IOP Conf. Series: Earth and Environmental Science **17** (2014) 012057 doi:10.1088/1755-1315/17/1/012057

Use of high resolution imagery and ground survey data for estimating crop areas in Mengcheng county, China

H Kerdiles[1], Q Dong[2], S Spyratos[1], FJ Gallego[1]

[1] Joint Research Centre (JRC) of the European Commission, Institute for Environment and Sustainability (IES), MARS Unit, Via Enrico Fermi 2749 21027 Ispra (Va) Italy

[2] Flemish Institute for Technological Research (VITO), Centre for Remote Sensing and Earth Observation, Boeretang 200, B-2400 Mol, Belgium

E-mail: herve.kerdiles@jrc.ec.europa.eu

Abstract. The use of remote sensing images in combination with ground survey data was assessed for deriving crop areas over Mengcheng County in 2011 in the North China Plain. First, a stratification of the county into arable land, permanent crops and non agricultural land was carried out by photo-interpreting a grid of points on Google Earth and a 2.5m Spot5 image from 2011. Then a sample of 83 segments was randomly selected in the arable stratum and surveyed with GPS. Two high resolution images (TM 30m and Spot5 10m) were acquired over the 2011 summer crop season and classified using maximum likelihood. The regression estimator was then applied using the surveyed segments and the classification and compared to the direct expansion estimate derived from the segments only; the calibration estimator was also tested using the same classification and the 83 arable points that served as seeds for the segments and compared to the estimate derived from the 83 points alone. The regression estimator proved to be the most efficient one in the North China Plain landscape. To reach the same variance of estimate as the regression estimator, the number of points to be surveyed for the calibration estimator should be multiplied by seven. Last pixel counting tested on the whole county and on the arable points of the grid resulted in biased estimates, in contrast to estimates based on ground data, in combination with remote sensing or not.

1. Introduction

Accurate agricultural statistics are the backbone of any agricultural and food security policy and yet, the overall quality and availability of agricultural statistics, especially in the developing world, has been declining since the early 1980s [1]. Developing accurate and cost efficient methods for estimating crop area at region or country level is therefore still relevant. For such purpose, area frame sampling is one of the methods that best guarantees unbiased estimates. It consists in dividing the region of interest into areal units (e.g. segments or points) and surveying a sample of these units on the ground. The accuracy of the estimation increases with the number of samples; however in practice, reaching the target accuracy may be hampered by budget constraints as the survey cost also increases with the number of samples. Because remote sensing imagery allows covering a whole region in a few "shots" and recognizing crops (to a certain extent), techniques based on high resolution images were developed to improve the accuracy of crop estimates derived from ground survey in the early 70s with the launch of Landsat-1 [2]. In practice, nowadays, these techniques are operationally used in few

countries only, mainly because of a lack of capacity as well as of financial resources. It is therefore of interest to demonstrate their use and efficiency in various contexts.

The objective of this study was to assess the combined use of ground survey data and high resolution imagery for estimating crop area in China. Two questions were targeted: since no regular ground survey data was available, which type of area frame / ground survey should we set up and test? Regarding the contribution of high resolution imagery to crop area estimation, which estimator would be the most accurate and cost-efficient?

Mengcheng county located in the north of Anhui province (North China Plain) was selected as its landscape is representative of the North China Plain, the main producing area for wheat and maize in China. The main crops in this county of 2150 km2 are winter wheat (harvested in May) followed by maize and soybean (harvested in September, October). According to the available official statistics, the maize area has grown from 55,000 ha in 2008 to 73,000 ha in 2010 while the soybean area has remained more stable around 32,000 ha (2009).

2. Data and methods

There are several ways for combining ground data and classified imagery and these depend on the type of ground data collected: the most classical technique is the regression estimator which is based on segments, i.e. blocks of land delimited by either physical or artificial (e.g. squares) boundaries (see for instance [2] [3] [4]); for a given crop, the relationship between its area proportion in the segments as derived from the ground survey and its proportion (or the proportion of a related class) in the classified imagery allows deriving an estimate of the crop area proportion with a reduced variance of estimate with respect to the variance derived from the ground survey alone. When points are surveyed, which is more cost efficient for certain types of landscapes, the calibration estimator technique has to be used to correct the area estimates derived from the classified imagery. Because the landscape of the North China Plain is made of regular blocks of thin land strips (typically 10 to 30m wide), we thought that surveying segments defined by physical boundaries would be efficient; therefore our primary objective was to assess the efficiency of the regression estimator for summer crops in a typical Chinese landscape. However, because of the way the segments were selected, we also had point survey data and could therefore compare the estimates of the calibration estimator to the ones derived from point data only.

2.1. Ground data collection

Our target was to survey 100 arable segments distributed over the whole county. In order to identify the segments as well as to determine the proportion of the arable stratum in the county, a 2 km grid was overlaid over the county and its 532 points were photo-interpreted using Google Earth imagery (which was made up of Very High Resolution imagery, mostly from 2004, for 80% of the county in July 2011). The following classes were considered: arable land, permanent crops and poplars, non-agricultural land (artificial, water), thematic doubt (i.e. doubt between arable land and non-arable land, mainly on the areas covered with TM 30m data) and geometric doubt for points falling on the border between arable and non-arable land. In order to identify 100 arable segments, a 4 km subgrid was extracted from the 2 km grid and all its arable points were selected; since 14 points were missing to reach 100 arable points, a second 4 km grid was extracted and 14 points were randomly drawn from its arable points. The segments corresponding to these 100 seeds were then identified by photo-interpretation of their boundaries on Google Earth (a segment being defined in most cases as a set of parallel strips of land)

Of these 100 segments, due to time constraints only 83 were surveyed with GPS by the Anhui Institute For Economic Research (AIFER) in August and September 2011 (3 additional segments could not be surveyed as their land cover changed from arable to non-agricultural between 2004 and 2011). On average the segment size was of 3.8 ha and 6 to 7 segments could be surveyed per day. The main crops were maize (76.8% of the total surveyed area) and soybean (19.8%); other summer crops

(cotton, peanuts, sesame, vegetables) and non-agriculture accounted for 3.0% and 0.4% of the surveyed area (i.e. the arable stratum) respectively.

2.2. Satellite data processing

Two satellite images were obtained: a Spot5 ordered on 15 July and acquired on 22 September 2011, close to the time of harvest for summer crops, and a Landsat 5 TM image (bands 3, 4 and 5, georeferenced) acquired on 1st June 2011. Two resolutions were requested for the Spot5 image: the standard 10m multispectral image, and the 2.5m pansharpened image for updating the stratification made with the 2004 to 2011 images from Google Earth. The 10m Spot5 image was georeferenced by AIFER, while the 2.5m image was orthorectifed by Spot Image. The images were classified with the maximum likelihood algorithm with no prior probability; 41 segments were used to train the classifier while the remaining 42 were kept for assessing the classification accuracy. For the classes "woodland", "water bodies" and "artificial surfaces" that were nearly not present in the arable segments, polygons were selected by photo-interpretation of the Spot 2.5m image (taking care of the overall proportion of these land cover types for the validation data set). Three combinations of the 10m Spot and the TM5 images were tested: the Spot image alone, the Spot image combined with TM band 4, the two images together (7 bands).

2.3. Crop area estimation methods

The maize and soybean areas were estimated using the following six methods: two based on ground data only viz. (1) the direct expansion estimator, i.e. the mean crop proportion derived from the segments, and (2) the crop mean percentage derived from the 83 points that served as seeds for the segments; two based on the classified images only through pixel counting over the whole county (3) or over the 379 arable points of the grid (4); and two based on the combination of ground data and remote sensing, viz. (5) the regression estimator based on the 83 segments and the classified images and (6) the (direct) calibration estimator based on the 83 points and the classified images. The last two estimators are described in detail in [5].

3. Results

3.1. Stratification

According to the photo-interpretation of the Spot5 2.5m image, 73.1% of the county total area (2148.8 km^2) is covered with arable land. This estimation results from the re-allocation of 50% of the 17 doubtful points (10 points assigned to geometric doubt, 1 to thematic doubt and 6 falling outside the image or on a cloud) to the arable stratum. None of the 532 points was interpreted as permanent crop. The estimated arable area is therefore of 157,000 ha with a maximum error of 2,200 ha due to the re-allocation of doubtful points. For comparison, the estimated arable land area derived from the photo-interpretation of Google Earth imagery was 164,000 ha (76.3%) after re-allocation of 50% of the geometric doubts (18 points). Out of the 50 points interpreted differently on Google Earth and Spot5, only 7 had a real land cover change between the two images (6 going from arable to non-agriculture and 1 from village to arable); the remaining discrepancies were due to differences in the registration of the two images. The official figure for arable land is 122,500 ha according to the 2010 statistical yearbook (considered to be the most reliable source of statistics), which is clearly below the two remote sensing estimates. This might be explained by the omission of the cultivated area on river beds. Interestingly Mengcheng county estimates its arable land at 153,000 ha, a figure which seems more compatible with the 140,000 ha of summer crops estimated for 2007.

3.2. Classification

Of the three image combinations tested, the classification of the spot + TM band 4 image (Figure 1) resulted in the highest overall accuracy (80.4%) followed by the Spot + TM all 3 bands (79.2%) and by the Spot image alone (74.6%). Table 1 shows the confusion matrix for this combination which

resulted in the highest producer accuracy for maize and soybean, with 79.3% and 72.5% of the maize and soybean pixels respectively being correctly identified. Confusions between summer crops appear to be relatively high with 20% of the maize pixels assigned to other summer crops, mainly soybean; 27% of soybean pixels assigned to maize and other crops and 73% of other crops pixels classified as maize and soybean. The classifier underestimated the majority class (maize) and overestimated the minority ones (e.g. soybean); this could be partly corrected through the use of prior probabilities (e.g. using crop statistics from the previous year); in such a case however, the opposite bias (overestimation of large classes, underestimation of small ones) is obtained [5].

Table 1. Confusion matrix (expressed in % of the number of pixels of a given "ground" class) for the Spot + TM band 4 image, derived from the 42 validation segments

	Ground truth (%)						
Class	Maize	Other crops	Soybean	Woodland	Water bodies	Artificial	Total (% classified pixels)
Maize	79.3	31.2	24.4	1.9	0	0.1	57.9
Other crops	3.3	23.1	2.5	1.9	0	0.1	3.3
soybean	16.9	41.9	72.5	0.7	0	0.3	22.6
Woodland	0.1	0	0	95.5	0	0	9.2
Water bodies	0	0	0	0	96.5	0	0.6
Artificial	0.4	3.8	0.7	0	3.5	99.5	6.5
Total (%)	100	100	100	100	100	100	100

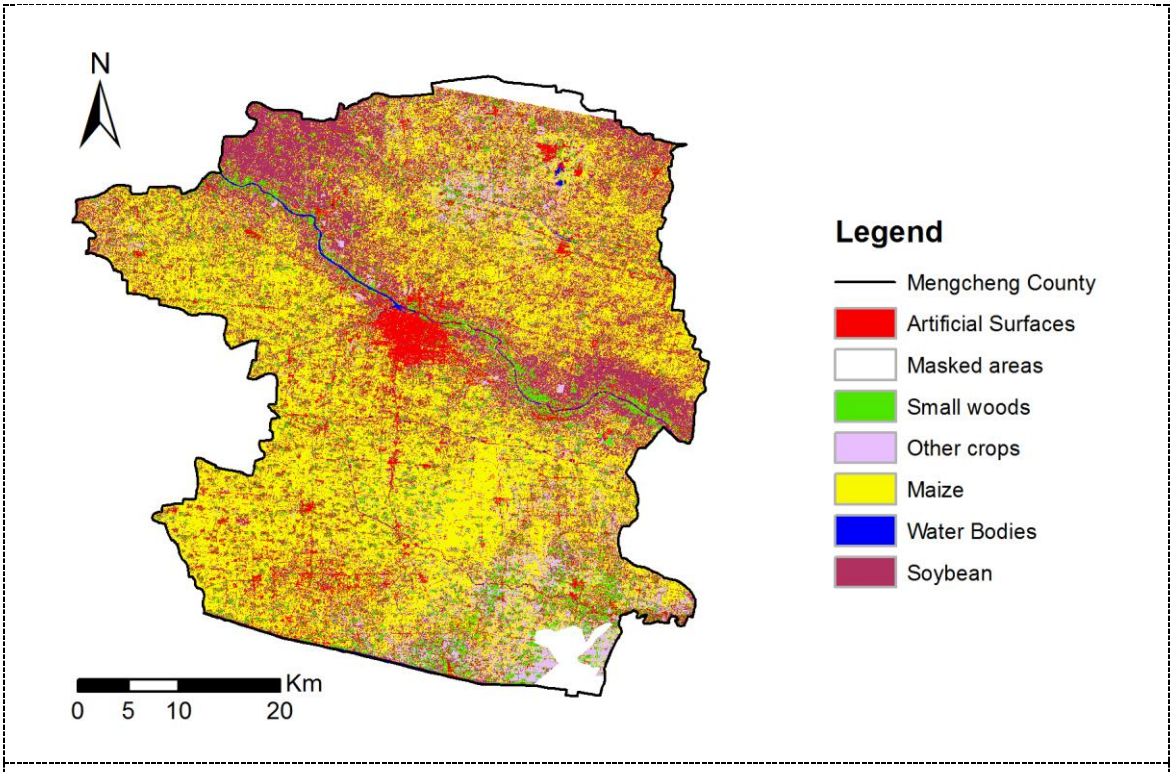

Figure 1. Classification of Spot5 + TM band 4 image (the masked areas correspond to the limit of the image (north) or to a cloud (south east)).

3.3. Crop area estimations

The regressions on all segments between the crop percentages derived from the classification and the ground survey percentages are shown for maize and soybean in figures 2 and 3 respectively.

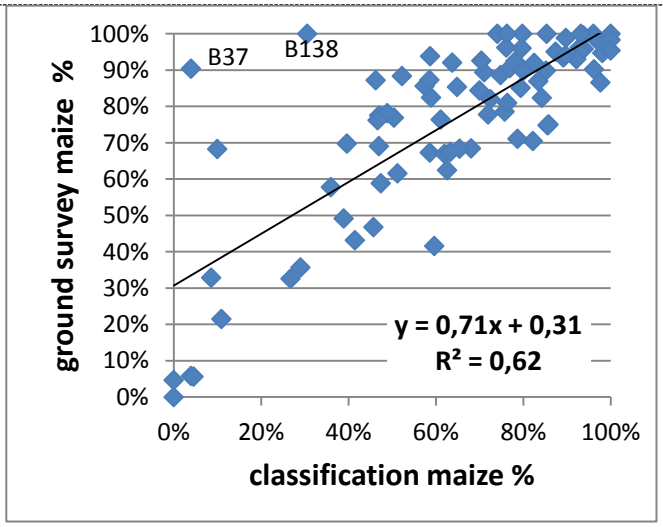

Figure 2. Relationship between the maize area percentage derived from the ground survey and from the classification in the 83 segments

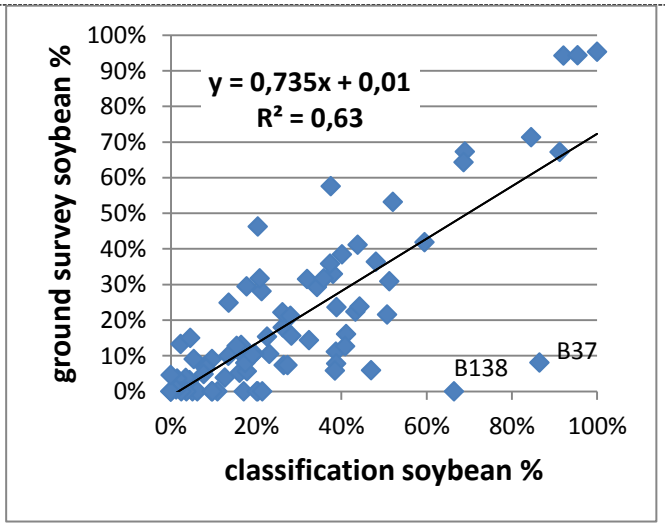

Figure 3. Relationship between the soybean area percentage derived from the ground survey and from the classification in the 83 segments

R^2 around 0.60 means a relative efficiency (i.e. the ratio of the variance of the direct expansion estimator over the variance of the regression estimator) of 2.6; in other words, the use of the classified image is equivalent to surveying about 220 segments, i.e. 1.6 more segments than the initial 83. In this context, remote sensing will be cost-efficient if the image acquisition and processing cost is less than the survey cost of the additional 135 segments (estimated at € 8,000 in this study).

Figures 4 and 5 show the maize and soybean area estimates with their standard error, except for the two pixel counting methods, derived by the six methods for Mengcheng County:

- Not surprisingly pixel counting resulted in biased estimates with underestimation of maize and overestimation of soybean (e.g. 97,600 and 61,700 ha for maize and soybean resp. by counting all pixels of the county), in contrast to the four estimates based on ground data.
- The regression estimator yielded the lowest standard error of estimate (2,600 and 2,400 ha for maize and soybean resp.) whereas both the point survey and the calibration estimator showed the highest error (around 7,000 and 6,000 ha for maize and soybean resp.); the calibration estimator error being slightly lower than the one of the point survey; this result is not a surprise as only 83 points were surveyed (this test being a by-product of the segment survey). To reach the same level of variance as the regression estimator, the number of points to survey should be multiplied by seven.
- The most recent official statistics indicate a maize area of 73,400 ha for 2010, in progress since 2007 (53,600 ha), and of 32,300 ha for soybean in 2009. If the 2009 soybean area is close to our 2011 estimates (31,400 ha with the regression estimator, 31,000 ha with the direct expansion, 29,100 ha with the calibration estimator and 28,400 ha with the point survey), the 2010 maize area seems to be well below the 2011 estimates (115,000 ha with the regression estimator, 120,700 ha with the direct expansion, 118,300 ha with the calibration estimator and 119,200 ha with the point survey).

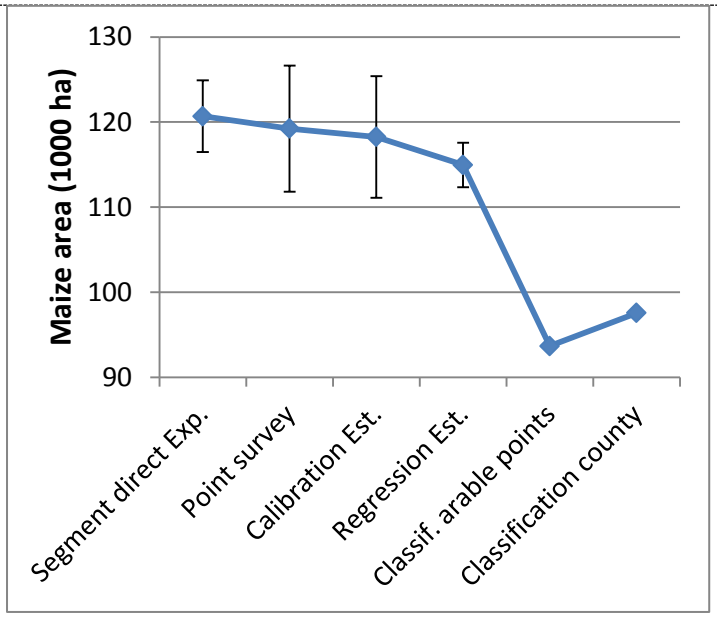

Figure 4. Maize area (with standard error of estimate) for Mengcheng county according to the six estimators.

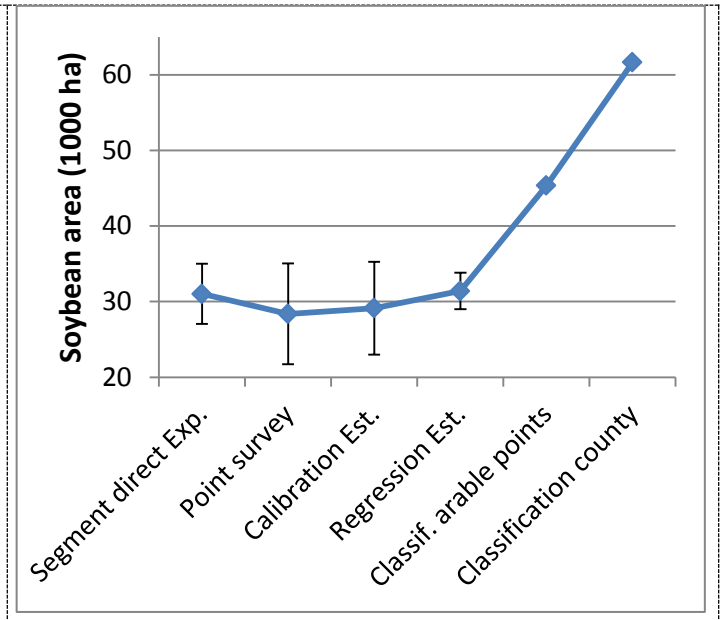

Figure 5. Soybean area (with standard error of estimate) for Mengcheng county according to the six estimators

4. Conclusions

Of the six estimators tested over Mengcheng county, the ones based on ground survey alone or combined with remote sensing resulted in more unbiased crop area estimates, in contrast to the "pixel counting" estimators. The regression estimator based on physical segments (identified on Google Earth imagery and then surveyed in the field) and a classified image yielded the lowest variance of estimate. The same level of accuracy could be reached with the calibration estimator by surveying a number of points equal to seven times the number of segments. No strict cost comparison was made between the two types of survey; however we believe that the segment approach is more efficient than the point survey in a regular pattern of arable fields, as the one of the North China Plain, in contrast to Europe. The cost efficiency of remote sensing depends on the relative costs of ground survey and of image acquisition and processing. Sentinel 2 with its 290 km swath, 10 m resolution and hopefully free access should be cost efficient in the North China Plain.

References

[1] The World Bank, FAO 2010 Global Strategy to Improve Agricultural and Rural Statistics report number 56719-GLB

[2] Hanuschak G A, Allen R D and Wigton W H 1982 Integration of Landsat data into the crop estimation program of USDA's statitiscal reporting service 1972-1982 in *Proc. 1982 Machine Processing of Remotely Sensed Data Symp* pp 1-11

[3] Gallego F J, Delincé J and Carfagna E 1994 Two-Stage Area Frame Sampling on Square Segments for Farm Surveys, *Survey Methodology* **20**, 2, 107-115

[4] Taylor C, Sannier C, Delincé J and Gallego F J 1997Regional Crop Inventories in Europe Assisted by Remote sensing: 1988-1993, Synthesis Report. EUR 16317 EN, Office for Official Publications of the EC, Luxembourg, 71 pp.

[5] Carfagna E and Gallego F J, 2005 Using Remote Sensing for Agricultural Statistics, *Int. Statistical review* **73**, 3, 389-404