Using Bayesian Statistical Methods to Determine the Level of Error in Large Spreadsheets

Leslie Bradley
Kevin McDaid

Software Technology Research Centre
Dundalk Institute of Technology, Ireland

This research is supported by the Institutes of Technology Ireland's Technological Sector Research: STRAND I Post-Graduate R&D Skills Programme.

Introduction - Spreadsheets

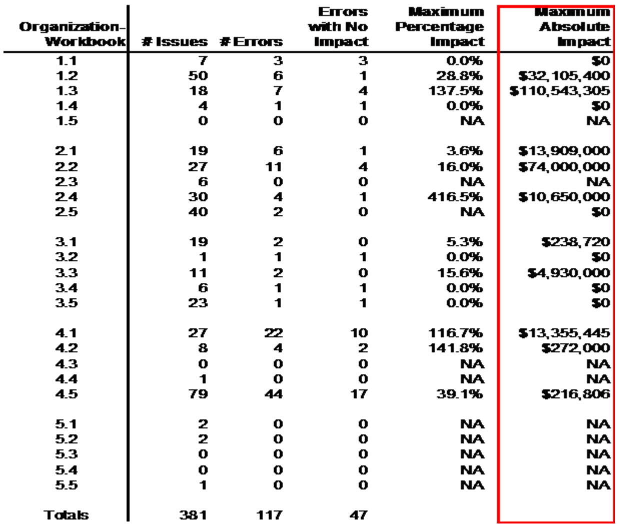
- Who - auditors, accountants, managers
- What - accounts, budgets, databases
- Why – easy to use, hold large amounts of data
- Where - finance sector, business, science
- When - everyday
- How – created with few controls and guidelines
- Result – a high dependence on an application with few controls, and a lack of guidelines and best practices.

Introduction – Impact of Spreadsheet Errors

- The Nevada city budget showed a deficit of $5 million dollars because the spreadsheet was not updated. January 2006
- A cell entry error cost Columbia Housing Authority $118,387, which was overpaid to landlords. February 2006
- Deliberate fraud, AIB losses of nearly $700 million dollars were hidden by a trader, John Rusnak. 2001

[2]

Introduction – Impact of Spreadsheet Errors



Organization-Workbook	# Issues	# Errors	Errors with No Impact	Maximum Percentage Impact	Maximum Absolute Impact
1.1	7	3	3	0.0%	$0
1.2	50	6	1	28.8%	$32,105,400
1.3	18	7	4	137.5%	$110,543,305
1.4	4	1	1	0.0%	$0
1.5	0	0	0	NA	NA
2.1	19	6	1	3.6%	$13,909,000
2.2	27	11	4	16.0%	$74,000,000
2.3	6	0	0	NA	NA
2.4	30	4	1	416.5%	$10,650,000
2.5	40	2	0	NA	$0
3.1	19	2	0	5.3%	$238,720
3.2	1	1	1	0.0%	$0
3.3	11	2	0	15.6%	$4,930,000
3.4	6	1	1	0.0%	$0
3.5	23	1	1	0.0%	$0
4.1	27	22	10	116.7%	$13,355,445
4.2	8	4	2	141.8%	$272,000
4.3	0	0	0	NA	NA
4.4	1	0	0	NA	NA
4.5	79	44	17	39.1%	$216,806
5.1	2	0	0	NA	NA
5.2	2	0	0	NA	NA
5.3	0	0	0	NA	NA
5.4	0	0	0	NA	NA
5.5	1	0	0	NA	NA
Totals	381	117	47		

[1]

Research Questions

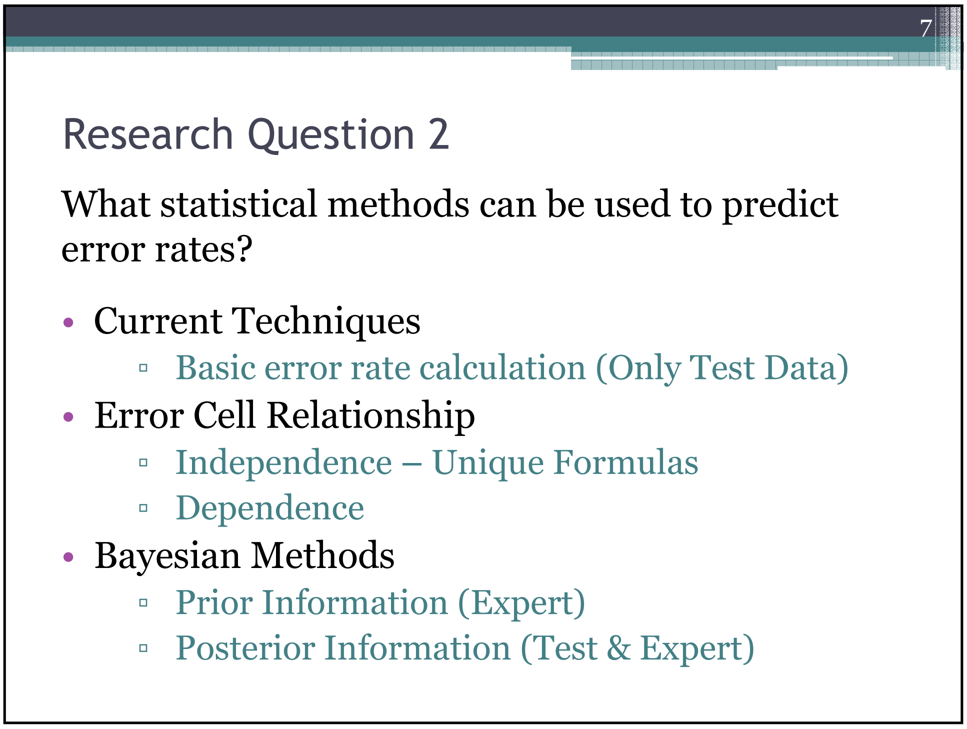
Can a model be established to predict the cell error rate of large spreadsheets, based on expert knowledge and any available test data, to aid the decision on whether to test the spreadsheet?

1. What does existing research say about the level of spreadsheet error and methods to discover errors in spreadsheets?
2. What statistical methods can be used to predict spreadsheet error rates?
3. Can a model be developed that combines prior knowledge and available test data to estimate the cell error rate (CER) for large spreadsheets?
4. How effective is the model at predicting the CER in spreadsheets

Research Question 1

What does existing research say about the level of spreadsheet error and methods to discover errors in spreadsheets?

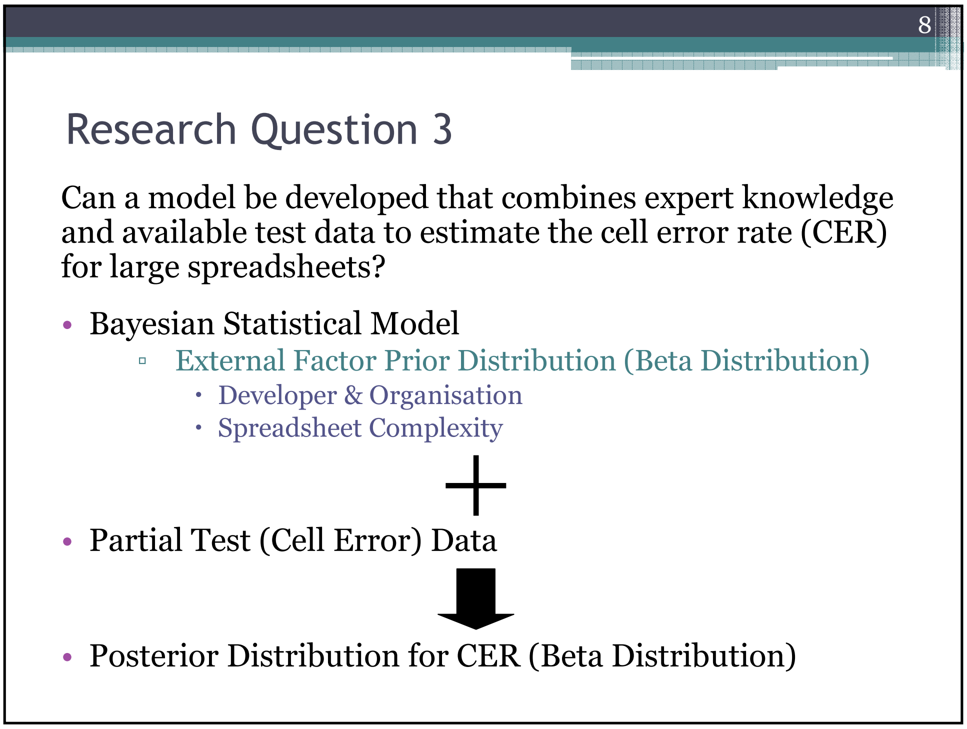
- Research into Spreadsheet Review, Taxonomies, Tools,....
- Error Rates
  - Panko – 5.2% [3, 4]
  - Powell – 1.79% (0.87% for Defects) [5]



Research Question 2

What statistical methods can be used to predict error rates?

- Current Techniques
  - Basic error rate calculation (Only Test Data)
- Error Cell Relationship
  - Independence – Unique Formulas
  - Dependence
- Bayesian Methods
  - Prior Information (Expert)
  - Posterior Information (Test & Expert)



Research Question 3

Can a model be developed that combines expert knowledge and available test data to estimate the cell error rate (CER) for large spreadsheets?

- Bayesian Statistical Model
  - External Factor Prior Distribution (Beta Distribution)
    - Developer & Organisation
    - Spreadsheet Complexity

+

- Partial Test (Cell Error) Data

⬇

- Posterior Distribution for CER (Beta Distribution)

Research Question 4

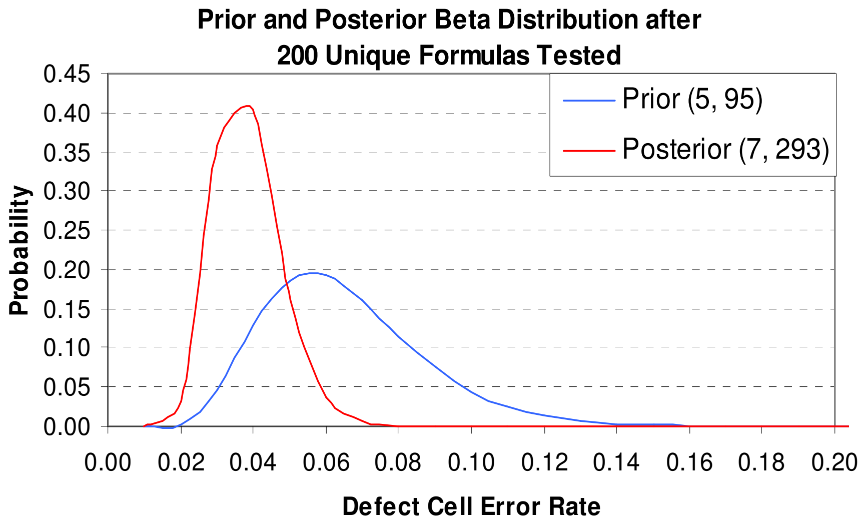
How effective is the model at predicting the cell error rate in spreadsheets?

- Study
  - Suite of real spreadsheets
    (split equally into 2 groups)
  - Test Group 1
  - Apply methodology sequentially to Group 2
  - Examine the quality of prediction and associated decision at each time point
  - Test Group 2 using frequentist and compare results with predictions.

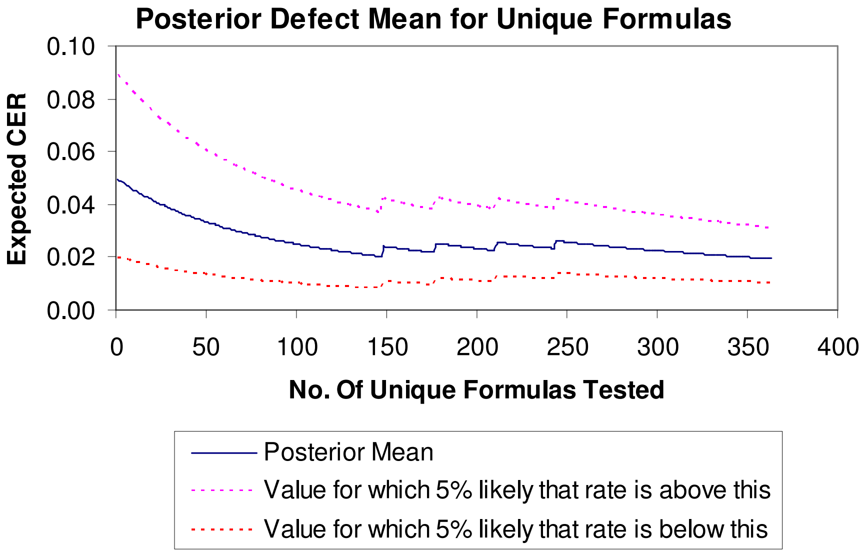
Example

- Suppose a spreadsheet contains 303 unique formulas.
- Expert information indicates a defect error rate of 0.05 and standard deviation of 0.0217
- The first 200 unique formulas are tested and results show 2 defect error cells.
- This test data is added to the prior to give posterior information which has mean 0.023 and standard deviation of 0.0087

Example Continued



Example Continued

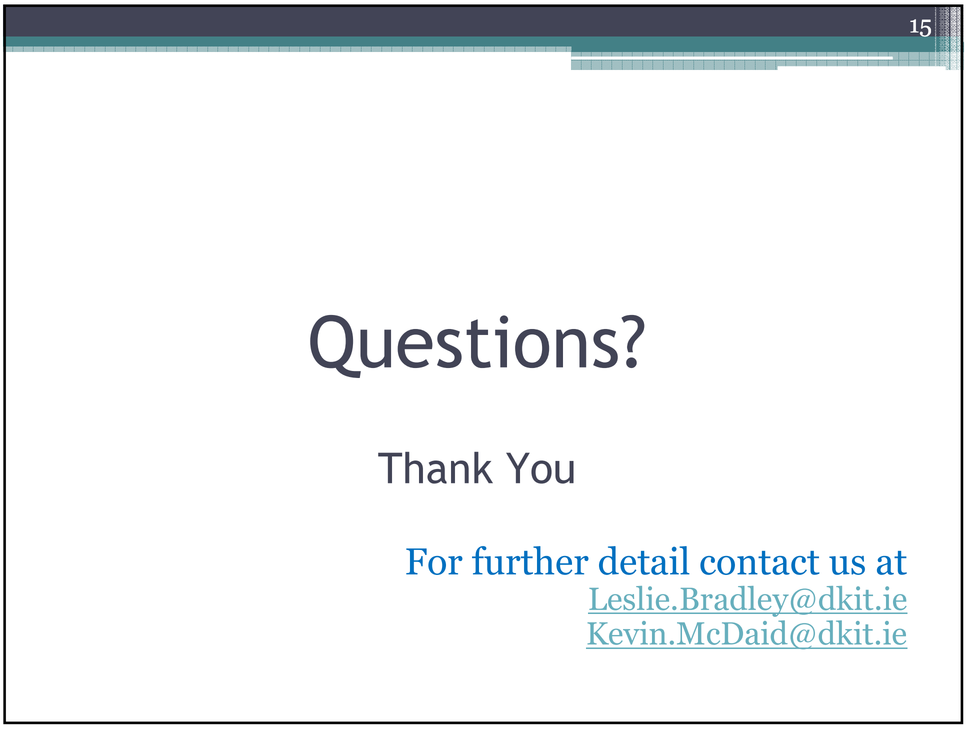


Conclusion

- Spreadsheet errors can have significant financial impact
- Spreadsheets audits are time consuming
- Bayesian method combining expert knowledge and test data can aid the decision on whether a complete audit is required

Bibliography

1. Powell, S.G., Baker, K. R., and Lawson, B. *Impact of Errors in Operational Spreadsheets*. in *Proceedings of the European Spreadsheets Risks Interest Group*. 2007. Greenwich, England.
2. EuSpRiG, 10:55 a.m. February 25, http://www.eusprig.org/stories.htm
3. Panko, R.R., *What we know about spreadsheet errors*. Journal of End User Computing Special issue on scaling Up End User Development, 1998. **10**(2): p. 15-21.
4. Panko, R.R., *What we know about Spreadsheet Errors Extended Version*. 2005.
5. Powell, S.G., Baker, K. R., and Lawson, B., *Errors in Spreadsheets*. 2007, Tuck School of Business at Dartmouth College.



Questions?

Thank You

For further detail contact us at
Leslie.Bradley@dkit.ie
Kevin.McDaid@dkit.ie