A Minimum Cross-Entropy Approach to Disaggregate Agricultural Data at the Field Level

António Xavier [1],*, Rui Fragoso [1] , Maria de Belém Costa Freitas [2], Maria do Socorro Rosário [3] and Florentino Valente [4]

[1] CEFAGE-UE (Center for Advanced Studies in Management and Economics), Management Department, Universidade de Évora, N° 2, Apt. 95, 7002-554 Évora, Portugal; rfragoso@uevora.pt

[2] ICAAM (Institute of Mediterranean Agricultural and Environmental Sciences), Sciences and Technology Faculty, Universidade do Algarve, Gambelas Campus, Edf. 8, 8005-139 Faro, Portugal; mbfreitas@ualg.pt

[3] Direção de Serviços de Estatística, GPP (Gabinete de Planeamento e Políticas), Praça do Comércio, 1149-010 Lisboa, Portugal; socorro.rosario@gpp.pt

[4] Direção Regional de Agricultura e Pescas do Algarve, Patacão, 8001-904 Faro, Portugal; fvalente@drapalg.min-agricultura.pt

* Correspondence: amxav@sapo.pt; Tel.: +351-289-800-900 (ext. 7391)

Abstract: Agricultural policies have impacts on land use, the economy, and the environment and their analysis requires disaggregated data at the local level with geographical references. Thus, this study proposes a model for disaggregating agricultural data, which develops a supervised classification of satellite images by using a survey and empirical knowledge. To ensure the consistency with multiple sources of information, a minimum cross-entropy process was used. The proposed model was applied using two supervised classification algorithms and a more informative set of biophysical information. The results were validated and analyzed by considering various sources of information, showing that an entropy approach combined with supervised classifications may provide a reliable data disaggregation.

Keywords: data disaggregation; supervised classifications; classification algorithms; minimum cross-entropy; land uses; Algarve; empirical validation

1. Introduction

Agriculture and forests are essential to preserve biodiversity and develop the economy in rural areas. They supply essential goods for human survival and well-being and hence need to be well managed [1]. Thus, information on the spatial distribution of land-use at a detailed level is crucial for models and applications on agro-forestry production that require a spatial representation [2,3]. For instance, in the European Union, agricultural statistics try to report information at the regional and sub-regional level. However, the Agricultural Census, which is the main territorial statistical operation in European Union, is carried out every 10 years and between this period, there is no available information at the municipality or parish levels. This lack of information is a worldwide problem since an updated knowledge of land-use contributes to ensure a judicious spatial planning that considers characteristics of interest [4–9].

Every 3 years the LUCAS survey (Land-Use/Cover Area frame Statistical survey) is carried out. Taking photographs, this survey collects land cover/land-use, agro-environmental, and soil data by field observation of referenced points [10]. Information is also available monthly via satellite imagery from LANDSAT and more recently SENTINEL 2. LANDSAT and SENTINEL 2 are multispectral satellites with high spatial resolution developed by the National Aeronautics and Space Administration

(NASA) and by the European Space Agency (ESA), respectively. All these sources can be used to update and predict the land-use at a detailed level where the information is incomplete. Thus, this paper is addressed to the development of a framework that is able to use several sources of information including satellite imagery to disaggregate agricultural data.

The traditional land-use models are generally based on econometrics tools and typically aim to assess the relationship between land use choices and a set of independent variables [11]. The land use literature already explicitly considers the spatial dependence, but only a few studies benefit from spatial econometric tools [12–14]. Recently, the proposed land use models tend to ignore spatial dependency since it raises some problems associated with econometric estimation, namely, hypothesis testing and prediction [15,16].

However, there is an increased demand for data disaggregation tools [17–20]. Several models have been developed, using different techniques to directly disaggregate data from statistical sources. A cross-entropy approach to disaggregate data at the pixel level, using previous estimates from land cover maps and suitability maps was developed [5,6]. An approach that uses a multinomial logistic regression model, taking advantage of the LUCAS survey and an entropy approach to disaggregate data from regional statistics has been presented [13]. The posterior density of the previous estimated was maximized to achieve consistency with administrative statistics, using information from the sampling points utilized to estimate the land-use choices [4]. In Europe, the Capri-Spat (Common Agricultural Policy Regional Impact) approach for disaggregating agricultural data from statistics, which uses the Highest Posterior Density estimator, was compared with the Dynamic Conversion of Land Use and its Effects model [21]). In Portugal, Reference [22] presented an approach which combines a previous estimate of an iterative algorithm and the HJ-Biplot methods. In this study, a general approach is also proposed, which uses supervised classifications as previous estimates in a cross-entropy model.

Remote sensing and satellite imagery offer new perspectives for analyzing land use, namely by performing supervised classifications in which sample areas ("training fields") are used to identify the spectral signatures of the various land uses automatically [23]. In the last 20 years, supervised classifications of satellite imagery have developed considerably [24–27] and several case studies can be found worldwide. Two agricultural land cover classifications using mono-temporal and multi-temporal Landsat scenes have been presented [28]. A land use map of the Galaudu watershed was produced in Nepal [29]. Satellite data and data from the Israeli Geographic Information System were used to create a land use map for the northern Negev [30]. Maximum likelihood supervised classification and post-classification change detection techniques were applied to Landsat images to map the land cover changes in the north-west coast of Egypt [31].

Over the last years, the agricultural data disaggregation approaches used several techniques: logistic regressions, expert knowledge, homogeneous units, Bayesian methods, entropy, and highest posterior density. These approaches allow taking advantage of the existing information (including biophysical information, land use maps, and the LUCAS survey), but not of the data presented on satellite imagery that are available periodically using automatic cartography techniques, such as supervised classifications.

Therefore, this study proposes a methodological approach for disaggregating agricultural data using supervised classifications and an entropy approach. This is the first study that provides an approachable method to take full advantage of the LUCAS survey, supervised classification techniques, biophysical data, experts' knowledge, and historical data for disaggregating data at the detailed pixel level from incomplete information. No other study has proposed the use of supervised classification techniques for data disaggregation or the use of the LUCAS survey and expert knowledge to carry out supervised classifications. Moreover, techniques for improving these estimates using entropy have not been considered. The proposed approach allows providing more up-to-date and reliable data due to the good periodicity of the available satellite imagery.

Thus, this study integrates a complete set of biophysical information, compares different classification algorithms and tests the proposed approach for the Algarve region and one pilot

municipality, in southern Portugal. Two different supervised classification algorithms are used in order to show the reliability of the approach. The aim is to include in the entropy model used in the disaggregation process for additional information through various restrictions.

The remainder of the paper is presented as follows: in section two, the methodological approach is presented; in section three, the data and model application scenarios are explained; sections four is dedicated to the results and analyses. Finally, section five presents the concluding remarks.

2. Methodological Approach

Recent research shows a variety of studies that use supervised classification techniques to produce thematic maps of land use [23,25,28,32]. The supervised classification method is an established classification from a training dataset, which contains the predictor variables measured in each sampling unit and assigns prior classes to the sampling units and, therefore, presents several advantages over unsupervised ones [26]. A comparison between different classification methods and their performance can found in References [27–30].

The methodological approach proposed combines several techniques, such as the classified supervisions of satellite images, cluster analysis, mapping, and cross-entropy minimization, and considers several sources of information. Several studies use entropy to estimate data when ordinary methods are not applicable since it overcomes some problems that hamper traditional econometric methods [6,33–36]. A generalized maximum entropy model to estimate multi-output production functions was adopted by References [37,38]. The maximum entropy can be used to estimate farm-level multi-input/multi-output production functions [39]. A dynamic approach for disaggregating agricultural data was presented by Reference [17]. The results of farm management models were disaggregated to the level required by natural science models [40]. Cross-entropy was also used to present the spatiotemporal dynamics of a maize cropping system in Northeast China [3]. In Portugal, several entropy models were also developed to disaggregate data [18,19,40–43].

The proposed methodological approach comprises two main steps, as shown in Figure 1. In the first one, prior information is previously estimated from a supervised classification of satellite imagery, the Lucas Survey, and experts' knowledge from the Ministry of Agriculture. In the second one, a cross-entropy model is applied to disaggregate the data from an aggregate level (for instance national or regional) to a detailed level (local or pixel level) with respect to the prior of information estimated. This procedure allows for a guaranteed consistency among the different sources of information and with the aggregate.

The supervised classification of satellite images is carried out to identify the distribution of land-use. This process comprises of the following steps: (i) collecting all the available information and selecting carefully the satellite imagery to be processed; (ii) defining the "training fields" using the LUCAS survey samples as references and empirical knowledge; (iii) defining the spectral signatures; (iv) implementing the supervised classification algorithms. As in other previous studies [4,8], full advantage was taken from the LUCAS Survey. It allows for using a set of samples that can work as "training fields" or sample areas in a "supervised classification" of the satellite images. To calculate the prior estimate, the Minimum Distance algorithm and the Maximum Likelihood algorithm algorithms were used and their results compared.

The Minimum Distance algorithm calculates the Euclidean distance (x, y) between the spectral signatures of the image pixels and the training spectral signatures, according to the following equation:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

where x is the spectral signature vector of an image pixel; y is the spectral signature vector of a training area; and n is the number of image bands. Therefore, the distance is calculated for every pixel in the image, assigning the class of the spectral signature that is closer, according to the following discriminant function:

$$d(x, y_k) < d(x, y_j), \ \forall k \neq j, \ x \in C_k \tag{2}$$

where C_k is the land cover class k; y_k is the spectral signature of class k; and y_j is the spectral signature of class j.
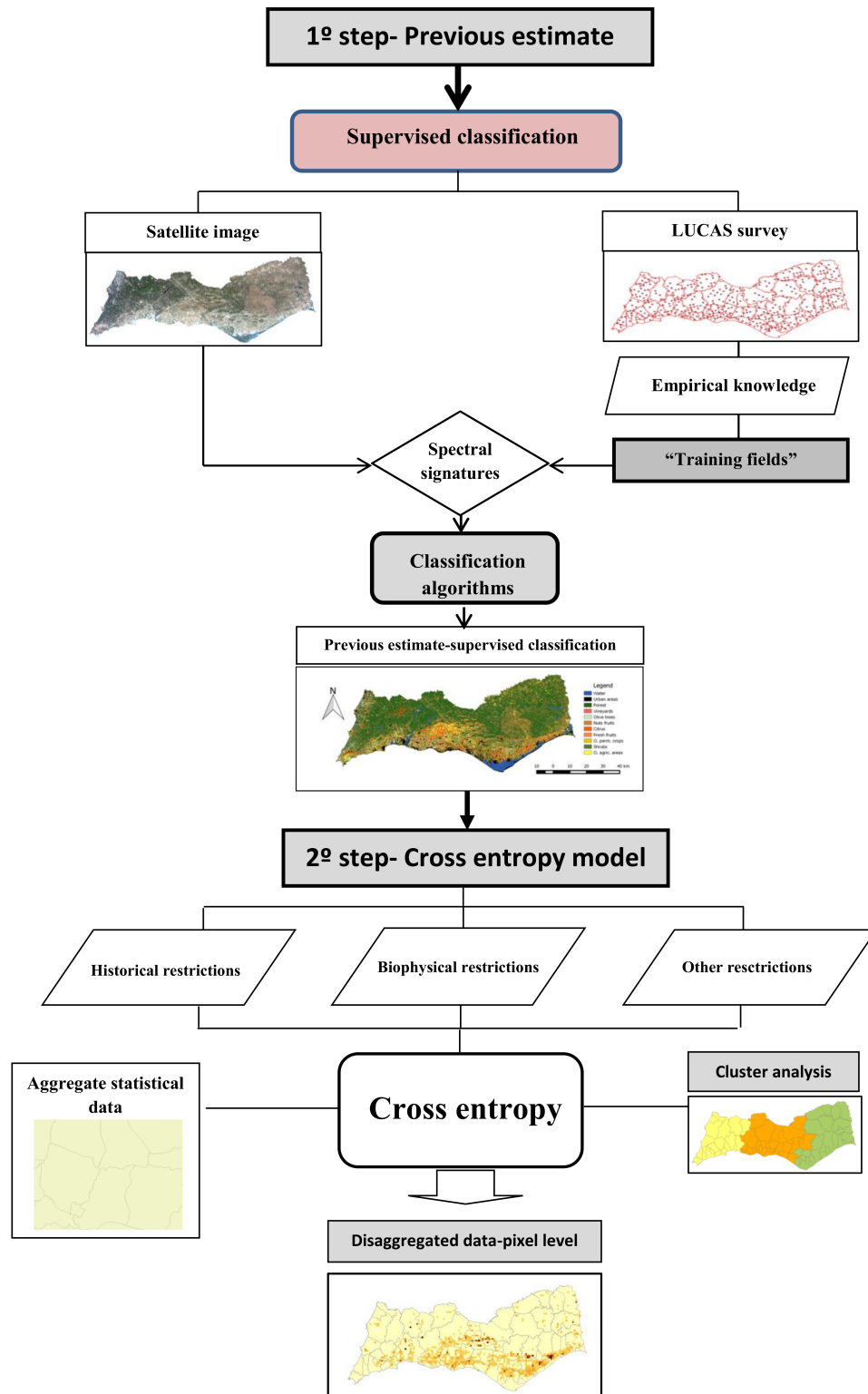

Figure 1. The methodological approach.

The Maximum Likelihood algorithm calculates the probability distribution for the classes related to Bayes' theorem, estimating if a pixel belongs to a land cover class. The discriminant function is calculated for every pixel as follows:

$$g_k(x) = ln\,p(C_k) - \frac{1}{2}ln\left[\left|\sum k\right| - \frac{1}{2}(x - y_k)^t\sum_{k}{}^{-1}(x - y_k)\right] \tag{3}$$

where $p(C_k)$ is the probability that the correct class is C_k; $\left|\sum_k\right|$ is the determinant of the covariance matrix of the data in class C_k; and $\sum_k{}^{-1}$ is the inverse of the covariance matrix.

Since the prior of information has been estimated, it can be used in the disaggregation process to guide a cross-entropy model. This procedure is very useful because it allows incorporating additional information in the disaggregation process, such as biophysical restrictions, historic restrictions, and so forth. In addition, a unique optimal solution is obtained to the disaggregation process and the consistency among different sources of information is guaranteed.

Thus, inspired by the studies of References [8,41], the following generalized cross-entropy model was developed

$$MaxH\left(\frac{x_k^i}{B_k^i}\right) = \sum_{i}^{I}\sum_{k}^{K} x_k^i \cdot log\left(\frac{x_k^i}{B_k^i}\right) + \sum_{k}^{K}\sum_{n}^{N} e_{kn} \cdot log(e_{kn}) \tag{4}$$

$$\sum_{k=1}^{K} x_k^i = 1 \text{ and } x_k^i \in [0,1] \text{ and } \sum_{n=1}^{N} e_{kn} = 1 \text{ and } e_{kn}(t) \in [0,1] \tag{5}$$

$$\sum_{i}^{I} x_k^i \cdot ST^i + \sum_{k}^{K}\sum_{kn}^{N} e_{kn} \cdot log(e_{kn}) = STAT_k \; \forall \text{ i and k} \tag{6}$$

$$x_k^i \cdot ST^i \leq LAND_k^i \; \forall \text{ i and k} \tag{7}$$

$$HM_k^i \leq x_k^i \leq HMX_k^i \; \forall \text{ i and k} \tag{8}$$

where, x_k^i is the probability of land-use k to be estimated in area i; B_k^i is the matrix of probabilities of each land-use k in area i resulting from prior estimates; ST^i is the area weight of each disaggregated unit i; $STAT_k$ are the regional statistics for land-use k; $LAND_k^i$ is the land use available for land-use k in disaggregated unit i; HM_k^i are the minimum historical limits and HMX_k^i are the maximum historical limits for each land-use by disaggregated unit i; and e_{kn} refers to a parameterized error term [36].

Equation (4) is the objective function, which minimizes the joint cross-entropy of the estimated probability distribution (x_k^i), the previous estimate (B_k^i), and the error distribution (e_{kn}). Equation (5) guarantees that x_k^i and e_{kn} have the characteristics of a probability distribution. Equation (6) ensures that the disaggregated shares x_k^i are compatible with the aggregate at the regional level. Equation (7) ensures that biophysical limits (restrictions of soils, climate, and slope) are respected. Equation (8) relates to the historical limits that must be respected for land use. These limits represent the maximum and minimum areas that a given land-use has achieved in the past. So, using this information, we can bound the model variables to more likelihood values concerning crop areas.

After having calculated the shares, it only remains to redistribute the regional data by using the following equation:

$$\hat{S}_k^i = x_k^i \cdot SA^i \tag{9}$$

where is the estimated area for land use k in unit i and SA is the area of unit i.

An important phase of our approach is the validation of the entropy model in order to test the coherency of the disaggregation process. To carry out this validation process, deviation measures and general statistical measures were used. As deviation measures, the Prescription Absolute Deviation (PAD) and the Weighted Prescription Absolute Deviation (WPAD) were considered. The PAD indicator measures the deviation between estimations and statistical data:

$$\mathrm{PAD}_{\mathrm{k}}^{\mathrm{c}} = \left| \frac{S_k^c - \hat{S}_{\mathrm{k}}^{\mathrm{c}}}{S_k^c} \right| \tag{10}$$

The *WPAD^i* indicator allows assessing the real deviation at the statistical unit *c* level and at the aggregate level and is obtained by the following:

$$\mathrm{WPAD}^{\mathrm{c}} = \sum_{k=1}^{k} P_k^c \mathrm{PAD}_{\mathrm{k}}^{\mathrm{c}} \tag{11}$$

Finally, at the aggregate level, WPAD is calculated as follows:

$$\mathrm{WPAD} = \sum_{\mathrm{c}=1}^{\mathrm{C}} \frac{\mathrm{s}^{\mathrm{c}}}{\mathrm{S}} \cdot \mathrm{WPAD}^{\mathrm{c}} \tag{12}$$

Regarding the general statistical measures, the correlation coefficient of Pearson *R*, the determination coefficient R^2, and the modeling efficiency (EF) were used to compare S_k^c and $\hat{S}_k^c$. The *R* coefficient is a measure of association among two variables while R^2 refers to how the variance of the dependent variable is explained by the independent variables and is used to measure the adjustment of a regression line. Thus, when R^2 is equal to 1, the estimated data are completely explained by the variance of real data. EF is a normalized measure to evaluate the model performance [6]. An EF indicator equal to 1 shows a total efficiency of the model, since there are complete information gains, while an indicator equal to 0 means the opposite. In cases where deviations between real and estimated data are high, this indicator may present negative values. These indicators were calculated as follows:

$$\mathrm{R} = \frac{\mathrm{cov}(S_k^c, \hat{S}_{\mathrm{k}}^{\mathrm{c}})}{\sqrt{\mathrm{var}(S_k^c) \cdot \mathrm{var}\left(\hat{S}_{\mathrm{k}}^{\mathrm{c}} \right)}} \tag{13}$$

$$R^2 = \left(\frac{\mathrm{cov}(S_k^c, \hat{S}_{\mathrm{k}}^{\mathrm{c}})}{\sqrt{\mathrm{var}(S_k^c) \cdot \mathrm{var}\left(\hat{S}_{\mathrm{k}}^{\mathrm{c}} \right)}} \right)^2 \tag{14}$$

$$EF = 1 - \frac{\sum \left(S_k^c - \hat{S}_k^c \right)^2}{\sum \left(S_k^c - \overline{S}_k^c \right)^2} \tag{15}$$

where S_k^c is the observed value; $\hat{S}_k^c$ is the model result, and $\overline{S}_k^c$ is the average of the S_k^c values.

3. Data and Application Scenarios

The Algarve region in the south of Portugal was selected to implement the proposed approach in order to disaggregate the data to a kilometric grid. (Figure 2). This region was selected to implement this study due to recent dynamics regarding permanent crops, namely, regarding irrigated ones (such as citrus) and the necessity of policy evaluation. Algarve is a region with an area of 4996.8 km^2 and, in 2010, was composed of 16 municipalities and 84 parishes, which was reduced to 67 a few years later. The Mediterranean climate predominates and there are several biophysical contrasts between the coastal and inland areas with less fertile areas and higher slopes.
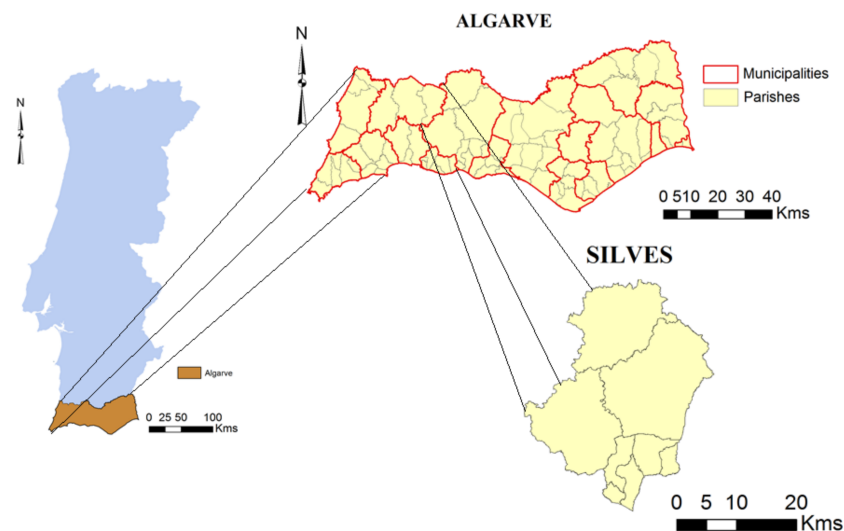
Figure 2. The study area.

There are municipalities where permanent crops are predominant and citrus areas are relevant. This is the case of Silves, which was chosen as the pilot municipality to disaggregate the data to a more detailed 25-hectare grid. This municipality covers an area of about 680.1 km² and, in 2009, was divided into 8 parishes which were later reduced to 6. It extends from the inland Algarve to the coast and in 2009, the agrarian census represented more than 17% of the permanent crop area and about 41% of the citrus area in the region.

In the Algarve region and Silves municipality, the predominant land-use is permanent crops, which were disaggregated as follows: fresh fruits, citrus, nuts, olive trees, vineyards, and other permanent crops.

The "training fields" (that is, the sample areas for defining the spectral signatures) for the supervised classification were defined using the 2012 LUCAS survey and the knowledge of experts from the Ministry of Agriculture, having been considered from a total of 191 sample areas in the whole Algarve region. For defining the training fields to implement the supervised classifications, the LUCAS survey may be a source of information, but if these observations are limited in number, the empirical knowledge of the area by technicians may be inserted. In Portugal, this empirical knowledge was easily obtained in the different regions to define the training fields and to carry out the supervised classifications. This approach has, therefore, the potential to be implemented in other areas if there is information on the satellite imagery. Regarding the satellite imagery, the LANDSAT 5 and 8 images with 30-m resolution were used in the supervised classification process. The LANDSAT 5 image used to disaggregate the 2009 data is from 4 July 2009 while the LANSAT 8 image used for Silves is from 29 June 2013. Both images used were obtained from the NASA system by using the Earth Explorer: http://earthexplorer.usgs.gov.

For the Algarve region, two simulations of the entropy model using the LANDSAT 5 2009 image were developed. One (SCMD2009) considers for the supervised classification (SC) the minimum distance algorithm (MD) and another (SCML2009) uses the maximum likelihood algorithm (ML). In the case of the Silves municipality, both simulations were also considered, but they were tested in the entropy model with historical restrictions of land use (SCMD2009 and SCML2009) and without historical restrictions (simulations SCMD2009WR and SCML2009WR). Besides these four simulations, an SC using the minimum distance algorithm and the LANDSAT 8- 2013 image was tested considering the entropy model with and without historical restrictions (simulations SCMD2013 and SCMD2013WR). The historical restrictions include general incomplete limits indicated by experts (that may not be available to all region) and limits regarding crops evolution.

Most approaches developed in recent years were applied at a scale comparable to a 1 × 1 km grid [4]. Thus, the entropy model was used to disaggregate variables to the pixel level using a kilometric grid, which allowed obtaining a total of 6832 disaggregated units for the Algarve region. In the case of the Silves municipality, the data were disaggregated considering a 25-hectare grid with a total of 3148 disaggregated units.

For the error definition in the entropy model, the three-sigma rule was used [8,18,19,22]. The error vectors considered were $v = \{-0.5, 0, 0.5\}$ and $v = \{-1, 0, 1\}$ in the Algarve region and $v = \{-0.5, 0, 0.5\}$ in the Silves municipality.

Technical implementation of the approach used QGIS and the Semi-Automatic Classification Plugin (SCP) [23], and the entropy models were implemented using the General Algebraic Modelling System (GAMS).

4. Results and Analysis

In a first step, a prior estimate of land-use was calculated using the supervised classification. Figure 3 presents examples of the results for 2009 using the LANDSAT 5 image in the Algarve region and the minimum distance algorithm and the maximum likelihood algorithm. Despite the differences between the two classification algorithms presented, some contrasts in the Region can be observed, namely, between the more forested areas in inland Algarve and the coastal areas with different uses. We also concluded that the minimum distance algorithm tends to identify agricultural areas in inland Algarve, which are identified as forestry areas with the maximum likelihood algorithm. However, according to the knowledge of experts from the Ministry of Agriculture in those areas, the Maximum Likelihood Algorithm tends to provide results more coherent with the observed reality.
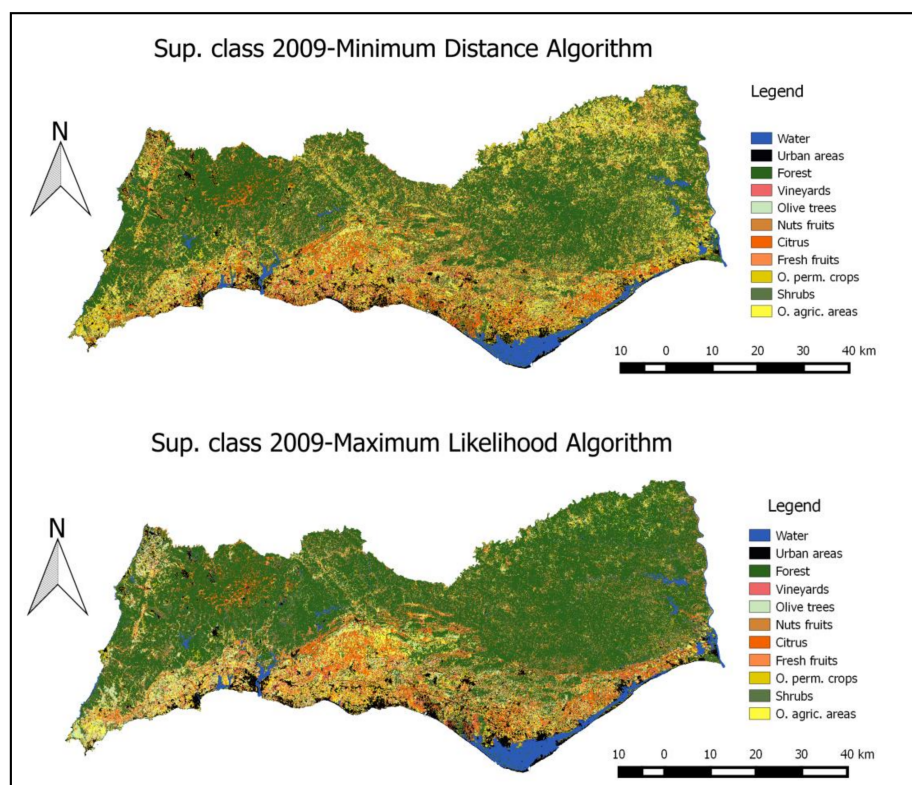

Figure 3. The supervised classification results for 2009 in the Algarve region; (source: model results).

These results are presented at a 30 m × 30 m pixel level and are aggregated in a grid at a kilometric level. Examples of some "pixel" level estimates using the kilometric grid are presented in Table 1.

Table 1. The examples of the prior estimates *i* at a pixel level using the Maximum Likelihood algorithm in the Algarve region in 2009.

Territorial Unit *i*	Nuts	Vineyards	Olive Trees	Citrus	Fresh Fruits	Permanent Crops	Other Areas
i1003	0.201	0.001	0.057	0.007	0.025	0.029	0.680
i1004	0.050	0.005	0.034	0.043	0.007	0.030	0.832
i1005	0.135	0.002	0.096	0.005	0.009	0.053	0.701
i1006	0.018	0.000	0.000	0.000	0.000	0.000	0.982
i1007	0.080	0.000	0.044	0.005	0.000	0.015	0.857
i1008	0.000	0.000	0.069	0.008	0.000	0.000	0.922
i1009	0.058	0.000	0.034	0.004	0.005	0.023	0.877
i1010	0.097	0.002	0.017	0.011	0.000	0.040	0.832
i1011	0.055	0.000	0.007	0.042	0.001	0.005	0.890
i1012	0.101	0.000	0.037	0.019	0.002	0.009	0.834
i1013	0.039	0.003	0.039	0.007	0.006	0.025	0.881
i1014	0.067	0.012	0.010	0.024	0.006	0.015	0.866
i1015	0.000	0.000	0.150	0.158	0.005	0.143	0.544
i1016	0.038	0.000	0.006	0.000	0.000	0.038	0.918
i1017	0.021	0.013	0.035	0.031	0.000	0.068	0.832

Source: model results.

In the second step of our approach, the data disaggregation process was carried out by applying the cross-entropy model in the Algarve region and Silves municipality. Figure 4 presents examples of results per disaggregated unit according to the algorithms tested for the Algarve region. They allow us to identify some of the major contrasts in the regional distribution of permanent crops. Several differences in allocation according to the classification algorithm used are also seen. The use of different algorithms allows providing the decision maker with different approaches which result in different spatial patterns that will be validated and analyzed.
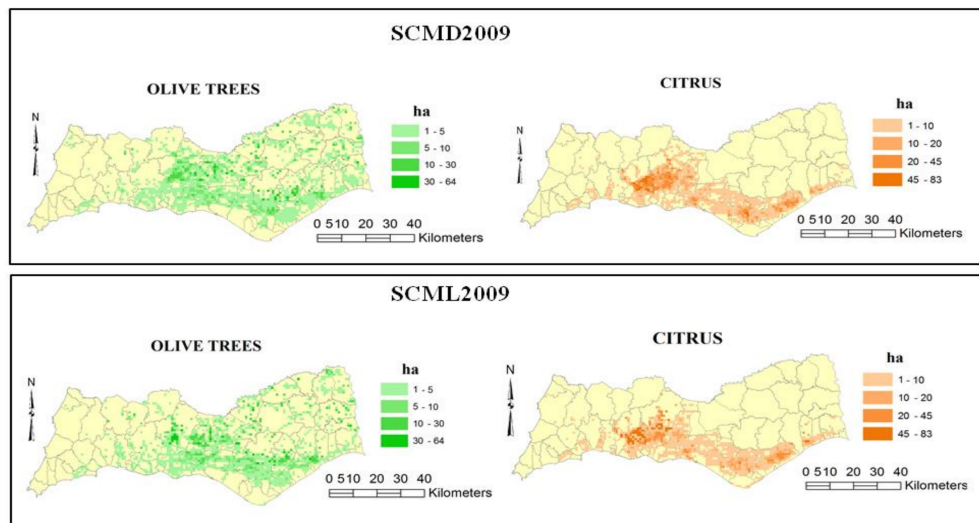

Figure 4. The examples of disaggregated units in the Algarve Region; source: model results.

For the Silves municipality, Figure 5 presents examples of the results for several distinct simulations using a finer grid (25 hectares). These different simulations are relevant to test the methodological approach at a more detailed pixel level. Spatially, the results let us identify some of the major contrasts in land use distribution, providing a more detailed "picture", with differences according to the simulations considered. For instance, the areas with the highest concentration of citrus are located in the parishes of S. B. Messines, Algoz, and Silves, while in the inner areas, they do not exist due to the unsuitable biophysical conditions. This is consistent with the knowledge held in this area.
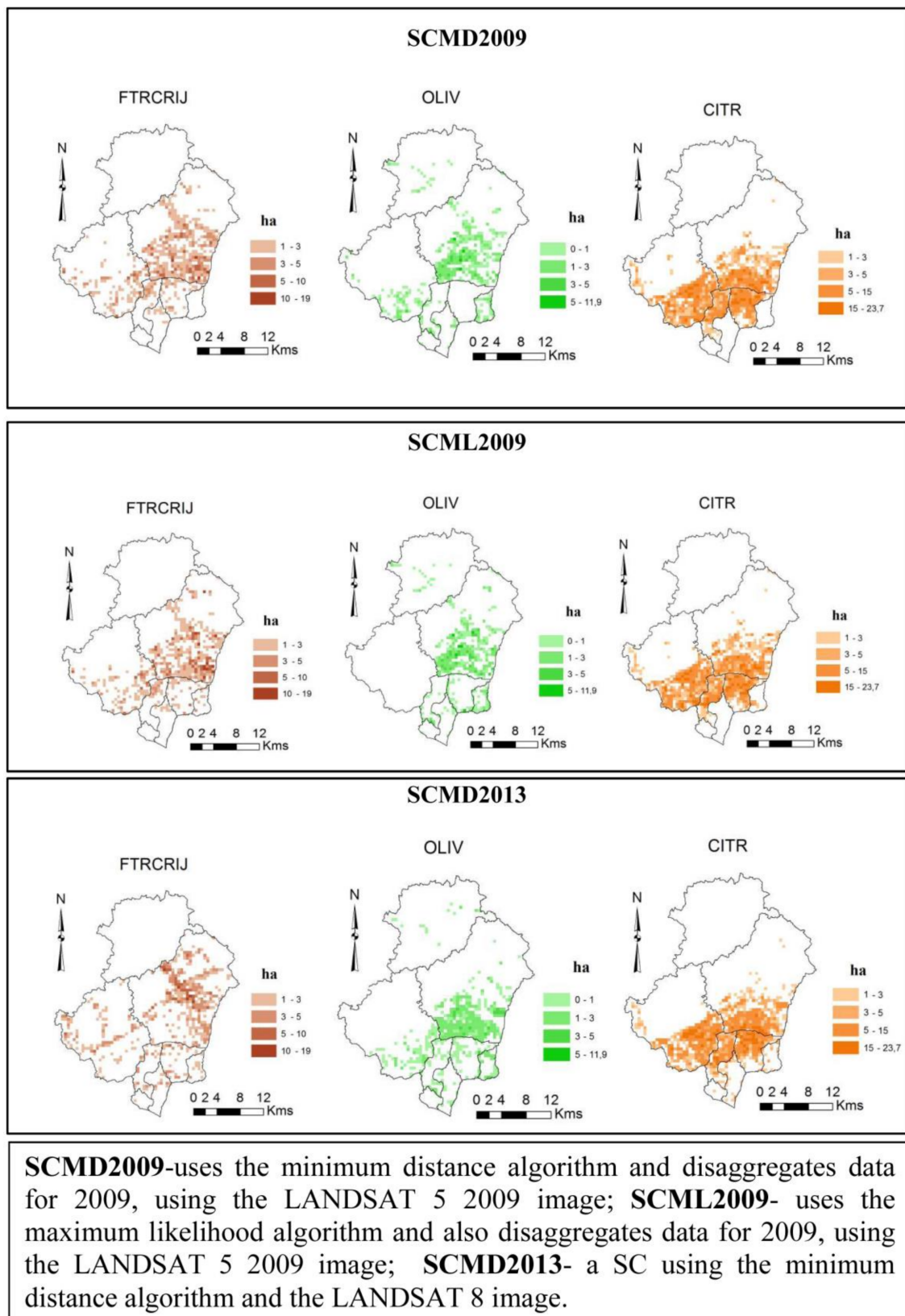
Figure 5. The examples of the data disaggregation results in the Silves municipality; source: model results.

The results of the cross-entropy model were validated using the deviation indicators mentioned before. The average and median PAD indicators are presented in Table 2 per crop type for the Algarve region and the Silves municipality. In general, the average and median PAD values are high, which may mean a weak consistency between the model results and observed statistical data. In the Algarve region, the lowest median values are obtained for olive trees and for other permanent crops. At the aggregate level, the WPAD indicator is 42.8% in simulation SCMD2009 and 41.1% in simulation SCML2009 (Table 3).

Table 2. The average and median PAD indicator per crop type and simulation in the Algarve region and Silves municipality.

Scenario		Nuts	Vineyards	Olive Trees	Citrus	Fresh Fruits	Other Permanent Crops
				Algarve Region			
SCMD2009	Average	87.7	87.1	110.6	77.8	90.4	26.1
	Median	53.8	44.6	31.3	56.5	72.6	0.0
SCML2009	Average	77.8	84.0	114.1	90.6	89.6	27.2
	Median	47.8	46.4	35.3	52.1	67.6	0.0
				Silves Municipality			
SCMD2009	Average	51.3	91.0	65.9	44.7	69.9	31.9
	Median	43.6	93.1	21.9	34.9	87.6	5.0
SCML2009	Average	51.6	91.0	62.4	78.4	82.6	31.9
	Median	50.1	93.1	19.0	32.6	93.6	5.0
SCMD2009WR	Average	65.3	70.9	201.6	1217.4	50.4	259.8
	Median	47.2	38.6	61.0	222.9	46.9	4.8
SCML2009WR	Average	61.2	75.2	217.6	1290.4	53.1	274.4
	Median	42.7	50.4	58.9	230.6	49.0	6.8
SCMD2013	Average	50.1	50.6	57.7	184.5	47.4	21.8
	Median	43.2	33.6	19.0	38.3	45.2	5.0
SCMD2013WR	Average	50.3	75.9	50.3	84.4	69.3	16.2
	Median	50.1	91.0	57.7	44.7	69.9	15.0

Source: model results.

For the Silves municipality, the PAD values are better than in the Algarve region and can be compared with those of previous studies [18,19,41,42]. In terms of results per crop, other permanent crops tend to present the lowest median values, but several crops have high median values, often reaching more than 50%. In aggregate terms, the lowest WPAD (20.86%) is recorded in simulation SCMD2013 (see Table 3). This result may be explained due to the more precise set of bands in the LANDSAT 8 image than in the LANDSAT 8 image. For 2009, SCMD2009 is the simulation that presents the best results (25.6%). Only the simulations without historical restrictions (SCMD2009WR, SCML2009WR, and SCMD2013WR) present WPAD values higher than 30%.

Table 3. The WPAD in the Algarve region Silves municipality.

Simulation	WPAD (%)
Algarve Region	
SCMD2009	42.80
SCML2009	41.10
Silves Municipality	
SCMD2009	25.60
SCML2009	26.12
SCMD2009WR	34.99
SCML2009WR	35.87
SCMD2013	20.86
SCMD2013WR	33.38

Source: model results.

Despite the PAD results tending to present some high values, we must highlight that they are summary results and there are territorial units with heterogeneous areas. Therefore, if we analyze the individual results of the parishes of Algoz, Silves, Alcantarilha, and S. B. Messines, which have the most relevant area of permanent crops (about 92%), we find that the PAD values are low in several relevant crops and hence a WPAD[i] lower than 30% can be observed in all simulations. The errors in other crops are of little importance and parishes with little relevance for the total area. Thus, the above values hide very satisfactory WPAD results.

The correlation coefficient (R) and the determination coefficient (R^2) are presented in Table 4.

Table 4. The correlation and determination coefficients R and R^2 in the Algarve region and Silves municipality.

Simulations		Nuts	Vineyards	Olive Trees	Citrus	Fresh Fruits	Other Permanent Crops
				Algarve Region			
SCMD2009	R	0.838	0.716	0.871	0.960	0.656	0.741
	R^2	0.703	0.513	0.758	0.922	0.430	0.549
SCML2009	R	0.896	0.732	0.863	0.946	0.643	0.743
	R^2	0.802	0.536	0.744	0.896	0.414	0.552
				Silves Municipality			
SCMD2009	R	0.937	0.863	0.985	0.992	0.686	0.873
	R^2	0.877	0.744	0.969	0.983	0.471	0.762
SCML2009	R	0.938	0.863	0.980	0.989	0.455	0.873
	R^2	0.880	0.744	0.960	0.979	0.207	0.762
SCMD2009WR	R	0.909	0.771	0.769	0.985	0.525	0.324
	R^2	0.827	0.594	0.591	0.971	0.276	0.105
SCML2009WR	R	0.923	0.709	0.705	0.988	0.484	0.262
	R^2	0.852	0.503	0.497	0.977	0.234	0.069
SCMD2013	R	0.938	0.766	0.991	0.995	0.760	0.963
	R^2	0.880	0.587	0.982	0.991	0.577	0.928
SCMD2013WR	R	0.780	0.743	0.695	0.994	0.621	0.423
	R^2	0.608	0.552	0.483	0.988	0.385	0.179

Source: model results.

In the Algarve region, all the R^2 indicators are above 0.5, except for fresh fruits, which present R^2 values of 43% in simulation SCMD2009 and of 41.4% in simulation SCML2009. In some crops, such as olive trees and citrus, the R^2 values are always higher than 0.7.

In Silves, the results are considerably better than those presented in previous studies since they are higher than 0.5 for most crops in all simulations and, in several cases, are higher than 0.8 or even 0.9. In the case of citrus, the most relevant permanent crop in the Silves municipality, the results are always above 0.9.

A similar validation process was implemented in Brazil in Reference [5] considering all the country's municipalities (more than 4000) and correlation coefficients between 0.4 and 0.65 were obtained. Other authors (from Reference [6]) validated their model using 4 crops and obtained for one crop an R^2 of 0.8 while the others presented values of R^2 between 0.40 and 0.45.

Finally, the results were tested using the Efficiency Indicator (EF), as shown in Table 5 for the Algarve region and Silves municipality.

In the case of the Algarve region, the EF is always higher than 0.45, except for fresh fruits. Citrus and olive trees tend to reveal results always above 0.7. Both algorithms (minimum distance algorithm–SCMD2009 and maximum likelihood algorithm–SCML2009) reveal similar results regarding the EF.

Table 5. The efficiency indicator in the Algarve region and Silves municipality.

Scenario	Nuts	Vineyards	Olive Trees	Citrus	Fresh Fruits	Other Permanent Crops
			Algarve Region			
SCMD2009	0.614	0.455	0.724	0.885	0.369	0.489
SCML2009	0.758	0.493	0.707	0.811	0.362	0.492
			Silves Municipality			
SCMD2009	0.544	−2.172	0.967	0.983	0.095	0.476
SCML2009	0.551	−2.172	0.954	0.979	−0.573	0.476
SCMD2009WR	0.782	0.541	0.440	0.915	0.019	0.088
SCML2009WR	0.822	0.373	0.375	0.910	−0.039	0.065
SCMD2013	0.552	0.285	0.960	0.988	0.535	0.900
SCMD2013WR	0.387	0.067	0.430	0.972	0.343	0.165

Source: model results.

For the Silves municipality, vineyard land-use present negative EF values and therefore, a null disaggregation efficiency of results in these cases. Despite using a good number of training fields, the best EF in vineyards is 0.541. Previous studies were also less successful in estimating vineyard areas due to the low number of training fields [22]. However, this study presented a detailed number of training fields and revealed low EF values in several cases. This may be due to the classes used in the supervised classification process, which requires a better revision of the macro-classes considered. Fresh fruits also tend to present EF values lower than 0.3 in most simulation. The reason for this is the fact that fresh fruits include a diversity of crops and it will also require a better revision of the macro-classes considered.

On the other hand, citrus always presents an EF above 0.9. All the other crops present EF values above 0.5 in most simulations. These results are not much different from those of previous studies [6], which obtained EF values between 0.23 and 0.71, and only one crop presented a value higher than 0.71.

The proposed approach allows disaggregating agricultural data at a detailed level, being relevant for agricultural economics analysis. As with other crop mapping studies, the quality of maps depends on the quality of data sources [7]. From an economic point of view, knowing land uses will allow for the identification of total yields and the economic output of agricultural areas. This model is also well suited when the data variation is great, such as the case of Mediterranean regions where crop acreage, farm economics, and production technologies have a high variability. The model's results can also be aggregated (up-scaled) into different spatial units, such as agro-ecological zones, providing another framework for analysis [6]. The model results allow for the revealing of contrasts related to different farms' strategies and different biophysical conditions and it provides information on the location of crops in the territory but does not differentiate them according to the productive system.

Therefore, the proposed model offers an effective way to disaggregate data at a detailed level, but further research is necessary to improve previous estimates and integrate the different layers of information. Also, changing the crop patterns over time is important as crop patterns change over space [7] and there must be efforts to estimate land-use continuously over time [17–19]. One line of research will be to test other classification algorithms. This research focused on two known ones to provide a detailed experience of the feasibility of these approaches because they are very known and used widely. Nevertheless, it may be tested further using other classification algorithms, such as the random forest algorithm.

5. Concluding Remarks

This study presented a methodological approach for disaggregating agricultural data at the pixel level, which may be useful to planning land-use, monitoring policy, and strategies of rural development. This approach is based on the supervised classifications to identify the distribution of land use and improve it through an entropy model. The study showed that full advantage of up-to-date satellite imagery can be made use of. These satellite images in combination with the LUCAS

survey and empirical knowledge allow for the development of more precise supervised classifications. The use an entropy model guarantees consistency among the different sources of information and allows for the correction of existing errors in a supervised classification. Therefore, this paper proposes a good alternative to the traditional econometric approaches to disaggregate data at a detailed level and recovery incomplete information. Further research is being made to improve the model results, such as the development of an approach to disaggregate yearly agricultural data at the pixel level and its implementation in more complex areas is being designed.

Author Contributions: A.X. conducted the bibliographic review, developed and implemented the model. R.F. contributed to the development of the model and analysed the results and reviewed the paper. M.d.B.C.F. helped conducting the bibliographic review, developing the model and the approach. M.d.S.R. and F.V. analysed the model results and reviewed the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Hajkowicz, S.; Collins, E.; Cattaneo, E. Review of Agri-Environment Indexes and Stewardship Payments. *Environ. Manag.* **2009**, *43*, 221–236. [CrossRef] [PubMed]
2. Fritz, S.; You, L.; Bun, A.; See, L.; McCallum, I.; Schill, C.; Perger, C.; Liu, J.; Hansen, M.; Obersteiner, M. Cropland for sub-Saharan Africa: A synergistic approach using five land cover data sets. *Geophys. Res. Lett.* **2011**, *38*. [CrossRef]
3. Tan, J.; Yang, P.; Liu, Z.; Wu, W.; Zhang, L.; Li, Z. Spatio-temporal dynamics of maize cropping system in Northeast China between 1980 and 2010 by using spatial production allocation model. *J. Geogr. Sci.* **2014**, *24*, 397–410. [CrossRef]
4. Kempen, M.; Heckelei, T.; Britz, W.; Leip, A.; Koeble, R.; Marchi, G. *Computation of a European Agricultural Land Use Map–Statistical Approach and Validation*; Discussion Paper; Institute for Food and Resource Economics: Bonn, Germany, 2005.
5. You, L.; Wood, S. An entropy approach to spatial disaggregation of agricultural production. *Agric. Syst.* **2006**, *90*, 29–347. [CrossRef]
6. You, L.; Wood, S.; Wood-Sichra, U. Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach. *Agric. Syst.* **2009**, *99*, 126–140. [CrossRef]
7. You, L.; Wood, S.; Wood-Sichra, U.; Wu, W. Generating global crop distribution maps: From census to grid. *Agric. Syst.* **2014**, *127*, 53–60. [CrossRef]
8. Chakir, R. Spatial downscaling of agricultural land use data: An econometric approach using cross–entropy. *Land Econ.* **2009**, *85*, 238–251. [CrossRef]
9. Xavier, A.; Costa Freitas, M.D.B.; Fragoso, R. Disaggregation of Statistical Livestock Data Using the Entropy Approach. *Adv. Oper. Res.* **2014**, 397675. [CrossRef]
10. EUROSTAT. *LUCAS 2012 (Land Use/Cover Area Frame Survey)*; EUROSTAT: Brussels, Belgium, 2013.
11. Chakir, R.; Lungarska, A. Agricultural rent in land-use models: Comparison of frequently used proxies. *Spatial Econ. Anal.* **2017**, *12*, 279–303. [CrossRef]
12. Chakir, R.; Le Gallo, J. Predicting land use allocation in France: A spatial panel data analysis. *Ecol. Econ.* **2013**, *92*, 114–125. [CrossRef]
13. Chakir, R.; Parent, O. Determinants of land use changes: A spatial multinomial probit approach. *Pap. Reg. Sci.* **2009**, *88*, 327–344. [CrossRef]
14. Ferdous, N.; Bhat, C.R. A spatial panel ordered-response model with application to the analysis of urban land-use development intensity patterns. *J. Geogr. Syst.* **2013**, *15*, 1–29. [CrossRef]
15. Anselin, L. Spatial econometrics in RSUE: Retrospect and prospect. *Reg. Sci. Urban Econ.* **2007**, *37*, 450–456. [CrossRef]

16.   Brady, M.; Irwin, E. Accounting for spatial effects in economic models of land use: Recent developments and challenges ahead. *Environ. Resour. Econ.* **2011**, *48*, 487–509. [CrossRef]

17.   Howitt, R.; Reynaud, A.A. Spatial disaggregation of agricultural production data using maximum entropy. *Eur. Rev. Agric. Econ.* **2003**, *30*, 359–387. [CrossRef]

18.   Fragoso, R.; Martins, M.B.; Lucas, M.R. Generate disaggregated soil allocation data using a Minimum Cross Entropy Model. *WSEAS Trans. Environ. Dev.* **2008**, *9*, 756–766.

19.   Martins, M.B.; Fragoso, R.; Xavier, A. Spatial disaggregation of agricultural data in Castelo de Vide, Alentejo, Portugal: An approach based on maximum entropy. *JP J. Biostat.* **2011**, *5*, 1–16.

20.   Louhichi, K.; Jacquet, F.; Butault, J.P. Estimating input allocation from heterogeneous data sources: A comparison of alternative estimation approaches. *Agric. Econ. Rev.* **2012**, *13*, 83–102.

21.   Britz, W.; Verburg, P.H.; Leip, A. Modelling of land cover and agricultural change in Europe: Combining the CLUE and CAPRI-Spat approaches. *Agric. Ecosyst. Environ.* **2011**, *142*, 40–50. [CrossRef]

22.   Xavier, A.; Freitas, M.B.; Fragoso, R.; Socorro Rosário, M. Agricultural data disaggregation at a local level: An approach using entropy and supervised classifications. In Proceedings of the 1st International Congress on Interdisciplinarity in Social and Human Sciences, Faro, Portugal, 5–6 May 2016; CIEO: Faro, Portugal, 2016.

23.   Congedo, L. Semi-Automatic Classification Plugin Documentation Release 4.8.0.1. 2015. Available online: https://semiautomaticclassificationmanual-v4.readthedocs.org/en/latest/ (accessed on 15 February 2016).

24.   Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [CrossRef]

25.   Lu, D.; Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870. [CrossRef]

26.   Xie, Y.; Sha, Z.; Yu, M. Remote sensing imagery in vegetation mapping: A review. *J. Plant Ecol.* **2008**, *1*, 9–23. [CrossRef]

27.   Perumal, K.; Bhaskaran, R. Supervised classification performance of multispectral images. *J. Comput.* **2010**, *2*, 2151–9617.

28.   Samaniego, L.; Schulz, K. Supervised classification of agricultural land cover using a modified k-NN technique (MNN) and landsat remote sensing imagery. *Remote Sens.* **2009**, *1*, 875–895. [CrossRef]

29.   Bahadur, K.C. Improving Landsat and IRS image classification: Evaluation of unsupervised and supervised classification through band ratios and DEM in a mountainous landscape in Nepal. *Remote Sens.* **2009**, *1*, 1257–1272. [CrossRef]

30.   Rozenstein, O.; Karnieli, A. Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Appl. Geogr.* **2011**, *31*, 533–544. [CrossRef]

31.   Shalaby, A.; Tateishi, R. Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the Northwestern coastal zone of Egypt. *Appl. Geogr.* **2007**, *27*, 28–41. [CrossRef]

32.   Jeon, Y.J.; Choi, J.G.; Kim, J.I. A study on supervised classification of remote sensing satellite image by bayesian algorithm using average fuzzy intracluster distance. In *Combinatorial Image Analysis*; Klette, R., Žunić, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 597–606. ISBN 978-3-540-30503-3.

33.   Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

34.   Jaynes, E.T. Information theory and statistical methods I. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]

35.   Good, I. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Stat.* **1963**, *34*, 911–934. [CrossRef]

36.   Golan, A.; Judge, G.; Miller, D. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*; John Wiley & Sons: New York, NY, USA, 1996; ISBN 978-0-471-95311-1.

37.   Lence, H.L.; Miller, D. Estimation of Multi-Output Production Functions with Incomplete Data: A Generalized Cross Entropy Approach. *Eur. Rev. Agric. Econ.* **1998**, *25*, 188–209. [CrossRef]

38.   Zhang, X.; Fan, S. Estimating crop-specific production technologies in Chinese agriculture: A generalized maximum entropy approach. *Am. J. Agric. Econ.* **2011**, *83*, 378–388. [CrossRef]

39.   Howitt, R.E.; Msangi, S. Entropy estimation of disaggregate production functions: An application to northern Mexico. *Entropy* **2014**, *16*, 1349–1364. [CrossRef]

40.   Aurbacher, J.; Dabbert, S. Generating crop sequences in land-use models using maximum entropy and Markov chains. *Agric. Syst.* **2011**, *104*, 470–479. [CrossRef]

41. Xavier, A.; Martins, M.B.; Fragoso, R. A mininum cross entropy model to generate disaggregated data at the local level. In Proceedings of the 122nd EAAE Seminar "Evidence-based agricultural and rural policy making: Methodological and empirical challenges of policy evaluation", Ancona, Italy, 17–18 February 2011.
42. Fragoso, R.M.; Carvalho, M.L. Estimation of joint costs allocation coefficients using the maximum entropy: A case of Mediterranean farms. *J. Quant. Econ.* **2012**, *10*, 91–111.
43. Fragoso, R.; Carvalho, M.L.D.S. Estimation of cost allocation coefficients at the farm level using an entropy approach. *J. Appl. Stat.* **2013**, *40*, 1893–1906. [CrossRef]