

Grid Enabled Data Fusion for Calculating Poverty Measures.

Simon Peters¹, Pascal Ekin², Anja LeBlanc², Ken Clark¹ and Stephen Pickles²

¹School of Social Sciences & ²Manchester Computing, University of Manchester

2006

Abstract

This article presents and discusses the motivation, methodology and implementation of an e-Social Science pilot demonstrator project entitled: Grid Enabled Micro-econometric Data Analysis (GEMEDA). This used the National Grid Service (NGS) to investigate a policy relevant Social Science issue: the welfare of ethnic minority groups in the United Kingdom. The underlying problem is that of a statistical analysis that uses quantitative data from more than one source. The application of grid technology to this problem allows one to integrate elements of the required empirical modelling process: data extraction, data transfer, statistical computation and results presentation, in a manner that is transparent to a casual user.

1. Introduction

Economic investigation of the experiences and prospects of Britain's ethnic groups paints a picture of multiple deprivation and disadvantage in areas such as earnings and employment. One consequence of this is that the welfare of such groups is of major policy concern. However, in order to evaluate minority welfare, a requirement for successful policy intervention, one needs to be able to produce appropriate statistical quantities, such as poverty measures.

The full modelling process required for this goes beyond an analysis using a single data set, and has components that suggest an e-Research approach is appropriate. The motivation for two of these components, data and statistical modelling, is discussed in section 2, with details of the methodology reported in section 3. Section 3 also deals with Grid adoption issues, and the perspective taken is from the economics discipline, where dedicated e-Science resources, such as equipment and specialised staff, are not readily available. This underlines one of the objectives of the demonstrator. We not only want to show that modern micro-econometric research can be implemented on the Grid, but that it can be done in a manner that builds on existing infrastructure investments (such as the National Grid Service), and develops them.

Another component of our modelling process, results visualization, is briefly discussed in section 4 along with other details of the Grid implementation. Section 5 presents a summary of the substantive analysis, and section 6 concludes.

2. Motivation

Social science researchers in the UK have access to a wide range of survey data sets. These are collected by numerous different agencies, including the government, with different purposes in mind and exhibit considerable variation along a number of dimensions. Only on rare occasions are the needs of social scientists uppermost in the minds of survey designers — hence the topics covered, the sample frames and sizes, the questionnaire formats, data collection methodologies and specific questions asked are extremely diverse. In order to obtain a more complete answer to their research questions, researchers frequently have to use more than one data set. The type of analysis possible, however, is often constrained by the fact that each data set possesses one desirable attribute while being deficient elsewhere.

2.1. A Grid Solution for the Data Problem?

The economic welfare of ethnic minority groups in the UK, raises data issues that require such a multiple data set approach. The basic problem is that non-whites account for a small proportion of the population and sample surveys typically yield minority samples that are too small for meaningful results to be obtained. To some extent the situation is improved when Census data are available. However, while these provide relatively large samples of minority individuals and households, they do not contain any direct measures of income. Other surveys do contain such information but have limited sample sizes when minorities are analysed, with

the problem being especially acute when reporting is required for small area geographies.

A major consequence of such data problems is that important questions about the welfare of minority groups have not been answered. For example, small sample sizes preclude useful measures of household welfare such as poverty rates or inequality measures at anything other than high levels of aggregation. Yet research suggests that disaggregation along two dimensions is crucially important when discussing the welfare of Britain's ethnic minority groups. First, it is clear that treating non-white, minority groups as a homogenous entity is not valid. There is considerable diversity between groups such as Caribbeans, Indians, Pakistanis, Bangladeshis and the Chinese (Leslie, 1998; Modood et al., 1997). This diversity is often quantitatively greater than the differences between non-whites, taken as a whole, and the majority white community. The second dimension where aggregation is important is geographical. Britain's ethnic minorities tend to live in co-ethnic clusters, or enclaves, and this clustering has important consequences for economic activity and unemployment (Clark and Drinkwater, 2002).

The Social Science micro-data sets that could be used to address the above problem are not large from an e-Science perspective. They do, however, tend to be messy and difficult to work with. This problem is compounded for repeated samples (variable definitions change), longitudinal samples (records require information from previous waves), and for the data combination approach considered in this article. Grid technology has the potential to integrate the tasks (data extraction, file transfer) associated with processing quantitative data of this type, by hosting the information in an appropriate manner on a data grid.

2.2. A Grid Solution for the Modelling Problem?

The empirical analysis, a micro-econometric one that relies upon the combination of two or more data sources, belongs to the broader group of modelling techniques associated with linking data sets. Following the terminology of Chesher and Nesheim (2006), who provide a review of this area, as the data linkage is performed with no or unidentifiable common records between the data sets, it falls into the class known as statistical data fusion.

The essence of the approach is to estimate a statistical model on one data set, the so-called donor sample (a relatively small scale but detailed survey), and then apply elements of the

fitted model (predicted responses for data imputation, residuals for simulation) to another data set, the so-called recipient sample (possibly larger-scale but less detailed), taking due account of the statistical issues surrounding both model assumptions and data matching as required, such as the potentially heterogeneous nature of the survey data.

Further, the underlying assumptions associated with standard statistical inference may well be violated in a combined analysis. As a consequence, it may be preferable to calculate statistical items such as poverty measure standard errors using re-sampling methods (variants of statistical bootstrapping). As noted by Doornik *et al.* (2004), such analyses have a component that allows for so-called embarrassingly parallelisable computations, and as such are well suited to implementation on the high performance computing (HPC) resources available on the NGS.

3. Methods

The initial plan was to extend an existing macro-econometric application, the SAMD project of Russell *et al.* (2003), to deal with multiple data sources and a different statistical analysis.

3.1. Grid Adoption Issues.

A review of current technologies early in the project resulted in the decision not to re-use software from the SAMD project. Although the broad design principles still applied, technology had moved on significantly since the earlier project was conceived. The intervening years have seen several trends become well established. In particular, Web Service technologies have become widely accepted in the e-Science community, the UK e-Science programme has made a significant investment in OGSA-DAI, and there has been a steady shift towards Web portals to avoid difficulties in deploying complex middleware stacks on end-users' computers. Fortunately, sacrificing re-use of SAMD's redundant technology was offset in part by several factors: the advent of the National Grid Service; the advent of the OGSA-DAI software; and the completion of other projects, which did provide components that we were able to re-use, such as the Athens authorisation software developed for ConvertGrid (Cole *et al.*, 2006). After due consideration, the availability and increased maturity of the NGS suggested efforts should be concentrated on their systems, namely Oracle for the data bases and MPI for parallelization.

3.2. The Data Sources

The project has grid enabled two data sources, the British Household Panel Survey (the BHPS) and the 1991 Census Samples of Anonymised Records (the SARs). One can now combine data from the smaller-scale, detailed, BHPS source with the larger sample sizes and geographical coverage of the Census SARs. This lets us provide poverty measures for ethnic minorities which are both broken down by particular ethnic group and geographically disaggregated.

The decision to use the 1991 data needs some comment. This decision was taken when the project was first proposed. The data needs to be readily available to accredited researchers to allow deployment on a data Grid, and it was felt at the time that the confidentiality restrictions envisaged for the so-called Licensed 2001 SARs (the public domain version of the 2001 SARs) would not contain variables suitable for the analysis required. The original release of the Licensed data was not scheduled to contain detailed information on ethnic minorities or on geographical details below regional level. There were also further restrictions (such as the grouping of age) on a variable's response categories. These restrictions are lifted for the Controlled Access (CAMS) version of the 2001 SARs. However, the access and confidentiality constraints imposed on the 2001 CAMS make them presently unusable from a data Grid perspective.

3.3. The Statistical Methodology

This follows an approach in the poverty mapping literature due to Elbers *et al.* (2003). The version of their methodology that is employed is presented here.

3.3.1. Calculations Using the Survey Data, the Donor Sample.

The survey data is used to estimate a model of the economic variable of interest, y_{ic} , which is income in this study. The index i refers to an individual in a sample cluster, which is indexed by c . Each cluster contains n_c observations and there are $N = \sum_{c=1}^C n_c$ observations over the C clusters. The economic variable of interest is specified as $\log(y_{ic}) = \beta' x_{ic} + u_{ic}$ where x_{ic} is a vector of suitably defined explanatory variables, individual idiosyncratic error terms: $u_{ic} = \eta_c + \varepsilon_{ic}$ where η_c and ε_{ic} are uncorrelated with x_{ic} , independent of each other and $\text{IID}(0, \sigma_\eta^2)$ and $\text{IID}(0, \sigma_{ic}^2)$ respectively.¹

First step estimation uses ordinary least squares (OLS) to obtain the coefficient estimates $\hat{\beta}$. The second step uses the fitted residuals, $\hat{u}_{ic} = \log(y_{ic}) - \hat{\beta}' x_{ic}$, to estimate a model of the variance components. Set $\hat{u}_{ic} = \hat{u}_{.c} + \hat{e}_{ic}$ where $\hat{e}_{ic} = \hat{u}_{ic} - \hat{u}_{.c}$ and proceed to model the idiosyncratic heteroscedastic component σ_{ic}^2 using a logistic style transformed equation

$$\frac{\hat{e}_{ic}^2}{(A - \hat{e}_{ic}^2)} = \alpha' z_{ic} + r_{ic} \quad \text{where } z_{ic} \text{ is a vector of}$$

appropriate explanatory variables, A is set to $1.05 \max(\hat{e}_{ic}^2)$ and r_{ic} is a suitable error term. Estimation of the α coefficients is done using OLS.

Once \hat{u} has been obtained the appropriate prediction for σ_{ic}^2 can be calculated. The remaining variance component, σ_η^2 can be calculated as $\hat{\sigma}_\eta^2 = \max(\hat{V}(\eta_c), 0)$ where

$$\hat{V}(\eta_c) = \frac{1}{C-1} \sum_c (\hat{u}_{.ic} - \hat{u}_{..})^2 - \frac{1}{C} \sum_c V(\hat{e}_{.c})$$

$$\text{and } V(\hat{e}_{.c}) = \frac{1}{(n_c-1)n_c} \sum_{i=1}^{n_c} \hat{e}_{.c}^2.$$

3.3.2. Calculations Using the Census Data, the Recipient Sample.

Once the above estimates have been obtained one can impute the economic variable of interest for any given set of comparable explanatory variables x_{kc} and z_{kc} . The index k indicates an individual in the Census data source. The Census based predictions can then be calculated, under an assumption of

Normality, as: $\hat{y}_{kc} = \exp(\hat{\beta}' x_{kc} + \frac{\hat{\sigma}_\eta^2 + \hat{\sigma}_{kc}^2}{2})$. The variance prediction $\hat{\sigma}_{kc}^2$ is calculated using $\hat{\alpha}' z_{kc}$.

One can also calculate a wide variety of poverty measures. The demonstrator uses the parametric (expected) head count (PHC), simulated head count (SHC), and simulated poverty gap (SPG). The PHC measure requires an assumption of Normality and is calculated as

$$\text{PHC}(p) = \frac{1}{n} \sum_{k=1}^n \Phi((\log(p) - \hat{\beta}' x_{kc}) / \sqrt{\hat{\sigma}_\eta^2 + \hat{\sigma}_{kc}^2})$$

where $\Phi(\cdot)$ is the standard Normal distribution function and p is a so-called poverty line. Summation is taken over the sub-sample of interest, ethnic group within geographic area.

If the parametric assumption of Normality was incorrect, this would cause misspecification problems that might affect the estimated measures and their associated standard errors. Simulation can be used to counter this possibility and is used for SHC and SPG. The SHC measure is obtained as the average of B simulated head count measures:

$$\text{SHC}(p) = \frac{1}{B} \sum_{b=1}^B \text{SHC}(p)_b \quad \text{where}$$

$$\text{SHC}(p)_b = \frac{1}{n} \sum_{k=1}^n I((\hat{\beta}_b' \mathbf{x}_{kc} + \tilde{u}_{k,b} + \tilde{e}_{k,b} \hat{\sigma}_{kc,b}) < \log(p))$$

Note that $I(\cdot)$ is an indicator function taking the value of 1 if the condition inside the parentheses is satisfied and zero otherwise. The standard error predictor, $\hat{\sigma}_{kc,b}$, is calculated using

$\hat{\alpha}_b' \mathbf{z}_{kc}$. The coefficients, $\hat{\beta}_b$ and $\hat{\alpha}_b$, are obtained from the b^{th} casewise resample of the survey data, error terms, $\tilde{u}_{k,b}$ & $\tilde{e}_{k,b}$, are drawn with replacement from the error vectors \tilde{u}_b & \tilde{e}_b .³

The SPG measure is obtained in a similar manner:

$$\text{SPG}(p) = \frac{1}{B} \sum_{b=1}^B \text{SPG}(p)_b$$

$$\text{where } \text{SPG}(p)_b = \frac{1}{n} \sum_{k=1}^n I(\hat{y}_{kc,b} < p) * \left(1 - \frac{\hat{y}_{kc,b}}{p}\right).$$

$$\text{Here } \hat{y}_{kc,b} = \exp(\hat{\beta}_b' \mathbf{x}_{kc} + \tilde{u}_{k,b} + \tilde{e}_{k,b} \hat{\sigma}_{kc,b}).$$

Standard errors are calculated in the usual fashion using the B simulated values of the chosen poverty measure: $\text{SPG}(p)_b$ or $\text{SHC}(p)_b$. A simulated version of the $\text{PHC}(p)$, $\text{PHC}(p)_b$ is used to calculate its standard error. This is based upon the $\text{PHC}(p)$ equation above, but with the $\hat{\sigma}_\eta^2, \hat{\sigma}_{jc}^2, \hat{\beta}$ and $\hat{\alpha}$ replaced by the values obtained from the b^{th} simulation. Elbers *et al.* (2003) suggest B can be set to 300.

4. The Grid Implementation

The demonstrator is designed for a researcher who wishes to investigate the welfare of ethnic minority groups in the UK. Specifically it allows the researcher to choose different ethnic groups, to specify a level of geography, and to pick from a limited set of poverty measures. The aim is to provide an easy-to-use (Web-based) interface which allows the user to make choices about the type of analysis to be performed and which then returns the results of

that analysis to the user. The details of the actual analysis and associated data management are invisible to the user.

The demonstrator service presently produces poverty measures using individual level data. Demonstrator options are the standard headcount measure, and the poverty gap measure. These can be calculated for two possible poverty lines, either 60% of the UK median income or 50% of the UK mean income. The poverty measures are then displayed on a GIS (Geographic Information System) style choropleth map display for UK regional and SARs area geographies. The display presents the poverty measure for the chosen ethnic minority group. The use of individual income data also allows calculation of poverty measures by gender, and, if this option has been chosen for the analysis, the resulting poverty measures can be displayed by ethnic group within gender. The display also produces a box-whisker style plot of the predicted income quantiles⁴ for all the ethnic groups associated with a region. This is available for the whole of the UK and at the level of geography displayed by the map (UK region or SARs area).

Other information, such as standard errors, and the results of the model fitted using the survey data, are available from the supporting files returned by the demonstrator. It is also possible to request classification based upon the pseudo-geography available for the 1991 SARs, although these results are not accessible via the visualization tool. The visualisation applet will not display information at a geographic level if the sample size for an ethnic minority group is deemed small. This mimics the type of confidentiality restrictions applied to the equivalent controlled access 2001 data.

4.1.1. The Web Service Client

The GEMEDA service stands at the heart of the architecture illustrated in Figure 1. It provides the services and generates the HTML sent to the light client (Web browser) through the handling of events (control). The Spring framework container enforces the MVC (Model View Control) design pattern, enforcing the separation of logic from content and greatly encouraging code re-use.

When a user first accesses GEMEDA he/she creates a user account. An Athens username/password combination is required which is verified on the fly by calling an XML-RPC Athens security interface.

The service downloads a proxy credential automatically through a MyProxy server during each step of the workflow. The proxy credential

lasts 24 hours. The longevity of the Proxy credential is verifiable at all times by clicking on the appropriate link. Connection to the front-end (Web client) is through an HTTPS connection.

The service generates separate OGSA-DAI SQL queries targeted at the SARs and BHPS datasets as a result of user input, and uploads necessary executables to the user's selected HPC computation node if they are not already present.

4.1.2. Security

A user needs login permissions for the GEMEDA service, an e-science certificate to access the NGS, and an Athens username to allow use of the SARs and BHPS data. The user is asked for his/her pass-phrase which is sent using HTTPS to the service. This automatically initiates the creation of a proxy certificate by calling a designated MyProxy server containing the user's certificate. Note that the user's certificate needs to have been uploaded to this server by an appropriate tool. The proxy credential is stored and used by the GEMEDA service throughout the lifetime of the session. A single sign-on mechanism allows the web service to query data through OGSA-DAI and to communicate with the HPC by means of the GSI (Grid Security Infrastructure). This provides message level encryption as well as authenticating and authorising the owner of the proxy credentials.

4.1.3. Grid-enabled Datasets

The SARs and BHPS data sets are stored in separate Oracle databases which occupy slightly over 1 gigabyte of storage space. Data access is done entirely through the OGSA-DAI grid middleware. The Oracle server hosting the SARs and BHPS datasets is a NGS resource administered by the University of Manchester. It should be noted that all the available waves of the BHPS were grid-enabled, along with both the individual and household SARs for 1991. Only the 1991 BHPS wave and individual SARs file are used in the present version of the demonstrator.

Using OGSA-DAI version 4 has proved to be unreliable when accessed securely with GSI. Hence, a local instance of OGSA-DAI accessing the NGS hosted datasets was installed on the Linux server hosting the GEMEDA service. OGSA-DAI query results (XML asynchronous data streams) are uploaded to an FTP server before being converted to a data format recognized by the GEMEDA logic.

Converted data sources are then uploaded to the HPC node using secure GridFTP.

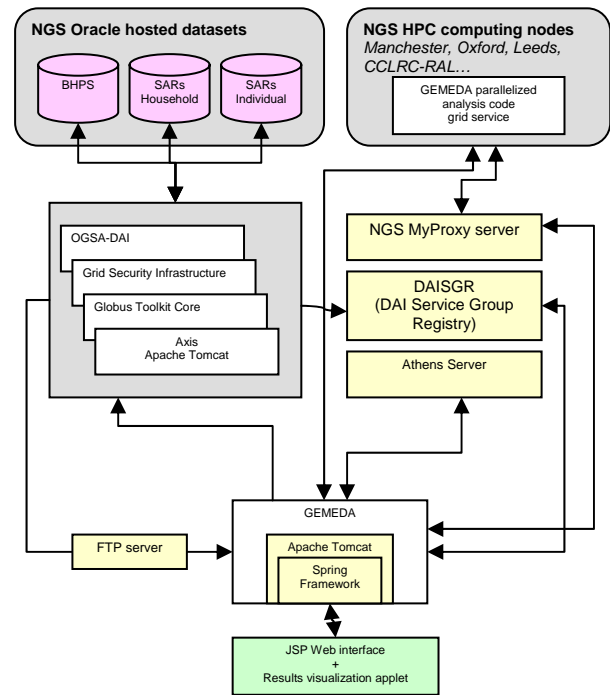


Figure 1: The GEMEDA Architecture

4.1.4. The Analysis Code

The GEMEDA business logic, i.e. the statistical/econometric analysis component, consists of four steps: command/data input, estimation & prediction, welfare measure calculation, and output of the results. These were developed in MPI-parallelized Fortran 95 and produce: model parameter and standard error estimates for the Survey data (the donor sample), imputed income quantiles, poverty measures and associated standard error estimates for the Census data (the recipient sample).

The GEMEDA service accesses the logic modules as a GT 2.4 service through the Globus CogKit running in a separate Java instance (to circumvent namespace clash with the Globus toolkit called by the OGSA-DAI client toolkit). Seen from the level of the GEMEDA Logic, access to the command file, BHPS & SARs data consists in simply reading the appropriate files. ALL files are made accessible to the GEMEDA service code through the GridFTP server.

Event notification is carried by the encapsulating grid service which periodically interrogates the status of the grid service (this does not return a percentage of completion but a status: running, pending, halted, etc).

Once the job is completed, the results are written to file and downloaded through GridFTP by the web service. These results are then converted to XML and spatially mapped by the GEMEDA service before being sent back to the user interface for viewing.

4.1.5. Visualisation.

Once the results files have been returned to the GEMEDA service, they may be viewed in the raw or processed by the service's visualisation applet. A C utility converts the raw data into a form (dbf) appropriate for the applet. This is done for both the regional and SARs area geographies. The applet, which runs under Java 1.4 and above, uses this information along with special mapping data files (shp files) to produce choropleth maps at regional and SARs area geographies for the selected ethnic group and gender category. The shp files are obtained from <http://edina.ac.uk/ukborders/> and the applet combines these to produce the map, along with the legend, linked plot, and buttons for gender/ethnic group/geography selection. The applet uses the GeoTools java library 2 to aid the reading and display of the information provided. GeoTools is open source, and was also used by the Hydra I Grid project (Birkin *et al.*, 2005). The data and maps remain on the GEMEDA service's server, and permission to use the mapping information is allowed if a user has Athens authentication.

area of interest. The map has a zoom facility which is useful when the finer SARs area geographies are displayed. Figure 2 presents a screenshot of the SARs area map for Indian Male simulated head count poverty measures using the half mean income poverty line. The linked plot displays predicted income quantiles for all the groups available from the SARs area last pointed to, which was Manchester in this case.

5. A Summary of the Substantive Analysis.⁵

Using data from 1991, models of individual income were estimated using the BHPS data separately for males and females. The specification of the regression equation included all of the variables available via the demonstrator.⁶ The results supported splitting the sample by gender as the signs of the parameters on some of the regressors were different for males and females (e.g. the variables indicating marital status and the presence of children in the household). Full regression results are not presented here; instead we note that the explanatory power of both prediction equations appeared reasonable, 53% and 40% for males and females respectively, though the functional form tests suggest there may be room for improvement in the female equation specification. Heteroscedasticity tests strongly rejected the null of homoscedasticity. This heterogeneity in the variance was modelled using the methodology of Elbers *et al.* (2003) described in section 3.3 above

The breakdown of poverty measures by region and ethnic group for males shows considerable diversity across each of these dimensions. In general non-whites have higher poverty measures than Whites and this conforms to what we know about the higher unemployment rates and lower earnings of ethnic groups in the UK. Some groups, particularly Black Africans, Pakistanis and Bangladeshis, do particularly badly while the Indians and, to a lesser extent, the Chinese have poverty rates closer to those of Whites. This broad ranking is similar to that in Berthoud (1998). 'Southern' areas of the country generally have lower poverty headcounts than other regions although it should be noted that we do not correct for regional price differentials here. Some regions have relatively high poverty rates for particular groups, for example Bangladeshis and Pakistanis in the North, Pakistanis in the East Midlands and Black Africans in Yorkshire and Humberside, the

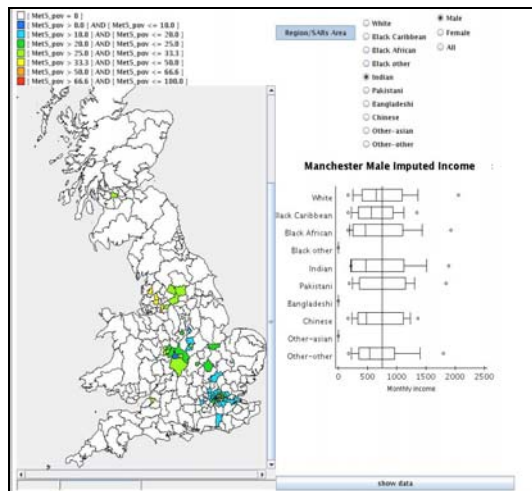


Figure 2: A Screenshot of the Visualization Applet.

The poverty measure displayed using the map's colour scheme is the one chosen by the user at job initiation. The applet allows the user to switch between different linked box-whisker style plots of predicted income quantiles by using the map cursor to point to the geographic

West Midlands, the North West and especially Wales.

The broad ranking of the groups and regions is similar for females but poverty rates are much higher. This is because these measures are based on individual income and females have lower participation rates in the labour market. Clearly this measure does not take account of intra-household income transfers. Pakistani and Bangladeshi females stand out as having extremely high poverty rates while, again, Indian and Chinese women are more comparable with their White counterparts.

An advantage of using SARs data is the ability to examine sub-regional geography. However at this level, small samples become a problem. While poverty rates can be estimated for Whites in all the areas, meaningful comparisons between different ethnic groups are only possible for urban areas in which ethnic minorities tend to cluster. There are a number of solutions to the problem of small samples. First, more detail is available using the 2001 Individual SARs which feature considerably more ethnic minority respondents than the 1991 data set. Alternatively, it is possible to capture something of the results for different 'types' of area. Tables 1 and 2 do this using the 'GB Profiles' area classifications attached to the 1991 SARs.⁷ All GB profiles that indicate categorisation based on one or more ethnic minorities are grouped together as *Enclaves*. Clark and Drinkwater (2002) suggest that enclaves are associated with worse outcomes for ethnic minorities. The remainder are split into *Poor* (based on housing tenure categorisation), and *The Rest*. The results do not indicate that there is a strong difference in ethnic minority wealth when comparing the SHCs in the *Enclave* profile grouping with the *Poor* profile grouping, however both do worse than *The Rest*.⁸

| UK Male. | Profile | | |
|------------------------|----------------|-------------|-----------------|
| <i>Ethnicity</i> | <i>Enclave</i> | <i>Poor</i> | <i>The Rest</i> |
| <i>White</i> | 23 (1.0) | 25 (0.6) | 17 (0.4) |
| <i>Black Caribbean</i> | 27 (1.5) | 33 (1.6) | 22 (1.1) |
| <i>Black African</i> | 39 (2.2) | 39 (2.5) | 31 (2.0) |
| <i>Indian</i> | 24 (1.3) | 25 (1.4) | 20 (0.9) |
| <i>Pakistani</i> | 36 (1.6) | 31 (1.7) | 29 (1.4) |
| <i>Bangladeshi</i> | 37 (2.9) | 38 (3.7) | 25 (2.1) |
| <i>Chinese</i> | 39 (2.7) | 30 (2.2) | 25 (1.3) |

Table 1: Male SHC Poverty Measures for Profiled Areas.⁹

| UK Female. | Profile | | |
|------------------------|----------------|-------------|-----------------|
| <i>Ethnicity</i> | <i>Enclave</i> | <i>Poor</i> | <i>The Rest</i> |
| <i>White</i> | 42 (1.4) | 54 (0.7) | 49 (0.6) |
| <i>Black Caribbean</i> | 39 (1.8) | 46 (2.0) | 39 (1.7) |
| <i>Black African</i> | 47 (2.3) | 51 (3.4) | 48 (2.5) |
| <i>Indian</i> | 57 (1.6) | 56 (2.3) | 51 (1.4) |
| <i>Pakistani</i> | 73 (1.8) | 72 (2.3) | 68 (1.9) |
| <i>Bangladeshi</i> | 71 (2.7) | 72 (3.9) | 73 (3.1) |
| <i>Chinese</i> | 48 (2.7) | 56 (3.3) | 48 (2.0) |

Table 2: Female SHC Poverty Measures for Profiled Areas.¹⁰

6. Concluding Comments

The NGS was used to aid the investigation of the welfare of ethnic minorities in the UK by grid enabling the required statistical analysis. This is a small scale problem compared to some science based applications, however, given the local level of resourcing available within the Social Sciences it would not have been possible to obtain the full benefits of a grid implementation without using the NGS.

Our project service arranges for the appropriate data sub-sets to be extracted in a coherent and consistent manner by running queries against OGSA-DAI enabled Oracle databases hosted on the NGS. It then transfers these sub-sets, along with the MPI parallelized code for the statistical analysis, and its associated command file, to a compute node on the NGS. The results of the statistical analysis are then returned to the service, and processed for presentation to the user via a GIS style visualization tool. Job initiation and results viewing are all performed on a web browser.

While the application to ethnic minorities addresses substantive research questions which cannot be addressed adequately using existing techniques, it should be noted that the methodology is general and its future development offers opportunities to social science researchers to address a wide variety of questions using a number of different, complementary data sets. Indeed, the use of OGSA-DAI now offers the potential to include data sets that are hosted as SQL Server databases. In the context of the present analysis, this could be further concurrent data sets, or later versions (2001 for instance) of the BHPS and Census SARs. Extensibility is not restricted to classical quantitative applications, however, and there exists the possibility of integrating appropriate qualitative information using the

techniques of Ahmad *et al* (2005), although this is outside the domain of the authors.

As this was a small project, evaluation of our objectives and usability issues have proceeded in a somewhat *ad hoc* manner and are still ongoing. One area of concern is the level of security required to access the service. The steps required for obtaining and processing an e-certificate are quite involved, and this, combined with problems of access caused by institutional firewalls, can be off-putting for potential users in the Social Sciences.

Notwithstanding this, and other issues such as compute resource brokerage, we regard the establishment of a critical mass of **compatible** Grid-enabled datasets and tools as a necessary condition for the success of e-Social Science in the UK, and hope that the GEMEDA service is at least a useful stepping stone towards this goal.

Acknowledgements

Research supported by ESRC grant number RES-149-25-0009, "Grid Enabled Microeconomic Data Analysis".

We benefited from discussion with members of the following projects: SAMD (Celia Russell, Mike Jones), ConvertGrid (Keith Cole), Hydra I Grid (Mark Birkin, Andrew Turner), and with locally based NGS staff (Matt Ford).

Additional support was provided in the form of an allocation of resources on the NGS itself.

References

Ahmad, K., L. Gillam, D. Cheng (2005), Society Grids, *Proceedings of the UK e-Science All Hands Meeting 2005*, EPSRC Sept. 2005

Berthoud, R. (1998), *The Incomes of Ethnic Minorities*, ISER Report 98-1, Colchester: University of Essex, Institute for Social and Economic Research.

Birkin, M., P. Dew, O. McFarland J. Hodrien. (2005), HYDRA: A Prototype Grid-enabled Decision Support System, *Proceedings of the First International Conference on e-Social Science*.

Chesher, A. and L. Nesheim (2006), Review of the Literature on the Statistical Properties of Linked Datasets, *DTI Economics Papers*, Occasional Paper No. 3.

Clark, K. and S. Drinkwater, (2002), Enclaves, neighbourhood effects and economic outcomes: Ethnic minorities in England and

Wales, *Journal of Population Economics*, **15**, 5-29.

Cole, K., L. Mason, P. Ekin, J. Maclaren (2006), ConvertGrid, *Proceedings of the Second International Conference on e-Social Science*.

Doornik, J., N. Shepherd, D. F. Hendry (2004), *Parallel Computation in Econometrics: A Simplified Approach*, Nuffield Economics Working Paper.

Elbers, C., J. O. Lanjouw and P. Lanjouw (2003), Micro-level Estimation of Poverty and Inequality, *Econometrica*, 355-364.

Leslie, D. (1998), *An Investigation of Racial Disadvantage*, Manchester University Press, Manchester.

Modood, T., R. Berthoud, J. Lakey, J. Nazroo, P. Smith, S. Virdee, and S. Beishon (1997), *Ethnic Minorities in Britain: Diversity and Disadvantage*, Policy Studies Institute, London.

Russell, C., K. Cole, M. A. S. Jones, S. M. Pickles, M. Riding, K. Roy, M. Sensier (2003), Grid Technology for Social Science: The SAMD Project. *IASSIST Quarterly*, 27#4.

Endnotes

1. IID: identically and independently distributed. ID: independently distributed.
2. The "." subscript means that an average has been taken over that index.
3. $\tilde{\epsilon}_b$ needs to be appropriately standardised.
4. Minimum, 10% quantile, 25% quantile, median, 75% quantile, 90% quantile, maximum.
5. Space restrictions preclude presentation of the regression results, plots, and tables at the regional and SARs levels of geography.
6. Constant, Gender, Age, Age squared, Children present, Marital status, Labour force position, Housing tenure, High qualifications, Immigrant, Region.
7. Results are reported to whole percentages, and one decimal place for the standard error.
8. *Enclave* refers to GBprofile codes 5, 13, 18, 22, 29 and 33. *Poor* refers to GBprofile codes 1, 7, 17, 21, 23, 35, 36, 27, 43, 44, 45, 49. *The Rest* refers to the remaining GBprofile codes
9. Black-other, Other-asian and Other-other omitted. SHC is the simulated head count. Standard errors are in parentheses.
10. As 9 above.