UNIVERSITY OF Southampton

Missing data in linked administrative registers – estimating the size of the Māori population in New Zealand

Paul A. Smith

joint work with Peter van der Heijden and Maarten Cruyff

Missing Data Workshop 11 December 2020

Outline

- Dual system estimation
- Missing information
  - single-source variables
- Extension to multiple systems
  - partial coverage
  - the same variable measured different ways
- Conclusions

Dual system estimation

- Population size estimation
  - ≥ 2 sources with imperfect information
  - capture-recapture (Lincoln-Peterson estimator)
- Many practical examples in human populations
  - Population census coverage
  - Modern slavery
  - Homeless
  - Unregistered persons

Dual system estimator (DSE)

		Source 2: In	Source 2: Out
Source 1	In	n_{11}	n_{10}
	Out	n_{01}	?

$$\widehat{N} = \frac{(n_{11} + n_{10})(n_{11} + n_{01})}{n_{11}} = \frac{n_{1\cdot}\, n_{\cdot 1}}{n_{11}}$$

$$\hat{n}_{00} = \widehat{N} - n_{11} - n_{10} - n_{01} = \frac{n_{10} n_{01}}{n_{11}}$$

DSE assumptions

- Relies on assumptions
  - Closed population
  - Perfect linkage
  - No overcoverage
  - Inclusion in one source independent of inclusion in a second
  - Probability of inclusion homogeneous in at least one source
- …but often they don’t hold

Adaptations to deal with broken assumptions

- Homogeneity
  - Add categorical explanatory variable(s)
  - Homogeneity within classes
- Independence
  - Add categorical explanatory variable(s)
  - Independence conditional on included variables
- Write as loglinear model
  - Contingency table of sources and explanatory variables

Fully observed covariates

- Sources (registers) A, B
- Covariate X_1

- Classical DSE
  - [A][B]
- Independence conditional on X_1
  - $[AX_1][BX_1]$

Collapsibility

- Marginalisation: add up entries for one variable (and therefore eliminate it)
  - Eg add up over X_1 returns us to A × B table
- Marginalisation affects estimates in general
- If marginalisation over X_1 does *not* affect population size estimate, we say the model is collapsible over X_1
- If a model is collapsible over X_1, we say that X_1 is a passive covariate; otherwise it is active

Example

- Netherlands population register check
- People of Afghan, Iraqi & Iranian nationality
  - A = population register, B = police register
  - Covariates X_1 = sex; X_2 = age (4 bands)

model id	model definition	deviance	df	AIC	population size estimate
models fitted to the A × B (sub)table					
M_0	[A][B]	0.0	0	28.9	6170.3
models fitted to the A × B × X_1 (sub)table					
M_1	[AX_1][B]	548.5	1	558.5	6170.3
M_2	[A][BX_1]	1.1	1	11.1	6170.3
M_3	[AX_1][BX_1]	0.0	0	12.0	5696.1
models fitted to the A × B × X_1× X_2 table					
M_4	[AX_1][BX_2]	617.6	13	639.6	6170.3
M_5	[AX_1][BX_1][X_2]	228.6	15	246.6	5696.1
M_6	[AX_1X_2][B]	718.2	7	752.2	6170.3
M_7	[AX_1][AX_2][X_1X_2][B]	725.6	10	753.6	6170.3
M_8	[AX_1][BX_2][X_1X_2]	588.6	10	616.6	6179.4
M_9	[AX_1][BX_1][BX_2]	69.1	12	93.1	5696.1
M_{10}	[AX_1][BX_1][X_1X_2]	200.2	12	224.2	5696.1
M_{11}	[AX_1][AX_2][BX_2][BX_1]	65.9	9	95.9	5837.1
M_{12}	[AX_1][BX_1X_2]	4.9	6	40.9	5696.1
M_{13}	[AX_1][BX_1][BX_2][X_1X_2]	34.4	9	64.4	5696.1
M_{14}	[AX_1X_2][BX_1X_2]	0.0	0	48.0	5910.1
M_{15}	[AX_1X_2][BX_1][BX_2]	23.3	3	65.3	6257.1
M_{16}	[AX_1][AX_2][BX_2][BX_1][X_1X_2]	31.2	6	67.2	5831.4

model id	model definition	deviance	df	AIC	population size estimate
models fitted to the A × B (sub)table					
M_0	[A][B]	0.0	0	28.9	6170.3
models fitted to the A × B × X_1 (sub)table					
M_1	[AX_1][B]	548.5	1	558.5	6170.3
M_2	[A][BX_1]	1.1	1	11.1	6170.3
M_3	[AX_1][BX_1]	0.0	0	12.0	5696.1
models fitted to the A × B × X_1× X_2 table					
M_4	[AX_1][BX_2]	617.6	13	639.6	6170.3
M_5	[AX_1][BX_1][X_2]	228.6	15	246.6	5696.1
M_6	[AX_1X_2][B]	718.2	7	752.2	6170.3
M_7	[AX_1][AX_2][X_1X_2][B]	725.6	10	753.6	6170.3
M_8	[AX_1][BX_2][X_1X_2]	588.6	10	616.6	6179.4
M_9	[AX_1][BX_1][BX_2]	69.1	12	93.1	5696.1
M_{10}	[AX_1][BX_1][X_1X_2]	200.2	12	224.2	5696.1
M_{11}	[AX_1][AX_2][BX_2][BX_1]	65.9	9	95.9	5837.1
M_{12}	[AX_1][BX_1X_2]	4.9	6	40.9	5696.1
M_{13}	[AX_1][BX_1][BX_2][X_1X_2]	34.4	9	64.4	5696.1
M_{14}	[AX_1X_2][BX_1X_2]	0.0	0	48.0	5910.1
M_{15}	[AX_1X_2][BX_1][BX_2]	23.3	3	65.3	6257.1
M_{16}	[AX_1][AX_2][BX_2][BX_1][X_1X_2]	31.2	6	67.2	5831.4

Properties of invariance

- Same estimates are obtained by different processes
  - Marginalisation/collapsing
  - Model choice on full table
- Invariant population size estimates not only for nested models

Applications

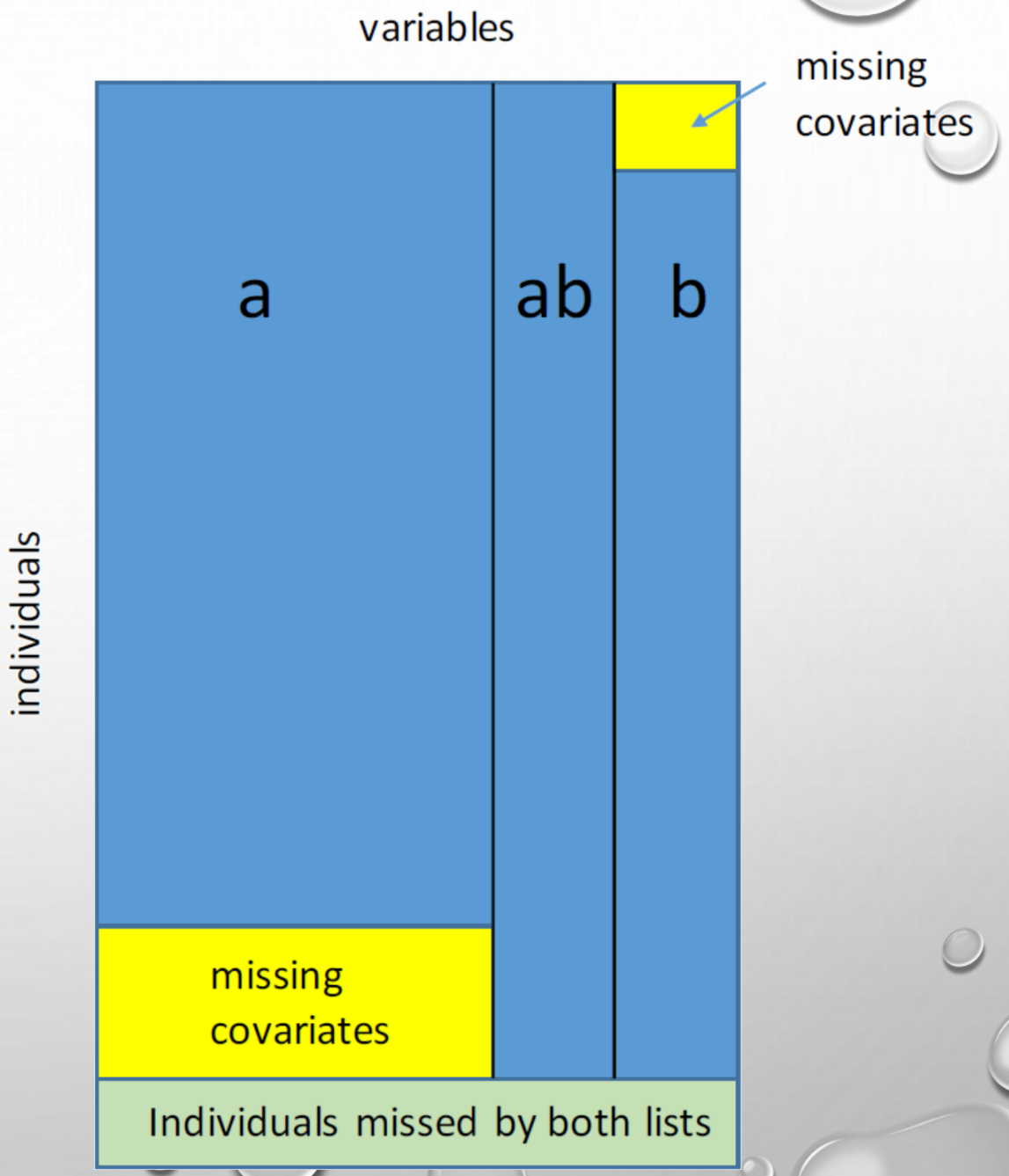
- Can introduce additional breakdown over a passive covariate without affecting total population size
- Subpopulation sizes vary by model
  - Choose best fitting model
  - M_2 lowest AIC

A	B	X_1	observed	M_0	M_1	M_2	M_3
1	1	1	972	1,085	629.2	976.5	972.0
1	1	0	113		455.8	108.5	113.0
0	1	1	234	255	234.0	229.5	234.0
0	1	0	21		21.0	25.5	21.0
1	0	1	14,883	26,254	15,225.8	14,883.0	14,883.0
1	0	0	11,371		11,028.2	11,371.0	11,371.0
0	0	1	0	6,170.3	5,662.2	3,497.9	3,582.9
0	0	0	0		508.1	2,672.5	2,113.2

Partially observed covariates

Partially observed covariates

- Where variables available in only one source, linkage generates missing values
- Number of completely missed individuals still to be estimated



Data and model fitting

- Need process to complete table
- EM algorithm
  - E-step: find missing values from current model
  - M-step: refit model to completed data
- Eg Afghan, Iranian & Iraqi data
  - X_3: marital status
  - X_4: region where apprehended
- Maximal model $[AX_4][X_3X_4][BX_3]$
  - Can examine reduced models
  - No more terms can be added

		B=1		B = 0
		$X_4 = 0$	$X_4 = 1$	X_4 missing
A = 1	$X_3 = 0$	259	539	13,898
	$X_3 = 1$	110	177	12,356
A = 0	X_3 missing	91	164	-

Completed table $[AX_4][X_3X_4][BX_3]$

		B=1		B = 0
		$X_4 = 0$	$X_4 = 1$	X_4 missing
A = 1	$X_3 = 0$	259	539	13,898
	$X_3 = 1$	110	177	12,356
A = 0	X_3 missing	91	164	-

		B=1		B=0	
		$X_4 = 0$	$X_4 = 1$	$X_4 = 0$	$X_4 = 1$
A = 1	$X_3 = 0$	259.0	539.0	4,510.8	9,387.2
	$X_3 = 1$	110.0	177.0	4,735.8	7,620.3
A = 0	$X_3 = 0$	63.9	123.5	1,112.4	2,150.2
	$X_3 = 1$	27.1	40.5	1,167.9	1,745.4

Assessing quality of estimates

Precision (Buckland & Garthwaite 1991)

- Parametric bootstrap
  - Generate multinomial probabilities from completed table
  - Draw sample of size $\widehat{N}$ with these probabilities
  - Aggregate to form of original table
  - Refit chosen model
  - Repeat

Sensitivity (Gerritse *et al*. 2015)

- Can't assess more complicated models
  - Insufficient observations
- Take *fixed* value of additional interaction
- Fit as offset in loglinear model
- Repeat for different (plausible) values

Comments on incomplete variables

- Easily extended to further variables on either or both sources
  - Variety of models grows quickly with number of variables
- Sometimes have same variable, but imperfectly measured
  - Treat as two separate variables
  - Decide on best solution
  - Or…

Multiple sources with common variables

Multiple sources

- Allow independence assumptions to be relaxed
  - In two source case always unable to fit [AB] – not enough data
  - With three sources can fit [AB]
  - Only [ABC] interaction unidentified with complete covariates
  - More complex models possible even with incomplete covariates

Population size of Māori in New Zealand

- Māori ethnicity key policy variable in New Zealand
- Estimates derived from population census
- Research to examine possibilities to use administrative data as basis for estimation
  - Linkage of sources in IDI-ERP
  - Four available sources (A, B, C, D)
    - Population census (2013)
    - DIA birth registrations data
    - MoE tertiary education enrolment
    - MoH National Health Index
  - Each source includes an ethnicity variable (a, b, c, d)
  - Available data matched population counts by in/out of source by Māori/non-Māori/missing
    - random rounding to base 3

Challenges

1. Population coverage
   - Sources cover defined parts of the population
   - How to deal with missing parts?
2. Extra level of missingness
   - People not in source have no recorded ethnicity
   - Some people in a source do not give their ethnicity
3. Ethnicity is not measured in the same way in each source
   - Consistent, single estimate?

1. Population coverage

- Census and MoH cover full population characteristics
- DIA covers only births since 1998 – people up to age 14 in 2013
- DoE covers tertiary education from late 1990s – approx. people aged 18-40 in 2013

- How to cope with parts of the population not in DIA and DoE?

Towards a general strategy

- Two sources (Zwane *et al*. 2004)
  - Algebraic solution
- A full coverage, B in north only
  - No matches of A to B in south
  - *If* homogeneous inclusion in A then can proceed without modification
- A north and centre, B centre and south
  - Bias if proceed without modification
  - Can use two-source approach to complete using interaction structure in centre region

- Let X_1 define region
- If collapsible over X_1 ($\Leftrightarrow$ X_1 is passive)
  - Incompleteness of B can safely be ignored

General result

- For any number of registers, and any combination of full and partial coverages of the population, the partial coverage will be ignorable if the model is collapsible over the covariate (or covariates) which defines the partial coverage(s)

Population coverage, Māori example

- Census + MoH + DIA
- Age defines coverage in DIA
  - Not related to coverage in census or MoH
  - Therefore model collapsible over age
  - Assumes homogeneous inclusion probabilities for age in census and MoH
- Similar argument for 4 sources
  - Age related to DIA and MoE
  - Model again collapsible over age

2. Extra level of missingness

	Census (A)	DIA (B)	MoH (C)	MoE (D)
non-Māori	3,225,804	574,077	3,527,874	1,763,463
Māori	560,427	236,673	617,205	405,063
-	20,619	6,045	188,781	20,424
x	595,140	3,585,195	68,130	2,213,040
Total	4,401,990	4,401,990	4,401,990	4,401,990

Maximal model
[Ac][ac][Ca]

Observed values

		C = 1		C = 0		Totals
		c = 0	c = 1	c = -	c = x	
A = 1	a = 0	3,004,335	31,995	150,840	38,634	3,225,804
	a = 1	108,189	435,465	12,405	4,368	560,427
	a = -	16,512	2,769	900	438	20,619
	a = x	398,838	146,976	24,636	-	570,450
Totals		3,527,874	617,205	188,781	43,440	4,377,300

Fitted values under [Ac][ac][Ca]

		C = 1		C = 0		Totals
		c = 0	c = 1	c = 0	c = 1	
A = 1	a = 0	3,170,294.8	33,797.9	38,616.0	411.6	3,243,110.3
	a = 1	111,242.5	448,084.8	877.6	3,534.9	563,739.8
A = 0	a = 0	402,709.4	10,770.8	4,905.2	131.2	418,516.6
	a = 1	14,130.7	142,839.1	111.5	1,126.8	158,208.1
Totals		3,698,377.4	635,482.6	44,510.3	5,204.5	4,383,574.8

Observed values

		C = 1			C = 0	Totals
		c = 0	c = 1	c = -	c = x	
A = 1	a = 0	3,004,335	31,995	150,840	38,634	3,225,804
	a = 1	108,189	435,465	12,405	4,368	560,427
	a = -	16,512	2,769	900	438	20,619
	a = x	398,838	146,976	24,636	-	570,450
Totals		3,527,874	617,205	188,781	43,440	4,377,300

Maximal model
[Ac][ac][Ca]

Fitted values under [Ac][ac][Ca]

		C = 1		C = 0		Totals
		c = 0	c = 1	c = 0	c = 1	
A = 1	a = 0	3,170,294.8	33,797.9	38,616.0	411.6	3,243,110.3
	a = 1	111,242.5	448,084.8	877.6	3,534.9	563,739.8
A = 0	a = 0	402,709.4	10,770.8	4,905.2	131.2	418,516.6
	a = 1	14,130.7	142,839.1	111.5	1,126.8	158,208.1
Totals		3,698,377.4	635,482.6	44,510.3	5,204.5	4,383,574.8

Parameter estimates, two registers

	Estimate	Std error	Z value	Pr
(intercept)	8.50	na	na	na
A	2.06	0.002	1248.1	< 0.001
C	4.41	0.005	865.1	< 0.001
a	-3.78	0.016	-233.7	< 0.001
c	-3.62	0.006	-585.2	< 0.001
A:c	-0.92	0.003	-268.9	< 0.001
C:a	0.43	0.016	27.1	< 0.001
a:c	5.94	0.007	903.7	< 0.001

Three sources

- Maximal model
  [ABc][ACb][BCa][Abc][Bac][Cab][abc]

	Census	DIA	MoH
non-Māori	3,690,122	3,648,027	3,776,521
Māori	729,123	771,217	642,724
full population	4,419,245	4,419,245	4,419,245

	Estimate	Std error	Z value	Pr
(intercept)	10.487	na	na	na
A	0.029	0.050	0.6	0.559
B	-4.001	0.033	-122.4	<0.001
C	2.254	0.050	45.1	<0.001
a	-5.756	0.201	-28.7	<0.001
b	-4.303	0.183	-23.5	<0.001
c	-4.866	0.025	-198.5	<0.001
A:B	0.332	0.005	67.9	<0.001
A:C	2.004	0.050	40.1	<0.001
B:C	2.030	0.032	62.7	<0.001
A:b	0.426	0.086	4.9	<0.001
A:c	-0.750	0.024	-31.7	<0.001
B:a	1.700	0.059	28.8	<0.001
B:c	1.080	0.014	79.4	<0.001
C:a	0.702	0.201	3.5	<0.001
C:b	0.661	0.182	3.6	<0.001
b:a	5.472	0.275	19.9	<0.001
c:a	5.075	0.022	232.1	<0.001
c:b	4.156	0.032	128.9	<0.001
A:B:c	-0.160	0.008	-19.7	<0.001
A:C:b	-0.857	0.086	-10.0	<0.001
B:C:a	-0.568	0.059	-9.6	<0.001
A:c:b	0.232	0.029	8.1	<0.001
B:c:a	-1.129	0.014	-81.7	<0.001
C:b:a	-0.249	0.275	-0.9	0.366
c:b:a	-2.136	0.029	-73.0	<0.001

Four sources

- Maximal model
  [ABCd][ABDc][ACDb][BCDa][ABcd][ACbd][ADbc][BCad][BDac][CDab][Abcd][Bacd][Cabd][Dabc][abcd]
  is unstable
- Restrict fifteen parameters to zero giving a stable model
  [ABcd][AC][ADbc][BCad][BDac][CDa][CDb][Abcd][Bacd][Dabc][abcd]

	Census	DIA	MoE	MoH
non-Māori	3,689,668	3,647,265	3,777,849	3,660,740
2.5%	3,687,698	3,643,429	3,775,988	3,658,421
97.5%	3,691,559	3,651,285	3,790,137	3,662,854
Māori	733,294	775,697	645,112	762,222
2.5%	731,608	772,236	643,533	760,103
97.5%	734,947	779,109	646,707	764,323
full population	4,422,962	4,422,962	4,422,962	4,422,962
2.5%	4,421,894	4,421,894	4,421,894	4,421,894
97.5%	4,424,080	4,424,080	4,424,080	4,424,080

3. One measure of Māori population

- Several options to proceed
  - Choose one measure as the best
  - Some combination – eg individuals recorded as Māori in 2 or more sources
  - Latent class analysis
- LCA
  - Assumes categorical latent variable
  - Observed variables independent conditional on latent variable

Latent class model

- $\pi_{rstu} = \sum_{x=1,2} \pi_x^X \pi_{r|x}^a \pi_{s|x}^b \pi_{t|x}^c \pi_{u|x}^d$
- Fitted to outcome table, model is [aX][bX][cX][dX]

		Census	DIA	MoH	MoE
	π_x	$\pi_{r=1\|x}^a$	$\pi_{s=1\|x}^b$	$\pi_{t=1\|x}^c$	$\pi_{u=1\|x}^d$
Class 1	0.827	0.004	0.016	0.003	0.015
Class 2	0.173	0.937	0.937	0.826	0.922

Embedded latent class model – Latent Class Multiple System Estimation (LCMSE)

- Replace [abcd] by [aX][bX][cX][dX]
- Delete terms involving two or more of a, b, c, d together – violates assumptions of latent class model
  - Eg [ABcd] – interaction between c and d is explained by X
- Thus maximal model is [ABCd][ABDc][ACDb][BCDa][aX][bX][cX][dX]

LCMSE

		Census	DIA	MoH	MoE
	π_x	$\pi^a_{r=1\|x}$	$\pi^b_{s=1\|x}$	$\pi^c_{t=1\|x}$	$\pi^d_{u=1\|x}$
Panel 1: Latent class analysis of MSE output					
Class 1	0.827	0.004	0.016	0.003	0.015
Class 2	0.173	0.937	0.937	0.826	0.922
Panel 2: Estimates for LCMSE					
Class 1	0.834	0.007	0.014	0.005	0.016
Class 2	0.166	0.957	0.958	0.847	0.959

Conclusions

- DSE suitable for estimating missed parts of datasets
  - Check assumptions and approaches to dealing with breakdowns
- Can use same procedure with extended model to complete covariates on only one source
  - EM algorithm
  - Extra checking for sources that do not cover the whole population
- Extend to multiple sources with different types of missingness
  - Treat different measurements as different variables
  - Heuristic choice of final variable, or
  - Latent class approach to combine information on different versions of the same variable
    - Estimates of measurement error rates on different versions

References

- This presentation is based on:

Van der Heijden, P.G.M., Cruyff, M., Smith, P.A., Bycroft, C., Graham, P. & Matheson-Dunning, N. (2020) Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Māori population in New Zealand. https://arxiv.org/abs/2007.00929.

- References in slides:

Buckland, S. & Garthwaite, P. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics* **47** 255-268. doi: 10.2307/2532510.

Gerritse, S.C., van der Heijden, P.G.M. & Bakker, B.F.M. (2015) Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of Official Statistics* **31** 357-379.

Zwane, E., van der Pal-de Bruin, K. & van der Heijden, P.G.M. (2004) The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine* **23** 2267-2281. doi: 10.1002/sim.1818

Contact

Paul Smith

p.a.smith@soton.ac.uk