

General Insurance Seminar

Insuring Tomorrow



**Actuaries
Institute**

17-18 November 2014 • Hilton Sydney



Comparing software options for actuaries

Hugh Miller

© Taylor Fry

*This presentation has been prepared for the Actuaries Institute 2014 General Insurance Seminar.
The Institute Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not
responsible for those opinions.*

Why benchmark software?

The software tools we have available intrinsically affects how we work:

- Our “paradigm” for data management
- Favour types of analyses that are quick & easy
- The efficiency of work
 - Programming time
 - Runtime
 - Degree of automation
- Cost of doing business
- The types of employee skills required

What should we benchmark?

Data management

- Loading
- Cleaning
- Sorting
- Merging
- Summarising

GLMs

- Pricing
- Reserving
- Demand modelling

Statistics/ Analytics capability

- Fast predictive models
 - Decision trees
- Insight generation
- Clustering
- Outlier detection

Types of issues: Speed (computer and human time), functionality, interface & ease of use, interoperability, hardware requirements

Can't I just use Excel?

Maybe, but some significant headwinds:

- Dataset size limitations
- Repeatability and auditability lower
- No in built GLM package, not much high end analytics
- Server operation
- Sharing data in a controlled way is hard

Is it possible to benchmark fairly?

Of course not:

- Familiarity bias
- Hardware complications
- Differing tastes
- All job descriptions are different
- Generic trade off between flexibility and usability
- Commercial bias

But we'll try anyway!

All opinions expressed are those of the author

Notes on hardware

We ran the tests:

- On a desktop
- Intel i7-2600 CPU @3.4GHz (4 cores + hyper-threading)
- 12 GB RAM
- All files on a Solid State drive (SSD), particularly helping SAS & WPS
- Using the in built timers in each piece of software
- >5 mins to complete equivalent to a “did not complete”

Mileage will vary depending on the setup. Most results applicable to server environments



Who should we benchmark?

TEST 1: Data management

- SAS
- WPS
- R
- Revolution
- SPSS
- SQL Database
- Python

TEST 2: GLMs

- SAS (x2)
- R
- Revolution
- SPSS
- Python
- EMBLEM

TEST 3: Statistics/ Analytics capability

- SAS-EM
- R / Revolution
- SPSS
- Salford Systems

SAS

A leader for much of insurance work:

- Designed to deal with large datasets (those too big for RAM)
- Active development stream with recent efforts in:
 - Multi-core computation
 - Visual analytics
- SAS Enterprise Miner (EM) an extra product for analytics, with point and click interface



Feature snapshot	
Popularity	●●●●●
Affordability	●
Enterprise grade	●●●●
Flexibility	●●●●
Data mgmt	✓
GLMs	✓
Analytics	SAS-EM

World Programming System (WPS)

Started as a “clone” for SAS, but has developed into an ecosystem in its own right

- Compatibility with SAS datasets and data steps, plus some procedures
- No GLM procedure
- Legality recently upheld in court against SAS, EU Court of Justice
- The Workbench GUI reviews well



Feature snapshot	
Popularity	●
Affordability	●●●●
Enterprise grade	●●●
Flexibility	●●●
Data mgmt	✓
GLMs	✗
Analytics	✗

R

Leading statistics package in academia

- Open source and free software
- Flexible and concise programming language
- The “package” functionality means that virtually every procedure ever invented is available
- Increasingly popular in non-academic settings



Feature snapshot

Popularity	●●●
Affordability	●●●●●
Enterprise grade	●●●
Flexibility	●●●●●
Data mgmt	✓
GLMs	✓
Analytics	✓

R variants

Revolution analytics (REVO)

“R for business”, includes all of R plus custom data structures and some functions for handling larger datasets and multi-core computing. Includes rewritten GLM package



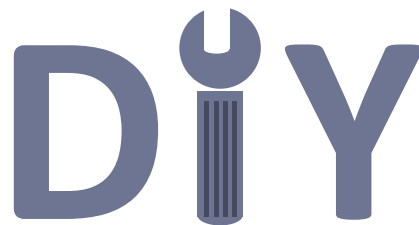
Tibco Enterprise Runtime for R (TERR)

An enterprise version of R with better multiprocessing and integration with other TIBCO software



Custom R

Anyone can write a package, or re-compile the software in ways to suit them



SPSS



Popular statistics package, particularly in social sciences, health and marketing

- Virtually all operations can be performed by “point and click”
- Acquired by IBM in 2009, part of their stable of software offerings

Feature snapshot	
Popularity	● ● ●
Affordability	● ●
Enterprise grade	● ● ● ●
Flexibility	● ● ● ●
Data mgmt	✓
GLMs	✓
Analytics	✓

SQL



Popular database language. It comes in many flavours – we've used Microsoft SQL Server, which comes in free and paid versions.

- Database paradigm encourages storage efficiency
- Requires more consideration of issues such as hardware, data types etc

Feature snapshot	
Popularity	● ● ●
Affordability	Varies
Enterprise grade	● ● ● ●
Flexibility	● ● ●
Data mgmt	✓
GLMs	✗
Analytics	✗

Python



Is a relatively easy and popular programming language

- Is a real programming language – so significantly more daunting to work with
- Many packages for getting things done, but few specifically written for statistics
- Python itself is incredibly broad – we've mainly used the pandas and statsmod packages
- Other languages (eg C/C++ are potentially faster but harder)

Feature snapshot	
Popularity	●●
Affordability	●●●●●
Enterprise grade	●●
Flexibility	●●●●
Data mgmt	✓
GLMs	✓
Analytics	✗

EMBLEM

Leading “Point and click” software for fitting GLMs in insurance, from Towers Watson.

Custom purpose software, and requires steps to get data in and models out.



Feature snapshot

Popularity	●●●
Affordability	???
Enterprise grade	●●●●●
Flexibility	●
Data mgmt	✕
GLMs	✓
Analytics	✕

Salford Systems



Specialist software for non-parametric data mining algorithms. Two of the more popular sub-products:

- CART (decision trees)
- TreeNet (decision trees + boosting)

GUI interface and good SAS integration

Feature snapshot	
Popularity	●●●
Affordability	●●◐
Enterprise grade	●●●◐
Flexibility	●●
Data mgmt	✕
GLMs	✕
Analytics	✓

TEST 1

Data Management

Data management

We tested a routine that included the following tasks:

- Read in data from CSV
- Two merges (one easy, one hard)
- Four sorts (of various types)
- Two sets of summary statistics by group
- Export data to CSV

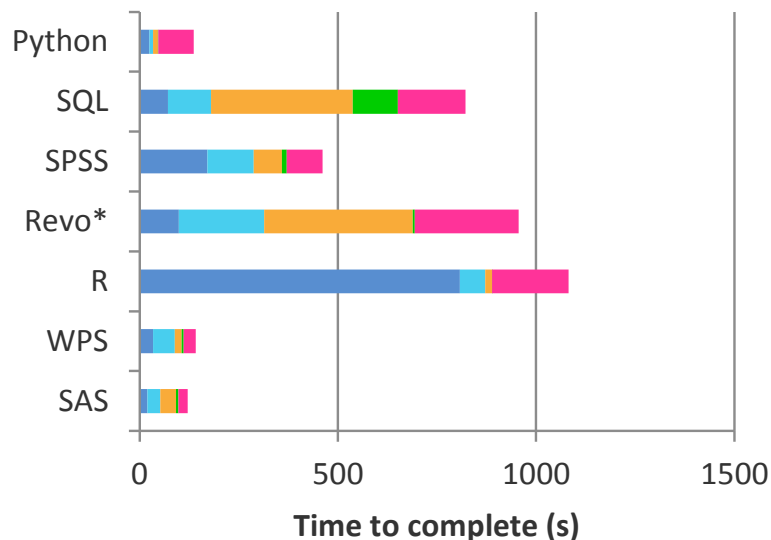
Run with various numbers of rows to see how it scales:

10k, 20k, 50k, 100k, 200k, 500k, 1m, 2m, 5m, 10m

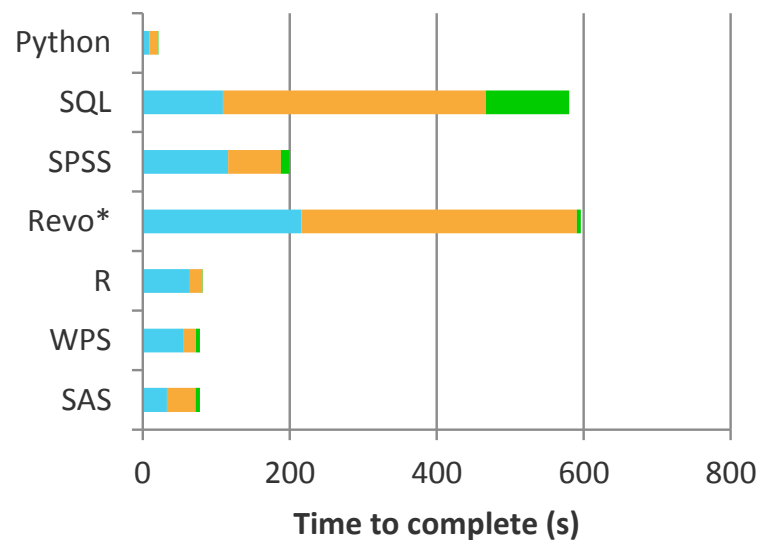
Data management - speed

Time to complete bench test with 5m observations

Including import/export



Excluding import/export

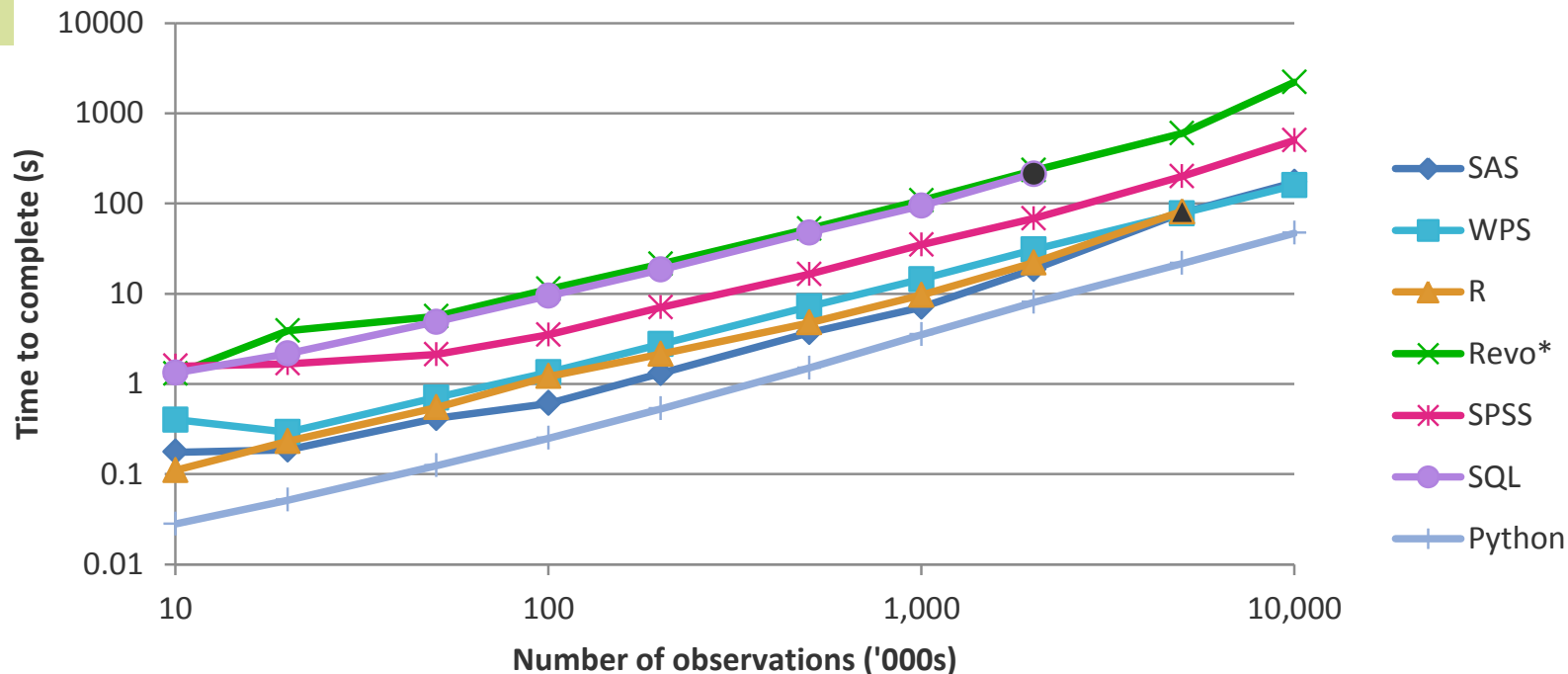


* Tested the 'big dataset' option for Revo – it can run at same speed as R

Data management - speed

Time to complete, excluding import/export

log-log
scale!



Data management – summary

	Speed	Hardware efficiency	Big datasets	Coding efficiency	Ease of use
SAS	●●●●●	●●●●●	●●●●●	●●	●●●
WPS	●●●●◐	●●●●◐	●●●●●	●●	●●●
R	●●●●	●●	●●	●●●●●◐	●●●●●
Revo	●●●●	●●●●	●●●●●	●●●●●	●●●●●
SPSS	●●	●●●●	●●	●●●●	●●●●●●
SQL	●◐	●●	●●●●	●●	●
Python	●●●●●◐	●●●●●	●●●●	●●●●●	●●

Data management – comments

- SPSS inferior on speed and programmability
- All others are plausible options:
 - WPS a genuine alternative to SAS for data management
 - Of the RAM based options, only Python has the corresponding speed boost over SAS/WPS
 - SQL fairly slow and cumbersome

TEST 2

Fitting GLMs

GLMs

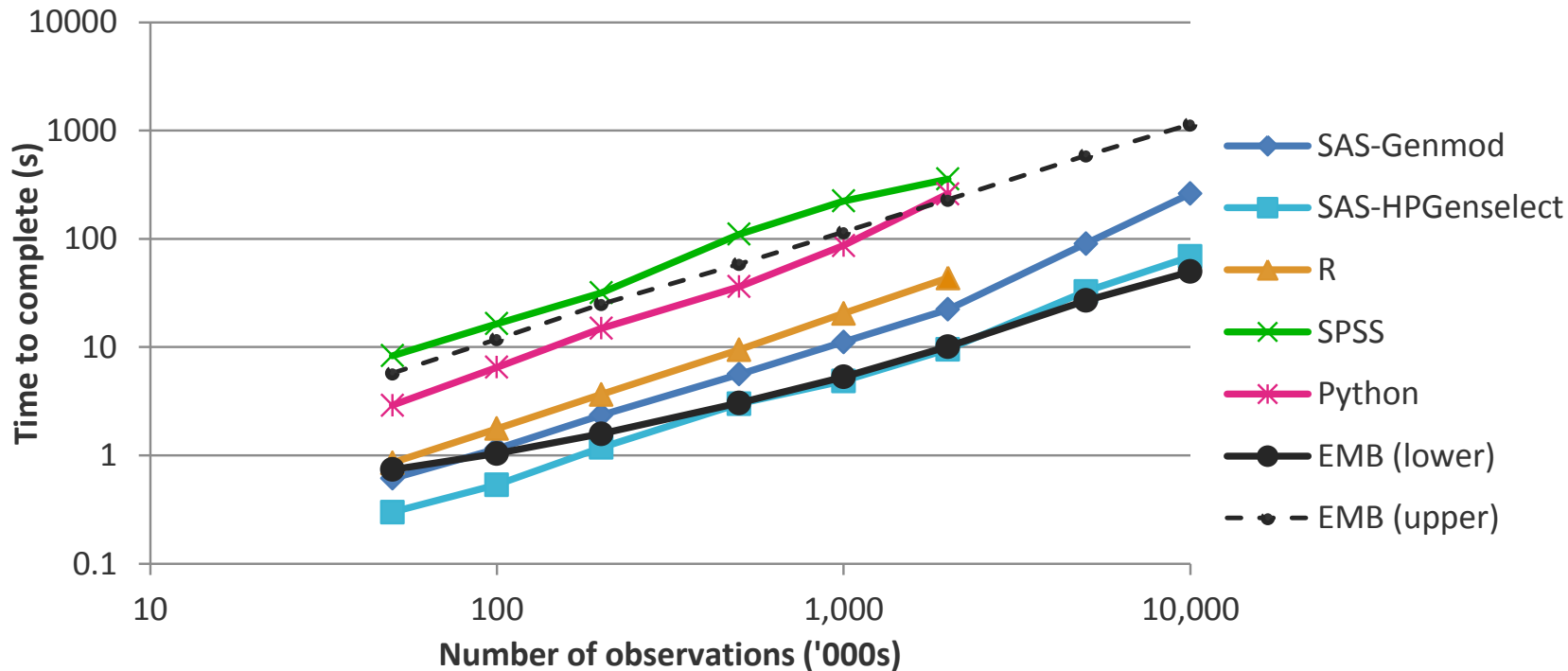
Tested a fit with 20 continuous effects and 5 categorical effects (each with 5 levels) , varying:

- Number of rows: 50k through to 10m
- Distributions: Binomial, Gaussian, Poisson, Gamma, Tweedie
- Note speed benchmarks often vary with different combinations of rows to columns

log-log
scale!

GLMs – speed

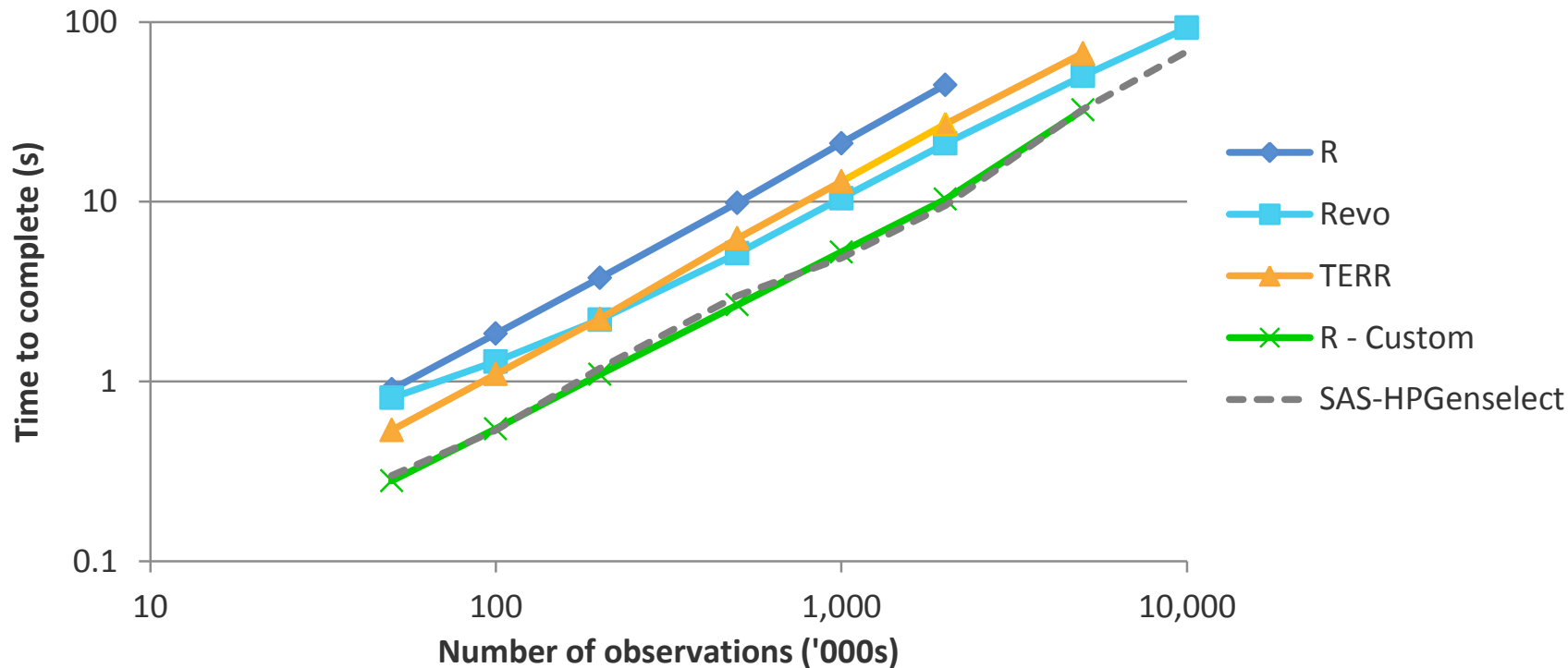
[Avg over all distributions]



GLMs – speed (R + variants)

[Avg over all distributions]

log-log
scale!



GLMs – Summary

* Possible with user effort

	Speed	Big datasets	Ease of use + Interactivity	Interoperability / implementation	Multi-core	Tweedie?	Offset & weights	Diagnostics	Relativities Charts
SAS Genmod	●●●	●●●●●	●●	●●●●●	N	Y*	Y	Y	N*
SAS HPGenselect	●●●●●	●●●●●●	●●	●●●●●	Y	Y	Y	Partial	N*
R	●●●	●●●	●●	●●●	N	Y*	Y	Y	N*
Revo	●●●●●	●●●●●	●●	●●●	Y	Y	Y	Y	N*
SPSS	●●	●●	●●●	●●	N	Partial	Y	Y	N*
Python	●●	●●	●●	●●●	Y	N	Y	Partial	N*
EMBLEM	●●●	●●●●●●	●●●●●	●●	Partial	Y	Y	Y	Y

GLM – comments

- All non-EMBLEM options require work to build support functions around the GLM procedure for dynamic modelling
- SAS and Revolution attractive options with regards to speed and scalability
- EMBLEM interface and usability good, probably most complete package, despite limitations
- R can be made to work well – buy plenty of RAM

TEST 3

Analytics functionality

What do we use analytics for?

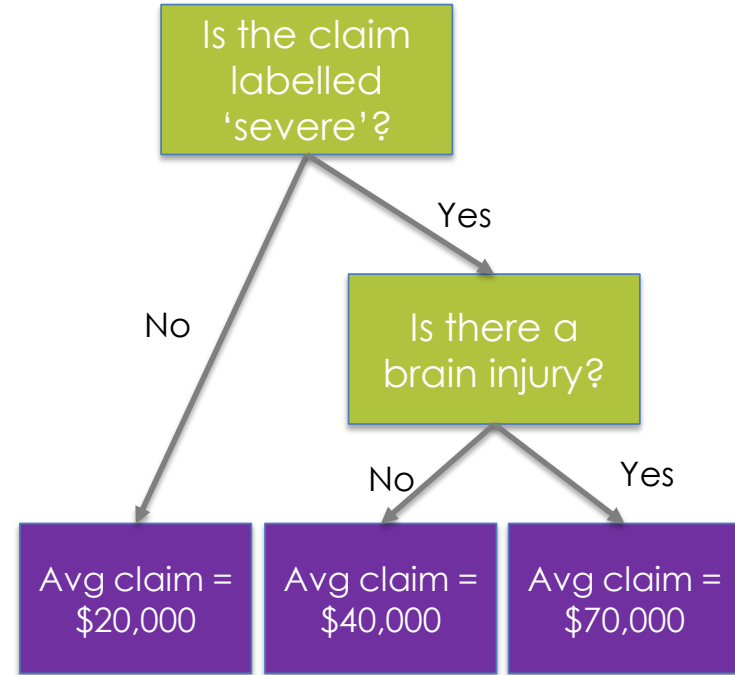
A GLM is a hand-crafted model with good interpretability and smoothness properties. However, sometimes a particular analysis might:

- Need to be done quickly
- Have too many predictor variables
- Need to generate other types of insights (e.g. important variables, or segments)
- Need to be super accurate (and not interpretable)

Data mining / analytics procedures fulfill this need

Analytics

- Tested decision tree functionality for SAS-EM, SPSS, Salford, R
- Looked at the range and quality of other functions included
- Comments are more centred around features and qualitative comparisons



Decision trees

All pretty good

	Speed	Diagnostics	Response types	Modelling options	Interactive output	Interoperability	Ease of use
SAS-EM	●●●●	●●●●◐	●●●	●●●●	●●●	●●◐	●●●
SPSS	●●●●	●●●●◐	●●●	●●●●	●●●	●●◐	●●●◐
Salford	●●●●◐	●●●●◐	●●●	●●●●◐	●●●●	●●●●	●●●●
R	●●●◐	●●●◐	●●●●	●●●	●●	●●◐	●●

Analytics – selected tools

	Hierarchical Clustering	PCA	Decision trees	Boosting with trees	GAM or MARS type regression	Mixed models / penalised regression	Neural Networks	SVM
SAS (base)	✓	✓			✓	✓		
SAS-EM	✓	✓	✓	~	✓	✓	✓	✓
SPSS	✓	✓	✓		✓	✓	✓	✓
Salford			✓	✓✓	✓	✓		
R	✓	✓	✓✓	✓✓	✓✓	✓	✓	✓

~ = Weak implementation ✓ = Included feature ✓✓ = Very good implementation

Analytics – comments

- Easiest to decide on the type of analyses commonly performed, and tailor software selection accordingly
- Of the choices tested, I've been most satisfied with Salford offerings
- Unless all your datasets are very big, having some R capability makes a lot of sense

Final thoughts

Where does this all leave us?

- SAS appears to be here to stay. WPS interesting
- The rise and rise of R (and variants)
- Some room in the market, particularly in the GLM space, or a compelling “integrated solution”
- A brave actuarial team could spend \$0 on statistical software