Generating high-resolution national crop distribution maps: Combining statistics, gridded data and surveys using an optimization approach

M. Van Dijk[1]; L. You[2]; P. Havlik[3]; A. Palazzo[3]; A. Mosnier[3]

1: Wageningen Economic Research, International Policy, Netherlands, 2: International Food Policy Research Institute, , United States of America, 3: IIASA, , Austria

Corresponding author email: michiel.vandijk@wur.nl

Abstract:

Detailed information on the location and size of crop area is essential for the assessment of agricultural production, food security and emissions resulting from land use change. Although, there exist several initiatives to produced spatially explicit crop distribution maps, these are generally too coarse for detailed country assessments, which require high resolution spatial maps. Using Malawi as a case-study and building on the Spatial Production Allocation Model (You et al. 2014), this paper presents an approach to produce high resolution crop distribution maps that incorporate all available information, including sub-national agricultural statistics, crop specific land use information and national irrigation surveys.

Acknowledegment:

JEL Codes: C61, Q13

#2058

Generating high-resolution national crop distribution maps: Combining statistics, gridded data and surveys using an optimization approach

January 15, 2018

Abstract

Detailed information on the location and size of crop area is essential for the assessment of agricultural production, food security and emissions resulting from land use change. Although, there exist several initiatives to produced spatially explicit crop distribution maps, these are generally too coarse for detailed country assessments, which require high resolution spatial maps. Using Malawi as a case-study and building on the Spatial Production Allocation Model (You et al. 2014), this paper presents an approach to produce high resolution crop distribution maps that incorporate all available information, including sub-national agricultural statistics, crop specific land use information and national irrigation surveys.

JEL: Q15, C61, Q13

Introduction

Information on the location and size of crop area is essential to assess future agricultural growth, biodiversity loss, food security and emissions from land use change. Not surprisingly, land cover and land use maps are key input for global and national agricultural models that address these issues. Unfortunately, for most countries and in particular for developing countries, detailed agricultural statistics, such as crop-specific area, output and yield are often not available or only presented at highly aggregate levels (e.g country, district or province). Only in very rare cases, data is available in some sort of spatial format and, if so, it is nearly always very coarse, which makes it unsuitable for land use modelling.

There have been several initiatives to produce global and spatially explicit land use maps for a large number of crops (Monfreda, Ramankutty, and Foley 2008, Portmann, Siebert, and Döll (2010), You et al. (2014)). Most these products use spatial information on land cover, suitability and irrigation to downscale (sub)national agricultural statistics to the grid level. A comparison revealed that crop-specific harvested area and yield can differ considerably and mostly result from variation in input data, in particular crop masks and downscaling methodologies (Anderson et al. 2015).

A major advantage of the above mentioned datasets is that they have global coverage and therefore are often used by global integrated assessment models as base year input to model future land use change (Leclère et al. 2014). On the other hand, due to their broad coverage, the global downscaling approaches mainly rely on coarse land cover masks and are restricted in their use of country specific data and expert input.[1] Although most initiatives use subnational statistics, none of them incorporate other types of country specific agricultural information that can improve the crop allocation process. Particularly interesting are (nationally representative) household and farm surveys that contain geo-coded information on the location of crop production. These surveys are increasingly available for a number of African countries (Carletto, Jolliffe, and Banerjee 2015).

Due to the lack of detail, global land use maps are useful, or can serve at best, as a starting point for national or more detailed land use assessments. This type of studies zoom in on subnational regions and often involve a process of participatory scenario development in which stakeholders are asked to validate land use input data (Rutten et al. 2014). For this type of assessments a flexible approach is needed that creates crop

[1] The aim is to create the crop distributions maps for 2 periods: 2000 and 2010. This paper presents preliminary results for 2000 only.

distribution maps at high resolution and is able to incorporate new information from expert, such as the location of plantations, updated land cover and land use maps and the location of irrigated areas.

An interesting new approach to create high resolution crop distribution maps are machine learning techniques, which can be used to identify crops on satellite imagery. Although promising, these techniques are still under development and at the moment it is only possible to create land use maps for crops such as soy bean, corn and palm oil, which can be relatively easily identified from satellite data by means of machine learning classification approaches (Zhong et al. 2016). Moreover, to train the models and increase the accuracy of the land classification, training data is needed that is often not easily available, in particular for developing countries.

In this paper, we present a flexible model to create high resolution crop distribution maps, which is able to incorporate all available and relevant information, including subnational agricultural statistics, high resolution land use information and detailed irrigation information. The model builds on the Spatial Production Allocation Model (SPAM) presented in You et al. (2006; 2009; 2014), which uses a cross entropy framework to allocate subnational land use information to 5 arcmin (~10x10 km) grid cells. We extent the model by increasing the resolution to 30 arc sec (~1x1 km) and add constraints that manage the allocation of additional spatially explicit information on crop area from satellite imagery as well as new irrigation map based on irrigation surveys and information from Open Street Map. Finally, similar to You and Wood (2006), we use additional statistics (in our case from household survey) to validate the results from the allocation procedure. The model is implemented and tested for Malawi using data for around 2000.[1^]

The structure of the paper is as follows. Section 2 describes the spatial allocation model that is used to create the crop distribution maps. Section 3 summarizes the various data sources that are used in the process. Section 4 presents the results, zooming in on key crops. In Section 5, the crop distribution maps are validated using a nationally representative household survey for Malawi. Finally, Section 5 concludes.

The spatial allocation model

Our spatial allocation model refines the most recent version of SPAM (You et al. 2014, Wood-Sichra, Joglekar, and You (2016)). SPAM uses an optimization approach to spatially allocate national and subnational agricultural statistics. To do this a cross-entropy objective function is minimized subject to a set of constraints that capture the available information on location of certain crops and crop systems (e.g. crop cover and irrigation maps) or the likelihood a crop is grown in a certain area (e.g. suitability maps and subnational land use statistics). Although previous work uses country specific data sources for subnational land use statistics, they mainly rely on global sources to define the land allocation constraints. In this study, we were able to collect more detailed information, which allows us to refine the allocation procedure. In particular, we obtained a recent high resolution land cover map for Malawi, which reveals the location of four crop groups. In addition, we created new crop specific irrigation map for Malawi. In this Section we present the adjusted SPAM model and describe the calculation of priors, which are needed to start the cross entropy optimization. To remain comparability with You et al. (2014), we use the same variable names, crop classification and crop systems (high-input irrigated, high-input rainfed, low-input rainfed and subsistence rainfed).

Cross entropy framework and constraints

Before the cross entropy framework can be implemented, the allocated crop area needs to be converted into a probability with value between 0 and 1. This is done as follows:

$$s_{ijl} = \frac{A_{ijl}}{CropArea_{jl}}$$

where, s_{ijl} is the area share allocated to grid cell identifier $i = 1, 2, 3, ...$ and crop identifier $j = maize, cassave, rice, ...$ at input level $l = irrigated, rainfed - highinput, rainfed - lowinputandsubsistence$.

A_{ijl} is the area allocated to grid cell i for crop j at input level l. $CropArea_{jl}$ is the total physical crop area in Malawi for crop j at input level l.

The spatial allocation model can then be defined as a non-linear optimization problem in which the error between prior information area shares ($\{\pi_{ij}$) and allocated area shares (s_{ijl}) is minimized, subject to a set of constraints. The objective function is defined as follows:

$$\min_{s_{ijl}} \quad CE(s_{ijl}, \pi_{ijl}) = \sum_i \sum_j \sum_l s_{ijl} \; lns_{ijl} - \sum_i \sum_j \sum_l s_{ijl} \; ln\pi_{ijl}$$

subject to:

(i) A constraint defining the range of permitted physical area shares:

$$0 \leq s_{ijl} \leq 1$$

(ii) A constraint, which ensures ensures that the sum of allocated physical area shares within grid cells sum to one. This ensures all physical crop area is allocated to grid cells.

$$\sum_i s_{ij} = 1 \qquad \forall j \quad \forall l$$

(iii) A constraint, which specifies that the sum of allocated physical area for each crop group, $c = rice, sugar\,cane, ...$, and production systems is lower or equal than the actual crop area ($Avail_{ij}$) that is available for each crop group c in grid cell i. This equation extends the one in You et al. (2014), by adding spatially explicit information on the location of specific crops in the country.

$$\sum_{j \in c} \sum_l CropArea_{jl} \times s_{ijl} \leq Avail_{ic} \qquad \forall i$$

(iv) A constraint, which ensures that the sum of allocated physical area over all production systems within a sub-national geopolitical unit,$k = 1, 2, 3, ...$ (in this case administrative zone 2 level) is equal to the physical area information ($SubCropArea_{jk}$) for commodities J for which information is available from sub-national land use statistics. Crops for which only national level information is available are not affected by this constraint.

$$\sum_{i \in k} \sum_l CropArea_{jl} \times s_{ijl} = SubCropArea_{jk} \qquad \forall k \quad \forall j \in J$$

(v) A constraint, which allocates spatially explicit information on irrigated areas ($IrrArea_{ijl}$) for irrigated crops (J) that belong to the irrigated crop system in grid cells I. In contrast to You et al. (2014), who do not have crop specific irrigation area, we have collected detailed and crop specific information on the location of irrigated areas, which we incorporate in the spatial allocation model.

$$CropArea_{jl} \times s_{ijl} = IrrArea_{ijl} \qquad \forall i \in I \quad \forall l = irrigated \quad \forall j \in N$$

Calculation of priors

To guide the optimization procedure, prior information on the location of crops is essential. The location of crop areas is determined by the combination of economic (e.g. market access and population density) and bio-physical factors (e.g suitability conditions). We broadly follow You et al. (2014) and calculate the priors in the following manner. For the High input-irrigated system, the priors can directly be derived from the new irrigation map we created for Malawi. For the High input-rainfed system, we assume that market access, approximated by travel time to the nearest town and population density as well as suitability drive the location decision. For the low input-rainfed and subsistence systems, we assume that farmers mainly grow crops for their own consumption and therefore the prior allocation is mainly determined by rural population density and suitability.

Data

We combine information from a number of different sources to run the model described above (see Table 1 for an overview). A key input for the analysis is a high resolution land cover map that provides detailed information on the location of agricultural production. Recently, several new land use and land cover maps have been produced for Malawi that cover the period 1990-2010 (see Haack, Mahabir, and Kerkering 2015 for a comparison). We decided to use the maps produced by the FAO (2013) that is available for 2010. The advantage of this map in comparison to others is the very high resolution of 30x30 meter as well as spatially explicit information on the location of four crop groups (rice, sugar cane, tea & coffee and other crops) in Malawi. We aggregated the map to a resolution of 30 sec and calculated the physical area share for each crop group per grid cell i.

Statistics on actual land use at the national level were taken from FAOSTAT and extended with data from Agro-maps, CountryStat and national statistics for the 26 regions at the administrative zone two level. We followed the approach of Wood-Sichra, Joglekar, and You (2016) and aggregate all crop statistics to a set of standardised crop classes. We compared and harmonized all agricultural statistics to create a consistent dataset in which subnational totals add up to the national total from FAOSTAT. We were able to collect subnational data for 12 out of the 30 crops that are cultivated in Malawi.

Similar to You et al. (2014), we use Global Agro-Ecological Zones (GAEZ) suitability maps, which can be mapped to the standardised crops (see Wood-Sichra, Joglekar, and You (2016)). Finally, we created a new irrigation map using detailed information from the National irrigation master plan and investment framework, which presents a geo-referenced inventory of irrigation systems in Malawi, including the size of the irrigated area and crop type, for the period around 2000-2010. Where needed we compared and extended the inventory with data from World Bank (2010), which provides similar but less detailed information, as well as data from Google Earth and Open Street Map. A comparison showed that the resulting area is similar to Siebert et al. (2013), which is used by You et al. (2014) as input layer. To create the final irrigation map, we compared the FAO land cover map with spatial irrigation data and, where possible, allocated the irrigation point data to grid cells with matching crops (i.e. coordinates of sugar cane areas to sugar cane grid cells). Finally, we combined the agricultural statistics data and irrigation data to estimate total crop area for each of the production systems. We decided to classify all data to three systems only and not to use the low input-rainfed system, for which we did not find any statistical information.

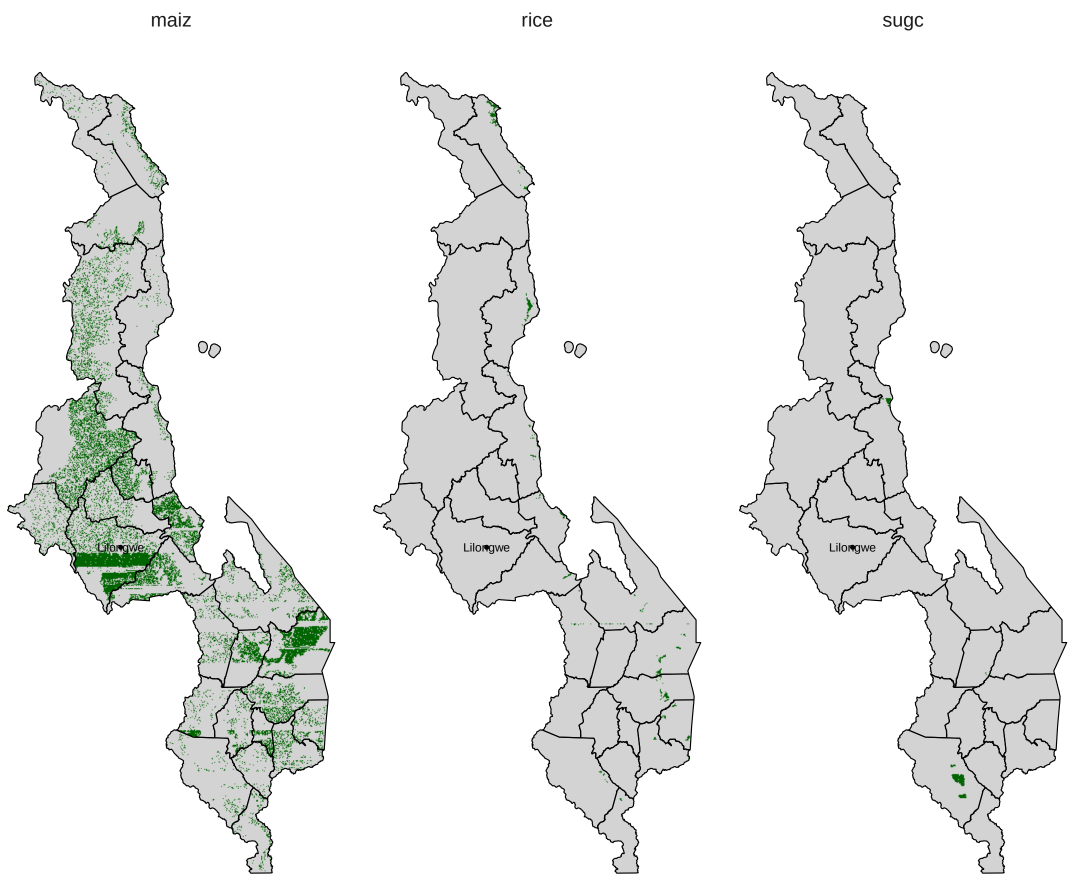
Table 1: Data sources

Data	Source
Malawi land cover map	FAO (2013)
National agricultural statistics	FAOSTAT
Subnational agricultural statistics	Agro-maps, CountryStat and national statistics
Irrigation survey	National irrigation master plan and investment framework, World Bank (2010), Google Earth and Open Street Map
Crop suitability	GAEZ
Travel time	Nelson (2010)
Population density	WorldPop
Validation data	LSMS-ISA

Results

In total 34 crop distribution maps are generated using the spatial allocation model: 28 Subsistence, 4 High input-irrigated and 2 High input-rainfed maps. Here we show only the maps for 3 crops as an illustration (Figure 1). Maize, which is only grown under subsistence conditions, is the largest food crop in Malawi and produced by the majority of smallholders. As depicted in the map, it is cultivated throughout the country. Rice is a much less important food crop but we include it here as we have detailed spatial information on its location. It is predominantly produced under subsistence conditions but we also found information that a small area can be considered as High input-irrigated production system. For convenience we pooled both results. Rice is mainly grown at the shore of lake Malawi and in the South of the country. Finally, the map for sugar cane is presented. Sugar cane is predominantly produced on plantations by large corporations under irrigated conditions (i.e High input-irrigated system). The map shows two key locations where sugar cane is grown, one in the centre of the Malawi and one in the South.

Figure 1: model results for selected crops



Validation

To assess the quality of spatial allocation procedure it is important to compare the results with information that is not used in the model. A number of approaches and data sources have been used to validate crop land maps in the literature. You and Wood (2006) use municipality level agricultural statistics to validate a land use map for Brazil, while S. Fritz et al. (2011) use high-resolution imagery from Google Earth to assess results. Pekkarinen, Reithmaier, and Strobl (2009) combine both approaches to validate a European forest map. As it is difficult to derive information on crop specific land use from Google Earth, we use the 2010 Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) to evaluate our model results. The LSMS-ISA is a nationally and regionally (administrative zone 2) representative household survey for Malawi. It specifically focuses on agriculture and contains information on the number of plots owned by farmers and the specific crops that are grown. A drawback of the LSMS-ISA is that it only presents the coordinates of the enumeration areas (e.g. communities) to which the farmers belong, not those of the fields itself where the crops are grown. Moreover, for privacy reasons, the locations of the enumeration areas have an offset of 2-10 km. To account for this, we aggregated the crop distribution maps to 5 arcmin before comparison. Finally, to make the data from both sources comparable, we mapped all the crop types listed in the LSMS-ISA to the SPAM crop classes.

Figure 2 compares the results of our model to the LSMS-ISA information on crop location for maize and rice. We also included information on the number of household per crop, which serves as a proxy for the importance of the crop. The figures show reasonable overlap between the model results and the survey for both rice and maize. For maize almost all LSMS-ISA locations are placed on the maize map, while for rice a substantial number of survey locations are located outside the rice map. The mismatch is particularly strong in the South of the country. This shows that the land cover map, which includes spatial information on the location of rice farmers is not completely accurate and introduces a bias in the crop distribution map.

Figure 2: Comparison between land use allocation and LSMS-ISA

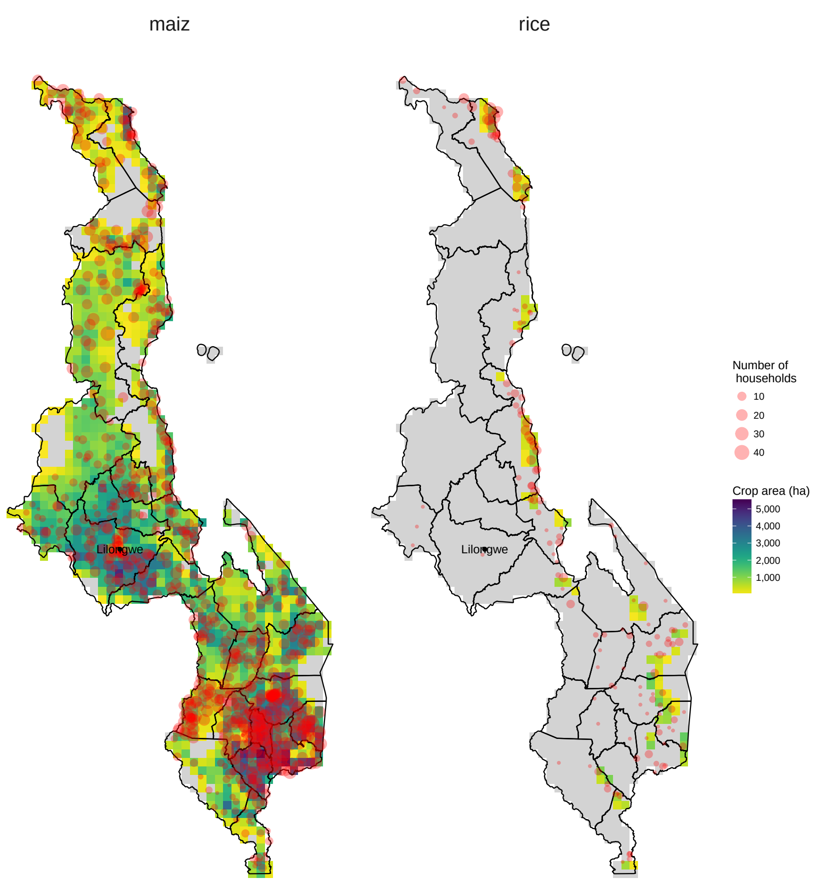


Table 2 provides information on the number of grid cells per crop from the crop allocation model. Note that in most areas, more than one crop is cultivated. It also shows similar information based on the LSMS-ISA. In order to compare both sources, the analysis is restricted to 460 out of 1,163 5 arcmin grid cells for which LSMS-ISA data is available. On the basis of this, we can calculate the share of cells per crop that are (in)correctly classified in comparison to the LSMS-ISA data, which can be considered as ground-truth information. The number of crops that are produced by farmers in a grid cell ranges from 2 to 15 different crops. This results in a total of 3,453 crop-gridcell combinations from the LSMS-ISA.

Table 2 shows that the overlap is surprisingly large, with a misclassification of only 4% on average. This is particularly striking taking into account that both sources cover different time periods (2000 for the crop distribution map and 2010 for the LSMS-ISA). Nonetheless, for some crops the overlap between the crop distribution maps and the LSMS-ISA is low, in particular for tea/coffee and sugar cane. We assumed that all production of these two cash crops can be regarded as High-input or High-irrigated production systems and allocated the crop area using information on the location of large plantations in Malawi from secondary sources. This finding indicates that at least some of the production of tea, coffee and sugar cane is undertaken by smallholders and our assumptions are not 100% accurate. In line with Figure 1, with 72% the misclassification for rice is substantial.

As mentioned above, the validation results suffers from a number of problems, most importantly the biased location of the survey data, which forces us to aggregate the crop distribution maps to a lower resolution. It is therefore not surprisingly to find a very high correspondence between the two sources of data. The result of the aggregation is that many crops occur in a large number of grid cells, while in reality production is much more location specific. This holds both for the data from the crop distribution maps and for the crop location from the LSMS-ISA survey. Hence, the results of the validation exercise are indicative only and have to be interpreted with care.

Table 2: Share of (in)correctly classified crops in comparison to the LSMS-ISA

short_name	Modelled land use (Number of cells)	LSMS land use (Number of cells)	Misclassified (%)
maiz	457	465	2
grou	384	391	2
trof	331	339	2
toba	234	239	2
bana	231	237	3
cass	222	228	3
vege	215	217	1
pige	193	194	1
bean	186	189	2
swpo	152	154	1
soyb	150	153	2
temf	141	142	1
sorg	131	133	2
mill	74	75	1
cott	67	70	4
chic	43	44	2
rice	32	113	72
pota	28	28	0
rest	24	25	4
sunf	8	8	0
teas_coff	3	6	50
sugc	0	3	100
total	3,306	3,453	4

Conclusions

This paper demonstrated an optimization approach to create high-resolution crop distribution maps at the national level. It extends the Spatial Allocation Model (SPAM) developed by You et al. (2014) by increasing the resolution of the maps and added flexibility to incorporate all relevant and available data that provides information on the location of crops in a country or region, including sub-national agricultural statistics, irrigation surveys and detailed spatially explicit information on the location of crops. The model was applied and tested for Malawi. In total 34 crop distribution maps were created, covering 28 crops and 3 production systems.

To validate the model spatially explicit data on the production of crops from a recent nationally representative household survey was compared with the crop distribution maps. Apart from a few crops, in particular rice, the results were surprisingly similar. Although, this is an encouraging outcome, the results have to be interpreted with care. Crop location coordinates in the LSMS-ISA are provided with a bias, which means the validation can only be done at a lower resolution. As many crops are grown throughout the country, this creates an upward bias in the overlap between the two sources of data. A proper validation demands detailed ground truth information for which the location is accurately known. One way forward to collect such data and improve the validation is to use information from a crowd-sourcing exercise (Lesiv et al. 2017).

The crop distribution maps presented in this paper can only be as good as the input data that goes into the spatial allocation model. Although, an effort has been made to clean and harmonise subnational agricultural statistics and create a new detailed irrigation map, there still might be a bias as a consequence of input data inaccuracies. Ultimately, the location of the crops is determined by the crop cover mask. Several papers have shown that land cover maps, even at high resolution, are still characterised by substantial inaccuracies (Lesiv et al. 2017) and may differ considerably, depending on the source and methodology used (Haack, Mahabir, and Kerkering 2015). One approach we aim to explore in the future to deal with this issue is to use the synergistic approach of S. Fritz et al. (2011) to create a new land cover map for Malawi that combines and harmonizes different sources.

Finally, to further test the model described in this paper, it would be interesting to create crop distribution maps for other Sub Saharan countries. In particularly interesting are countries like Zambia, Tanzania and Ethiopia for which similar nationally representative households surveys exists, which can be used for validation.

References

Anderson, Weston, Liangzhi You, Stanley Wood, Ulrike Wood-Sichra, and Wenbin Wu. 2015. "An analysis of methodological and spatial differences in global cropping systems models and maps." *Global Ecology and Biogeography* 24 (2): 180–91. doi:10.1111/geb.12243.

Carletto, Calogero, Dean Jolliffe, and Raka Banerjee. 2015. "From Tragedy to Renaissance: Improving Agricultural Data for Better Policies." *The Journal of Development Studies* 51 (2): 133–48. doi:10.1080/00220388.2014.968140.

FAO. 2013. *Atlas of Malawi Land Cover and Land cover Change: 1990-2010.* Rome: FAO.

Fritz, Steffen, Liangzhi You, Andriy Bun, Linda See, Ian McCallum, Christian Schill, Christoph Perger, Junguo Liu, Matt Hansen, and Michael Obersteiner. 2011. "Cropland for sub-Saharan Africa: A synergistic approach using five land cover data sets." *Geophysical Research Letters* 38 (4): n/a–n/a. doi:10.1029/2010GL046213.

Haack, Barry, Ron Mahabir, and John Kerkering. 2015. "Remote sensing-derived national land cover land use maps: a comparison for Malawi." *Geocarto International* 30 (3): 270–92. doi:10.1080/10106049.2014.952355.

Leclère, D, P Havlík, S Fuss, E Schmid, A Mosnier, B Walsh, H Valin, M Herrero, N Khabarov, and M Obersteiner. 2014. "Climate change induced transformations of agricultural systems: insights from a global model." *Environmental Research Letters* 9 (12). IOP Publishing: 124018. doi:10.1088/1748-9326/9/12/124018.

Lesiv, M, S Fritz, N Herold, J.-F Buchhorn, and Van De Kerchove. 2017. "Evaluation of ESA CCI prototype land cover map at 20m." IIASA Working Paper. International Institute for Applied Systems Analysis (IIASA). http://pure.iiasa.ac.at/14979/1/WP-17-021.pdf.

Monfreda, Chad, Navin Ramankutty, and Jonathan A. Foley. 2008. "Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000." *Global Biogeochemical Cycles* 22 (1). doi:10.1029/2007GB002947.

Pekkarinen, Anssi, Lucia Reithmaier, and Peter Strobl. 2009. "Pan-European forest/non-forest mapping with Landsat ETM+ and CORINE Land Cover 2000 data." *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (2): 171–83. doi:10.1016/j.isprsjprs.2008.09.004.

Portmann, Felix T., Stefan Siebert, and Petra Döll. 2010. "MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling." *Global Biogeochemical Cycles* 24: 1–24. doi:10.1029/2008GB003435.

Rutten, Martine, Michiel van Dijk, Wilbert Van Rooij, and Henk Hilderink. 2014. "Land use dynamics, climate change, and food security in Vietnam: A global-to-local modeling approach." *World Development* 59: 29–46.

Siebert, Stefan, Verena Henrich, Karen Frenken, and Jacob Burke. 2013. "Global Map of Irrigation Areas version 5." Bonn, Germany: Rheinisch Friedrich-Wilhelms-University.

Wood-Sichra, Ulrike, Alison B. Joglekar, and Liangzhi You. 2016. "Spatial Producion Allocation Model (SPAM) 2005: Technical Documentation." Harvest Choice Working Paper. Washington, D.C.: International Food Policy Research (IFPRI).

World Bank. 2010. "The Zambezi river basin: A multi-sector investment opportunities analysis." Washington, DC.

You, Liangzhi, and Stanley Wood. 2006. "An entropy approach to spatial disaggregation of agricultural production." *Agricultural Systems* 90 (1): 329–47. doi:10.1016/j.agsy.2006.01.008.

You, Liangzhi, Stanley Wood, and Ulrike Wood-Sichra. 2009. "Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach." *Agricultural Systems* 99 (2): 126–40. doi:10.1016/j.agsy.2008.11.003.

You, Liangzhi, Stanley Wood, Ulrike Wood-Sichra, and Wenbin Wu. 2014. "Generating global crop distribution maps: From census to grid." *Agricultural Systems* 127: 53–60. doi:10.1016/j.agsy.2014.01.002.

Zhong, Liheng, Lina Hu, Le Yu, Peng Gong, and Gregory S. Biging. 2016. "Automated mapping of soybean and corn using phenology." *ISPRS Journal of Photogrammetry and Remote Sensing* 119 (September). Elsevier: 151–64. doi:10.1016/J.ISPRSJPRS.2016.05.014.