

Estimating risk through data

Richard Jarrett
Stream Leader, Quantitative Risk Management
CSIRO Mathematical and Information Sciences

Outline

Purpose of the study:

- To determine failure rates in water pipes
- To predict forward in time
- To decide appropriate $ for replacement each year
- To compare various replacement strategies
- To determine which pipes to replace each year

We will look at:

- “Risk”
- Getting the data “right”
- Getting the modelling “right”
- Creating “value” for the client

How is “Risk” defined here?

Risk:

- Measured in terms of a combination of
  - the likelihood of an event, and
  - the consequences of that event

In this case, it is represented as

$$\textbf{Risk} = \Sigma_{\text{events}}\ \textbf{Probability} \times \textbf{Consequence}$$

AS/NZS 4360:2004

- We will deal today mainly with the “probability”
- Then combine it with the “consequences”
- Aim to “minimise expected costs”

The challenge

Water distribution systems in major cities

We need to know

- How much to spend of replacing, rather than just repairing, pipes that fail?
- Which pipes to replace and when?

Traditional approach

- Cohort of pipes
  - define a “lifetime” and replace at end of useful life.
- Individual pipes
  - wait till each pipe “goes bad” and replace it,
  - for example, if ”3 failures in a year”, replace

Getting the data right

"Data collected for other purposes"…

Data is generally from two sources:

- Asset database, generally in a GIS system
  - so we know where pipes are,
  - which valves to turn off

- Failure database, generally simple database
  - starts in a call centre
  - is completed by field workers after the pipe is repaired

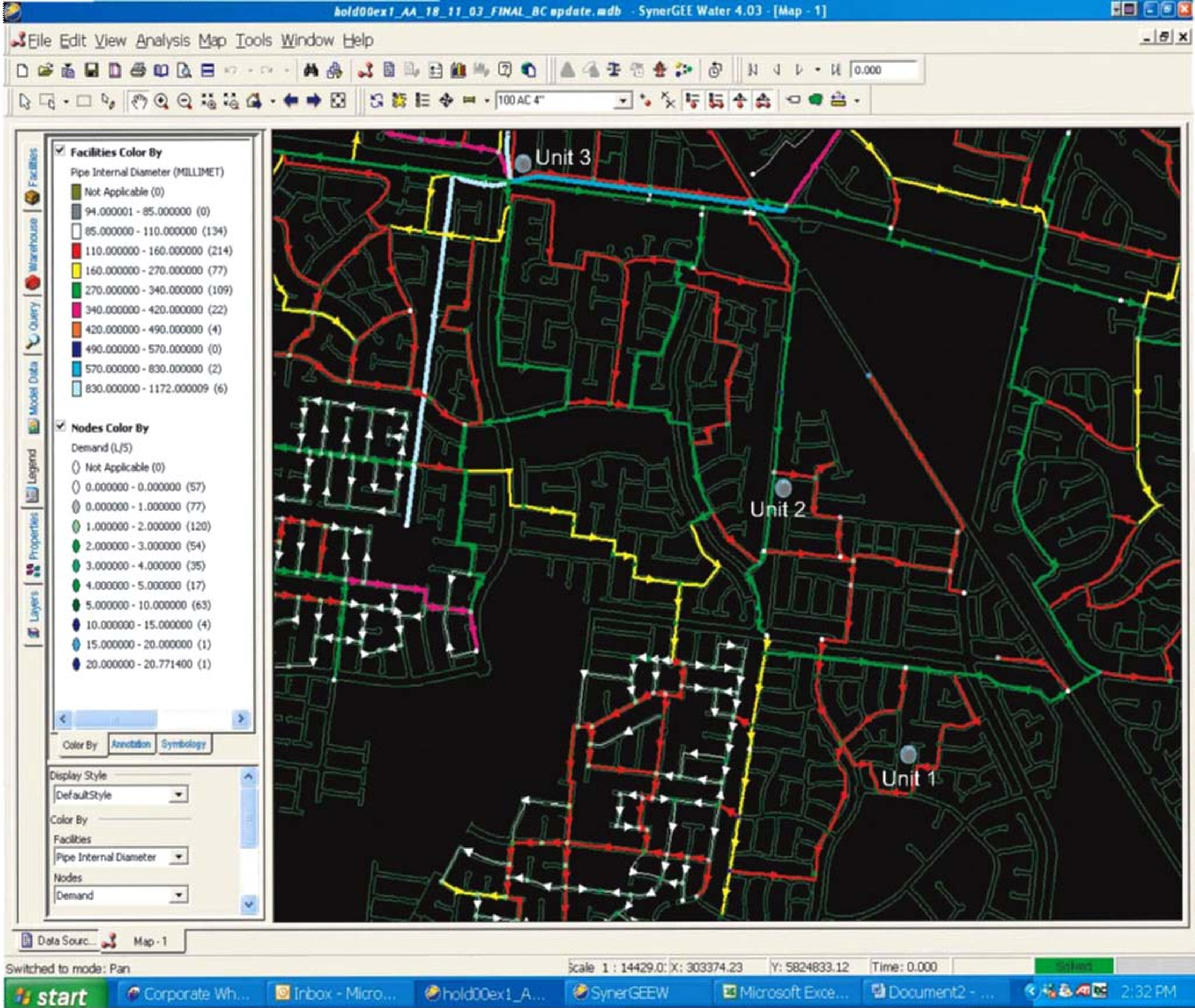
Data issues: Assets

Assets typically have:

- Material
- Diameter
- Length
- Date laid/abandoned/rehabilitated
- Pressure, Soil, Traffic condition, …

Problems:

- Missing fields, particularly older pipes
- “Pretend” pipes added to get connectivity



hold00ex1_AA_18_11_03_FINAL_BC update.mdb - SynerGEE Water 4.03 - [Map - 1]
File Edit View Analysis Map Tools Window Help
Facilities Color By
Pipe Internal Diameter (MILLIMET)
Not Applicable (0)
94.000001 - 85.000000 (0)
85.000000 - 110.000000 (134)
110.000000 - 160.000000 (214)
160.000000 - 270.000000 (77)
270.000000 - 340.000000 (109)
340.000000 - 420.000000 (22)
420.000000 - 490.000000 (4)
490.000000 - 570.000000 (0)
570.000000 - 830.000000 (2)
830.000000 - 1172.000009 (6)
Nodes Color By
Demand (L/S)
Not Applicable (0)
0.000000 - 0.000000 (57)
0.000000 - 1.000000 (77)
1.000000 - 2.000000 (120)
2.000000 - 3.000000 (54)
3.000000 - 4.000000 (35)
4.000000 - 5.000000 (17)
5.000000 - 10.000000 (63)
10.000000 - 15.000000 (4)
15.000000 - 20.000000 (1)
20.000000 - 20.771400 (1)
Color By
Annotation
Symbology
Display Style
DefaultStyle
Color By
Facilities
Pipe Internal Diameter
Nodes
Demand
Unit 3
Unit 2
Unit 1
Data Sourc...
Map - 1
Switched to mode: Pan
Scale 1 : 14429.0
X: 303374.23
Y: 5824833.12
Time: 0.000
start
2:32 PM

Data issues: Failures

Failure data base typically has:

- Failure date
- Location, often street address
- Failure type, failure cause

Problems:

- Typically details are incomplete
- A proportion of failures cannot be matched to assets (10-40%)
- Recording rate may vary over time
- Relatively short period of matched data

Data issues: organising the data

For each failure:

- Match each failure to an asset, where possible
- Determine "match rate" for each year
- Determine "recording rate" for each year (except for first and last)

For each asset:

- Determine number of failures for each year in the period

Year Laid	Fail Year					Tot	Len
	1992	1993	…	2000	2001		
Pipe 1	1	0	…	0	1	2	
Pipe 2	3	1	…	2	3	9	
…	…	…	…	…	…		
Pipe n-1				4	0	4	
Pipe n					2	2	

Getting the modelling right

Darroch and Constantine (1990s):

- Nonhomogeneous Poisson process (NHPP) where failure rate increases with age, with cumulative intensity function for ith asset at age t :

$$H_i(t) = L_i \exp(\alpha' x_i)\, t^{\beta}$$

where L_i = length, β =shape parameter, x_i =covariates.

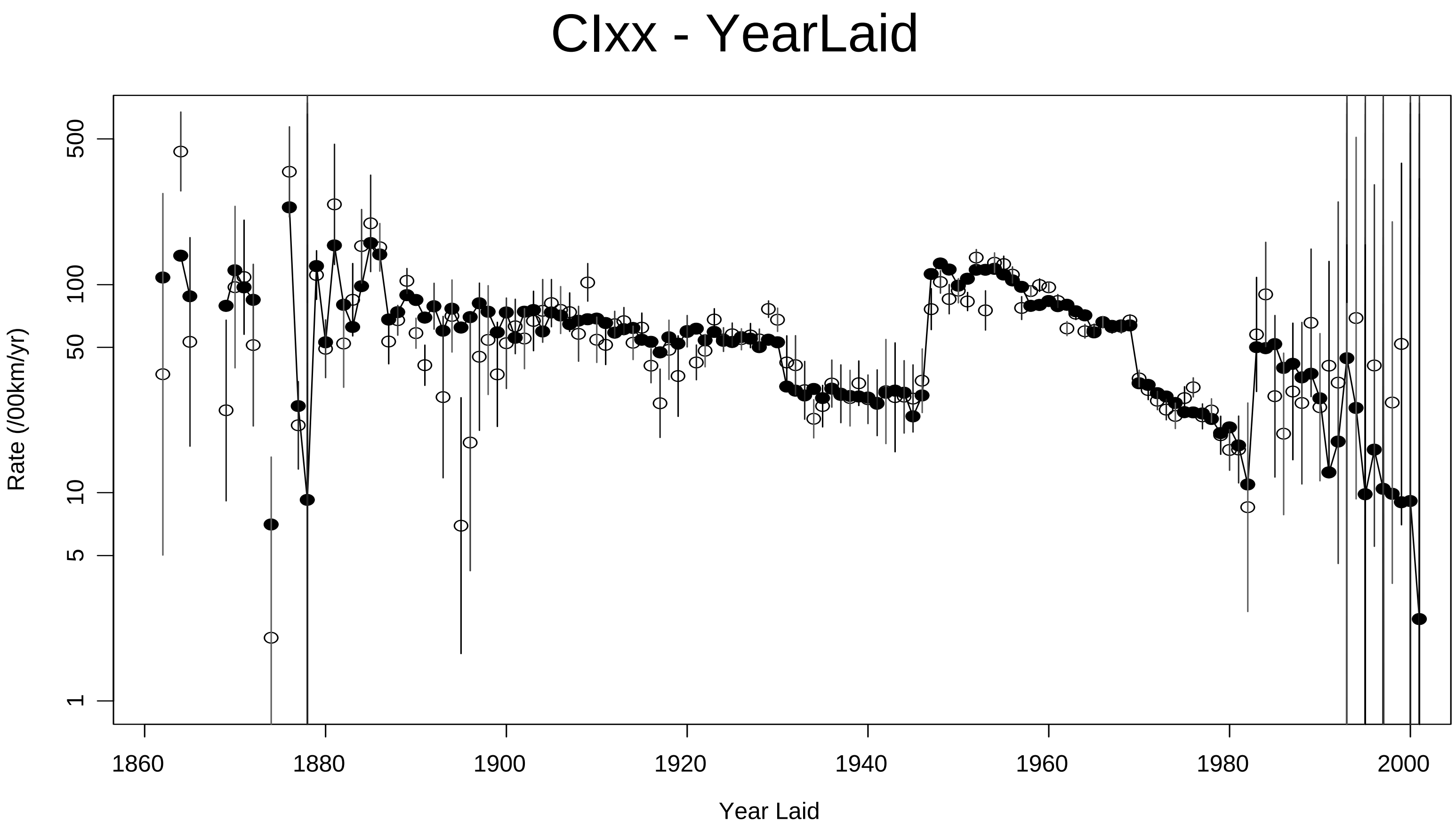
- covariates were just material, diameter.

When we reviewed the methodology in 2000, we found:

- forward predictions not ‘realistic’
- no real ‘goodness of fit’ measures
- material properties varied according to date laid
- “early failures” change shape of age curve
- reporting/recording/matching rates were not consistent over time

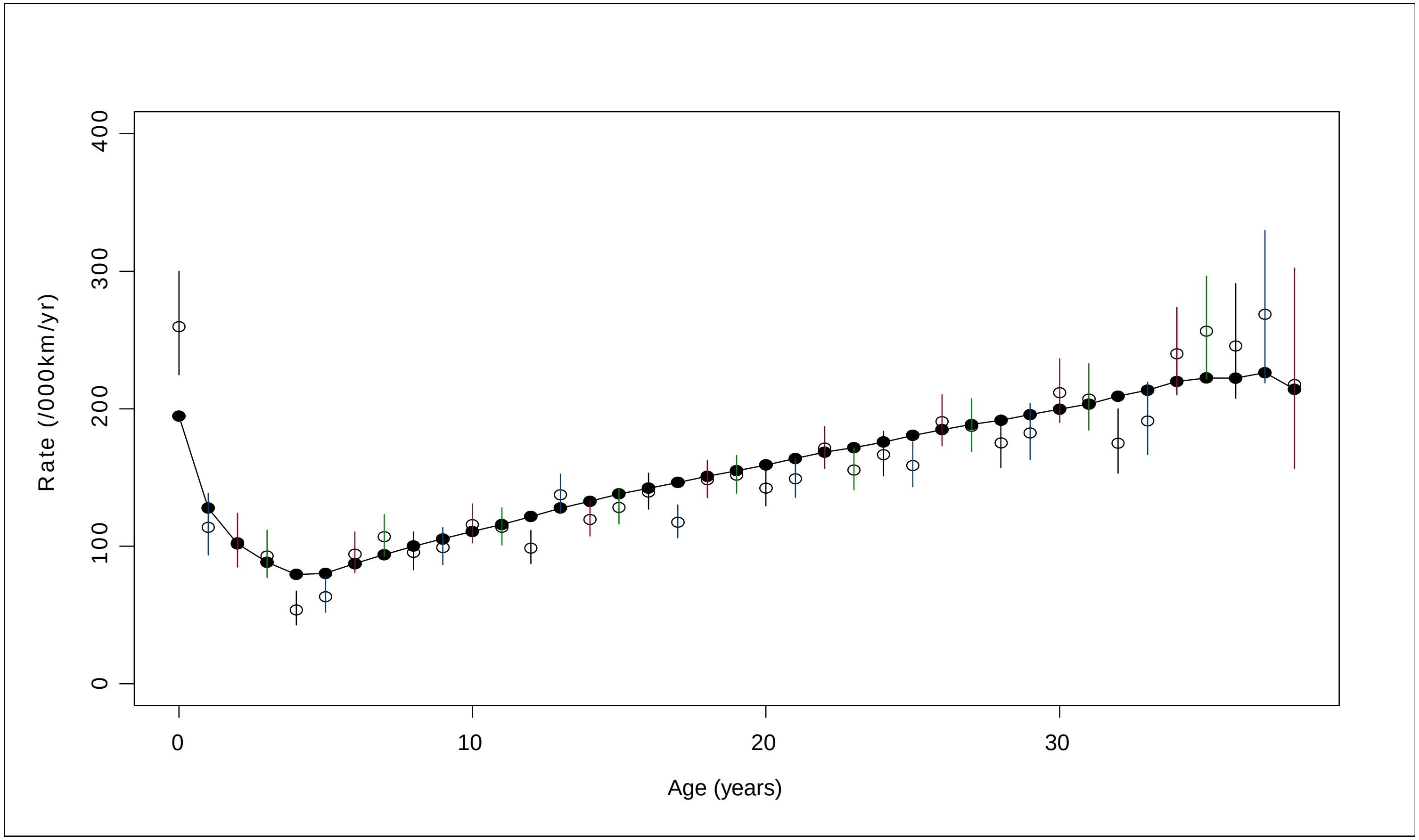
Failure rate vs Year Laid

- Failure Rate against Year Laid for cast iron pipes
- Material properties clearly vary with Year Laid



Early failures

- Failure rate vs Age for one material
- Problems through first 4 years.



A revised model

The final model takes into account many issues including:

- Matching rates, recording rates,
- Identifying sets of “Year Laid” whose assets have higher or lower rate of failure
- Burn-in problems (early failures)
- Goodness of fit, and allowance for extra-variation
- Incorporating spatial information

"Expected" failures

For asset *i,* the number of failures we see in year *j* is Y_{ij}, which we suppose initially is Poisson with mean

$(t_{ij} + 0.5)^{\beta - 1}$

Fail Year effect

$$\mu_{ij} = L_i^{\theta}\, e^{\alpha' x_i + \phi_j + z_j} \{(t_{ij} + 1)^{\beta} - t_{ij}{}^{\beta}\}$$

Length

Effect of age

Effect of attributes:

Material, Diameter, Soil, Pressure..

Proportion matched or recorded

Age-period-cohort designs

Data can be represented in an (incomplete) two-way table

- Note that

$$\text{Age} = (\text{Fail Year} – \text{Year Laid})$$

and the 3 linear effects of these are confounded.

Year Laid	Fail Year					Tot	Len
	1992	1993	…	2000	2001		
1881	1	0	…	0	1		
1882	3	1	…	2	3		
…	…	…	…	…	…		
2000				4	0		
2001					2		

Likelihoods

How do we fit the model? There are typically 100,000 pipes x 10-15 years of record

Marginal likelihood: total failures for pipe *i*

$$Y_i \sim Poisson\ [\ L_i^{\theta} exp(\alpha' x_i)\ \Sigma_j\ f_j\ \{(t_{ij}+1)^{\beta} - t_{i,j-1}{}^{\beta}\}]$$

where $f_j = exp(\phi_j + z_j)$.

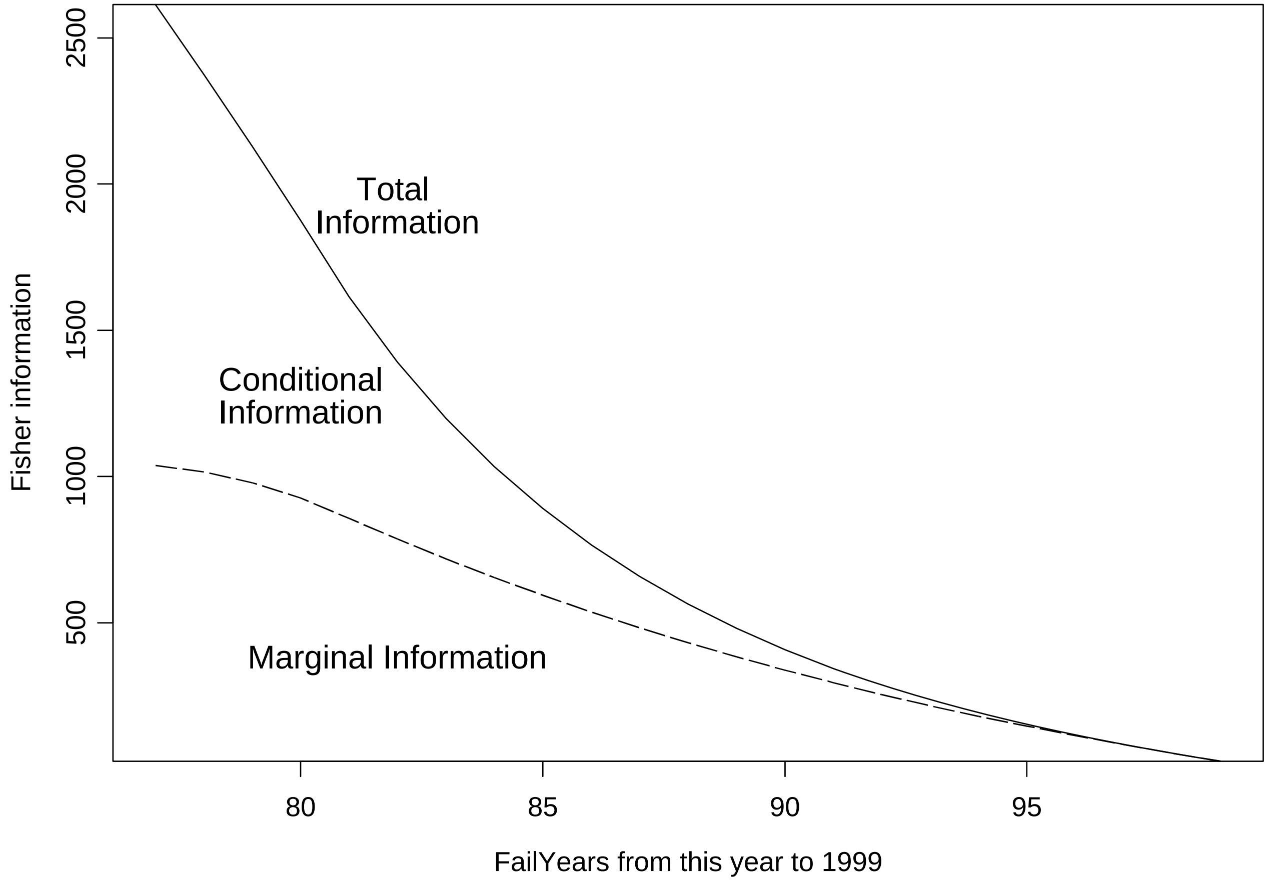
Conditional likelihood: where do the failures occur in time?

$$Y_{ij}\ |Y_i \sim Multinomial[Y_i\ ;\ f_j\{(t_{ij}+1)^{\beta} - t_{i,j-1}{}^{\beta}\}\ /\ \Sigma_k\ f_k\ \{(t_{ik}+1)^{\beta} - t_{i,k-1}{}^{\beta}\}]$$

for those with $Y_i > 0$.

Can estimate β from the conditional distribution, then use that value as given in the marginal distribution. Is there information about β in the marginal likelihood?

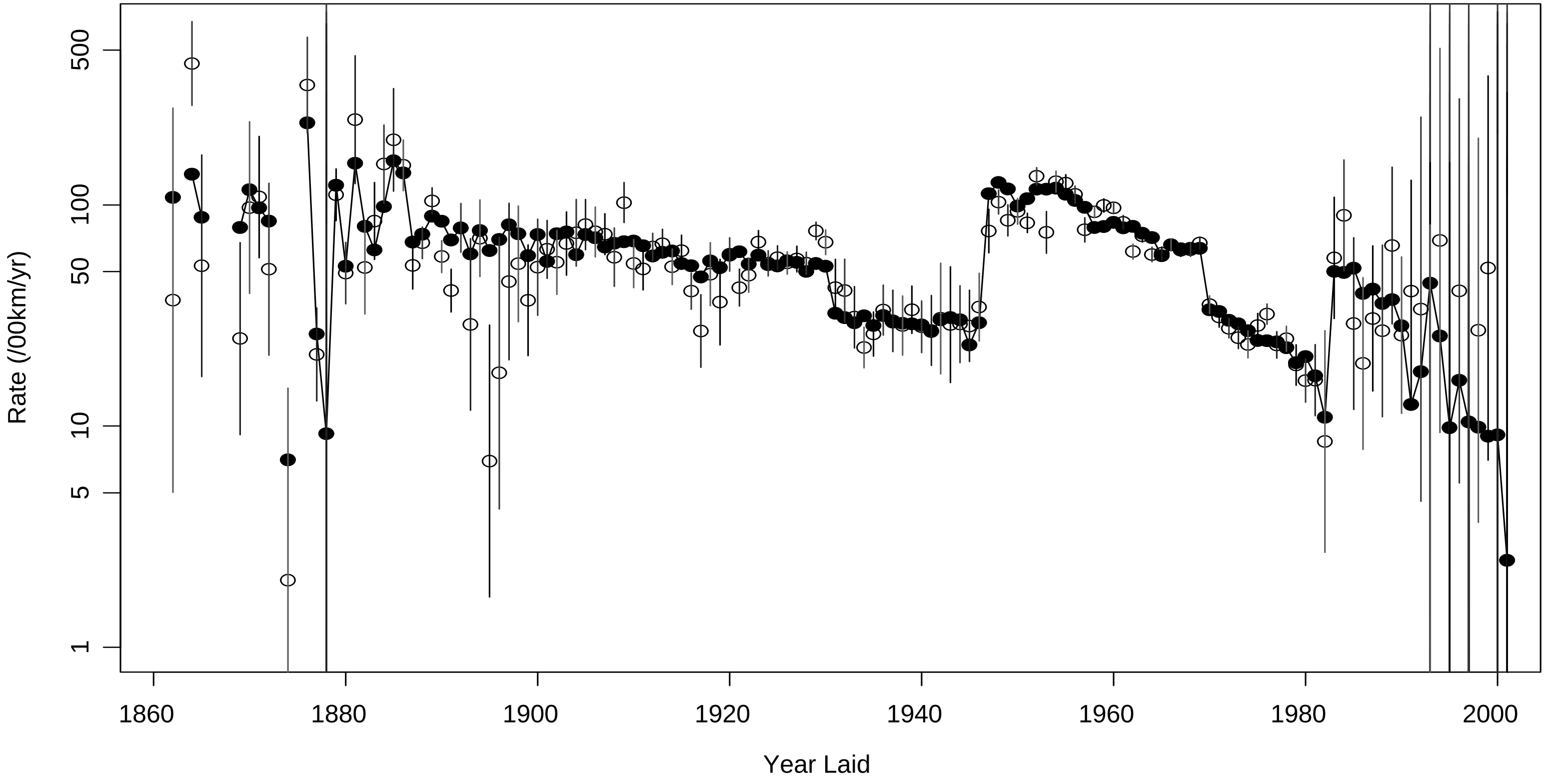
Is there information about β in the marginal likelihood?



Failure rate vs Year Laid

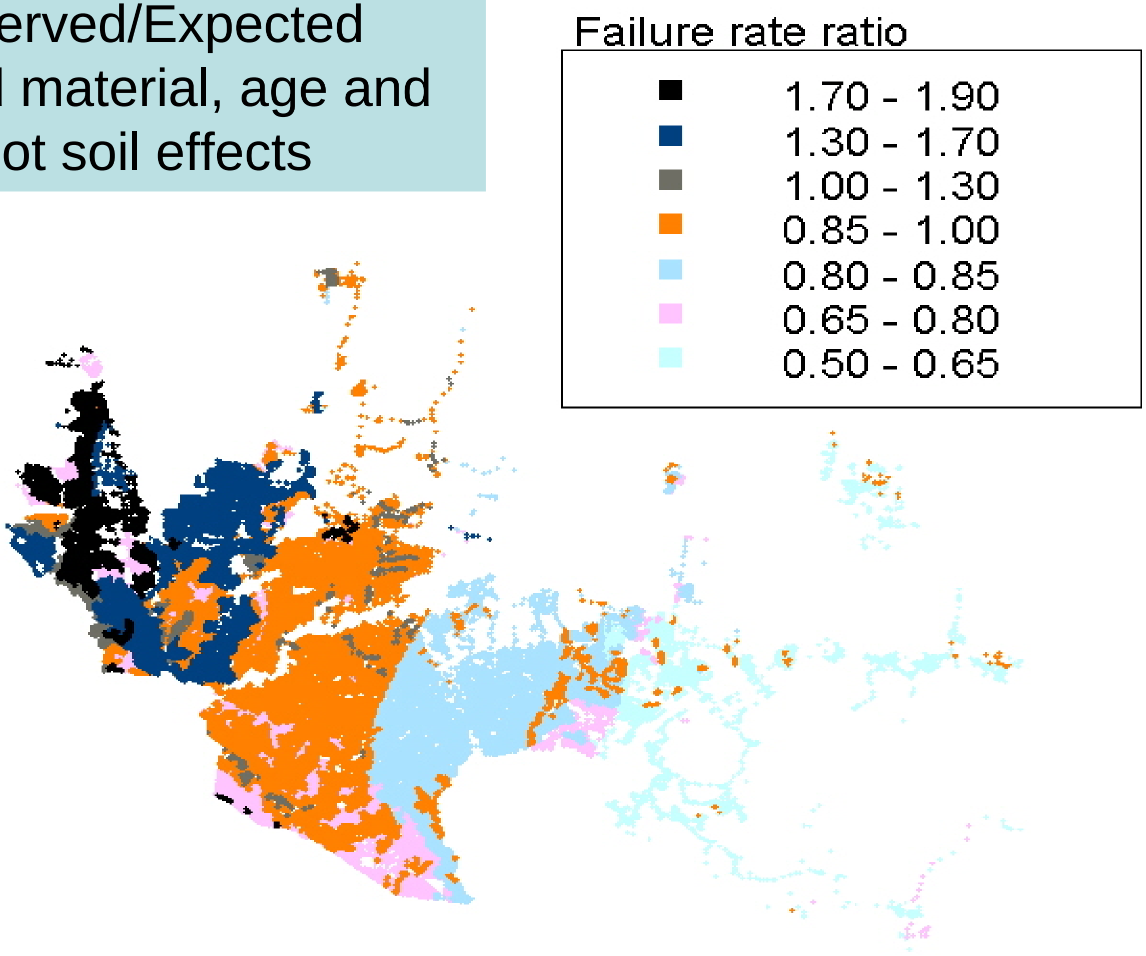
- Failure Rate against Year Laid for cast iron pipes
- Model now captures the step changes in quality of cast iron pipes

CIxx - YearLaid



Soil and pH effects

FR ratio = Observed/Expected
Model included material, age and diameter, but not soil effects



How well does this do?

Residual deviance << DF for both the marginal and conditional likelihoods, but that is not enough!!

- Pearson Chisquare shows ~15% extra-variation

	DF	Pearson Chisq	Mean PChisq
Between assets	91028	172966	1.9000
Within assets	665685	742675	1.1157
Total	756721	915641	1.2100

- Poisson predicts too few zeros and not enough multiple failures

This causes problems in delivering to the client:

- We need to identify which pipes are “worst” in order to find those that should be replaced
- We need to predict their failures in an unbiased way if we are to predict the gains to be made from replacement
- Current replacement strategies are often based on “replace any pipe which has 3 failures in a year”

"Predicted" failures (BLUPs)

Use a "random effects" model to deal with extra-variation

1. Random effect for each pipe *i* :

- Suppose there is an unobserved "z_i" which is Gamma with mean 1, variance $1/\gamma$, and then we observe Poisson($\mu_i z_i$)
- Unconditionally, the moments are:

Count	Poisson Model		One random effect	
	Mean	Variance	Mean	Variance
Y_i	μ_i	μ_i	μ_i	$\mu_i(1+\mu_i/\gamma)$
Y_{ij}	μ_{ij}	μ_{ij}	μ_{ij}	$\mu_{ij}(1+\mu_{ij}/\gamma)$

- The conditional distribution of Y_{ij} given Y_i is still multinomial

"Predicted" failures (BLUPs)

2. Random effect for each pipe *i* in each year *j*:

- Suppose there is an unobserved "z_{ij}" which is Gamma with mean 1, variance $1/\omega$, and then we observe Poisson($\mu_{ij} z_{ij}$)
- Unconditionally, the moments are:

Count	Poisson Model		Two random effects	
	Mean	Variance	Mean	Variance
Y_i	μ_i	μ_i	μ_i	$\mu_i\{1+\mu_i/\gamma+\sum_j \mu_{ij}^2/(\mu_i\omega)\}$
Y_{ij}	μ_{ij}	μ_{ij}	μ_{ij}	$\mu_{ij}(1+\mu_{ij}/\gamma+\mu_{ij}/\omega)$

- The conditional distribution of Y_{ij} given Y_i is no longer multinomial

"Predicted" failures (BLUPs)

We define $1/\gamma$ and $1/\omega$ relative to the underlying expected number of failures $\xi_{ij} = \mu_{ij} / f_j$; that is, before it gets reduced by matching and recording losses, and adjusted by FailYear effects.

If we assume that the extra-variation is a constant multiple, we get:

$$1/\gamma = 0.0902/\xi, \quad \text{and} \quad 1/\omega = 0.1129/\xi.$$

- This gives:

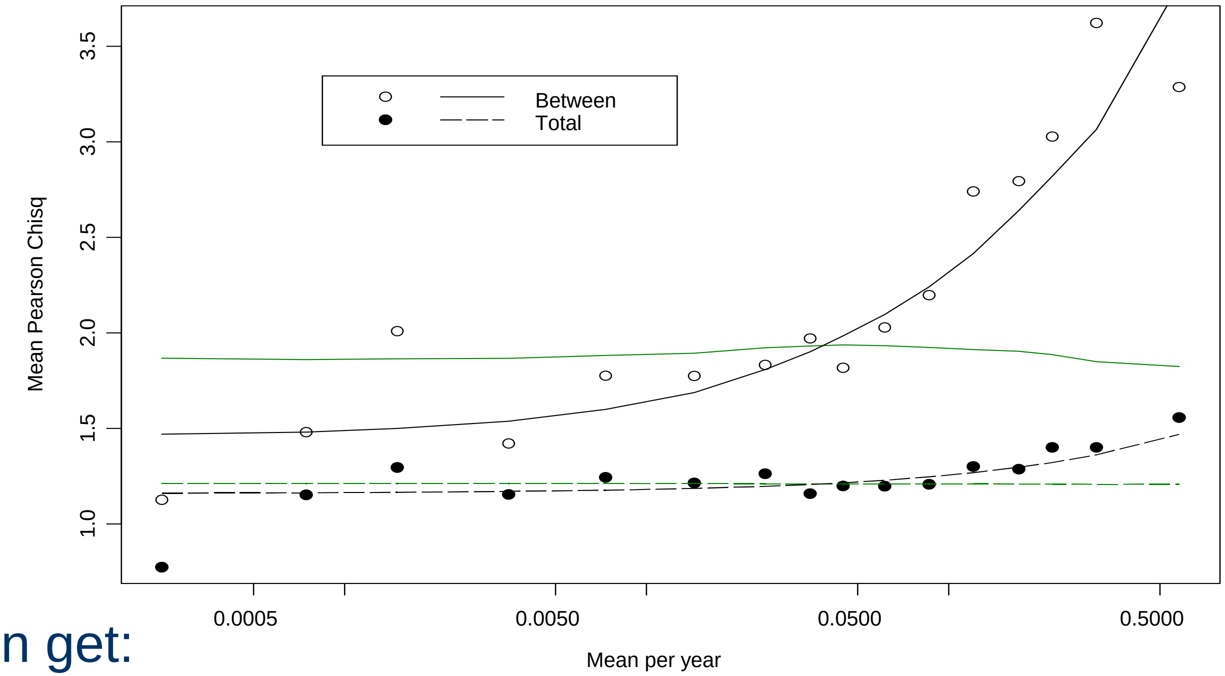
$$Var(Y_{ij}) = \mu_{ij} (1+0.2031\, f_j),$$

while, if a pipe is present for all 9 years, we get:

$$Var(Y_i) = \mu_i (1.9065)$$

However, the Pearson Chisquare shows that it is not constant.

"Predicted" failures (BLUPs)



We then get:

$$1/\gamma = 0.04(1+11\xi^{2/3})/\xi, \quad \text{and} \quad 1/\omega = 0.1129/\xi$$

- This gives

$$Var(Y_{ij}) = \mu_{ij}\{1+0.1529 f_j + 0.44 f_j \xi_i^{2/3}\}.$$

while, if a pipe is present for all 9 years, we get:

$$Var(Y_i) = \mu_i(1.494+4.086\,\xi_i^{2/3})$$

"Predicted" failures (BLUPs)

Expected value for a past year:

$$\log(\mu_{ij}) = \theta \log(L_i) + \alpha' x_i + \log[f_j \{(t_{ij}+1)^{\beta} - t_{ij}^{\beta}\}],$$

- which is the number expected allowing for matching, recording and FailYear effects across the years 1994-2002.

- If we suppose 100% matching/recording and an "average" FailYear, we can remove f_j.

Expected number of failures for a future year, ξ_{ij}, is given by

$$\log(\xi_{ij}) = \theta \log(L_i) + \alpha' x_i + \log[\{(t_{ij}+1)^{\beta} - t_{ij}^{\beta}\}]$$

”Predicted” failures (BLUPs)

Predicted number of failures:

- This is a best linear unbiased predictor (BLUP), essentially a posterior mean given the data.

$$BLUP_{ij} = \xi_{ij} \times \frac{1+\xi_i B_i/\omega + \boldsymbol{y}_i/\gamma}{1+\xi_i B_i/\omega + \xi_i A_i/\gamma}$$

where $A_i = \Sigma f_j$ and $B_i = \Sigma f_j^2/A_i$, where the sum is over those years where the *i*th asset is present in the data.

Delivering value for the client

Water utilities want a management tool:

- How much should we spend on asset replacement?
- Where should we spend it?
- What are good strategies to determine what to replace?
- How do we combine the probability models with the consequences/costs?

Consequences

For each failure, we can

- Determine cost of repair
- Develop “social costs”, such as disruption to supply, loss of business, traffic delays,… ,

and hence obtain

$$\textbf{Risk} = \Sigma_{\textbf{events}}\ \textbf{Probability} \times \textbf{Consequence}$$

We can then consider

- Strategies for replacement
- Costs of replacement/repair

and hence compare different strategies

How do our predictions perform?

For 2002, we looked at two ways in which one might have replaced 9.5 km of assets:

- Ranking by “number of bursts in 2001”
- Ranking by “Predicted failure rate for 2002”

Option	Obs in 2001	Obs in 2002	Pred in 2002
Replace if ≥3f in 2001	**163**	**25**	**26.0**
Replace if in top 9.5km for Pred FR *	**82**	**43**	**55.7**

* With small penalty against “short” pipes

Lessons:

- Rough and ready rule: 100 years life implies replace 1% p.a.
- We can do a lot: eg reduce pressure by 10m,…
- Statistics can’t help much with new materials…

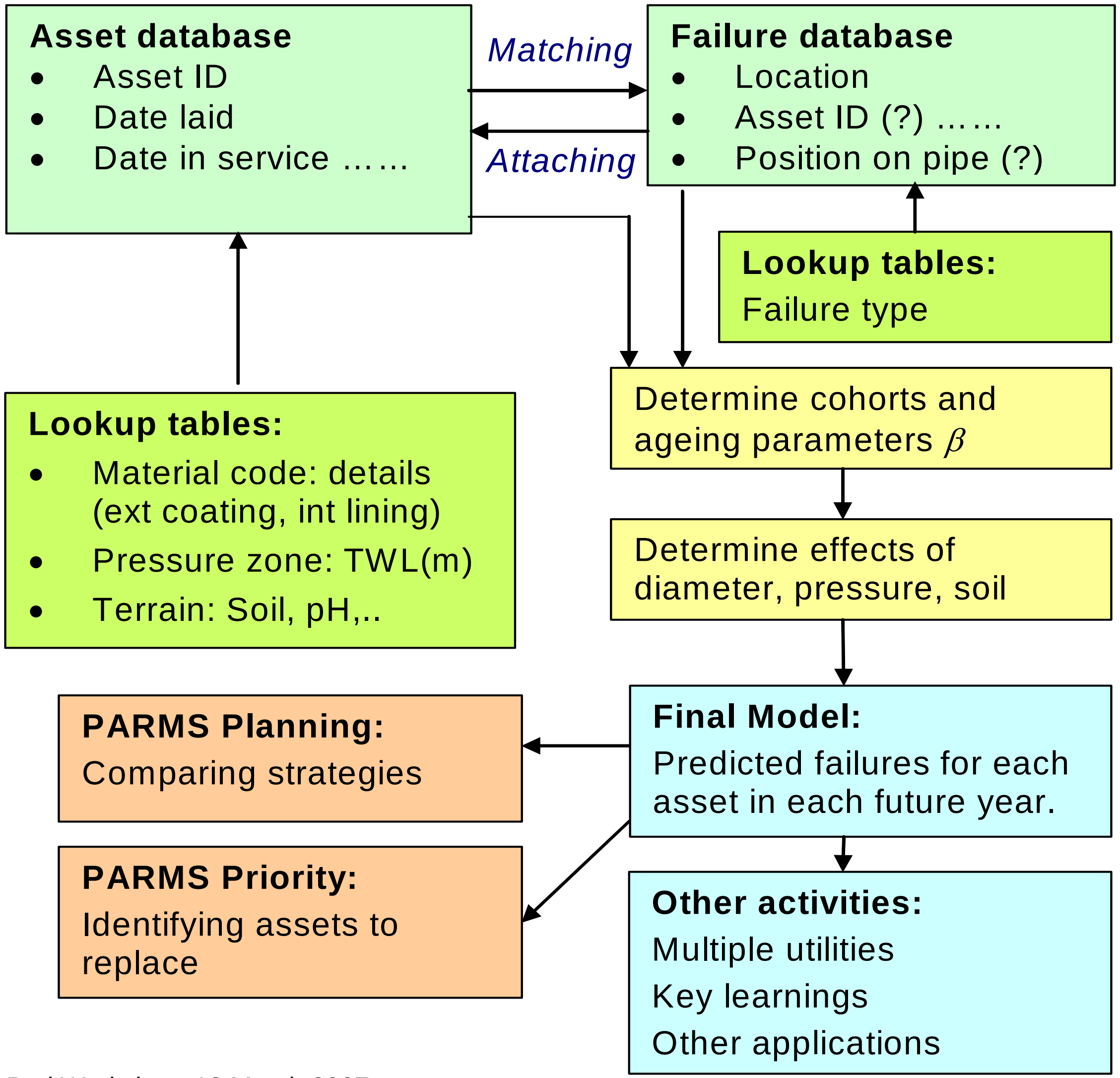
How do our predictions perform?

Compare two strategies:

- Just continue to repair failures as they occur (Old)
- Spend $1M/yr replacing the “worst” pipes with the best of the new materials

Year	E(B.Old)	E(B.New)	d(E(B))	E(BCostOld)	E(BCostNew)	d(E(BCost))
2001	2520.72	2464.8	55.91	3,949,629	3,863,953	85,676
2002	2566.07	2468.3	97.77	4,020,592	3,872,395	148,197
2003	2611.82	2477.31	134.52	4,092,177	3,887,153	205,024
2004	2657.99	2487.91	170.08	4,164,400	3,904,580	259,820
2005	2704.52	2498.94	205.58	4,237,198	3,923,350	313,848
...	...	...	...	...	...	...
2020	3739.48	2891.64	847.84	5,877,212	4,577,438	1,299,774
2021	3819.45	2920.5	898.95	6,004,435	4,625,383	1,379,052
2022	3900.57	2951.91	948.66	6,133,537	4,675,801	1,457,736
2023	3982.85	2980.63	1002.21	6,264,538	4,724,427	1,540,111
2024	4066.3	3011.69	1054.61	6,397,458	4,774,745	1,622,713
2025	4150.93	3041.08	1109.85	6,532,316	4,824,147	1,708,169
				127,631,432	106,825,879	20,805,553

Software called "PARMS":



What have we learnt?

- A third of the time goes in getting the data right
- A third of the time doing the modelling
- A third of the time goes in making it useful for the client

- Models for “risk” for the client need both probabilities and consequences
- A model is only “good” if it makes sense to the client and produces results that they believe

What binds the “risk” work together?

Water distribution systems

- Utilities have “asset classes” (pipes), then model prob of failure, and how it changes over time.
- Modelling done on individual pipes, with a variety of predictors.
- Pipes fail, are “repaired” and may fail again.
- There are costs associated with each failure
- Seasonal and long term weather changes mean that prob of failure changes over time in a rather smooth way.

Finance

- Banks have “asset classes” (accounts), then model prob of failure, and how it changes over time.
- Modelling done on individual accounts, with a variety of predictors.
- Accounts default, are “repaired” and may default again.
- There are costs associated with each default
- Economic cycles imply that prob of default changes dynamically over time.