Some Initiatives on Combining Data to Support Small Area Statistics and Analytical Requirements at ONS-UK

Denise B. N. Silva and Philip Clarke[1] – ONS-UK

1. Introduction

Like any other National Statistical Institute (NSI) around the world, the Office for National Statistics (ONS) faces the challenge of producing comprehensive, accurate and reliable information under financial and time constraints. Although data is frequently requested for detailed geographical areas or domains of study and there is an increasing demand for these figures to be up-to-date, resources for data collection for small areas are very limited. In addition, the pressure for reducing sample sizes and respondent burden reveals the need for methods to produce small area statistics from combined data sources. Small area estimation covers a variety of methods used to produce survey based estimates for geographical areas or domains of study in which the sample sizes are too small to provide reliable direct estimates. This occurs when the survey is not designed for estimation at the required level. In order to obtain reliable estimates, additional datasets are generally brought to bear upon the process. In the last 10 years, ONS has experienced noticeable progress in the implementation of small area estimation methods[2] in which the use of administrative data plays a central role (Clarke, 2005). Based on these successful initiatives, familiarity with model based[2] estimates has grown among users and so have calls for a greater variety of outputs. Model based estimates of sample survey variables at small areas are accepted as a part of ONS established statistical outputs and are exploited by different users including government departments for policy making and research.

This paper introduces the historical background to small area estimation initiatives at ONS, reviews the well established projects that provide model based unemployment estimates for Local Authorities (LA)[3] and income estimates at lower geographical levels, and summarizes the ongoing work to produce estimates for all labour market states simultaneously and also unemployment figures at cross cutting geographies. In addition, it outlines the role of small area estimation methods in the UK Census and also addresses the more recent users' requirements for local area data and the ONS corresponding development projects to meet these demands. The aim is to present an overview of the initiatives in which specialised statistical techniques for small area estimation have already been implemented at ONS together with our views for future advance. The paper does not offer a comprehensive account of the production of small area statistics at ONS. Methods employed for obtaining post-census population and migration figures and the development of the Neighbourhood Statistics (NeSS) programme, for example, are not addressed. However, NeSS played a dual role in the provision of small area estimates by ONS by supplying the necessary auxiliary data and also by demanding and stimulating the production of small area estimates of socioeconomic indicators (see Teague, 2008).

2. In the beginning

Small Area Estimation at ONS was initiated as a research project in the 1990s in response to calls not only for locally focussed information in social, business and environmental subjects but also for more general domain estimation. At that time, ONS expertise came from projects carried out at the Office for Population Census and Surveys (OPCS) and at the Central Statistical Office (CSO), prior to the formation of ONS in 1996. A review of the development of the statistical methods for small area estimation is out of the scope of this paper. However it is worth noting that the research and academic literature on this subject experienced a visible expansion in the late eighties as well as during the

[1] denise.silva@ons.gov.uk and philip.clarke@ons.gov.uk - Office for National Statistics – Methodology Directorate.

[2] Small area estimation methods rely on the use of statistical models that relate the survey data with auxiliary information and produce the so-called model based estimates.

[3] Local Authorities are the main tier of local government in UK. There are 434 LAs with an average population of 140,000. However they vary widely in size from around 60,000 to over 1 million people.

following decade. It is currently a well established research topic with motivation driven mostly by user needs, although the innovative methods are not always simple to implement. This valuable link between theoretical development and practical experiences is evident, for example, in a book on small area statistics published in 1987 (Platek at al., 1987) containing the invited papers from the 1985 International Symposium on Small Area Statistics held in Ottawa. It shows how relevant practical problems attracted not only the interest of official statisticians but also of well-known academic researchers. This allowed NSIs to benefit from fruitful working relationships with academia. In 1992, the Central Statistical Office of Poland hosted the International Scientific Conference on Small Area Statistics and Survey Designs and, once more, researchers from the academia and those working in several NSIs (including OPCS) attended the meeting.

OPCS also took this promising route and commissioned the University of Southampton to examine the potential for small area estimation. Results of this investigation were published in 1993 (Skinner, 1993). From 1995 to 1997, projects to obtain small area estimates of demand for general practitioner services and incidence of psychiatric morbidity were carried out for the Department of Health. In addition, an investigation into the use of multilevel models for small area estimation was conducted jointly with another academic consultant (Heady at al., 1997). In the same period, survey statisticians from CSO took part in a course promoted by the University of Southampton under the auspices of Eurostat in which domain estimation was one of the topics discussed. This knowledge contributed to the preparation of a report discussing the need for regional business statistics and reviewing some available techniques for small area/domain estimation (Baskerville at al., 1995).

In 1998, a Small Area Estimation Project (SAEP) was formally established at ONS as a research and development programme attracting support from other government departments. SAEP had the ambitious aim of "*developing generalised statistical methodology and operational system for deriving estimates to known precision from variables contained in social surveys, for areas defined by a variety of boundary systems*" (ONS, 2003). Although experience now shows that there is no such thing as a *generalised* methodology for small area estimation, SAEP proved to be a successful project. In conjunction with SAEP, a Eurostat SUP.COM[4] project was conducted in partnership with Statistics Finland, University of Jyvaskyla and the University of Southampton (section 3.1 provides a brief discussion on the SAEP and related projects). At the same time, ONS was heavily involved in the preparation of the 2001 Census and the development of the One Number Census (ONC) methodology. It is important to note that small area estimation methods were introduced within the ONC project that aimed *"to provide Local Authority District level population estimates by age-sex groups that have been adjusted for the undercount in the 2001 Census and to adjust the census database for this undercount at an individual level so that all census outputs add up to One Number"* (ONS, 1998). Small area estimation techniques were implemented to improve the precision of LA population estimates obtained from the 2001 Census Coverage Survey by incorporating auxiliary information and assuming relationships between the undercount pattern at LA level and broader areas. Section 3.2 outlines the 2001 census coverage adjustment and the role of small area estimation in this process.

Besides the two initiatives mentioned above, ONS was also engaged in conducting research to develop improved estimation methods for local area labour market indicators. Once again a partnership between academic consultants and official statisticians took place (see Ambler at al. 2001). Combining Labour Force Survey data with administrative data on Job Seekers Allowance[5], the proposed method provides model based LA estimates for unemployment level and rates that have already being accredited as National Statistics[6]. A summary of this project is presented in Section 3.3. The efforts in developing small area estimation methods at ONS received an extra boost in 2001 with the establishment of the EURAREA[7] project, funded by the European Community. The project activities

[4] Activities of scientific and technical SUPport of a COMpetitive nature. Eurostat funded research projects.

[5] Jobseeker's Allowance (JSA) is the main benefit for people of working age who are out of work but actively seeking work.

[6] The term "National Statistics" is an accreditation kitemark which stands for a range of qualities such as relevance, integrity, quality, accessibility, value for money and freedom from political influence (from www.statistics.gov.uk).

[7] http://www.statistics.gov.uk/eurarea/download.asp

took place from 2001 to 2004 and involved NSIs and universities from seven European countries (ONS and the University of Southampton from UK). The aim of the project was "*to assess the potential of applying standard small area estimators in the European context*" (Heady and Hennell, 2001). Pulling together the experience of applying small area estimation techniques to the One Number Census and the years dedicated to research, in 2001 ONS entered an implementation phase in which small area estimates were produced for income and unemployment. The following sections focus on the implementation of the projects, describing the techniques, data used as well as the outputs produced.

3. Successful ONS experiences on small area estimation

In order to set up a small area estimation framework it is necessary to define the target for estimation (comprised of the variable of interest plus the corresponding summary measure and the geographical level or domain requirements) and also to identify the potential auxiliary data. There are a variety of geographic unit types (administrative, health, electoral, postal, etc.) in the UK and their boundaries do not always align[8]. This means that linkage of survey and administrative data is not always straightforward imposing some restrictions on the choice of small area techniques. Throughout this paper, the reader will encounter references to some of the following geographic units: postcode units and sectors, wards, middle layer super output areas, parliamentary constituencies and local authorities. Those that have not been mentioned yet will be defined when quoted.

The initiatives introduced in this paper are related to social statistics and the production of small area estimates for income, unemployment and population estimates in the 2001 ONC. In these cases, survey data comes from the Family Resources Survey, the Labour Force Survey and the Census Coverage Survey, respectively. Auxiliary information is usually obtained from Census and from socioeconomic administrative databases which are only made available at aggregated levels, as highlighted by Clarke (2005). This is because: there is no unique system of individual statistical registers in the UK; the administrative systems that hold individual and household data use distinct identifiers and data sharing/matching activities operate under very controlled conditions in UK due to the 1998 Data Protection Act[9] (see ONS website for protocols on data access, confidentiality and data matching and also for a guide to legal framework on data sharing for statistical purposes).The problem of small area estimation refers not only to the production of reliable estimates but also on how to assess the estimation error. In numerous cases the latter constitutes the most complex part of the estimation procedure. The methods currently employed by ONS are based on explicit models that relate survey direct estimates for small areas or domains to supplementary data.

3.1 The small area estimation project (SAEP) and the production of small area estimates of income

The SAEP project followed the initial ONS work in small area estimation and was a co-ordinated effort to introduce a methodological *system* which could be applied for household sample survey variables within the framework of UK statistics. Broadly speaking, the framework concerns the following components.

Sample social surveys: Are designed to obtain good estimates at national and sometimes regional level. They are based on samples of households drawn from an address register owned by the postal service (postcode[10] address file, PAF). Clustered sample designs are generally adopted with postcode sectors[11] as the primary sampling units (PSUs). These are variable in size but on average would be roughly 2,500 households. The sampled PSUs are designed to give representative samples at the national level but are not suitable for small area estimation – in particular the clustering will mean that many small areas will have no sample.

[8] For details consult http://www.statistics.gov.uk/geography/default.asp.

[9] It is a UK Act of Parliament that regulates the processing and disclosure of information relating to individuals.

[10] A postcode is associated with each address in the UK and they are assigned by the Royal Mail. They are also a key means of providing locational references for statistical data. There are approximately 1.78 million postcode units in the UK.

[11] A postcode sector is formed by a set of contiguous postcode units. There are roughly 11,600 in the UK.

National census: Has many variables which would be good for explanatory purposes however it is only carried out every 10 years and the auxiliary information becomes out of date over time.

Administrative systems: Can hold many useful variables for explanatory purpose and similarly to the census is not subject to sampling error. Such data includes claimants of various social benefits, data on dwelling categories for local taxation etc. Administrative records however are not collected for statistical purposes so care must be taken when utilising the data regarding the concepts, definitions, target population, collection mode and the purpose for data collection.

Lack of statistical registers: There is no national register of households/persons in UK. This means that there is no straightforward way of linking individual survey responses to their census or administrative records.

Preponderance of incompatible and ever changing boundary systems: Unlike in many other countries, boundary systems in UK exist for varying purposes and very often do not co-exist well with each other. Moreover they often change over time. A particular problem for small area estimation is that the sample design is based on postcode sectors while the usual requirement for small area estimation is for administrative areas such as local authorities or smaller nesting units. These two systems exhibit no commonality!

The SAEP project was designed to address the difficulties described above. Within ONS, unit level census and sample records were available with an address georeference allowing placement in any of the available digitised boundary systems (this includes postal and administrative geographies). Administrative data were available at aggregated levels for administrative boundaries so the models were designed with a **unit level response** plus **area level covariates** and catered for the varying sampling and estimation geographies by separating the fixed effect and the random effect structures. The underlying model is a two level model $y_{id} = \underline{X}_d \beta + u_j + \varepsilon_{ij}$ where y_{id} is the survey variable for household *i* in the estimation area *d* and *j* represents sampling area. In addition, $u_j \sim N(0, \sigma_u^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ are random terms at sampling area level and household level respectively and $\underline{X}_d$ is a vector of values for area *d* of a set of covariates.

Models are fitted just on the sample data and using covariates in areas for which a sample is present. However covariates are available for all areas. This allows SAEP to determine a synthetic estimator of the mean for all areas given by $\hat{\bar{y}}_d^{synth} = \underline{X}_d^T \hat{\beta}$. However, the SAEP framework has a limitation in the order of area size at which a precision can be determined. For a model where sampling and estimation areas are the same, an estimator of the mean squared error of the estimated area mean would be given by $\hat{\sigma}_u{}^2 + \underline{X}_d^T Var(\hat{\beta}) \underline{X}_d$. Empirical testing using an unclustered survey has shown that the values of σ_u^2 are comparable when the estimation areas and the areas that form the random effects are of comparable size but not when their sizes vary considerably. For this reason, SAEP models are used only for estimation at areas of similar size to the sampling postcode sectors. In terms of administrative areas these are electoral wards[12] or the more recently established middle layer super output areas (MSOAs)[13].

The SAEP is however more than just the modelling framework. It involved the establishment of a set of potentially useful area level covariates from the census and administrative systems, georeferencing of survey respondents to required areas and then matching to appropriate covariate values. Auxiliary data sources and corresponding covariates considered in the SAEP system are obtained from: the 2001 Census, the Department for Work and Pensions (social benefit claimants), the Valuation Office Agency (council tax banding data), the Land Registry (house sale price index) and the

[12] Electoral wards/divisions are the key building block of UK administrative geography, being the spatial units used to elect local government councillors. There are approximately 11,600 electoral wards in the UK.

[13] Super Output Areas (SOAs) is a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales. There are 3 layers in the hierarchy and 7193 middle layer super output areas (MSOAs).

HM Revenue and Customs (HMRC) (Income tax data). These are all aggregate data at area level, e.g. the number of individuals in an area with income above a certain threshold (from HMRC data).

The model fitting is an important part of the SAEP system. As there are a large number of possibilities, a forward stepwise procedure is used starting with a null model, i.e. just an intercept. The accepted model is that produced when all included covariates are significant and no others are significant enough to enter. Finally model diagnostic procedures are performed and, when satisfactory, estimation is carried out. As noted in Section 2, a partnership was conducted with Statistics Finland and others during the development of SAEP which formed the SUP.COM project. Essentially the project compared model based and model assisted design based methods using samples drawn from Finland's statistical registers. This enabled the quality of resulting estimates to be empirically tested against the register population value. It formed a major advance in the validation of the SAEP methodology. The SAEP system has been used to publish estimates of mean household income for wards using the Family Resources Survey in 1988/89 (Longhurst at al., 2004) and 2001/02 (Goldring at al., 2005) and for MSOAs in 2004/05 (estimates and report are available at the NeSS website).

3.2 Small area estimation for the 2001 One Number Census

The use of model based small area estimation methods in the UK Census was introduced in 2001 within the ONC project. A very simplified description of the ONC identifies the following key elements/stages in the process: the Census itself; the Census Coverage Survey (CCS); the matching of Census and CCS databases to estimate undercount at national and sub-national levels[14]; **the process to obtain model based population estimates for Local Authorities**; the production of a database with individual and household level records consistent with LA population level estimates. In the 2001 Census, small area estimation techniques based on standard regression models were employed to produce population estimates (and respective quality measures for each LA by hard to count index (HtC)[15] and age-sex groups using data from the CCS and the Census. Direct survey estimates for LAs could have been obtained from the 2001 CCS but with corresponding large standard errors due to small sample sizes[16] (by LA) and the survey design. The use of small area estimation techniques was introduced to improve the precision of the LA estimates incorporating auxiliary information by assuming relationships between the undercount pattern at LA level and broader areas (*design groups*[16]). The underlying idea of the method was to exploit similarities in order to borrow strength over areas.

The small area estimation technique implemented[17] consisted of fitting simple linear regression models to relate CCS counts with the (unadjusted) population counts from the Census in each design group, allowing the model coefficient (the slope of the regression line) to vary according to LAs. The heterogeneity of the slopes accounted for differences in census coverage between LAs within a specific design group (see Abbot at al., 2000). The chosen **area level model**[18], used a simple linear regression model through the origin with the CCS count as the response/target variable and the unadjusted Census count as the explanatory variable with different coefficients (slopes) across LAs. For each design group, the model specification for the small area estimation procedure is given by:

$$Y_{kadl} = \left(\theta_{cd} + \gamma_{dl}\right) X_{kadl} + \varepsilon_{kadl}\sqrt{X_{kadl}} \quad \text{with} \quad a \in c \quad (\text{c defined as collapsed category levels}) \ \text{and} \ \sum_{l \in g} \gamma_{dl} = 0 \ ,$$

where Y_{kadl} and X_{kadl} are the CCS count and the unadjusted Census count, respectively, for postcode *k*, age-sex group *a*, HtC stratum *d*, Local Authority *l* in a given design group *g*. The model was fitted separately for each HtC stratum within each design group using age-sex by postcode level data.

[14] *Design groups* or *estimation areas* are mutually exclusive groups of LAs.

[15] The HtC index, defined for 1991 Census enumeration districts (ED), was used as a stratification variable for the 2001 CCS. It was obtained by ranking the EDs according to non response rates and other characteristics (Brown at al., 1999).

[16] 2001 CCS sampled 320,000 households in England and Wales.

[17] For a description of the different small area estimation approaches considered see Abbot at al. (2000).

[18] Area level models relate the small area quantities of interest to area-specific auxiliary variables.

Therefore, the model based estimator for the population total by LA, HtC stratum and age-sex group is defined as: $\hat{T}_{adl} = \sum_{k \in s_{dl}} Y_{kadl} + \sum_{k \notin s_{dl}} \left(\hat{\theta}_{cd} + \hat{\gamma}_{dl}\right) X_{kadl}$ for $a \in c$ with s_{dl} the sample in HtC stratum d and LA l ·

The estimates were then calibrated to the "gold standard" population figures by 5 year age-sex groups for approximately one hundred design groups (more information can be found at the ONS website[19]).

3.3 Small Area Estimates for Unemployment and Unemployment rate in GB

Since 1999 ONS has been continuously devoting efforts to develop and improve a small area estimation methodology to provide local authority estimates of unemployment based on annual Labour Force Survey[20] data. The survey is the key source of national information on the labour market however it is not able to deliver direct estimates of unemployment with adequate precision for every local authority in Great Britain because the sample size in many areas is too small, leading to large sampling variability. A project was established to produce LA estimates of ILO[20] unemployment levels and rates using model based methodology jointly developed by the ONS and the University of Southampton. The model based approach relies on determining a strong relationship between ILO unemployment (as measured by the Labour Force Survey) and auxiliary information. This relationship is then used to provide more reliable estimates of ILO unemployment for LAs. The main source of this auxiliary information is the register of the number of recipients of job seekers allowance (the 'claimant count').

The first attempt to produce unemployment small area estimates within this framework is reported in Ambler at al. (2001) and Brown at al. (2001). In this case, a model based approach was formulated using **an area level** logistic model to relate the probability of being unemployed for an individual of a particular sex and age group within a LA with the corresponding claimant count[21] information, incorporating additional explanatory variables accounting for age, sex and regional differentials. The model based estimates were published in 2003 as experimental statistics[22] (as reported in Hastings at al., 2003) and after that the methodology was improved to incorporate LA random effects to better capture unexplained sources of variation and area heterogeneity (see Saei and Chambers, 2003). This allows the relationship to be different for different LAs recognising that there may be between area differences that are not explained by the auxiliary data. The model framework comprises the specification of a logistic linear mixed model to relate the unemployment probabilities in a given age-sex group in each LA with the auxiliary information, given by

$$\text{logit}\left(p_{di}\right) = \ln\left\{\frac{p_{di}}{1-p_{di}}\right\} = \boldsymbol{X}_{\boldsymbol{di}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_d \quad \text{such that}_{di} = \frac{exp\left(\boldsymbol{X}_{\boldsymbol{di}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_d\right)}{1 + exp\left(\boldsymbol{X}_{\boldsymbol{di}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_d\right)}$$

where p_{di} is the probability that an individual in (age-sex) group $i = 1,\ldots,6$ from LA *d* is unemployed, $\boldsymbol{X}_{di}$ is the column vector of group indicators and covariates for age-sex group *i* in LA *d*, $\boldsymbol{\beta}$ is the vector of fixed effect coefficients and $u_d \sim N(0,\phi)$ is the random area effect that accounts for between LA variability beyond that explained by the auxiliary variables included in the model. The model covariates are indicators of age-sex groups (male/female for age groups: 16 to 24; 25 to 49; 50 and over), of the 12 government office regions[23] of GB and of the ONS socioeconomic family classification for LAs plus the logit of the claimant count proportion in each age-sex group within each LA and the logit of the claimant count proportion in the LA. An estimate of the unemployment level in LA *d* is

[19] www.statistics.gov.uk/census2001/onc.asp and www.statistics.gov.uk/census2001/onc_evec_eval.asp.

[20] The Labour Force Survey is a continuous survey, with a sample of around 60,000 households in each three-month period, and measures unemployment according to the International Labour Organisation (ILO) definition.

[21] Claimant count measures the number of people seeking work who claim unemployment benefits (job seekers allowance).

[22] Term used to define statistics that are in the testing phase and are not fully developed.

[23] "The nine Government Office Regions (GORs) are the primary statistical subdivisions of England and each GOR covers a number of local authorities" (http://www.statistics.gov.uk/geography/glossary/g.asp).

given by $\hat{\theta}_d = \sum_{i=1}^{6} \hat{\theta}_{di} = \sum_{i=1}^{6} \{ y_{sdi} + (N_{di} - n_{di}) \hat{p}_{di} \}$ where y_{sdi} is the LFS sample count of number of unemployed, N_{di} and n_{di} are the population and sample size for group i in LA d, respectively, and $(N_{di} - n_{di}) \hat{p}_{di} = \hat{y}_{rdi}$ is the predicted value for y_{rdi} - the LFS non-sample count of number of unemployed obtained from the modelling procedure fitted just on the sample data.

The work progressed towards the implementation of this logistic mixed model to produce annual small area estimates (they use an average of the previous twelve months claimant count totals and twelve months of survey data) that are published every quarter at the ONS website[24]. To ensure that the model based estimates are consistent with the Labour Force Survey published estimates at high geographical levels, they are constrained to the corresponding direct LFS estimates of unemployment.

3.4 Lessons learned from these experiences

A key aspect of a small area estimation system is the ability to use auxiliary data to improve on the direct estimates obtained from sample surveys. Small area estimation techniques are used to overcome the problem of small sample sizes. However, although more precise, the resulting model based estimates are often biased[25] (in a statistical sense). The aim of the estimation procedure is to balance the trade-off between variance and bias, producing estimates with good precision and with as little bias as possible. This depends on the adequacy of the modelling procedures and on the availability of covariates with good prediction power. Although successful experiences are reported here, ONS was involved in other projects and feasibility studies that did not result in publishable outputs because of the lack of adequate auxiliary data. These initiatives however were vital for capacity building providing the means for acquiring a practical understanding of the underlying statistical principles involved in the model based estimation activity.

ONS became aware on how imperative it is to validate the model based estimates prior to publication. In order to compare different models and evaluate their performance as well as to check whether a small area model was producing adequate estimates, several diagnostic tests were developed and employed. Results were compared against other sources of data available (proxies) and users' consultations were carried out to guarantee that estimates were plausible and fit for purpose. In addition, the methodologies were subjected to academic review.

The implementation of the reported initiatives proved how crucial the validation stage is and how much effort has to be devoted to transfer the project from the development area to the production (business) area, to prepare the indispensable documentation and to communicate the outputs to the general public. Overall, the main challenges for NSIs when producing small area estimates is the ability to master the complexities of the required statistical theory (e.g. the assessment of the estimation error is recognised as a complex problem in the small area estimation context), the availability of relevant administrative data and the capacity to overcome internal and external barriers for the acceptance of model based estimates as trustworthy official statistics outputs.

4. When we find the answers then they change the questions

Following the developments reported in Section 3, ONS has been facing new and unavoidable requests for small area estimates. More familiar now with the concept of small area estimation, both the users and ONS recognise the potential for fulfilling previous unmet demands for small area statistics as well as some more recent ones. Examples are the production of unemployment estimates for parliamentary constituencies (PCAs)[26] and also model based estimates for all labour market categories/states; the request from the Department for Work and Pensions and other analysts for estimates of local area income distribution instead of mean income and the estimation of proportion of

[24] Local area labour markets: statistical indicators – (www.statistics.gov.uk/STATBASE/Product.asp?vlnk=14160).

[25] Bias refers to a systematic error contributing to the difference between the sample estimate and true population value.

[26] The UK is currently divided into 646 PCAs, each of which is represented by one MP in the House of Commons.

households with income below a threshold; and moreover the call for adjusting the population estimates, based on small area models, at a lower level than local authority in the 2011 Census. This section presents some development work currently being carried out at ONS to meet these demands.

ONS publishes LA model based estimates for unemployment and unemployment rates and in addition there is a need for reliable labour market unemployment data for parliamentary constituencies. This has been expressed by the Librarian of the House of Commons Library whose staff and Members of Parliament (MP) will be the main audience. There are 35 parliamentary constituencies which are identical areas to local authorities and there are also LA areas which are coterminous to a number of PCAs. Therefore, this is a case in which estimates are requested for two non-nested geographies. The original model developed for LA estimation (Section 3.3) was taken as the basis for the parliamentary constituency model which in turn is supplied with the same sort of survey and administrative data (social benefits) for the constituencies. The results show good performance but with less explanatory power than the local authority model.

Issues arise from having alternative estimates for the same areas from the local authority and the parliamentary constituency models where areas coincide and also because PCA model estimates may not add up to LA model based published figures for cases where several parliamentary constituencies exactly make up one local authority (this affects 48 LAs and 150 PCAs). The PCA estimates have to be constrained to direct survey estimates of regional totals and, in addition, where a PCA is the same area as a LA, the proposed estimation procedure constrains the model based PCA estimate to be the same as the model based estimate for the LA. Moreover, it has to be considered if the PCA estimates should also be calibrated to LA figures for all other local authorities that are conterminous with multiple PCAs. The new estimates are still subjected to further evaluation, the issue of how the parliamentary constituency estimates align with the model based estimates for the local authorities and the need for extra constraining/calibration are being addressed. The discussion exposes the duality between producing estimates based on statistical procedures that are optimised for a given geography and the need for calibrating/constraining the estimates for presentation purposes. It highlights that consistency is an important feature of any statistical system and the establishment of a small area estimation framework certainly has to address the issue of consistency between model based estimates produced for different geographies besides the consistency of model based estimates and direct surveys estimates published for higher/broader geographies.

Coherence is also relevant when dealing with multipurpose analysis and corresponding estimation procedures. Once again focussing on producing small area estimates to inform labour market analysis, ONS publishes estimates not only of unemployment indicators, but also covers all the labour market states and related variables. The current state of local labour market status estimates is that the estimate for unemployment (levels and rates) is model based while estimates of employment and not economically inactive are direct survey estimates. This implies that the sum of the modelled unemployed level and employed level will not match a survey estimate of the total economically active population. ONS is developing a project on multivariate estimation in which the variable of interest has a multinomial response and the proportions of individuals classified in each of its categories are proportions of a whole subject to a unity-sum constraint.

The analytical motivation for this project is the need for information on how the relationship between benefit take up and worklessness varies at small area level, which can be used to direct interventions to satisfy government targets. The term worklessness describes all those without work (and includes not only those actively seeking work but also those who are economically inactive). The aim is to build an estimation procedure in which the three labour market categories are modelled simultaneously in a coherent way. The method builds on the previous work (Section 3.3) and extends the binomial model to a multinomial one that relates the probability of a person being unemployed, employed or inactive with auxiliary information. In this case a multivariate random effects model is employed and the logit for two of the categories is defined in terms of the third one. This model has been developed theoretically (Saei and Chambers, 2005) however there is work remaining before an operational system can be set up for the publication.

Estimation of household income using the SAEP system has currently targeted the mean income. However, for the purposes of social support and regeneration, it is the distribution of income in each area (in particular that of its lower tail in representing poverty) which is of major interest. For the measures of equivalised household income which allow for household size, a poverty level exists defined as 60% of national median income. SAEP methods can be used to estimate proportions of a binary variable – in this case of household income under this threshold value – by modelling under a logistic transformation. However, results of such modelling to date have not been sufficiently explanatory to permit publication. Further investigations are currently in progress.

The development work reported above indicates that ONS is committed to provide good quality small area estimates whilst responding to relevant users' requirements. The more we invest in meeting the demands for small area statistics and the more successful we are in obtaining small area estimates, the more complex the estimation system becomes as care has to be taken to ensure comparability over time (when repeat estimation takes place) as well as consistency over area/domains and coherence over variables. Queries on how to compare or analyse change over time and on how to aggregate small area estimates to broader areas or domains have to be addressed due to soaring users' calls. Besides this development work, ONS has also started studying the problem of small area estimation for business surveys (Hidiroglou and Smith, 2005; Merad and Brodie, 2008) and is examining models to improve migration small area statistics (Heasman, 2008). In addition, the office also carries out research projects (for an outline of the current ones see Clarke at al., 2007).

5. Preparing for future challenges

Model based estimation methods are powerful tools not only for dealing with small area problems but also as a promising route to allow ONS to keep delivering relevant statistics while facing pressures to reduce sample or questionnaire sizes and to optimise survey costs. On the other hand, the use of model based methods is not a panacea and its success relies on good knowledge of statistical theory, on the availability of correlating variables from administrative data sources and the potential for data linkage, and on efforts to educate users about the output limitations. Focussing on gathering the indispensable auxiliary information and recognising the existing challenges for producers of official statistical data, Methodology Directorate at ONS has established a more recent strategy for developing expertise on data matching and data sharing. The aim of the project is to promote capacity building, improving the current knowledge on methods for producing statistics from combined data sources and related quality measures (Cruddas, 2007). There are also other related initiatives to improve ONS' capability in record linkage and data sharing, as for example the collaborative work with the Department for Children, Schools and Families reported in Phillips (2008).

None of developments reported in this paper are free from criticism (some already known and addressed by ONS). At the moment, ONS produces small area estimates using cross-sectional models and repeated independent estimation procedures are conducted over time. If comparability over time and consistency over geographies are important contraints then it is essential to keep investigating models that can account for these complexities - models with time series components (that borrow strength over time) or multilevel models to consistently estimate for different area geographies. In addition, there is also need to experiment resampling[27] methods for calculating precision of the estimates. The approach taken by the office is to keep the dialogue with external experts, especially those in academia, to be able to continuously enhance the methods.

Acknowledgement

The authors wish to thank Jane Longhurst and Marie Cruddas for their comments and suggestions and Minda Phillips for the helpful discussion.

[27] Simulation methods for estimating the precision of sample statistics by recomputing the statistics several times using subsets of available data or drawing randomly with replacement from a set of data points.

6. References

Abbott, O., Brown, J., Chambers, R. and Cruddas, M. (2000) *One Number Census Local Authority Estimation*. Paper submitted to the One Number Census Steering Committee. (www.statistics.gov.uk/census2001/onc.asp)

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M., Wang, S. (2001) *Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method.* Proceedings of the 53th ISI Session, Seoul.

Baskerville, D., Byrne, M. , Cruddas, M. and Smith, P. (1995) *Regional Statitics – a leadership management project.* CSO unpublished report.

Brown, G., Chambers, R., Heady, P. Heasman, D. (2001) *Evaluation of Small Area Estimation Methods – An Application to Unemplyment Estaimtes from UK LFS.* Proceedings of Statistics Canada Symposium 2001.

Brown, J.J., Diamond, I.D., Chambers, R.L., Bucker, L J., Teague, A. D. (1999) *A methodological strategy for a one-number census in UK.* JRSS A, 162, Part 2, pp.247-267.

Clarke, P. (2005) *Small Area Estimation: UK Experiences.* Paper presented at the 55th ISI Session, Sydney.

Clarke, P., McGrath, K., Chandra, H. and Tzavidis, N. (2007) *Developments in Small Area Estimation in UK with Focus on Current Research Activities.* Paper presented at IASS Satellite Conference on Small Area Estimation. 56th ISI Session.

Cruddas, M. (2007). *Combining Data: Developing a Centre in MD to Meet the Challenges.* Paper presented at the 12th Meeting on the National Statistics Methodology Advisory Committee. ONS.

Hastings, D., Maine, N., Brown, G. Cruddas, M. (2003). *Development of improved estimation methods for local area unemployment levels and rates.* Labour Market Trends, vol. 111, no 1. ONS publication.

Haworth, M. and Cruddas, M. (2003) *Developing small area estimates in the UK – a practitioners' perspective.* Proceedings of Statistics Canada Symposium.

Heady, P. and Hennell, S. (2001) *Enhancing Small area estimation techniques to meet European needs.* Statistics in Transition, vol.5, no.2, pp.195-203.

Heady, P., Ruddock, V. and Goldstein, H. (1997) *An investigation into the possible use of multi-level models based on survey data to update census estimates for small areas.* Proceedings of the 97 Symposium. Statistics Canada.

Heady, P., Clarke, P. Brown, G., Ellis, K., Heasman, D., Hennel, S., Longhusrt, J. and Mitchell, B. (2003) *Model-based small area estimation series. No.2 – Small area estimation project repor*t. ONS publication.

Heasman, D. (2008) *A local area model for emigration.* Paper presented at the 13th GSS Methodology Conference.

Hidiroglou, M. and Smith, P. (2005) *Developing small area estimates for business surveys at the ONS.* Statistics in Transition, 7, 527-539.

Goldring, S., Longhurst, J. and CruddasM. (2005) *Model-based estimates of income for wards, 2001/02 – technical report.* ONS publication

Longhurst, J., Cruddas, M., Goldring, S. and Mitchell, B. (2004) *Model-based estimates of income for wards, 2001/02 – technical report.* ONS publication

Merad, S. and Brodie, P. (2008) *Small domain estimation in business surveys.* Paper presented at the 14th Meeting of the National Statistics Methodology Advisory Committee.

ONS(1998). *2001 - A One Number Census. Consultation Paper.* (www.statistics.gov.uk/census2001/onc.asp)

Phillips, M. and Sinclair, P. (2008) *Using Administrative Data to Improve Social Statistics - An Example of Collaborative Work.* Paper presented at IAOS conference on Reshaping Official Statistics.

Platek, R., Rao, J.N.K., Sarndal, C.E. and Singh, M.P. (1987). *Small Area Statistics – An International Symposium.* John Wiley & Sons.

Saei, S., Chambers, R. (2003) *Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects.* S3RI Methodology Working Paper M03/15, University of Southampton.

Saei, A. and Chambers, R. (2005) *Small area estimation methods for multi-category response variables (applications to LFS Data).* S3RI unpublished technical report to ONS.

Skinner, C. (1993) *The use of synthetic estimation techniques to produce small area estimates.* New Methodology Series 18. London: OPCS (Social Survey Division).

Teague, A. (2008) - *A strategy for making better use of administrative data for statistics in the UK* . Paper presented at IAOS conference on Reshaping Official Statistics.