RECOVERING LOCALIZED INFORMATION ON AGRICULTURAL STRUCTURE UNDERLYING DATA CONFIDENTIALITY REGULATIONS - POTENTIALS OF DIFFERENT DATA AGGREGATION AND SEGREGATION TECHNIQUES

Alexander Gocht und Norbert Röder

Institut für Ländliche Räume
des Johann Heinrich von Thünen-Institut (vTI),
Bundesforschungsinstitut für Ländliche Räume, Wald und Fischerei

Bundesallee 50
D-38116 Braunschweig

Norbert.Roeder@vti.bund.de

2010

Vortrag anlässlich der 50. Jahrestagung der GEWISOLA
„Möglichkeiten und Grenzen der wissenschaftlichen Politikanalyse“
Braunschweig, 29.09. – 01.10.2010

RECOVERING LOCALIZED INFORMATION ON AGRICULTURAL STRUCTURE WHILE OBSERVING DATA CONFIDENTIALITY REGULATIONS - POTENTIALS OF DIFFERENT DATA AGGREGATION AND SEGREGATION TECHNIQUES

Summary

The modelling and information system RAUMIS is used for policy impact assessment to measure the impact of agriculture on the environment. The county level resolution often limits the analysis and a further disaggregation at the municipality level would reduce aggregation bias and improve the assessment. Although the necessary data exists in Germany, data protection rules (DPR) prohibit their direct use. With methods such as the Locally Weighted Averages (LWA), and with aggregation singling production activities into larger groups of activities, the data at the municipality level can be made publicly available. However, this reduces the information content and introduces an additional error. This paper's aim is to investigate how much information is necessary to satisfactorily estimate Germany-wide production activity levels at the municipality level and whether the data requirements are still in compliance with the DPR. We apply Highest Posterior Density (HPD) estimation, which is easily able to include sample information as prior. We tested different prior information content at the municipality level. However, the goodness of the developed estimation approach can only be evaluated having knowledge about the population. Because the real population is not known to us, we took advantage of the special situation in Bavaria and derived a pseudo population for that region. This is used to draw information conforming to DPR for our estimation and to evaluate the resulting estimates. We found that the proposed approach is capable of adequately estimating most activities without violating the DPR. These findings allow us to extend the approach towards the Germany-wide municipality coverage in RAUMIS.

Keywords: Highest Posterior Density estimator (HPD), RAUMIS, locally weighted average (LWA)

1 Introduction

Frequently the impact of agricultural activities on the environment can only be properly assessed if the underlying distribution is well covered. For instance, the likely impact of new pests like the western corn rootworm (*Diabrotica virgifera ssp. virgifera* LeConte), which has a high relevance in the debate on bT-maize, depends on the share of maize in the crop rotation. Especially if its share exceeds 50 %, the western corn rootworm can have a serious impact (CARRASCO et al., 2009). Analysing the cultivated areas shares for 2007 at the county (326 regions in Germany) level, the cultivation of maize in Germany is not affected at all (FDZ, 2010). However, if the analysis is done on a municipality level (over 9,000 regions in Germany) almost 13 % of the maize area is concerned. This illustrative example shows that the utilization of wider regional averages to model these specific situations can be misleading, as agricultural land use and its dynamics are also site dependent (e.g., OSTERBURG *et al.*, 2009, p. 40 ff.). The agricultural and environmental modelling and information system RAUMIS (HENRICHSMEYER *et al.*, 1996), a mathematical programming, modelling and information platform for Germany‘s agricultural sector used to analyse agricultural and agri-environmental policy instruments currently operates at the county resolution. As in other economic models such as CAPRI (BRITZ and WITZKE, 2008), the RAUMIS model simulates an aggregate over all farms for a region. To overcome the problem of too aggregated analysis, the underlying heterogeneity of the farming pattern has to be better covered. Different approaches exist in the literature to disaggregate regional models. One example is a specifically tailored component in the CAPRI model, disaggregating crop shares, stocking densities and fertilizer application rates from the about 250 administrative regions across Europe to clusters of 1x1 km grid cells (LEIP et al., 2008) based on Homogeneous Spatial Mapping Units (KEMPEN *et al.*, 2005). Another approach is the disaggregation of regional production levels in farming groups (GOCHT, 2010). Both approaches have drawbacks with respect to the RAUMIS requirements. In the case of farm groups the missing territorial representation does not allow spatially geo-referenced data, an important feature for regional models, to be added. If data is spatially disaggregated to clusters of grid cells, problems emerge from the fact that the borders of these clusters do not necessarily coincide with administrative boundaries. Therefore the option, elaborated in the paper, is to disaggregate the county data to municipality level using Agricultural Census data. However, the provision of data is limited by legal constraints. In particular, the problem is that many production activities at the municipality level underly DPR and are not reportable, because too few observations exist. Currently, DPR is ensured in a two step procedure. In the first step, aggregated data are censored if they are derived from less than three observations or if a single observation contributes more than 80% to the aggregate. In a second step additional aggregates are censored to ensure that data censored in step one could not be exactly retrieved by applying arithmetic on published data. This implies that the higher the resolution regarding both topographically and thematically the higher is the likelihood that is censored due to DPR.

To increase the number of observation per municipality we applied Locally Weighted Averages (LWA) (cf. ANSELIN et al., 2006, p. 24 ff) to blur the production activity levels at municipality with the neighbour regions and, in addition, we aggregated the blurred activity levels into activity groups.

The aim of the paper is to investigate how much information for aggregated activities (e.g., arable land, main forage area ...) is necessary to satisfactorily recover Germany-wide production activity levels at municipality level, under the condition these aggregated data are blurred using LWA. To recover the activity levels, the production activity group levels at the municipality level are treated as random variables comprising the errors introduced by LWA. By assuming properties of the error distribution of the prior data, it is possible to estimate the most probable activity levels at the municipality level while ensuring regional consistency at the county level. Traditional estimation methods, however, would fail due to problem’s underdetermined nature. To solve such an undefined system, prior information must be utilized in the form of the LWA activity aggregates.

In order to evaluate the estimation results it would be necessary to know the real data’s distribution at the municipality level. For Germany as a whole, we have no access to those real observations. To test the approach, we take advantage of the special situation in Bavaria. Here, the high number of farms per municipality limits the loss of information regarding activity levels at municipality level and allows the construction of a pseudo population.

The remainder of the paper is organized as follows. In Section 2 we introduce the estimation framework, firstly summarizing the data information applied as constraints in the model, afterwards the estimation approach is described and the procedure how the prior information is drawn from our pseudo population. A subsection provides an overview about the test statistics used to validate the estimation results with our pseudo population. Section 3 describes briefly some key characteristics of the used data. Section 4 presents results and the final section concludes.

2 Methods

In this section we discuss in some detail the layout of the estimator, starting with the data constraints and introducing the estimation method Highest Posterior Density (HPD). We present the inclusion of prior information in the estimator and explain how the LWA approach is used to derive this prior information for each municipality level subject to the DPR. Finally, we explain the different experiments and introduce the test statistics.

2.1 Estimator

The estimation approach wants to identify production activity levels (j) with $j = 1,2,..\mathrm{N}$ with $\mathrm{N} = 36$ for each Bavarian municipality (m) with $\mathrm{m} = 1,2,..\mathrm{M}$ and $\mathrm{M} = 2012$ using prior information in form of LWA for designated groups of activities under the data constraints, that the sum of the production activity levels equals the observed and given county level production activities,

(1) $$x^{o}_{j,c} = \sum_{m} x_{j,m} \;\; \forall\; m \in c$$

where (x) is the activity level to be estimated and (x^{o}_{c}) is the activity level observed in the pseudo data set at county (c). These constraints alone do not allow a *unique* solution to be found, as there are the M x N unknown vectors of cropping hectares and livestock herd sizes to be estimated, which by far exceed the number of linear equality constraints in (1). Therefore prior information has to be included in combination with a penalty function. For such an approach Generalized Maximum Entropy (Golan *et al.*, 1996) has been used frequently. We use, however, the HPD estimation, allowing for a direct and transparent formulation of prior information and considerably reducing the computational complexity of the model (HECKELEI *et al.*, 2005). This can be done considering the LWA aggregated activity groups (q) and the production activity levels (x) on the municipality level as realisation of a random variable (**z**). It is assumed that prior belief on the possible realisation of the true values (Ψ) exists and can be expressed in the form of a prior density function indicated by $p(\Psi)$. The prior density summarizes information collected from non-sample information and the likelihood function $L(z,\Psi)$ linking the true values with the outcomes of the data generating process. Based on (ZELLNER, 1971) the combination of the prior density and the likelihood function results in $h(\Psi \mid z) \propto p(\Psi)L(z \mid \Psi)$ where $\propto$ denotes proportionality and the object function (h) can be interpreted as the joint posterior density of the model parameters and is defined via the prior density $p(\Psi)$. $p(\Psi)$ is multiplied with the likelihood function $L(z \mid \Psi)$ assigning zero weights to values of Ψ that violate model constraints and positive constant weights to values of Ψ that are compatible with the data and the model relationships (HECKELEI *et al.* , 2008).

Hence, the value for Ψ that maximizes h is the Highest Posterior Density (HPD) estimate of Ψ $\max_{\Psi}\{h(\Psi \mid z) \propto p(\Psi)L(z \mid \Psi)\}$. The highest posterior density values (h) can be obtained when the arguments are found that maximize the prior beliefs $p(\Psi)$ subject to the likelihood function, this can be formulated as maximization problem in the following way $\max p(\Psi) \quad s.t. \quad g(\Psi, x^{o})$. We assume a multivariate normal density function for $p(\Psi)$ in the form of

(2) $$p(\Psi) = \frac{1}{(2\pi)^{n/2} \mid \mathrm{V} \mid^{1/2}} \exp\left[-\frac{1}{2}(\Psi - \Psi^{\mathrm{p}})'\mathrm{V}^{-1}(\Psi - \Psi^{\mathrm{p}}) \right].$$

As the value that maximizes the function also maximizes its natural logarithm, the logarithm of the objective function is taken. Cancelling those elements that are irrelevant for the maximization of the objective function, we derive the following minimisation problem

(3) $\min \; vec(\Psi - \Psi^{\mathrm{p}})' \times \mathrm{V}^{-1}\, vec(\Psi - \Psi^{\mathrm{p}})$

where (V) is the covariance matrix,(Ψ) are the true values of the data generating process (our estimates) and Ψ^{p} are the moments of the prior information. To obtain an applicable model (3) is added to the constraints in (1). The cropping acreages and livestock herd sizes and the aggregated activity group information are assumed to be distributed around the true, but unknown observations which are characterised by the above defined data constraints. We assume that the error terms around (x) and (q) are white noise with co-variance zero. This leads to the following estimator.

(4) $\min vec(x - x^{\mathrm{p}}, q - q^{\mathrm{p}})' \times \Sigma^{-1} vec(x - x^{\mathrm{p}}, q - q^{\mathrm{p}})$

subject to the equations in (1) and the production activity groups (q^{p}) are defined as

(5) $q^{\mathrm{p}}_{m,n} = \sum_{i_n} x^{*}_{i_n,m} \;\; \forall\, i \in j\,,$

where the index (i) is a subset of the activities of (n) different aggregates drawn from the municipality level pseudo population using different sampling methods (later called experiments) indicated by the asterisk. The estimation framework, combing the estimator and the data constraints, can be interpreted as the search for the production activity levels and activity groups which minimize the deviation between the prior information on levels (x^{p}) and the prior information on activity groups (q^{p}) for each individual municipality with respect to the consistency constraints at county level.

2.2 Prior Information and Locally Weighted Averages (LWA)

We now discuss how the prior information (x^{p}) and (q^{p}) are derived. Due to the DPR we have no information at individual production activity level for Germany[1]. We can have only naïve expectations, distributing the county level ($x^{o}_{j,c}$) equally to municipalities where production occurs in reality (pseudo population). The prior information on aggregated activity groups (q^{p}) is derived based on LWA (ANSELIN et al., 2006). Generally, LWA values are weighted averages of production activity levels observed in municipality (m) and in their neighbours (g). For all (M) municipalities we calculate the production activity groups ($q^{\mathrm{p}}_{m,n}$) using (5) by:

(6) $x^{*}_{i_n,m} = \; \mathrm{w}^{\mathrm{E}}\, x^{o}_{i_n,m} + \sum_{g} \left(1 - \mathrm{w}^{\mathrm{E}}\right) x^{o}_{i_n,g} GWF_{g}^{\;\mathrm{E}}\,,$

where (w^{E}) is a dispersion factor and (GWF^{E}) is a weighting factor depending on the experiment, described in the remainder. Note that because the official statistic unit (FDZ) only provides data at aggregated LWA levels, we can only use aggregated groups of activities as prior for (q^{p}), in order to be in compliance with the DPR. The question now is how to define the aggregates derived from the sampling methods (E) and how this information can recover the underlying distribution of the pseudo population.

2.3 Experimental Design

In all experiments the county production activity levels as used in formula (1) are given and we know whether or not production activities are observed at a particular municipality. We define a reference point (*STD*) as a naïve experiment where no other prior information on aggregated activity levels, besides the real UAA from the pseudo population, are provided for the estimation, and the prior information on (x^{p}) is calculated, distributing the county level

[1] We justify this in Section 3.

($x^{o}_{j,c}$) equally to municipalities[2] where production occurs in reality. In addition to a reference (*STD*), we conduct three experiments (E = [*EXACT*, *MW_N*, *MW_A*]) to evaluate the impact of different algorithms for calculating the priors ($q^{\mathrm{p}}_{m,n}$). In experiment *EXACT*, we assume that we obtain the real values for the priors on the municipality level calculated as (5) where the activity levels from the pseudo population are used ($x^{*}_{i_n,m} = x^{o}_{i_n,m}$), implying formula (6) with $\mathrm{w}^{\mathrm{E}} = 1$. Here, the observed pseudo population production activity levels equal our activity level for defining the prior (q^{p}). This experiment serves as a benchmark to assess the information loss induced using LWA algorithms ($\mathrm{w}^{\mathrm{E}} < 1$) which take into account neighbouring information to increase the sample size at municipality (*m*). The following LWA sampling methods are distinguished. In order to investigate the impact of different weighting schemes we define experiment *MW_N* (Moving windows weighted by number the neighbours) and experiment *MW_A* (Moving windows weighted by neighbours' used agricultural land (UAA)). For both experiments the dispersion factor (w^{E}) is set to 0.5, whereas GWF^{MW_N} is set to the reciprocal value of the number of (g_m)'s neighbours. GWF^{MW_A} is based on the UAA of (g_m)'s neighbours instead of their number.

For each experiment, we evaluate six scenarios differing in the number of aggregated activity groups as prior information. In all scenarios the UAA is given (Table 1). Scenario *area_raw* provides the respective acreage of arable land and grassland as additional prior information. Scenario *area_fine* supplements *area_raw* with information on the acreage of permanent and annual specialised crops and the main forage area. The *anim_fine* scenario differs from *anim_raw* that the total stocking is not the only additional prior information but also the stocking of the granivores is used in the calculations. *Aran_raw* combines the information of *area_raw* and *anim_raw*, while *aran_fine* includes all seven priors.

Table 1: Overview of the prior information used in the scenarios

Prior	area_raw	area_fine	anim_raw	anim_fine	aran_raw	aran_fine
UAA (ha)	X	X	X	X	X	X
Grassland (ha)	X	X			X	X
Permanent crops (ha)	X	X			X	X
Annual specialised crops (ha)		X				X
Main forage area (ha)		X				X
Stocking (LU)			X	X	X	X
Granivores (LU)				X		X

Source: own presentation, LU = Livestock Units

The relation between the experiments' fit and *STD*'s fit of the original data serves as indicator for the value of the additional prior information and is defined in the following section.

2.4 Test statistics

To validate the estimation results we compare the production activity levels at municipality level of the pseudo population with the estimates obtained from the experiments and their scenarios. Besides the known Pearson correlation coefficient (Pearson's r) we will use as a goodness of fit measure the aggregated Residual Sum of Squares (aRSS). We calculate it, subscripted with "*L*" the absolute deviation whereas "o" is the subscript for the pseudo population by

[2] All livestock and crop activities on county level are equally distributed using total LU and UAA obtained from different sampling methods (E) as weights.

(7) $$\text{aRSS}_L = \sum_{j,m} (x^o_{j,m} - x_{j,m})^2 \ .$$

In addition, and to capture the structure and composition of land use and livestock husbandry, the aRSS based on the activities respective relative shares on UAA and on LU are calculated by

(8) $$\text{aRSS}_S = \sum_{t,m} \left(\frac{x^o_{t,m}}{LU^o_m} - \frac{x_{t,m}}{LU_m}\right)^2 + \sum_{f,m} \left(\frac{x^o_{f,m}}{UAA^o_m} - \frac{x_{f,m}}{UAA_m}\right)^2 \ \ t \in j; f \in j \ .$$

The index (t) depicts all RAUMIS livestock activities and (f) the respective cropping activities. Then we calculate the aRSS relative to the reference point *STD* for both, the absolute activity levels (FIT_L) and the cropping and livestock shares (FIT_S) calculated by

(9) $$\text{FIT}_S = 1 - \frac{\text{aRSS}_S}{\text{aRSS}_S^{STD}} \ \text{ and } \ \text{FIT}_L = 1 - \frac{\text{aRSS}_L}{\text{aRSS}_L^{STD}} \ .$$

3 Data

We use data of the Bavarian agricultural census for the year 1999 to evaluate different options to recover local information on agricultural land use (BayLStaD, 1999). This data is based on the total population of Bavarian farms and differentiates among other 95 different cropping and livestock husbandry activities. We aggregate these 95 codes into the 36 activities defined in RAUMIS (Table 2). We choose the Bavarian data set for several reasons:

1. Due to the high number of farms per municipality most information on land use is published at the municipality level. For comparison, for each of the 95 statistical codes in Eastern Germany 84% of the data are kept confidential on the municipality level (FDZ, 2010). In comparison, for each of the 95 statistical codes in the Bavarian sample, 21% of the data are kept confidential on the municipality level. The gaps and inconsistencies of 21% were removed using statistical methods as applied by BRITZ and WITZKE (2008), which resulted in a very close to reality data set in the paper named as *pseudo population*. Please note, the pseudo population at municipality level is not used in the estimation framework, but only used to sample the prior information on aggregated activity groups in (5) using the LWA methods and, of course, it is used as benchmark to evaluate the estimation results.
2. Roughly a fifth of all German municipalities and of German's UAA is located in Bavaria. Therefore, it should be feasible to draw some extrapolations regarding the selected indicator's behaviour to the rest of Germany.
3. Bavaria shows a quite high regional diversity regarding the dominant types of agricultural production, e.g.: low input livestock husbandry in the Alps, intensive dairy farming on grassland in the county of *Unterallgäu*, intensive dairy farming and arable land in the county of *Ansbach*, intensive granivore production in the county of *Rottal Inn*, intensive cash cropping with sugar beets and vegetable production in the *Straubinger Gäu*, extensive cash cropping in the *Münchner Schotterebene* and wine production along the river *Main*.
4. Due to the undulated relief in the low mountain ranges and Alps the conditions for agricultural production are quite heterogeneous even within one county. While, e.g., the southern part of the county of *Rosenheim* is dominated by low input dairy farming with a significant share of rough grazing, the middle part is characterized by intensive grassland based dairy farms, and arable forage cropping and cash cropping are typical for the Northern part of the county.
5. We have some activities with a very high degree of concentration (e.g., 9% of the vineyards and 7% of the laying hens can be found in one municipality).

Table 2: Activities in RAUMIS and their respective extent in Bavaria (1999)

RAUMIS	Description	n°	Extent	avg.	max.	σ
				(LU or ha)		
KALB	Calves	1,955	185,895	93.9	1,070	3.1
BULL	Male cattle > 6 month; stock bulls	1,944	332,870	185.1	3,246	7.2
FAER	Heifers	1,977	814,038	306.9	2,873	9.6
MIKU	Dairy cows	1,950	1,453,902	734.6	7,262	23.7
AMMU	Suckler and fattening cows	1,963	87,504	171.2	1,259	5.4
SCHA	Sheep	1,765	38,399	23.1	324	1
SOTI	Other livestock (horses)	1,901	82,203	41.7	465	1.4
SAUH	Sows for piglet production	1,537	154,900	87	1,328	4.2
SMAS	Pig fattening	1,903	279,105	116.6	3,045	5.9
LEHE	Laying hens	1,898	22,502	11.8	1,653	1.6
SOGE	Poultry fattening (broiler, turkeys, …)	1,223	19,379	15.8	1,559	2.7
WWEI	Winter wheat, spelt	1,801	378,003	205.3	1,929	7.9
SWEI	Summer wheat, durum wheat	1,623	35,715	21	292	1
WGER	Winter barley	1,790	276,889	151.6	1,212	5.3
SGER	Summer barley	1,814	180,317	97.2	1,507	4.2
ROGG	Rye, and winter cereal mixes	1,475	45,886	30	699	1.5
HAFE	Oats and summer cereal mixes	1,787	74,323	40.8	454	1.4
KMAI	Grain maize (incl. CCM)	1,183	94,776	78.7	2,786	5.9
SGET	Other cereals, trititcale	1,654	71,347	42.2	489	1.6
RAPS	Rape and turnip rape	1,583	176,642	109.1	923	4
HUEL	Pulses	1,303	16,686	12.5	124	0.5
SHAN	Other oilseeds and industrial crops (hops, tobacco, …)	895	35,832	39.3	1,844	4.1
SKAR	Potatoes	1,799	55,476	29.5	1,321	2.3
ZRUE	Sugar beet	849	77,703	89.4	1,281	6.4
SHAC	Other root crops (fodder beet,…)	1,048	2,850	2.3	32	0.1
SMAI	Green and silage maize	1,857	301,420	160.1	1,842	5.5
KLEE	Alfalfa / clover and mixtures with grass	1,838	116,792	62.7	1,070	2.5
FEGR	Grass on arable land (incl. all other fodder on arable land)	1,570	77,703	10.6	282	0.6
WIES	Meadow	2,012	1,066,796	521	6,050	17
WEID	Pasture	1,821	52,025	27.7	552	1.4
HUTU	Rough pastures	1,355	62,417	43	1,624	3.6
FLST	Set aside	1,662	17,023	74.4	681	2.6
GEMU	Vegetables, strawberries	1,360	13,998	9.6	643	1
SOPF	Other plant production (flowers, nurseries, …)	1,464	5,888	3	86	0.2
OBST	Fruits (without strawberries)	1,380	7,454	5.3	353	0.6
REBL	Wine	163	5,809	33	538	6.1

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

4 Results

This section is structured as follows. The first part in Section 4.1 analyzes the question, to which degree the RSS can be reduced, while complying with DPR. Therefore, we analyze how the overall model fit is affected by the different LWA experiments and the available prior information (scenarios). In addition, we present the dependence of the RSS on different activities, and we map the RSS, to investigate the spatial distribution of the error. In Section 4.2 we focus on the topic, whether the achieved reduction allows a reasonable coverage of the underlying land use. Hence, we analyse the average correlation between the observed and estimated values. We conclude using the introductory example with the distribution of silage maize to exemplify the quality of the estimation.

4.1 Error Distribution

Table 3 shows, that in experiment *EXACT,* FIT_L (cf. (9)) rises from 18% for scenario *area_raw* to 68% for *aran_fine*. All experiments rank the scenarios in a comparable order and additional information on livestock husbandry elevates FIT_L stronger than on cropping. The LWA experiments reduce the quality of the estimation markedly, especially if only the number of neighbours is considered. The values of FIT_L of *MW_N's* scenarios are 35% to 41% lower than the respective figures in *EXACT*. If no information on livestock husbandry is provided the fit of the activity levels is even worse than in *STD* (negative values of FIT_L). *MW_A* performs clearly better than *MW_N* and the respective values for FIT_L are on average 23% higher, but still 12% to 18% below EXACT.

Table 3: Absolute and relative fit of the original data in the experiments in dependence of the provided prior information (FIT_L / FIT_S)

	Cropping		**Livestock husbandry**		**Both**	
	area_raw	**area_fine**	**anim_raw**	**anim_fine**	**aran_raw**	**aran_fine**
EXACT	18% / 24%	24% / 31%	25% / 0%	43% / 16%	43% / 24%	68% / 46%
MW_N	-17% / 21%	-12% / 25%	-11% / -1%	6% / 11%	5% / 21	27% / 36%
MW_A	6% / 24%	12% / 28%	12% / 0%	29% /13%	28% / 24%	50% / 41%

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

For a given scenario the differences between the experiments of the FIT_S values are generally smaller than 10%. This implies that *MW_A* and *MW_N* recover the composition of the production activity shares nearly as well as if unblurred priors (*EXACT*) were to be used. The relative importance of the LWA algorithm and the amount of prior information differs between FIT_L and FIT_S. In contrast to FIT_S, FIT_L is strongly negatively affected if only blurred information regarding the UAA is available. For FIT_L this negative impact can hardly be compensated by additional prior for cropping activities (cf. scenarios *area_raw* and *area_fine* in experiment *MW_N*). If information is provided for one of the two domains only, livestock husbandry or cropping, additional information on livestock husbandry improves FIT_L and FIT_S stronger than additional information on cropping.

Figure 1 shows, that the cattle stock (KALB, BULL, FAER, MIKU, AMMU) contributes nearly 50% to the aRSS_L, and another 20% result from the mapping of the grassland (WIES, WEID, HUTU) in *STD*. The aRSS_L is strongly influenced by the estimation of these two groups of activities. This high dependence of the aRSS_L is not surprising if one considers that cattle account for over 82% of the Bavarian stock and grassland covers more than 36% of the UAA. These activity levels must be distributed into five and three categories respectively without any other prior information on municipality level. In experiment *MW_A* adding information on the municipalities' stock (*anim_raw*) halves the RSS_L over the cattle activities. If additionally data on the granivore stock is provided (*anim_fine*) the respective RSS_L drops to a third of *STD*'s value. Providing information on the total grassland acreage at the municipality level (*area_raw*), more than half of the RSS_L is aggregated over the grassland activities. As expected, providing additional information on other forms of land use, e.g., main forage area or the area of permanent crops (*area_fine*) does not influence grassland's RSS_L.

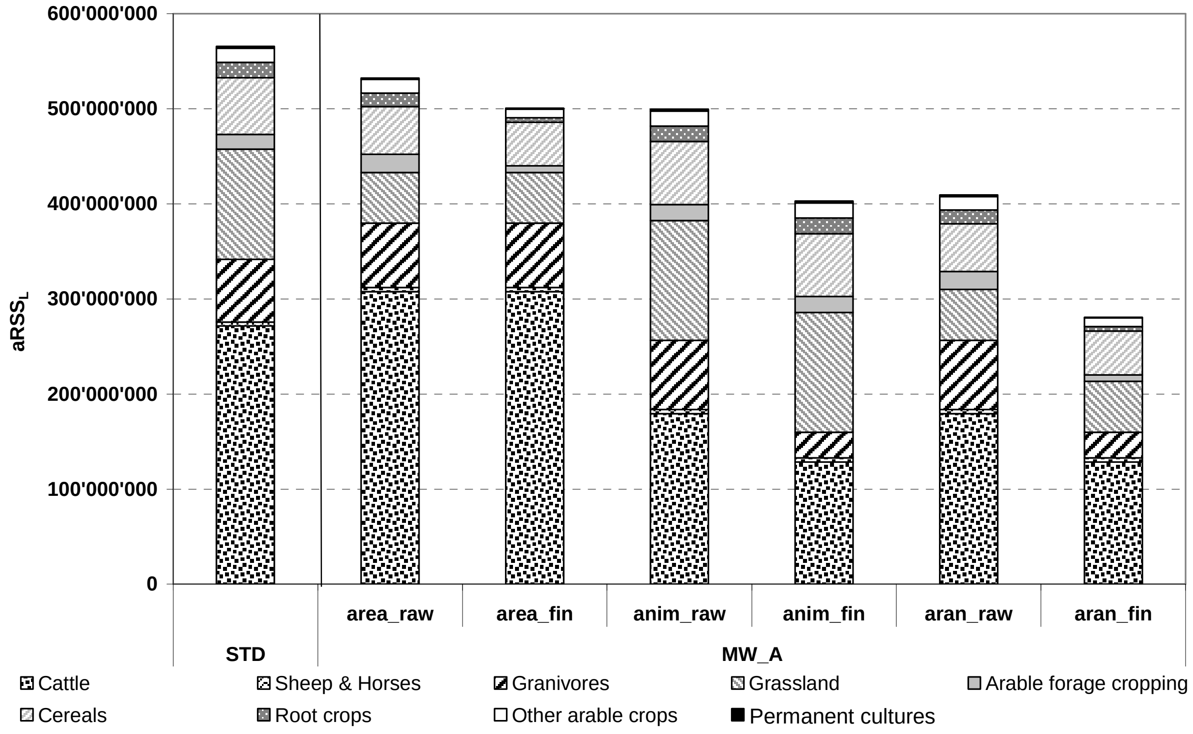


Figure 1: Effect of the scenarios on the composition of aRSS$_L$ for experiment MW_A

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

Analysing the shares, the inclusion of information on the stocking level only (*anim_raw*) does not improve the aRSS$_S$ even for livestock husbandry activities (Figure 2).

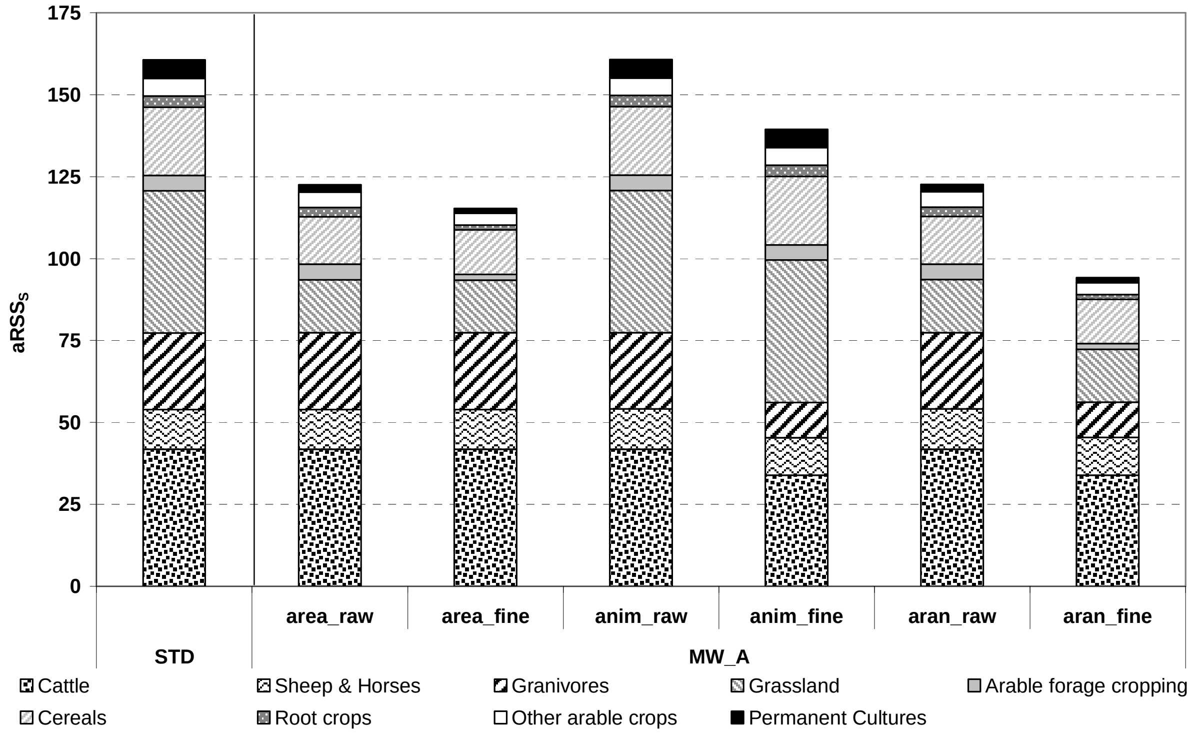


Figure 2: Effect of the scenarios on the composition of aRSS$_S$ for experiment MW_A

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

The explicit coverage of aggregated cropping activities (grassland, arable forage cropping, permanent cultures) in *area_fine* reduces the aggregates' RSS$_S$ to 25% to 40% of the respective *STD*'s value. The relative importance of the different activity groups differs between aRSS$_S$ and aRSS$_L$. The most pronounced difference is that aRSS$_L$ is much less influenced from the cattle sector than aRSS$_S$ and that the cattle activities' RSS$_S$ is only slightly improved by addition of information. On the other hand, fairly negligible activities, like horses and sheep, that account for 3.5% of the Bavarian stock, contribute roughly 10% to the aRSS$_S$.

In the last part of this section we analyze how well the best LWA algorithm (*MW_A aran_fine*) works throughout Bavaria (Figure 3). As the number of activities and the UAA varies among the municipalities, we aggregate for each municipality the RSS$_S$ of the activities and divide it by the respective number of activities. The number of municipalities with high RSS$_S$ values declines markedly. One can clearly see that the RSS$_S$ is not randomly distributed[3]. These high RSS$_S$ values are induced by a high degree of heterogeneity regarding conditions for agricultural production. Concentrations of high values can be found in the Alps, the *Rhön* and *Spessart* and the counties along the *Danube* in *Lower Bavaria*. The Alps, the *Rhön* and the *Spessart* are mountainous areas characterized by low input grassland based forage cropping systems. The high RSS$_S$ are due to the bad fit of the estimation of sheep, horses and rough pastures. In case of the counties along the *Danube* the high RSS$_S$ in *STD* can be mainly attributed to the fact that these counties can be divided into two parts. The first compromising areas with intensive cash cropping on the southern bank of the *Danube* and areas with intensive grassland based forage cropping on the northern bank. The additional prior information used in *aran_fine* is capable to reduce this error.

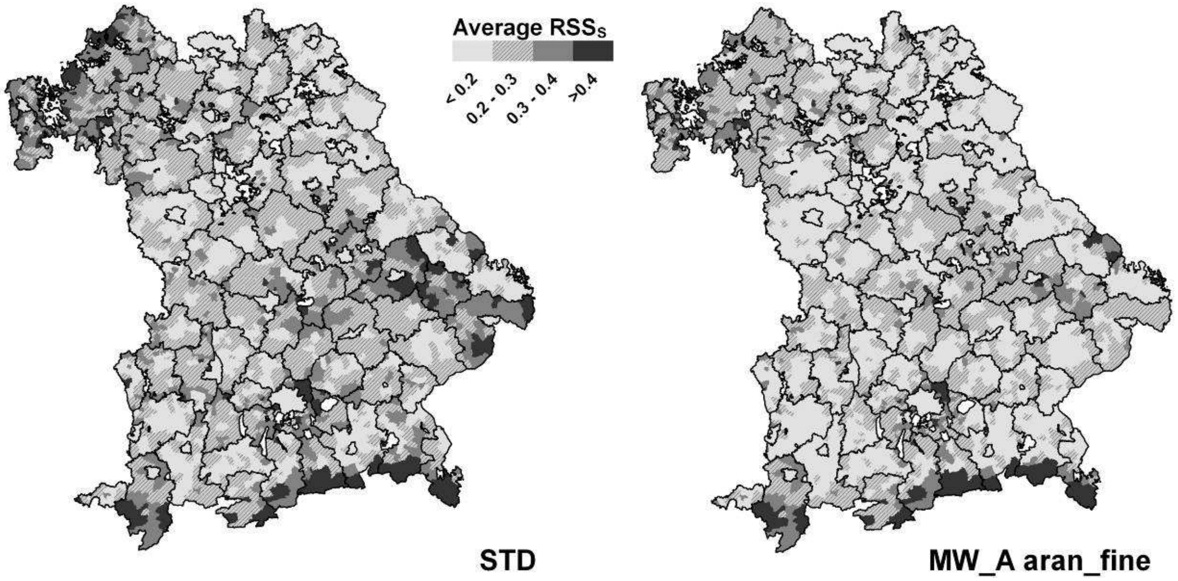


Figure 3: Comparison of the spatial distribution of the average RSS$_S$ between *STD* and *MW_A aran_fine*

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

4.2 Fit of the estimation

As shown in the previous section, the more information provided, the better the fit of the data. We now analyze whether we achieved a suitable fit for the different RAUMIS activities. To analyze this, we use Pearson's r. Table 4 depicts the average and smallest correlation coefficients based on absolute activity levels for all experiments and scenarios. Pearson's r

[3] The respective Moran's I, as measure of spatial correlation, confirm the presence of spatially clustered patterns for *STD* and *MW_A aran_fine* at significance level of 1%.

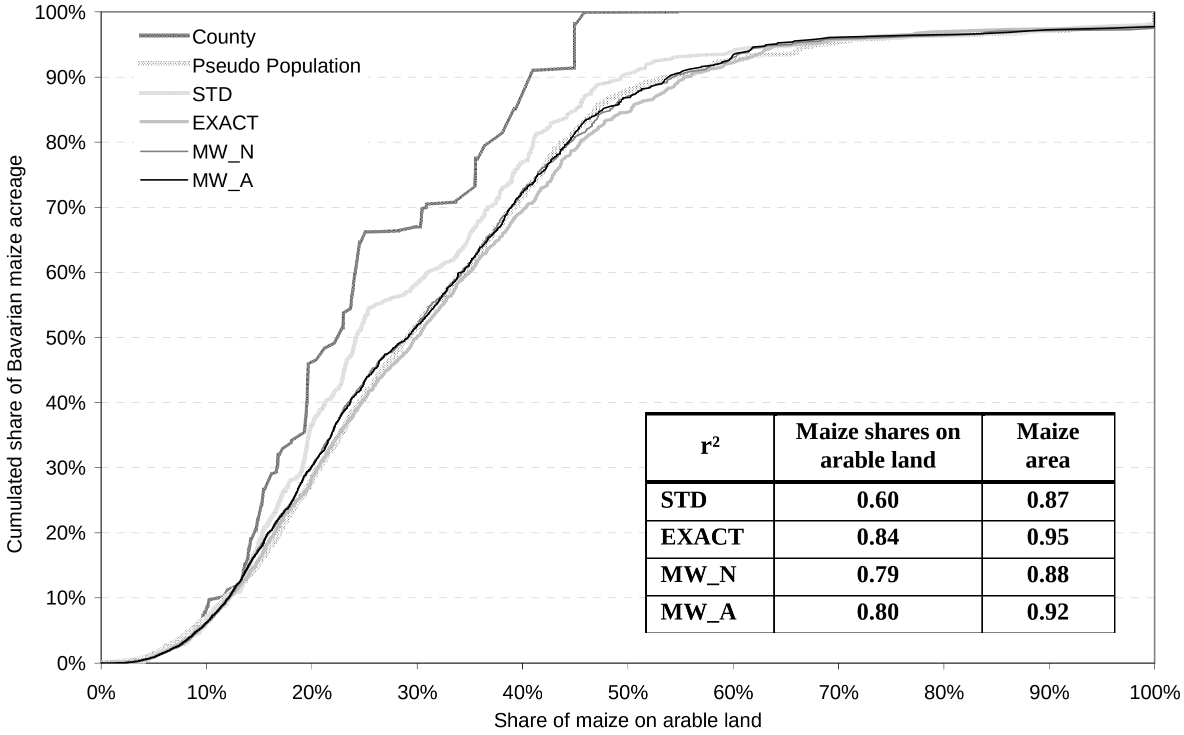
varies more across the scenarios (type and number of prior information) and than across the experiments (LWA algorithm). Generally speaking, a higher number of prior information increases the fit, but the average of the coefficients varies only in a fairly small bandwidth (0.71-0.84). It can be seen that detailed information on cropping and livestock husbandry must be provided (*aran_fine*) to lift the minimum Pearson's r (i.e., improve the worst match of the considered activities). The respective Pearson's r based on relative activity shares convey the same general picture (not shown).

Table 4: Average Pearson's r of all RAUMIS activities for STD and the experiments based on absolute activity levels (in brackets smallest observed correlation coefficient)

		Cropping		Livestock husbandry		Both	
	STD	**area_raw**	**area_fine**	**anim_raw**	**anim_fine**	**aran_raw**	**aran_fine**
STD	0.73 (0.34)						
EXACT		0.74 (0.34)	0.80 (0.34)	0.76 (0.34)	0.80 (0.34)	0.74 (0.34)	0.84 (0.50)
MW_N		0.74 (0.33)	0.77 (0.33)	0.71 (0.30)	0.73 (0.35)	0.74 (0.30)	0.80 (0.43)
MW_A		0.75 (0.35)	0.78 (0.35)	0.72 (0.32)	0.75 (0.33)	0.76 (0.32)	0.81 (0.51)

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

Figure 4 depicts the cumulative distribution of the Bavarian maize acreage in dependence of maize's share in the crop rotation. First, *STD* allows a much better recovery of the distribution at the municipality level than using county averages only. Nevertheless, the LWA algorithms are able to improve the fit further, although the information on municipality level is blurred.



r²	**Maize shares on arable land**	**Maize area**
STD	**0.60**	**0.87**
EXACT	**0.84**	**0.95**
MW_N	**0.79**	**0.88**
MW_A	**0.80**	**0.92**

Figure 4: Cumulative density distribution of the Bavarian maize acreage in dependence of maize's share on the respective arable land (the data for the scenarios *EXACT*, *MW_N* and *MW_A* use scenario *aran_fine*)

Source: BayLStaD, Landwirtschaftszählung (agricultural census), 1999, own calculation.

For all three experiments the respective r² for the relative crop shares and absolute activity levels are of a comparable magnitude. Despite the fact that for maize the r² are quite satisfactory for Bavaria as a whole, the fit in particular regions can be quite devastating. E.g., in the county of *Passau* the r² for the mapping of maize based on its relative share in the crop rotation lie depending on the chosen experiment between (0.02 and 0.07).

5 Conclusions & Outlook

The presented disaggregation, combining Highest Posterior Density (HPD) and Locally Weighted Averages (LWA), is able to improve the estimate of the land use on the municipality level while complying with data protection rules (DPR). Acceptable results could be found for LWA based on weights derived from the municipalities' UAA (*MW_A*) and a limited amount of prior information on aggregated activities (*aran_fine*). An implementation of the LWA (*MW_N*), considering the number of neighbours, implies an information loss of up to 40% compared to a situation where prior information for the municipality is available with certainty (*EXACT*). In *MW_A* this loss could be reduced to levels below 20%. While the fit of the absolute activity levels is strongly influenced by the selected LWA algorithm, its impact on the estimation of the relative shares is relatively small. It remains open, how the results of the different algorithms respond to situations (e.g., East Germany) when the data are characterised by large fluctuation on local level, e.g., in case the local aggregates are derived from few farms.

If the correlation between the observed and predicted values is analyzed for the entire data set, we can conclude that the proposed approach is capable to adequately depict for most RAUMIS activities both their spatial and density distribution while complying with DPR. It could also be shown that the correlation coefficient is hardly affected by the chosen LWA algorithm. This can be intuitively explained as the LWA affects only the local variation but this variation is prevailed by the high regional variation in the pseudo population.

The proposed procedure can be extended and improved in different directions. The goodness of fit of the approach is low if the counties are very heterogeneous, and if this is not covered by the aggregated variables used as priors on municipality level. In the current implementation the key data at the county level are equally distributed among the respective municipalities if no additional information is provided. However, increasing the number of priors is no solution, as one would soon face the same problems due to data confidentiality restrictions as on the level of the individual activity. How can this problem be tackled? First, we could replace the counties by units that better reflect the differences in the conditions for agricultural production using, e.g., *Bodenklimaräume* (ROßBERG et al., 2007) or units resulting from a tailor made regionalization, in order to reduce the heterogeneity of the municipalities belonging to a certain regional unit. Second, we could run cluster analyses based on the activity data on the municipality level, and use for each activity the respective cluster medians as prior information in a given municipality. Third, we could improve the fit by integrating geo-referenced data sources, e.g., CORINE or ATKIS.

References

ANSELIN, L., N. LOZANO and J. KOSCHINSKY (2006): Rate Transformation and Smoothing. URL:*geodacenter.asu.edu/pdf/smoothing_06.pdf.* University of Illinois: p. 87.

BAYLSTAD (Bayerisches Landesamt für Statistik und Datenverarbeitung) (1999): Landwirtschaftszählung (agricultural census), München.

BRITZ W. and P. WITZKE (2008): CAPRI model documentation (2008): Available at http://www.capri-model.org/docs/capri_documentation.pdf, pp. 181.

CARRASCO L. R., T. D. HARWOOD, S. TOEPFER, A. MACLEOD, N. LEVAY, J. KISS, R. H. A. BAKER, J. D. MUMFORD and J. D. KNIGHT (2009): Dispersal kernels of the invasive alien western corn rootworm and the effectiveness of buffer zones in eradication programmes in Europe. Annals of Applied Biology 156 (1): 63-77.

FDZ (Forschungsdatenzentrum der Statistischen Ämter des Bundes und der Länder) (2010): AFiD-Panel.

GOCHT, A. (2010): Methodes in Economic Farm Modeling. Ph.D. Thesis, Universität of Bonn. Available at http://hss.ulb.uni-bonn.de /diss_online/, pp. 11.

GOLAN A., G. JUDGE and D. MILLER (1996): Maximum entropy econometrics, Robust Estimation with Limited Data. John Wiley, New York.

HECKELEI T., R. MITTELHAMMER and W. BRITZ (2005): A Bayesian Alternative to Generalized Cross Entropy. Paper presented at the 89th European Seminar of the European Association of Agricultural Economists (EAAE), Parma, Italy, February 3-5.

HECKELEI T., T. JANSSON, and R. MITTELHAMMER R. (2008): A Bayesian Alternative to Generalized Cross Entropy Solutions for Underdetermined Econometric Models Discussion Paper 2008:2, University of Bonn, Available at http://www.ilr1.uni-bonn.de/agpo/publ/dispap/download/dispap08_02.pdf

HENRICHSMEYER, W., CYPRIS, C., LÖHE, W., MEUDT, M., SANDER, R., VON SOTHEN, F., ISERMEYER, F., SCHEFSKI, A., SCHLEEF, K.-H., NEANDER, E., FASTERDING, F., HELMCKE, B., NEUMANN, M., NIEBERG, H., MANEGOLD, D., and MEIER, T. (1996): Entwicklung eines gesamtdeutschen Agrarsektormodells RAUMIS96. Endbericht zum Kooperationsprojekt. Forschungsbericht für das BML (94 HS 021), 1996, vervielfältigtes Manuskript Bonn/Braunschweig.

KEMPEN M., BRITZ W. and HECKELEI T. (2005): A Statitical Approach for Spatial Disaggregation of Crop Production in the EU, In: Arfini Filippo (ed.). Modelling agricultural policies: state of the art and new challenges; proceedings of the 89th European Seminar of the European Association of Agricultural Economists (EAAE), Parma, Italy, February 3rd-5th, 2005. Parma : Monte Universita Parma Editore, pp. 810-830.

LEIP A., MARCHI G., KOEBLE R., KEMPEN M., BRITZ W. and LI C. (2008): Linking an economic model for European agriculture with a mechanistic model to estimate nitrogen and carbon losses from arable soils in Europe. Biogeosciences 5(1), 73-94.

OSTERBURG B., H. NITSCH, B. LAGGNER and W. ROGGENDORF (2009): Auswertung von Daten des Integrierten Verwaltungs- und Kontrollsystems zur Abschätzung von Wirkungen der EU-Agrarreform auf Umwelt und Landschaft. Arbeitsberichte aus der vTI-Agrarökonomie 07/2009. Braunschweig.

ROßBERG D., V. MICHEL, R. GRAF and R. NEUKAMPF (2007): Definition von Boden-Klima-Räumen für die Bundesrepublik Deutschland. Nachrichtenbl. Deut. Pflanzenschutzd., 59 (7): 155–161.

ZELLNER, A. (1971): An introduction to Bayesian inference in econometrics. New York: Wiley.