The creation of a national agricultural land use dataset: combining pycnophylactic interpolation with dasymetric mapping techniques

Alexis Comber¹, Chris Proctor², Steve Anthony²

¹Department of Geography
University of Leicester,
Leicester,
LE1 7RH, UK
ajc36@le.ac.uk

²Environmental Systems,
ADAS,
Wolverhampton,
WV6 8TQ, UK
chris.procter@adas.co.uk
steve.anthony@adas.co.uk

Key words: land use, volume preserving, dasymetric, agricultural census

Abstract

Agricultural Census data is summarised over spatially coarse reporting units for reasons of farm confidentiality. This is problematic for research at a local level. This paper describes an approach combining dasymetric and volume preserving techniques to create a national land use dataset at 1km² resolution. The results for an English county are compared with contemporaneous aggregated habitat data. The results show that the accurate estimates of local agricultural land use (Arable and Grass) patterns can be estimated when individual 1km squares are combined into blocks of >9 squares, thereby providing local estimates of agricultural land use. This in turn allows more detailed modelling of land uses related to specific livestock and cropping activities. The dataset created by this work has been subject to extensive external validation through its incorporation into a number of other national models: nitrate leaching (e.g. MAGPIE, NEAP-N), waste and pathogen modelling related to agricultural activity.

1. Introduction

In this paper we demonstrate a method for generating local land use estimates at 1km² from aggregated agricultural census data for England and Wales. The June Agricultural Census (JAC) summarises land use over reporting units that are spatially coarser than the level at which the data was originally collected for reasons of confidentiality or expediency. The JAC is an annual survey of agricultural activity which records information from agricultural holdings (farms) relating to land use, crops, livestock, labour and horticulture. In order to preserve farm confidentiality only aggregate data are released. The way that the aggregate data are reported has changed over time and a number of reporting units have been used: Parish (or Parish Groups), District, County and

Country Summaries. From October 2004 NUTS level 5 areas based on “Super Output Areas (middle layer) replaced those based on wards as the standard lower-level unit¹.

Periodic changes in reporting unit is the norm for the JAC as for many other census data collection activities. Changes in the spatial units over which detailed data is summarised result from changes to political geographies. Evans (1996) noted that issues associated with changing geographies, such as quantifying land use change, have been exacerbated by recent changes in the structure of agriculture at the individual farm level.

Modernisation and expansion in the late 1980s made it difficult to conceal information about individual farms using previous Parish summaries. The agricultural census literature is littered with research examining ways in which data reported using one type of reporting unit can be transformed into another using various areal interpolation approaches, all of which seek to manage the modifiable areal unit problem (MAUP) as described by Openshaw (1984). The MAUP is the interaction between the scale of data reporting and effects of data aggregation. Traditional approaches for generalising agricultural (and other census) data within reporting units (e.g. parishes, parish groups) distribute the information evenly across those units. There are two major problems with this approach. First, the assumption of within-unit homogeneity is problematic as there is often considerable internal variation in land use patterns and practice (Allenson and Moxey, 1996). Second, and specific to agricultural census data is that the inclusion of any one holding within the reporting unit is subject to uncertainty as the aggregation of farm returns is done on the basis of the location of the farm office. Hence, field areas reported under a civil parish may not be located within the parish boundaries (Coppock, 1960) and may lead to under- or over-reporting of the true agricultural land area within a reporting unit, and a distortion of the relative mix of land uses within a parish (Hooper, 1989). Within the context of the disaggregating JAC data from coarse *source zones* to spatially finer *target zones* any assumption of homogeneity is extremely unrealistic. It is this counterpoint in the MAUP that drives much research into agricultural census data analysis.

In this paper agricultural census data is disaggregated from lower level reporting units (source zones) into a 1km² grid (target zones) using a combination of dasymetric and volume preserving techniques. The approach seeks to tackle issues caused by the MAUP using dasymetric techniques combined with iterative refinement whilst preserving and smoothing the data ‘volumes’ (areas of arable land). The context for this work is the need to be able to generate spatially consistent land use data across time to be able to support science-led agricultural research and policy developments. Whilst, historically these activities have been hindered by changing reporting units and the desire to preserve anonymity, the data generated in the work described in this paper has been used as input into a range of spatially explicit models of Ammonia emissions (Webb et al., 2006), Nitrate leaching (Smith et al., 2003), and Agricultural impacts on water quality (Lord et al., 2002), amongst others. These are required to report at local, catchments and national scales (Lord and Anthony, 2000) in order to fulfil various EU directives relating to the environmental impacts of agriculture.

¹http://www.defra.gov.uk/esg/work_html/publications/cs/farmstats_web/Publications/data_documents/data_notes.htm#nuts

2. Background

Historically long term data collection activities such as censuses are subject to changes in reporting units. In order to answer questions that relate to alternative geographies and to make comparisons of the data across time, data from one set of spatial units have to be translated into another set of spatial units.

2.1 Agricultural census data in England and Wales

The Agricultural Census has been collecting data from individual farms since 1866 by the ministry responsible for agriculture, currently the Department for the Environment, Food and Rural Affairs (DEFRA). Information on individual agricultural holdings is confidential and only aggregate data are released in the form of Parish, County, and latterly NUTS summaries. Haines-Young and Watkins (1996) provide a brief review of the evolution of the nature of agricultural census data collection. There are significant and acknowledged problems in summary agricultural census data, especially related to the nature of the statistical reporting units (Coppock 1955, 1960). Farms may not sit precisely in those parishes only their 'headquarters', usually location of the farmhouse. This results in both under- and over-estimation when the data for those units are analysed. Geddes et al. (2003) note that amount of error varies depending on the relative size and spatial arrangement of the farm holdings and reporting units. Allanson and Moxey (1996) identify further two problems associated with confidential time series data collected over such a long period. First, data aggregated in this form does not lend itself to local analysis. Second, the administrative units are arbitrary and frequently change due to political re-organisation. However, spatially distributed agricultural data is useful for agricultural policy evaluation, development and scenario testing. In 1964 Terry Coppock recognised this and produced an agricultural atlas which pioneered the use of electronic digital computers in manipulating census data for mapping purposes (Coppock, 1964). Evans and Morris (1997) describe an increasing demand for policy-oriented and theory-led research, which takes account of the on-farm responses to new policies in terms of farm adjustment strategies and farm business restructuring. The recent focus of much agricultural research has been on the role of the farm as a point of production within a wider, global food system (e.g. Whatmore, 1995; Sullivan et al., 2004); on the nature of farm diversification activities (e.g.; Evans and Ilbery, 1993; Leff et al., 2004) and the increased interest in agri-environmental issues (e.g. Cook and Norman, 1996; Chamberlain et al., 2000; Vaughan et al., 2003).

Analysis of land use (e.g. change monitoring) at anything other than a very coarse spatial scale is hampered by changes in the way that data is collected and reported. This problem is commonly overcome by spatially reworking or transforming the agricultural census data. One of the most commonly used techniques to do this is areal interpolation (Goodchild et al., 1993) and recent studies considering changes in agricultural land use include Allanson et al. (1992), Allanson and Moxey (1996) and Geddes et al. (2003). The work of Allanson and Moxey (1996) is typical of this body of research, integrating land cover data with agricultural census statistics. Agricultural census data is spatially

transformed without taking into account all of what is already known about the area under consideration.

2.2 Areal Interpolation

Areal interpolation techniques generate values for a set of standardized units or ‘target’ zones (Langford et al. 1991). Gregory and Ell (2005) review spatial interpolation methodologies that reallocate population census data from one set of administrative units onto another using within geographical information systems. They note that “*the true potential of these data has not been realized, the main reason being that the boundaries of the administrative units that are used in publishing the data change over time*” (Gregory and Ell, 2005, p.419). Areal interpolation in its simplest form overlays the source units onto the target units. Data for the new areal units, target zones, are usually estimated as a weighted average (or weighted sum) of the data of the original units (source zones) with which they intersect and the weighting is in proportion to the intersecting area. That is, the values of the variables of interest for the target zones are calculated from the sum of values in the source zones by the areal proportion of the source zones in the intersecting areas (Goodchild and Lam 1980). This approach has the advantage of simplicity but it based on the assumption that the variables of interest (e.g. population) are evenly distributed across the source zones. A number of areal interpolation approaches seek to overcome the assumption of within zone homogeneity by including ancillary information about the distribution of the values of interest. These dasymetric approaches refine the distribution of the variable of interest to the target zones by incorporating additional data to provide a more realistic estimate of the actual distribution of the process under investigation. The incorporation of additional data relevant to the study breaks down the artificial spatial structure imposed by the changing political boundaries. Regression techniques have been used to estimate the values of interest in the target zones from classified satellite data, under an assumption of a positive relationship between population density and the amount of urban area development (Langford, et al., 1991; Langford and Unwin 1994; Mennis 2003). Similarly Eicher and Brewer (2001) used urban land use data to refine their analysis of six socio-economic variables. Control zones have been used as intermediaries between source and target zoned in the interpolation process (Goodchild et al., 1993) where the set of control zones are assumed to have a homogeneous distribution of the variable of interest. The control zone values provide ancillary information to constrain the regression used to estimate the values in the target zones. Another approach, using ancillary data for the target zones to constrain the allocation of the values, is an iterative technique called the EM algorithm (Dempster et al. 1977). Flowerdew and Green (1994) discuss how the EM algorithm can be applied in a number of forms to a variety of different circumstances in order to allocate proportionally the variable of interest across the target zones.

A second, and sometimes overlapping, set of approaches to overcome the assumption within zone homogeneity when interpolating between source and target zones have been based on “pynophylactic interpolation” first proposed by Tobler (1979). This is a process that generates a smooth surface from polygon based data, whilst preserving the overall mass or volume of the data. The process starts similarly to areal weighting – dividing the aggregated data into smaller units, typically raster cells – but smoothes the

raster volumes iteratively with the weighted average of nearest neighbours. It transforms data from source to target zones and the volumes are distributed within the original source zone not to neighbouring ones. The volume is adjusted at each iteration to maintain the mass of the original polygon (volume preserving). That is, the data totals for the source zones are preserved during the transformation to target zones. Pycnophylactic interpolation computes a continuous surface (of target zones) from source zone data (often polygons) in 2 stages:

- 1) Source zones are converted to a regular grid with height values (e.g. the area of agricultural land use) assigned to the grid points;
- 2) The height values are increased or decreased individually to make the surface smooth whilst simultaneously enforcing volume preservation.

The step 2 is repeated until the remaining “roughness” – i.e. the deviation from the ideal smoothness – has reached a user-defined threshold or the maximum iteration is reached. The number of nearest neighbours used and the number of iterations determines the overall level of smoothing and is a subjective process (Hay et al., 2005). Pycnophylactic interpolation is an elegant solution to the problem of generating a continuous surface from discontinuous data, although it does assume that no sharp boundaries exist in the distribution of the data (Hay et al., 2005), which may not always be the case. However pycnophylactic interpolation has been used in many applications including modelling the distribution of roe deer (Dray et al., 2003), creating population surfaces to model malaria transmission (Hay et al., 2005) and to improve the visualisation and modelling of population data for regional planning (Rase, 2001). Kyriakidis (2004) proposed a geostatistical framework for supporting the differences between source areal data and target point predictions that yields pycnophylactic (ie mass-preserving) predictions. Tobler presents a verbal overview of the approach which is illustrated in Figure 1.

[Insert Figure 1 about here]

3. Method

The objective was to develop a database describing general classes of non-overlapping land use. Data integration was through a series of discrete stages. The approach taken was to identify areas of non-agricultural land use within each 1km², to iteratively sum the various land use areas over groups of 1km² squares and then to compare those totals with the reported totals at the higher level geographies (parishes, output areas, etc.). The data used to identify the areas of non-agricultural land use was as follows:

1. Strategi data (Ordnance Survey);
2. Land Cover Map of Great Britain (Centre for Ecology and Hydrology);
3. Common land data (Rural Surveys Research Unit, University of Wales).

The agricultural census data has historically been subject to a number of changes in reporting units as described above. Data from different censuses were modelled to produce time series land use estimates. In the text below the various agricultural censuses reporting units are referred to as “source zones” (e.g. SOAs, Parishes, etc) and their higher aggregations as “higher level source zones” (e.g. Parish Groups, Parish Districts). In this work the values for the target zones were constrained to be equal to the values of

their component source zones: the pycnophylactic criterion. The source data values of the area of arable crops were treated as the volume to be smoothed and preserved.

Stage 1. Identification of non-agricultural areas

A raster dataset of non-agricultural areas was produced by combining the vector maps of urban land, woodland (Strategi), common land, rivers, canals, railways, and roads (Strategi) with a 1km² grid. All of the linear features used were buffered by an appropriate amount prior to integration so as to fully represent the land area that they occupy. The size of the buffers were determined in consultation with the relevant organisations responsible for the administration of the UK communication infrastructure - The Highways Agency, Department for the Environment Transport and the Regions, Railtrack, and British Waterways. Each buffered vector data set was intersected with the 1km² grids. For each cell in the grid this identified the areas of each type of (non-agricultural) land use (Figure 2).

[Insert Figure 2 about here]

Stage 2. Comparison with the PAC at the district level

The estimates of the non-agricultural land use at 1km² as generated in Stage 1 were grouped and compared with agricultural census data at higher level source zones (e.g. Parish Groups, Parish Districts, etc) as in Figure 3. There was an under-estimation of non-agricultural land area in rural areas due to incomplete vector data for woodlands and small hamlets. Urban areas did not suffer this problem.

[Insert Figure 3 about here]

Stage 3. Improvement of the non-agricultural land estimates

The non-agricultural land uses were allocated into 5 land classes for comparison with aggregated Land Cover Map of Great Britain (LCMGB) classes: Sea, Inland Water, Woodland, Urban and Rough Grass. A full description of LCMGB can be found in Fuller et al., (2004). The LCMGB aggregation table for agricultural and non-agricultural land use classes is shown in Table 1.

[Insert Table 1 about here]

In the comparison with LCMGB, the initial land use estimates from Stage 1 for each 1km square were treated as being incomplete or as a lower bound. The amounts of each of the 5 land classes was increased by *adding in* equivalent values from the aggregated LCMGB classes until they matched the total amount of non-agricultural land reported by the JAC at the source zone level. That is, the additional amounts of land use were incrementally distributed around the 1km squares until the total amount matched that reported by the JAC for the higher level source zones. Each non-agricultural land cover type was assessed in turn in the strict order noted above, as this was the order in which most confidence may be placed in the correct classification of pixels in LCMGB. Where the LCMGB had more land in any of these classes than the estimate, and the total non-

agricultural land was still under-reported, the non-agricultural land use value was increased accordingly.

If at the end of this constraint, the non-agricultural land total was still under-estimated, the additional land required was added to the Rough Grass category, as this is the category shown to have most confusion associated with it in the LCMGB classification. The sub-stages in this were:

- a) Taking each 1km² cell contained within the higher level source zones in turn, the difference with the LCMGB data in the amount of non-agricultural land under the current land type (Sea, Inland Water, Woodland, Urban, Rough Grass) was calculated. The differences for each 1km² in the source zone (e.g. the Parish) were summed to give a total amount of the current land type that was to be used to increase the initial non-agricultural estimate. A check was made to ensure that any increase would not take the total area of a 1km² above the maximum 100 hectares.
- b) If the area calculated in the preceding sub-stage was greater than the target area of non-agricultural land of the source zone, then a scalar was derived to prevent the revised estimates for agricultural and non-agricultural land exceeding the physical area of the higher level source zones (e.g. Parish District).
- c) For each 1km² cell contained within the higher level source zones in turn, the differences with the LCMGB non-agricultural land types were calculated again. Using the scalar derived above along with the difference calculated, the estimated value was increased and the target amount of non-agricultural land being distributed across the whole higher level source zones was reduced accordingly.
- d) Within the current higher level source zones an assessment was made to determine whether enough non-agricultural land had been added. If more land was still required then a fraction was calculated to express the remaining target land area as a proportion of the current estimate of agricultural land in the higher level source zones.
- e) For each 1km² in the higher level source zones, the area of non-agricultural land was calculated. If this area was less than the physical area of the cell (100 ha), the area under Rough Grass was increased by the proportion calculated above.

Stage 4. Improvements in the agricultural land estimates

After Stages 1-3 the total crop areas (Arable) were within $\pm 5\%$ of those reported by the census totals. In order to ensure the total areas for each crop matched those of the census totals completely the crop areas were adjusted in the following fashion:

- Initial crop areas were estimated from the proportions reported in the source zone data;
- The total area under each crop was then summarised by higher level source zones, from the original census data and the land use database;
- The differences between these estimates at higher level source zones were used to smooth the crop proportions at a source zone level in such a way that the total area of each crop in the target zones matched the census at higher level source zones (e.g. Parish Districts).

Hence, the proportions of crops continued to be held on a per-source zone (e.g. district, SOA) basis. These were smoothed to ensure that the totals, when calculated at a higher

level source zones from the disaggregated 1km² data, matched those of the higher level source zones (e.g. parish district) census totals. That is, the crop totals or “volumes” were preserved in this smoothing process.

4. Results

A national dataset of land use was created at 1km² for England and Wales based on aggregated JAC data for 2000, reported at higher geographies. In order to illustrate and assess the estimated land use distributions, the results were compared with alternative land information for Kent, a county in the south East of England. The objective of this work was to provide local measures of Arable and Grass land uses, as these underpin much finer redistribution of other information in the agricultural census (specific crops areas, livestock numbers, on-farm employment etc). A comparison was made with the Kent Wildlife Habitat Survey (Wain, 2002), conducted in 2000 in a partnership of Kent County Council, English Nature, the Environment Agency, the Defence Estate, the Kent Wildlife Trust and the district councils within Kent. It used a combination of aerial photography interpretation and field survey. The habitat classification used the Integrated Habitat Classification System developed by the Somerset Environmental Records Centre². This includes habitat definitions of over 400 categories of habitat. For the purposes of comparison the habitat data was aggregated to 7 general land classes and 1km² squares.

The data for Kent is noteworthy, because it is comprehensive, represents a snapshot in time and was done as a single survey. For most counties in England, their habitat data is often incomplete, may have been done by different teams at different times and may incorporate different habitat classifications. The Habitat data were recoded into 7 land classes (Sea, Inland Water, Woodland, Urban, Rough Grass and Arable) as shown in Table 2 and for each 1km square, the constituent proportions of the 7 classes were calculated. Rough was used as a catch-all class where there was uncertainty about land class allocation.

[Insert Table 2 about here]

The land use estimated from the agricultural census data at 1km² was compared with the habitat generated data at a range of local aggregation scales: 1km², 4km², 9km², 16km² and 25km² square blocks. It was not expected that the estimates would be particularly reliable at the 1km² level due to the nature of the processing described in Section 3. However, it was expected that as the 1km² were combined into blocks, a truer, more representative picture of local agricultural practice would emerge. In order to test this Root Mean Square Error (RMSE) was calculated using the formula suggested by Gregory (2000) for situations where the values for the target zone populations (in this case proportions of Arable land use) vary greatly:

² <http://www.somerc.com/cdrom.php>

$$RMSE = \left[\frac{1}{n} \sum_n \left(\frac{X_t - \hat{X}_t}{X_t} \right)^2 \right]^{1/2} \quad (\text{Equation 1})$$

where n is the number of target zones X_t is the estimated land use value in target zone t and $\hat{X}_t$ is the corresponding land use from the habitat data.

Figure 4 shows the scatterplots of modelled (y-axis) against habitat derived (x-axis) values for Arable and Grass land uses for the different spatial aggregations. The expected improvement in the RMSE values and R^2 (in Table 3) are illustrated as the size of the spatial aggregation increases by the plots. The errors converge at 16km² as the improvement by increasing the aggregated area to 25km² reduces.

Figure 5 shows the 1km² maps of arable and grassland distributions in Kent. Arable and grass distribution and areas are reasonably accurate when compared with thematically and spatially finer scaled information as well as summary unit total. The differences are for both land uses are randomly distributed

[Insert Table 3 about here]

[Insert Figure 4 about here]

[Insert Figure 5 about here]

5. Discussion

The objective of this work was to generate local estimates of agricultural land use at 1km² from data aggregated to coarser units for both summary reporting and confidentiality purposes. The approach presented in this work uses techniques for overcoming a number of traditional data problems relating to census data. It sought to avoid the assumption of within-unit homogeneity by allowing internal source zone (i.e. within original reporting unit) variation to emerge in the process of allocating data allocation to lower level (target zone) geographies. First, other data were incorporated into the analysis in a manner similar to what Langford and Higgs (2006) call ‘binary’ dasymetric approaches, where ancillary data on land use were used to partition source zones of interest into excluded (in this case non-agricultural land use) and included areas similar to Langford & Unwin (1994) and Mennis (2003). Second the initial estimates were compared with data at the higher level source zone. Third, these initial estimates of non-agricultural and agricultural areas were iteratively refined by comparing their totals at a higher level source zones (e.g. Parish Group figures). The total areas of the different types of non-agricultural land for the target zones were improved by adding in values from the LCMGB and applying a scalar. Fourth, the target zone values for different crops were iteratively smoothed by comparing their totals at higher source zones. The difference between the sum of the target zones within each higher level source zone were

used to adjust source zone values in such a way that the total area of each crop in the target zones matched the census at higher level source zones (e.g. Parish Districts).

Overall the approach is similar to the iterative refinement method as contained in the maximum likelihood estimation described in Flowerdew and Green (1994) and the EM algorithm in Flowerdew and Green (1991) where target zone ancillary variables relating to different properties (in this case aggregated land use types) were used to improve interpolation accuracy. The approach of Flowerdew and Green (1994) assumed that the ancillary data for the target zone provides information about the within target zone distribution of the variable of interest. The EM algorithm (Dempster et al., 1997) determines the relative allocation of the variable of interest. It computes an estimation of the missing data for target zones by their likelihood from the observed data or the latent variables as if they were observed (the Expectation stage or E step). Then the model is fitted using a maximum likelihood function which computes maximum likelihood estimates of the variable: it maximizes the expected likelihood found on the E step over the entire data set including the data that were estimated in the E step. The values computed in the M step are then used to begin another E step, and the process is repeated until the algorithm converges. Gregory and Ell (2005) provide an excellent illustration of how this process can be implemented for geographical data.

There is a wider issue of whether like is being compared with like when the data being compared come from very different sources. In this case the agricultural data were generated by farmers completing census forms and the habitat data from the manual interpretation of aerial photographs. However, the comparison between the two datasets was at aggregated levels of thematic detail not at detailed levels. For instance the data for 30 or so crop types from the agricultural census were aggregated into one land use class. The habitat data has a conservation focus and is concerned with natural and semi-natural habitats. It has 2 agricultural classes 'Arable and Horticulture' 'Hops'. The former is a catch-all class and the latter was essentially to demarcate a land use that is culturally significant locally. Therefore, whilst the scale and the epistemology of data collection are different, the process of aggregation into a super-class and analysis at that super-class level obviates the issue of semantic heterogeneity. Comber et al., (2004) include a more detailed discussion on this topic and the issues associated with combining data of different ontologies at a refined level of thematic detail.

The most powerful validation of this output of this approach and the data originates from the number of applications that have defensibly incorporated the agricultural land use data generated in this work. These include a number of agri-environmental models relating to nitrate losses, spatial modelling of waste and faecal borne pathogens and phosphorous and sediment mobilisation modelling. A sample of this work is reviewed below. Lord and Anthony (2000) report on the development of a national agri-environmental databases and nitrate modelling system. This provides local measures of soil nitrate concentrations and leached nitrate based on spatially distributed land use (cropping patterns), soils, livestock distributions and to develop a decision support system for policy makers and farmers (MAGPIE). Silgram et al. (2001) reviewed nitrogen loss models including one based on the land use data described in this paper,

NEAP-N. It uses the spatially re-allocated agricultural census information to relate land use and crop type to dominant soil type and hydrologically effective rainfall, calculating soil nitrate as a function of the balance of nitrogen inputs to and nitrogen offtake by the preceding crop. They state that that “nitrogen losses modelled using this approach have been found to compare favourably with independent field measurements using porous pots in fields within the Nitrate Sensitive Areas scheme (Lord, 1992; Anthony et al., 1996) and with stream nitrate fluxes measured in several contrasting catchments” (Silgram et al., 2001, p191). A further area of application using this data is pathogen modelling. CSAT (Coliform Source Apportionment Tool) evaluates the impact on shellfisheries of runoff from land receiving organic waste (Food Standards Agency, 2004). The land use data described in this work provided the land use parameters needed for the catchment level models. The land use data identified the numbers, types and distribution of livestock around different agricultural activities and was used as input to calculate faecal coliform loadings to agricultural land, fresh and stored manures. The results of this work found a good correspondence between the monthly profiles of predicted concentrations of faecal coliforms at the tidal limit in two selected study areas and the actual contamination found in shellfish.

6. Conclusions

The results of this analysis show that land use data reported at coarse levels of geographic detail for reasons of confidentiality (such as Parishes) can be reliably disaggregated into finer detailed target zones if the interpolation is appropriately constrained. In this case non-agricultural land use information from a number of sources was iteratively added into the solution, constrained by the other data on the actual amount of agricultural land use. Agricultural land use data for the target zones were iteratively adjusted and smoothed, constrained by the totals at higher level source zones (such Parish Districts).

The results of the comparison with external data emphasise that at the individual 1km² should not be considered on the own, but as a part of a larger aggregation: as 1km² blocks are combined the variations resulting from redistributing land use values start to reduce. The results indicate that the modelled land use data is accurate for the purposes of local land use estimations at 3km by 3km blocks or greater.

The approach of combining dasymetric approaches that are iteratively refined with a range of external ancillary data with iterative smoothing and volume preservation (pseudophylactic) approaches provides a method for overcoming differences in areal reporting units. It allows local measures of different types of arable land use to be generated from aggregated agricultural census data. Validation of the data is made easier and more reliable when aggregated classes are used as they obviate issues relating to semantic and ontological dissimilarity between the data being compared. The land use data generated in this way for England and Wales has been subject to a number of implicit quality assurance assessments:

- It has been successfully used for a range of modelling activities, all of which have produced validated results and whose outputs have been scrutinised by policy makers, academics and farmers’;

- The models have been heavily cited in literature by other scientist;
- The results of these models have been accepted by stakeholders and clients;
- The results of the models have validated against measured data.

For much research using agricultural census data analysis there is a tension between managing the differences in spatial units and the desire to preserve farm confidentiality. The approach presented in this paper seeks to address these two fundamental problems of using aggregated census data by combining dasymetric techniques with pycnophylactic techniques.

Acknowledgements

The authors wish to acknowledge the funding provided by Defra under projects NT2202, NT2203 and NT2206 for the development and implementation of this methodology. The authors would also like to thank Kent County Council for permission to use their habitat data and the anonymous reviewers of for their suggestions and comments relating to earlier drafts of this work.

References

- Allanson, P.F. and Moxey, A. 1996. Agricultural Land Use Change in England and Wales. *Journal of Environmental Planning and Management*, 39(2) 243-254
- Allanson, P.F., Savage, D. and White, B. 1992. Areal interpolation of parish Agricultural Census data, in: M.C. Whitby (Ed) *Land Use Change: The Causes and Consequences (ITE Symposium no. 27)* (London, HMSO).
- Anthony S., Quinn P. and Lord E.I. 1996. Catchment scale modeling of nitrate leaching. *Aspects of Applied Biology*, 46: 23–32.
- Chamberlain D.E., Fuller R.J., Bunce R.G.H., Duckworth J.C., Shrubbs M., 2000. Changes in the abundance of farmland birds in relation to the timing of agricultural intensification in England and Wales. *Journal of Applied Ecology* 37 (5): 771-788
- Comber, A., Fisher, P. and Wadsworth, R., 2004. Integrating land cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science*, 18(7): 691-708.
- Cook, H.F. and Norman, C., 1996. Targeting agri-environmental policy: An analysis relating to the use of Geographical Information Systems. *Land Use Policy*, 13(3): 217-228
- Coppock J. T., 1955, The relationship between farm and parish boundaries. *Geographical Studies*, 1, 12–25.
- Coppock, J. T., 1960, The parish as a geographical statistical unit. *Tijdschrift voor Economische en Sociale Geografie*, 51(12), 317–326.
- Coppock, J. T., 1964, *An Agricultural Atlas of England and Wales* (London: Faber and Faber).
- Dempster, A., Laird, N. and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39: 1–38.
- Dray, S., Pettorelli, N. and Chessel, N., 2003. Multivariate Analysis of Incomplete Mapped Data. *Transactions in GIS*, 7(3): 411-422.

- Eicher, C. L., and Brewer, C.A., 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28 (2):125–38.
- Evans, N. J. (1996) Evaluating recent changes to the publication of the UK agricultural census data. *Geography* 81: 225-234.
- Evans, N. and Ilbery, B. 1993. The pluriactivity, part-time farming and farm diversification debate. *Environment and Planning A* 25, 945-959.
- Evans, N.J., and Morris, C., 1997. Towards a Geography of Agri-Environmental Policies in England and Wales. *Geoforum*, 28(2): 189-204.
- Flowerdew, R., and M. Green. 1994. Areal interpolation and types of data. In *Spatial analysis and GIS*, A. S. Fotheringham and P. Rogerson (eds.), pp. 121–45, Taylor and Francis, London.
- Flowerdew, R., and Green M. (1991). *Data integration: Statistical methods for transferring data between zonal systems*. In *Handling Geographical Information*, I. Masser, and M. Blakemore (eds.), pp.38–54, Longman, New York.
- Food Standards Agency, 2004. *Evaluation of the impact on shellfisheries of runoff from land receiving organic waste (B05006/B05007)*.
<http://www.foodstandards.gov.uk/science/research/researchinfo/foodborneillness/organicwasteresearch/b17programme/b17projlist/b05006and07/b05006and07r>. Available 19 September 2007.
- Fuller, R.M., G.B. Groom, and Jones, A.R., 1994. The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data, *Photogrammetric Engineering & Remote Sensing*, 60:553–562.
- Geddes A., Gimona A., and Elston D.A., 2003. Estimating local variations in land use statistics. *International Journal of Geographical Information Science* 17 (4): 299-319.
- Goodchild, M.F. and Lam, N.S.-N., 1980, Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing*, 1: 297–312.
- Goodchild, M.F., Anselin, L. and Deichmann, U., 1993. A framework for the areal interpolation of socio-economic data, *Environment and Planning A*, 25, pp. 383±397.
- Gregory, I.N., 2000. An evaluation of the accuracy of the areal interpolation of data for the analysis of long-term change in England and Wales. In *Proceedings of the 5th International Conference on GeoComputation*, University of Greenwich, UK
UK, GeoComputation
- Gregory, I.N. and Ell, P.S. (2005). Breaking the boundaries: geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society A*, 168(2): 419-437.
- Haines-Young, R. and Watkins, C., 1996. The rural data infrastructure. *International Journal of Geographical Information Systems*, 10(1): 21-46.
- Hay, S.I., Noor, A.M., Nelson, A. and Tatem, A.J., 2005. The accuracy of human population maps for public health application. *Tropical Medicine and International Health*, 10(10): 1073–1086
- Hooper, A., 1989. Assessment of the Suitability of CPH Numbers for Mapping the General Distribution of Set-Aside Holdings. Unpublished ADAS Report, Leeds.
<http://reports.eea.eu.int/tech40add/en/tech40add.pdf> (available 14 April, 2005).

- Kyriakidis P.C., 2004. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis* 36 (3): 259-289.
- Langford, M., Maguire, D. and Unwin, D.J., 1991, The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In *Handling Geographical Information*, I. Masser, and M. Blakemore (eds.), pp.55–77, Longman, New York).
- Langford, M. and Higgs, G., 2006. Measuring Potential Access to Primary Healthcare Services: The Influence of Alternative Spatial Representations of Population. *The Professional Geographer*, 58(3): 294-306.
- Langford, M. and Unwin, D.J., 1994. Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal* 31:21–26.
- Leff B, Ramankutty N and Foley JA, 2004. Geographic distribution of major crops across the world. *Global Biogeochemical Cycles* 18 (1): Art. No. GB1009
- Lord, E.I., Anthony, S.G., Goodlass, G., 2002. Agricultural nitrogen balance and water quality in the UK, *Soil Use and Management*, 18 (4): 363-369.
- Lord E.I., 1992. Modelling of nitrate leaching: Nitrate Sensitive Areas. *Aspects of Applied Biology*, 30: 19–28
- Lord, E. and Anthony, S. 2000. MAGPIE: A modelling framework for evaluating nitrate losses at national and catchment scales. *Soil Use and Management*, 16: 167-174.
- Mennis, J., 2003. Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1): 31-42.
- Openshaw, S., 1984. *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography, vol. 38. Norwich: Geobooks.
- Rase, W-D., 2004. Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, 3(2): 199-213
- Silgram, M., Waring, R., Anthony, S. and Webb, J. 2001. Intercomparison of national and IPCC methods for estimating nitrogen loss from agricultural land. *Nutr. Cycl. AgroEcosystems*, 60, 189-195.
- Smith, K.A., Anthony, S.G., Henderson, D., Jackson, D.R., 2003. Critical drainage and nitrate leaching losses from manures applied to freely draining soils in Great Britain, *Soil Use and Management*, 19(4): 312-320.
- Sullivan A., Ternan .JL. and Williams A.G., 2004. Land use change and hydrological response in the Camel catchment. *Cornwall Applied Geography* 24 (2): 119-137
- Tobler, W. R., 1979. Smooth Pycnophylactic Interpolation for Geographical Regions, *Journal of the American Statistical Association* 74: 519–36
- Vaughan N., Lucas E.A., Harris S., and White P.C.L., 2003. Habitat associations of European hares *Lepus europaeus* in England and Wales: implications for farmland management. *Journal of Applied Ecology* 40(1): 163-175
- Wain, P.S., 2002. Use of GIS in Kent Wildlife Habitat Survey. *ESRI News*, 9, 10-11.
- Webb, J., Ryan, M., Anthony, S. G., Brewer, A., Laws, J., Aller, M. F. and Misselbrook, T. H., 2006. Cost-effective means of reducing ammonia emissions from UK agriculture using the NARSES model. *Atmospheric Environment* 40(37): 7222-7233.
- Whatmore, S. 1995. From farming to agribusiness: the global agro-food system. In *Geographies of Global Change*, eds R. Johnston, P. Taylor and M. Watts. Blackwell, Oxford.

List of tables and Figures

Table 1. Aggregation of CEG LCMGB Target classes, a full description of each class can be found at http://192.171.153.203/data/lcm/LCMclassInfo_p3.shtm (October 7th 2007)

Table 2. The aggregation of the Phase 1 habitats into 7 land use classes

Table 3. The differences between the observed (from aggregated habitat data) and predicted (from the modelled agricultural census data) for Kent for Arable and Grass land uses

Figure 1. The process of pycnophylactic interpolation from

http://www.csiss.org/streaming_video/csiss/tobler_pycno.htm

Figure 2. The generation of initial non-agricultural land use estimates at 1km² from buffered vector data of urban land, woodland, common land, rivers, canals, railways, and roads.

Figure 3. Comparison of the estimates of the non-agricultural land use at 1km² generated in Stage 1 with agricultural census data at the level of higher level source zones.

Figure 4. Comparison of 1km² land percentages (y-axis) with aggregated habitats for Kent (x-axis), with RMSE

Figure 5 shows the 1km² maps of arable and grassland distributions in Kent, classified into quintiles (lightest is < 20%, darkest >80%), differences in intervals of $\pm 10\%$ (white) $\pm 30\%$ (grey) and $\pm 85\%$ (black)

<u>Land Use Class</u>	<u>Aggregated LCMGB Classes</u>
Sea	1,3,4
Inland Water	2
Woodland	15,16,23
Urban	20,21
Rough Grass	5,8,9,10,11,12,13,14,17,19,22,24,25
Arable	18
Improved / Managed Grass	6,7

Table 1. Aggregation of CEG LCMGB Target classes, a full description of each class can be found at http://192.171.153.203/data/lcm/LCMclassInfo_p3.shtm (October 7th 2007)

Land use	Habitats
Arable	Arable and horticulture, Hops
Grass	Improved grassland, Other lowland hay meadows
Rough	Other continuous bracken, Fen, marsh and swamp, Swamp, Reedbeds, Other swamp vegetation, Marginal and inundation vegetation, Inundation vegetation, Fens [lowland], Other fens, transition mires, springs and flushes, Other purple moor grass and rush pastures, Lowland dry acid grassland, Lowland calcareous grassland, Neutral grassland, Other neutral grassland, Dwarf-shrub heath, European dry heaths, Northern Atlantic wet heaths with <i>Erica tetralix</i> , Other wet heaths, Chalk reefs, Inshore sediment, Littoral rock, Littoral chalk, Other littoral rock, Littoral sediment, Saltmarsh, <i>Salicornia</i> [Glasswort] & other annuals colonising mud & sand, <i>Spartina</i> swards [Cord grass] [Spartinion], Atlantic salt meadows [Glauco-Puccinellietalia], Continental salt meadows [Puccinellietalia distantis], Other saltmarsh, Mudflats and sandflats not covered by sea water at low tide, Other littoral sediment, Inland rock, Natural rock exposure features, Inland cliff, Artificial rock exposures and waste, Quarry, Spoil heap, Refuse tip, Supralittoral rock, Maritime cliff & slopes, Boulders and rock above high tide mark, Other supralittoral rock, Supralittoral sediment, Embryonic shifting dunes, Shifting dunes along the shoreline with Marram, Fixed dunes with herbaceous vegetation [grey dunes], Decalcified fixed dunes, Dunes with Sea buckthorn [<i>Hippophae rhamnoides</i>], Humid dune slacks, Shingle above high tide mark, Coastal vegetated shingle, Perennial vegetation of stony banks, Other shingle above high tide mark, Strandline vegetation, Other supralittoral sediment
Urban	Built-up areas and gardens
Water	Rivers and streams, Standing open water and canals
Wood	Intensively managed orchards, Mixed woodland, Scrub woodland, Broadleaved woodland, Lowland beech and yew woodland, <i>Asperulo-Fagetum</i> beech forests, <i>Taxus baccata</i> woods, Other lowland beech and yew woodland, Wet woodland, Residual alluvial forests, Other wet woodland, Other coniferous woodland
Sea	Sea

Table 2. The aggregation of the Phase 1 habitats into 7 land use classes

	R²		RMSE	
	Arable	Grass	Arable	Grass
1km²	0.5373	0.2952	1.743458	22.03354
4km²	0.6816	0.4444	0.668428	8.404142
9km²	0.7576	0.5481	0.308588	1.410581
16km²	0.8236	0.667	0.248299	0.855839
25km²	0.8588	0.6806	0.221276	0.695869

Table 3. The differences between the observed (from aggregated habitat data) and predicted (from the modelled agricultural census data) for Kent for Arable and Grass land uses

Zero Iterations

Five Iterations

Twenty-five Iterations

Figure 1. The process of pycnophylactic interpolation from
http://www.csiss.org/streaming_video/csiss/tobler_pycno.htm

Figure 2. The generation of initial non-agricultural land use estimates at 1km² from buffered vector data of urban land, woodland, common land, rivers, canals, railways, and roads.

Figure 3. Comparison of the estimates of the non-agricultural land use at 1km² generated in Stage 1 with agricultural census data at the level of higher level source zones.

Figure 4. Scatterplots of the predicted land uses estimates (y-axis) against the observed aggregated habitats for Kent (x-axis), at different spatial aggregations for Arable and Grass land uses.

Figure 5 shows the 1km² maps of arable and grassland distributions in Kent, classified into quintiles (lightest is < 20%, darkest >80%), differences in intervals of ±10% (white) ±30% (grey) and ±85% (black)