Estimating migration flows during the inter-censal period in Northern Ireland by health characteristics measured in the Census

John W McDonald

Outline

- migration of British residents born in New Commonwealth - 4 by 4 table
- migration in NI by health status measured in the Census - 3 by 3 table(s)
- features of migration data & QI model
- Bayesian framework for inference
- results, discussion and model extensions

Residence in 1966 & 1971

Residence in 1966	Residence in 1971 CC	L&Y	WM	GL	Total
CC	118	12	7	23	160
L&Y	14	2 127	86	130	2 357
WM	8	69	2 548	107	2 732
GL	12	110	88	7 712	7 922
Total	152	2 318	2 729	7 972	13 171

CC: Central Clydeside L&Y: Lancashire & Yorkshire
WM: West Midlands GL: Greater London

Migration from 1966 to 1971 (flows)

Residence in 1966	Residence in 1971 CC	L&Y	WM	GL	Total
CC		12	7	23	42
L&Y	14		86	130	230
WM	8	69		107	184
GL	12	110	88		210
Total	34	191	181	260	666

CC: Central Clydeside L&Y: Lancashire & Yorkshire

WM: West Midlands GL: Greater London

Model origin-destination flows

$$\log \mathsf{flow}_{ij} = \mu + \alpha_i + \beta_j, + \lambda_{ij} \qquad i \neq j$$

μ - overall level of migration

α_i - pushing factor for i th origin

β_j - pulling factor for j th destination

λ_{ij} - origin i and destination j interaction

quasi-independence (QI) if $\lambda_{ij} = 0$ for $i \neq j$

Adjusted residuals: QI off diagonal

Residence in 1966	Residence in 1971 CC	L&Y	WM	GL
CC		-.27	-1.38	1.47
L&Y	.81		1.25	-1.71
WM	-.36	-.56		.74
GL	-.47	.72	-.44	

$L^2 = 4.367$ on 5 df, p $= .50$

Features of migration flow data

- interested in off-diagonal cells (flows)
- often flows partially observed, i.e. some/all off-diagonal cell counts missing
- accurate external estimates of the marginal totals may be available

Only marginal totals of flows known

Residence in 1966	Residence in 1971 CC	L&Y	WM	GL	Total
CC					42
L&Y					230
WM					184
GL					210
Total	34	191	181	260	666

CC: Central Clydeside   L&Y: Lancashire & Yorkshire

WM: West Midlands   GL: Greater London

LS, SLS & NILS have differences

ONS Longitudinal Study (LS)

Scottish Longitudinal Study (SLS)

Northern Ireland Longitudinal Study (NILS)

LS & SLS - initial sample from the census

NILS - initial sample from health card registration data

NILS linked twice yearly to health card registration data

	E&W LS	SLS	NILS
Number Dates	4	20	104
Size	500,000 (1%)	300,000 (5%)	500,000 (28%)
Start Date	1971	1991	2001
Number of Censuses	4	2	1
Data on all 3 LSs	Births to sample mothers		
	Births of sample members		
	Deaths of sample members		
	Immigrants		
	Embarkations		
	Widow(er)hoods		

	E&W LS	SLS	NILS
Births to Sample Fathers			✓
Stillbirths/ Infant Deaths	✓		✓
Internal Migration			✓
Cancer Registrations	✓	✓	
Hospital Episodes		✓	
Education		✓	
Marriages		✓	✓
Claimant Count	✓		

NILS data

access from NISRA safe setting

research must relate to health

all primary data analysis in safe setting

outputs must be cleared so they do not raise confidentiality or disclosure issues

must work with aggregated categories so all cell counts $>$ disclosure value of 10

NILS flows crossing LGD boundaries

Residence July 2007	Residence June 2008 Belfast	Eastern	Western
Belfast	0		
Eastern			
Western			

by long term limiting illness reported in Census: Yes, No, non-Census NILS member

Source: Northern Ireland Longitudinal Study

NILS flows crossing LGD boundaries

Residence July 2007	Residence June 2008 Belfast	Eastern	Western	Total
Belfast	0	186	25	211
Eastern	111	263	64	438
Western	27	60	147	234
Total	138	509	236	883

long term limiting illness = Yes

Source: Northern Ireland Longitudinal Study

Adjusted residuals: QI not (1,1) cell

Residence July 2007	Residence June 2008 Belfast	Eastern	Western
Belfast	0	7.31	-7.31
Eastern	4.22	3.50	-7.08
Western	-4.22	-10.37	15.41

long term limiting illness = Yes

$L^2 = 223.4$ on 3 df

Source: Northern Ireland Longitudinal Study

Adjusted residuals: QI off diagonal

Residence July 2007	Residence June 2008 Belfast	Eastern	Western
Belfast	0	1.88	-1.88
Eastern	-1.88	0	1.88
Western	1.88	-1.88	0

long term limiting illness = Yes

$L^2 = 3.496$ on 1 df, p= .062

Source: Northern Ireland Longitudinal Study

NILS flows crossing LGD boundaries

Residence July 2007	Residence June 2008 Belfast	Eastern	Western	Total
Belfast	0	896	218	1114
Eastern	581	1550	366	2497
Western	231	340	707	1278
Total	812	2786	1291	4889

long term limiting illness = No

Source: Northern Ireland Longitudinal Study

Adjusted residuals: QI not (1,1) cell

Residence July 2007	Residence June 2008 Belfast	Eastern	Western
Belfast	0	10.18	-10.18
Eastern	3.67	12.44	-16.63
Western	-3.67	-22.76	29.19

long term limiting illness = No

$L^2 = 814.1$ on 3 df

Source: Northern Ireland Longitudinal Study

Adjusted residuals: QI off diagonal

Residence July 2007	Residence June 2008 Belfast	Eastern	Western
Belfast	0	4.29	-4.29
Eastern	-4.29	0	4.29
Western	4.29	-4.29	0

long term limiting illness = No

$L^2 = 18.32$ on 1 df, p= .000

Source: Northern Ireland Longitudinal Study

Notation

y_{ij} – true flow from region i to region j

z_{ij} – reported flow from region i to region j

reported flows $\boldsymbol{z} = \{\, z_{ij},\ i \neq j \,\}$ complete

true unknown flows $\boldsymbol{y} = \{\, y_{ij},\ i \neq j \,\}$

assume marginal totals of $\boldsymbol{y}$ are known

Bayesian inference

parameters random with distributions

- prior to data
- posterior given data

apply Bayes Theorem

posterior $\propto$ likelihood $\times$ prior

posterior distribution used for inference

Bayesian framework

lognormal approximates discrete distribution

$$\log z_{ij} \overset{\text{ind}}{\sim} N\left(\log y_{ij}, \sigma^2\right) \qquad i \neq j$$

on the log scale, reported values are

- □ unbiased
- □ of the same accuracy

Inference for model unknowns

$f(\boldsymbol{y}, \sigma^2)$ is *prior* distribution for $(\boldsymbol{y}, \sigma^2)$

inference based on *posterior* distribution

$$f(\boldsymbol{y}, \sigma^2 \mid \boldsymbol{z}) = \frac{f(\boldsymbol{z} \mid \boldsymbol{y}, \sigma^2)\, f(\boldsymbol{y}, \sigma^2)}{f(\boldsymbol{z})}$$

inference about a given flow obtained from its marginal distribution $f(y_{ij} \mid \boldsymbol{z})$

Inference for model unknowns . . .

assume true flows follow prior distribution centred on quasi-independence (QI)

$$\log y_{ij} \sim N(\mu + \alpha_i + \beta_j, \tau^2) \qquad i \neq j$$

Inference for model unknowns . . .

main idea: produce estimates which best reflect the observed values given that

- they must also satisfy known margins
- quasi-independence fitted values often provide good estimates of the true flows

Hierarchical model

$$\log z_{ij} \;\sim\; N\left(\log y_{ij},\, \sigma^2\right) \qquad i \neq j$$

$$\log y_{ij} \sim N(\,\mu + \alpha_i + \beta_j,\, \tau^2\,) \times I[\,\bullet\,]$$

indicator $I\left[\sum_j y_{ij} = y_{i+}\,, \sum_i y_{ij} = y_{+j}\right]$
maintains the marginal constraint

Hierarchical model . . .

hyperparameters: set by considerations of data quality and belief in QI model

posterior distribution too complex for direct computations so use simulation methods to learn about the posterior

Markov chain Monte Carlo methods

hybrid Gibbs & Metropolis-Hastings steps

Gibbs steps: conditional conjugacy of priors
$\Rightarrow$ sampling from conditionals easy

$\mathbf{Y}$ must always satisfy the marginal constraints so exact generation is difficult and we use a Metropolis-Hastings step

Metropolis-Hastings step

$$\text{proposal} \quad \mathbf{Y}' = \mathbf{Y} + \begin{pmatrix} & +\epsilon & -\epsilon & 0 \\ 0 & & 0 & 0 \\ 0 & 0 & & 0 \\ 0 & -\epsilon & +\epsilon & \end{pmatrix}$$

$\epsilon \sim$ uniform with bounds set by counts in selected rows/columns to ensure that no entries in the proposal can be negative

Metropolis-Hastings step . . .

proposal $\mathbf{Y}'$ is accepted with probability

$$\min\{1, f(\mathbf{Y}')/f(\mathbf{Y})\}$$

where f is given by full joint density with all other parameters at their current values

do not insist that ϵ is integer, which makes construction of an irreducible Markov chain to sample from much more straightforward

Metropolis-Hastings step . . .

use QI fitted values as initial $\mathbf{Y}$ values for the MCMC simulation

QI fitted values are useful for comparison with the results of the MCMC analysis

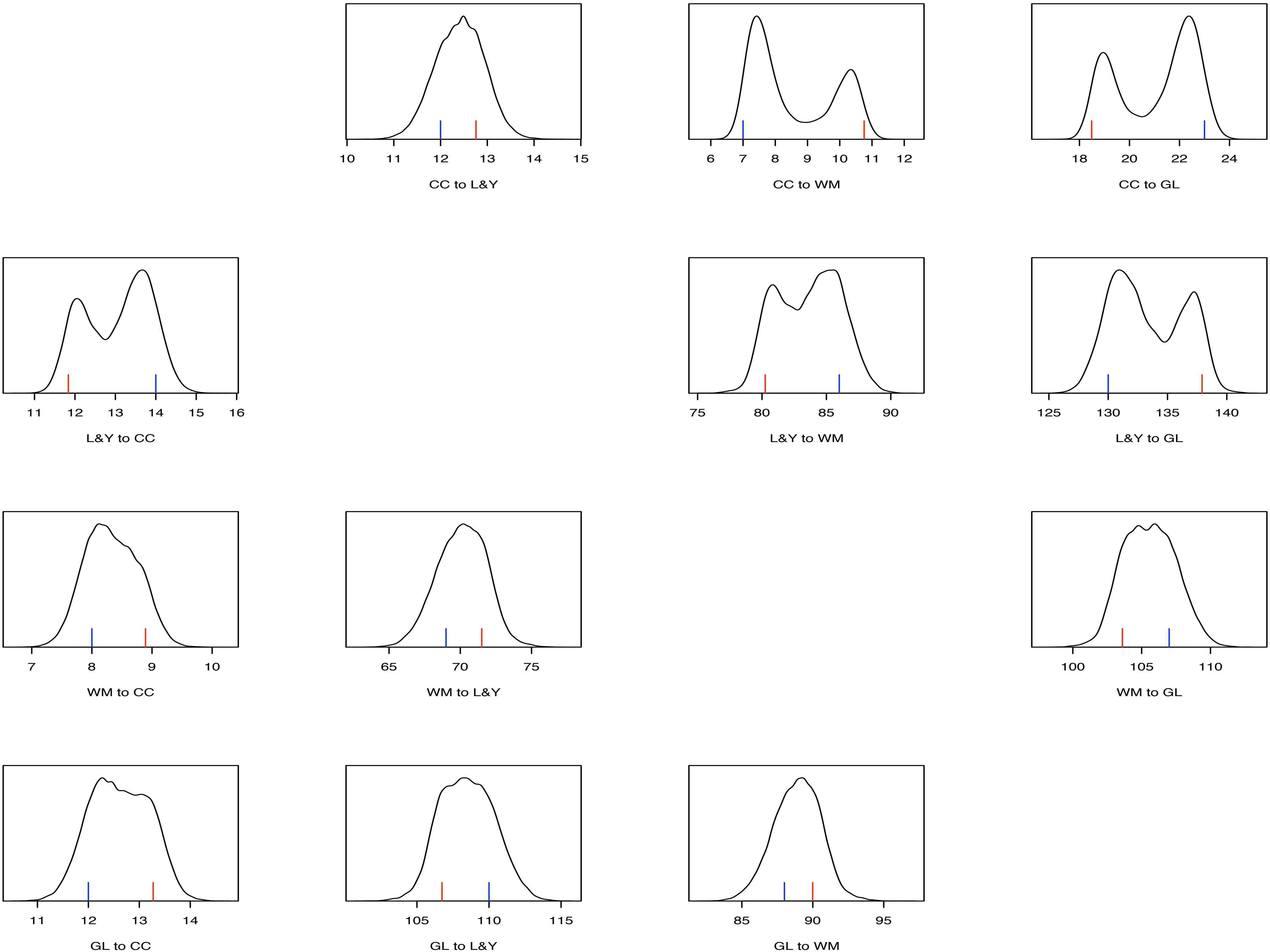
Migration from 1966 to 1971 (flows)

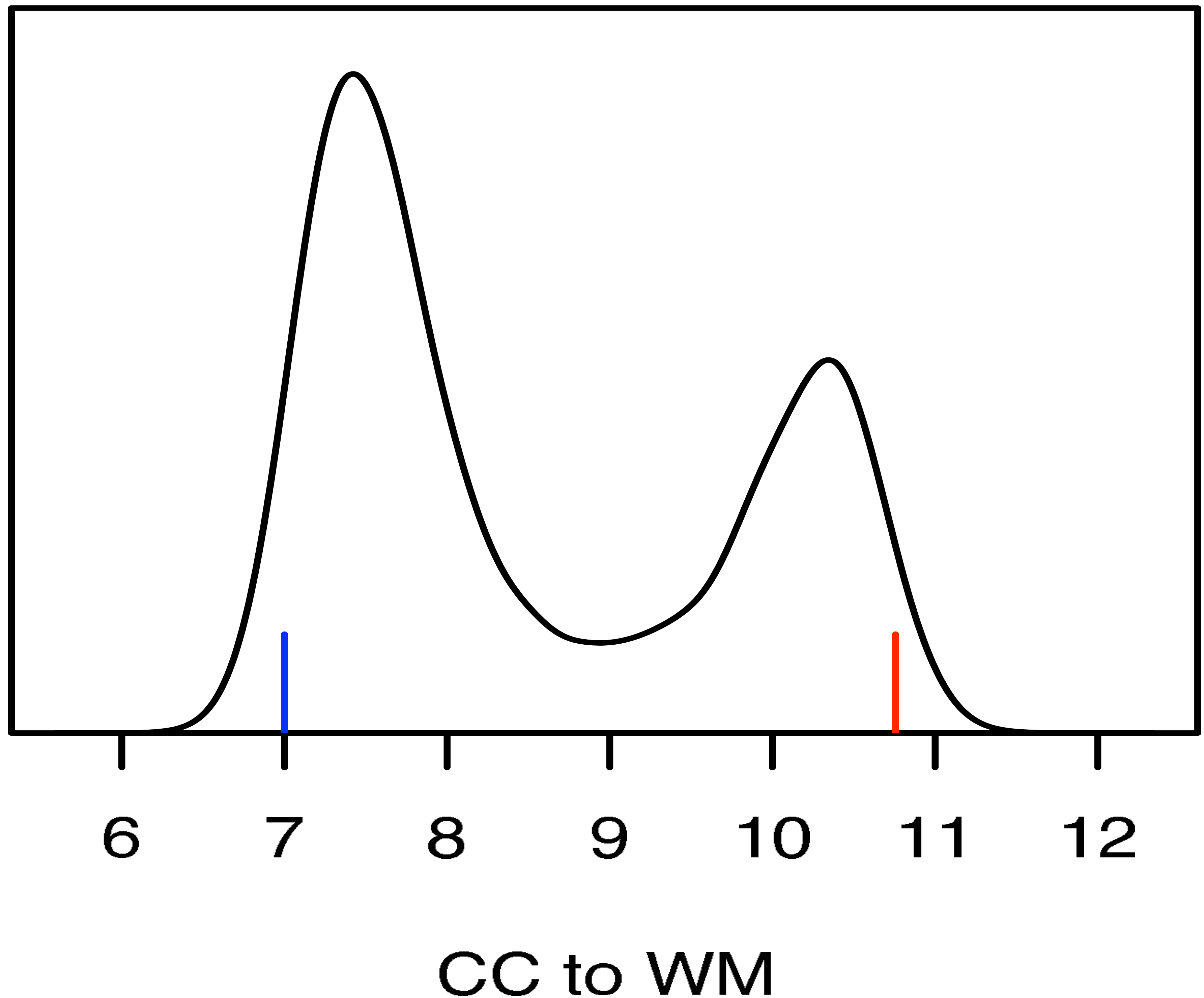
Residence in 1966	Residence in 1971 CC	L&Y	WM	GL	Total
CC		12	7	23	42
L&Y	14		86	130	230
WM	8	69		107	184
GL	12	110	88		210
Total	34	191	181	260	666

margins of flow table ($\mathbf{Z}$) were used as true known margins of $\mathbf{Y}$ for MCMC simulation

MCMC simulation results

- 500 000 iterations
- estimated posterior densities for each entry of true migration flow matrix $\mathbf{Y}$
- posteriors maximized close to z_{ij}
- bimodal posterior when discrepancy between observed & QI values $\Rightarrow$ considerable uncertainty as to the magnitude of the true flow


10 11 12 13 14 15
CC to L&Y
6 7 8 9 10 11 12
CC to WM
18 20 22 24
CC to GL
11 12 13 14 15 16
L&Y to CC
75 80 85 90
L&Y to WM
125 130 135 140
L&Y to GL
7 8 9 10
WM to CC
65 70 75
WM to L&Y
100 105 110
WM to GL
11 12 13 14
GL to CC
105 110 115
GL to L&Y
85 90 95
GL to WM


6
7
8
9
10
11
12
CC to WM

MCMC simulation results

different priors lead to different behavior

if τ^2 takes larger values, representing lower confidence in QI model, then secondary mode around the QI estimate diminishes

correspondingly, credible interval estimates become narrower

Discussion and model extensions

framework assumes that the margins of migration flow matrix are known

assumption of known margins can be relaxed to allow for error in the margins, presumably with a smaller variance than the errors in the cell counts

Discussion and model extensions . . .

prior belief: QI structure of true flows

QI structure likely to be oversimplistic

regions sharing borders are likely to have greater flows than the QI model predicts, since there will be an interaction term which the QI model lacks

Discussion and model extensions . . .

possible to add a parameter for contiguity
into the modelling process

$$\log y_{ij} = \mu + \alpha_i + \beta_j + \gamma\, d_{ij} \qquad i \neq j$$

$$d_{ij} = \begin{cases} 1 & i,\, j \text{ share a border} \\ 0 & \text{otherwise} \end{cases}$$

Bayesian modelling framework

individual migration flow estimates

with associated measures of precision

natural way to incorporate prior information about migration process and data quality

can combine multiple data sources

calibration to known marginal totals

Acknowledgement

"The help provided by staff of the Northern Ireland Longitudinal Study (NILS) and NILS Research Support Unit is acknowledged. NILS is funded by the HSC R&D function of the Public Health Agency. The authors alone are responsible for the interpretation of the data."