

Grid Enabled Data Fusion for Calculating Poverty Measures

Social Science Applications
UK e-Science ALL HANDS MEETING 2006



Background

Social Science Problem *And Policy Issue*

- Researchers frequently have to use more than one data set in order to obtain a more complete answer to their questions

What do we know about ethnic minority economic welfare when it is disaggregated by group and geography

- One data set may provide a large sample of the target population, but offer incomplete coverage of the topics of interest

Census data can lack direct measures of income

- Another data set with coverage of the topics of interest may not sample the target population adequately

Survey data yield minority samples that may be too small for meaningful results to be obtained

- The statistical analysis involved belongs in the group of data linkage methods.
- As there are no identifiable common units in the problem considered here, it is one of statistical data fusion.
- The actual methodology used is taken from the poverty mapping literature.
- The **survey data** can be viewed as the **donor data set**, with the *census based data* being the *recipient data set*.

Data

- The British Household Panel Survey (BHPS) provides the small scale survey data.
 - BHPS is a longitudinal (panel) study with yearly waves.
- The Sample of Anonymised Records (SARs) provides the large scale Census data.
 - SARs are a random sample of individuals and households from the UK Census
- Uses 1991 data because of projected confidentiality restrictions on the publicly available version of the 2001 SARs.
 - 2% sample of individuals, 1% sample of households.

Data Issues

- It is time consuming dealing with different data sets when they are available in a wide variety of user unfriendly formats.
- Need common and coherent variable definitions for the donor and recipient data sets.

Addressed by the Data Grid?

- The issues suggest a work practice which becomes messy when dealing with communication between the steps in the overall analysis: data extraction, computation, and results presentation.
- This is alleviated by hosting the data on a data grid.

Empirical Analysis

1. estimate a statistical model using the BHPS data
 - taking account the heterogeneous nature of the survey data
2. use the results to provide income predictions for the SARs data
 - uses parameter estimates from 1.
3. use the income predictions along with other results to estimate poverty measures and their standard errors.
 - headcount (% below a given poverty line)
 - poverty gap (% distance from a poverty line)
4. present these poverty measures
 - by UK regions, SARs areas, GB profiled areas
 - for ethnic groups (by gender if using individuals)

Empirical Issues

- Statistical inference may be difficult in combined data problems. Underlying theoretical assumptions may be too strong. Calculation may be difficult.
- Simulation techniques may address these difficulties, however, these can be computational intensive.
- Here we use a so-called *casewise bootstrap*: resample from the BHPS data, repeating steps 1 & 2 and SARs sub-sample poverty measure calculation B times.

Addressed by the Computational Grid?

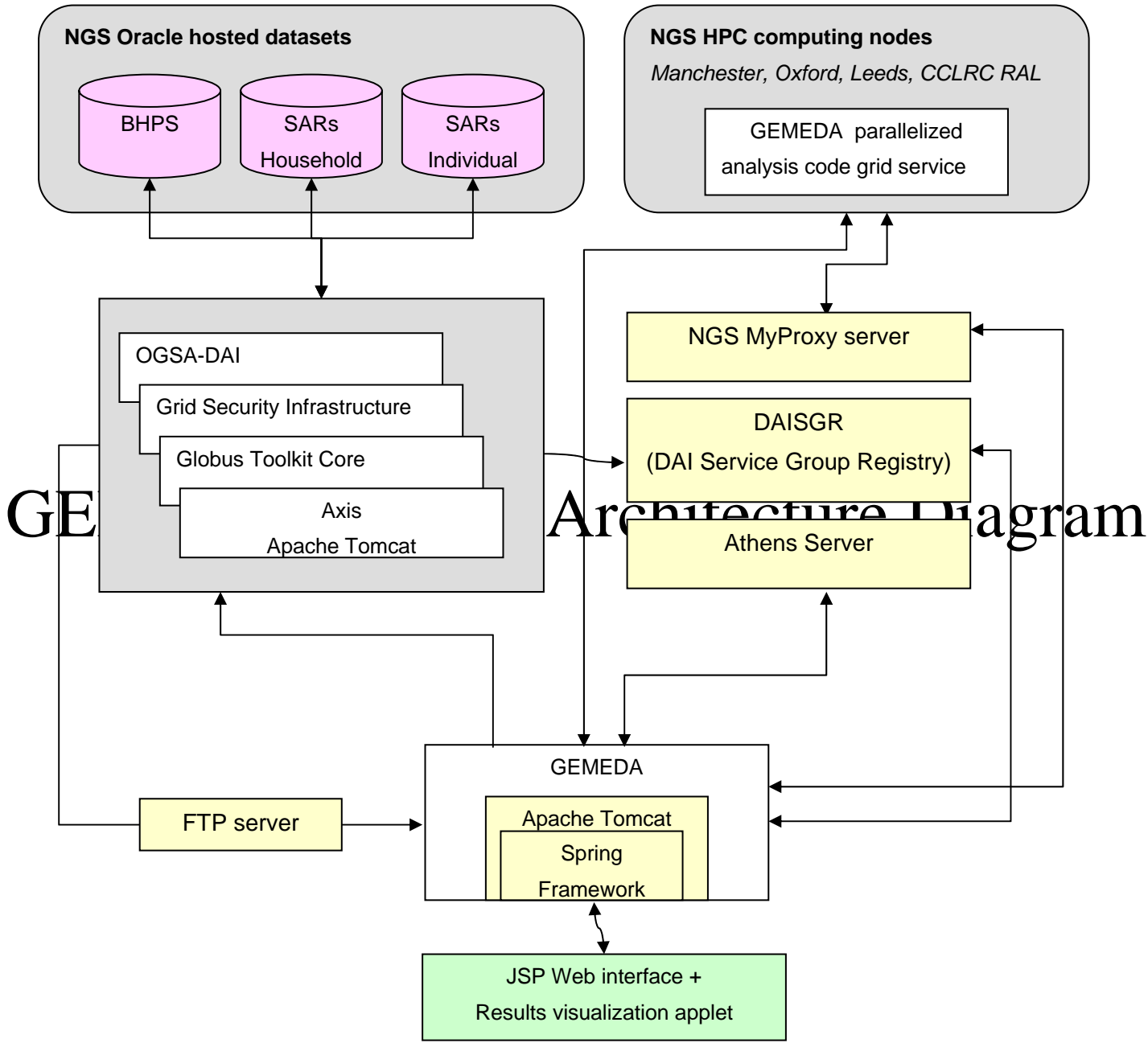
- Yes. These techniques are embarrassingly parallel and well suited to implementation on High Performance Computers (HPC).
- *Scatter* the data over P processors, do B/P bootstrap replications per processor, *gather* the results.

Implementation Issues

1. Social scientists rarely have easy access to HPC or HTC resources: **staff, human capital, equipment**
2. Need to avoid deploying complex middleware stacks on end-user's computers.

Suggested the following solution

- Tap into established trends & ongoing investments in UK e-Science
 - National Grid Service (NGS), solves 1.
 - Web service technologies (portals), solves 2.



Male EHC, %	Ethnic Group									
	Region	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese	Other Asian
North	21.0	.	.	.	17.9	40.0	42.8	22.9		
Yorkshire & Humber	17.6	26.6	40.4	20.4	24.6	32.0	36.2	31.7		
East Midlands	17.0	24.2	17.9	26.1	21.2	41.4	26.3	25.0		
East Anglia	17.0	19.7	.	15.0	18.0	33.2	.	15.2		
Inner London	17.5	25.1	34.8	34.3	20.1	30.8	34.5	33.7		
Outer London	13.3	17.4	28.7	25.5	16.1	23.8	25.4	19.4		
Rest of S.E.	12.9	17.9	24.7	18.1	15.0	21.9	21.9	20.1		
South West	17.9	26.9	30.4	27.5	20.8	23.4	.	.		
West Midlands	17.4	29.8	38.4	33.7	23.2	33.2	34.9	.		
North West	17.8	27.8	37.4	37.8	26.7	29.6	31.2	.		
Wales	20.6	.	53.9	.	19.5	29.6	15.5	.		

Male EHC, %	Ethnic Group								
	Area	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese
Bolton	18.2	.	.	.	30.8	36.7	.	.	.
Bury	14.1	20.1	.	.	.
Manchester	25.3	33.0	39.7	.	38.7	27.5	.	34.9	.
Oldham	16.5	29.1	41.4	.	.

Female EHC, %	Ethnic Group								
	Area	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese
Bolton	36.2	.	.	.	52.4	49.4	.	.	.
Bury	32.6	47.2	.	.	.
Manchester	40.6	46.5	48.1	.	49.3	48.6	.	43.8	.
Oldham	34.3	50.1	64.2	.	.
Rochdale	33.4	52.0	.	.	.
Sheff Hallam	34.9	.	.	.	41.8
Sheff Hallam	34.9	54.7	.
Sheff Hallam	34.9	53.8	54.3	.	.
Sheff Hallam	34.9	51.2	.	.	.

Female EHC, %	Ethnic Group							
	Region	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi
North	37.7	.	.	.	31.0	53.4	43.0	.

UK Male	Ethnic Group								
	Profile	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese
Enclave	20.4	24.2	35.0	36.2	22.0	32.9	.	.	.
Poor	21.6	29.6	34.9	31.4	22.7	27.5	.	.	.
The Rest	15.1	19.6	28.0	22.2	17.6	26.3	.	.	.
NW Male	Ethnic Group								
Profile	White	Black Caribbean	Black African	Black Other	Indian	Pakistan			
Enclave	29.2	27.4	.	.	33.1	32.2	.	.	
Poor	25.0	38.0	46.4	46.2	32.8	29.3	.	.	
The Rest	15.3	20.1	23.8	33.4	22.1	27.4	.	.	

UK Female	Ethnic Group							
	Profile	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi
Enclave	30.0	33.0	41.8	39.4	41.8	54.7	50.7	44.7
Poor	39.7	40.5	43.3	40.2	40.5	48.7	52.9	40.9
The Rest	34.1	34.3	38.4	33.3	33.9	46.7	38.5	37.0
NW Female	Ethnic Group							
Profile	White	Black Caribbean	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese
Enclave	45.8	44.4	.	.	54.4	54.5	62.4	.
Poor	42.7	51.4	51.4	54.5	51.0	50.6	.	52.8
The Rest	34.0	36.2	33.3	41.9	38.0	49.6	34.6	41.6

Visualization/ Results Presentation

- GIS style choropleth map,
 - *area colouring represents range of poverty measure*
- with linked plot
 - *boxplot style graphic of income for main category of interest for a chosen area.*
- Requires mapping data
 - *from EDINA, Athens authenticated*
- Implemented as a Java applet
 - *uses opensource GeoTools java library 2.*

GEMEDA map - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://pascal.mvc.mcc.ac.uk:8443/gemeda/gemedaMap/map.jsp?id=1138296800988

Google Met Office

[Met2_pov = 0]
 [Met2_pov > 0.0] AND [Met2_pov <= 10.0]
 [Met2_pov > 10.0] AND [Met2_pov <= 20.0]
 [Met2_pov > 20.0] AND [Met2_pov <= 25.0]
 [Met2_pov > 25.0] AND [Met2_pov <= 33.3]
 [Met2_pov > 33.3] AND [Met2_pov <= 50.0]
 [Met2_pov > 50.0] AND [Met2_pov <= 66.6]
 [Met2_pov > 66.6] AND [Met2_pov <= 100.0]

Region/SARs Area

White
 Black Caribbean
 Black African
 Black other
 Indian
 Pakistani
 Bangladeshi
 Chinese
 Other-asian
 Other-other

Male
 Female
 All

gender buttons

area toggle

ethnic group buttons

UK Male Imputed Income

Ethnic Group	Monthly Income (Approximate)
White	500 - 2500
Black Caribbean	500 - 1500
Black African	500 - 1500
Black other	500 - 1500
Indian	500 - 1500
Pakistani	500 - 1500
Bangladeshi	500 - 1500
Chinese	500 - 1500
Other-asian	500 - 1500
Other-other	500 - 1500

Monthly income

show data

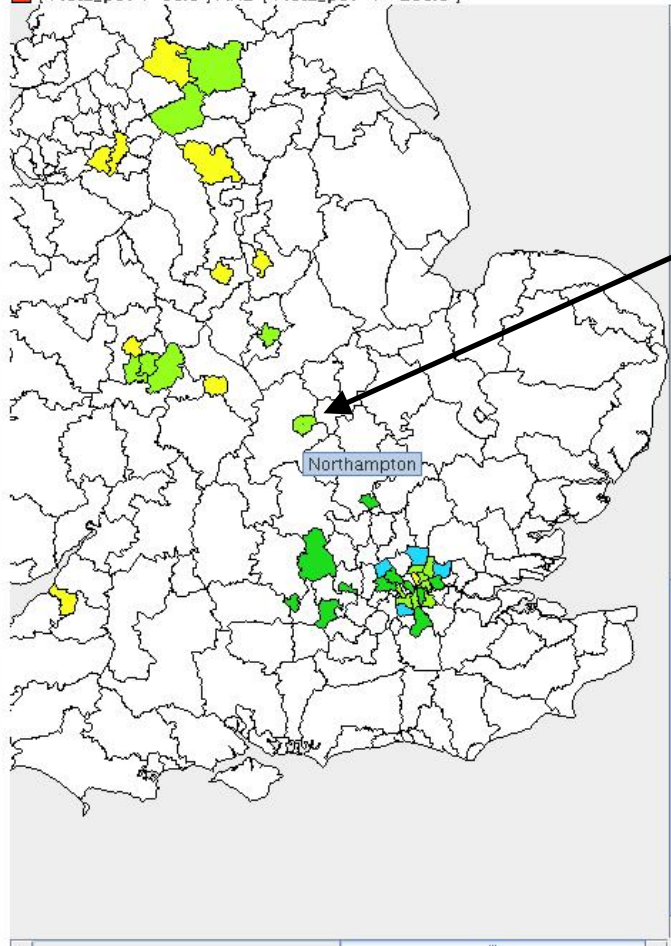
Applet MapViewer started

pascal.mvc.mcc.ac.uk:8443

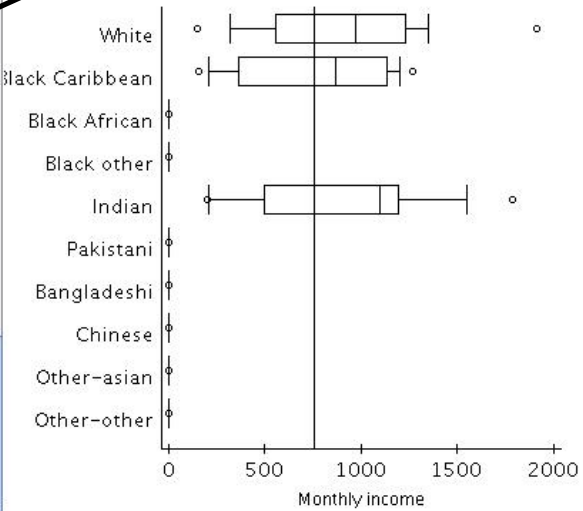
- [Met2_pov = 0]
- [Met2_pov > 0.0] AND [Met2_pov <= 10.0]
- [Met2_pov > 10.0] AND [Met2_pov <= 20.0]
- [Met2_pov > 20.0] AND [Met2_pov <= 25.0]
- [Met2_pov > 25.0] AND [Met2_pov <= 33.3]
- [Met2_pov > 33.3] AND [Met2_pov <= 50.0]
- [Met2_pov > 50.0] AND [Met2_pov <= 66.6]
- [Met2_pov > 66.6] AND [Met2_pov <= 100.0]

Region/SARs Area

- White
 - Black Caribbean
 - Black African
 - Black other
 - Indian
 - Pakistani
 - Bangladeshi
 - Chinese
 - Other-asian
 - Other-other
- Male
 - Female
 - All



Northampton Male Imputed Income



An Empirical Worker's Perspective

- **What has an e-Research solution provided?**
 - ✓ Equipment.
 - middle range HPCs (and software) are accessible to Social Scientists.
 - ✓ Staff.
 - maintenance, development and implementation of middleware.
 - ✓ Human capital.
 - knowledge and experience gained by researchers stays relevant and remains within the Social Science.
 - ✓ No output clutter.
 - Visualization allows output filtering.

- **What are the weaknesses of the present implementation?**

- **Sustainability.**

- staff, maintenance, ...

- re-engineering (Globus WSRF, OGSA-DAI & SQL Server, GridSphere, GROWL ...)

- **Reliability.**

- error reporting for distributed architecture, documentation

- robustness of compute nodes, interoperability of middleware

- **Security.**

- Athens authentication, e-certificates, ...

- firewalls, ...

- **Modeling.**

- more variables, more data sets, repeat data sets (2001)

- interactivity, visualization, ...

- **Implementation Limitations**

- **Having grid-enabled data does not mean you have the process of data cleaning or data manipulation grid-enabled.**

- **Computational grids require compute resource brokerage (NGS is batch).**

- **Supporting code much larger than business code, specialist elements of the project were easiest to develop.**

- **Other Issues Related to User Engagement**

- **Data disclosure controls.**

- **Confidentiality restrictions.**

- **Proprietorship.**

- **Expertise of researchers within discipline areas.**

- **Technical rather than substantive research agenda**

Project Team

Simon Peters, Ken Clark SoSS (economics)

Pascal Ekin, Anja Le Blanc, Stephen Pickles Manchester
Computing

Project portal and service: <http://pascal.mvc.mcc.ac.uk/gemeda/>

Acknowledgements

Celia Russell, Mike Jones

SAMD

Mark Birkin, Andy Turner

Hydra I Grid (now MoSeS)

Keith Cole

ConvertGrid

Matt Ford

NGS



RES-149-25-0009

