UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Statistical Modelling of Population Processes

Peter W. F. Smith
pws@soton.ac.uk

Southampton Statistical Sciences Research Institute (S3RI)
University of Southampton

Recent advances in the analysis of categorical and count data
LSE, December 2013

Contents

Rationale for Bayesian methods

- Increasing need for population estimates and forecasts to include uncertainty

Rationale for Bayesian methods

- Increasing need for population estimates and forecasts to include uncertainty
- Methods for integrating uncertainty are underdeveloped in these areas

Rationale for Bayesian methods

- Increasing need for population estimates and forecasts to include uncertainty
- Methods for integrating uncertainty are underdeveloped in these areas
- The Bayesian approach offers a natural, flexible and probabilistic framework to integrate:
    - variability in data
    - uncertainty in model parameters
    - uncertainty in model choice
    - expert judgements, including their uncertainty

Example 1: Cohort component population projections

- **Objective**: To explore the consequences of choosing different specifications of age-specific fertility and mortality forecasting models in a closed cohort projection model

Example 1: Cohort component population projections

- **Objective**: To explore the consequences of choosing different specifications of age-specific fertility and mortality forecasting models in a closed cohort projection model

- **Acknowledgements**
Research undertaken by the Centre for Population Change (CPC) funded by the Economic and Social Research Council (ESRC)

Thanks to other members of the CPC Modelling Team: Jakub Bijak, Jon Forster, Xiaoling Ou, James Raymer and Arkadiusz Wiśniowski

Forecasting age-specific mortality

- Model M1: Poisson-log-normal extension of the Lee-Carter model

$$D_{x,t} \sim Po(E_{x,t} m_{x,t})$$

$$\log(m_{x,t}) = \alpha_x + \beta_x k_t + \xi_{x,t}$$
$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$

Forecasting age-specific mortality

- Model M1: Poisson-log-normal extension of the Lee-Carter model

$$D_{x,t} \sim Po(E_{x,t}m_{x,t})$$

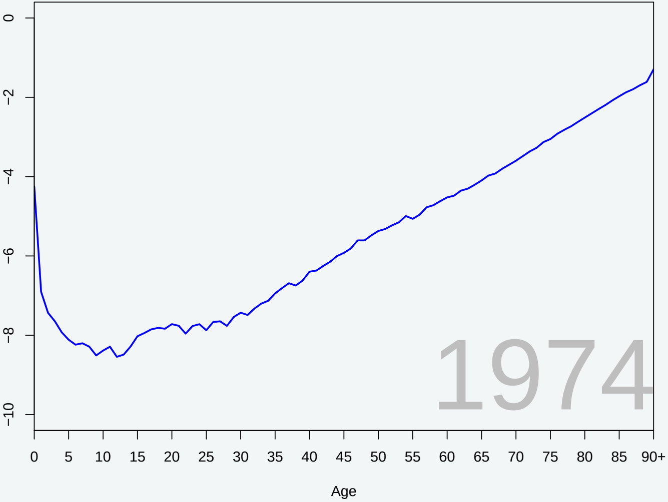
$$\log(m_{x,t}) = \alpha_x + \beta_x k_t + \xi_{x,t}$$
$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$

- Model M2: Poisson-log-normal extension of the Lee-Carter model with cohort component (Haberman and Renshaw, 2006)

$$D_{x,t} \sim Po(E_{x,t}m_{x,t})$$

$$\log(m_{x,t}) = \alpha_x + \beta_x k_t + \gamma_x c_{t-x} + \xi_{x,t}$$
$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$
$$c_t = p_3 + p_4 c_{t-1} + \xi_t^c$$

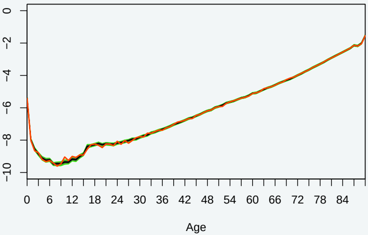
Logarithm of female $m_{x,t}$ in the UK, 1974 to 2007



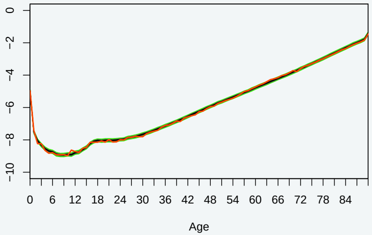
Model fit, 1990 and 2007

1990, Model M1

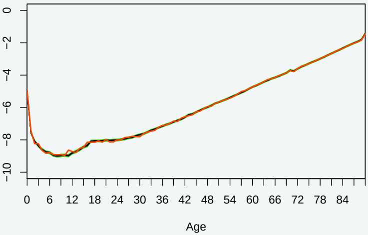
1990, Model M2



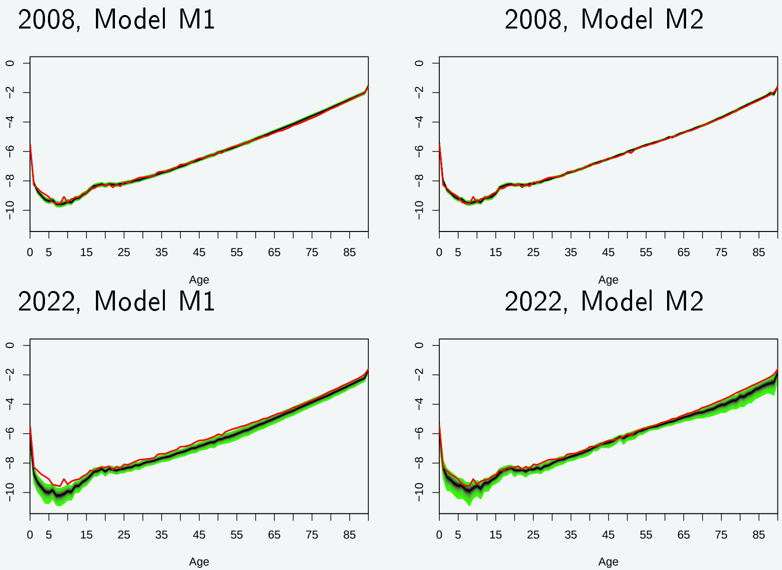
2007, Model M1



2007, Model M2



Forecasted female mortality, 2008 and 2022



- Model F1: Lee-Carter model

$$\log(f_{x,t}) = \alpha_x + \beta_x k_t + \xi_{x,t}$$

$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$

Forecasting age-specific fertility

- Model F1: Lee-Carter model

$$\log(f_{x,t}) = \alpha_x + \beta_x k_t + \xi_{x,t}$$

$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$

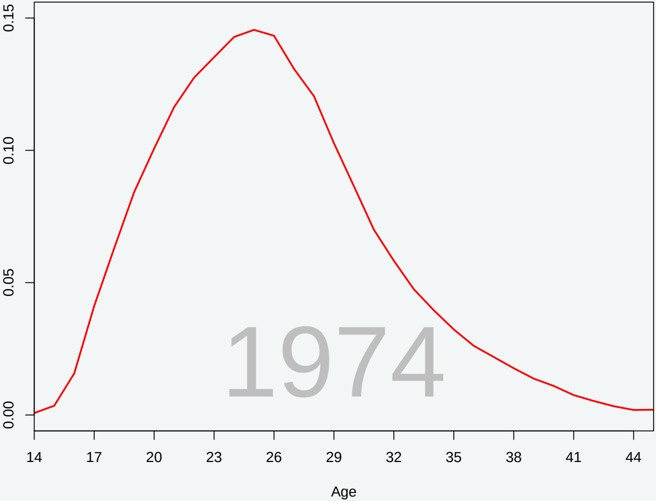
- Model F2: Lee-Carter model with cohort component

$$\log(f_{x,t}) = \alpha_x + \beta_x k_t + \gamma_x c_{t-x} + \xi_{x,t}$$

$$k_t = p_1 + p_2 k_{t-1} + \xi_t^k$$

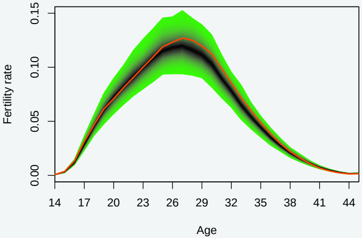
$$c_t = p_3 + p_4 c_{t-1} + \xi_t^c$$

Fertility $f_{x,t}$ in the UK, 1974 to 2007

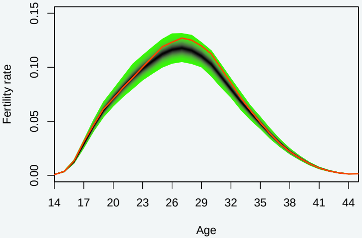


Model fit, 1990 and 2007

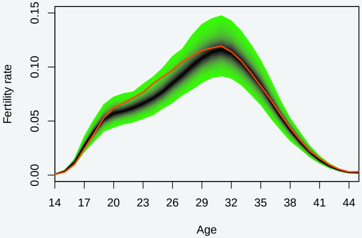
1990, Model F1



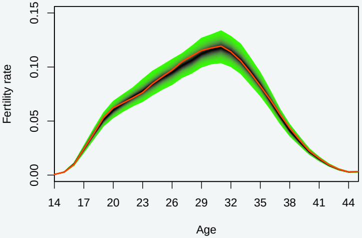
1990, Model F2



2007, Model F1

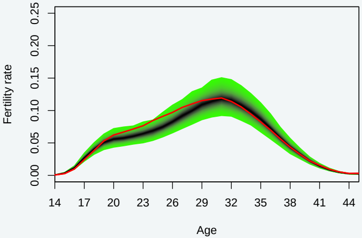


2007, Model F2

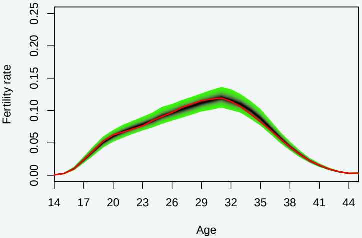


Forecasted fertility, 2008 and 2022

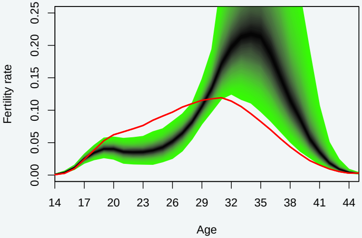
2008, Model F1



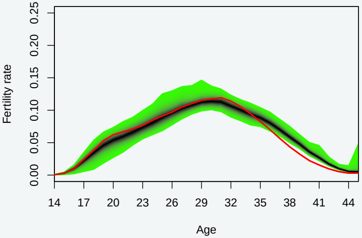
2008, Model F2



2022, Model F1



2022, Model F2



Female population projection model

$$
\begin{bmatrix} P_{0,t+1} \\ P_{1,t+1} \\ \\ \vdots \\ \\ P_{89,t+1} \\ P_{90+,t+1} \end{bmatrix}
=
\begin{bmatrix}
0 & \dots & b_{14,t} & \dots & b_{45,t} & \dots & 0 \\
s_{0,t} & 0 & 0 & & \dots & & 0 \\
0 & s_{1,t} & 0 & & \dots & & 0 \\
\vdots & & \ddots & & & & \vdots \\
\vdots & & & \ddots & & & \vdots \\
0 & 0 & \dots & & s_{88,t} & 0 & 0 \\
0 & 0 & \dots & & 0 & s_{89,t} & s_{90+,t}
\end{bmatrix}
\begin{bmatrix} P_{0,t} \\ P_{1,t} \\ \\ \vdots \\ \\ P_{89,t} \\ P_{90+,t} \end{bmatrix}
$$

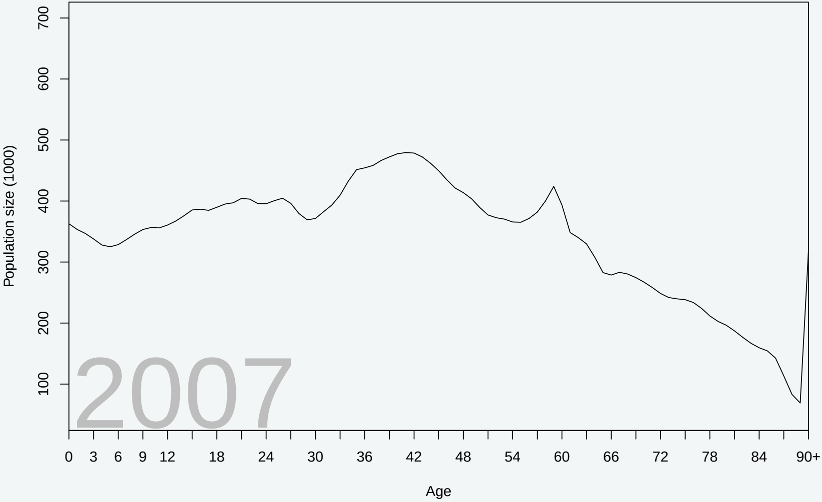
where life table birth rates come from the fertility model:

$$b_{x,t} = g(f_{x,t})$$

and survivorship rates come from the mortality model:

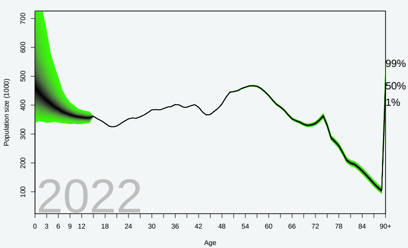
$$s_{x,t} = g(m_{x,t})$$

Forecasted female population by age, M1+F1

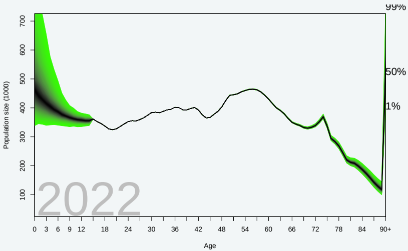


Forecasted female population by age, 2022

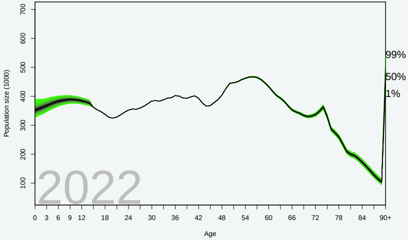
Model M1+F1



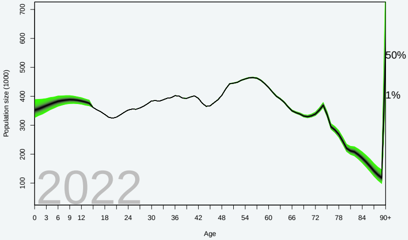
Model M2+F1



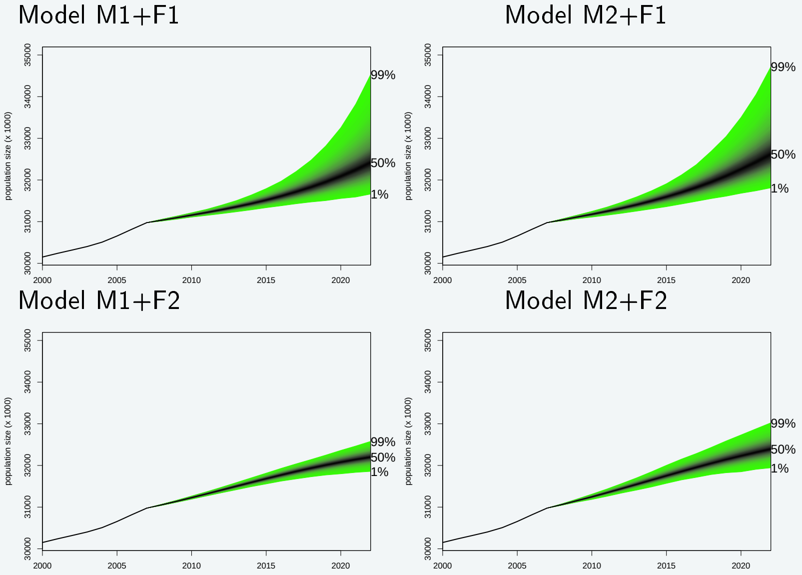
Model M1+F2



Model M2+F2



Forecasted female population totals



Example 2: Integrated Modelling of European Migration

- Project funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)

Example 2: Integrated Modelling of European Migration

- Project funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)
- Brings together expertise in modelling, data and uncertainty

Example 2: Integrated Modelling of European Migration

- Project funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)
- Brings together expertise in modelling, data and uncertainty
- Southampton Statistical Sciences Research Institute (S3RI)
  - James Raymer (PI), Jonathan J. Forster, Peter W.F. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy J. Abel

Example 2: Integrated Modelling of European Migration

- Project funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)
- Brings together expertise in modelling, data and uncertainty
- Southampton Statistical Sciences Research Institute (S3RI)
    - James Raymer (PI), Jonathan J. Forster, Peter W.F. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy J. Abel
- Netherlands Interdisciplinary Demographic Institute (NIDI)
    - Rob van der Erf and Joop de Beer

Example 2: Integrated Modelling of European Migration

- Project funded by New Opportunities for Research Funding Agency Co-operation in Europe (NORFACE)
- Brings together expertise in modelling, data and uncertainty
- Southampton Statistical Sciences Research Institute (S3RI)
    - James Raymer (PI), Jonathan J. Forster, Peter W.F. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy J. Abel
- Netherlands Interdisciplinary Demographic Institute (NIDI)
    - Rob van der Erf and Joop de Beer
- University of Oslo
    - Nico Keilman and Solveig Christiansen

Objectives

- To provide a general framework for modelling migration flows between countries in the world in the context of inconsistent, inadequate and missing data

Objectives

- To provide a general framework for modelling migration flows between countries in the world in the context of inconsistent, inadequate and missing data
- Focus is on recent international migration flows between European Union and European Free Trade Association (EFTA) countries, using primarily publicly available sources, as well as qualitative information from experts

Double-entry matrix for selected countries, 2003

Origin		BE	CZ	DK	DE	EE	GR	ES	FR	IE	IT
		Destination									
BE	I		80	587	4291	...	...	3037	...	...	1959
	E		...	...	...	...	...	...	...	...	...
CZ	I	...		232	9258	...	...	388	...	...	915
	E	78		47	950	2	66	70	283	31	197
DK	I	...	65		2693	...	...	764	...	...	281
	E	511	180		2540	133	229	1720	1333	264	782
DE	I	...	1228	3221		...	...	13746	...	...	12902
	E	4623	8909	2712		597	18106	16236	19060	2415	33802
EE	I	...	4	169	947		...	60	...	...	103
	E	...	...	...	...		...	...	...	...	...
GR	I	...	57	278	12959	...		273	...	...	638
	E	...	...	...	...	...		...	...	...	...
ES	I	...	103	1665	14647	...	...		...	...	2051
	E	647	34	130	2109	4	38		2474	487	801
FR	I	...	462	1488	18133	...	...	8847		...	4647
	E	...	...	...	...	...	...	...		...	...
IE	I	...	45	306	2046	...	...	1649	...		292
	E	...	...	...	...	...	...	...	...		...
IT	I	...	274	895	23702	...	...	5796	...	...	
	E	1414	20	155	9778	1	211	895	2933	130	

I = receiving country's reported flow; E = sending country's reported flow; ... = no reported data available

Double-entry matrix for selected countries, 2003

Origin		BE	CZ	DK	DE	EE	GR	ES	FR	IE	IT
		Destination									
BE	I		80	587	4291	...	...	3037	...	...	1959
	E		...	...	...	...	...	...	...	...	...
CZ	I	...		232	9258	...	...	388	...	...	915
	E	78		47	950	2	66	70	283	31	197
DK	I	...	65		2693	...	...	764	...	...	281
	E	511	180		2540	133	229	1720	1333	264	782
DE	I	...	1228	3221		...	...	$\boxed{13746}$	...	...	12902
	E	4623	8909	2712		597	18106	$\boxed{16236}$	19060	2415	33802
EE	I	...	4	169	947		...	60	...	...	103
	E	...	...	...	...		...	...	...	...	...
GR	I	...	57	278	12959	...		273	...	...	638
	E	...	...	...	...	...		...	...	...	...
ES	I	...	103	1665	14647	...	...		...	...	2051
	E	647	34	130	2109	4	38		2474	487	801
FR	I	...	462	1488	18133	...	...	8847		...	4647
	E	...	...	...	...	...	...	...		...	...
IE	I	...	45	306	2046	...	...	1649	...		292
	E	...	...	...	...	...	...	...	...		...
IT	I	...	274	895	23702	...	...	5796	...	...	
	E	1414	20	155	9778	1	211	895	2933	130	

I = receiving country's reported flow; E = sending country's reported flow; . . . = no reported data available

Double-entry matrix for selected countries, 2003

		Destination									
Origin		**BE**	**CZ**	**DK**	**DE**	**EE**	**GR**	**ES**	**FR**	**IE**	**IT**
BE	I		80	587	4291	...	...	3037	...	...	1959
	E		...	...	...	...	...	...	...	...	...
CZ	I	...		232	9258	...	...	388	...	...	915
	E	78		47	950	2	66	70	283	31	197
DK	I	...	65		2693	...	...	764	...	...	281
	E	511	180		2540	133	229	1720	1333	264	782
DE	I	...	1228	3221		...	...	13746	...	...	12902
	E	4623	8909	2712		597	18106	16236	19060	2415	33802
EE	I	...	4	169	947		...	60	...	...	103
	E	...	...	...	...		...	...	...	...	...
GR	I	...	57	278	12959	...		273	...	...	638
	E	...	...	...	...	...		...	...	...	...
ES	I	...	103	1665	14647	...	...		...	...	2051
	E	647	34	130	2109	4	38		2474	487	801
FR	I	...	462	1488	18133	...	...	8847		...	4647
	E	...	...	...	...	...	...	...		...	...
IE	I	...	45	306	2046	...	...	1649	...		292
	E	...	...	...	...	...	...	...	...		...
IT	I	...	274	895	23702	...	...	5796	...	...	
	E	1414	20	155	9778	1	211	895	2933	130	

I = receiving country's reported flow; E = sending country's reported flow; . . . = no reported data available

Double-entry matrix for selected countries, 2003

		Destination									
Origin		BE	CZ	DK	DE	EE	GR	ES	FR	IE	IT
BE	I		80	587	4291	...	...	3037	...	...	1959
	E		...	...	...	...	...	...	...	...	...
CZ	I	...		232	9258	...	...	388	...	...	915
	E	78		47	950	2	66	70	283	31	197
DK	I	...	65		2693	...	...	764	...	...	281
	E	511	180		2540	133	229	1720	1333	264	782
DE	I	...	1228	3221		...	...	13746	...	...	12902
	E	4623	8909	2712		597	18106	16236	19060	2415	33802
EE	I	...	4	169	947		...	60	...	...	103
	E	...	...	...	...		...	...	...	...	...
GR	I	...	57	278	12959	...		273	...	...	638
	E	...	...	...	...	...		...	...	...	...
ES	I	...	103	1665	14647	...	...		...	...	2051
	E	647	34	130	2109	4	38		2474	487	801
FR	I	...	462	1488	18133	...	...	8847		...	4647
	E	...	...	...	...	...	...	...		...	...
IE	I	...	45	306	2046	...	...	1649	...		292
	E	...	...	...	...	...	...	...	...		...
IT	I	...	274	895	23702	...	...	5796	...	...	
	E	1414	20	155	9778	1	211	895	2933	130	

I = receiving country's reported flow; E = sending country's reported flow; ... = no reported data available

Solutions

- A Bayesian model has been developed to harmonise and correct the inadequacies in the available data and to estimate the completely missing flows

Solutions

- A Bayesian model has been developed to harmonise and correct the inadequacies in the available data and to estimate the completely missing flows
- The methodology is integrated and provides a synthetic data base with measures of uncertainty for international migration flows and other model parameters

Solutions

- A Bayesian model has been developed to harmonise and correct the inadequacies in the available data and to estimate the completely missing flows
- The methodology is integrated and provides a synthetic data base with measures of uncertainty for international migration flows and other model parameters
- There are two key design aspects of the methodology:
  - The development of the underlying statistical model
  - The specification and elicitation of relevant expert prior information

Modelling framework

- Scope: flows between 31 EU and EFTA countries, 2002-2008

Modelling framework

- Scope: flows between 31 EU and EFTA countries, 2002-2008
- Adopted United Nations 1998 recommended definition

Modelling framework

- Scope: flows between 31 EU and EFTA countries, 2002-2008
- Adopted United Nations 1998 recommended definition
- Data problems:

  Availability $\Rightarrow$ Migration model (including economic, demographic and geographic explanatory variables)

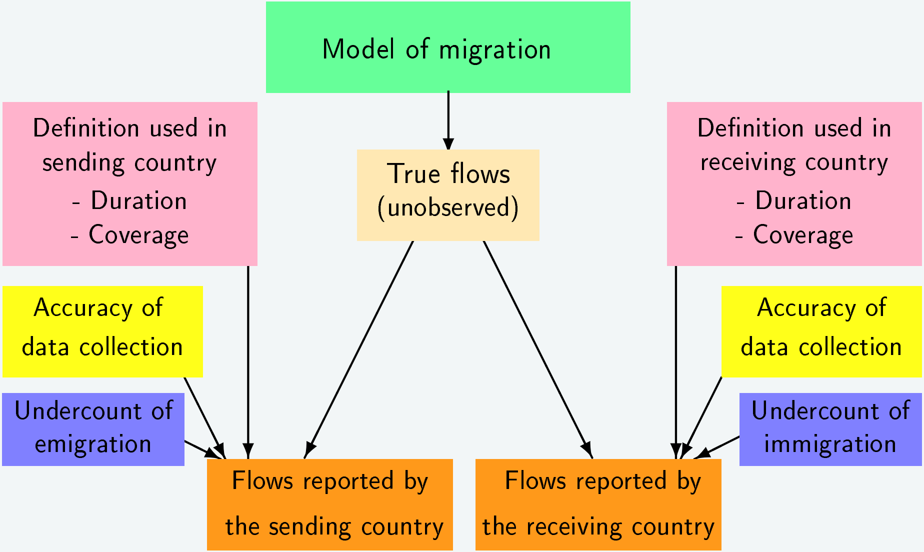
Modelling framework

- Scope: flows between 31 EU and EFTA countries, 2002-2008
- Adopted United Nations 1998 recommended definition
- Data problems:

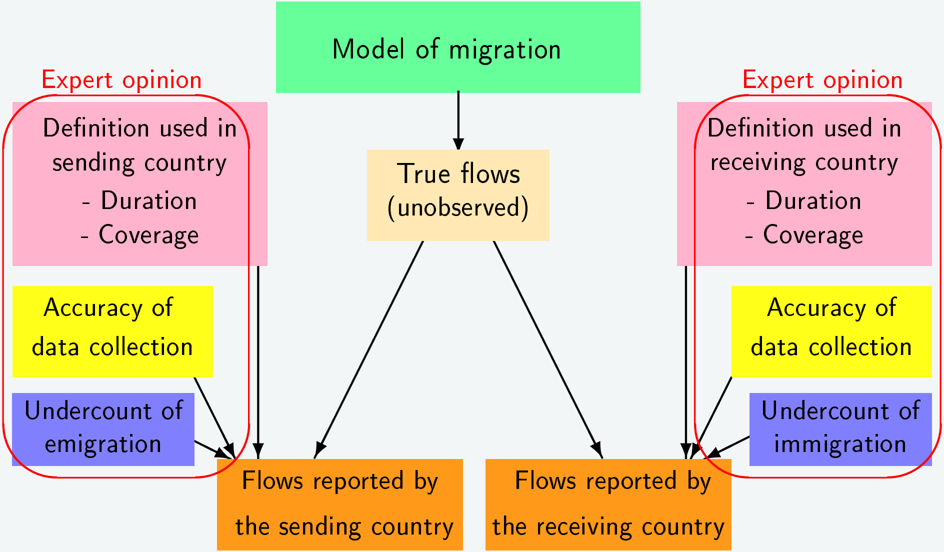
Availability $\Rightarrow$ Migration model (including economic, demographic and geographic explanatory variables)

$$\left.\begin{array}{l}\text{Duration of stay criterion}\\ \text{Coverage of subpopulations}\\ \text{Undercount of emigrants and}\\ \quad\text{immigrants}\\ \text{Accuracy of data collecting systems}\end{array}\right\} \Rightarrow \text{Measurement error model}$$

Modelling framework . . .



Modelling framework . . .



Modelling framework - model equations

S - sending, R - receiving country, z - observed flows, y - true flows

Data equations:

$$\begin{aligned} z_{ijt}^{S} &\sim \mathrm{Po}(\mu_{ijt}^{S}) \\ z_{ijt}^{R} &\sim \mathrm{Po}(\mu_{ijt}^{R}) \end{aligned}$$

Modelling framework - model equations

S - sending, R - receiving country, z - observed flows, y - true flows

Data equations:

$$z_{ijt}^{S} \sim \mathrm{Po}(\mu_{ijt}^{S})$$
$$z_{ijt}^{R} \sim \mathrm{Po}(\mu_{ijt}^{R})$$

Measurement error model:

$$\log \mu_{ijt}^{S} = \log y_{ijt} + \delta_{m(i)} + \kappa_i + \lambda_{f(i)} + \varepsilon_{ijt}^{S} \qquad \varepsilon_{ijt}^{S} \sim N(0, \tau_{c(i)}^{S})$$
$$\log \mu_{ijt}^{R} = \log y_{ijt} + \delta_{m(j)} + \kappa_j + \lambda_{g(j)} + \varepsilon_{ijt}^{R} \qquad \varepsilon_{ijt}^{R} \sim N(0, \tau_{c(j)}^{R})$$

$$mean\ obs\ flow = true\ flow \times duration \times coverage \times undercount \times error$$

Modelling framework - model equations

S - sending, R - receiving country, z - observed flows, y - true flows

Data equations:

$$z_{ijt}^S \sim \mathrm{Po}(\mu_{ijt}^S)$$
$$z_{ijt}^R \sim \mathrm{Po}(\mu_{ijt}^R)$$

Measurement error model:

$$\log \mu_{ijt}^S = \log y_{ijt} + \delta_{m(i)} + \kappa_i + \lambda_{f(i)} + \varepsilon_{ijt}^S \qquad \varepsilon_{ijt}^S \sim N(0, \tau_{c(i)}^S)$$
$$\log \mu_{ijt}^R = \log y_{ijt} + \delta_{m(j)} + \kappa_j + \lambda_{g(j)} + \varepsilon_{ijt}^R \qquad \varepsilon_{ijt}^R \sim N(0, \tau_{c(j)}^R)$$

$$mean\ obs\ flow = true\ flow \times duration \times coverage \times undercount \times error$$

Migration model:

$$\begin{aligned}\log y_{ijt} = {} & \alpha_1 + \alpha_2 \log P_{it} + \alpha_3 \log P_{jt} + \alpha_4 \log(G_{it}/G_{jt}) + \alpha_5 \log T_{ijt} \\ & + \alpha_6 C_{ij} + \alpha_7 A_{it} + \alpha_8 A_{jt} + \alpha_9 A_{ijt} + \alpha_{10} S_{ij} + \alpha_{11} S_{ji} \\ & + \alpha_{12} N_{ij} + \alpha_{13} M_{ijt} + \alpha_{14} F_{ijt} + \beta_t E_t + \nu_{ij} + \xi_{ijt}\end{aligned}$$

Results

- Results incorporate expert opinions

Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

- For low undercount of emigrants countries:

$$mean\ obs\ emigration\ flow = 77\% \times true\ flows$$

Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

- For low undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 77\% \times true\ flows$$
- For high undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 38\% \times true\ flows$$

Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

- For low undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 77\% \times true\ flows$$
- For high undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 38\% \times true\ flows$$
- For low undercount of immigrants countries:
$$mean\ obs\ immigration\ flow = 86\% \times true\ flows$$

Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

- For low undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 77\% \times true\ flows$$
- For high undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 38\% \times true\ flows$$
- For low undercount of immigrants countries:
$$mean\ obs\ immigration\ flow = 86\% \times true\ flows$$
- For high undercount of immigrants countries:
$$mean\ obs\ immigration\ flow = 70\% \times true\ flows$$

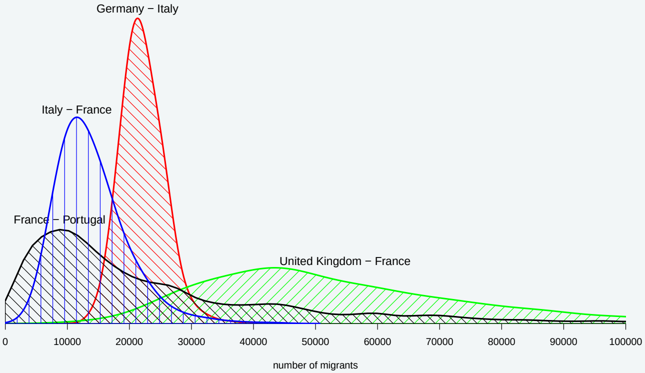
Results

- Results incorporate expert opinions
- Duration of stay factors: $mean\ obs\ flow = true\ flow \times duration$

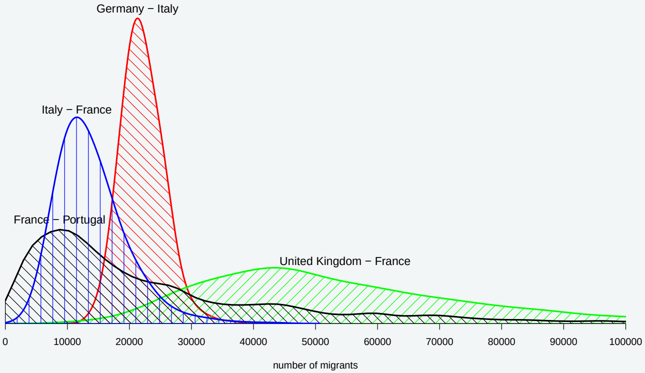
no time limit	3 months	6 months	12 months	permanent
188%	159%	137%	100%	44%

- For low undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 77\% \times true\ flows$$
- For high undercount of emigrants countries:
$$mean\ obs\ emigration\ flow = 38\% \times true\ flows$$
- For low undercount of immigrants countries:
$$mean\ obs\ immigration\ flow = 86\% \times true\ flows$$
- For high undercount of immigrants countries:
$$mean\ obs\ immigration\ flow = 70\% \times true\ flows$$
- Accuracy: immigration data $>$ emigration data
Scandinavian systems most accurate, survey-based least accurate

Posterior densities of true migration flows



Posterior densities of true migration flows



flow	mean	std.dev	rel.std.dev	median
FR-PT	29,400	43,100	147%	17,000
IT-FR	14,000	5,600	40%	12,800
DE-IT	22,100	3,300	15%	21,800
UK-FR	55,400	24,900	45%	49,600

Conclusions

- Have illustrated how Bayesian methods for modelling population processes can:
  - combine models for different components, such as mortality and fertility
  - combine a measurement model and a structural model
  - include expert opinion
  - provide coherent measures of uncertainty

Conclusions

- Have illustrated how Bayesian methods for modelling population processes can:
    - combine models for different components, such as mortality and fertility
    - combine a measurement model and a structural model
    - include expert opinion
    - provide coherent measures of uncertainty
- Bayesian methods can also incorporate model uncertainty

Conclusions

- Have illustrated how Bayesian methods for modelling population processes can:
  - combine models for different components, such as mortality and fertility
  - combine a measurement model and a structural model
  - include expert opinion
  - provide coherent measures of uncertainty
- Bayesian methods can also incorporate model uncertainty
- Further work
  - currently including immigration and emigration in the cohort component projections
  - Bayesian model averaging planned for the cohort component projections
  - have extended the European migration model to estimate flows by age and sex