesc != "Moderately Aggressive - sulpahtes locally present" AND CorrosionDesc != "Highly Aggressive" AND CorrosionDesc != "Disturbed (High/Very High) " AND CorrosionDesc != " " AND CorrosionDesc != "Impermeable Rock" AND CorrosionDesc != "Highly Aggressive - sulphates locally present") AND (U_DIAMETER IS MISSING OR (U_DIAMETER > "76.35000")) AND (Age IS MISSING OR (Age <= "55.0"))
8	(CorrosionDesc != "Moderately Aggressive - sulpahtes locally present" AND CorrosionDesc != "Highly Aggressive" AND CorrosionDesc != "Disturbed (High/Very High) " AND CorrosionDesc != " " AND CorrosionDesc != "Impermeable Rock" AND CorrosionDesc != "Highly Aggressive - sulphates locally present") AND (U_DIAMETER IS MISSING OR (U_DIAMETER > "76.35000")) AND (Age NOT MISSING AND (Age > "55.0"))

Node 1: **y = 378.73**
Node 3: **y = 140.85**
Node 4: **y = 582.92**
Node 5: **y = 189.22**
Node 7: **y = 2.178 x Age + 138.090**
Node 8: **y = 7.485 x Age – 227.004**

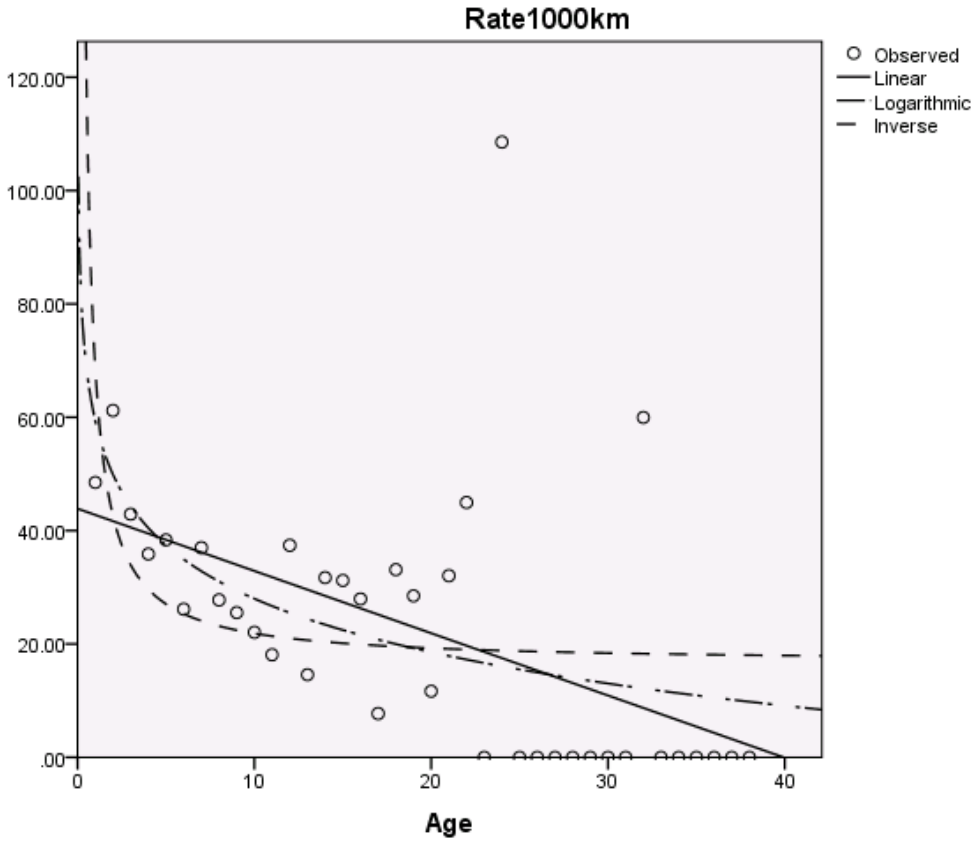
Descriptives

Rate1000Km

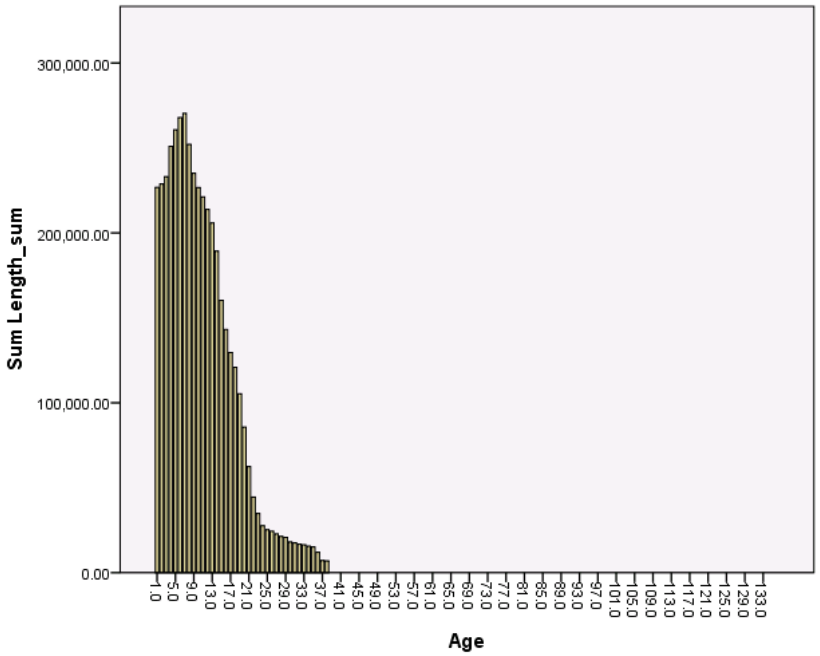
		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
1		751	378.7360	1417.00226	51.70715	277.2280	480.2439	.00	16460.63	
2		666	140.8588	897.55466	34.77954	72.5678	209.1497	.00	9102.33	
3		450	582.9287	2351.37046	110.84467	365.0899	800.7674	.00	28981.20	
5		1765	189.2281	961.96877	22.89753	144.3189	234.1372	.00	12430.16	
7		2286	241.1748	1006.92192	21.05997	199.8762	282.4735	.00	22764.68	
8		846	377.5644	1491.47404	51.27793	276.9173	478.2115	.00	17422.74	
Total		6764	272.8109	1242.46443	15.10714	243.1962	302.4257	.00	28981.20	
Model	Fixed Effects			1237.72230	15.04948	243.3092	302.3127			
	Random Effects				60.24021	117.9585	427.6633			15162.02626

6.1.3 DI

Analysis by Age shows some correlation, although after first 15 years, results get a bit more variable:

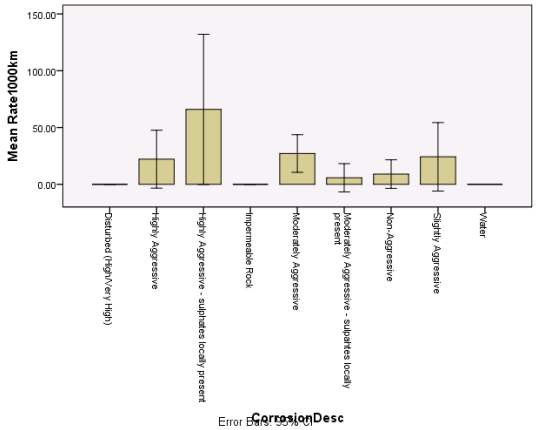


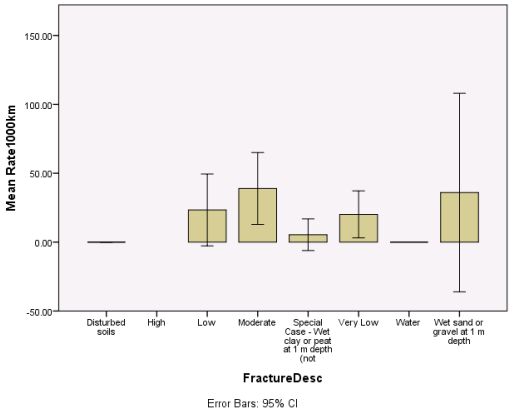
This increase in variance is probably due to exposure falling away sharply after 21yrs, as shown on the following graph:

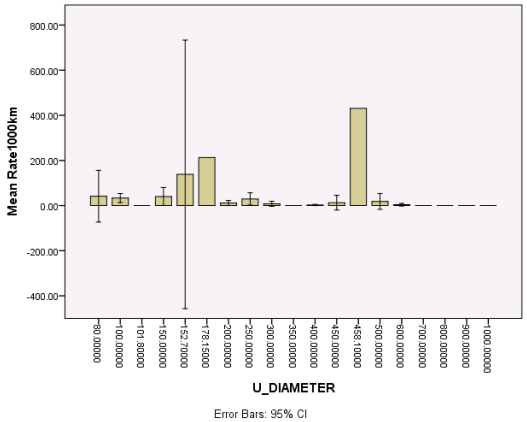


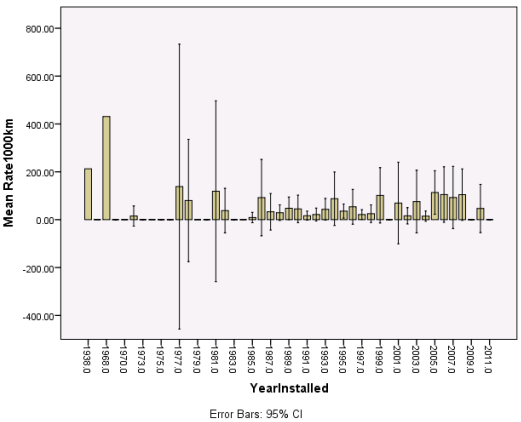
Other predictors

The other predictors were then assessed again using the regression tree. This didn't show any other significant predictors. The following charts showing means and 95% confidence intervals demonstrate some difference between indicators, but large confidence intervals prove excessive variance:

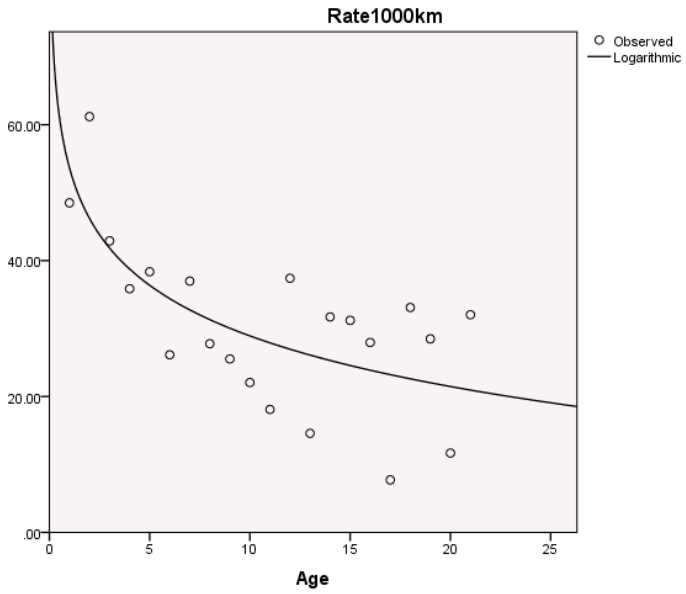








The most appropriate model for DI is probably a logarithmic curve up until 21yrs age, and then a flat rate from that point onwards:



ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	20998.316	2	10499.158
Residual	1482.243	19	78.013
Uncorrected Total	22480.559	21	
Corrected Total	3025.277	20	

Dependent variable: Rate1000km

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .510.

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	-10.743	2.416	-15.799	-5.687
b2	53.653	5.565	42.006	65.300

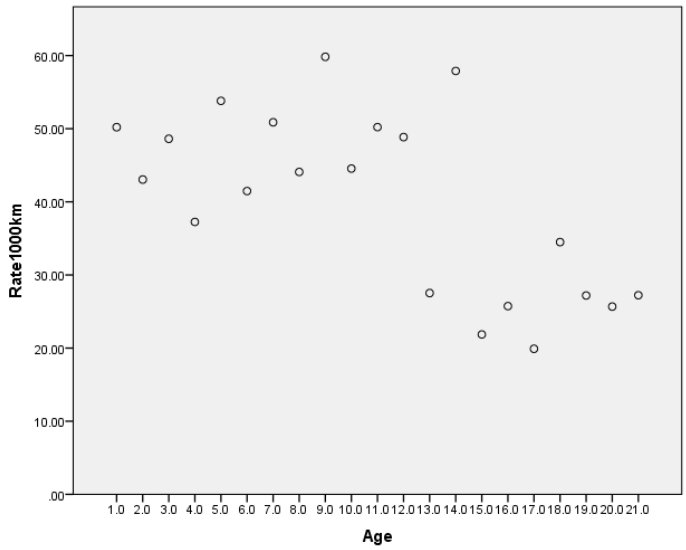
y = -10.743 * log(Age) + 53.653

Rate after 21yrs of age is: 53.653-10.743 * ln(21)
y = 20.95

There may be a need to assume some future deterioration for this material similar to the approach used for HPPE. This has not been attempted at this stage due to lack of data for older pipes and the expectation that coated ductile is likely to be a very stable and long life asset.

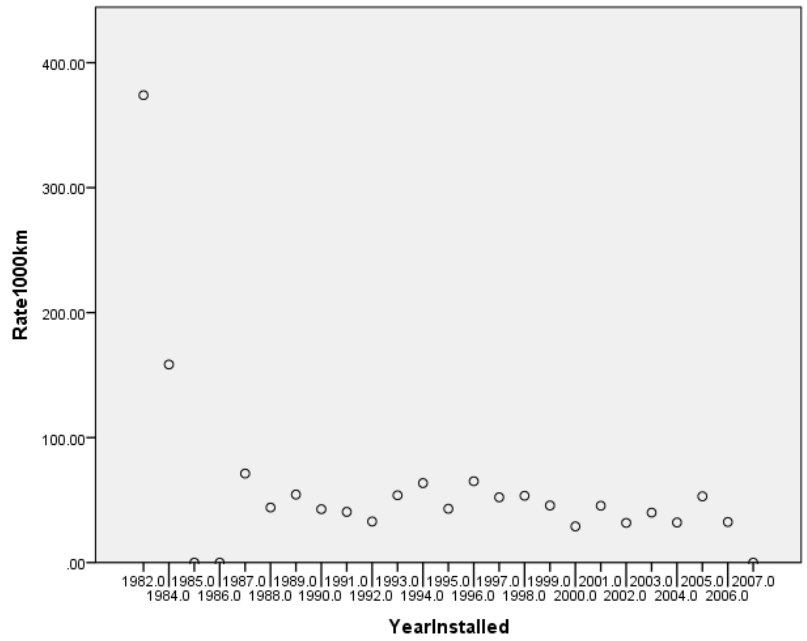
6.1.4 MDPE

MDPE shows an age related pattern, although the data seems to exhibit two clusters:



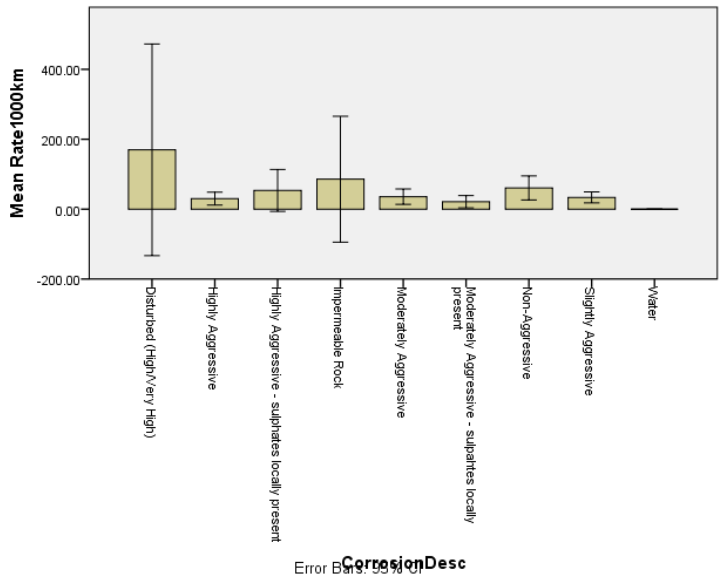
This could be due to a step change during a particular year, either an introduction of material changes, or more likely, jointing technique.

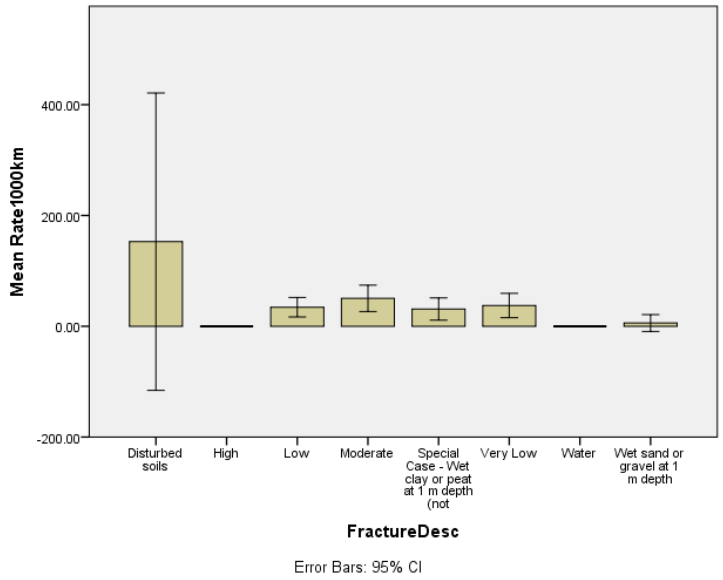
Repeating similar analysis on year installed reveals that this doesn't seem to be the case:

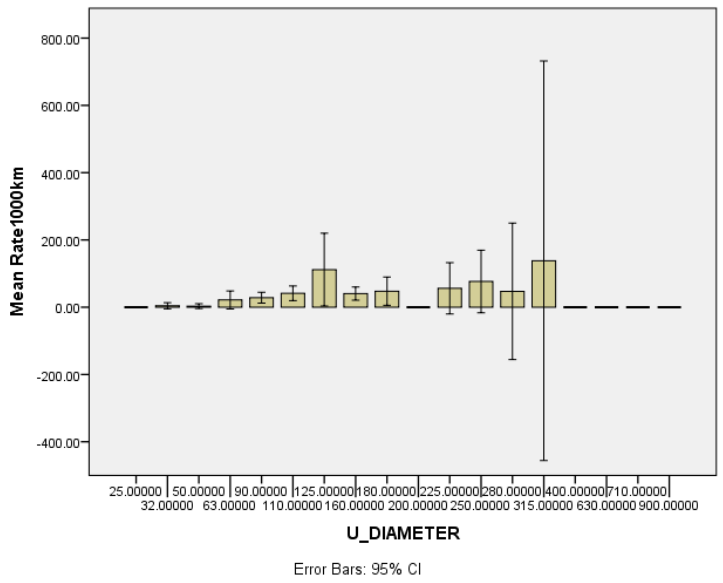


There are no apparent step changes for different eras of pipe.

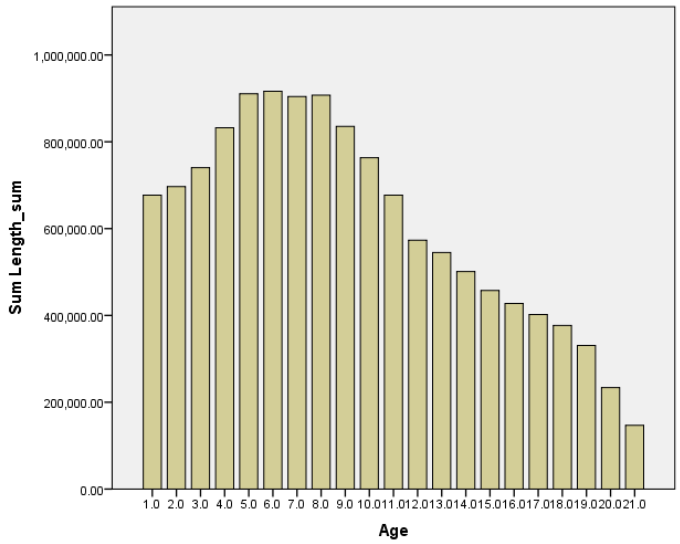
Analysis of the static predictors using a regression tree doesn't reveal any significant variables. This is backed up by looking at mean bursts rates individually for each predictor. As before, there are some patterns, but each showing a high level of variance:


Mean Rate1000km
400.00
200.00
0.00
-200.00
Disturbed (High/Very High)
Highly Aggressive
Highly Aggressive - sulphates locally present
Impermeable Rock
Moderately Aggressive
Moderately Aggressive - sulphates locally present
Non-Aggressive
Slightly Aggressive
Water
CorrosionDesc
Error Bars: 95% CI


Mean Rate1000km
400.00
200.00
0.00
-200.00
Disturbed soils
High
Low
Moderate
Special Case - Wet clay or peat at 1 m depth (not
Very Low
Water
Wet sand or gravel at 1 m depth
FractureDesc
Error Bars: 95% CI

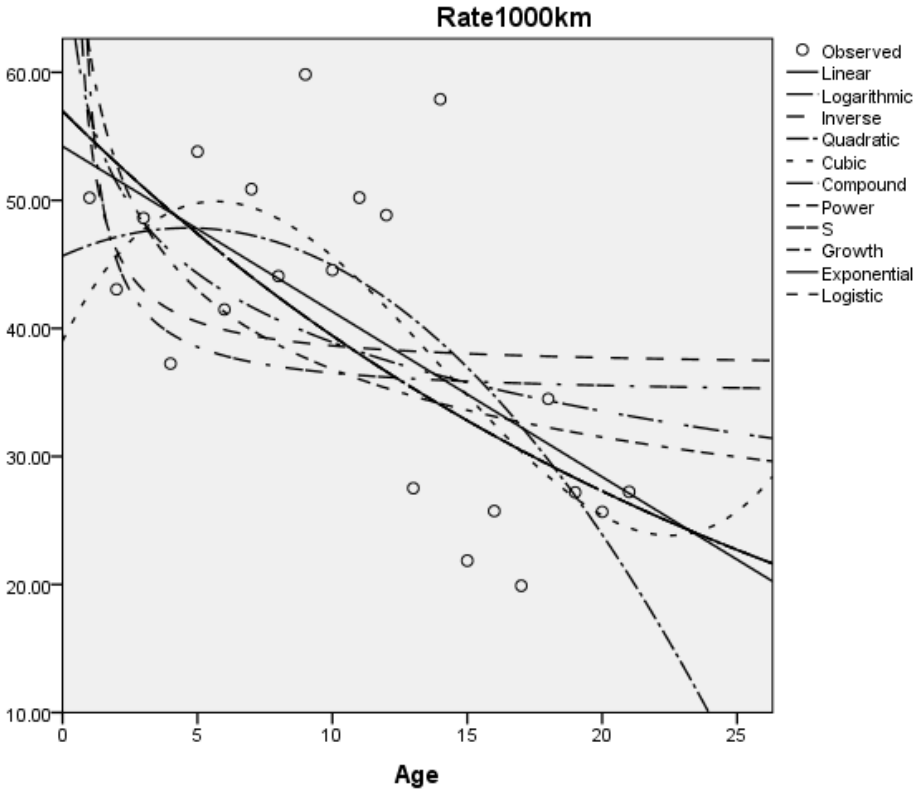

Mean Rate1000km
800.00
600.00
400.00
200.00
0.00
-200.00
-400.00
25.00000
32.00000
50.00000
63.00000
90.00000
110.00000
125.00000
160.00000
180.00000
200.00000
225.00000
250.00000
280.00000
315.00000
400.00000
630.00000
710.00000
900.00000
U_DIAMETER
Error Bars: 95% CI

The length of pipes at each age was reviewed, to see if the step change in the age graph could be due to a step change in exposure:



The length in each band changes quite gradually, so this is unlikely to be a significant problem. Also, any high influence outliers (generally with very low exposure) have already been removed.

There have been no new MDPE pipes installed since 2007. The most appropriate model is probably one which mimics the performance of the pipes up until 20 years age, and then stays flat until 40-50 years age, as per HPPE. A rate of age related deterioration then needs to be assumed for long term forecasting. This, like HPPE, is difficult to predict, but could be improved over coming years when more information is available.



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.415	13.455	1	19	.002	54.214	-1.291		
Logarithmic	.260	6.669	1	19	.018	56.773	-7.755		
Inverse	.108	2.308	1	19	.145	36.782	18.630		
Quadratic	.489	8.618	2	18	.002	45.659	.941	-.101	
Cubic	.513	5.975	3	17	.006	39.021	4.190	-.462	.011
Compound	.452	15.682	1	19	.001	56.961	.964		
Power	.293	7.860	1	19	.011	61.747	-.225		
S	.123	2.656	1	19	.120	3.544	.541		
Growth	.452	15.682	1	19	.001	4.042	-.037		
Exponential	.452	15.682	1	19	.001	56.961	-.037		
Logistic	.452	15.682	1	19	.001	.018	1.037		

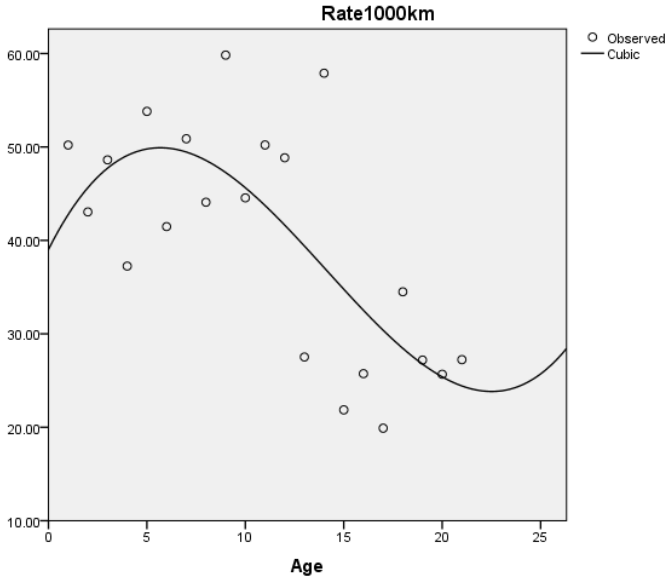
The independent variable is Age.

Final 3 x models for MDPE

Cubic polynomial up until 21yrs age, then flat mean rate until 50yrs age, then linear from 50yrs.

<= 21yrs:

Y = b1 + b2 * x + b3 * x^2 + b4 * x^3



Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	39.021	9.933	18.063	59.978
b2	4.190	3.820	-3.868	12.249
b3	-.462	.399	-1.303	.379
b4	.011	.012	-.014	.036

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	35214.304	4	8803.576
Residual	1506.281	17	88.605
Uncorrected Total	36720.585	21	
Corrected Total	3094.477	20	

Dependent variable: Rate1000km

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .513.

$y = 39.021 + 4.190 * Age – 0.462 * Age^2 + 0.011 * Age^3$

<u>>21yrs and <=40yrs:</u>

Modelled mean rate (b/1000km) at 21yrs age

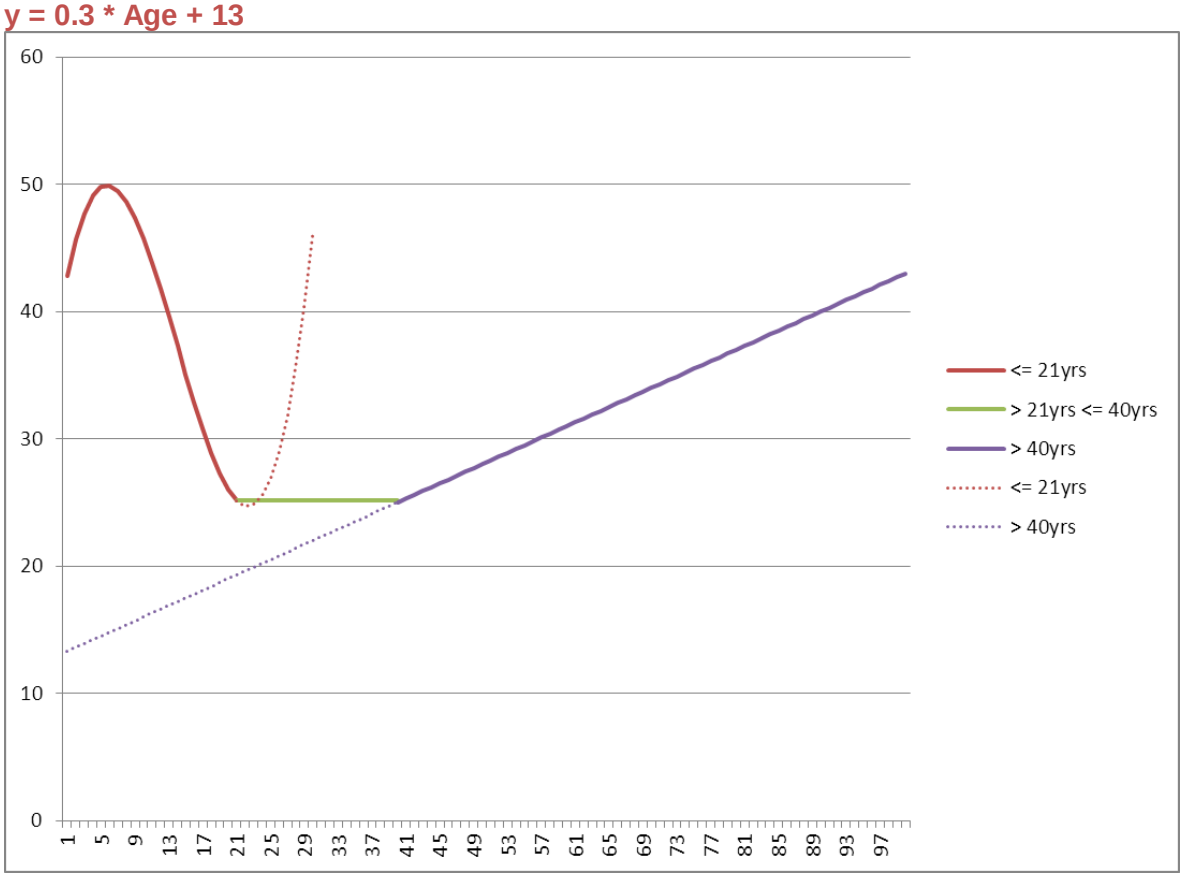
y = 25.14

Note: this number has been validated by measuring the mean of the lower cluster (13-21yrs). The mean works out at 29.72. The more conservative of these has been taken for the model – i.e. the modelled rate of 25.14, because a) it's based on more data and b) the residuals for the mean approach are relatively large giving 95% confidence limits of 7.55 – 51.9

<u>>40yrs:</u>

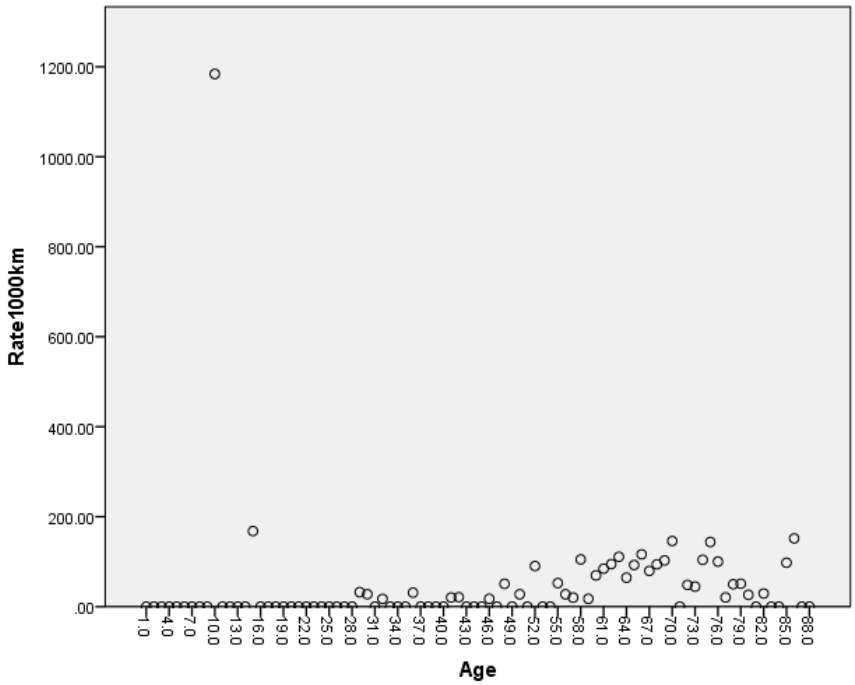
Designed to deteriorate linearly at same rate as HPPE, crossing over previous model at 40yrs age:

y = 0.3 * Age + 13

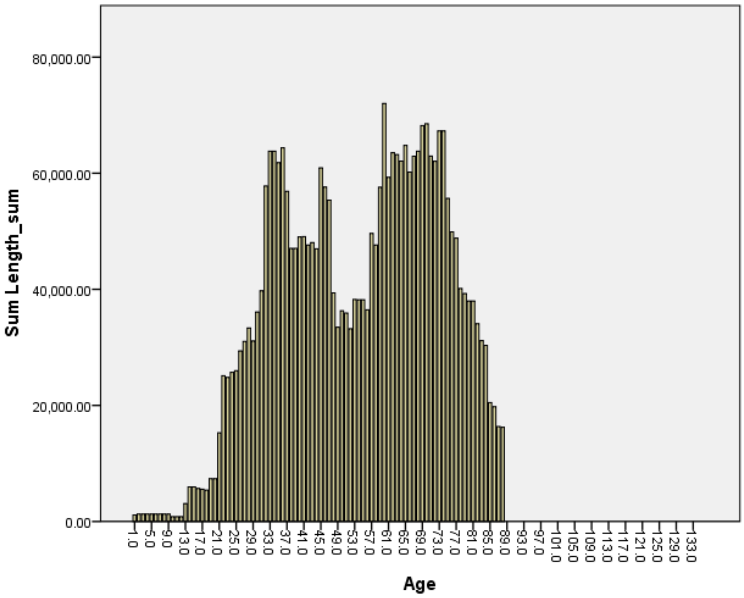


6.1.5 Steel

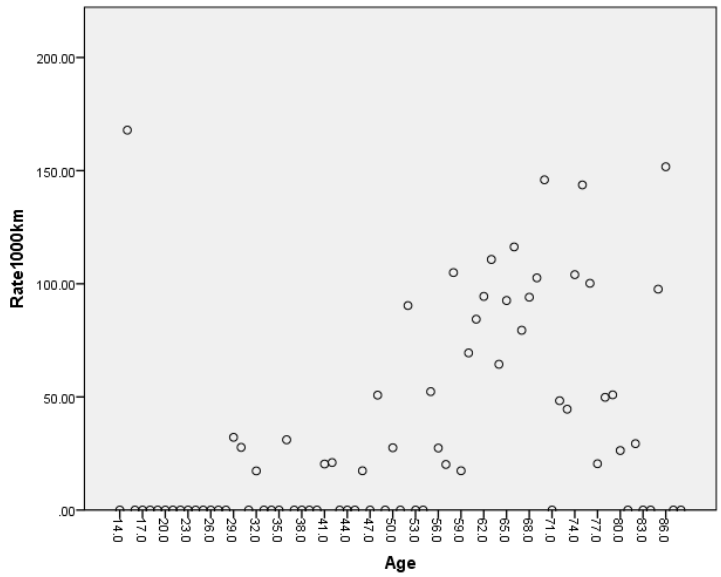
There is a slight increase in failure rate as steel ages, although the data seems to show inconsistent performance:



This could be due to a strange distribution of length across the age bands:

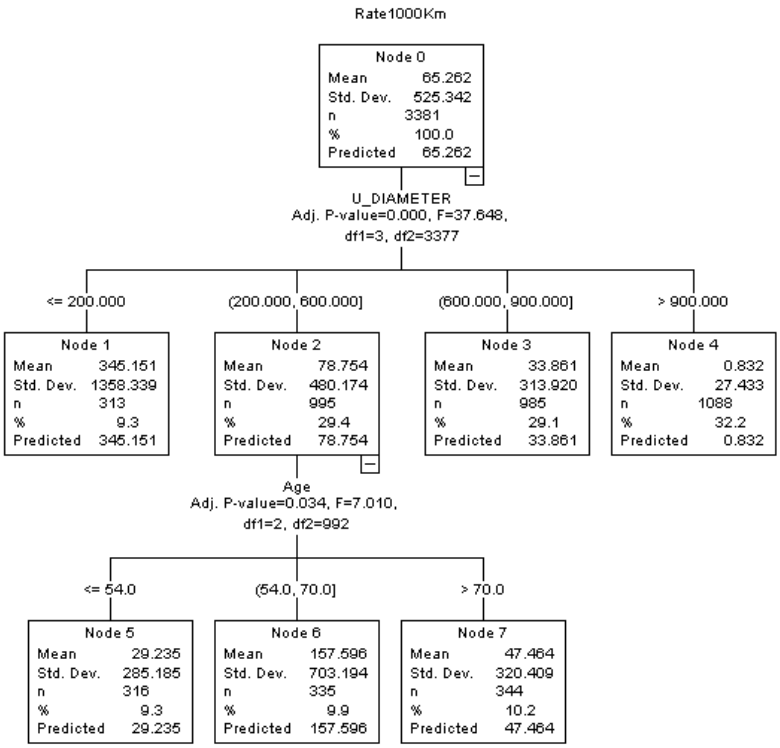


Removing outlying ages with exposure < 5km gives the following chart:

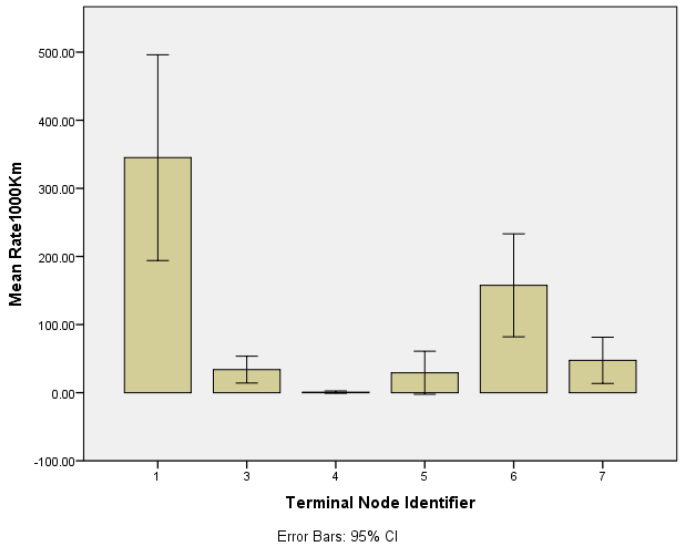


Review of other predictors

Regression trees indicate that it is appropriate to break the Steel assets into sub-groups, which is understandable considering the large range of diameters in use, which as expected, have vastly different failure rates:



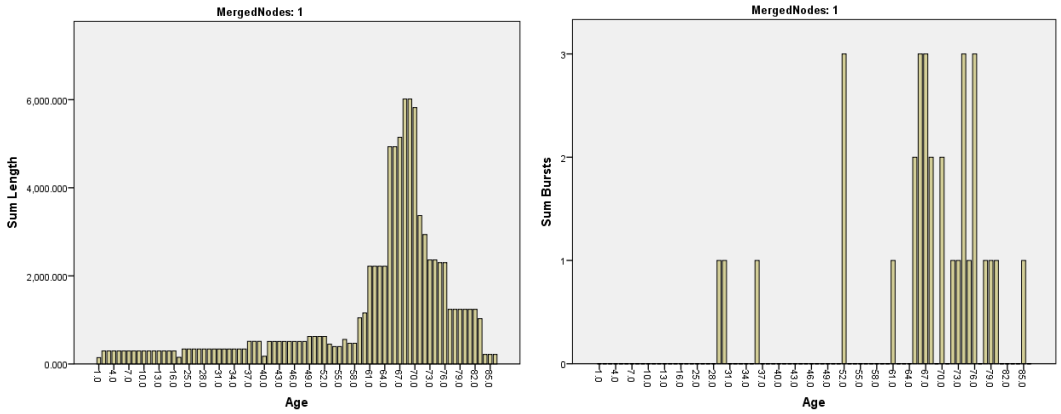
Mean burst rates for the above terminal nodes are as follows:



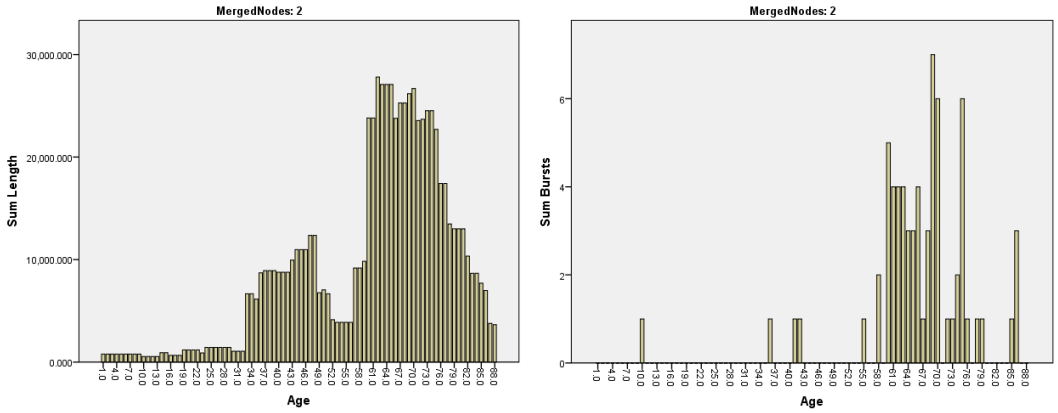
'Stand out' cases seem to be node 1 (<=200mm) and node 6 (>200mm, <=600mm and aged 55 to 70yrs).

It seems reasonable to take node 1 as a separate model because of the disparate performance. The remaining nodes are large diameter, but as predicted by the tree, there is some evidence of age related deterioration.

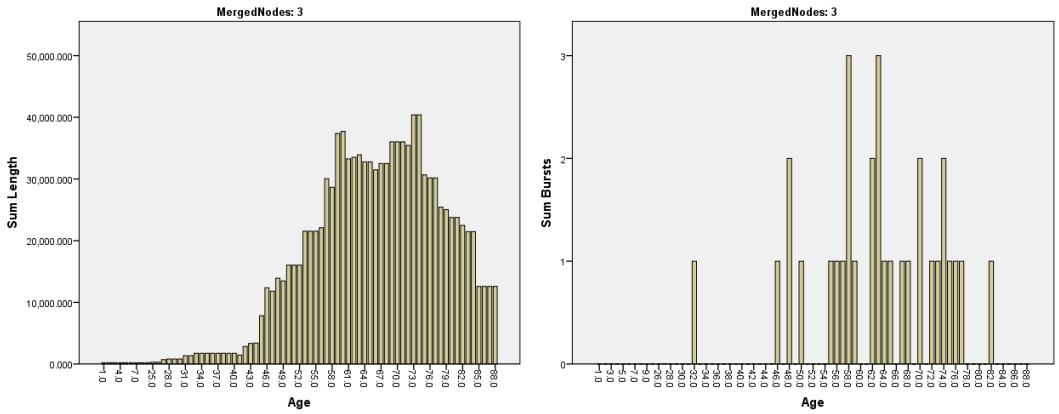
Node 1:



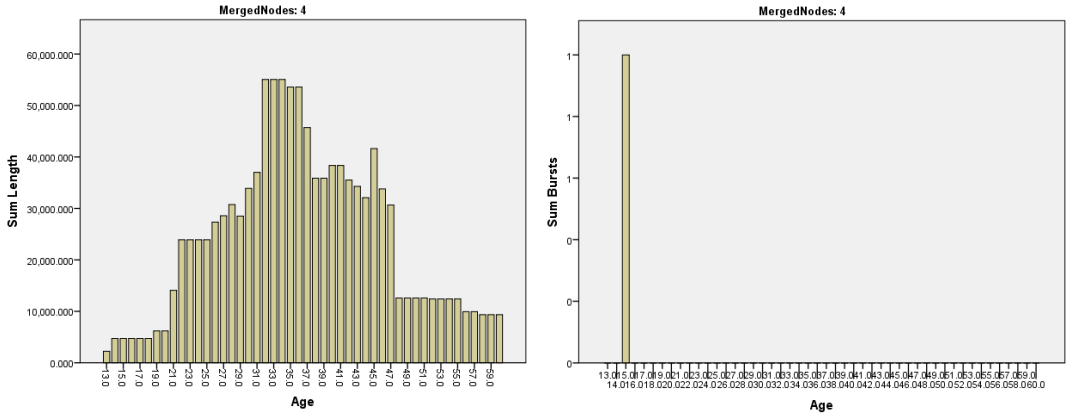
Node 2:



Node 3:



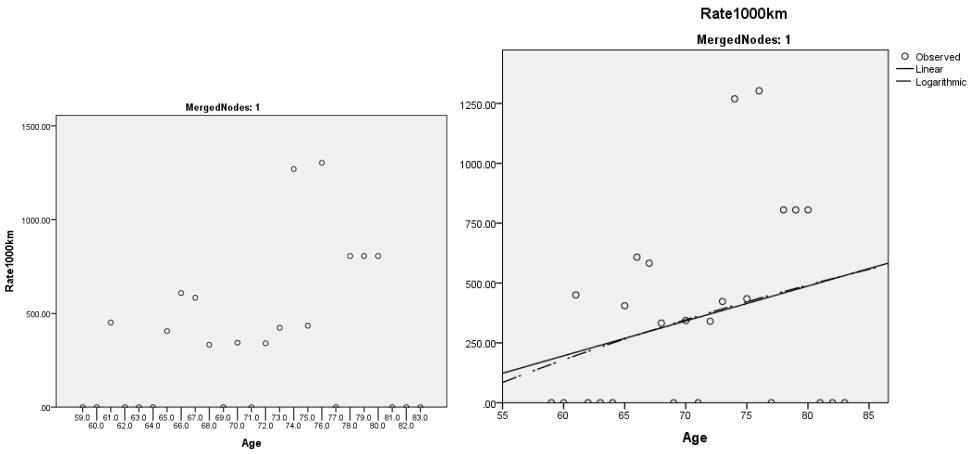
Node 4:



Above charts demonstrate lack of failure data for some of the Steel groups. There are very few recorded failures for groups 3 and 4, whilst there are more failures for groups 1 and 2, but it's still a relatively small group of assets compared to some of the other materials. The next step, considered reasonable due to the small number of annual bursts on steel, is to group into 3 diameter bands for modelling (Node1, Node2 and Node3/4). Outliers and low exposure ages can then be removed and it should be possible to estimate an approximate age based model for at least one of these groups.

Trimmed model groups are as follows, note disparate burst rates (y-scale) for each group:

Node1 (<=200mm):



Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

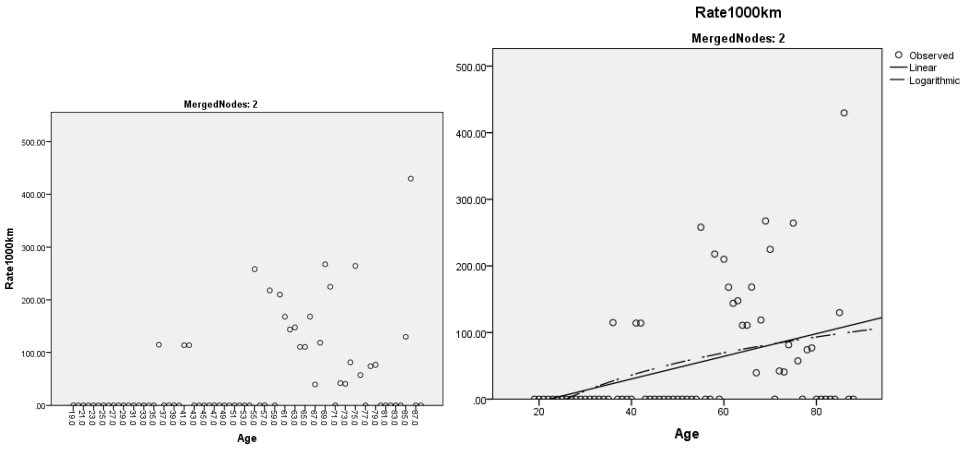
Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.071	1.765	1	23	.197	-677.805	14.567
Logarithmic	.080	1.994	1	23	.171	-4269.128	1086.462
Growth[b]	.	.	.	.	.	.	.

The independent variable is Age.

a. MergedNodes = 1

b. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Node2 (>200mm and <=600mm):



Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

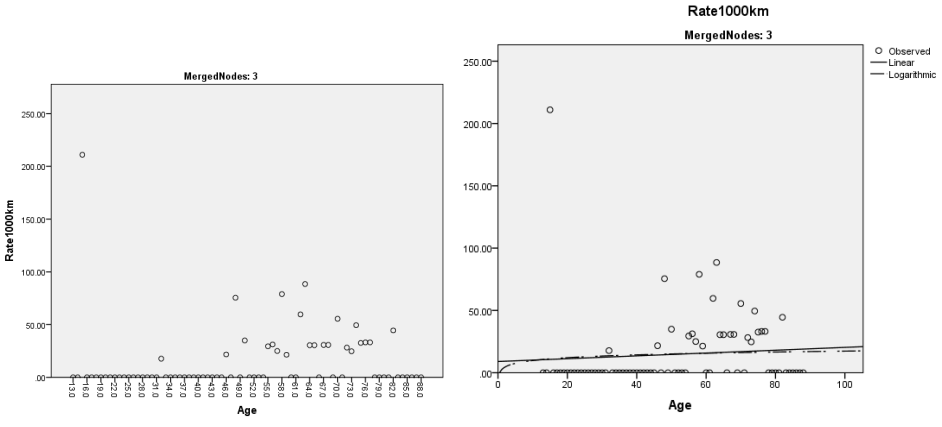
Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.148	11.846	1	68	.001	-37.721	1.700
Logarithmic	.155	12.509	1	68	.001	-268.435	82.548
Growth[b]	.	.	.	.	.	.	.

The independent variable is Age.

a. MergedNodes = 2

b. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Node3 (>600mm):



Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.007	.485	1	74	.489	8.999	.113
Logarithmic	.003	.223	1	74	.638	2.249	3.278
Growth[b]	.	.	.	.	.	.	.

The independent variable is Age.

a. MergedNodes = 3

b. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Final 3 x Models for Steel:

Node1 (<=200mm) > 50yrs:

y = a * log(Age) + c

y = 1086.462 * log(Age) – 4269.128

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
ln(Age)	1086.462	769.358	.282	1.412	.171
(Constant)	-4269.128	3276.468		-1.303	.205

a. MergedNodes = 1

Node2 (>200mm and <=600mm) > 30yrs:

y = a * Age + c

y = 1.7 * Age – 37.721

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age	1.700	.494	.385	3.442	.001
(Constant)	-37.721	28.246		-1.335	.186

a. MergedNodes = 2

Node3 (>600mm):

y = a * Age + c

y = 0.113 * Age + 8.999

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age	.113	.163	.081	.696	.489
(Constant)	8.999	8.965		1.004	.319

a. MergedNodes = 3

6.1.6 'Other' materials

This group of assets contains the least commonly occurring material groups.

Despite grouping all of these together, there are still too few failures to enable any kind of sensible modelling:

		Length	Bursts
		Sum	Sum
U_MATERIAL	BLACK POLY	23171.961	1
	CONCRETE	24369.895	0
	COPPER	57850.364	0
	GI	2401022.594	5
	GRP	11589.172	0
	HDP	990931.467	10
	LEAD	135199.611	0

Just 16 failures on 3,644km of pipes

The most appropriate model for this group is to take the mean failure rate:
y = 4.39 b/1000km

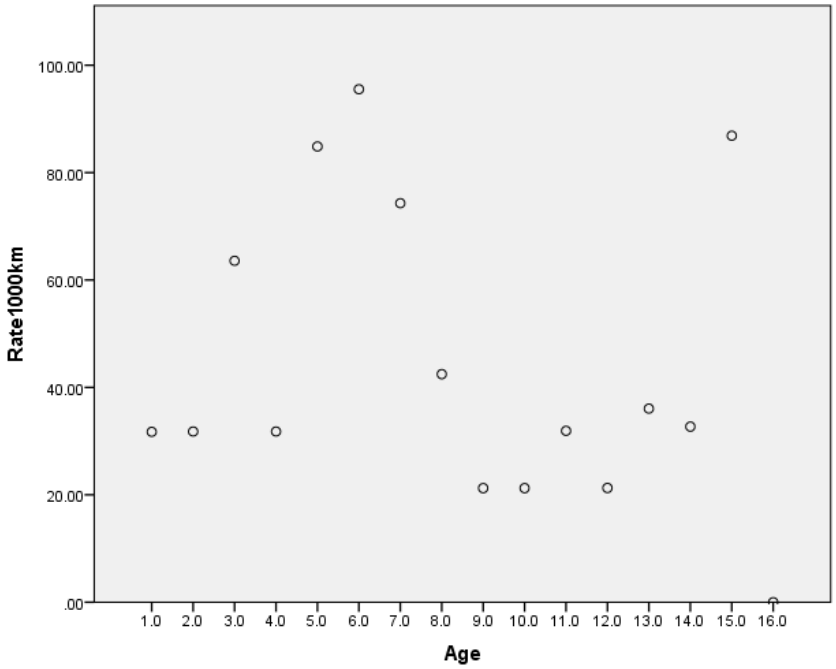
6.1.7 mPVC

As discussed in the preliminary analysis section - this is another small group of assets, with just 59 recorded failures against 1,300km of pipe:

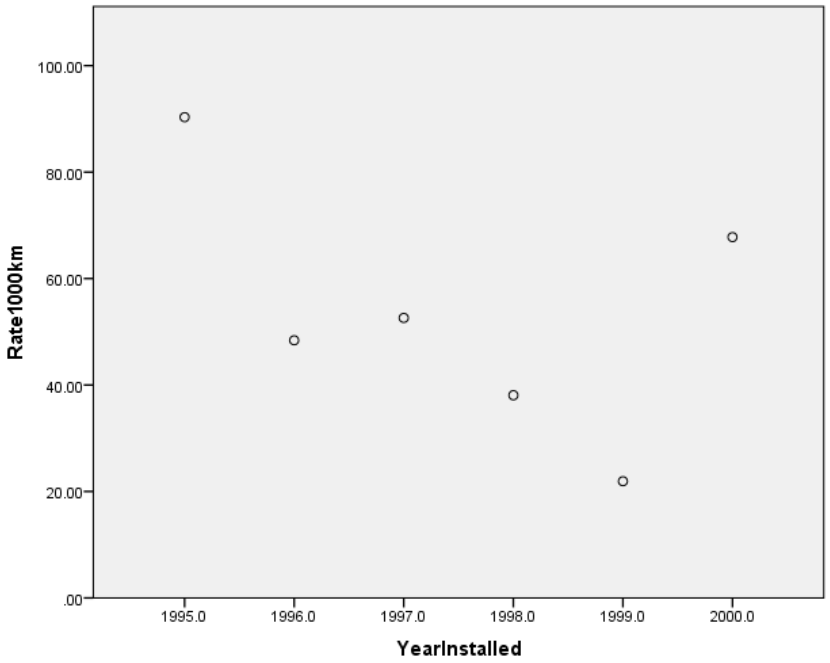
		Bursts	Length
		Sum	Sum
U_MATERIAL	mPVC	59	1300171.968

Initial investigations using decision trees doesn't show any significant predictors.

Aggregating by age and measuring the failure rate shows no clear pattern:

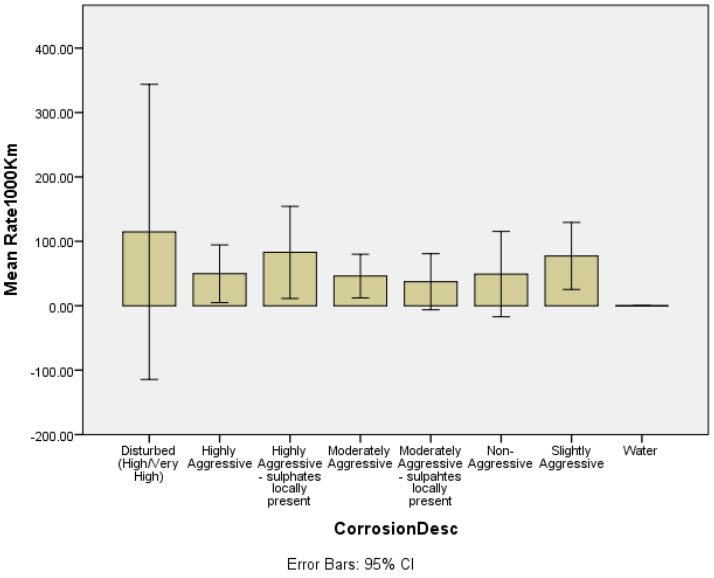


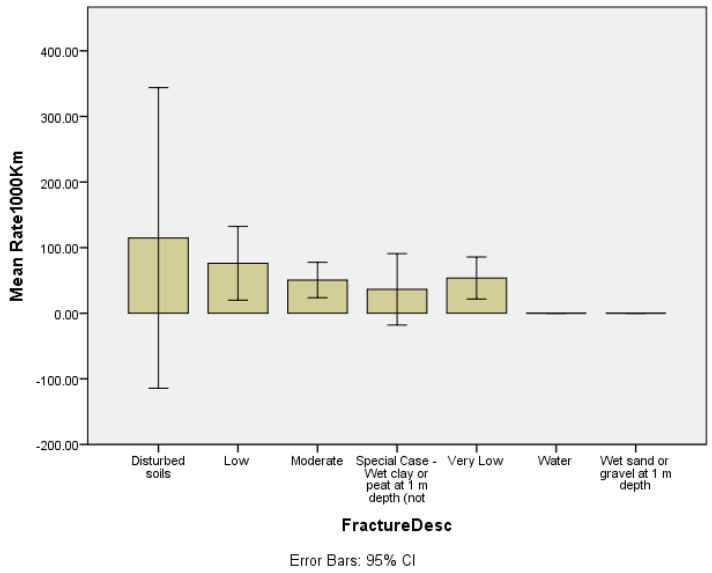
Analysis by year installed reveals similar disparity:

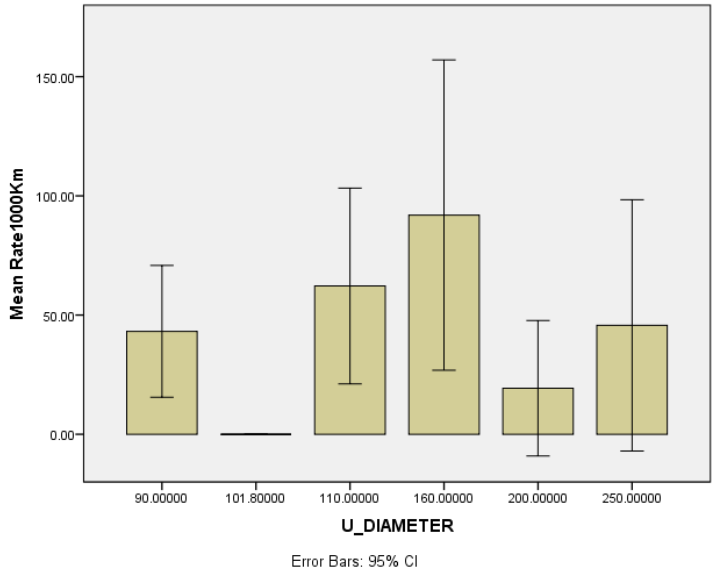


Note that mPVC was only installed during a 6 year period from 1995 to 2000.

Graphs of the other predictors reinforce the regression tree, each showing relatively low failure rates, similar means and large confidence intervals:







Final Model for mPVC

The most appropriate final model for mPVC is a simple mean based on the limited number of failures experienced to date:

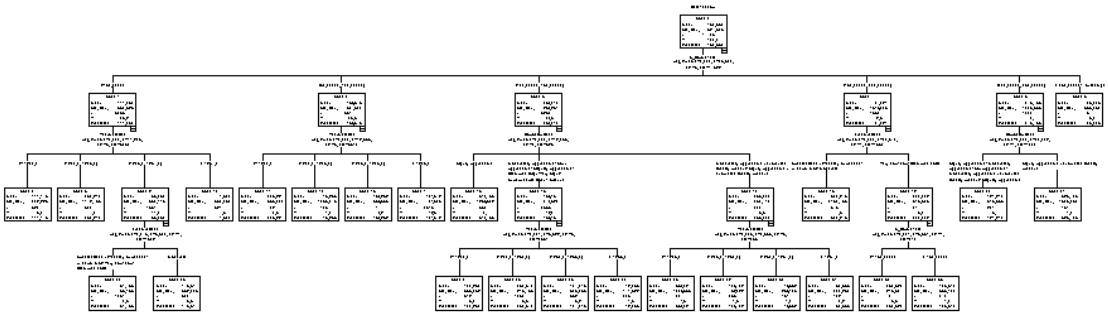
y = 45.379 b/1000km

6.1.8 PVC

There is a more significant amount of data to analyse for PVC:

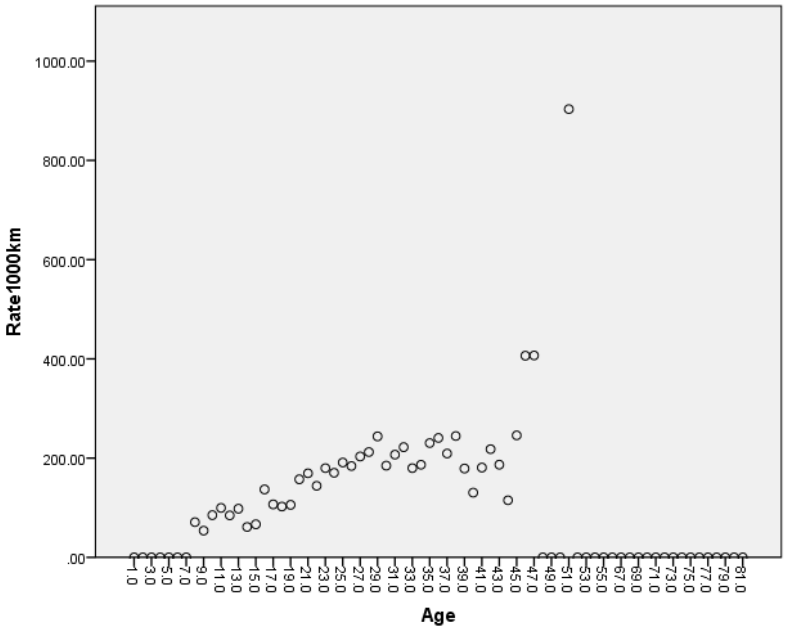
Bursts	Length
Sum	Sum
1894	11192913.22

Initial investigation using regression trees shows significant relationships between burst rate and diameter, corrositivity, fracture potential and year installed.

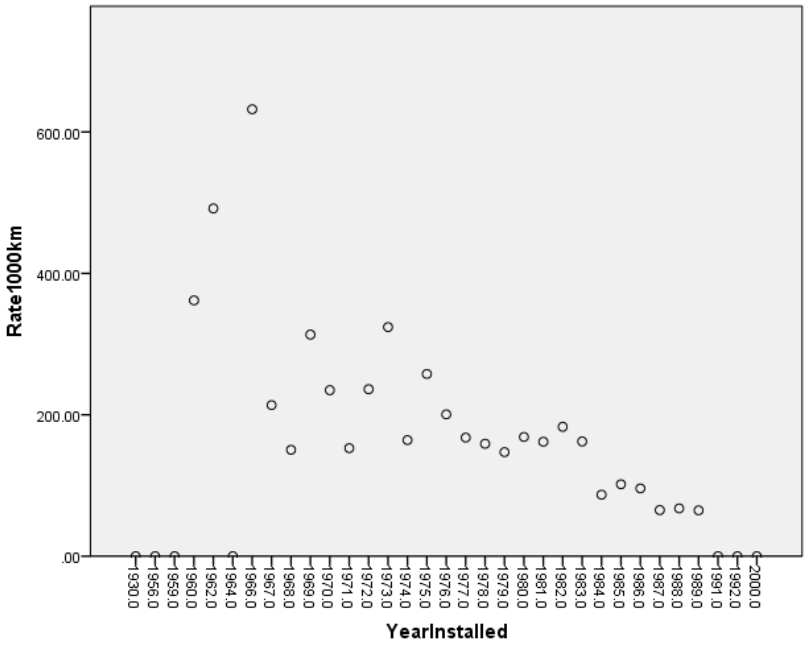


The above tree is too complicated to build into final models and is likely to be over-fitted, requiring further analysis.

Aggregating by age also reveals a relationship with burst rate:



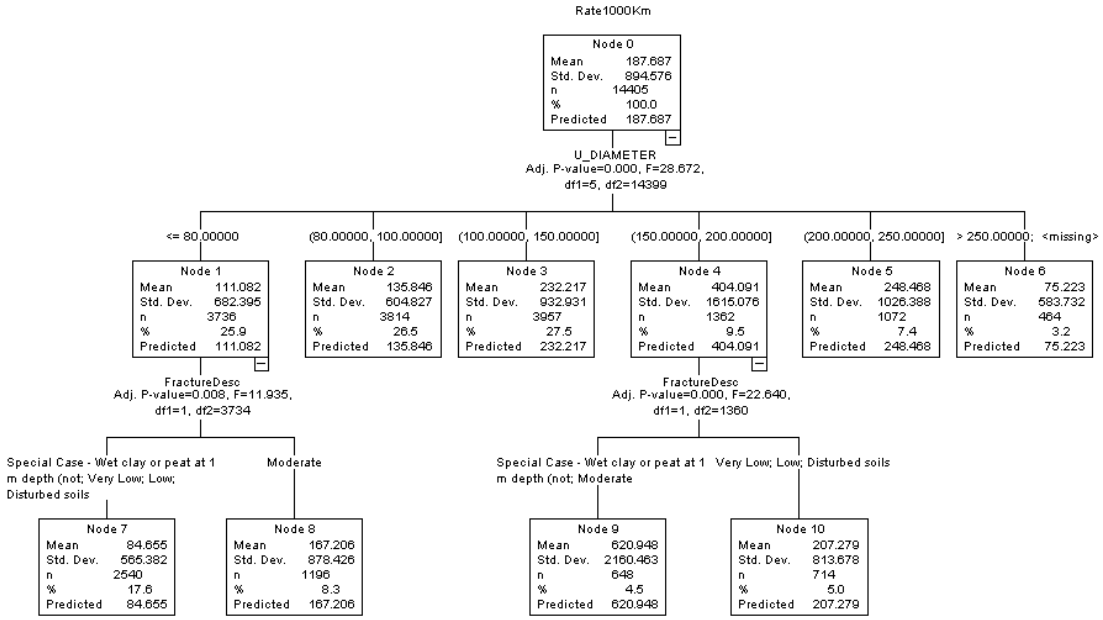
The year installed shows a reverse of the above, because all PVC was installed before the beginning of the bursts observation window (1996 to 2012), the two predictors are consequently inversely proportional/colinear:



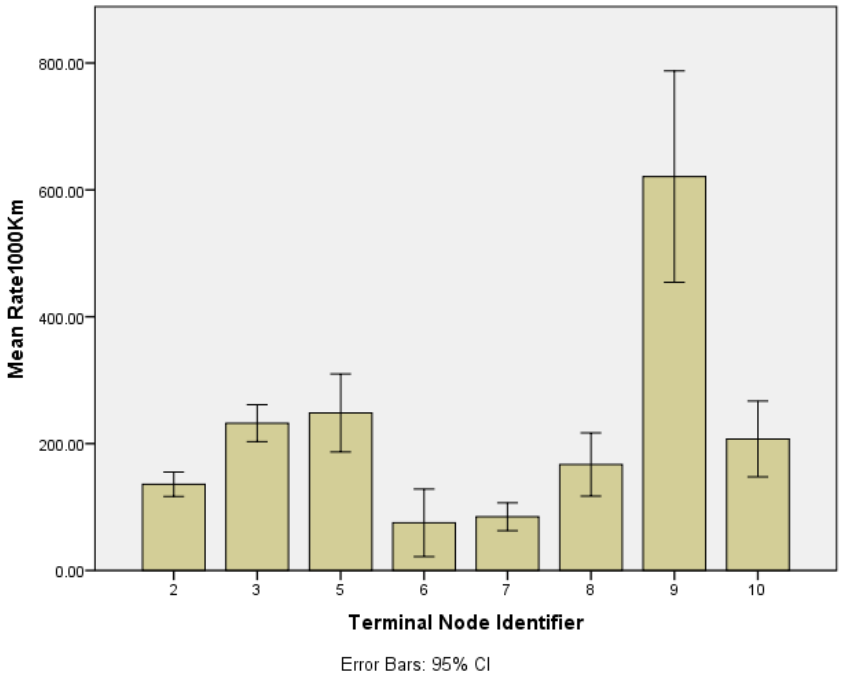
Linear regression is likely to be inaccurate because the assumption of normal data is, in this case, violated. However, stepwise regression techniques have still been used to determine order of significance for predictors:

Fracture Desc – 0.40 significance
Age – 0.32 significance
Diameter – 0.28

Re-executing the regression tree with a more limited set of predictors gives a more manageable number of terminal nodes:



It also appears that the predictors used are significant in their ability to affect the burst rate. Individually scoring the mean burst rate and 95% confidence limits for each terminal node gives the following output:



Node 9 seems a particular standout case, yet still accounts for over 6.5% of the 1894 observed failures (i.e. it's not an exceptionally low exposure outlier).

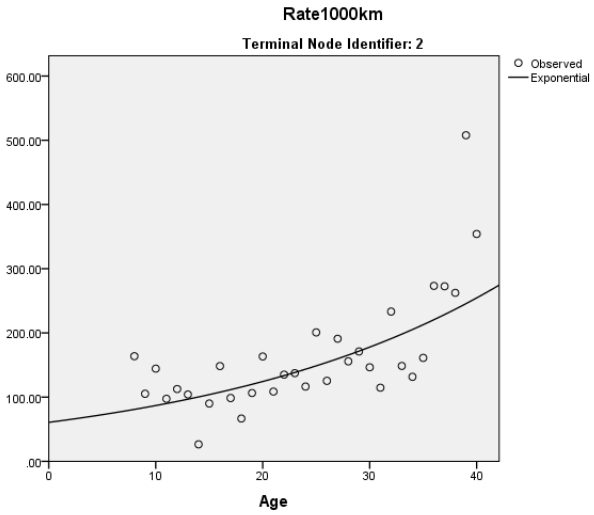
Node 2: >80mm and <=100mm
Node 3: >100mm and <=150mm
Node 5: >200mm and <=250mm
Node 6: >250mm
Node 7: <=80mm in anything other than 'Moderate' fracture
Node 8: <=80mm in 'Moderate' fracture
Node 9: >150mm and <=200mm in other than 'Very Low','Low' or 'Disturbed' fracture
Node 10: >150mm and <=200mm in 'Very Low','Low' or 'Disturbed' fracture

		Bursts_sum	Length_sum
		Sum	Sum
Terminal Node Identifier	2	590.00	4236938.97
	3	680.00	3199617.04
	5	129.00	514450.95
	6	13.00	120887.48
	7	121.00	1470312.92
	8	128.00	835163.27
	9	124.00	263649.46
	10	85.00	394242.26

Nodes 1 & 2 are likely to be the most consistent/reliable models, due to higher quantity of data.

The next stage is to review the age related pattern for each of the 8 terminal nodes. If any nodes demonstrate similar patterns and means, then there may be scope to prune the models, merging nodes back together;

Node 2: >80mm and <=100mm
This node shows a good pattern and an exponential curve was a good fit for the observed points:



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.667	.445	.427	.393

The independent variable is Age.

a. Terminal Node Identifier = 2

Coefficients[a]

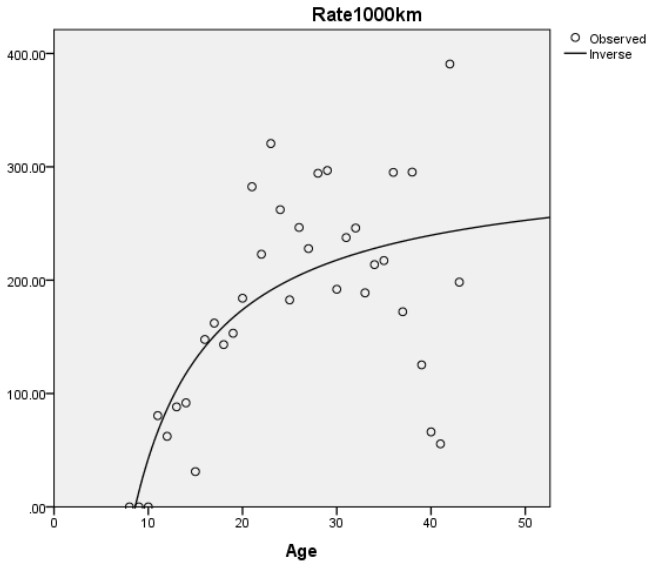
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age	.036	.007	.667	4.987	.000
(Constant)	60.756	11.264		5.394	.000

The dependent variable is ln(Rate1000km).

a. Terminal Node Identifier = 2

Node 3: >100mm and <=150mm

This node shows a strange age relationship, although an inverse curve provides a fairly robust fit for the earlier years. Excessive variance is a problem towards the (lower exposure) right hand side of the chart:



Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.696	.484	.469	72.678

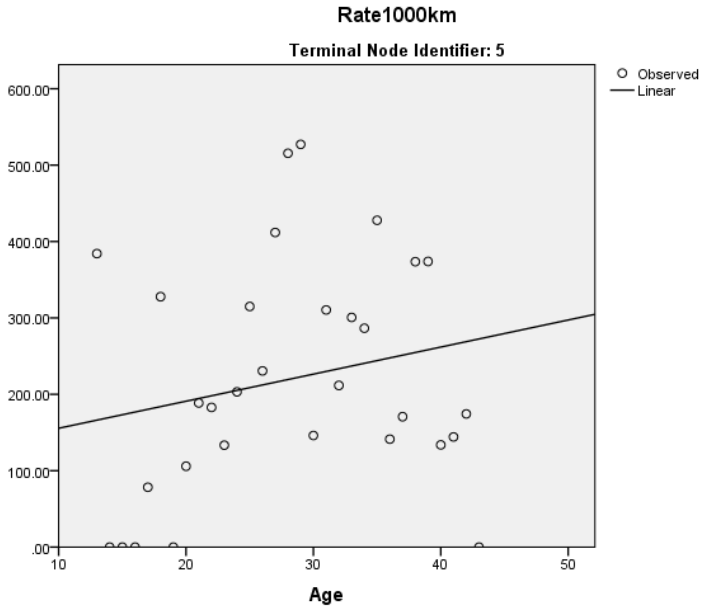
The independent variable is Age.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
1 / Age	-2627.690	465.423	-.696	-5.646	.000
(Constant)	305.313	25.745		11.859	.000

Node 5: >200mm and <=250mm

This node contains a relatively small amount of the PVC length and consequently shows a high amount of scatter. However, the plot does seem to indicate an increasing trend; so despite a relatively poor fit, a linear model has been selected:



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.212	.045	.012	150.772

The independent variable is Age.

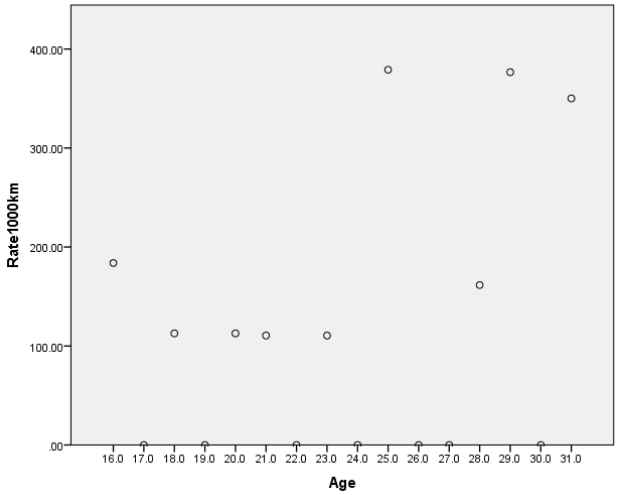
a. Terminal Node Identifier = 5

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Age	3.543	3.028	.212	1.170	.251
(Constant)	120.123	88.992		1.350	.188

Node 6: >250mm

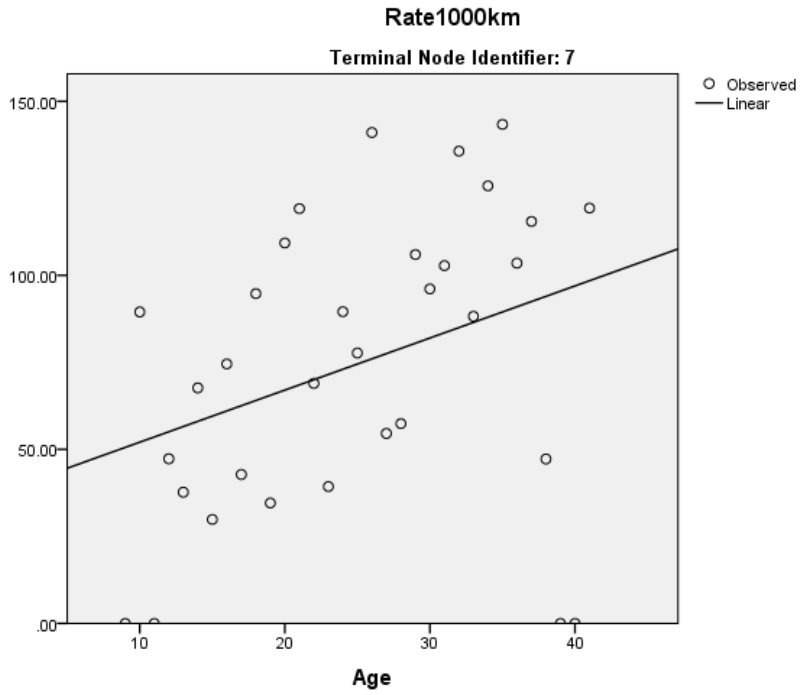
Again, this node contains only a small amount of the PVC data.



The amount of data is insufficient to draw conclusions, therefore a mean of 107.54 b/km will be used for this model.

Node 7: <=80mm in anything other than 'Moderate' fracture

Node 7 shows a high amount of scatter, but also shows a reliable increasing trend in failures. There are some low scores on the right hand side of the chart that seem to overly influence the fitted line, but any ages with particularly low exposure have already been addressed. The apparently conservative linear trend has therefore been kept.



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.339	.115	.087	40.768

The independent variable is Age.

a. Terminal Node Identifier = 7

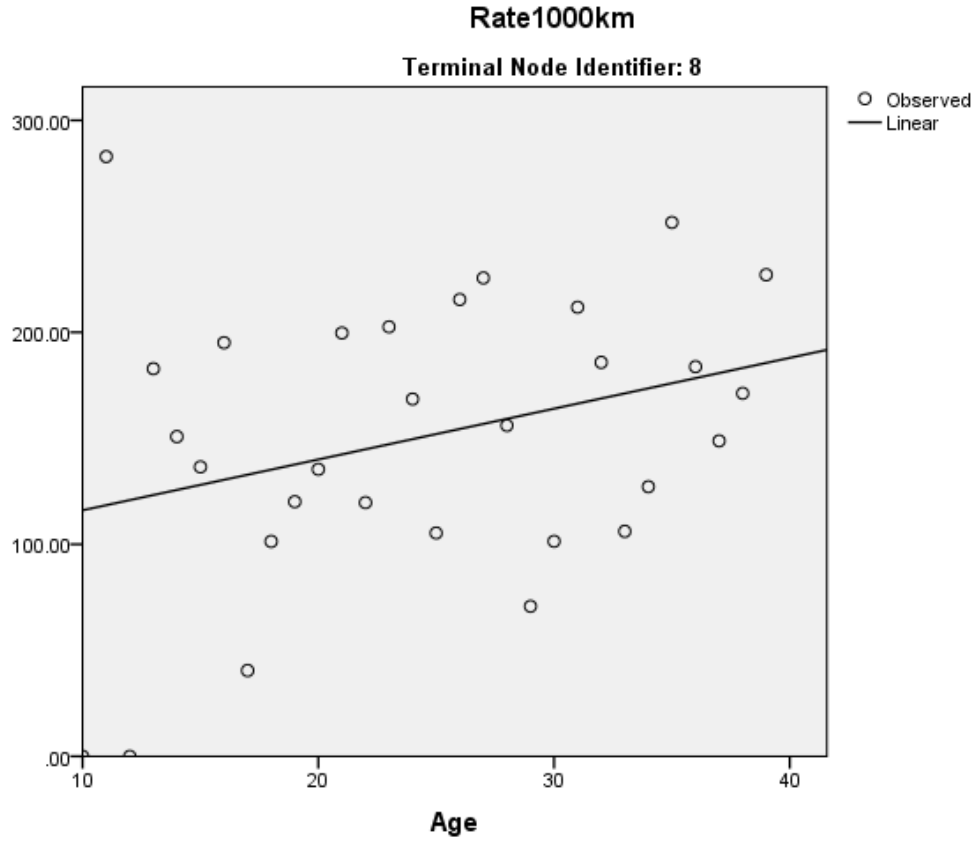
Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Age	1.496	.745	.339	2.008	.053
(Constant)	37.103	19.938		1.861	.072

a. Terminal Node Identifier = 7

Node 8: <=80mm in 'Moderate' fracture

As above, node 8 also shows a lot of scatter, but a clear pattern. A linear model has also been selected:



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.310	.096	.064	65.777

The independent variable is Age.

a. Terminal Node Identifier = 8

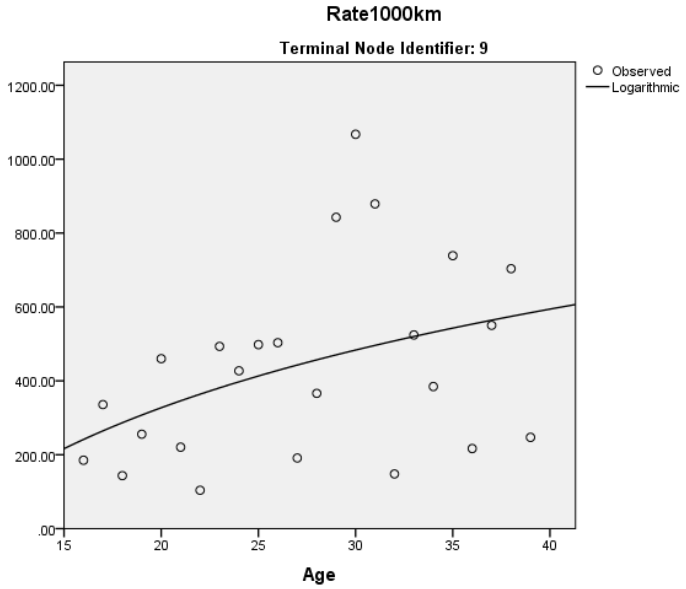
Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Age	2.397	1.387	.310	1.727	.095
(Constant)	92.087	36.052		2.554	.016

a. Terminal Node Identifier = 8

Node 9: >150mm and <=200mm in other than 'Very Low','Low' or 'Disturbed' fracture

This node contains a small amount of length and therefore generates a lower confidence model. However, there does seem to be an underlying increasing trend within the data, so a conservative approach has been adopted by fitting a levelling logarithmic curve:



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.402	.161	.123	242.218

The independent variable is Age.

a. Terminal Node Identifier = 9

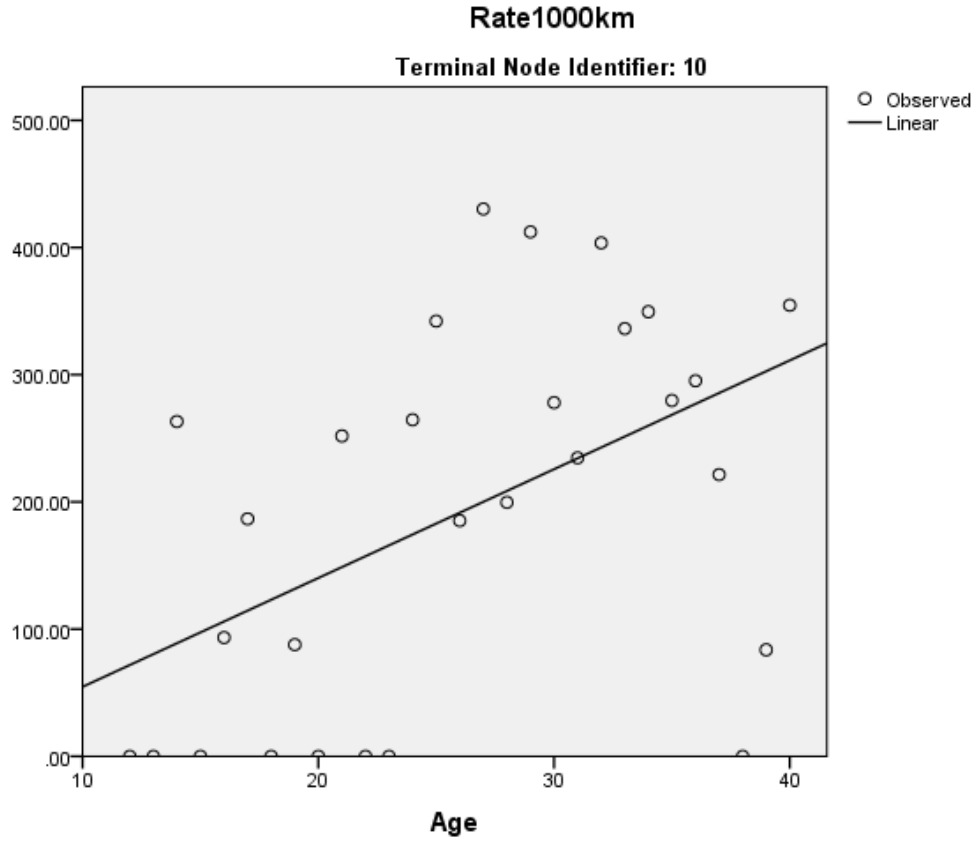
Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
ln(Age)	385.174	187.297	.402	2.056	.052
(Constant)	-826.675	616.417		-1.341	.194

a. Terminal Node Identifier = 9

Node 10: >150mm and <=200mm in 'Very Low','Low' or 'Disturbed' fracture

Similar to node 9, this is also a low exposure node. A linear model has been selected and shows a reasonably good fit to the data points:



Model Summary[a]

R	R Square	Adjusted R Square	Std. Error of the Estimate
.492	.242	.214	131.350

The independent variable is Age.

a. Terminal Node Identifier = 10

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Age	8.554	2.915	.492	2.934	.007
(Constant)	-30.899	79.625		-.388	.701

a. Terminal Node Identifier = 10

Summary of above models

Some of the above models are similar means, but different shaped graphs. It's therefore not possible to reduce the number of models by re-combining nodes.

Also, some of the above curves are not particularly good fits to the observed data, and branching the data bin into 8 separate nodes does arguably sub-divide the data too far. However, outliers have only been removed where the exposure is less than 5km, so the basic shapes of the scatter plots haven't been artificially affected too much. Also the regression tree does calculate the loss of model stability in addition to the increase in model accuracy when deciding where to sub-divide/merge nodes.

It should also be noted that the analysis above does reflect SSW's experience of managing PVC assets; whereby the performance of individual lengths of pipe can be unpredictable, with notably high 'in-cohort' disparity.

Final 8 x Models for PVC

Node 2: >80mm and <=100mm
$y = a * e^{(b * Age)}$
Where a = 60.756 and b = 0.036
$y = 60.756 * e^{(0.036 * Age)}$

Node 3: >100mm and <=150mm
y = a + b / Age
Where a = 305.313 and b = -2627.690
y = 305.313 – 2627.690 / Age

Node 5: >200mm and <=250mm
y = a * Age + b
Where a = 3.543 and b = 120.123
y = 3.543 * Age + 120.123

Node 6: >250mm
y = c
Where c = mean rate = 107.54
y = 107.54

Node 7: <=80mm in anything other than 'Moderate' fracture
y = a * Age + b
Where a = 1.496 and b = 37.103
y = 1.496 * Age + 37.103

Node 8: <=80mm in 'Moderate' fracture
y = a * Age + b
Where a = 2.397 and b = 92.087
y = 2.397 * Age + 92.087

Node 9: >150mm and <=200mm in other than 'Very Low','Low' or 'Disturbed' fracture
y = a + b * log(Age)
Where a = -826.675 and b = 385.174
y = -826.675 + 385.174 * log(Age)

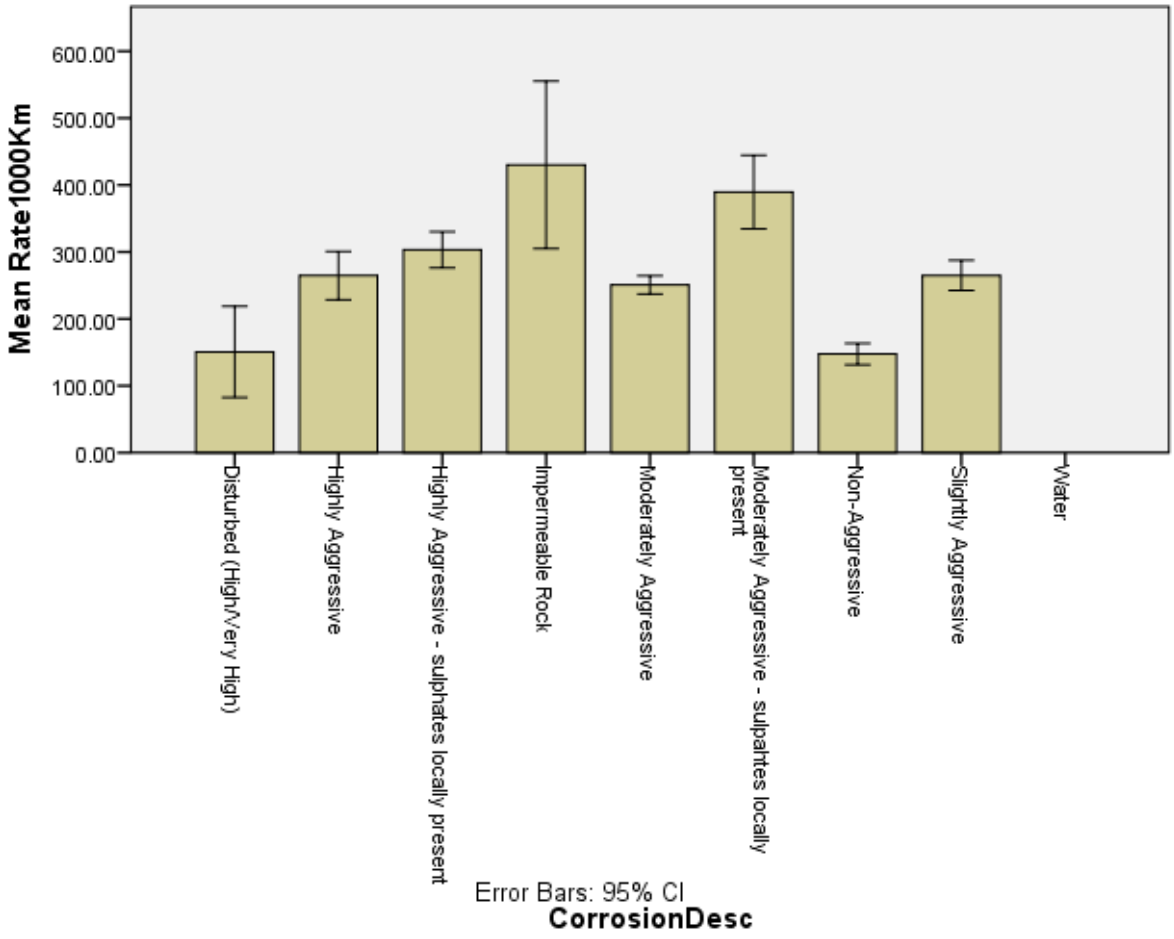
Node 10: >150mm and <=200mm in 'Very Low','Low' or 'Disturbed' fracture
y = a * Age + b
Where a = 8.554 and b = -30.899
y = 8.554 * Age – 30.899

6.1.9 CI

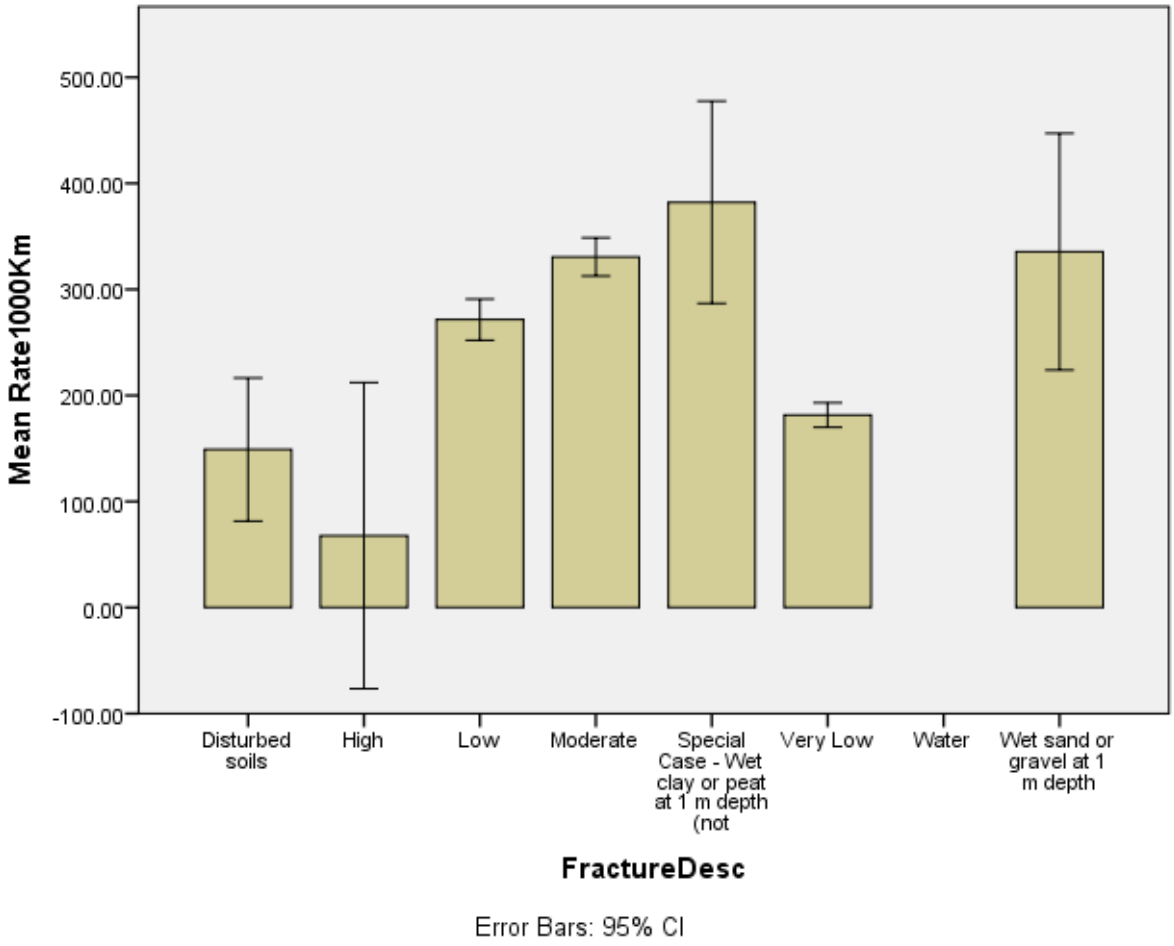
Cast Iron is the largest group within this analysis project, having the highest burst rate and the longest length. As assessed earlier, there are also complications introduced by the fact that there are clear era's of CI, where manufacturing techniques and material composition was changed, resulting in differing performance and reliability.

The following sections show some initial analysis of the predictor variables, assessing their significance against the failure rate:

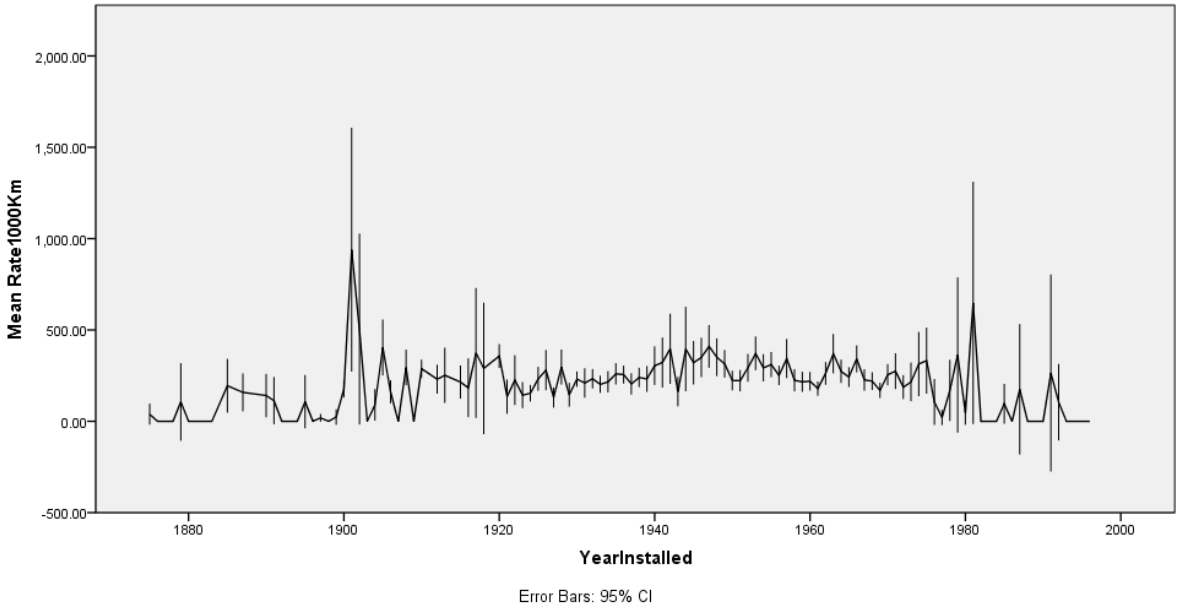
CI – By Corrosion



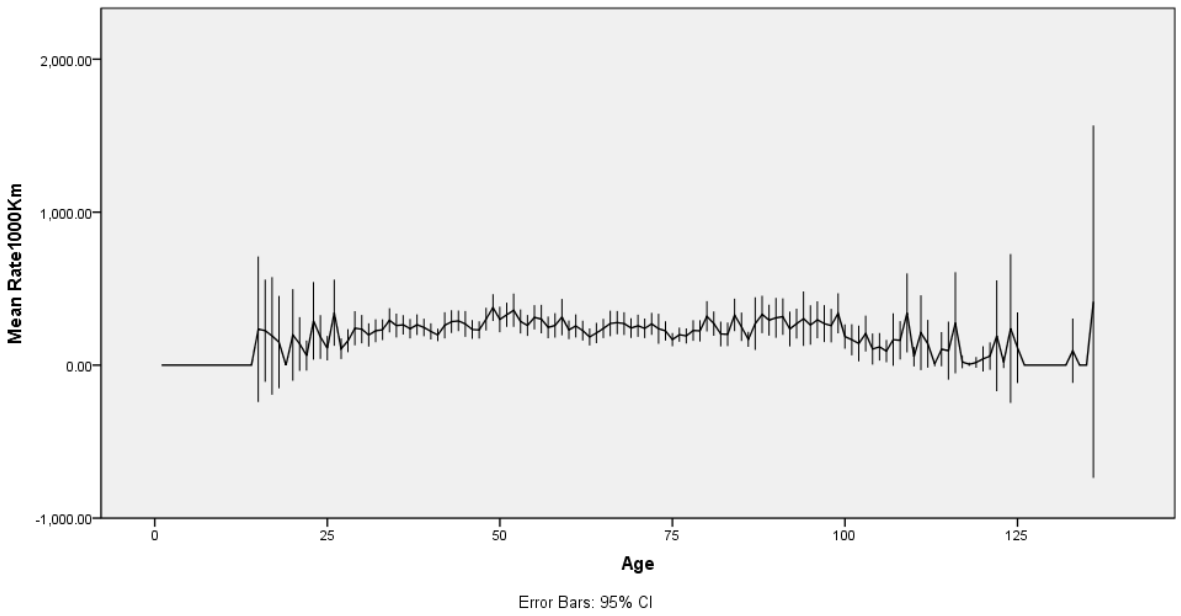
CI – By Fracture



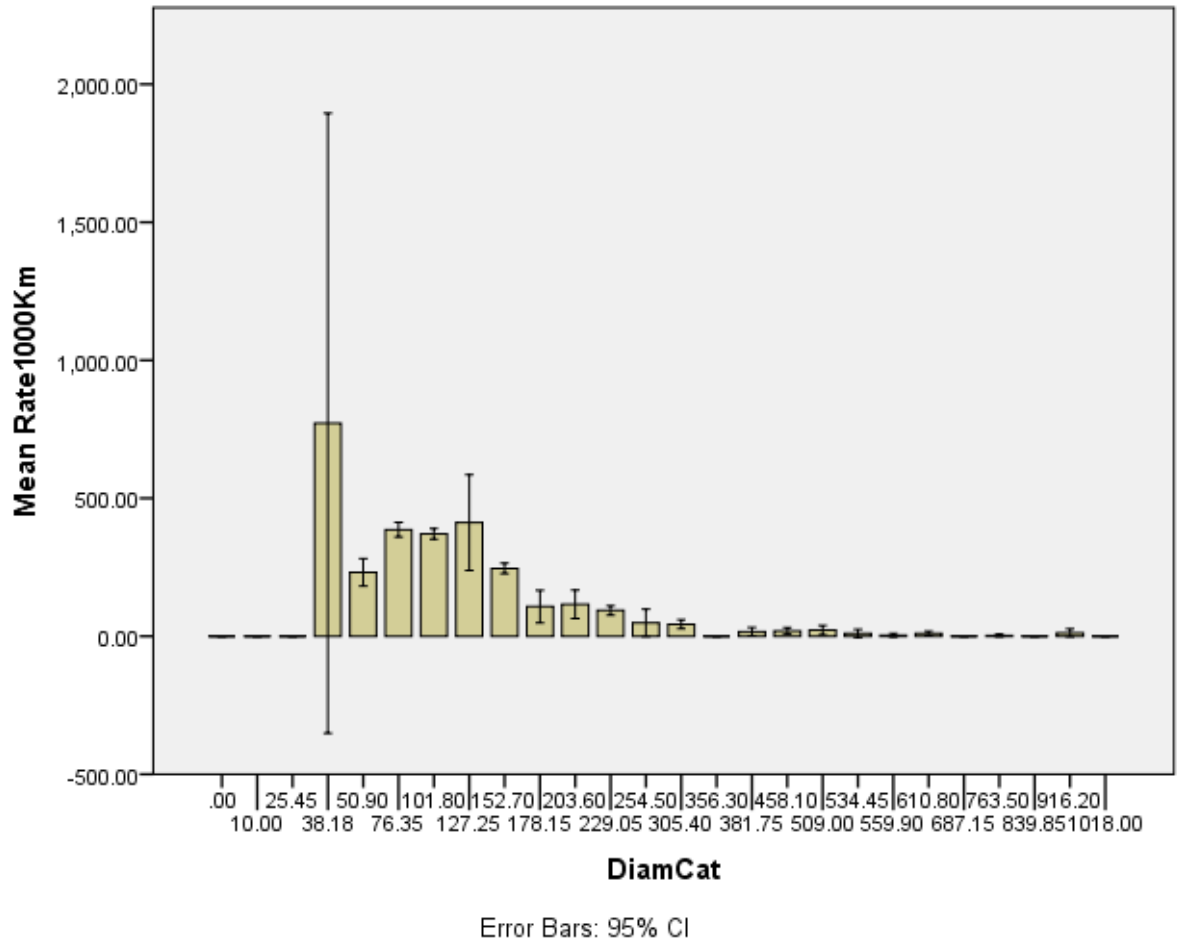
CI – By Year Installed



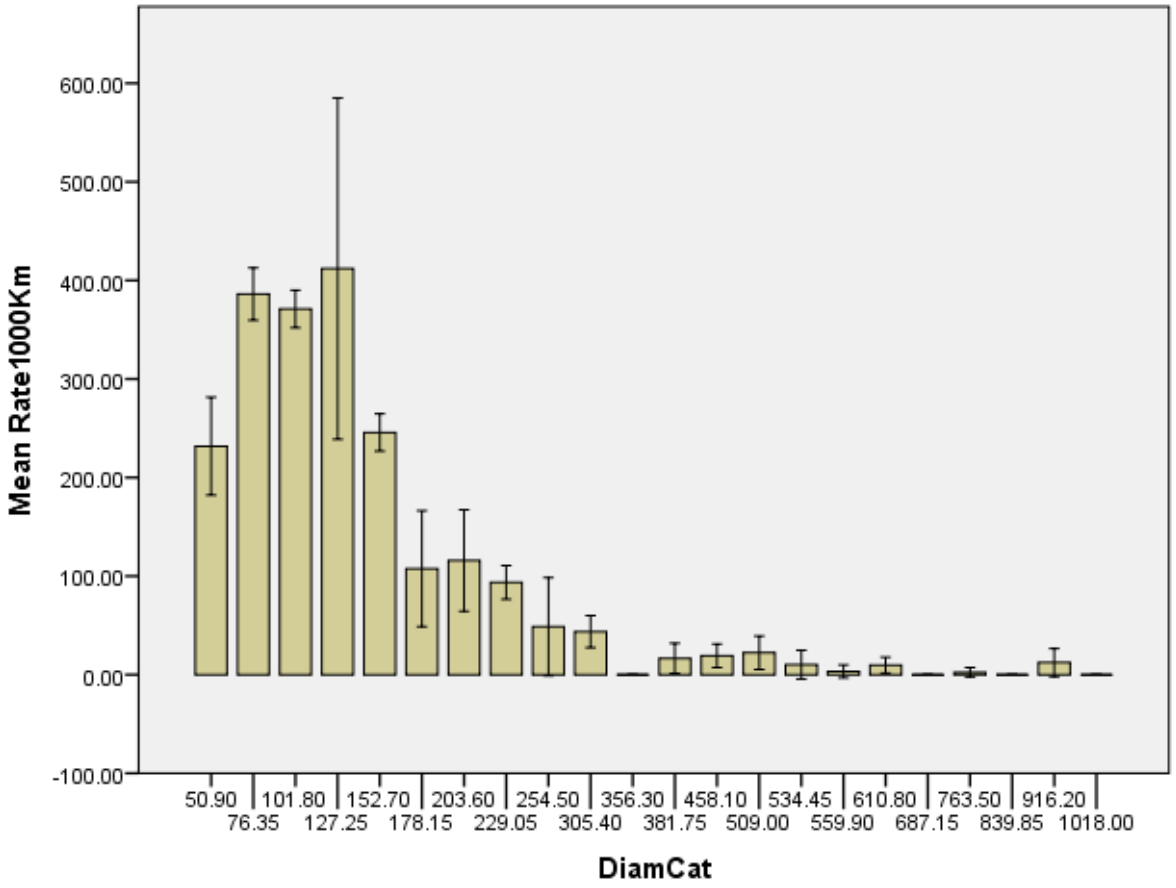
CI – By Age



CI – By Diameter



And again without <50.9:

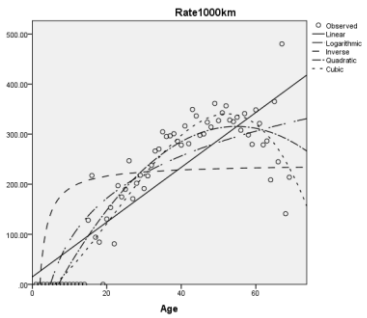


CI – Investigating Era Splits

Preliminary analysis indicated age related anomalies within the cast iron data. This is thought to be a consequence of step changes in manufacturing technique when spun iron was introduced. Efforts have been made to model the deterioration of cast iron holistically and segregated into various 'eras'. The following charts demonstrate some of this analysis, but in summary, the best results seemed to be achieved by splitting into two eras; this balanced over-division of the data with adequate reflection of the differing performance of each era:

1942

Greater than or equal to:



Model Summary and Parameter Estimates

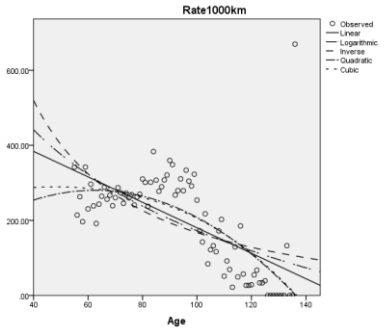
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.699	155.628	1	67	.000	14.886	5.471		
Logarithmic	.713	166.793	1	67	.000	-193.498	121.979		
Inverse	.268	24.551	1	67	.000	240.779	-492.794		
Quadratic	.840	173.767	2	66	.000	-99.542	15.141	-.138	
Cubic	.870	145.436	3	65	.000	-34.460	4.369	.244	-.004
Compound[a]	.	.	.	.	.	.	.	.	
Power[a]	.	.	.	.	.	.	.	.	
S[a]	.	.	.	.	.	.	.	.	
Growth[a]	.	.	.	.	.	.	.	.	
Exponential[a]	.	.	.	.	.	.	.	.	
Logistic[a]	.	.	.	.	.	.	.	.	

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

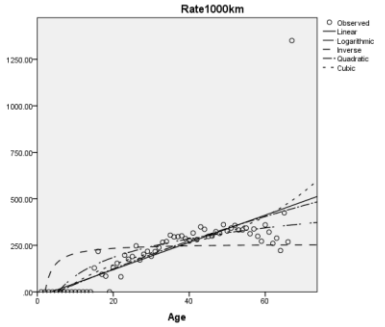
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.396	52.512	1	80	.000	519.494	-3.392		
Logarithmic	.357	44.404	1	80	.000	1524.435	-293.576		
Inverse	.307	35.435	1	80	.000	-67.383	23483.520		
Quadratic	.437	30.685	2	79	.000	78.669	6.444	-.051	
Cubic	.428	29.579	2	79	.000	253.888	1.099	.000	.000
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1944

Greater than or equal to:



Model Summary and Parameter Estimates

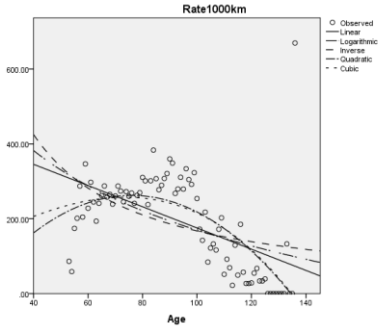
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.556	81.470	1	65	.000	-27.225	7.325		
Logarithmic	.468	57.094	1	65	.000	-246.897	144.228		
Inverse	.156	11.987	1	65	.001	260.461	-540.595		
Quadratic	.558	40.398	2	64	.000	-45.929	8.951	-.024	
Cubic	.567	27.521	3	63	.000	-98.645	17.923	-.351	.003
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

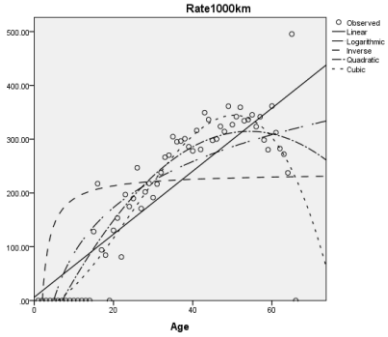
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.295	34.269	1	82	.000	458.781	-2.835		
Logarithmic	.244	26.458	1	82	.000	1237.520	-231.885		
Inverse	.187	18.825	1	82	.000	-4.455	17174.114		
Quadratic	.397	26.624	2	81	.000	-182.491	11.694	-.077	
Cubic	.378	24.577	2	81	.000	63.366	3.958	.000	.000
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1946

Greater than or equal to:



Model Summary and Parameter Estimates

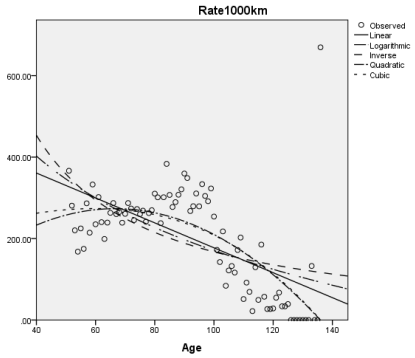
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.683	138.173	1	64	.000	6.063	5.862		
Logarithmic	.676	133.273	1	64	.000	-196.875	123.422		
Inverse	.251	21.403	1	64	.000	237.501	-484.723		
Quadratic	.799	124.920	2	63	.000	-101.215	15.328	-.141	
Cubic	.842	109.914	3	62	.000	-19.942	1.296	.378	-.005
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

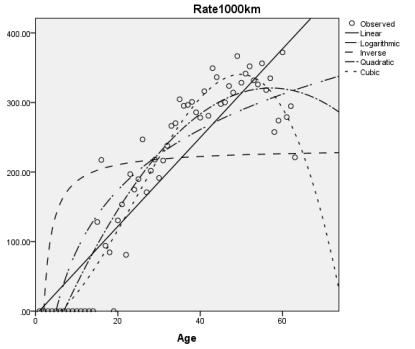
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.367	48.657	1	84	.000	483.382	-3.061		
Logarithmic	.320	39.468	1	84	.000	1336.751	-253.218		
Inverse	.264	30.097	1	84	.000	-23.608	19098.929		
Quadratic	.430	31.247	2	83	.000	20.027	7.602	-.057	
Cubic	.419	29.882	2	83	.000	194.961	1.975	.000	.000
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1948

Greater than or equal to:



Model Summary and Parameter Estimates

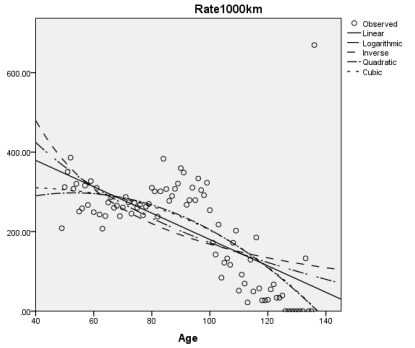
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.802	246.406	1	61	.000	-5.836	6.385		
Logarithmic	.754	186.866	1	61	.000	-202.234	125.589		
Inverse	.274	22.996	1	61	.000	234.278	-477.075		
Quadratic	.885	231.081	2	60	.000	-93.727	14.498	-.127	
Cubic	.923	235.979	3	59	.000	-20.182	1.228	.388	-.005
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

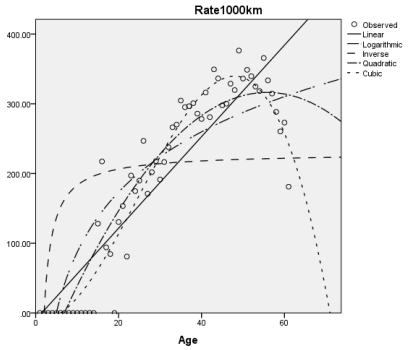
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.438	66.914	1	86	.000	512.117	-3.320		
Logarithmic	.394	56.005	1	86	.000	1439.833	-275.226		
Inverse	.339	44.073	1	86	.000	-39.083	20757.920		
Quadratic	.474	38.315	2	85	.000	178.104	4.491	-.042	
Cubic	.468	37.320	2	85	.000	304.503	.366	.000	.000
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1950

Greater than or equal to:



Model Summary and Parameter Estimates

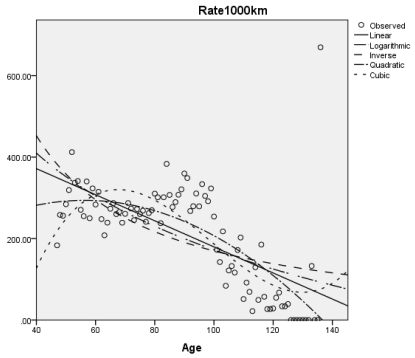
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.794	227.980	1	59	.000	-9.627	6.562		
Logarithmic	.740	167.518	1	59	.000	-200.017	124.637		
Inverse	.269	21.658	1	59	.000	229.674	-466.048		
Quadratic	.874	201.175	2	58	.000	-95.526	14.743	-.132	
Cubic	.926	236.137	3	57	.000	-9.630	-1.236	.507	-.007
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

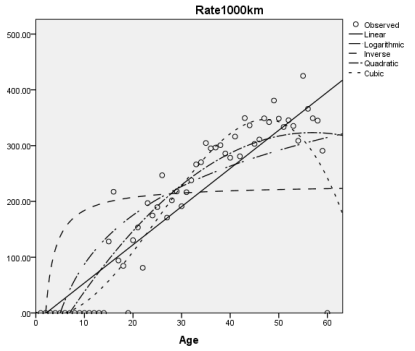
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.428	65.866	1	88	.000	499.864	-3.203		
Logarithmic	.379	53.646	1	88	.000	1365.587	-259.082		
Inverse	.317	40.778	1	88	.000	-19.052	18874.786		
Quadratic	.475	39.327	2	87	.000	149.226	5.133	-.046	
Cubic	.526	31.800	3	86	.000	-1218.687	55.057	-.619	.002
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1952

Greater than or equal to:



Model Summary and Parameter Estimates

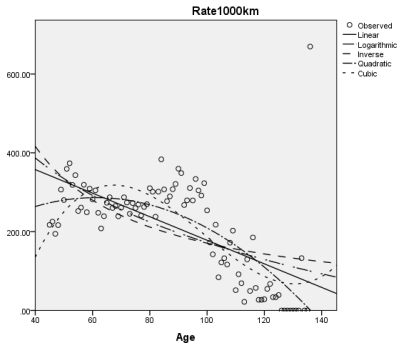
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.741	166.014	1	58	.000	-14.598	6.843		
Logarithmic	.677	121.804	1	58	.000	-204.657	126.845		
Inverse	.243	18.651	1	58	.000	230.620	-467.977		
Quadratic	.805	117.519	2	57	.000	-96.295	14.749	-.130	
Cubic	.864	118.390	3	56	.000	1.472	-3.724	.621	-.008
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

Dependent Variable: Rate1000km

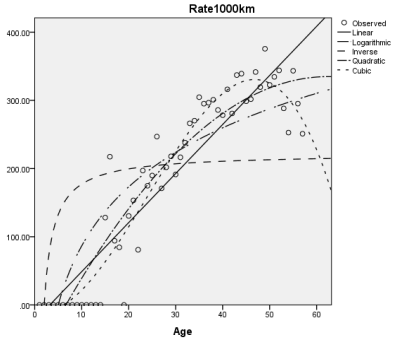
Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.403	60.688	1	90	.000	477.059	-2.990		
Logarithmic	.346	47.630	1	90	.000	1252.603	-234.642		
Inverse	.278	34.669	1	90	.000	6.728	16394.228		
Quadratic	.468	39.202	2	89	.000	96.675	6.208	-.051	
Cubic	.523	32.158	3	88	.000	-1148.957	52.548	-.591	.002
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

1954

Greater than or equal to:



Model Summary and Parameter Estimates

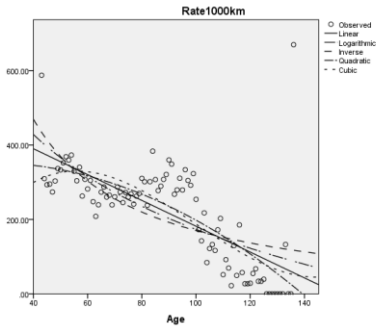
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.843	295.044	1	55	.000	-23.157	7.195		
Logarithmic	.743	159.082	1	55	.000	-201.088	124.921		
Inverse	.265	19.819	1	55	.000	221.868	-447.827		
Quadratic	.884	205.246	2	54	.000	-84.591	13.443	-.108	
Cubic	.922	209.879	3	53	.000	-10.110	-1.329	.523	-.007
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Less than:



Model Summary and Parameter Estimates

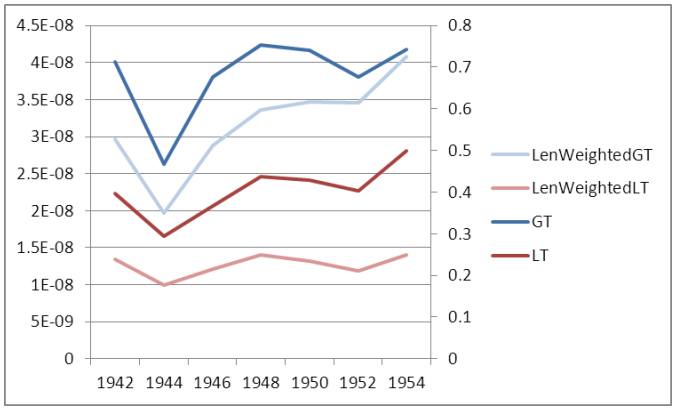
Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.499	91.739	1	92	.000	528.829	-3.472		
Logarithmic	.462	79.120	1	92	.000	1454.964	-278.342		
Inverse	.409	63.770	1	92	.000	-29.590	19950.102		
Quadratic	.516	48.457	2	91	.000	339.907	1.177	-.026	
Cubic	.525	33.182	3	90	.000	-154.345	19.933	-.248	.001
Compound[a]	.	.	.	.	.	.	.		
Power[a]	.	.	.	.	.	.	.		
S[a]	.	.	.	.	.	.	.		
Growth[a]	.	.	.	.	.	.	.		
Exponential[a]	.	.	.	.	.	.	.		
Logistic[a]	.	.	.	.	.	.	.		

The independent variable is Age.

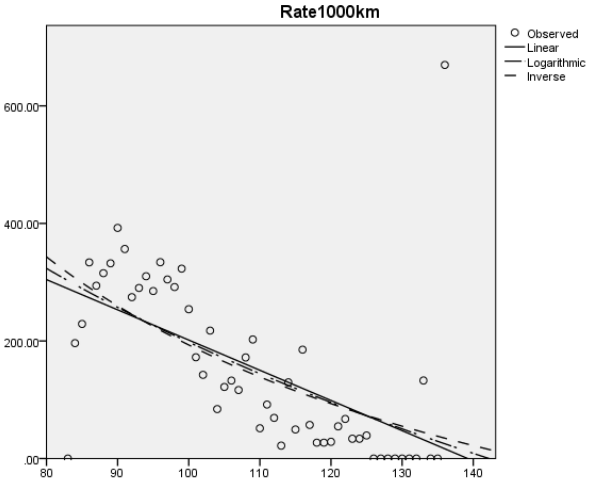
a. The dependent variable (Rate1000km) contains non-positive values. The minimum value is .00. Log transform cannot be applied. The Compound, Power, S, Growth, Exponential, and Logistic models cannot be calculated for this variable.

Results – Best date split:

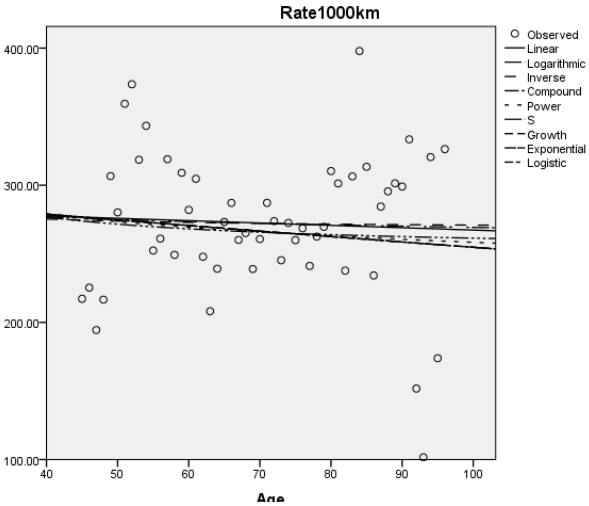


The lowest level of pre/post residual error seems to occur at either 1944 or 1952. Although the best method of choosing the appropriate date ranges may be visual, this is because the 'rehab effect' is still interfering. The object of this analysis is to try to determine background trends in deterioration; therefore the CI scatter plots demonstrating improving condition with age are inherently suspicious. Whilst 1944 has the lowest apparent level of model error, this is based on polynomial curves showing the 'improving condition' phenomenon for the older asset group (installed before 1944); the 1952 split is more realistic, showing clearer deterioration of the older assets and improvement in the younger asset group, as would be expected on a typical 'bathtub' deterioration model. Further analysis is therefore grouped using this 1952 split in year installed, assuming notably different asset performance either side of this estimated introduction date for spun iron.

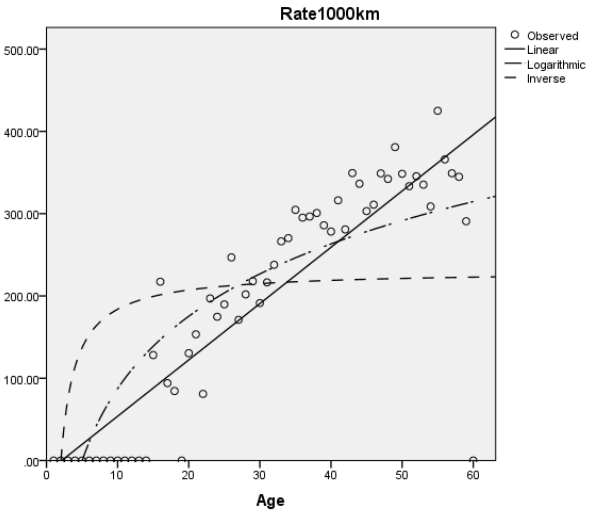
Testing three eras



Pre 1914
R^2 0.330



1914 – 1952
R^2 0.01



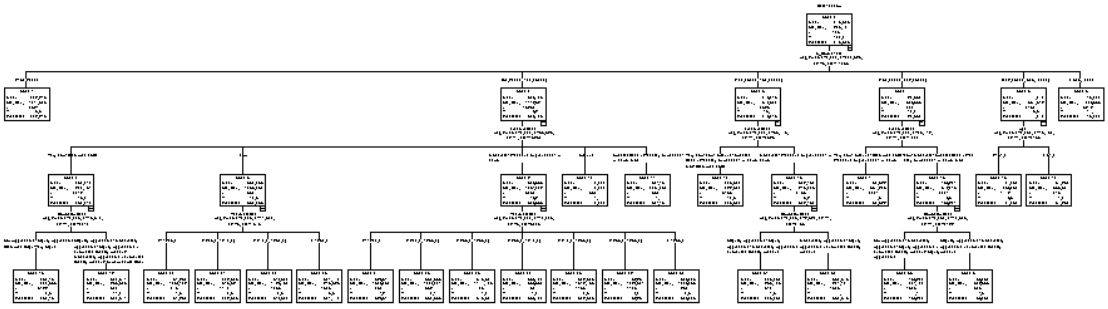
Post 1952
R^2 0.677

Splitting into 3 eras does give better fits on the extremes, but gives a flat period in between. This doesn't make much engineering sense, so hasn't been progressed. It is also likely to over divide the data/overfit the model.

CI Final Analysis for Models

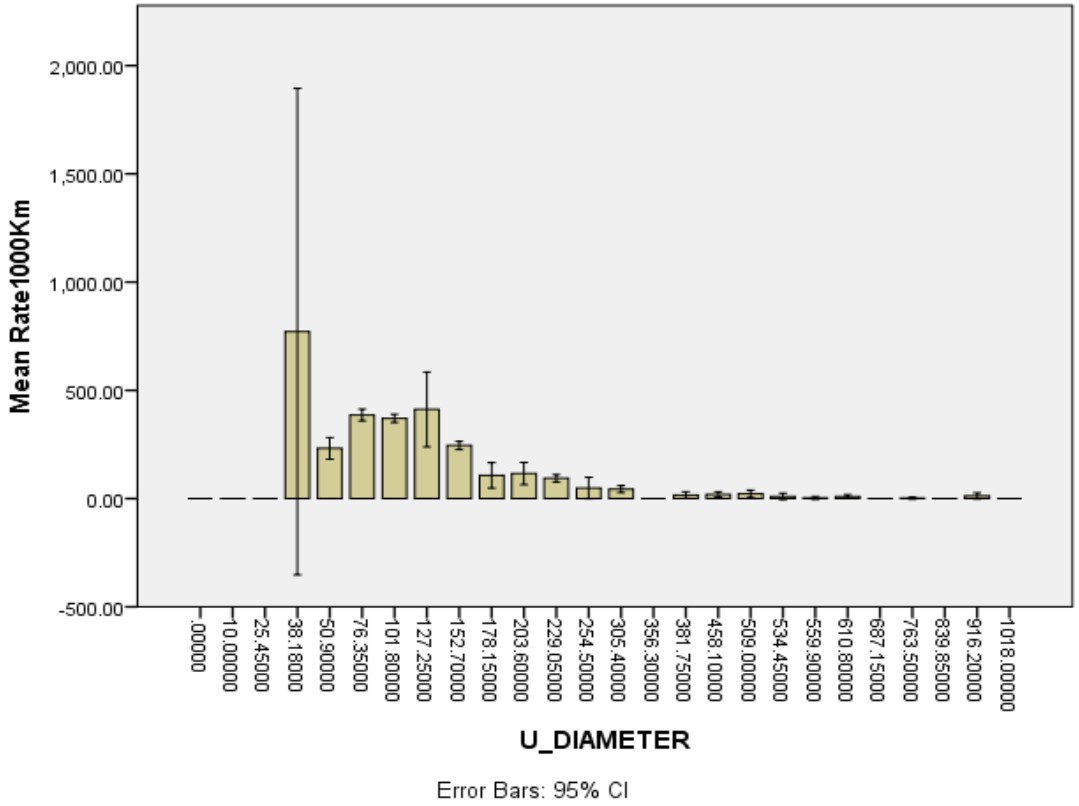
Initial analysis has been centred on trying to assess eras and patterns in age. This has shown too many seemingly significant relationships. The intention is to now go back to using regression trees to highlight significance order for predictors. Efforts will be made to limit the total number of models – over-fitting the tree. There is also an issue that restricting the tree from branching excessively is achieved by limiting the minimum number of cases allowed within each child node; this is fine for categorical predictors where each have thousands of constituent cases, but may prevent the emergence of smaller, yet still significant, predictor combinations.

Initial run of tree shows:



Shows initial branch of diameter, followed by fracture potential, then either corrosion, age or year installed.

Analysis of means by diameter gives:



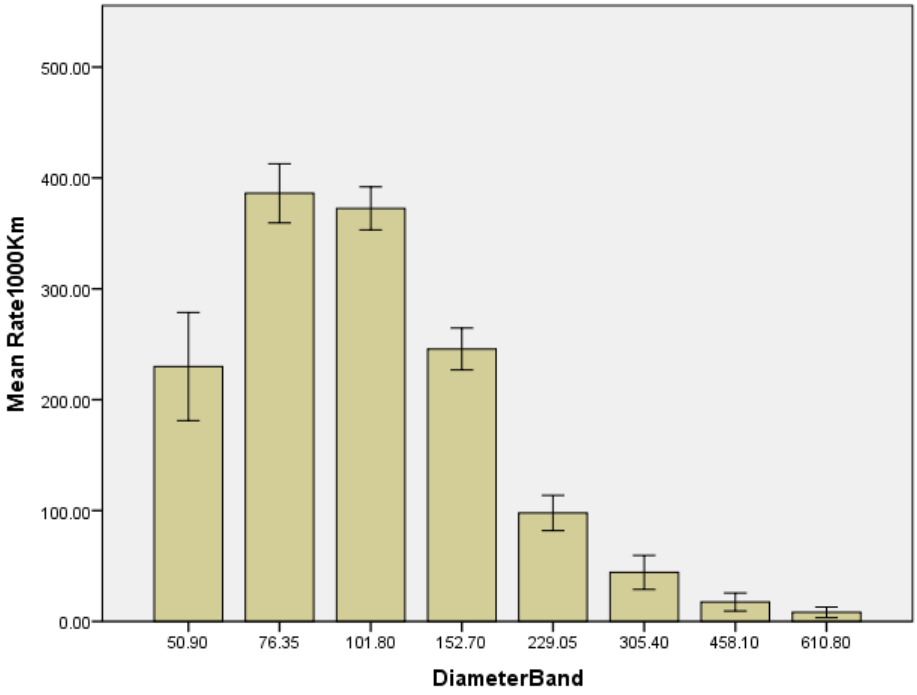
It's apparent that some of the diameters are notably different and that the confidence intervals are reasonably reliable in some cases.

		Length	Length	Rate1000Km	Rate1000Km
		Sum	Column Sum %	Mean	Count
U_DIAMETER	.00000	4231.216	0.0%	.00	16
	10.00000	4730.636	0.0%	.00	9
	25.45000	15136.717	0.0%	.00	32
	38.18000	7418.097	0.0%	771.57	16
	50.90000	952584.229	1.8%	231.83	2318
	76.35000	10791630.75	20.1%	386.17	8167
	101.80000	23488511.76	43.8%	370.98	10383
	127.25000	190493.334	0.4%	411.96	407
	152.70000	8538363.659	15.9%	245.71	7598
	178.15000	282242.474	0.5%	107.50	435
	203.60000	401617.032	0.7%	115.75	531
	229.05000	2088571.469	3.9%	93.63	3236
	254.50000	187508.286	0.3%	48.72	243
	305.40000	1545163.096	2.9%	43.67	1924
	356.30000	178864.362	0.3%	.00	176
	381.75000	1154181.421	2.2%	16.52	1341
	458.10000	1493379.310	2.8%	19.17	1552
	509.00000	360424.567	0.7%	22.30	432
	534.45000	368444.986	0.7%	10.08	304
	559.90000	155014.440	0.3%	3.33	160
	610.80000	907709.665	1.7%	9.75	1136
	687.15000	23746.412	0.0%	.00	64
	763.50000	74055.344	0.1%	2.45	208
	839.85000	46942.715	0.1%	.00	96
	916.20000	292675.264	0.5%	12.27	400
	1018.00000	98903.469	0.2%	.00	80

The above graph is complicated by the vast number of different diameters of cast iron. Some of them have so little length that it makes sense to group them in bands to make for easier analysis/interpretation, and to avoid over-dividing the data.

- < 2” become 2”
- 5” merged into 4”
- 7” and 8” merged into 9”
- 10” merged into 12”
- 14”, 15” and 20” merged into 18”
- Everything over 20” merged into 24”

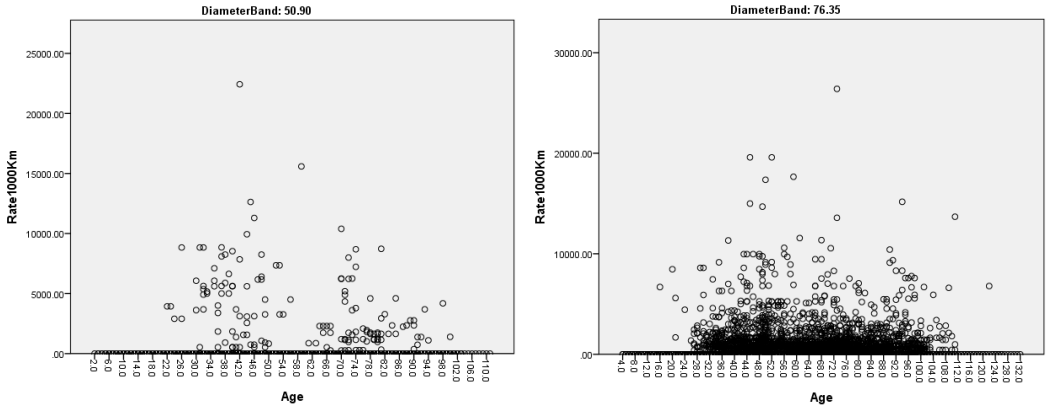
This gives this revised chart:

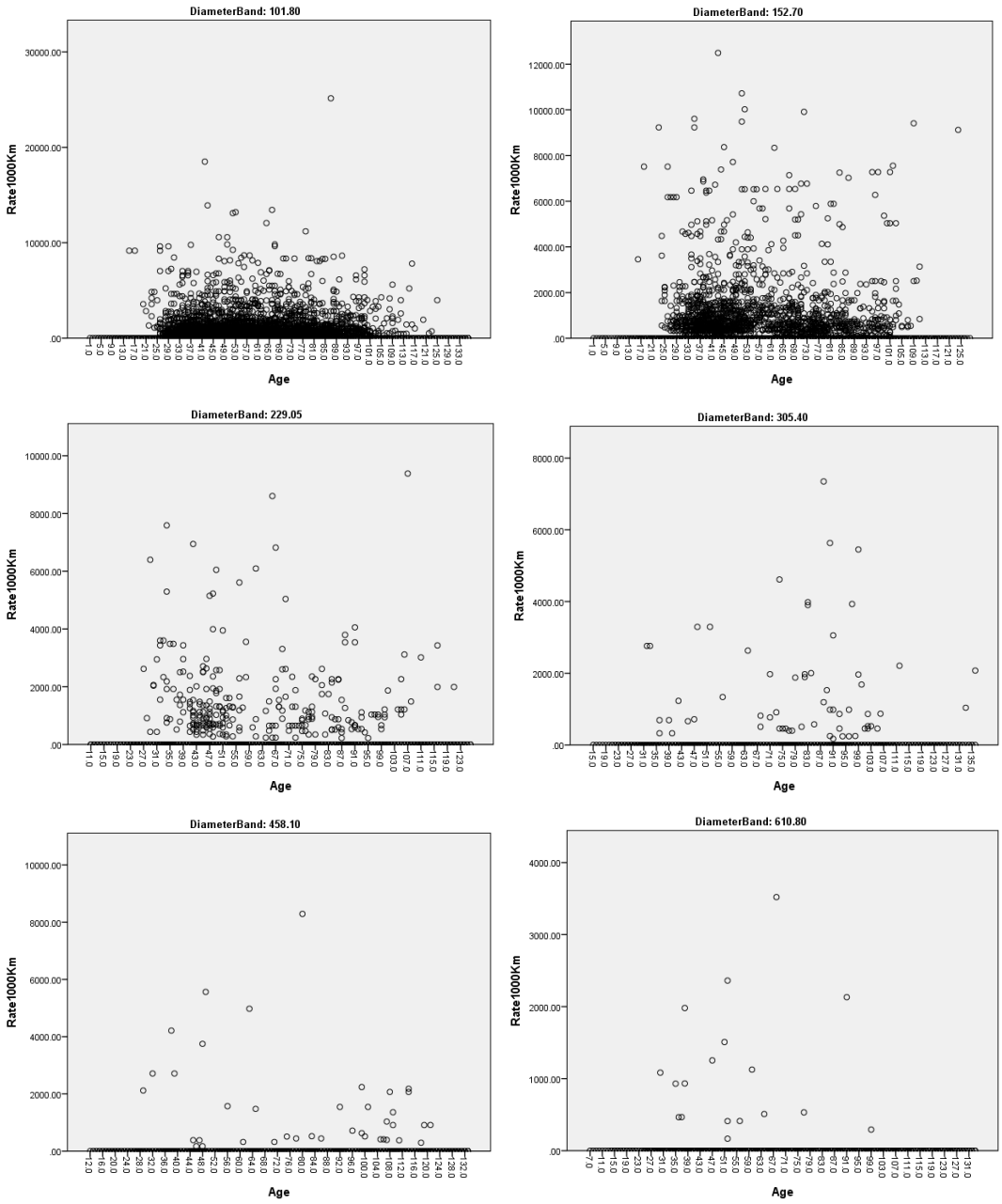


Error Bars: 95% CI

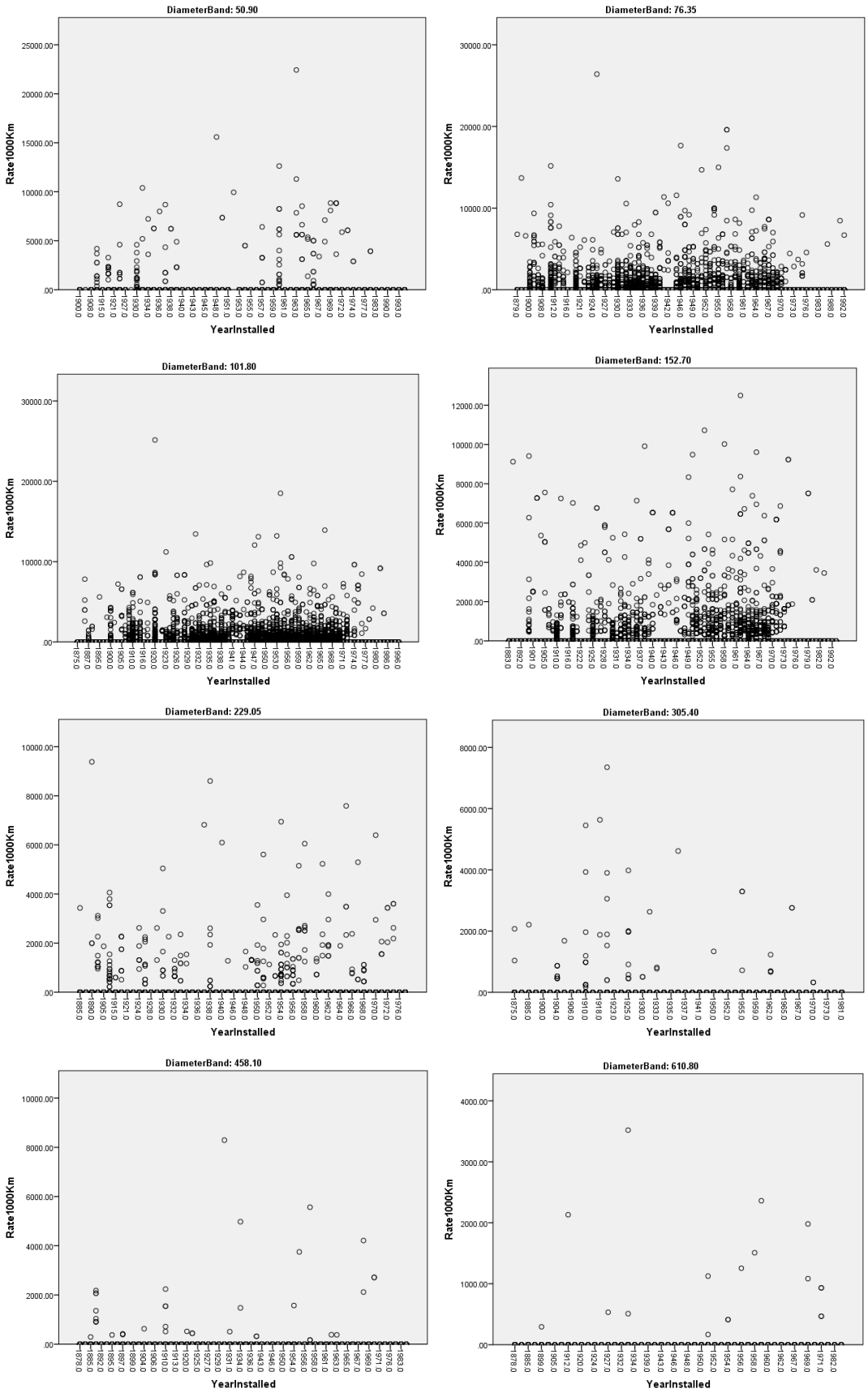
		Length	Length	Rate1000Km	Rate1000Km
		Sum	Column Sum %	Mean	Count
DiameterBand	50.90	984100.895	1.8%	229.91	2391
	76.35	10791630.75	20.1%	386.17	8167
	101.80	23679005.09	44.1%	372.52	10790
	152.70	8538363.659	15.9%	245.71	7598
	229.05	2772430.974	5.2%	97.86	4202
	305.40	1732671.382	3.2%	44.24	2167
	458.10	3186849.659	5.9%	17.58	3501
	610.80	1967492.295	3.7%	8.20	2448

Repeating age/year profiles for new diameter bands shows no obvious patterns with age:

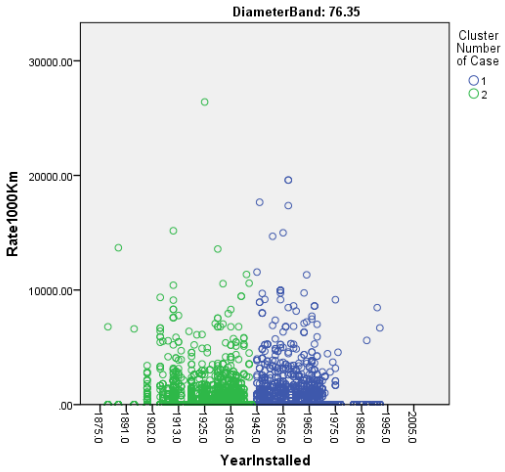




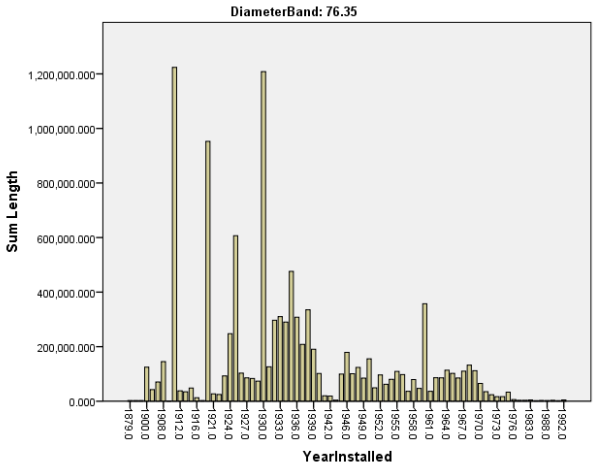
However, running the same graphs for year installed reveals some grouping/clustering:



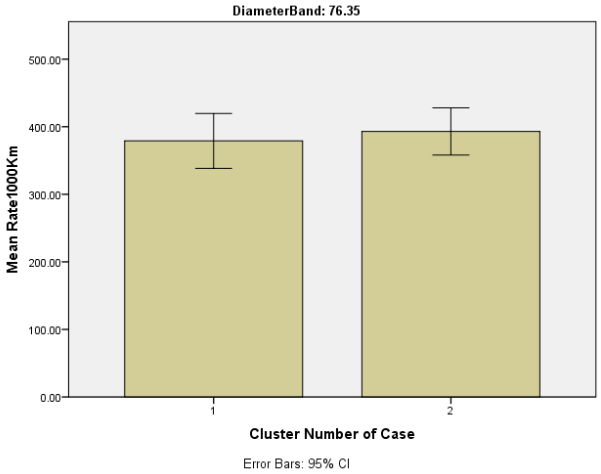
Most of the above graphs display two clear clusters, running a k-means cluster analysis algorithm highlights sensible groups, e.g these for 3":



But it seems that the clusters are affected by changes in the length exposure between groups:

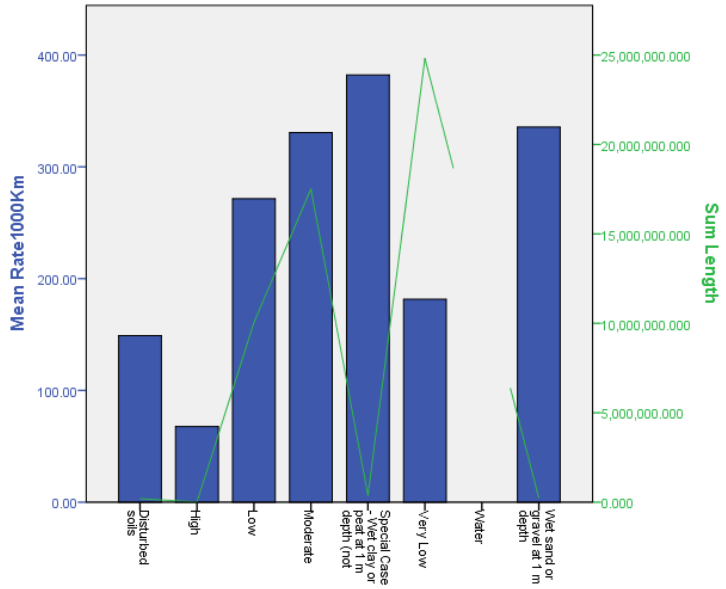


For most diameters, the mean failure rates of each cluster are similar:



So that when the cluster labels (eras) are re-processed through the regression tree, they are not found to be significant predictors. Likewise, when each cluster of each diameter band is analysed for age related patterns, there is still excessive variance in the lines.

The next stage is to follow the regression tree another branch and investigate the next most significant predictor – soil fracture potential:



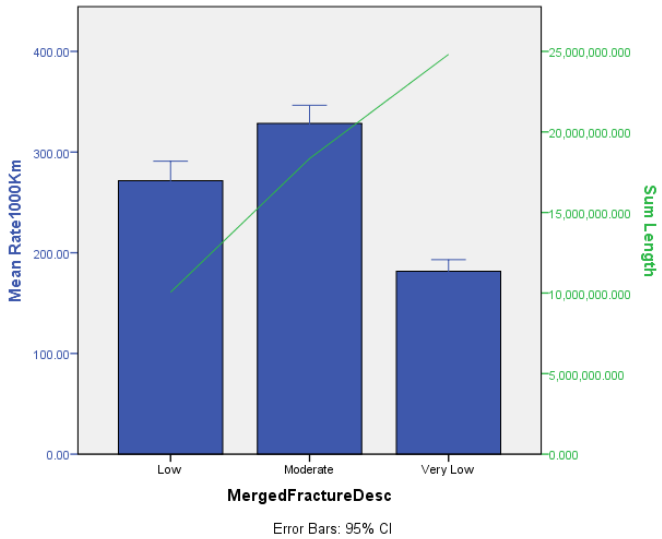
The fracture potential is found by the regression tree to be a significant predictor because the mean rate for each value is significantly different. However, the values that have notably high or notably low mean rates are generally very rare within the dataset.
Note: graphs assessing fracture input for each diameter band have also been reviewed and generally exhibit similar patterns, but with lower mean failures for larger diameters.

		Length	Length	Rate1000Km	Rate1000Km
		Sum	Column Sum %	Mean	Count
FractureDesc		435935.648	0.8%	.00	1113
	Disturbed soils	195271.991	0.4%	149.03	499
	High	14765.330	0.0%	67.73	16
	Low	10051388.08	18.7%	271.55	8496
	Moderate	17515695.95	32.6%	330.61	12013
	Special Case - Wet clay or peat at 1 m depth (not	380157.455	0.7%	382.14	1187
	Very Low	24816582.64	46.3%	181.58	17203
	Water	.	0.0%	.	0
	Wet sand or gravel at 1 m depth	242747.603	0.5%	335.53	737

The ‘Low’, ‘Moderate’ and ‘Very Low’ groups account for 97.6% of the assets. It therefore makes sense to group the others into these categories.

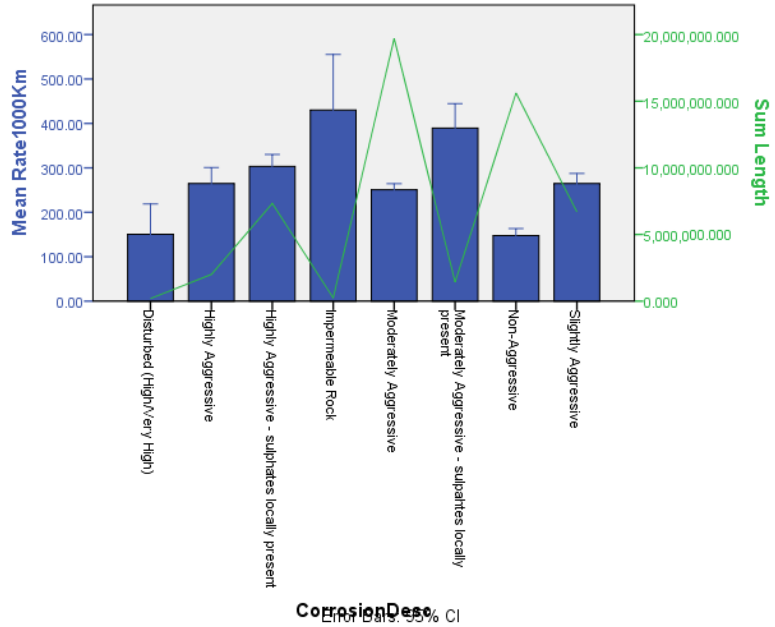
Special case, Wet Sand, High and Disturbed will all be merged into ‘Moderate’
Water will be merged into ‘Very Low’

New merged fracture values are as follows:



		Length	Length	Rate1000Km	Rate1000Km
		Sum	Column Sum %	Mean	Count
MergedFractureDesc		435935.648	0.8%	.00	1113
	Low	10051388.08	18.7%	271.55	8496
	Moderate	18348638.33	34.2%	328.53	14452
	Very Low	24816582.64	46.3%	181.58	17203

A similar process of factor reduction was also applied for corrosion grades:



		Length	Length	Rate1000Km	Rate1000Km
		Sum	Column Sum %	Mean	Count
CorrosionDesc		435935.648	0.8%	.00	1113
	Disturbed (High/Very High)	194471.573	0.4%	150.54	494
	Highly Aggressive	2007692.621	3.7%	264.61	4138
	Highly Aggressive - sulphates locally present	7332823.642	13.7%	303.27	5374
	Impermeable Rock	240510.576	0.4%	430.15	527
	Moderately Aggressive	19707372.35	36.7%	250.88	14074
	Moderately Aggressive - sulpahtes locally present	1427946.235	2.7%	389.59	2167
	Non-Aggressive	15617156.66	29.1%	147.44	6519
	Slightly Aggressive	6688635.394	12.5%	264.90	6858

There is significant cross correlation between the soil corrosion grades and fracture grades, this is because they follow natural changes in soil types between regions, so where fracture potential might change along a geographical line, corrositivity is also likely to change; and stable soils with low fracture potential also tend to exhibit low corrosivity.

Regression Tree

The decision trees module was used again with the newly merged fracture attributes and the original corrosion attributes. Again, the cross correlation seemed to be a problem, but the tree was still over-fitting to include both fracture and corrosion predictors; this created too many distinct 'bins' of data, some with minimal length of mains.

In an attempt to solve this problem a similar approach was taken to that used in the previous deterioration model. The most corrosive soils were grouped into one category, and the least corrosive soils into another. This creates a low number of categories, similar to that of the newly merged fracture variable:

- High group = Highly aggressive, Highly aggressive with sulphates and Moderately aggressive with sulphates
- Low group = All others

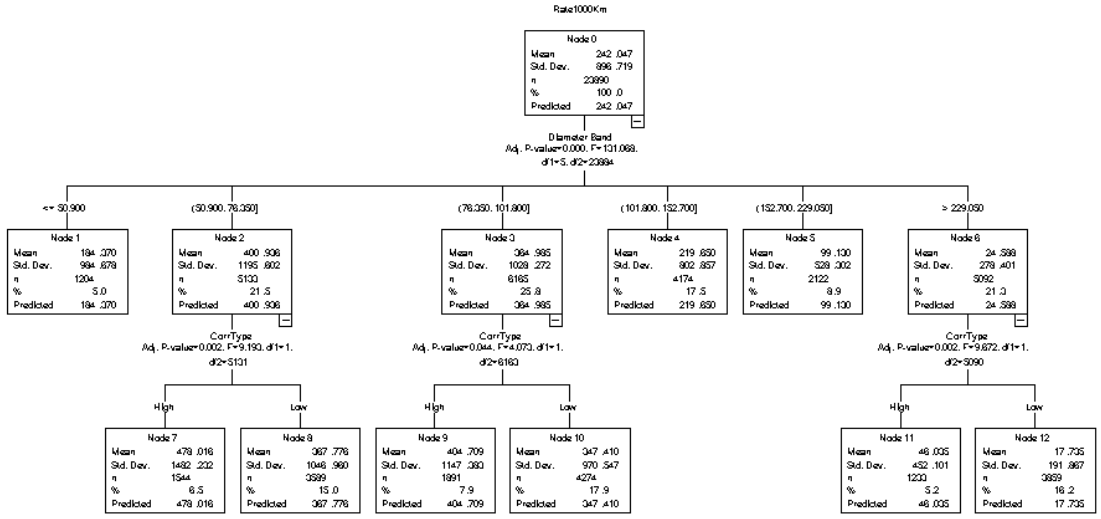
The repeated decision trees were then creating much more sensible numbers of 'branches' and the cross correlated fracture potential was no longer being selected.

Final Regression Trees

The final regression trees, also applying the proposed 1952 year installed split, are shown in the sections below. This immediately cuts the data into two, but is deemed necessary because of different levels of performance, and suspected different levels of deterioration, between vertically cast and spun iron.

6.1.9.1 CI where year installed less than 1952:

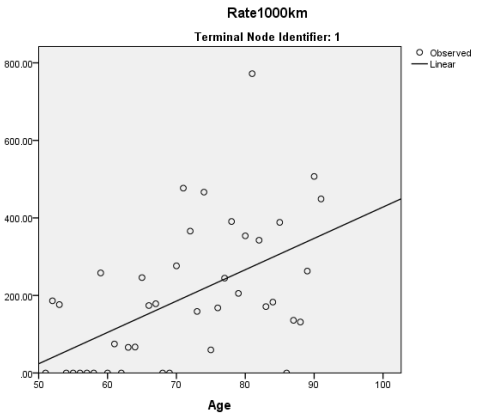
The older group of cast iron mains is analysed first. This again shows diameter as the most significant predictor, followed by the merged corrosion factor:

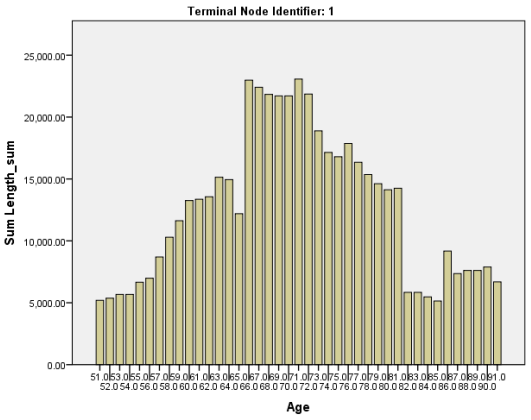


Node1: <=2”
Node4: 6”
Node5: >6” <=9”
Node7: >2” <=3” High Corrosion
Node8: >2” <=3” Low Corrosion
Node9: >3” <=5” High Corrosion
Node10: >3” <=5” Low Corrosion
Node11: >9” High Corrosion
Node12: >9” Low Corrosion

In each of the following nodes, attempts have been made to fit an age related model wherever possible. As seen with PVC, some of the nodes display excessive scatter resulting in some poor statistical fits; however, models have been selected where there is a clear visual pattern and where that pattern matches engineering expectations for how the node might perform:

Node1: <=2”





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.285	15.559	1	39	.000	-380.487	8.085

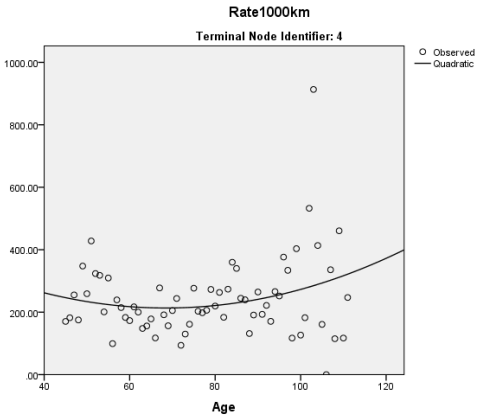
The independent variable is Age.

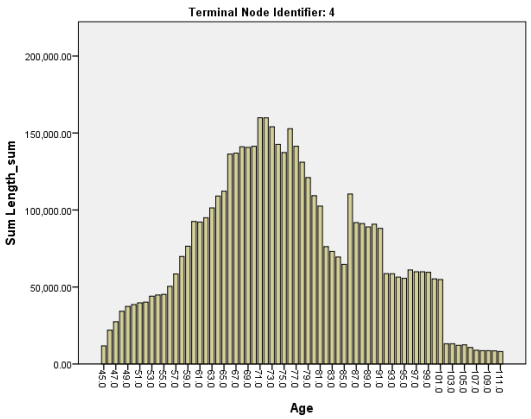
a. Terminal Node Identifier = 1

Chosen model:

y = 8.085 * Age – 380.487

Node4: 6”





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.056	1.902	2	64	.158	494.430	-8.210	.060

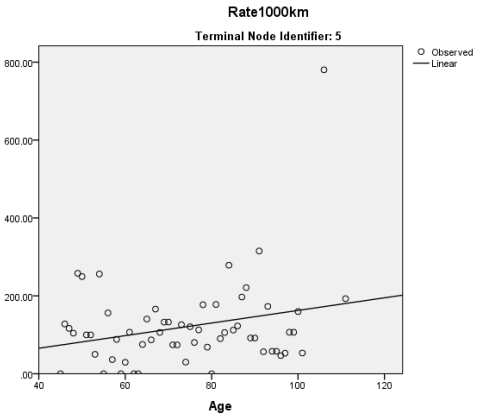
The independent variable is Age.

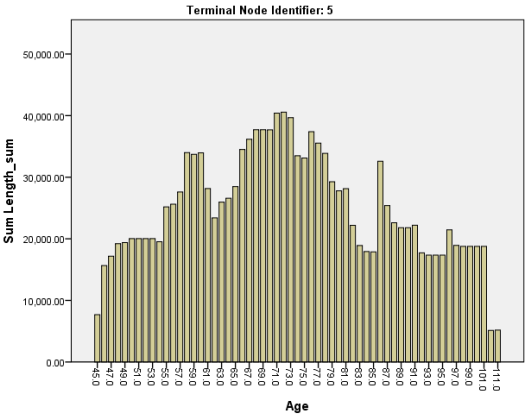
a. Terminal Node Identifier = 4

Chosen model:

y = -8.210 * Age + 0.060 * Age^2 + 494.430

Node5: >6” <=9”





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.063	3.806	1	57	.056	.697	1.619

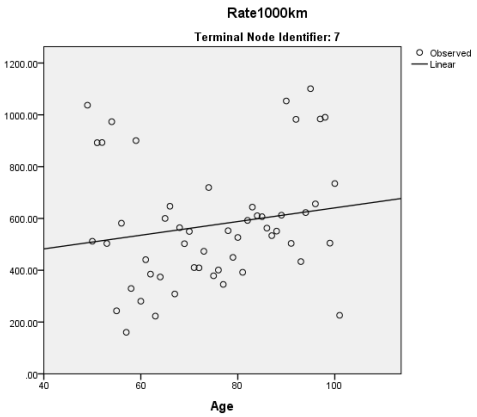
The independent variable is Age.

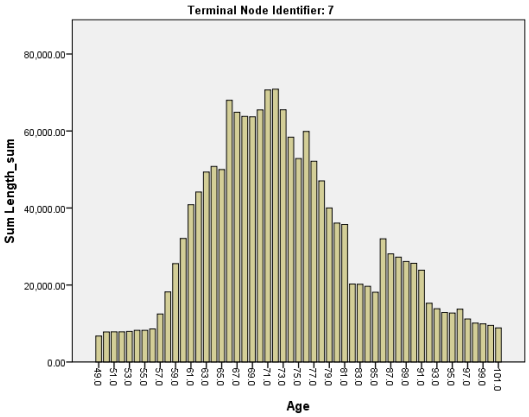
a. Terminal Node Identifier = 5

Chosen model:

y = 1.619 * Age + 0.697

Node7: >2" <=3" High Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.030	1.573	1	51	.215	376.837	2.640

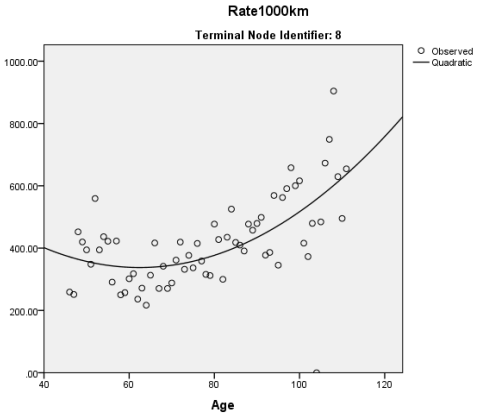
The independent variable is Age.

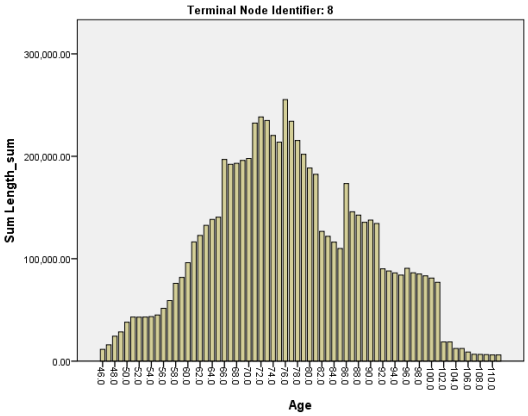
a. Terminal Node Identifier = 7

Chosen model:

y = 2.640 * Age + 376.837

Node8: >2" <=3" Low Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.378	19.122	2	63	.000	831.694	-15.828	.127

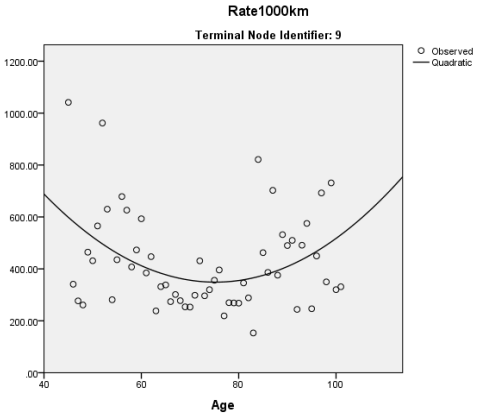
The independent variable is Age.

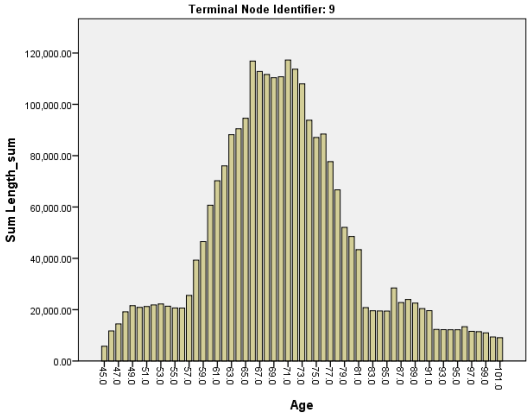
a. Terminal Node Identifier = 8

Chosen model:

y = -15.828 * Age + 0.127 * Age^2 + 831.694

Node9: >3" <=5" High Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.142	4.454	2	54	.016	1897.304	-41.182	.274

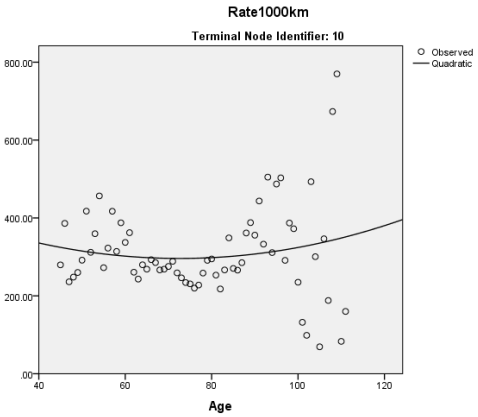
The independent variable is Age.

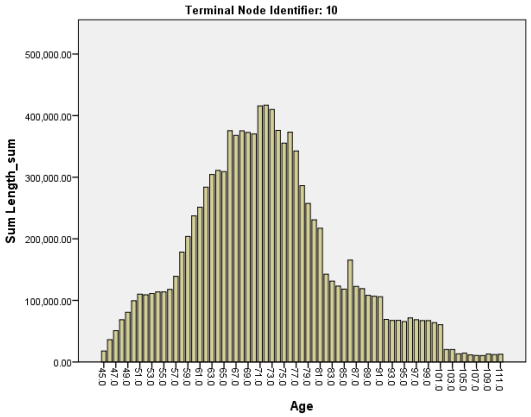
a. Terminal Node Identifier = 9

Chosen model:

y = -41.182 * Age + 0.274 * Age^2 + 1897.304

Node10: >3" <=5" Low Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.016	.528	2	64	.592	493.173	-5.433	.037

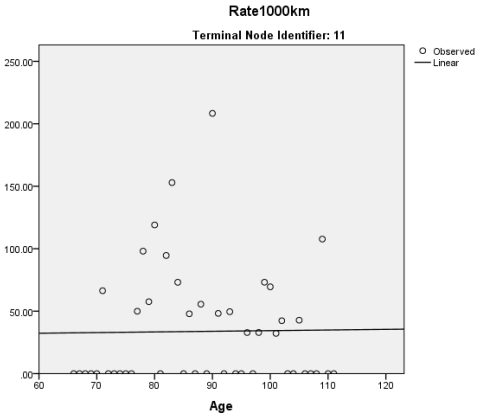
The independent variable is Age.

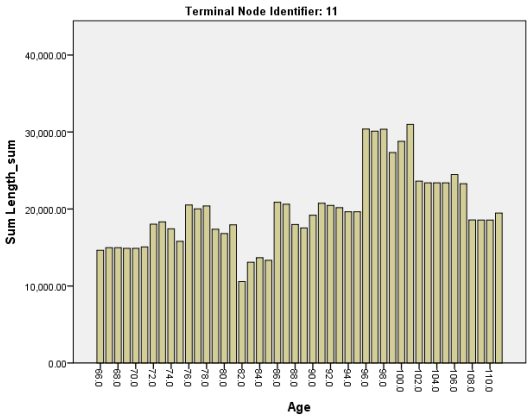
a. Terminal Node Identifier = 10

Chosen model:

$y = -5.433 * Age + 0.037 * Age^2 + 493.173$

Node11: >9" High Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.000	.009	1	44	.924	29.259	.051

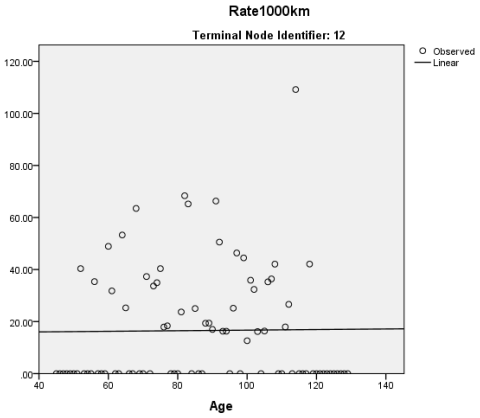
The independent variable is Age.

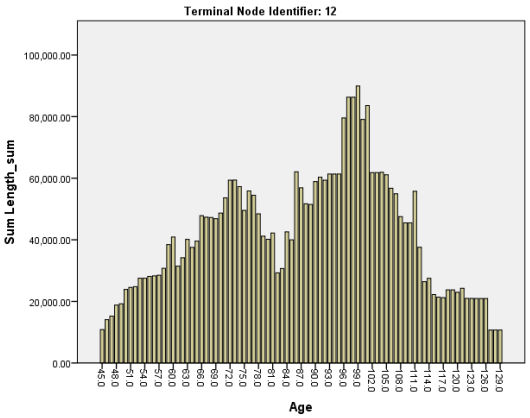
a. Terminal Node Identifier = 11

Chosen model:

$y = 0.051 * Age + 29.259$

Node12: >9" Low Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.000	.013	1	83	.910	15.580	.011

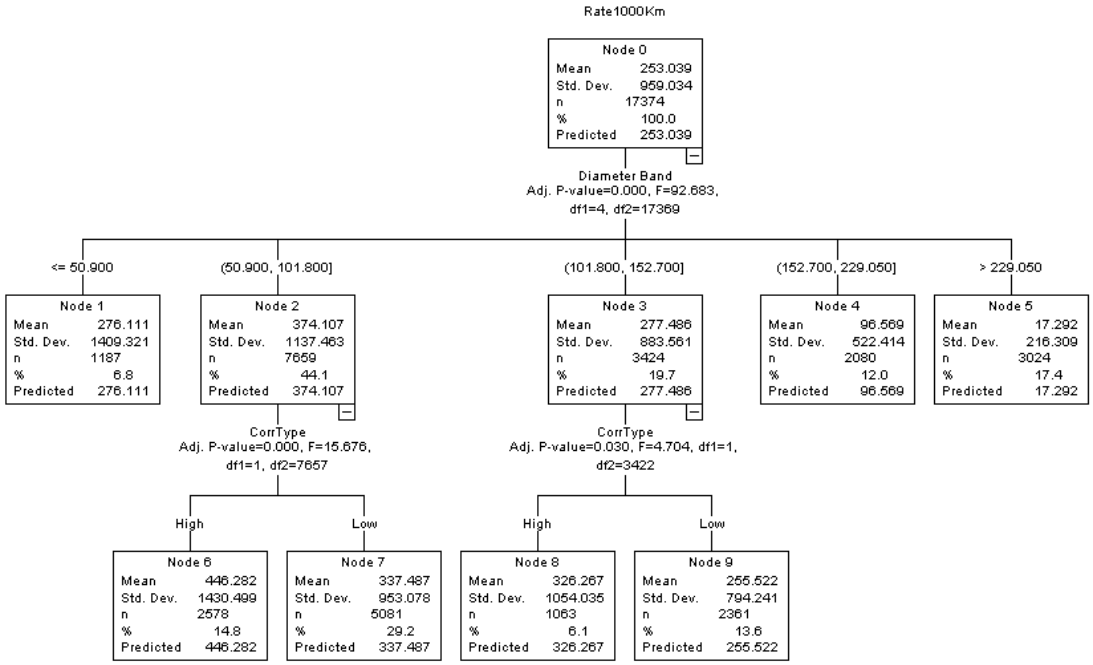
The independent variable is Age.

a. Terminal Node Identifier = 12

Chosen model:

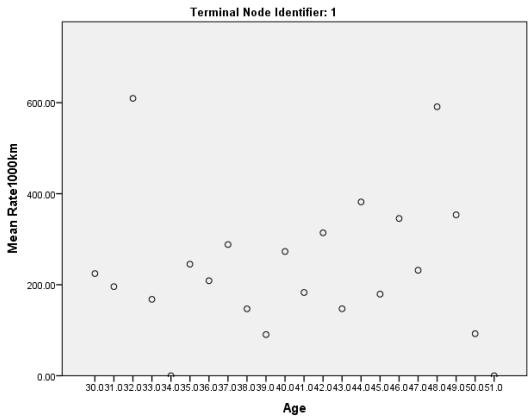
y = 0.011 * Age + 15.580

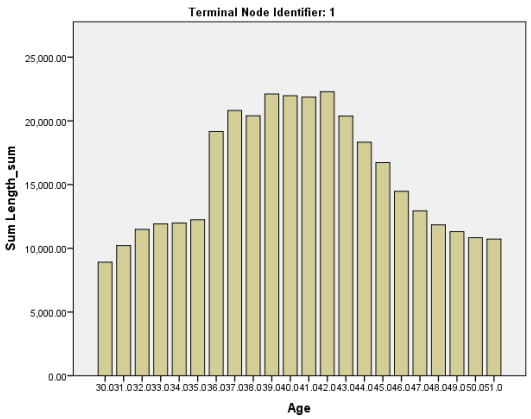
6.1.9.2 CI where year installed greater than or equal to 1952:

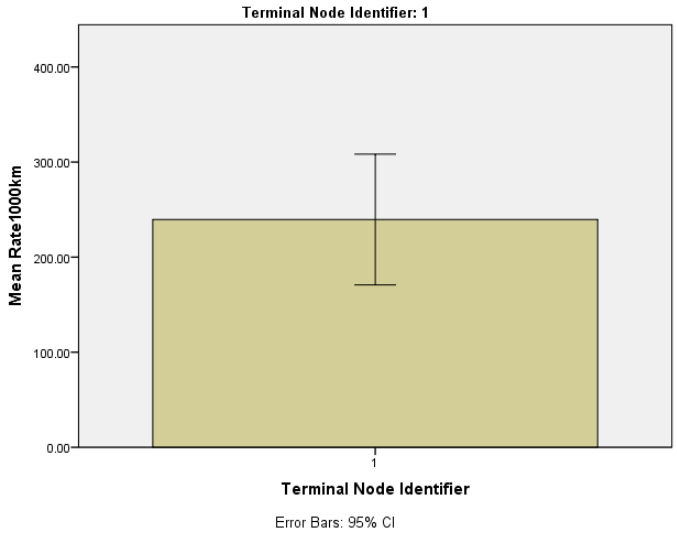


Node1: <=2”
Node4: >6” <=9”
Node5: >9”
Node6: >2” <=5” High Corrosion
Node7: >2” <=5” Low Corrosion
Node8: >5” <=6” High Corrosion
Node9: >5” <=6” Low Corrosion

Node1: <=2”



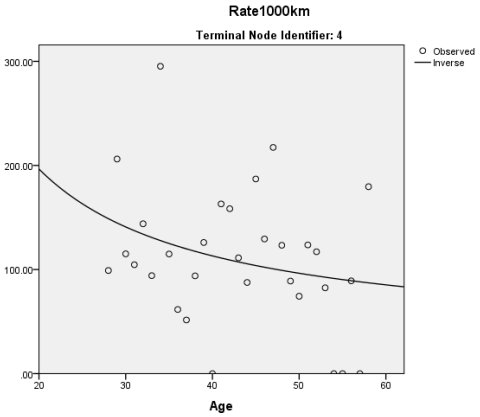


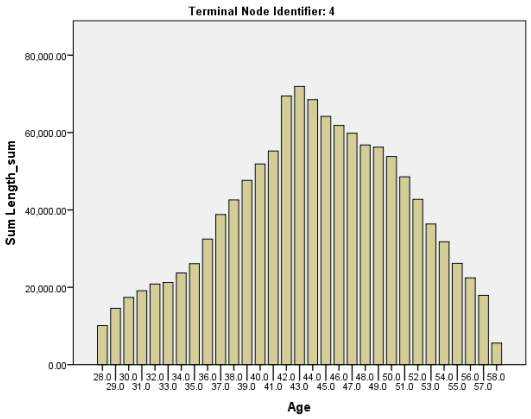


Mean value taken = 239.55

y = 239.55

Node4: >6" <=9"





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Inverse	.075	2.361	1	29	.135	29.633	3338.297

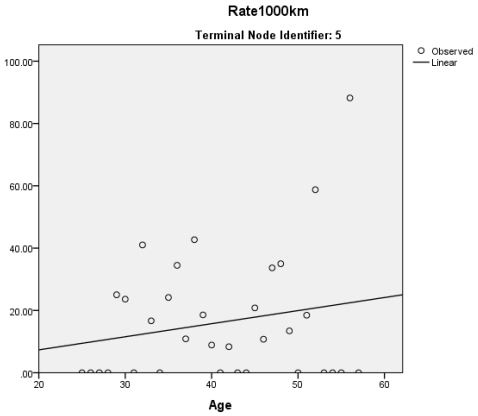
The independent variable is Age.

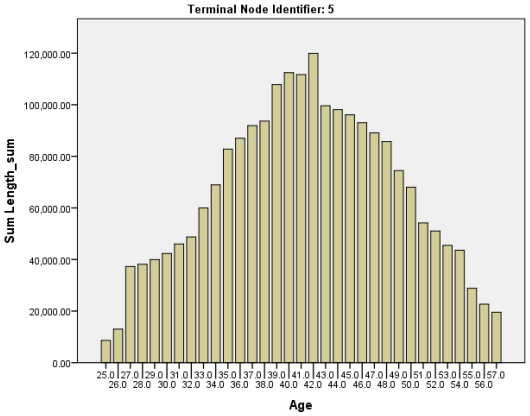
a. Terminal Node Identifier = 4

Chosen model:

y = 29.633 + 3338.297 / Age

Node5: >9"





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.039	1.271	1	31	.268	-1.051	.420

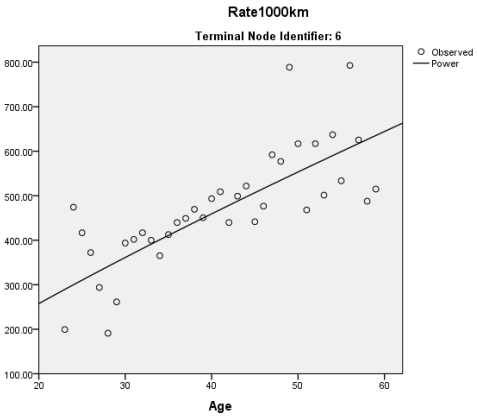
The independent variable is Age.

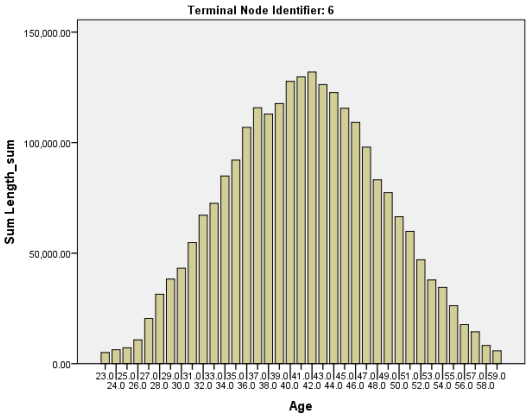
a. Terminal Node Identifier = 5

Chosen model:

y = 0.420 * Age – 1.051

Node6: >2" <=5" High Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Power	.579	48.129	1	35	.000	21.122	.835

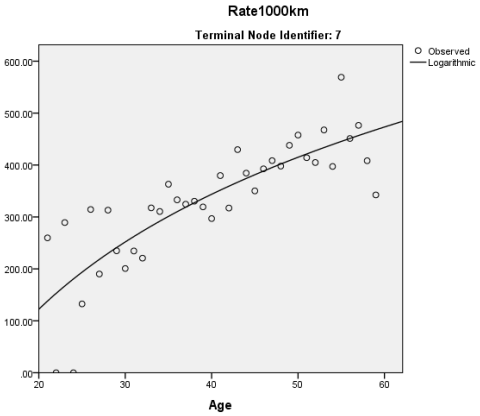
The independent variable is Age.

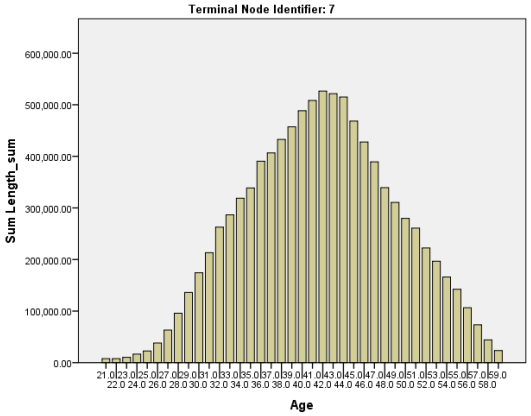
a. Terminal Node Identifier = 6

Chosen model:

$y = 21.122 * Age^{0.835}$

Node7: >2" <=5" Low Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Logarithmic	.674	76.597	1	37	.000	-834.491	319.395

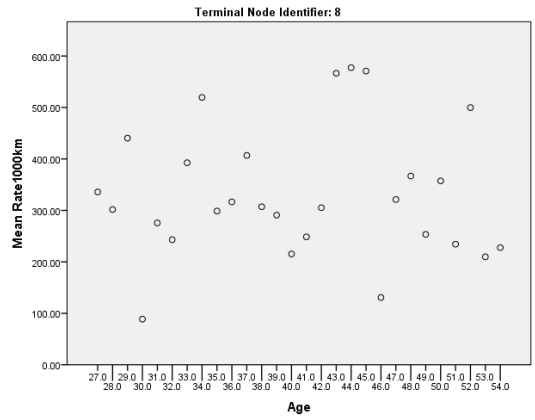
The independent variable is Age.

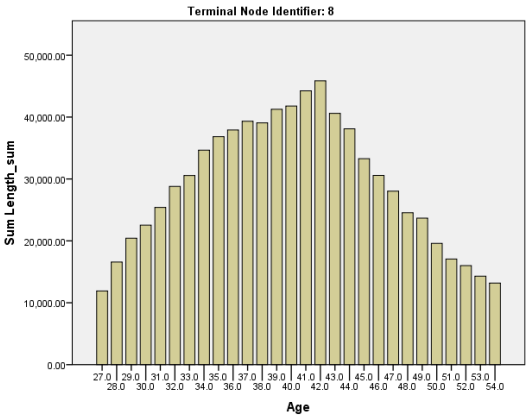
a. Terminal Node Identifier = 7

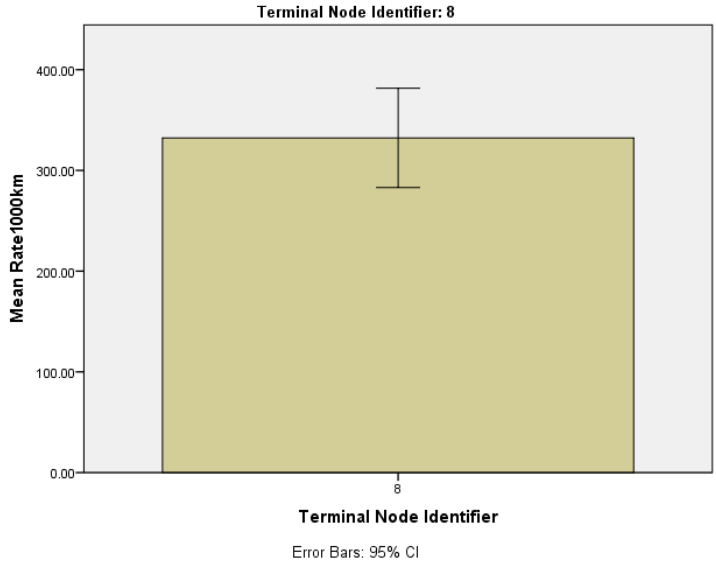
Chosen model:

y = 319.395 * ln(Age) – 834.491

Node8: >5” <=6” High Corrosion



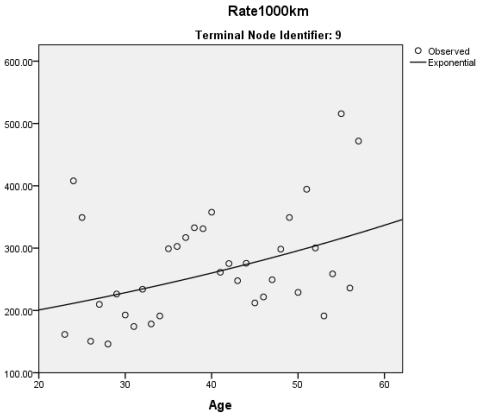


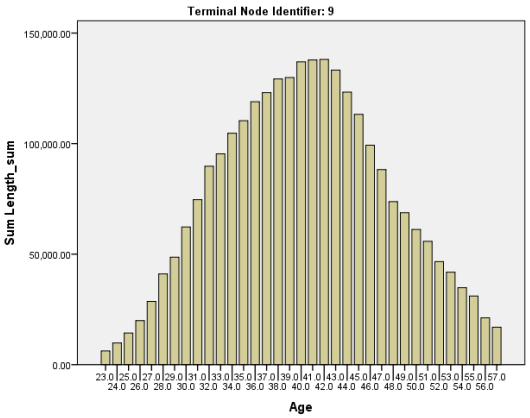


Mean value taken = 332.268

y = 332.268

Node9: >5" <=6" Low Corrosion





Model Summary and Parameter Estimates[a]

Dependent Variable: Rate1000km

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Exponential	.179	7.187	1	33	.011	155.022	.013

The independent variable is Age.

a. Terminal Node Identifier = 9

Chosen model:

$y = 155.022 * e^{0.013 * Age}$

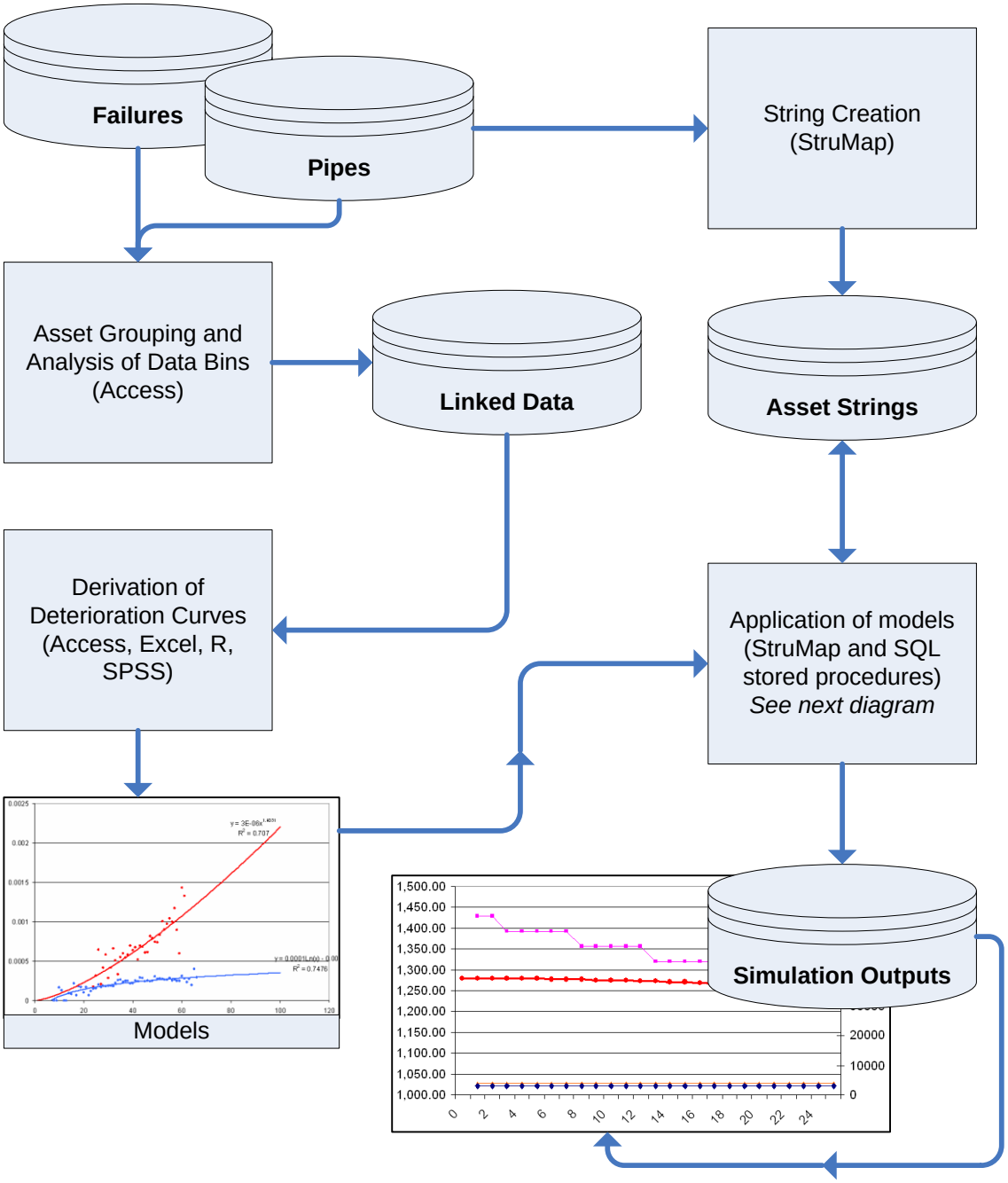
6.2 Final models for all cohorts

where **y = rate** (bursts/1000km) and **x = Age**
percentages shown are by length at start date:

HPPE:	
1. $y = 67.669 – 4.649 * \log(x)$	HPPE < 50 (9.45%)
2. $y = 0.2x + 39.5$	HPPE ≥ 50 (0%)
AC:	
3. $y = 378.73$	AC – Mod w/sulp (0.23%)
4. $y = 140.85$	AC – High (0.19%)
5. $y = 582.92$	AC – High w/sulp (0.18%)
6. $y = 189.22$	AC – Non to Mod ≤ 3” (0.74%)
7. $y = 2.178x + 138.090$	AC – Non to Mod > 3” <= 55 (1.21%)
8. $y = 7.485x – 227.004$	AC – Non to Mod > 3” > 55 (1.41%)
DI:	
9. $y = 53.653 – 10.743 * \log(x)$	DI ≤ 21yrs (4.82%)
10. $y = 20.95$	DI > 21yrs (1.68%)
MDPE:	
11. $y = 4.190x – 0.462x^2 + 0.011x^3 + 39.021$	MDPE ≤ 21 (13.35%)
12. $y = 25.14$	MDPE >21 ≤ 40 (3.63%)
13. $y = 0.3x + 13$	MDPE > 40 (0%)
STEEL:	
14. $y = 10$	STEEL outside following groups (0.06%)
15. $y = -4269.128 + 1086.462 * \log(x)$	STEEL >50yrs ≤ 200mm (0.07%)
16. $y = 1.7x - 37.721$	STEEL >30yrs > 200mm ≤ 600mm (0.67%)
17. $y = 0.113x + 8.999$	STEEL > 600mm (2.39%)
OTHERS:	
18. $y = 4.39$	OTHER Mats (0.23%)
mPVC:	
19. $y = 45.379$	mPVC (1.55%)
PVC:	
20. $y = 60.756 * e^{0.036x}$	PVC > 80mm ≤ 100mm (4.15%)
21. $y = 305.313 - 2627.690 / x$	PVC > 100mm ≤ 150mm (3.06%)
22. $y = 3.543x + 120.123$	PVC > 200mm ≤ 250mm (0.52%)
23. $y = 107.54$	PVC > 250mm (0.12%)
24. $y = 1.496x + 37.103$	PVC ≤ 80mm ≠ Mod Fracture (1.47%)
25. $y = 2.397x + 92.087$	PVC ≤ 80mm = Mod Fracture (0.86%)
26. $y = -826.675 + 385.174 * \log(x)$	PVC 150-200mm High Fracture (0.22%)
27. $y = 8.554x - 30.899$	PVC 150-200mm Low Fracture (0.38%)
CI:	
28. $y = 8.085x - 380.487$	Early CI ≤ 2” (0.16%)
29. $y = -8.210x + 0.060x^2 + 494.430$	Early CI 6” (4.63%)
30. $y = 1.619x + 0.697$	Early CI > 6” ≤ 9” (1.47%)
31. $y = 2.640x + 376.837$	Early CI > 2” ≤ 3” High Corr (1.41%)
32. $y = -15.828x + 0.127x^2 + 831.694$	Early CI > 2” ≤ 3” Low Corr (5.93%)
33. $y = -41.182x + 0.274x^2 + 1897.304$	Early CI > 3” ≤ 5” High Corr (2.29%)
34. $y = -5.433x + 0.037x^2 + 493.173$	Early CI > 3” ≤ 5” Low Corr (9.64%)
35. $y = 0.051x + 29.259$	Early CI > 9” High Corr (1.13%)
36. $y = 0.011x + 15.580$	Early CI > 9” Low Corr (3.68%)
37. $y = 239.55$	Late CI ≤ 2” (0.18%)
38. $y = 29.633 + 3338.297 / x$	Late CI > 6” ≤ 9” (1.18%)
39. $y = 0.420x - 1.051$	Late CI > 9” (2.29%)
40. $y = 21.122 * x^{0.835}$	Late CI > 2” ≤ 5” High Corr (1.99%)
41. $y = 319.395 * \log(x) - 834.491$	Late CI > 2” ≤ 5” Low Corr (8.20%)
42. $y = 332.268$	Late CI > 5” ≤ 6” High Corr (0.79%)
43. $y = 155.022 * e^{0.013x}$	Late CI > 5” ≤ 6” Low Corr (2.37%)

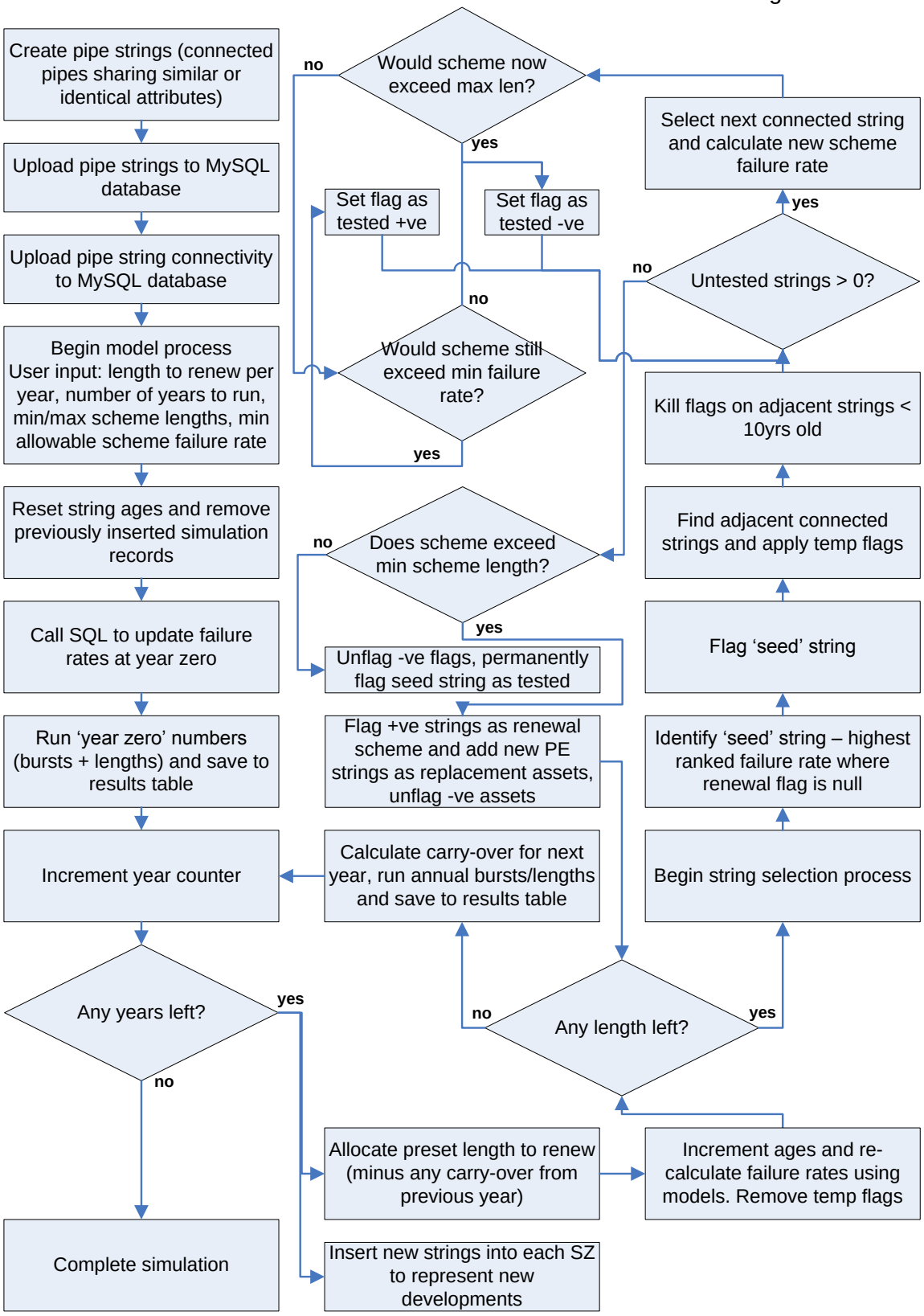
7. Application of model curves

7.1 Flow Diagram



7.2 Scheme selection logic diagram

This diagram shows how the models are applied iteratively to the asset data. Most of this process is handled within PL/SQL stored procedures, with GIS being used to post initial asset data and to download resulting schemes:



7.3 Stochastic effects

There are numerous potential random effects that could result in burst mains occurring above or below the mean values predicted by the model curves. One of these reviewed during the initial project is weather/seasonal impact.

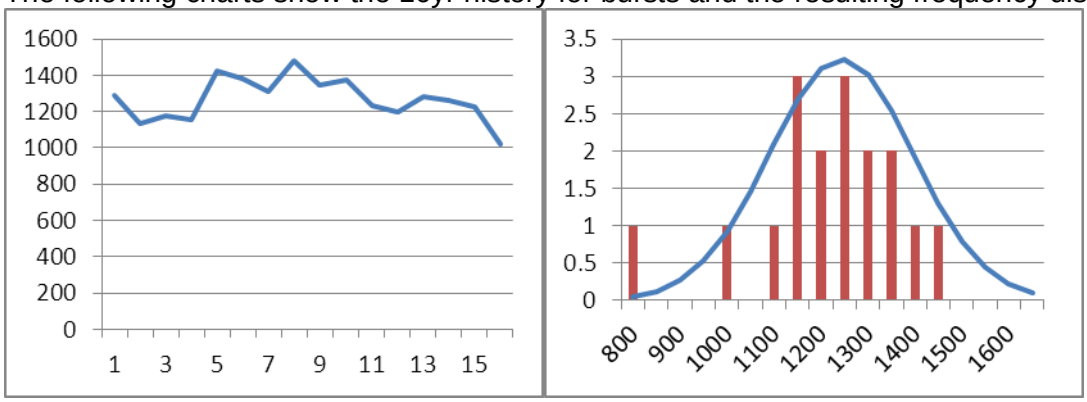
The intention was to generate the deterioration curves from the 16 years of available burst data, but by assessing these at each pipe age we can effectively 'iron-out' the spikes related to adverse seasonal weather; i.e.

- pipe 1 might be 50 years old during a bad winter (therefore the 50yr age bracket gets disproportionately large number of bursts)
- but pipe 2 doesn't reach 50yrs old until the following year which is a benign winter (therefore the 50yr age bracket gets a more realistic number of bursts)
- then pipe 3 reaches 50yrs old during the next year which is an exceptionally quiet year (therefore the 50yr age bracket gets a disproportionately low number of bursts)

Over the 16 year 'bursts window' any weather effects are likely to be 'averaged out', this gives more confidence that the core deterioration curves are as accurate as possible and not skewed by weather.

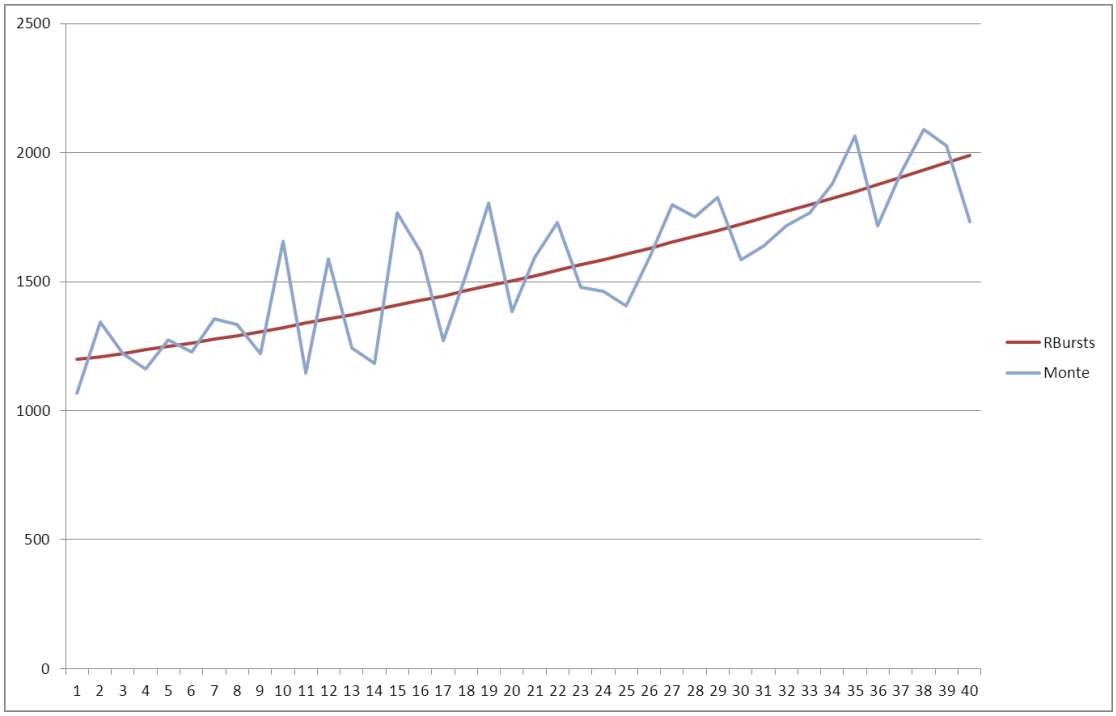
Monte Carlo techniques were then used to assess the expected variability around the curve for the overall numbers of burst mains.

The following charts show the 16yr history for bursts and the resulting frequency distribution:

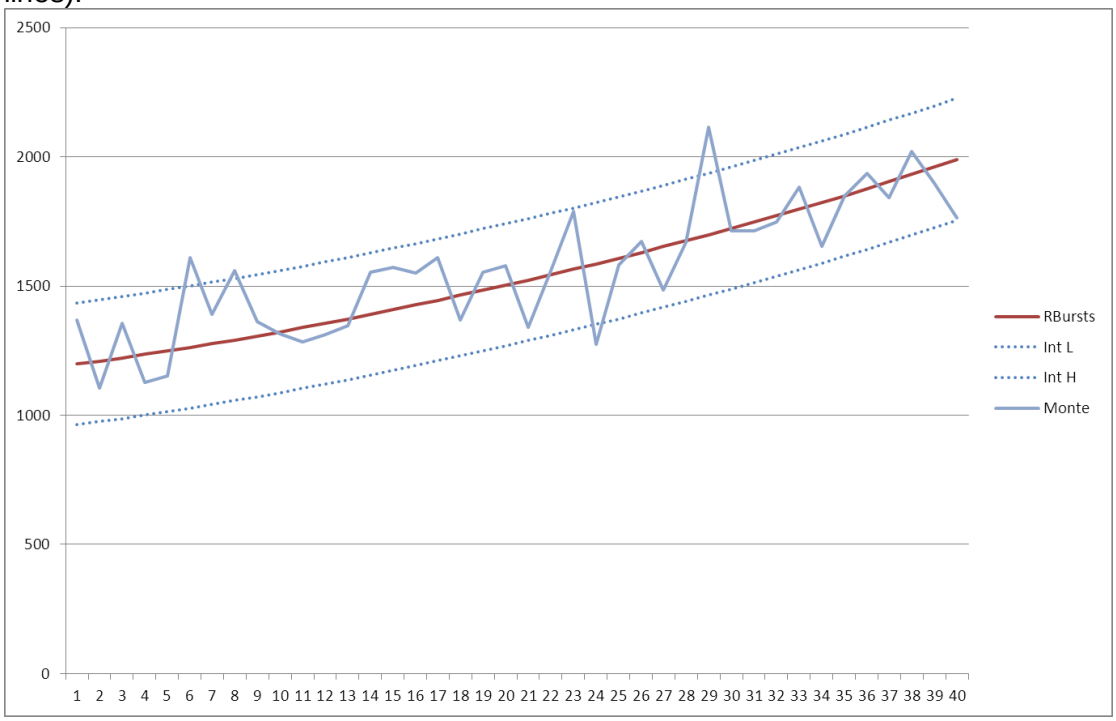


Assuming that all of the variance is associated with weather (which is false, but more on this below), then the number of bursts in a typical year should be the observed mean of 1243. The standard deviation around this mean (153.6) can then be used to determine the likelihood of this occurring within random weather conditions.

This level of variance can then be applied for the next 40 years to assess the potential minimum and maximum number of bursts likely to be achieved (again, this is based on assumption of constant variance, which is incorrect):



Every time the simulation is repeated, there is a new randomly selected line generated from the normal distribution shown above. Repeating this hundreds of times and observing the results allows for the generation of expected percentile limits for bursts, i.e. this chart showing the 5% and 95% percentiles (90% confidence that each year will be within the lines):



There are two inaccurate assumptions mentioned above that need to be addressed for the final version of the weather effects model:

- Historic observed variance is all due to weather
    - There is an element of asset improvement within the historic data due to the level of renewal activity undertaken since 1996. This should be measurable

and it's a reasonably simple activity to correct for this improvement before generating the sample distribution.
    - There will also be a reduction in the observed bursts due to large scale pressure management applied during the introduction of district metered areas. This will be more difficult to measure, but is thought to be secondary in magnitude when compared to asset improvements.
- Future variance will remain constant
    - Other analysis projects have proven that materials are affected differently by extreme seasons; i.e. cast iron reacts to cold weather ground movement much more than polyethylene. As the mix of materials moves away from AC/CI/PVC and towards PE/DI, then the level of seasonal variance can be expected to lessen.
    - It should be possible to apply this by undertaking individual Monte Carlo analyses at material level, where it's expected the CI will have a differently shaped probability distribution to PE. This would require more effort, but there should be enough data available for the main material groups to do this successfully.
    - Accounting for changes in material mix should give a wider confidence range at year 1, getting gradually narrower by year 40.

8. Validating outputs

This section of the report describes some of the work undertaken to validate the model and prove or disprove that the outputs are correct.

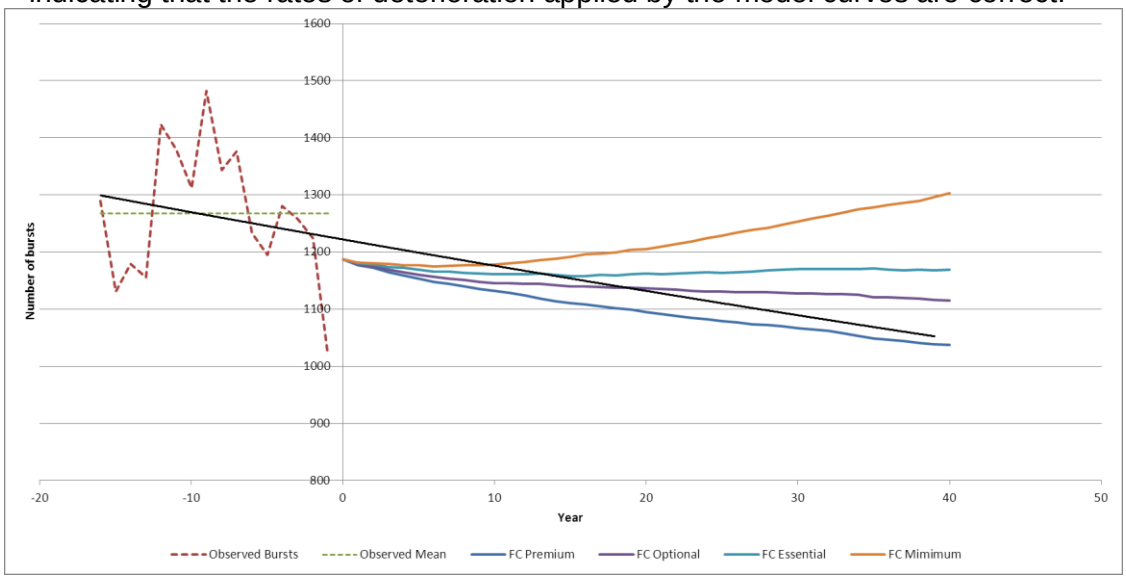
8.1 Sense check – scale and rates

The model curves and scheme selection algorithm were used to simulate four proactive scenarios for potential investment over the next 40 years. These scenarios were then compared against the recently observed burst trends (without any effort to correct for weather or any other stochastic influences).

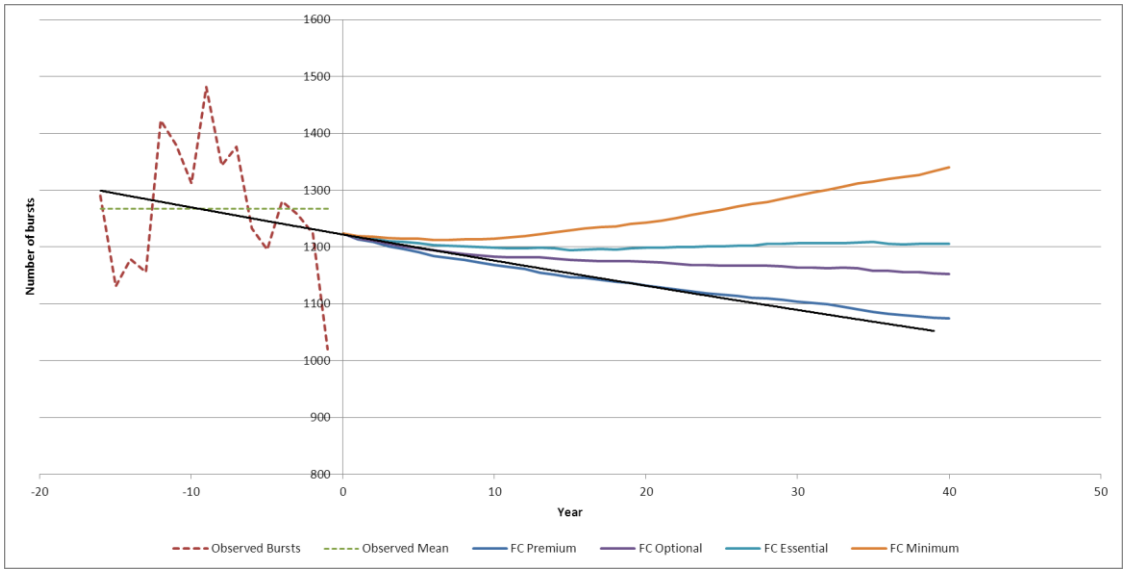
The scenarios were:

	km of assets replaced per year of each period:							
	AMP6	AMP7	AMP8	AMP9	AMP10	AMP11	AMP12	AMP13
Minimum	45	45	40	40	35	35	35	35
Essential	50	50	50	45	45	45	45	45
Optional	55	55	50	50	50	48	45	45
Premium	60	60	60	55	55	50	50	50

In the graph below, the trend line for the current bursts trend is shown in black and the proactive scenarios are drawn out on the right hand side of the y-axis. The premium scenario, shown in dark blue, is the closest investment level to that applied during the past few years. Note that the gradient for the forecast bursts matches that of the historic trend line – indicating that the rates of deterioration applied by the model curves are correct.



This chart shows the same proactive scenarios, but each one has been scaled upwards by 37 bursts to ensure that they cross the y-axis at the same point as the historic trend line.



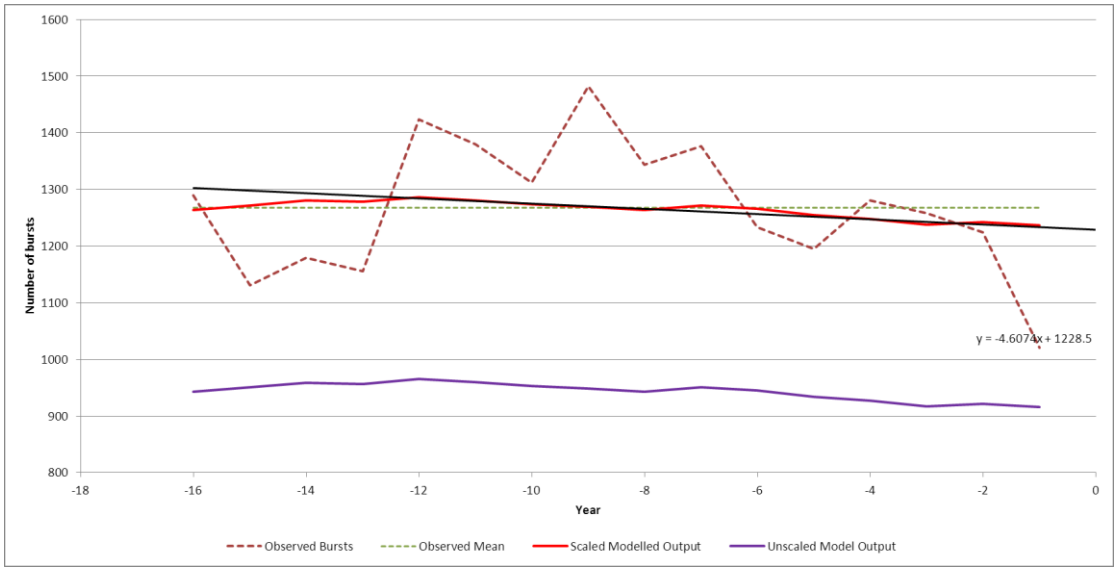
8.2 Backcast analysis

An attempt has been made to backcast the asset data to the beginning of 1996/97 year (start of reliable bursts data – the 'bursts envelope'). The asset data was then regressed again to 2012 using the model curves and the following known lengths of renewal undertaken in each year:

1996-97	3.1 km
1997-98	6.8 km
1998-99	36.3 km
1999-00	11.1 km
2000-01	47.7 km
2001-02	55.4 km
2002-03	55 km
2003-04	51.5 km
2004-05	10.3 km
2005-06	44.1 km
2006-07	71 km
2007-08	60.5 km
2008-09	67.2 km
2009-10	27.7 km
2010-11	49.7 km
2011-12	59.7 km

If the deterioration curves are accurate, then reapplying the regression curves using the above lengths should give modelled bursts that closely follow the gradient of the observed failures.

The following chart shows the results of this backcast analysis:



The modelled outputs have been scaled upwards to meet the same level as the observed trend so that the level of error between the trend line and the models can be ascertained. This is important because the accuracy of the deterioration gradient is difficult to assess without the scale matching.

The unscaled output is low because the backcast has been analysed using less data than the total asset stock. This was necessary because of difficulties reconstructing strings for the data which had been decommissioned during the 1996-2012 period (decommissioned assets have no stored connectivity on the GIS). It is therefore impossible to apply the models to this data effectively because of the string/scheme selection algorithm relying on the connectivity to 'stitch' string together – and it was considered more appropriate to leave this data out of the backcast calculations.

Allowing for the obvious scale issues, the backcast analysis has generated a reassuring set of results, showing that the gradient of the modelled output (bright red line) is very close to that of the observed failures during the same period. Scaling the modelled output to achieve a minimum average root squared error means that the average model error when compared against the observed trend (black line) across the 16 years is 0.65%.

8.3 Sensitivity checks

The model application relies on three main parameters, minimum scheme size (m), maximum scheme size (m) and minimum allowable scheme failure rate (bursts/km). As should be expected, varying these settings does affect the nature of the selected schemes, which in turn affects the number of bursts forecast - i.e. lowering the minimum scheme size results in a 'hot spotting' programme of work – lots of short schemes with high burst savings per km.

The default parameters have been set to reflect the mix of work that SSW has undertaken in recent years. This has proven efficient for contract management, because the number and duration of schemes is practical and allows for reasonable deployment charges. Doing too many schemes would increase costs, but doing fewer longer schemes would reduce the bursts savings.

As shown in section 8.2, the default parameters seem to accurately reflect the gradient of the observed deterioration/improvement for known renewal lengths. This section explores the variance introduced by varying the parameters away from the default positions.

The default parameters are:

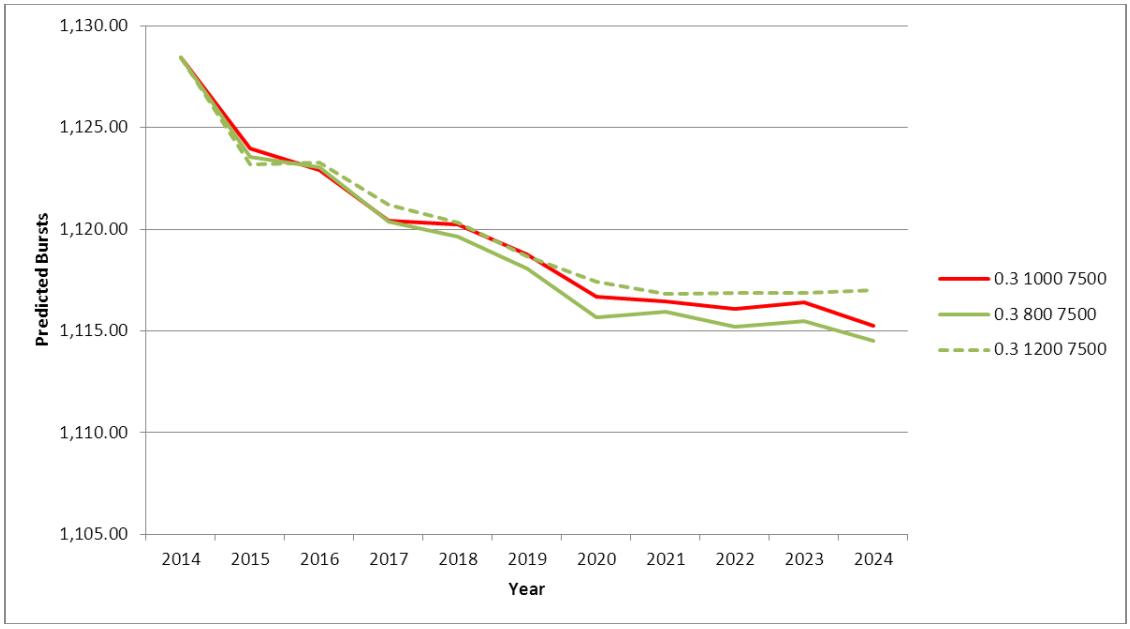
- Minimum scheme size = 0.8km
- Maximum scheme size = 7.5km
- Minimum scheme failure rate = 0.3 b/km

8.3.1 Minimum scheme size

Varying the minimum scheme size has a small effect of the predictions made by the model. In the following example the minimum size was increased and decreased by 20% (200m) away from the default setting of 1km.

Cumulative difference in predicted bursts over 10 years:

Test	Bursts	% Difference in bursts
Control – 1km	14,535.11	
Decrease to 800m	14,529.42 scenario saves 5.7 additional bursts	-0.04%
Increase to 1,200m	14,539.55 scenario saves 4.4 less bursts	+0.03%

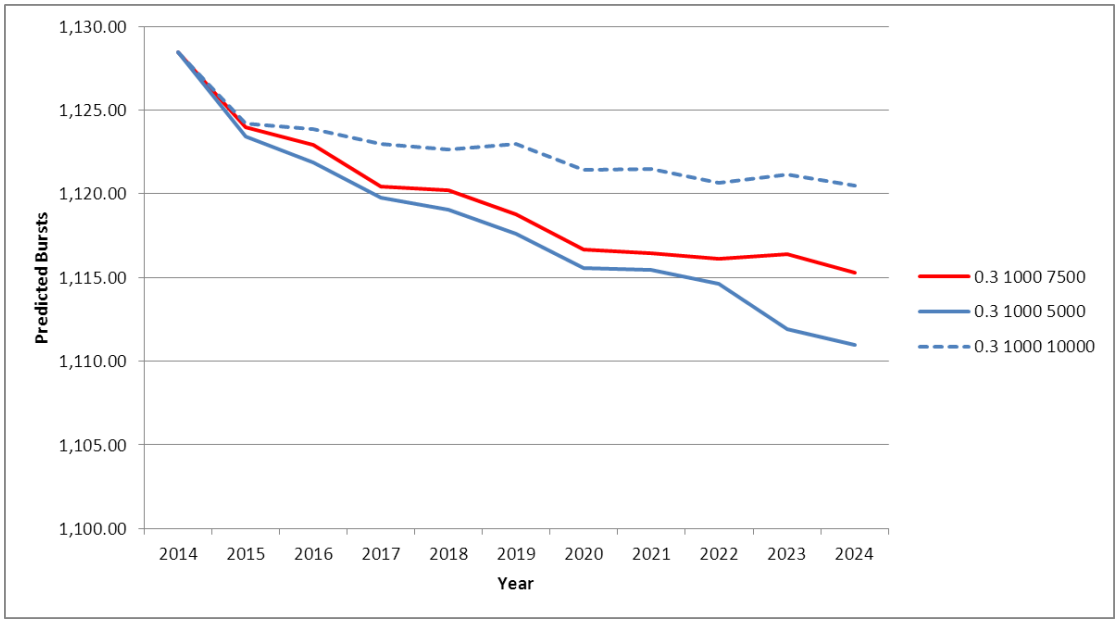


8.3.2 Maximum scheme size

Varying the maximum scheme size also has a fairly minor effect on the predictions made by the model. In the following example the maximum size was increased and decreased by 2.5km away from the default setting of 7.5km.

Cumulative difference in predicted bursts over 10 years:

Test	Bursts	% Difference in bursts
Control – 7.5km	14,535.11	
Decrease to 5km	14,518.15 scenario saves 16.96 additional bursts	-0.12%
Increase to 10km	14,569.90 scenario saves 34.8 less bursts	+0.24%

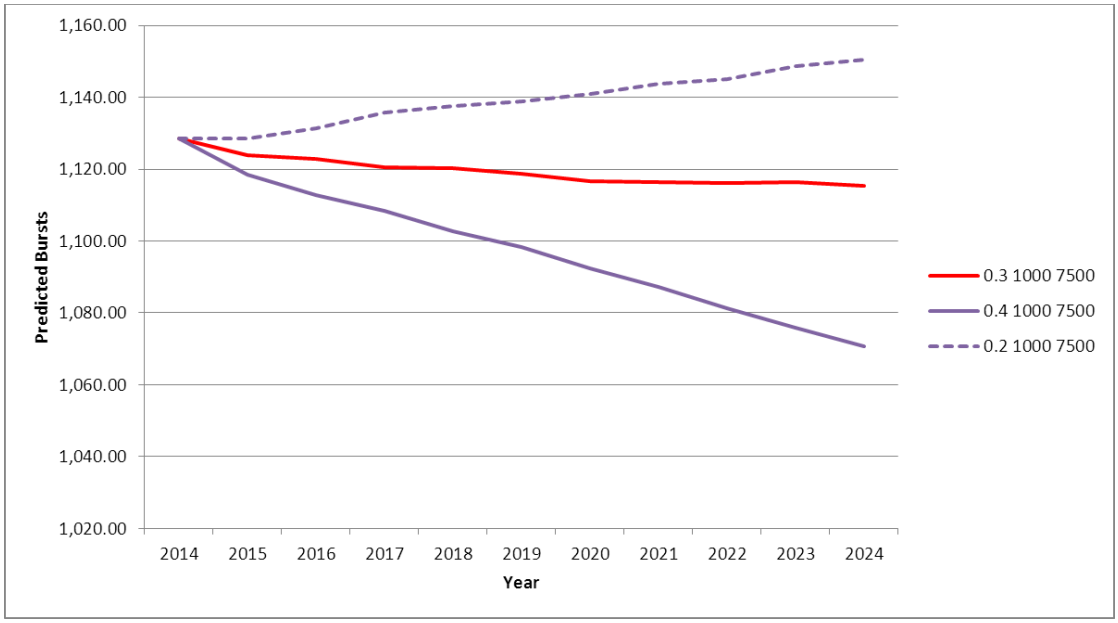


8.3.3 Minimum allowable failure rate

Varying the minimum allowable failure rate for a scheme has a much larger effect on the predictions made by the model. In the following example the rate was increased and decreased by 0.1b/km/yr away from the default setting of 0.3.

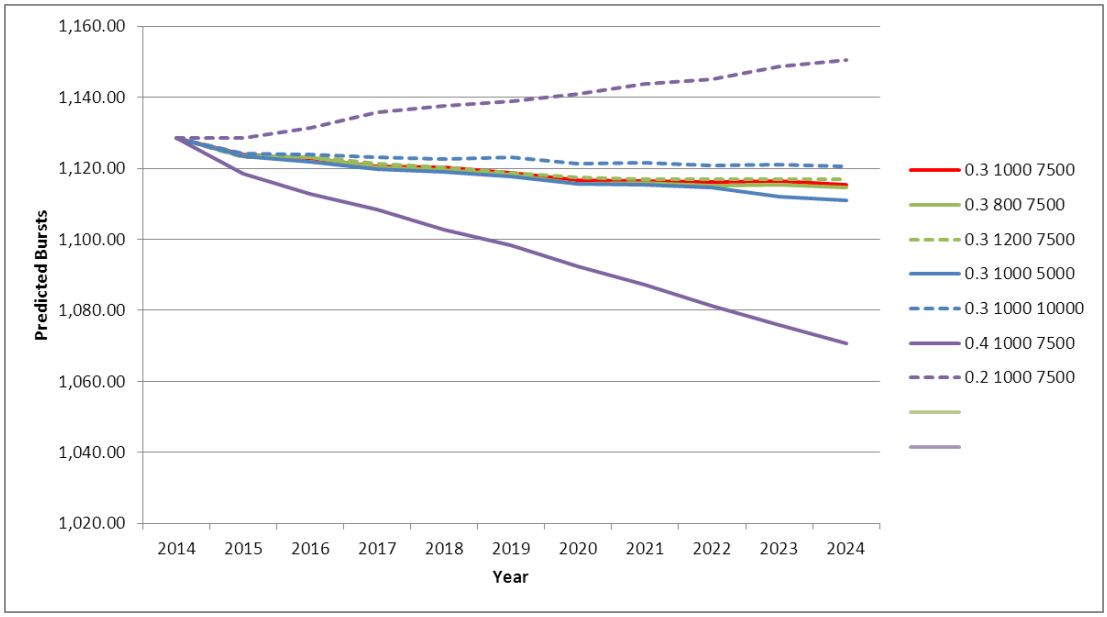
Cumulative difference in predicted bursts over 10 years:

Test	Bursts	% Difference in bursts
Control – 0.3b/km/yr	14,535.11	
Decrease to 0.2b/km/yr	14,748.96 scenario saves 213.85 less bursts	+1.47%
Increase to 0.4b/km/yr	14,296.35 scenario saves 238.76 additional bursts	-1.64%



8.3.4 Summary

As expected, varying the allowable failure rate has the largest impact on the predictions made by the model:



It should be noted that the axis scale used on the above graph exaggerates the difference introduced by varying the parameters. The largest cumulative difference predicted across the 10 year test was just 1.65% of the total bursts.

It should also be noted that varying the parameters always has the expected effect i.e. lowering the allowable failure rate generates less schemes of longer average scheme length, whilst lowering the minimum scheme length generates a higher number of schemes with very efficient burst savings etc.

8.4 Benchmarking against other companies

< insert section here showing comparative renewal rates from JR datashares >
< include section stating higher than industry average deterioration >
< any solid evidence of network sensitivity to weather? >

8.5 Expert review and critical appraisal

The following sections cover the main points raised from each of the reviews. Full versions of these reports are included within appendices.

8.5.1 Mott MacDonald – Statistics Review

Key points:

- "the Water Statistics User Group's (WSUG) *Good Practice Guide* to statistical analysis approaches has been followed"
- "3 Assumed models have been created for MDPE and HPPE for future deterioration as both currently show an improving failure rate (perhaps the initial downward trend of a bathtub curve) and there is no data on the deterioration phase of these pipes. A conservative rate of deterioration has been assumed. Again this will make little impact on modelling as these pipes are unlikely to be replaced in the near future as they are newer and more reliable."
- "The number of models/cohorts of pipes has been derived statistically (using SPSS) and with engineering judgement."
- "The deterioration models are then used in an investment tool to predict the effect different renewal strategies will have on the future burst rate. The deterioration modelling and investment tool accomplish SSWs initial aim of deriving new models and linking top-down and bottom-up analysis."
- "This is a good first step for in-house data derived deterioration models for SSW."

Future improvements:

- "No analysis of variability/uncertainty in the modelled burst rate has been undertaken, I believe this is not necessary at present for SSWs initial goals and how the models are to be used but this should be the next step of analysis. In many models there is a large range of uncertainty around the fitted curve."
- "Related to the above point GLM models using different probability distributions may better fit the data as well as model uncertainty, this is a more complex approach that should be explored in the future but the current approach is a very good first step."

8.5.2 Seams Analytics – Approach Review

Key points:

- "In summary, a relatively high score is given *[for descriptive analytics]* which represents the good quality data that SSW holds on their assets and bursts. The score is tempered somewhat by the data used in other model components, for example, costs and consequence. After discussions with SSW it was noted that in this first evolution of the model it was not designed to be all inclusive of many KPI and that this will be built upon in business as usual throughout AMP6."
- "It is clear where the priorities within the model development have been, with all data, analytics and validation around bursts scoring very highly (8+). The SPSS analysis was excellent and a number of investment scenarios have been carried out utilising an intelligent scheme prioritisation method."
- "The key weaknesses lie in the lack of risk and the simple consequence model but this was a known issue when discussed with SSW. This can be remedied throughout AMP6 by developing an appropriate risk algorithm for inclusion within the model."
- "In summary, the model is an excellent representation of bursts and gives SSW some sound investment planning scenarios for PR14."

Future improvements:

- "*[It is]* highly recommended that a robust sensitivity analysis is carried out to evaluate the level of certainty in the data and how important changes in the inputs are to the outputs."
- "There are a number of quick win improvements that SSW can carry out whereas others will be more complex and difficult."
- "It is recommended that more complex interruptions modelling be carried out. At the minimum the number of customers interrupted can be analysed by diameter. The more complex option would be to incorporate the outputs of hydraulic models to understand how many customers will be interrupted if a given pipe bursts. This can then be woven into the overall water infra model and used as part of investment scenarios."
- "As the model becomes more complex and different KPI are added, then additional interventions should be considered. For example, the effect of flushing on the number of discolouration events can be quantified via a CUSUM analysis and this could potentially be added to the model as a new intervention."
- "it is highly recommended that a thorough review of the costs be carried out in the next evolution of the model and that a greater amount of granularity be applied."; "Adding cost into the model as a common currency becomes increasingly important if the model evolves to include multiple serviceability measures and SSW wants to run scenarios that minimise the WLC of the network."
- "Section 3.2 provides details on how a risk algorithm could be included within the water infra model and how it would benefit the investment scenarios. This should be considered a high priority change to the model to ensure it is more risk focussed as opposed to driven by bursts."

8.5.3 Birmingham University

8.5.4 Summary of comments

Overall feedback about the deterioration model has been quite positive. It seems that the version of the model assessed is deemed fit-for-purpose for SSW's strategic asset management requirements.

There are numerous suggestions for future improvements, particularly from Seams. It should be noted that this was a requested part of the review, to gain an external opinion of areas which would benefit from further development. SSW intends to carry out continuous development of both the deterioration model and the processes for applying the model; the feedback from the reviews will help to target this development more effectively.

9. Communicating results and identifying improvements

- Present to network strategy and production strategy teams followed by critical assessment of methods
- Document potential future developments – i.e. with improved data…

9.1 Presentation of results

9.2 Future improvements

9.3 Action plan