



**AgEcon** SEARCH  
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*



## **A Statistical Approach for Spatial Disaggregation of Crop Production in the EU**

Markus Kempen<sup>1</sup>, Thomas Heckeley<sup>1</sup>, Wolfgang Britz<sup>1</sup>, Adrian Leip<sup>2</sup>, Renate Koeble<sup>2</sup>

<sup>1</sup> University of Bonn, Institute for Agricultural Policy, Market Research, and Economic Sociology, Bonn, Germany

<sup>2</sup> Joint Research Centre, Climate Change Unit, Ispra, Italia

**Contribution appeared in Arfini, F. (Ed.) (2005) “Modelling Agricultural Policies: State of the Art and New Challenges”, proceedings of the 89<sup>th</sup> EAAE Seminar, pp. 810 - 830**

February 2-5, 2005

Parma, Italy



**UNIVERSITA' DEGLI  
STUDI DI PARMA**

*Copyright 2005 by Markus Kempen, Thomas Heckeley, Wolfgang Britz, Adrian Leip and Renate Koeble. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.*

# A Statistical Approach for Spatial Disaggregation of Crop Production in the EU<sup>1</sup>

Markus Kempen, Thomas Heckelei, Wolfgang Britz

University of Bonn, Institute for Agricultural Policy, Market Research, and Economic Sociology,  
Bonn, Germany

Adrian Leip, Renate Koeble

Joint Research Centre, Climate Change Unit, Ispra, Italia

## *Abstract*

In this paper we describe a procedure for disaggregating agricultural land use choices at NUTS 2 level to about 18.000 homogeneous spatial units completely covering the usable agricultural area of the EU. The disaggregation procedure uses 40.000 sampling points and aggregate data from administrative regions and requires two steps: First, we employ crop specific, spatial binary choice models to regress land use decisions on local natural conditions (soil, relief, climate) based on the sample information. Results allow predicting crop shares in each spatial unit. Second, consistency with data from administrative regions is achieved by maximising the posterior density of crop shares subject to aggregating equations using the forecast distributions as prior information. Comparison with actual crop shares shows the validity of the procedure.

## *1. Introduction*

Not at least due to the so-called multi-functional model of European agriculture, there is growing interest in modelling environmental effects of the agricultural sector in the EU. In many cases, results beyond rather crude passive indicators can only be obtained linking biophysical models to economic models for policy impact analysis. An important methodological problem in this context is “bridging” the scales: whereas most bio-physical models work on field scale, comprehensive EU-wide economic models generally work on large administrative regions.

---

<sup>1</sup> The research is supported by the European Commission in the context of the CAPRI-DynaSpat project (Project number 501981).

Within these administrative boundaries the natural conditions of soil, relief and climate usually differ in such a manner, that the assumption of identical cropping pattern, yields or input use cannot be maintained. Simulations with bio-physical models thus require breaking down results from the economic models into a smaller regional scale. This paper proposes a statistical approach combining a logit model with a Bayesian highest posterior density estimator to break down production data of 30 crops in about 150 European administrative regions for EU15 (NUTS 2) to 18.000, so called, Homogeneous Spatial Mapping Units (HSMUs).

The approach is based on two steps. The first step regresses cropping decisions in each HSMU on geographic factors (soil, climate etc.), using results of the Land Use / Cover Area Frame Statistical Survey (LUCAS) providing observations on agricultural crops at approximately 40.000 sampling points all over the EU territory. Spatial statistical techniques are used to allow for spatial heterogeneity of the coefficients using a locally weighted logit model. In the second step of the disaggregation procedure, simulated or given data for the administrative Nuts II regions are broken down to HSMU level by Bayesian methods. Two possible ways to introduce prior information from the logit regression step are discussed: (1) using means and variances of the predicted shares in each HSMU, or (2) using the estimated coefficients and their covariance matrix in the Nuts II region. In the first case, we search for shares at HSMU level consistent with Nuts II results maximizing the posterior density of the predicted shares. The second approach selects the most probable set of regression coefficients producing data consistent shares over all HSMU maximizing the posterior density of the coefficients.

The basic approach – estimating prior information and achieving consistency between scales afterwards is in line with previously suggested disaggregation procedures (Howitt and Reynaud, 2003). While different estimation procedures are motivated by data availability, the proposed method contributes to the literature in the following respects: (1) Lower level units are defined by homogeneous production conditions rather than administrative boundaries; (2) Functional relationships between location factors and land use are identified explicitly using spatial statistical techniques. This allows to discern prior information on crop shares even under scarce data information for some lower level units; (3) The applied Bayesian method fully and transparently accounts for the available prior information – prior distributions – when searching for consistency between the scales.

The paper is organised as follows. Chapter 2 describes the database and the definition of the HSMUs. Chapter 3 explains the two step disaggregation procedure in detail. Chapter 4 presents and discusses selected results.

## *2. Database and Definition of Homogeneous Spatial Mapping Units (HSMUs)*

The description of the database is divided in three main parts: (1) sources and definitions of the natural location factors, (2) the construction of HSMUs, and (3) a characterisation of the agricultural production data in administrative regions.

*Maps of Natural Location Factors*

The relative competitiveness of an agricultural crop at a certain location is determined by natural factors, technology, and market conditions. While market conditions and the generally available technology are assumed to be rather invariant within an administrative region, differences in natural conditions will lead to heterogeneity regarding the optimal crop mix between different locations inside the Nuts II region. Therefore, this study concentrates on natural location factors.

**Table 1.** Relevant Maps of Natural Conditions

| <b>Factors</b>                   | <b>Indicators</b>               |
|----------------------------------|---------------------------------|
| <i>Soil quality</i> <sup>2</sup> | Sand content                    |
|                                  | Clay content                    |
|                                  | Organic Carbon Content          |
| <i>Relief</i> <sup>3</sup>       | Slope                           |
|                                  | Elevation                       |
| <i>Climate</i> <sup>4</sup>      | Annual Rainfall                 |
|                                  | Length of the vegetation period |
|                                  | Cumulative temperature sum      |
|                                  | Bio-geographical region         |

According to plant production literature (e.g. Heyland 1994), yield potentials of agricultural crops are mostly affected by soil quality, relief and climate conditions. Small scale information on location factors stems from different sources and was prepared with the help of geographical information systems (GIS). The bio-geographical region characterizes the ecological system at a certain location (alpine, alpine boreal, alpine pyrenees, anatolian, arctic, atlantic, atlantic north, black sea, boreal, continental, continental south, macaronesia, mediterranean, pan-nonian, stepic) and modifies a scheme proposed in Roekaerts (2002).

<sup>2</sup> Hiederer R., Jones B. and Montanarella L. (2003): European Soil Raster Maps (1km by 1km) for Topsoil Organic Carbon Content, Texture, Depth to Rock, Soil Structure, Packing Density, Base Saturation, Cation exchange. Developed under the EC-JRC-Action 2132: Monitoring the state of European soils (*MOSES*).

<sup>3</sup> EuroLandscape/Agri-Environment Catchment Characterisation and Modelling Activity, Land Management Unit, Institute for Environment and Sustainability, EC-Joint Research Centre. 250 Meter DEM, compiled on the basis of data acquired from data providers and national mapping agencies over Europe for internal use.

<sup>4</sup> Interpolated meteorological data. Source, JRC/MARS Data Base – European Commission – JRC; Van der Goot E. & Orlandi S. (1997/2003).

*CORINE Land Cover Map (CLC)*

The general distinction of different land cover classes is based on the CORINE land cover map (European Topic Centre on Terrestrial Environment, 2000) describing land cover (and partly land use) according to a nomenclature of 44 classes, based on the visual interpretation of satellite images and ancillary data (aerial photographs, topographic maps etc.).

The CORINE classification system distinguishes 11 agricultural classes (Non-irrigated arable land, permanently irrigated land, rice fields, vine yards, fruit and berry plantations, olive groves, annual crops associated with permanent crops, complex cultivation, pasture, marginal areas and forestry). Some of the classes as “Rice fields”, “Olive groves”, “Vineyard”, “Pasture” or “Arable Land” clearly indicate a special agricultural use. A minimum of 25 ha of homogeneous land cover is defined to build one CORINE mapping unit. That definition of the minimum mapping unit leads to two effects. Firstly, “pure” classes such as “Arable land” may in reality comprise small parcels of other land cover classes as well if these are smaller than 25 ha. Secondly, so-called heterogeneous agricultural areas as e.g. “Land principally occupied by agriculture with significant areas of natural vegetation (marginal area)” comprise no pre-dominant land use >25 ha and give only limited information about the type of agricultural use. The 25 ha limit results from the mapping conventions and the interpretative limits set by the spatial and spectral resolution of the satellite images.

In this study we assume that only the agricultural classes are suitable for farming. The reader is reminded that agricultural classes may comprise small parcels of non-agricultural uses, as agricultural use may be found in non-agricultural classes.

*Motivation and Construction of Homogeneous Spatial Mapping Units (HMSU)*

The aim of building HMSU is the definition of areas inside an administrative region where approximate homogeneity according location factors may be assumed. The HMSU serve then as simulation units for the bio-physical models and are constructed by overlaying different maps (land cover, soil map, climatic factors etc.). In order to allow for a manageable number of HSMUs, the most important factors must be selected, and continuous parameters must be grouped in classes. The CORINE land cover map was used here in combination with three further main factors relating to soil (sand content in 4 classes), relief (slope in 5 classes) and climate (“biogeographical region”). Each HSMU has identical values for these four items, other parameters (such as clay content) may differ inside the HSMU. Weighted averages are defined for the parameters shown in Table 1 above for each HSMU using GIS techniques.

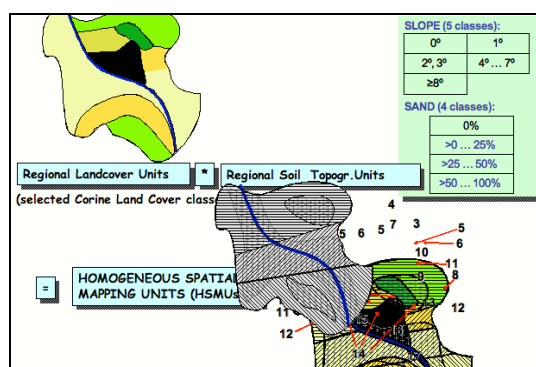


Figure 1. Homogeneous Spatial Mapping Units

The HMSU approach was deemed superior to a grid layout, especially as factors determining optimal cropping patterns may be identical across very large regions (say Northern Finland) so that grid units would be “wasted”, whereas in other regions especially such which high relief changes, the grid units may already comprise huge differences in natural conditions. Further on, the units can be defined so that they do not cross administrative borders, and grid data may be redefined based on the HSMUs.

#### *Agricultural Production Data*

The agricultural sector model CAPRI uses complete and consistent statistics for Nuts II regions, based on EUROSTAT. The second step of the disaggregation procedure adjusts land use choices in the HSMUs until the summed areas over the HSMU match the observed statistics for Nuts II regions. Another important database in the LUCAS survey, which allows direct assignment of land use choice and natural conditions.

#### *Land Use / Land Cover Frame Statistical Survey (LUCAS)*

In opposite to mapping approaches, area frame surveys based on a common statistical sampling method gather land cover and land use data (EUROSTAT, 2000) at specific sample points, only, and extrapolate from these to the entire area under investigation (European Commission, 2003a). LUCAS covers the territory of all EU Member States and all kinds of land uses, and is based on a two-stage sampling design: at the first level, so-called Primary Sampling Units (PSUs) are defined as cells of a regular grid with a size of  $18 \times 18$  km, while the Secondary Sampling Units (SSUs) are 10 points regularly distributed (in a rectangular of  $1500 \times 600$  m side length) around the centre of each PSU (Figure 2) resulting in approximately 10.000 PSUs for the whole EU (European Commission, 2003).

Due to possible measurement errors regarding the geo-references in the CORINE maps (Gallego 2002), about 30% of the LUCAS points closer than 100 m to the border of a CORINE class were not considered in here. The 38 agricultural classes found in LUCAS (36 crop land, 2 permanent grassland classes) were re-grouped according to the crops found in CAPRI as shown in Table 2. All other classes (artificial areas, woodland, water, etc.) are aggregated in a residual classed termed “OTHER”.

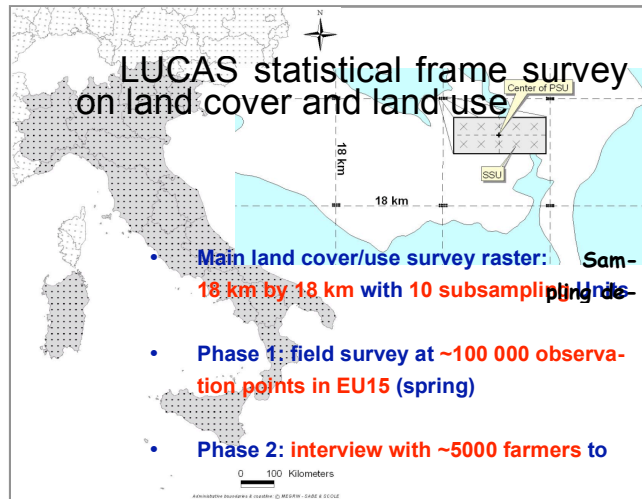


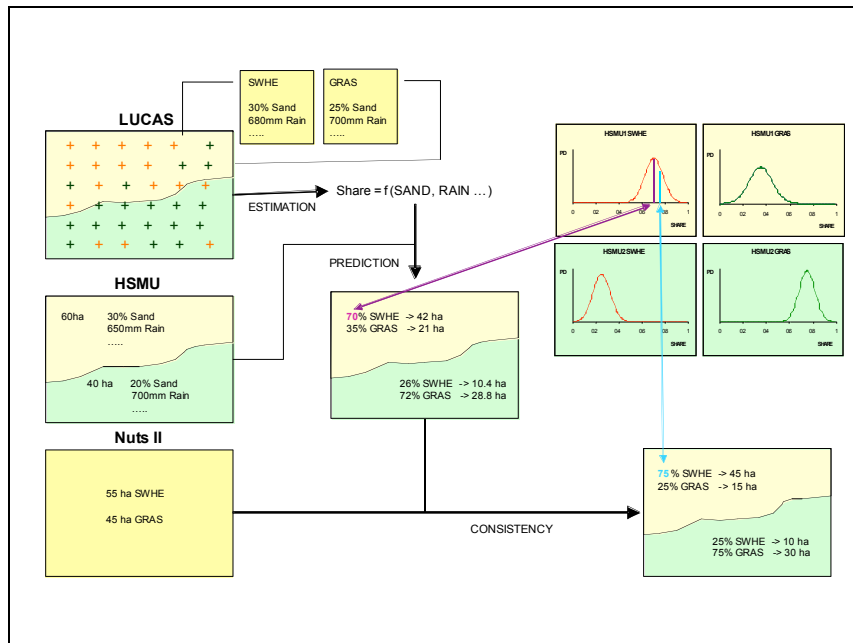
Figure 2. Design of the LUCAS survey



**Table 2.** Considered crops

| Group                  | Crops                                                                                                                          |
|------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Cereals                | Soft wheat (SWHE), Durum wheat (DWHE), Reye (REYM), Barley (BARI), Oats (OATS), Corn (MAIZ), Rice (PARI), Other cereals (OCER) |
| Oilseed and Pulses     | Rapeseed (RAPE), Sunflowers (SUNF), Soybean (SOYA), Pulses (PULS)                                                              |
| Industrial crops       | Potatoes (POTA), Sugar beet (SUGB), Texture (TEXT), Other industrial crops (OIND)                                              |
| Labour intensive crops | Tomatoes (TOMA), Other vegetable (OVEG), Flowers (FLOW), Tobacco (TOBA)                                                        |
| Permanent crops        | Olive grows (OLIV), Citrus fruits (CITR), Other fruits (FRUI), Nurseries (NURS), Vine (TWIN)                                   |
| Fodder Production      | Grass (GRAS), Food from arable land (OFAR), Fodder root crops (ROOF)                                                           |
| Fallow Land            | Set aside or fallow land (FALL)                                                                                                |
| Other Land Cover       | Other crops (OCRO), non agricultural land cover (OTHER)                                                                        |

### 3. The Disaggregation Procedure



**Figure 3.** Disaggregation Procedure

Before describing the crucial steps in detail the general approach of the disaggregation procedure is illustrated in Figure 3. Suppose there is a Nuts II region divided in only two HSMUs each comprising two crops – grassland (GRAS) and soft wheat (SWHE). Combining the LUCAS survey with digital maps provides us with several observations of crops grown at a defined point with a set of natural conditions. Using an adequate estimation model we can regress the probabilities of finding a crop at a certain location on the natural conditions. As this probability can be interpreted as the share of the crop in a homogeneous region, applying these estimated coefficients to the average natural conditions in a certain HSMU yields normally distributed predictions of crop shares for this HSMU under corresponding assumptions on the stochastic processes governing crop choice. These a priori information on cropping shares are generally not consistent with the “known” cropping area in the Nuts II region. The “best” set of data-consistent shares given the prior information is identified by a Bayesian *highest posterior density* approach.

### 3.1 Locally Weighted Binomial Logit Estimation

Generally, shares for each crop  $\hat{Y}_C$  are regressed on the following explanatory variables describing natural conditions:

- Sand content (SAND)
- Clay content (CLAY)
- Organic carbon content (OCTO)
- Slope (SLOP)
- Elevation (ELEV)
- Rainfall (RAIN)
- Length of vegetation period (VEGP)
- Sum of temperature in vegetation period (TSUM)

The regressions were estimated independently for each crop  $c$  in each CORINE class  $clc$ :

$$\hat{Y}_{c,clc} = f(SAND, CLAY, OCTO, SLOP, ELEV, RAIN, VEGP, TSUM) \quad (1)$$

The arguments for using specific coefficients for each CORINE class are as follows. Assume grass land parcels are found in the LUCAS survey in the “non-irrigated land” CORINE class. We would assume that slope has a positive effect on the probability to find grass. In the “pasture” class of CORINE, we would eventually find the opposite effect: with increasing slope, grass land could be replaced by forest (see also chapter 4.1). For convenience the indices  $c$  and  $clc$  are omitted in the following.

The LUCAS survey reports one point in time observations and hence does not deliver cropping shares (or rotations), but requires a binary choice model. Both logit and probit models (see e.g. Green 2000) were originally tested, with the logit approach giving slightly better

results. The likelihood function of finding crop  $c$  at a specific LUCAS point  $i$  for the binomial logit model is defined as:

$$\Lambda(\beta' \mathbf{x}_i) = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} \quad (2)$$

$$\log L = \sum_{i=1}^n [y_i \log \Lambda(\beta' \mathbf{x}_i) + (1 - y_i) \log (1 - \Lambda(\beta' \mathbf{x}_i))]$$

where  $\mathbf{Y}$  is a dummy vector indicating whether a certain crop was observed at a location  $i$  ( $y_i=1$ ),  $\mathbf{x}_i$  is the design matrix containing data on natural conditions and  $\Lambda(\beta' \mathbf{x}_i)$  is the probability that a specific crop is grown at location  $i$ .

Applying the estimated  $\hat{\beta}$  to the average natural conditions in a HSMU ( $\mathbf{x}_h$ ) give us a prior estimate for the share of a specific crop in a certain HSMU:

$$\hat{Y} = \Lambda(\hat{\beta}' \mathbf{x}_h) = \frac{e^{\hat{\beta}' \mathbf{x}_h}}{1 + e^{\hat{\beta}' \mathbf{x}_h}} \quad (3)$$

### *Binomial versus Multinomial Regression*

The approach discussed above examines the crops independently from each other and thus neglects the information that crops compete for the available land, with two possible effects. Firstly, the error terms for the different crops are probably correlated, and secondly, the individual estimated shares don't add up to unity. The multinomial probit model would be ideal as it allows for an unrestricted variance covariance structure of the error terms and satisfies the additivity condition, but is computationally infeasible for 30 crops and 10.000 points. The assumption of an identity matrix for the variance covariance matrix underlying the multinomial logit model was deemed as too inflexible (Nelson et al. 2004), albeit it is easier to solve. The way out might be a nested logit model, a possible expansion in further analysis.

However, both problems were not deemed crucial for the application at hand. Given the large number of observations, the possible gain of taking correlations between the error terms across crops into account is most probably small. Furthermore, the violation of the adding up condition for the shares is explicitly accommodated in the second step of the disaggregation procedure, where the estimated shares serve as prior information, only.

### *Local versus Global Regressions*

The assumption of European wide invariant relationships between the share of each crop and a limited number of location factors describing natural conditions may be problematic if other omitted explanatory factors are not randomly distributed in space, but "clustered". Suppose, for example, two HSMUs with identical natural conditions, the first one close to a sugar refinery, and the second one far way from the next sugar plant. The share of sugar beets in the first unit will be probably much higher, an effect not linked to the natural conditions. Clearly, omit-

ted variables as the effect of sugar refineries could lead to seriously biased parameter estimates. Adding more explanatory variables would certainly help, but it is simply impossible to collect information on all probably relevant factors (market points, transport infrastructure, environmental legislation, etc.). Instead, spatial econometric techniques are applied to overcome the problem of omitted variables that are correlated over space.

The basic idea behind Locally Weighted Regression, which was proposed by Cleveland and Devlin (1988), is to produce site specific coefficient estimates using Weighted Least Squares to give nearby observation more influence than those far away. Further on, the estimation for any specific site is limited to a number of observations within a certain bandwidth around the site. Locally Weighted Regression are mostly found combined with Least Squares estimators, but application to Maximum Likelihood Estimation as needed in the case of discrete dependent variables are described as well (Anselin et al. 2004).

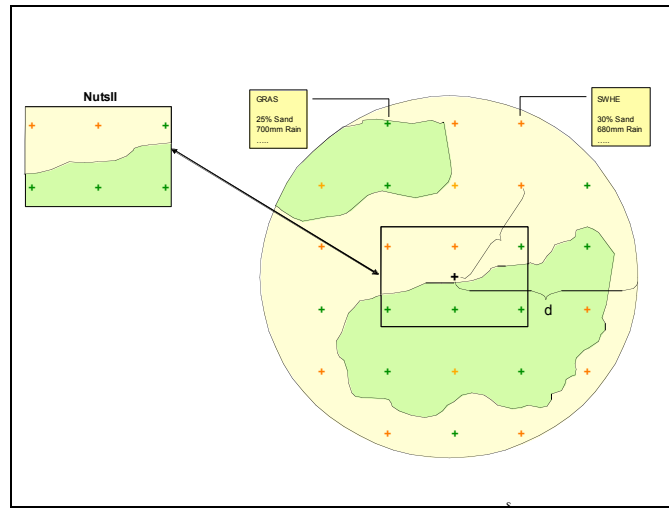


Figure 4. Scheme of locally weighted maximum estimation

The weight given to any observation  $i$  in constructing the estimate for site  $j$  is given by  $\omega_{ij}$ . The tri-cube is a commonly used weighting function:

$$\omega_{ij} = \left[ 1 - \left( \frac{\delta_{ij}}{d_j} \right)^3 \right]^3 I(\delta_{ij} < d_j) \quad (4)$$

Where  $\delta_{ij}$  is the distance between site  $i$  and observation  $j$ .  $d_j$  is the bandwidth and  $I(\cdot)$  is an indicator function that equals one when the condition is true. The effect of any one location in space on near points thus falls depending on the distance and becomes zero once the distance exceeds the bandwidth. There are other common weighting schemes like the Gaussian

function or several Kernel weighting functions (see: Anselin et al. 2004 or Fotheringham et al. 2002). But it has been shown that opting for a proper bandwidth is more significant than choosing a certain spatial weighting function.

When there is no prior justification for applying a particular bandwidth, an appropriate bandwidth can be found by the minimising either the cross-validation score (CV), the Akaike Information Criterion (AIC) or the Schwartz Criterion (SC). The AIC and the SC are offered by most software packages. The CV is calculated as:

$$CV = \sum_{i=1}^n (y_i - \hat{y}_{i \neq i})^2 \tag{5}$$

where  $n$  is the number of data points and the prediction for the  $i$ th data point  $\hat{y}_{i \neq i}$  is obtained with the weight for that observation set to zero. Each of the criteria can be minimised by a golden section search (see Press et al. 1989). In our study all criteria led to similar results. We opted to minimise the Schwartz Criterion, because according to Boots et al. (2002) it seems to have better large sample properties.

In typical applications, sites and observations would be identical. In our context, that would require estimates per crop and CORINE class for each LUCAS point, which is computational impossible. Instead, the NUTS II regions were chosen as sites. When estimating for a particular NUTS II region, all LUCAS point inside that NUTS II region received uniform unity weight, and points in neighbouring NUTS II regions weights equal or smaller unity according to (4). That still leads to a large number of possible estimations: 150 Nuts II regions times 10 agricultural CORINE classes times 30 crops, but fortunately, many of the combinations do not comprise any observations. Weighting each likelihood contribution with  $\omega_{ij}$  gives (Fotheringham et al. 2002):

$$\log L = \sum_{i=1}^n \omega_{ij} \left[ y_i \log \Lambda(\beta_j' x_i) + (1 - y_i) (1 - \Lambda(\beta_j' x_i)) \right] \tag{6}$$

*Robust Covariance Matrix Estimation*

Calculating accurate variance and covariance matrixes for the coefficients is essential to ensure proper a priori density functions for the Bayesian methods in the second step of the disaggregation procedure. Given the non-linear character of the estimations, the variance-covariance matrices offered by the statistical packages are not analytically calculated, but are instead numerically approximated which proved to be not suitable. Quite small predicted mean values in combination incredibly high variances led to shaky final results. It became therefore necessary to calculate the asymptotic covariance matrix analytically (see White (1982)) as:

$$Cov[\hat{\beta}] = \hat{\mathbf{H}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{H}}^{-1} \tag{7}$$

where for the weighted logit model the elements of Hessian  $\mathbf{H}$  and the Brendt, Hall, Hall and Hausman matrix  $\mathbf{B}$  are given by (Green 2000):

$$\mathbf{H} = \frac{\partial^2 \text{Log}L}{\partial \beta \partial \beta'} = - \sum_i \omega_i \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i' \quad (8)$$

$$\mathbf{B} = \sum_i \omega_i (y_i - \Lambda_i)^2 \mathbf{x}_i \mathbf{x}_i' \quad (9)$$

As insignificant parameter estimates might influence the efficient calculation of a robust covariance matrix although they do not influence the forecasted value, insignificant variables were removed from the estimations. The variance of  $\hat{\mathbf{Y}}$  builds upon the calculated covariance matrix  $\mathbf{V}_\gamma = \text{Cov}[\hat{\beta}]$ .

$$\mathbf{V}_Y = \text{Var}[\hat{\mathbf{Y}}] = [\Lambda_i (1 - \Lambda_i)]^2 [I + (1 - 2\Lambda_i) \beta \mathbf{x}'] \mathbf{V}_\gamma [I + (1 - 2\Lambda_i) \mathbf{x} \beta'] \quad (10)$$

Using specific  $\mathbf{x}_{HSMU}$  yields variances of the predicted land use share in each HSMU (Green 2000).

### 3.2 Data-consistent Disaggregation

The second step of the disaggregation procedure identifies crop shares in each HSMU using the prior information on the estimated crop shares from the first estimation step under two data constraints: Firstly, adding up the areas per crop in each HSMUs must recover the cropping areas  $CA$  for that crop at NUTS II level. Secondly, the posterior shares in each HSMU must add to unity, including all non-agricultural land use from the LUCAS survey aggregated to the category ‘‘OTHER’’. In opposite to the first step this requires simultaneous accounting for all crops  $c$  in all relevant HSMUs  $h$ . The notation is therefore extended, e.g. from  $Y$  to  $Y_{c,h}$ .

The crop areas in each HSMU are defined by multiplying the posterior shares  $Y_{c,h}^{con}$  with the entire area  $A_h$  thus

$$\sum_{h \in N_2} Y_{c,h}^{con} A_h = CA_{c,N_2} \quad (11)$$

and the adding up to unity

$$\sum_c Y_{c,h}^{con} = 1 \quad (12)$$

must be imposed.

As the predicted unrestricted shares will typically violate the constraints, a penalty function is necessary to define the optimal deviations from the predictions. Generalized Maximum Entropy (GME) techniques (Golan, Judge and Miller 1996) have often been used for this type of data balancing exercises in recent times. Here, however, a *Bayesian highest posterior density (HPD) estimator* is applied allowing for a direct and transparent formulation of prior information and considerably reducing the computational complexity compared to the GME approach (Heckelei et al. 2005). The prior information is expressed either as normal densities of predicted shares, with mean vector  $\hat{\mathbf{Y}}_{c,h}$  and variance  $\mathbf{V}_{\mathbf{Y}_{c,h}}$ , or as prior multivariate normal densities of parameters  $\hat{\mathbf{a}}_{c,clc}^{con}$  with mean vector  $\hat{\mathbf{a}}_{c,clc}$  and covariance matrix  $\mathbf{V}_{\hat{\mathbf{a}}_{c,clc}}$  .. After taking logs, the prior density function for the consistent shares  $Y_{c,HSMU}^{con}$  is:

$$-\sum_c \sum_h \left[ \log(\sqrt{2\pi} \mathbf{V}_{\mathbf{Y}_{c,h}}) + \frac{(\mathbf{Y}_{c,h}^{con} - \hat{\mathbf{Y}}_{c,h})^2}{2\mathbf{V}_{\mathbf{Y}_{c,h}}} \right] \quad (13)$$

In order to define the HDP solution for the coefficients, an additional equation has to be imposed. We assume that the coefficients should be the same in all HSMU belonging to a certain Corine class *clc*:

$$\mathbf{Y}_{c,h}^{con} = \Lambda(\beta_{c,clc}^{con} x_h) = \frac{e^{\beta_{c,clc}^{con} x_h}}{1 + e^{\beta_{c,clc}^{con} x_h}} \quad h \in clc \quad (14)$$

The multivariate prior density function equals:

$$-\frac{1}{2} \sum_c \sum_{clc} \left[ n \log(2\pi) + \log|\mathbf{V}_{\beta_{c,clc}}| + (\beta_{c,clc}^{con} - \hat{\beta}_{c,clc})^T \mathbf{V}_{\beta_{c,clc}}^{-1} (\beta_{c,clc}^{con} - \hat{\beta}_{c,clc}) \right] \quad (15)$$

Where *n* is the number of coefficients.

#### 4. Results

The framework of the binary choice models does not provide us with such a meaningful measure of fit like the R<sup>2</sup> value in a standard linear least squares regression. Other measures of fit are proposed but checking them for about 15.000 estimated equations is a time-consuming process and they do not tell us much about the quality of the final result. Therefore we concentrate on interpreting estimated land use shares in the light of prior knowledge on crop cultivation and on comparing aggregated estimation results with observed statistics at NUTS III level.

The following figures illustrate the estimated land use patterns under changing conditions in selected HSMUs of a French Nuts II region (FR71 – Rhone-Alpes). Figure 5 and Figure 6 show different land use choices on “Non Irrigated Land” depending on different slope, respectively sand classes. The cultivation of maize (MAIZ, shown in red) comes with a high erosion potential and is therefore likely to decrease significantly in steep areas. Equally, with increasing slope the share of high yield crops (soft wheat - SWHE, rape seed - RAPE) is expected to drop whereas more robust cereals (BARL - barley, OCER – other cereals) should increase. The main land use alternatives on high slopes according to the estimation are grass land (GRAS) and non-agricultural use (OTHER), e.g. forestry, both with low erosion risk. The estimated land use changes thus seem to be in line with common expectations.

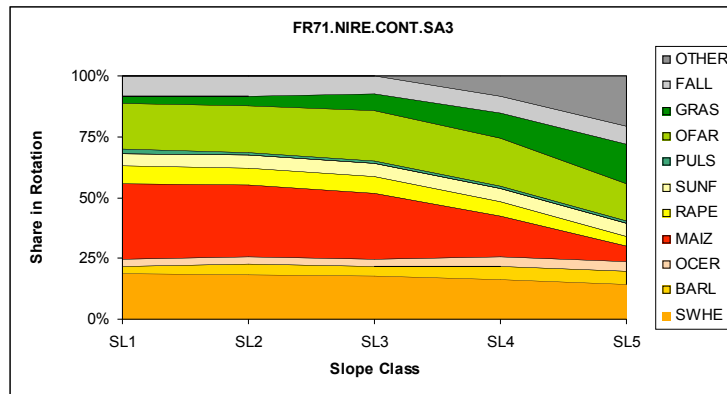


Figure 5. Changing Land Use on Arable Land (increasing slope)

The changes in the land use choice regarding changing sand content are less pronounced but comprehensible since the poor storage capacity of sandy soils can be compensated by rainfall and fertilizer applications in this region.

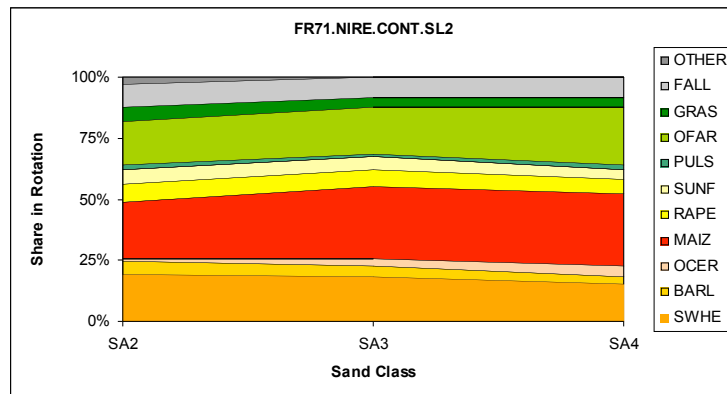
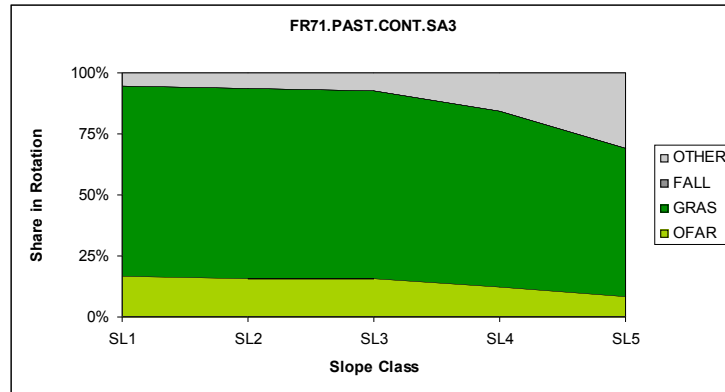


Figure 6. Changing Land Use on Arable Land (increasing sand content)



Figure 7 shows land use choices in the CORINE class PASTURE. As expected, this “pure” class consists mostly of grassland. With increasing slope the grassland is gradually replaced by non agricultural land cover as forestry. The response in the share of grassland to slope increases is fundamentally different from the reaction observed on arable land. This shows the necessity of estimating distinct models for different CORINE classes.



**Figure 7.** Changing Land Use in PASTURES (increasing slope)

#### 4.1 Comparing Estimated Results with Observations

For some European regions, land use statistics at a lower administrative level, called Nuts III, are available from the farm structure survey (FSS; EUROSTAT, 2002). This information is used as out-of-sample observation to validate the results of the disaggregation algorithm, which predicts cropping shares for the HSMUs consistent to NUTS II. Those predicted shares at HSMU are then aggregated to NUTS III level and compared to the observed data.

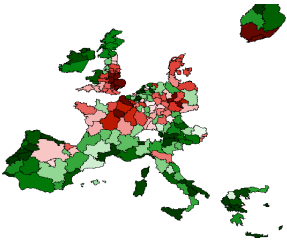
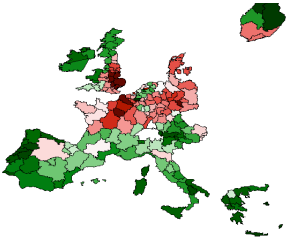
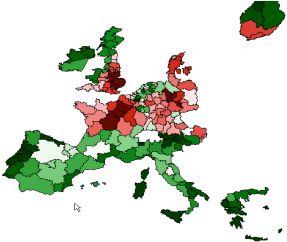
Another quick check is to run the second step of the disaggregation procedure without recovering the Nuts II statistics (without constraint (11)) and compare the “quasi multinomial” forecasts with the Nuts II statistics. This comparison allows a first evaluation whether adjusting “shares” or “coefficients” is the preferable method. Table 1 shows the R2 calculated from the deviation between the forecasted cropping area and the Nuts II statistics for common crops. Two different methods of predicting the cropping areas are compared with the cropping areas in a Nuts II region derived from the share of LUCAS points with a certain crop within this region. The fairly high R2 for the areas calculated from the LUCAS survey show that the point observations reflect the observed cropping areas quite well. At least the “adjusting

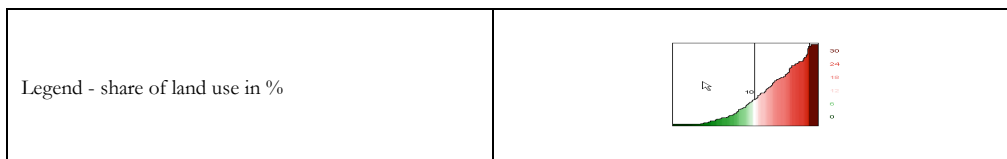
shares” prediction comes up in most cases with higher R2. This shows that the disaggregation procedure is more precise than a simple statistical method.<sup>5</sup>

**Table 3.** R<sup>2</sup> values - forecasted cropping area compared to observed production level in Nuts II regions of EU-15

| Crops | R <sup>2</sup> Values                                  |                                           |                                       |
|-------|--------------------------------------------------------|-------------------------------------------|---------------------------------------|
|       | Staitic                                                | Locly weighted Logit Estimation           |                                       |
|       | LUCAS survey                                           | "Quasi Multinomial"                       |                                       |
|       | shares calculated from observations in Nuts II regions | consistency by adjusting predicted shares | consistency by adjusting coefficients |
| SWHE  | 0.92                                                   | <b>0.94</b>                               | 0.71                                  |
| DWHE  | <b>0.87</b>                                            | 0.80                                      | 0.65                                  |
| RYEM  | 0.37                                                   | 0.73                                      | 0.69                                  |
| BARL  | 0.83                                                   | <b>0.90</b>                               | 0.09                                  |
| OATS  | 0.48                                                   | <b>0.78</b>                               | -0.04                                 |
| OCER  | 0.06                                                   | <b>0.28</b>                               | 0.12                                  |
| CERE  | 0.89                                                   | <b>0.94</b>                               | 0.63                                  |
| MAIZ  | <b>0.91</b>                                            | 0.79                                      | 0.54                                  |
| PARI  | 0.91                                                   | <b>0.93</b>                               | 0.93                                  |
| SUNF  | 0.84                                                   | <b>0.85</b>                               | 0.15                                  |
| PULS  | 0.62                                                   | <b>0.68</b>                               | -0.10                                 |
| RAPE  | <b>0.88</b>                                            | 0.71                                      | -3.50                                 |
| POTA  | <b>0.56</b>                                            | <b>0.55</b>                               | -2.70                                 |
| SUGB  | <b>0.84</b>                                            | 0.62                                      | 0.24                                  |
| TOMA  | <b>0.52</b>                                            | 0.50                                      | 0.24                                  |
| OVEG  | <b>0.30</b>                                            | 0.22                                      | -0.51                                 |
| CITR  | <b>0.97</b>                                            | <b>0.96</b>                               | 0.81                                  |
| FRUI  | 0.57                                                   | <b>0.64</b>                               | 0.41                                  |
| OLV   | <b>0.97</b>                                            | <b>0.95</b>                               | 0.80                                  |
| TWIN  | <b>0.91</b>                                            | <b>0.93</b>                               | 0.96                                  |
| TEXT  | <b>0.71</b>                                            | 0.62                                      | 0.48                                  |
| OFAR  | 0.23                                                   | <b>0.44</b>                               | 0.21                                  |
| GRAS  | 0.87                                                   | <b>0.93</b>                               | 0.65                                  |
| GRAL  | 0.88                                                   | <b>0.96</b>                               | 0.86                                  |
| FALL  | 0.61                                                   | 0.61                                      | -0.28                                 |

<sup>5</sup> Besides this the simple statistical method could not be applied to HSMU level, since there a sparse observations within a HSMU (in average 4 sampling points per HSMU, often even none).

|                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| <p>derived from <b>LUCAS</b> observation:</p> <p><math>\frac{\text{observation SWHE}}{\text{total observations}}</math> (in each NutsII region)</p> <ul style="list-style-type: none"> <li>□ provides realistic shares for NutsII regions (allows no breakdown to shares in HSMU)</li> <li>□ adequate database for locally weighted estimation</li> </ul>                                                                                   |    |
| <p>Weighted average over <b>forecasted</b> HSMU-shares</p> <ul style="list-style-type: none"> <li>• “Adding up to unity” is imposed</li> <li>• Data consistency <b>not</b> yet ensured</li> </ul> <ul style="list-style-type: none"> <li>□ allows for first evaluation of the prior information estimated using locally weighted maximum likelihood</li> <li>□ reduces in general deviation between LUCAS and NUTS II statistics</li> </ul> |   |
| <p>Complete and Consistent (Coco) Database from the <b>CAPRI</b> sector model</p>                                                                                                                                                                                                                                                                                                                                                           |  |



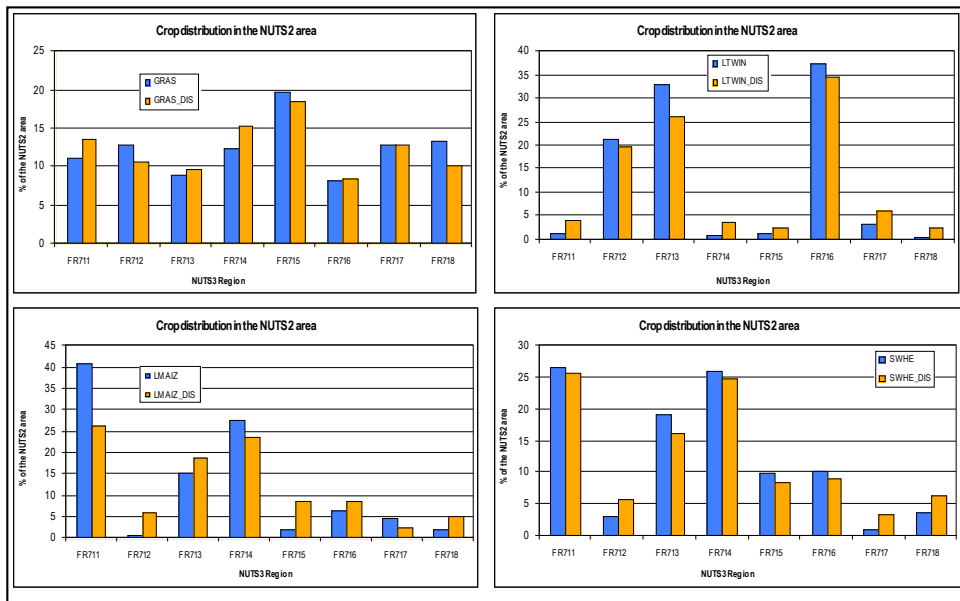
**Figure 8.** Cropping shares for soft wheat in NutsII regions derived from different methodologies and sources.

Compared to “adjusting shares”, the “adjusting coefficients” method suffers from a small feasible space, since the number of adaptable parameters is lower. The poor  $R^2$  for several crops indicates that this problem is present. Although further investigation is necessary we prefer achieving consistency by adjusting the predicted shares

The  $R^2$  of the estimated cropping areas resulting from the disaggregation procedure with full consistency imposed are generally better than those calculated only from the LUCAS ratios, proving the usefulness of the disaggregation procedure. Only for some crops (potatoes, sugar beet and vegetables) the  $R^2$  of the LUCAS areas is higher. The estimation may suffer from strong spatial heterogeneity unrelated to soil and climate but arising from heterogeneous market conditions.

The low  $R^2$  of “other fodder on arable land” (OFAR) indicates that the reclassification between LUCAS and CAPRI should be reviewed. OFAR is in fact mostly grass grown on arable land. The information contained in the LUCAS points might not perfectly discriminate between grass grown on pastures (GRAS) or on arable land (OFAR), as this classification is predominantly an administrative issue. Aggregating these two crops might help. Similar problems could be present among cereals, where distinguishing between types of grain is sometimes difficult (see e.g. Bavaria: significantly lower shares of soft wheat (SWHE) whereas shares of other cereals (OCER) are relatively high).

The validation with out-of-sample data (EUROSTAT, 2002) at NUTS III level (see Figure 9) shows that the land use shares predicted with the complete disaggregation procedure follow the actual distribution quite well. However, we can also see, that the disaggregation procedure is not able to capture the full variation between the NUTSIII regions at this point.



**Figure 9.** Comparison of estimated and observed shares in NUTS III region for different crops (Region Rhone-Alpes )

## 5. Conclusions

This paper introduced a procedure of disaggregating cropping shares at NUTS II level to cropping shares in homogeneous spatial mapping units (HSMU) within the NUTS II regions in order to provide appropriate input data to lower scale bio-physical models calculating environmental indicators. The methodology involves two steps: (1) Estimation of cropping shares depending on location factors based on a spatial maximum likelihood estimator and using observations on cropping choices at a high resolution grid-level. (2) Application of a Bayesian methodology to ensure consistency of predicted cropping shares at HSMU-level with NUTS II statistics. Variations of this computationally intensive methodology were discussed and evaluated. The procedure was applied to data from the EU-15 territory. Selected results show that the methodology provided results superior to crop shares directly calculated from the grid-level sample in capturing observed aggregate shares. An out-of-sample validation with NUTS III data indicated the procedure's ability to represent the distribution of cropping shares within a NUTS II region.

There are several limitations of the approach to be addressed in further research: First, the methodology needs to be extended to allow for simulation of changes in cropping patterns at lower scale. Second, the approach tends to somewhat underestimate the variation of cropping shares between smaller scale regional units. Third, the set of explanatory location factors is lim-

ited by the available data and the implied functional form is rather restrictive. Forth, other variables such as yield and related input use are also relevant as data for biophysical models. To overcome some of these problems, the study will be extended to introduce the new LUCAS survey and additional location factors becoming available soon. In addition, non-parametric techniques or non-linear transformations of explanatory variables are envisaged to make the approach more general. Finally, the study will be supplemented by making yield estimates and input use spatially varying across NUTS2 regions. At a later stage, we envisage to use time series data at NUTS II level in combination with spatial disaggregation to allow for the simulation of land use and input changes at lower scale in response to changing economic conditions.

### References

- Anselin L., Florax R.J.G.M. and Rey S.J. (2004): *Advances in Spatial Econometrics*, Springer Verlag, Berlin.
- Boots B., Okabe A. and Thomas R. (2002): *Modelling Geographical Systems*, Kluwer Academic Publishers, Dordrecht.
- European Commission, (2003): The Lucas survey. European statisticians monitor territory. Theme 5: Agriculture and fisheries, Series Office for Official Publications of the European Communities, Luxembourg.
- European Topic Centre on Terrestrial Environment (2000): Corine land cover database (Version 12/2000).
- EUROSTAT (2000): Manual of Concepts on Land Cover and Land Use Information Systems. Theme 5: Agriculture and Fisheries: Methods and Nomenclatures, Series Office for official Publications of the European Communities, Luxembourg.
- EUROSTAT (2002): Spatial redistribution of statistical data from the Farm Structure Survey. Series G.I.M. Geographic Information Management.
- Fahrmeier L., Tuts G. (1994): *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer Verlag, New York.
- Fotheringham A.S., Brusdon C. and Charlton M. (2002): *Geographically Weighed Regression*, John Wiley & Sons, Chichester.
- Gallego J. (2002): *Fine scale profile of CORINE Land Cover classes with LUCAS data*, in European Commission (ed.): *Building Agro Environmental Indicators. Focussing on the European area frame survey LUCAS*, Vol. EUR Report 20521 EN.
- Golan A., Judge G. and Miller D. (1996): *Maximum Entropy Econometrics*, Chichester UK, Wiley.
- Greene W.H. (2000): *Econometric Analysis*, Macmillian Publishing Company, New York
- Heckelei T., Mittelhammer R.C. and Britz W. (2005): "A Bayesian Alternative to Generalized Cross Entropy Solutions to Underdetermined Models", Contributed paper presented at the 89th EAAE Symposium *Modelling agricultural policies: state of the art and new challenges*, February 3-5, Parma, Italy.
- Hiederer R. Jones B. and Montanarella L. (2003): European Soil Raster Maps (1km by 1km) for Topsoil Organic Carbon Content, Texture, Depth to Rock, Soil Structure,

- Packing Density, Base Saturation, Cation exchange. Developed under the EC-JRC-Action 2132: Monitoring the state of European soils (MOSES).
- Howitt R. and Reynaud A. (2003): "Spatial Disaggregation of Agricultural Production Data by Maximum Entropy", *European Review of Agricultural Economics*, 30(3): 359-387.
- Nelson G., De Pinto A., Harris V. and Stone S. (2004): "Land Use and Road Improvements: A Spatial Perspective", *International Regional Science Review*.
- Mulligan D.T. (2004): Regional modelling of nitrous oxide emissions from fertilised agricultural soils within Europe. Series PhD thesis, submitted to the University of Wales, Bangor.
- Roekaerts, M. (2002): The Biogeographical Regions Map of Europe, The data and document can be found on <http://dataservice.eea.eu.int/dataservice/>.
- Van der Goot E & Orlandi S (1997/2003) Technical description of interpolation and processing of meteorological data in CGMS. The documentation can be found under [http://mars.jrc.it/marsstat/Crop\\_Yield\\_Forecasting/cgms.htm](http://mars.jrc.it/marsstat/Crop_Yield_Forecasting/cgms.htm).
- White, H. (1982): Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 53: 1-16.