


Calibration and Combination of Expert's Dependence Estimates

Oswaldo Morales Napoles[1] Daniël Worm[1] Anca M. Hanea[2] Ivo Kalkman[1]

1. TNO Oswaldo.MoralesNapoles@tno.nl

2. TU Delft

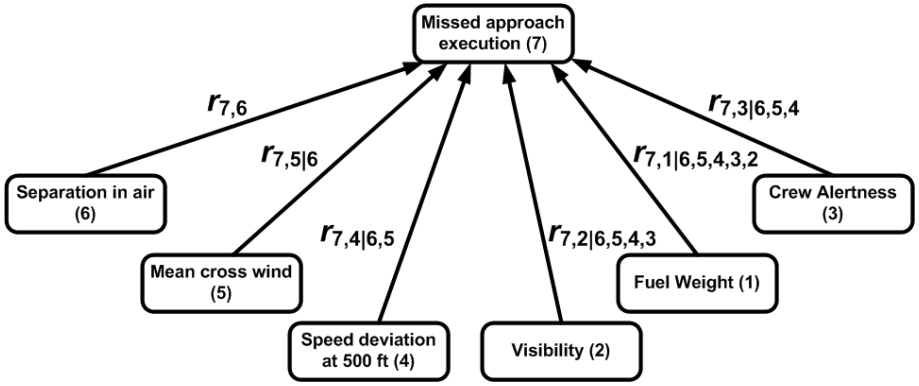

TNO innovation for life

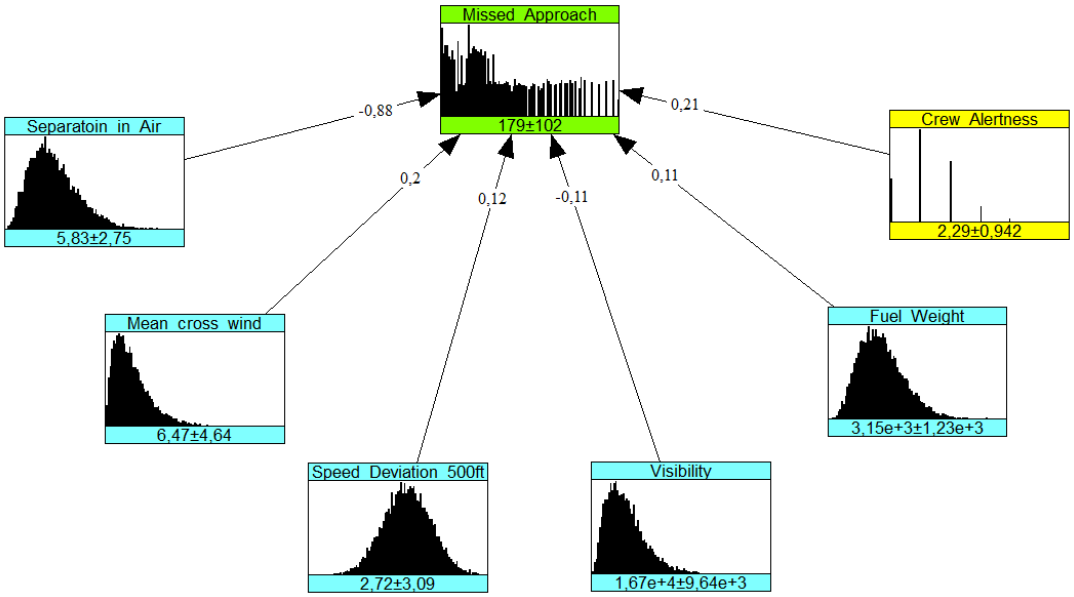
Outline

- Motivation
  - Previous Work
- The exercises
- Results
- Conclusions

Controlled Flight into Terrain

- 7 (conditional) rank correlations
- Conditional probabilities of exceedence

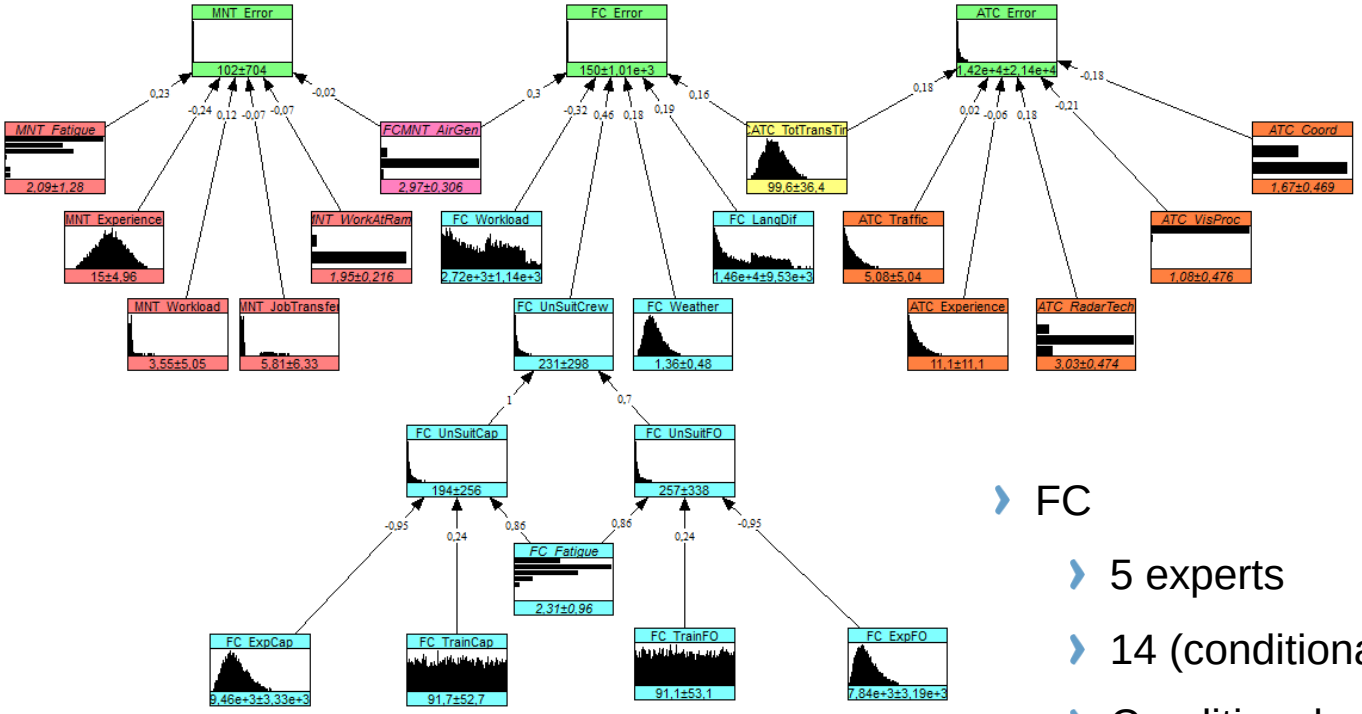




- P(7>med| 6>med) …
- P(7>med| 1>med, 2>med, 3>med, 4>med, 5>med, 6>med)

- r(7,6)
- r(7,5)/r(7,6) …
- r(7,1)/r(7,6)

Flight, Maintenance & ATC Crew Error

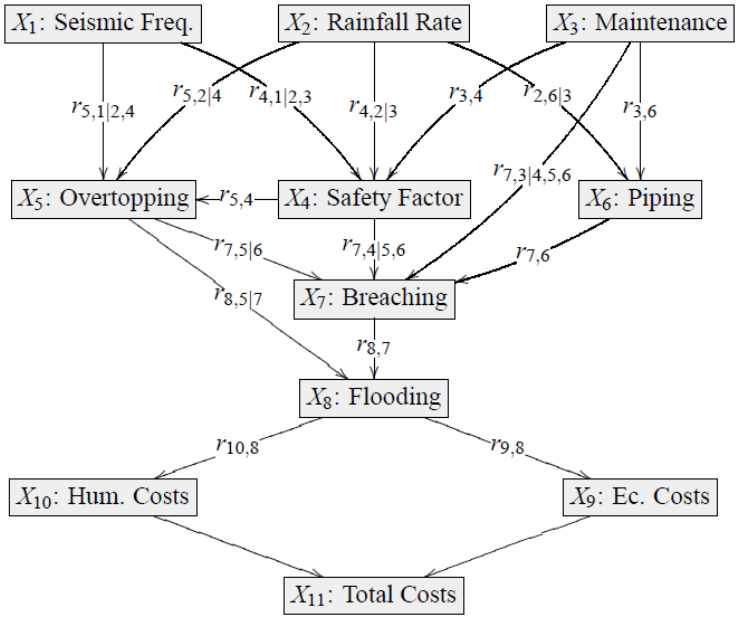


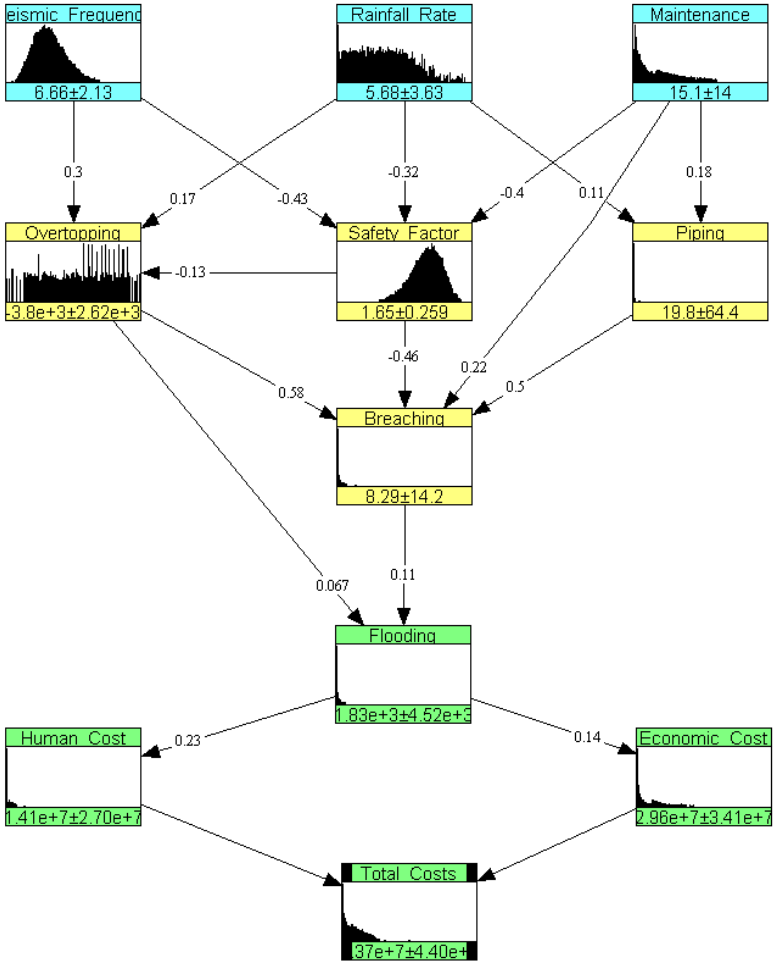
- MNT
  - 1 expert
  - 6 (conditional) rank correlations
  - Ratios of rank correlations

- FC
  - 5 experts
  - 14 (conditional) rank correlations
  - Conditional probabilities of exceedence
- ATC
  - 5 experts
  - 6 (conditional) rank correlations
  - Ratios of rank correlations


Earth Dams in Mexico

- 4 experts
- 16 (conditional) rank correlations
- Ratios of rank correlations



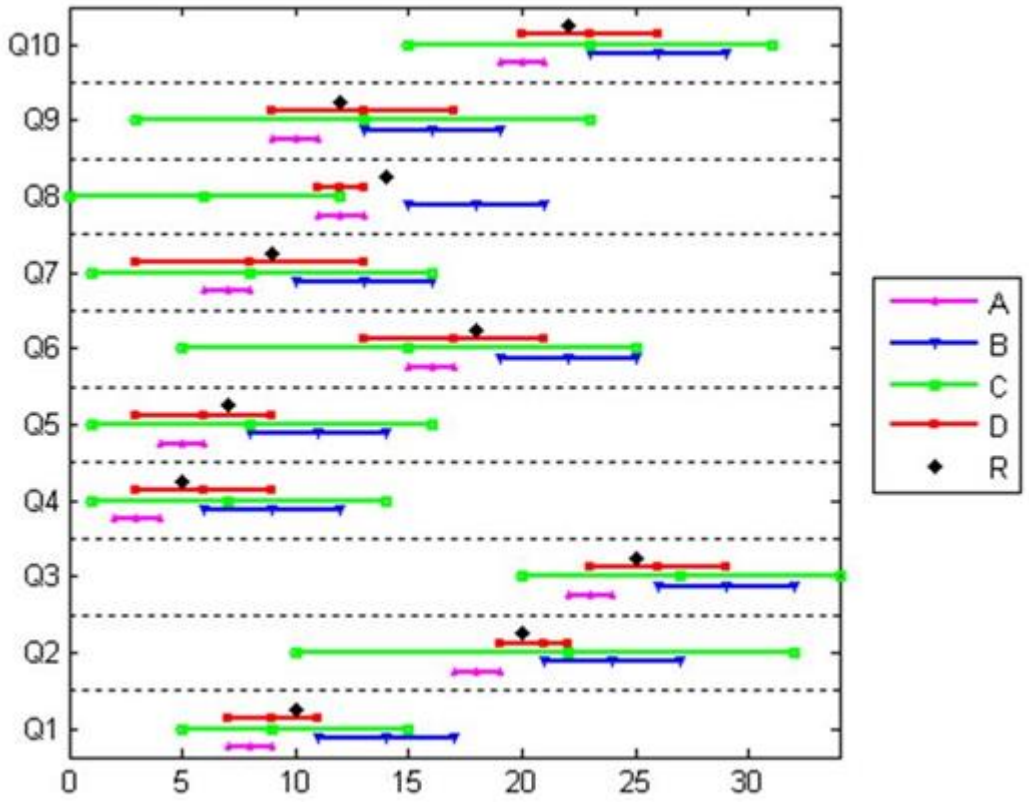



Question

- Which method would render more accurate answers?
- Can experts provide meaningful estimates?
- TNO Project GAMES2R: GrAphical ModEls for Systems Risk and Reliability

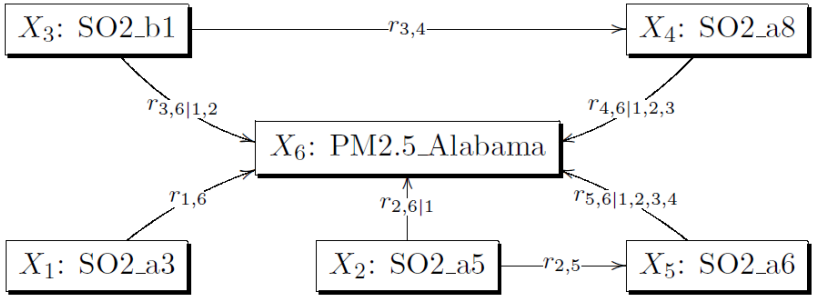

Parenthesis: Cooke's classical model

- Seed variables
  - Analyst knows the answer post hoc
- Calibration:
  - Accuracy in a statistical sense
- Information:
  - How uncertain are well calibrated experts?
- Weight experts based on their performance



Exercise 1: The Models

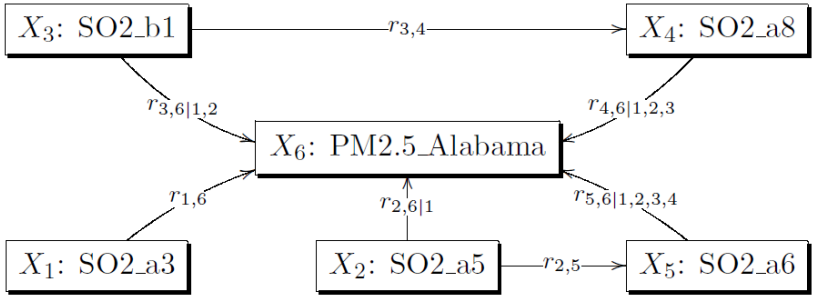
- SO_2 emissions and $PM_{2.5}$ concentrations
- 7 (conditional) rank correlations

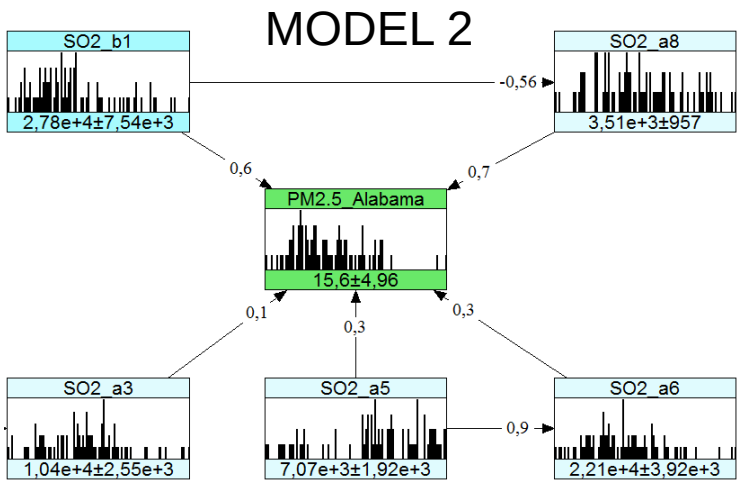


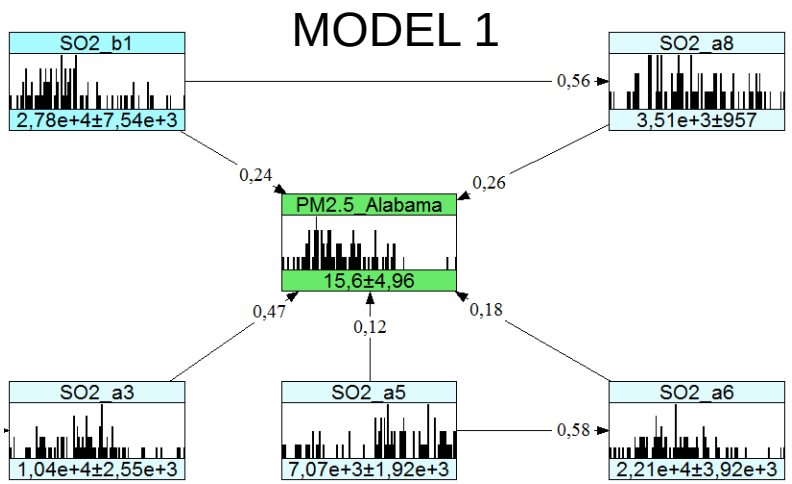
- Air pollution in the US
- Sometimes used in epidemiology
- Workshop December 2012
- Preliminary results presented in August 2013 in Strathclyde University

The Models

- SO_2 emissions and $PM_{2.5}$ concentrations
- 7 (conditional) rank correlations
- MODEL 1 → Original Data
- MODEL 2 → Fictitious Data

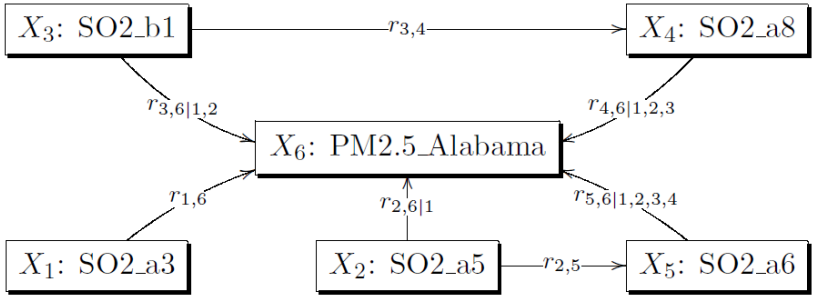







The Models

- 14 experts
  - 9 grad. students (TU Delft)
  - 5 researchers (TU Delft & TNO)
- 500k samples / model sent 1 week before
- Background information
  - data
  - type of questions
- Half day workshop (TU Delft)
- Two groups of 7 experts each
  - M1CPE & M2RRC
  - M1RRC & M2CPE



1. Consider model i. There are $N_{1,i}$ samples (out of 500,000) for which variable SO2_a3 is at least 10,466 (median). Consider the indices of all variables corresponding to this subset. In other words, conditionalize on this subset. In how many of these indices will the value of PM2.5_Alabama be at least 14.82 (median)?

5. Consider model i. There are $N_{5,i}$ samples (out of 500,000) for which variable SO2_a3 is at least 10,466 (median), SO2_a5 is at least 7,256 (median), SO2_b1 is at least 26,091 (median), SO2_a8 (median) is at least 3,429 (median) and SO2_a6 is at least 21,908 (median). Consider the indices of all variables corresponding to this subset. In other words, conditionalize on this subset. In how many of these indices will the value of PM2.5_Alabama be at least 14.82 (median)?

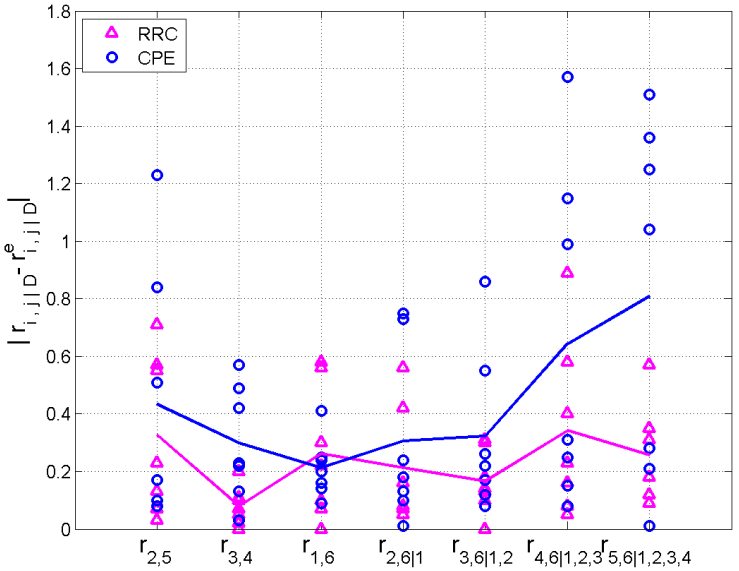
6. Consider model i. What is the rank correlation between SO2_a3 and PM2.5_Alabama?

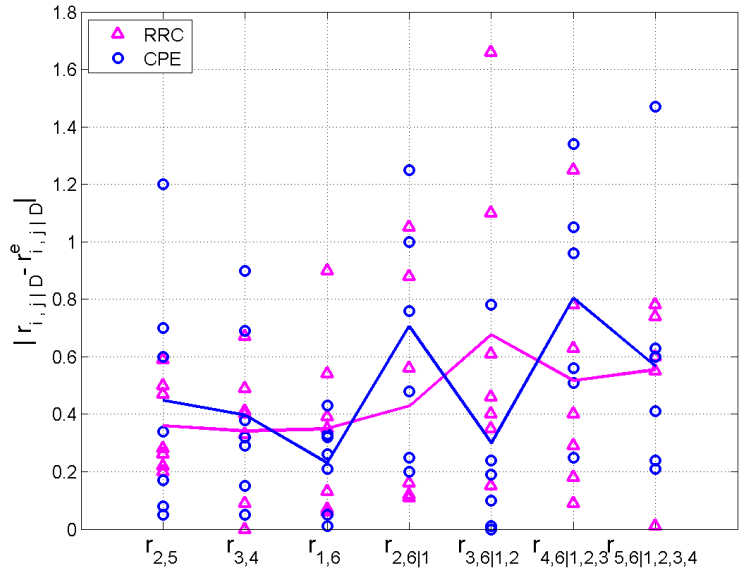
10. Consider model i. What is the ratio of the rank correlation between SO2_a6 and PM2.5_Alabama to the rank correlation between SO2_a3 and PM2.5_Alabama?

Results (Individual estimates)

$$|r_{i,j|D} - r^{e}_{i,j|D}|$$

$$\bar{\delta}_{M1RRC} = 0.23$$
$$\bar{\delta}_{M1CPE} = 0.43$$
$$\bar{\delta}_{M2RRC} = 0.46$$
$$\bar{\delta}_{M2CPE} = 0.49$$






Results (Individual estimates) ANOVA

$$H_0 : \overline{\delta}_{\mathrm{M1CPE}} = \overline{\delta}_{\mathrm{M2RRC}} = \overline{\delta}_{\mathrm{M2CPE}} = \overline{\delta}_{\mathrm{M1RRC}}$$

- Total Sum of Squares
  - Between Groups (Treatments)
  - Within Group (Error)
  - P-val (F statistic is actually larger) = 0.0016 → reject H_0
  - Which means are different?

$$Total\ SS = SST + SSE$$

$$\frac{SST/(k-1)}{SSE/(n_1+\ldots+n_k-k)} \sim \mathrm{F}$$

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Columns'	1,98	3	0,6612	5,291	0,0016
'Error'	23,99	192	0,1250	[]	[]
'Total'	25,98	195	[]	[]	[]


Results (Individual estimates) Tukey's allowances

$$\mathrm{H}_0 : \delta_i - \delta_j$$

$$(\overline{\delta}_i - \overline{\delta}_j) - q_{\alpha,k,(n-1)k}\sqrt{\frac{SSE/(n_1+\ldots+n_k-k)}{n}} \leq \delta_i - \delta_j \leq (\overline{\delta}_i - \overline{\delta}_j) + q_{\alpha,k,(n-1)k}\sqrt{\frac{SSE/(n_1+\ldots+n_k-k)}{n}}$$

- Randomized design (ANOVA)
- The probability that all $\binom{k}{2}$ pairs $\delta_i - \delta_j$ simultaneously satisfy the inequalities above is $1 - \alpha$.
- $q_{\alpha,k,v}$ is the upper α critical value of the Studentized range distribution

$$\bar{\delta}_{M1RRC} \neq \bar{\delta}_{M1CPE}$$
$$\bar{\delta}_{M1RRC} \neq \bar{\delta}_{M2RRC}$$
$$\bar{\delta}_{M1RRC} \neq \bar{\delta}_{M2CPE}$$

1	2	-0,38	-0,20	-0,01
1	3	-0,41	-0,23	-0,04
1	4	-0,44	-0,26	-0,07
2	3	-0,21	-0,03	0,15
2	4	-0,24	-0,06	0,12
3	4	-0,22	-0,03	0,15


TNO innovation for life

Results (Individual Models)

- H_0: $BN_e = BN_{true}$
- Sample from BN_{true}
- Empirical distribution of $det(R_{true})$
- Accept if $det(R_e)$ is within 5th and 95th percentiles distribution of $det(R_{true})$
- $det(\sum)$ is a measure of dependence
- Motivated from data driven applications. However…
  - VERY different correlation matrices might lead to the same determinant
  - Proof in our paper ESREL 2013

1	0,907981	0,517638		1	0,907981	0,517638
0,907981	1	0,795522		0,907981	1	0,144489
0,517638	0,795522	1		0,517638	0,144489	1
			0,022565			
1	-0,90798	-0,90798		1	0	0
-0,90798	1	0,733548		0	1	-0,98865
-0,90798	0,733548	1		0	-0,98865	1

- Instead → measure of distance
- Heillinger distance

$$dCal(e) = 1 - d_H(N_{true}, N_e)$$

$$d_H(N_{true}, N_e) = \sqrt{1 - \eta(N_{true}, N_e)}$$

$$\eta(N_{true}, N_e) = \frac{\det(\sum_{true})^{\frac{1}{4}} \det(\sum_e)^{\frac{1}{4}}}{\det(\frac{1}{2}\sum_{true} + \frac{1}{2}\sum_e)^{\frac{1}{2}}}$$


d-Cal Properties

- d_H is a metric:
  - d_H is symmetric
  - d_H satisfies the triangle inequality

- *d*Cal(e) = 1 iff $\sum_e = \sum_{true}$

- *d*Cal(e) = 0 if
  - $\sum_{true}$ = I and $\sum_e$ = perfect dep. linear combination of RV or
  - Viceversa

- **Capture magnitude right but direction wrong.**
- *d*Cal(e) → 0 if
  - $\sum_e$ → 2I - $\sum_{true}$ while
  - det($\sum_e$)→0 and det($\sum_{true}$) →0
  - $_{ij}(\sum_{ture}) \approx -\ _{ij}(\sum_e)$
  - Proof: paper in preparation
- **Capture magnitude and direction "close enough"**
- dCal(e) → 1 if
  - $_{ij}(\sum_{ture}) \approx\ _{ij}(\sum_e)$
  - Entry-wise equal
  - Proof: paper in preparation


Results Exercise 1

- Group 1
  - B is best with both methods
- Group 2
  - G, D, M: d-Cal >0,7
  - D high both methods
- Performance based combination best → Analogy with Cooke's method

Group 1					Group 2				
Id.	Calibr.	Inform.	d-Cal M1CPE	d-Cal. M2RRC	Id.	Calibr.	Inform.	d-Cal M1RRC	d-Cal. M2CPE
A	0.0139	2.092	0.46	0.13	D	0.0357	2.745	0.71	0.60
B	0.0013	1.662	0.65	0.52	E	0.0063	1.497	0.51	0.32
C	l.o.	1.89	0.32	0.10	F	0.7069	0.7571	0.12	0.09
H	l.o.	2.336	0.28	0.32	G	l.o.	1.86	0.87	0.49
I	l.o.	1.474	0.64	0.18	J	l.o.	2.49	0.32	0.17
K	0.0011	1.209	0.62	0.17	L	0.0028	1.169	0.09	0.16
N	l.o.	2.378	0.62	0.31	M	0.00131	3.84	0.75	0.32
Eq.	0.2282	0.0263	0.74	0.37	Eq.	0.5503	0.3009	0.66	0.37
Gl.	0.8283	1.459	0.76	0.52	Gl.	0.7069	0.7571	0.95	0.60

Table 1 Calibration, Information and d-Calibration scores for air pollution NPBN experts.

Illustration dCal scores

- G, M1RRRC → 0.87
- Gl., M1RRC → 0.95
- D, M2CPE → 0.60 = Gl M2CPE

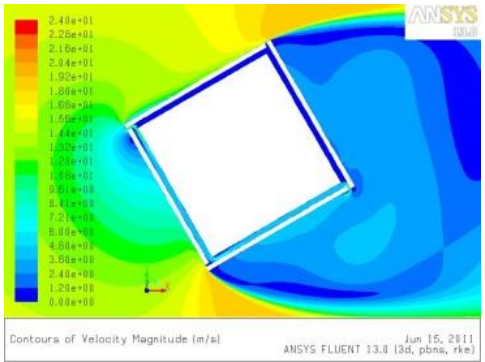
$$\Sigma_{M1}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.49 \\ & 1 & 0 & 0.58 & 0 & 0.21 \\ & & 1 & 0 & 0.59 & 0.10 \\ & & & 1 & 0 & 0.31 \\ & & & & 1 & 0.19 \\ & & & & & 1 \end{bmatrix} \quad \Sigma_{G,M1RRC}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.41 \\ & 1 & 0 & 0.48 & 0 & 0.12 \\ & & 1 & 0 & 0.45 & 0.20 \\ & & & 1 & 0 & 0.33 \\ & & & & 1 & 0.12 \\ & & & & & 1 \end{bmatrix} \quad \Sigma_{Gl,M1RRC}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.46 \\ & 1 & 0 & 0.61 & 0 & 0.16 \\ & & 1 & 0 & 0.60 & 0.16 \\ & & & 1 & 0 & 0.33 \\ & & & & 1 & 0.21 \\ & & & & & 1 \end{bmatrix}$$

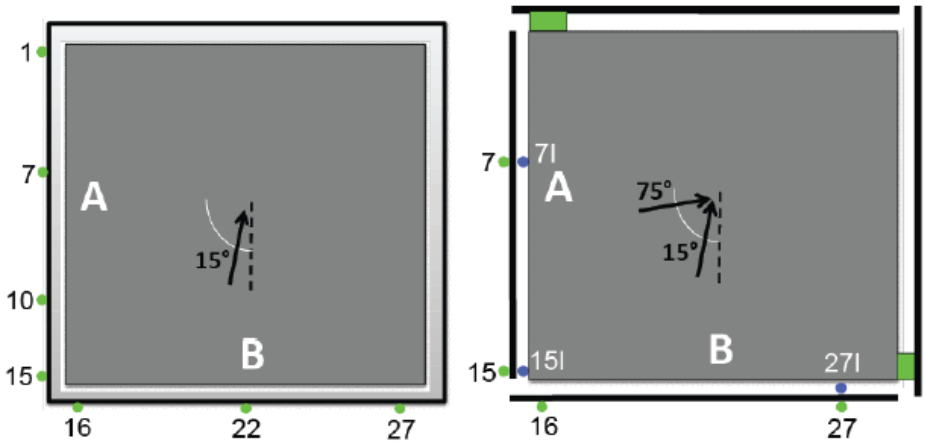
$$\Sigma_{M2}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.10 \\ & 1 & 0 & -0.57 & 0 & 0.58 \\ & & 1 & 0 & 0.90 & 0.30 \\ & & & 1 & 0 & 0.10 \\ & & & & 1 & 0.34 \\ & & & & & 1 \end{bmatrix} \quad \Sigma_{D,M2CPE}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.31 \\ & 1 & 0 & -0.51 & 0 & 0.33 \\ & & 1 & 0 & 0.92 & 0.31 \\ & & & 1 & 0 & -0.12 \\ & & & & 1 & 0.23 \\ & & & & & 1 \end{bmatrix} \quad \Sigma_{Gl,M2CPE}=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.31 \\ & 1 & 0 & -0.51 & 0 & 0.33 \\ & & 1 & 0 & 0.92 & 0.31 \\ & & & 1 & 0 & -0.12 \\ & & & & 1 & 0.23 \\ & & & & & 1 \end{bmatrix}$$

Table 2 Correlation matrices for Models 1 and 2, best individual expert per model and best combined expert per model.

Exercise 2

- Wind pressure coefficients measured in wind tunnel
- Use CFD models
- Pressure compromises structural integrity of building elements
- Interest in net forces over facade panels
- Correlation between pressures external & internal to the wooden model → net forces



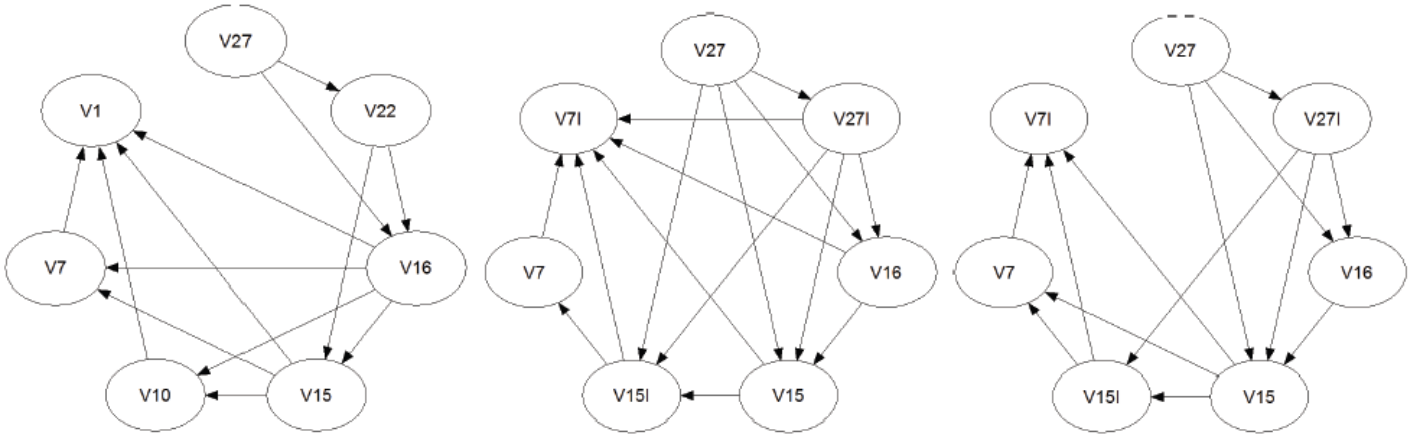


(a) Closed cube (b) Opened cube

Figure 2 Cross section of wooden cube models used in wind tunnel experiments.

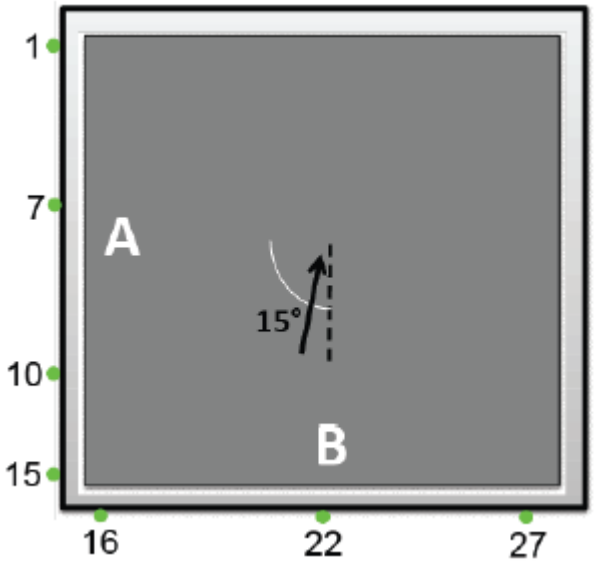
Exercise 2

- Workshop October 2013
- 9 TNO experts
- 3 models



(a) Closed cube 15^o NPBN (b) Opened cube 15^o NPBN (c) Opened cube 75^o NPBN

- Low calibration scores
- Low d-calibration scores
  - Negative correlations between A & B
  - Not the case (Capture magnitude right but direction wrong)
  - Big improvement when looking at separate sides
- Performance based combination best → Analogy with Cooke's method

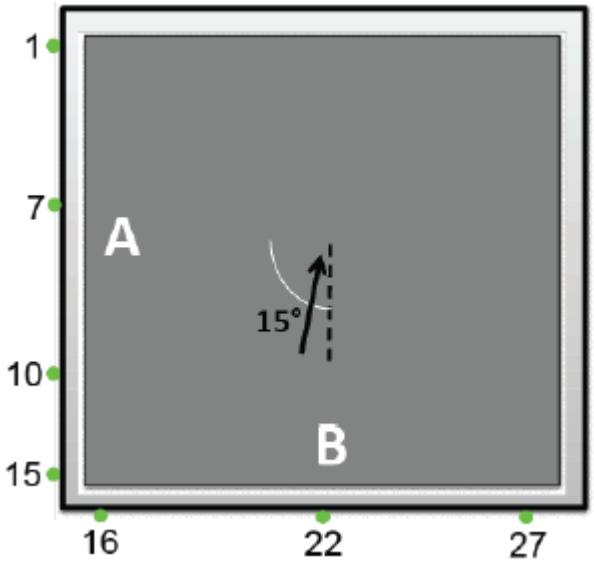


Group 1						Group 2					
Id.	Calibr.	Inform.	d-Cal M3CPE	d-Cal. M3CPEB	d-Cal. M3CPEA	Id.	Calibr.	Inform.	d-Cal M3RRC	d-Cal. M3RRCB	d-Cal. M3RRCA
A	l.o.	0.7587	0.24	0.58	0.45	D	l.o.	1.637	0.26	0.47	0.58
B	0.0026	0.8698	0.20	0.48	0.38	E	l.o.	1.444	0.18	0.66	0.45
C	0.015	0.5493	0.19	0.69	0.41	F	l.o.	0.5965	0.36	0.45	0.82
H	0.138	0.3332	0.17	0.49	0.50	I	l.o.	0.9946	0.31	0.44	0.65
G	l.o.	0.8242	0.09	0.95	0.28						
Eq.	0.265	0.242	0.33	0.67	0.61	Eq.	0.005	0.1989	0.35	0.50	0.82
Gl.	0.265	0.242	0.35	0.95	0.69	Gl.	0.0126	0.2947	0.36	0.66	0.82

Table 3 Calibration, Information and d-Calibration scores for Closed Cube 15° NPBN experts.

Illustration dCal scores

- Gl, M3CPEB → 0.95
- Gl, M3RRCB → 0.66
- Gl, M3CPEB → 0.69
- Gl, M3RRCB → 0.82



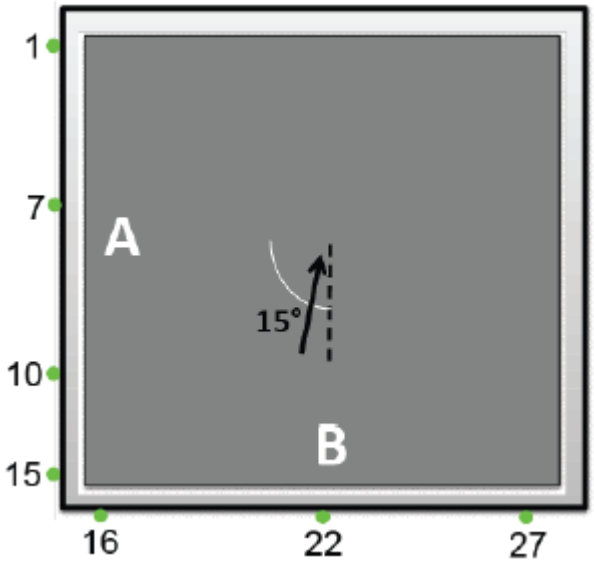
$$\Sigma_{M3B}=\begin{bmatrix}1 & 0.94 & 0.90\\ & 1 & 0.93\\ & & 1\end{bmatrix} \quad \Sigma_{Gl.,M3CPEB}=\begin{bmatrix}1 & 0.94 & 0.90\\ & 1 & 0.94\\ & & 1\end{bmatrix} \quad \Sigma_{Gl,M3RRCB}=\begin{bmatrix}1 & 0.89 & 0.76\\ & 1 & 0.94\\ & & 1\end{bmatrix}$$

$$\Sigma_{M3A}=\begin{bmatrix}1 & 0.43 & 0.18 & 0.28\\ & 1 & 0.11 & 0.16\\ & & 1 & 0.15\\ & & & 1\end{bmatrix} \quad \Sigma_{Gl.,M3CPEA}=\begin{bmatrix}1 & 0.70 & 0.57 & 0.33\\ & 1 & 0.46 & 0.24\\ & & 1 & 0.62\\ & & & 1\end{bmatrix} \quad \Sigma_{Gl,M3RRCA}=\begin{bmatrix}1 & 0.48 & 0.19 & 0.10\\ & 1 & 0.09 & 0.20\\ & & 1 & 0.48\\ & & & 1\end{bmatrix}$$

Table 4 Correlation matrices for Model 3 (15^o closed), and best combined expert per submodel (A and B sides).

Illustration dCal scores

- Gl, M3CPEB → 0.95
- Gl, M3RRCB → 0.66
- Gl, M3CPEB → 0.69
- Gl, M3RRCB → 0.82



$$\Sigma_{M3B}=\begin{bmatrix}1 & 0.94 & 0.90\\ & 1 & 0.93\\ & & 1\end{bmatrix} \quad \Sigma_{Gl.,M3CPEB}=\begin{bmatrix}1 & 0.94 & 0.90\\ & 1 & 0.94\\ & & 1\end{bmatrix} \quad \Sigma_{Gl,M3RRCB}=\begin{bmatrix}1 & 0.89 & 0.76\\ & 1 & 0.94\\ & & 1\end{bmatrix}$$

$$\Sigma_{M3A}=\begin{bmatrix}1 & 0.43 & 0.18 & 0.28\\ & 1 & 0.11 & 0.16\\ & & 1 & 0.15\\ & & & 1\end{bmatrix} \quad \Sigma_{Gl.,M3CPEA}=\begin{bmatrix}1 & 0.70 & 0.57 & 0.33\\ & 1 & 0.46 & 0.24\\ & & 1 & 0.62\\ & & & 1\end{bmatrix} \quad \Sigma_{Gl,M3RRCA}=\begin{bmatrix}1 & 0.48 & 0.19 & 0.10\\ & 1 & 0.09 & 0.20\\ & & 1 & 0.48\\ & & & 1\end{bmatrix}$$

Table 4 Correlation matrices for Model 3 (15^o closed), and best combined expert per submodel (A and B sides).


Final comments

- Other two models behave similarly
- Can experts provide meaningful estimates? Yes, but it is not easy.
- Which method would render more accurate answers?
  - Both would do the job
  - Experts d-Cal is more or less robust to RRC & CPE
- Higher order cond. rank correlation less accurate
- Calibration and d-Calibration do not correlate perfectly
  - M3M4M5, M3M4M5B and M3M4M5A CPE r(Cal,dCal) ≈ 0.05
- Many interesting theoretical questions
  - Distribution of d-Cal
  - d-Cal & Set of all correlation matrices on n variables
  - Combination schemes
  - Method works nicely for Correlation Matrices → other dependence measures? Tail dependence for example?

Questions?