Published online April 4, 2007

Deriving Comprehensive County-Level Crop Yield and Area Data for U.S. Cropland

Erandathie Lokupitiya,* F. Jay Breidt, Ravindra Lokupitiya, Steve Williams, and Keith Paustian

ABSTRACT

Ground-based data on crop production in the USA is provided through surveys conducted by the National Agricultural Statistics Service (NASS) and the Census of Agriculture (AgCensus). Statistics from these surveys are widely used in economic analyses, policy design, and for other purposes. However, missing data in the surveys presents limitations for research that requires comprehensive data for spatial analyses. We created comprehensive county-level databases for nine major crops of the USA for a 16-yr period, by filling the gaps in existing data reported by NASS and AgCensus. We used a combination of regression analyses with data reported by NASS and the AgCensus and linear mixed-effect models incorporating county-level environmental, management, and economic variables pertaining to different agroecozones. Predicted yield and crop area were very close to the data reported by NASS, within 10% relative error. The linear mixed-effect model approach gave the best results in filling 84% of the total gaps in yields and 83% of the gaps in crop areas of all the crops. Regression analyses with AgCensus data filled 16% of the gaps in yields and crop areas of the major crops reported by NASS.

CROP STATISTICS are widely used in decision making and policy formulation, and they provide important information for economists and researchers in agricultural and environmental fields. Crop statistics are useful indicators of economic performance, technological advances in crop breeding, and improvements in overall agricultural management practices. Crop statistics have also been used in environmental research to estimate net primary productivity (Prince et al., 2001). Such data can also be used as input to other models to analyze spatial patterns of regional scale C dynamics and/or for validation purposes; for example, to assess satellite-derived information on crop production (Lobell et al., 2002). Thus, availability of complete and comprehensive crop data at subregional scales, such as the county level, enhances the usability of these data for a variety of purposes.

Currently, county-level crop yields in the USA are collected in two surveys: one by NASS and the other by the AgCensus. The NASS crop yield data are produced annually using a survey done on selected farms, which is extrapolated statistically to estimate county-level crop yields. AgCensus estimates are produced every 5 yr during the years ending in “2” and “7”; AgCensus focuses on collecting data by contacting every farmer within a county; thus AgCensus data are generally more complete and comprehensive for the years when the survey is conducted (Pawel and Fecso, 1988; USDA, 1998; R. Korkosh, personal communication, 2004).

However, missing data (i.e., gaps) in some counties in certain years, and reporting of yield and crop area information only at the state level for certain states limits the usability of these data for research that requires comprehensive analyses involving single or multiple years. Therefore, the aim of our study was to derive complete county-level crop yield and crop area databases by filling the gaps in the yield and crop area data reported by NASS during the period 1982 to 1997, using AgCensus data and statistical models incorporating appropriate county-level environmental, management, and economic variables.

Studies to estimate crop yields thus far include models incorporating various agro-meteorological variables (e.g., Berka et al., 2003) or combinations of agrometeorological, hydrological, management, and economic variables such as the EPIC model (e.g., Cavero et al., 2001; Tan and Shibasaki, 2003). Some studies have used combinations of ground-based and satellite-based information (Lobell et al., 2002; Doraiswamy et al., 2003, 2004, 2005; Yang et al., 2004; Tao et al., 2005) to estimate yields. In some of these studies, yields have been estimated through combining agrometeorological variables with remotely sensed information in statistical models (e.g., Rudorff and Batista, 1990, 1991; Smith et al., 1995). In certain other studies, remotely sensed information has been combined with crop models such as EPIC (e.g., Doraiswamy et al., 2003; Yang et al., 2004) and FAO Crop Specific Water Balance Model (CSWB; Reynolds et al., 2000) to estimate yields. These models have been mostly used in field- or regional-scale estimation of crop yields.

In studies for crop area estimation, either remotely sensed information (Bauer et al., 1978; Hixson et al., 1981; MacDonald and Hall, 1980; Csornai et al., 2002) or purely statistical models (Griffith, 1999) have been used. Remotely sensed information has also been used in improving the precision of ground-sampled data for area estimates (Gonzalez-Alonso et al., 1997; Allen et al., 2002).

In this study, we first evaluate the existing national crop yield and area databases of NASS and AgCensus, their characteristics, and their compatibility. We describe the methods used for the imputation of missing data for crop yields and area and evaluate the appropriateness

E. Lokupitiya, S. Williams, and K. Paustian, Natural Resource Ecology Lab.; E. Lokupitiya and K. Paustian, Dep. of Soil and Crop Sci.; F.J. Breidt, Dep. of Statistics; and R. Lokupitiya, Dep. of Atmospheric Sci., Colorado State Univ., Fort Collins, CO 80523. Received 8 May 2006. *Corresponding author (erandi@atmos.colostate.edu).

Published in Agron. J. 99:673–681 (2007).
Soil & Crop Management
doi:10.2134/agronj2006.0143
© American Society of Agronomy
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: AIC, Akaike Information Criterion; AgCensus, Census of Agriculture; CRP, Conservation Reserve Program; FIPS, Federal Information Processing Standard, codes to identify U.S. counties; ITA, irrigated/total crop area ratio; LRR, land resource region; MST, mean monthly summer temperature; NASS, National Agricultural Statistics Service; NASSus, the final database created after filling the gaps in NASS data using AgCensus and linear mixed effect models incorporating environmental and economic data; NRI, National Resources Inventory; P, precipitation; PET, potential evapotranspiration.

of these methods in imputing long-term data gaps in national crop statistics.

METHODS

Evaluation of the Available National Crop Statistics for Major Crops in the USA

Yields and crop area of alfalfa (*Medicago sativa* L.) hay, barley (*Hordeum vulgaris* L.), corn (*Zea mays* L.) for grain, corn for silage and green chop, oat (*Avena sativa* L.), other hay (hay other than alfalfa; i.e., tame hay, small grain hay, wild hay), sorghum (*Sorghum bicolor* L.), soybean (*Glycine max* L.), and wheat (*Triticum aestivum* L.) were included in this study. Collectively, these crops make up $>90\%$ of the harvested cropland area of the USA.

NASS has reported crop statistics at the state level for >100 yr and at the county level for >70 yr, for most of the USA. The AgCensus was started in 1840, and has collected county-level information every 5 yr since 1920. When our study was initiated, data were available in digital form for 1982, 1987, 1992, and 1997 from AgCensus and digital data were available annually from NASS, so that 1982 to 1997 was chosen as the time period for constructing the new database. Data reported include planted and harvested crop area, yield and total production, and management practices (e.g., irrigation, summerfallowing). The data were organized using MS Access (Microsoft Corp., Redmond, WA) and all statistical analyses were performed using SAS v. 9.1 (SAS Institute, Cary, NC).

Compatibility of the crop yield and crop area estimates by NASS and AgCensus were evaluated by mapping (in ArcGIS v. 8.2, ESRI, Redlands, CA) the number of years AgCensus and NASS have each reported data for each crop for the period 1982 to 1997, and by mapping the absolute differences and percentage differences (e.g., the difference in NASS crop yield as a percentage of the yield reported by AgCensus) in the crop yields and crop area for the two databases. Percentage differences were used to find the distribution of data representing extreme differences (i.e., outliers) between the NASS and AgCensus databases.

Synthesis of Comprehensive Crop Yield and Area Databases

Following the preliminary analyses of discrepancies between the data reported by NASS and AgCensus, a step-wise procedure was used to fill data gaps in crop yields and areas (Fig. 1). Because NASS data is collected each year, it was chosen as the foundation database and data from AgCensus served in adjusting and filling certain missing data, as described below. The final synthetic database we produced by filling all the gaps for the relevant major crops in the NASS database is referred to as NASSus.

Filling Gaps in Crop Yields and Areas Reported by NASS Where AgCensus Data Were Available

In earlier work, leave-one-out and leave-*k*-out procedures were used to determine the most suitable statistical method for imputing NASS data (Lokupitiya et al., 2006). Regression analyses between NASS and AgCensus crop yield data and multiple imputation technique were found to be the best methods. Spatial statistical analyses such as the Kernel regression and kriging were less suitable (Lokupitiya et al., 2006). Therefore, regression analyses between NASS and AgCensus yield data were performed to replace extreme data or outliers in NASS yields and fill in the gaps. To detect outliers, a criterion based on the lower quartile (Q1; 25th percentile), upper quartile (Q2; 75th percentile), and interquartile distance (IQ) was used; any value $< (Q1 - 3 \times IQ)$ and any value $> (Q2 + 3 \times IQ)$ was removed. Regression analyses of AgCensus and NASS yields for each crop and each of the years 1982, 1987, 1992, and 1997 were used to replace outliers and fill the gaps in NASS yield data. Crop areas reported in AgCensus that were missing for corresponding years and counties in the NASS data were added to the NASSus database without adjustment.

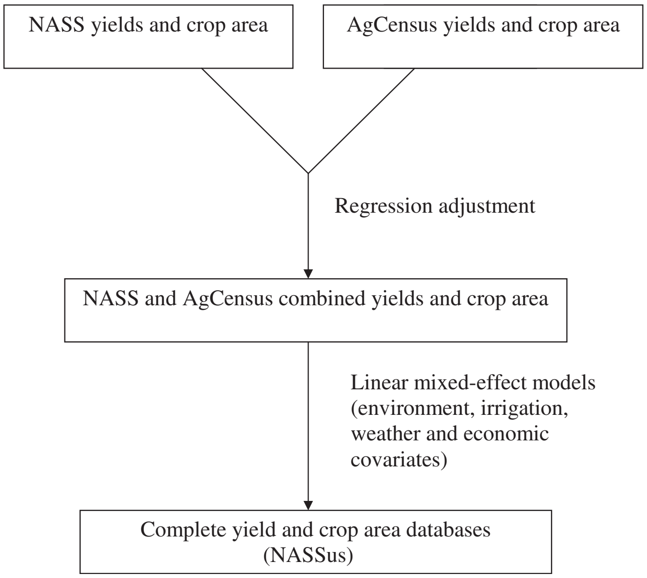
NASS yields and crop area
AgCensus yields and crop area
Regression adjustment
NASS and AgCensus combined yields and crop area
Linear mixed-effect models (environment, irrigation, weather and economic covariates)
Complete yield and crop area databases (NASSus)

Fig. 1. Flowchart outlining the database construction steps.

Using Environmental Variables to Fill Remaining Gaps in Crop Yields

Linear mixed-effect models (Littell et al., 1996) were used for filling remaining gaps in the yields in NASSus data, utilizing environmental and management factors such as irrigation to predict yields. County-level weather and irrigation data were chosen as the covariates in the mixed models, with yields as the dependent variable. Mean monthly summer temperature (MST), annual precipitation (P), precipitation/potential evapotranspiration ratio (P/PET), and irrigated/total crop area ratio (ITA) were the variables that represented fixed effects in the models. County FIPS (Federal Information Processing Standard) code was the only variable representing any random effects, where the random effect represents variation among counties due to factors other than the fixed effects.

Determination of Environmental Data for Linear Mixed Effect Model Runs

Annual PET was calculated for 1982 to 1997 using the method by Thornthwaite (1948), with weather data from the gridded PRISM (Daly et al., 1994) dataset (www.ocs.orst.edu/prism/, verified 2 Feb. 2007) for the conterminous USA (Fig. 2).

The USA has a total of 20 Land Resource Regions (LRRs), which delimit contiguous areas with similar geographical, climate, and land use conditions (Fig. 3). County-averaged yield data from different major crops and county weather variables were grouped by LRRs across the entire time series and sorted by year and county. In addition, ITA for a county was used as a conditional variable to determine whether moisture limitations would be included in the model. If a county had an ITA

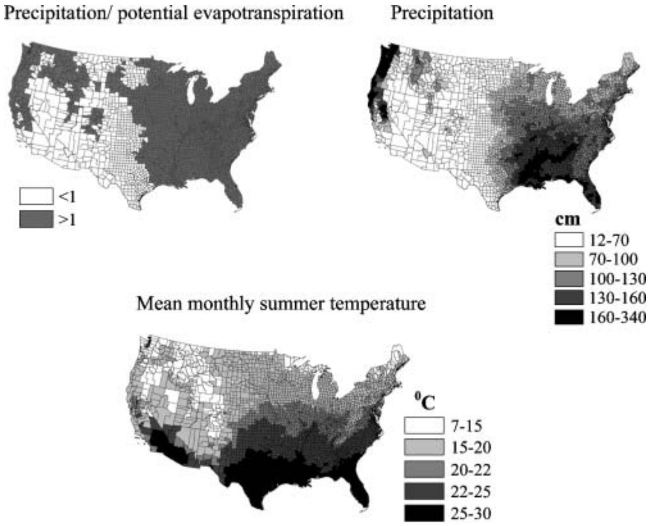

Fig. 2. The main environmental variables used in the mixed models for filling the gaps in yields (only the data from 1982 are shown here).

greater than 0.5 (i.e., majority of cropland is irrigated), then P and P/PET were not used in the prediction of yield.

NASS typically reports separate explicit categories for irrigated and nonirrigated cropland where both are present as significant land area fractions. In some arid counties, the entire area for a particular crop is likely to be irrigated, and hence NASS may only report total area for that crop. Similarly, in many eastern U.S. counties where irrigation is minimal, NASS may only report total crop area. In such instances, crops were designated as primarily irrigated or nonirrigated based on location, type of crop, and long-term climate averages.

Selection Criteria for the Best Linear Mixed Effect Models and Quality Control Measures for the Predicted Crop Yields

Equation [1] gives the basic model used in the linear mixed-effect models for crop yields:

$$Y = \mathbf{X} \times \boldsymbol{\beta} + \mathbf{Z} \times \mathbf{u} + \boldsymbol{\varepsilon}, \qquad [1]$$

where Y = yield; $\mathbf{X}$ = design matrix of covariates (or fixed effects, including the intercept, P, P/PET, ITA, MST, and interactions between these variables); $\boldsymbol{\beta}$ = vector of coefficients corresponding to the fixed effects; $\mathbf{Z}$ = design matrix of 0s and 1s for the random effects for each FIPS, with 1 in column j indicating that observation is from county j; $\mathbf{u}$ = vector of coefficients corresponding to the random effects for each FIPS; and $\boldsymbol{\varepsilon}$ = error vector (which may be autocorrelated with time).

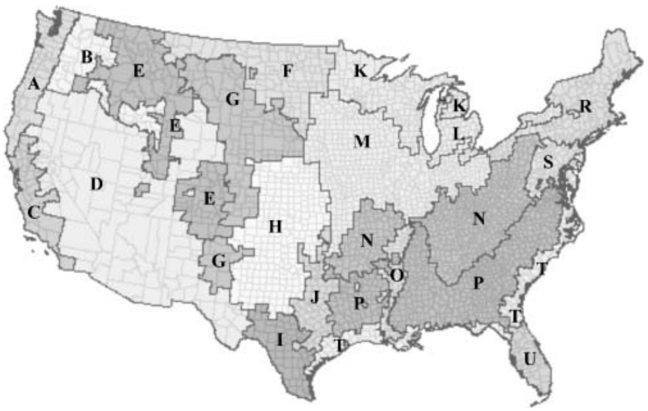

Fig. 3. U.S. Land Resource Regions (LRRs; USDA, 1981). The boundaries of LRRs have been rectified to county boundaries in the map.

The linear mixed-effect models were run separately for each LRR with autoregressive order 1 (AR1) covariance structure, with time as repeated measures, to fill in the remaining yields in counties. Thus, for each crop, several models with different combinations of the above covariates and the crop yield as the response variable were run on each LRR. The model with the lowest Akaike Information Criterion (AIC) was chosen as the best model for each LRR; the chosen model was used in filling the county-level yield gaps in different LRRs under different crops. Altogether, 10 models were attempted on each LRR per crop; if convergence criteria were not met, or the final Hessian was not positive definite, either AR(1) or random effects had to be dropped to satisfy the convergence criteria.

The predicted yields from the linear mixed-effect models were compared against a default method that used the mean of the observed values within a county as the predicted value. A few counties had only one observed (reported) yield or crop area value across the entire time series in an LRR under certain crops, while the other counties had several observations with relatively large variation in crop area or yields over the time series. Hence, no single standard statistical method could be adopted to screen for outliers. Therefore, the following procedure was adopted for identifying outliers. First, where predicted values fell outside plus/minus three times the mean of the observed values, they were given an initial designation as potential outliers. Where potential outliers (in predicted values) also fell outside the range of the observed values for the particular counties, they were given a final designation as outliers, and were replaced with the mean of the observed values (i.e., the default option). Final data screening was done considering the occurrence of the crop at county level; that is, imputed data were removed from counties that had no reported occurrence of the crop.

Using Environmental and Economic Variables in Gap-Filling for Crop Area Data

The mixed models for crop area data included economic and weather variables from the previous year as fixed-effect predictor variables: P, crop price, fertilizer cost (unit cost of anhydrous ammonia), and diesel cost. County area and cropland area set aside in the Conservation Reserve Program (CRP) for the current year were also used as fixed effect variables. County FIPS served as the only variable for random effects.

Data Preparation

Crop price data for the previous year were obtained from NASS (www.nass.usda.gov/Data_and_Statistics/index.asp, verified 2 Feb. 2007). Available state-level prices of the crops were extracted for the period 1981 to 1996. Where state-level price data were not available, the mean price of the multistate crop production region was used. Since no price data were available for corn for silage in NASS, corn for silage price per ton was estimated by multiplying the per-bushel price of corn grain by nine (Barkley, 2002). All crop prices were adjusted for inflation using the Gross Domestic Product-Implicit Price Deflator (GDP-IPD; S.R. Koontz, 2005, personal communication).

Fertilizer and diesel price data of the previous year were extracted from the USDA Agricultural Prices Annual Summary reports for the period 1981 to 1996 (USDA, 1982–1997), and adjusted for inflation using the GDP-IPD. Previous year's P was extracted from the PRISM data grid described above.

Cropland area enrolled in CRP since the beginning of the program in 1986 to 1997 was obtained from the Economic Research Service (2005) and A. Barbarika (2005, personal communication). County-level crop area data and the data from all the predictor variables for the entire time series for the period 1982 to 1997 were then compiled and organized in a similar structure as detailed above for yield data.

Selection Criteria for the Best Linear Mixed-Effect Models and Quality Control Measures for the Predicted Crop Area

Linear mixed-effect model analyses were performed with auto regressive order 1 (AR1) covariance structure with crop area as the dependent variable, county FIPS as the random effect, and different combinations of the following variables as the fixed effects: previous year's P, previous year's crop price, previous year's fertilizer price, previous year's diesel price, CRP crop area, and county area. The analyses were performed under the following model options: (i) crop area as the response (y) variable, and diesel price, fertilizer price, crop price, CRP crop area, and county area as regression (x) variables; (ii) crop area/county area as the y variable, and the rest of the variables as x; (iii) crop area/county area as the y variable, with rest of the variables standardized (by dividing the value of each variable by the standard deviation of each variable), as x variables; (iv) taking all the variables standardized including the y variable in (iii) above; (v) log-transformed crop area as the y variable, log-transformed county-area as an x variable, and all the other x variables standardized (by dividing by the standard deviation); predicted y values were exponentiated back to get the predicted crop area values from the models.

Linear mixed effect models were run on each different crop at LRR level, considering the entire time series. The model with the lowest AIC was selected as the best model for filling the gaps in crop area data.

Detection of any outliers and quality control of the predicted crop area were performed in the same way as for the yields. The total of the crop area aggregated at state-level was compared against the state-level cropland crop area based on the information collected by National Resources Inventory (NRI), as an additional quality control measure.

RESULTS

Preliminary Analysis of Data

Except for a few outlying values, AgCensus and NASS crop data were very similar, although some significant deviations were found. When all the crops were considered together across all the years, 80 to 99% of counties reported crop yields with <30% difference between NASS and AgCensus and 15 to 70% of counties had differences < 5% in reported crop yields (Table 1).

The number of outliers was relatively small (<1 to 5%) for each crop within each year. For instance, Fig. 4 shows the deviation of NASS wheat yields from those reported by AgCensus for 1997. Since NASS collects information from a sample of farmers within a county and extrapolates that information to the entire county, it may create occasional anomalous values that were obvious during the years when both NASS and AgCensus data are reported. Extreme differences between AgCensus and NASS for the same crop were infrequent even when the absolute differences between the yields were compared. For instance, the corn yields reported by NASS and AgCensus for most counties had an average difference of <0.6 Mg ha^{-1}.

Existing Gaps in the Crop Yields Reported by NASS

Although NASS reports annual data, NASS does not report crop yields in certain states and counties that are known to contain particular crops, especially hay crops. For instance, NASS does not report alfalfa crop yields in the counties of 21 states, while AgCensus does (Fig. 5). Similarly, AgCensus has not reported county-level yields and/or crop area of barley, corn for grain, corn for silage, oat, other hay, sorghum, soybean, and wheat for some counties where NASS has reported data. Even in the years that both AgCensus and NASS have reported data, we found that there are missing values in some counties.

Synthesis of Comprehensive Databases of Crop Yields And Crop Area

Filling Initial Gaps in NASS Data Using AgCensus

For alfalfa hay, barley, corn for grain, oat, sorghum, soybean, and wheat, close to 90% of the variation in the data reported by NASS were explained by AgCensus data (Table 2). However, corn for silage and nonalfalfa hay showed a weaker relationship between NASS and AgCensus (lower R^2 values). For all the crops, the slope of the regression was close to 1.0 and the intercept was close to zero in the majority of crops (except for corn for silage and green chop, oat, and sorghum, which could be due to the underrepresentation of the data in either of the databases). These regression models were used to replace outliers in the NASS data, and fill in the gaps during the above 4 yr when both NASS and AgCensus have reported crop yield and area. Using this

Table 1. Percentage of counties having < 30% and < 5% difference, respectively, in NASS yield data compared with AgCensus.

	1982		1987		1992		1997	
Crop	<30%	<5%	<30%	<5%	<30%	<5%	<30%	<5%
Barley	92	37	93	40	91	23	99	60
Wheat	94	50	93	29	93	26	98	42
Alfalfa hay	82	19	80	15	78	13	92	37
Corn	95	37	96	46	95	35	99	71
Corn for silage	94	46	91	35	88	24	96	48
Sorghum	89	45	91	35	84	17	91	39
Other hay	77	24	81	19	80	17	88	36
Soybean	97	59	97	60	97	47	99	77
Oat	95	50	93	51	92	23	98	40

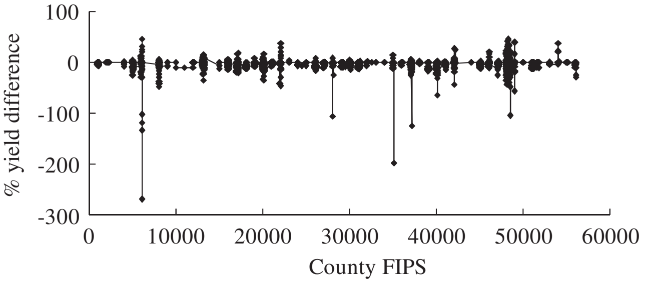

Fig. 4. Difference in NASS wheat yield as a percentage of AgCensus data for 1997. FIPS = Federal Information Processing Standard.

Table 2. Regression models obtained for AgCensus and NASS crop yields for 1997.

Crop	Regression model	R^2
Alfalfa hay	NASS = 0.2497 + 1.006 AgCensus	0.86
Barley	NASS = −0.1863 + 1.030 AgCensus	0.97
Corn for grain	NASS = 0.9819 + 1.009 AgCensus	0.97
Corn for silage and green chop	NASS = 9.7973 + 0.865 AgCensus	0.73
Oat	NASS = 4.7180 + 0.974 AgCensus	0.90
Other hay	NASS = 0.3297 + 0.907 AgCensus	0.71
Sorghum	NASS = 5.8482 + 0.967 AgCensus	0.92
Soybean	NASS = 0.2880 + 1.018 AgCensus	0.98
Wheat	NASS = 0.3128 + 1.051 AgCensus	0.96

process, 16% of the gaps in NASS yields and crop areas were filled.

Environmental Variables as Covariates in Linear Mixed-Effect Models for Filling Remaining Gaps in County-Level Crop Yields

Quality Control/Quality Assurance of Final Yields

The number of outliers in the predicted data was extremely low (i.e., <0.1% across all crops) and only three crops (i.e., alfalfa hay, barley and corn) contained them; these few outliers were replaced with the mean of the observed values for the county. All predicted yields had relative errors < 10%. During the 16-yr period, the majority of the missing data in NASS were for alfalfa hay (21% of the total gaps), other hay (25%), and corn for silage (11%); the percentage missing data in the rest of the crops ranged between 5 and 10% of the total gaps in the yields reported by NASS. Eighty-four percent of the total gaps in NASS (and 99.998% of the gaps in the NASS and AgCensus combined data) were filled using the mixed models. Only 0.02% of the gaps in NASS were filled with the default option (i.e., where the imputed values were designated as outliers).

Figure 6 shows the alfalfa yields from initial NASS database, NASS and AgCensus combined, and the completed alfalfa yields, with the imputed values (NASSus database). About 21% of the gaps in alfalfa yields reported by NASS were filled using the AgCensus information, and ≈78% of the gaps were filled using the linear mixed effect models; the remaining 1% of the gaps were filled with the default option. Figure 7 illustrates the yield trend across time with inclusion of the predicted values for corn in two counties, depicting the compatibility of the predicted (for the missing years) and the observed yields.

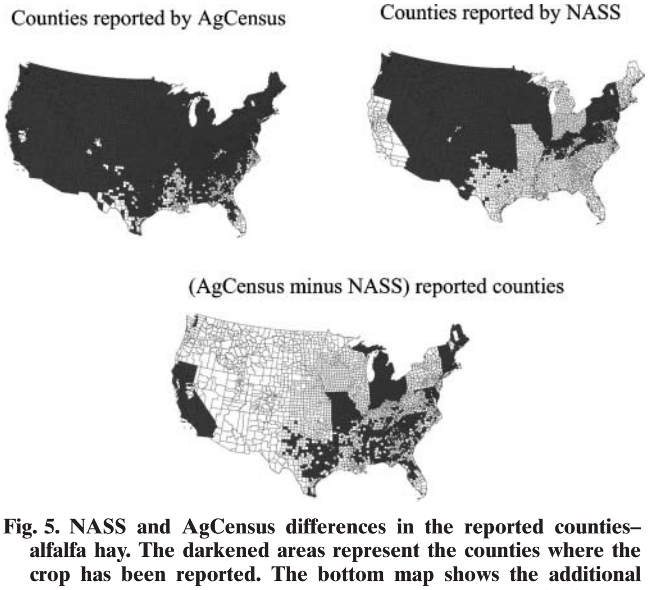
Counties reported by AgCensus
Counties reported by NASS
(AgCensus minus NASS) reported counties

Fig. 5. NASS and AgCensus differences in the reported counties–alfalfa hay. The darkened areas represent the counties where the crop has been reported. The bottom map shows the additional counties that AgCensus has reported compared with NASS.

Using Environmental and Economic Variables in Filling Remaining Gaps in Crop Area Data

Out of the mixed model options attempted, the final option (i.e., log-transformed crop area as *y* variable and all the other *x* variables standardized with log-transformed county-area as an *x* variable) gave the best results (Table 3). No universal trends could be observed in the response of the crop area to the fixed effect variables, but the crop area in the majority of the LRRs under each crop seem to have the best models with the combinations of three variables: diesel price, fertilizer price, and crop price of the previous year. When the importance of each single fixed effect variable is concerned, the diesel price of the previous year seems to be the most important predictor, being the sole predictor (other than the county area) for at least one LRR in a majority (five out of nine) of the crops. The area under CRP, either alone or in combination with the other variables, seemed

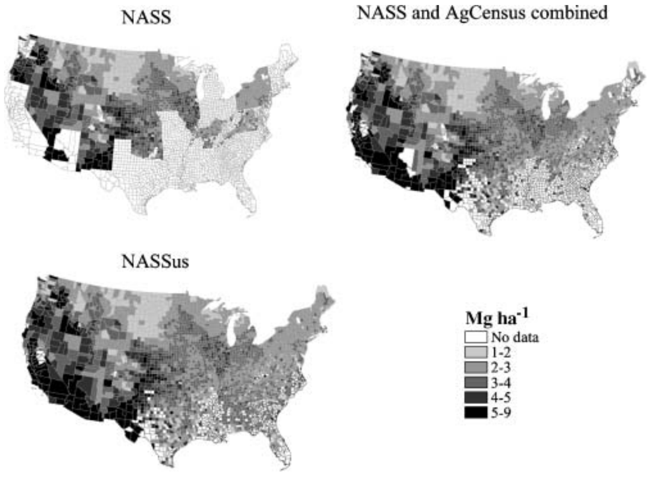

Fig. 6. Original NASS, NASS and AgCensus combined, and final gap-filled database (i.e., NASSus) of crop yields in 1997 alfalfa hay.

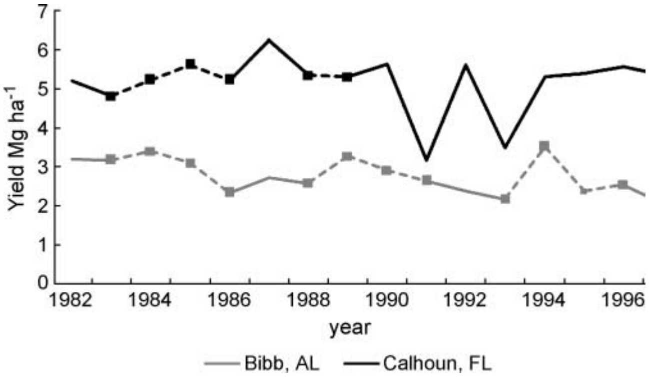

Fig. 7. Trend in corn yields with observed (continuous line) and predicted (broken line and filled squares for the values; for years with no data) values, for two counties in two different states during the 16-yr period.

to be more powerful in predicting the crop area when the area under CRP was high. This was particularly true for LRR G, having close to 30% of the cropland in CRP during most of the 16-yr period, and >95% of counties in the LRR containing land in CRP. However, other LRRs (e.g., B, F, and H) having an even greater percentage of counties with CRP, did not have CRP area as a predictor in most of the best models for any of the crops. The reason could be that CRP area was about ≤20% of the total cropland area of those LRRs.

Quality Control/Quality Assurance of Final Crop Area

The relative errors produced from the mixed models were much lower than the default option (i.e., if the gaps were simply filled with the mean of the observed crop area; Table 4). Occasionally, using the best mixed model yielded outliers, especially if only one or very few observed data were present during the entire time series. This trend was obvious for certain counties in LRR U for alfalfa hay, LRR O for corn for silage, and LRR K for sorghum. However, the number of outliers in each crop was <1% of the total predicted values.

Table 3. Akaike Information Criterion (AIC) values from different model options for wheat area in some land resource regions (LRRs). The AIC values of the best models are shown in italics.

Model†	LRR A	LRR B	LRR C	LRR D	LRR E
All variables	419.38	316.36	539.72	3408.63	2506.07
Diesel$, County_area	435.48	412.98	567.88	*3404.77*	2506.39
Fert$, County_area	445.97	418.20	567.67	3409.26	*2502.34*
Crop$, County_area	420.17	324.38	542.40	3414.19	2509.87
P, County_area	445.06	420.50	566.93	3426.76	2510.64
CRP, County_area	445.50	414.99	566.84	3428.41	2514.13
Diesel$, Fert$, Crop$, County_area	417.47	*313.08*	536.78	3406.57	2505.17
Diesel$, Fert$, Diesel$ × Fert$, Crop$, County_area	419.05	313.72	*536.05*	3408.50	2506.90
Crop$, P, CRP, County_area	421.95	328.38	545.46	3415.79	2510.45
Diesel$, Fert$, Crop$, CRP, County_area	419.42	314.49	537.80	3408.34	2505.24
Diesel$, Fert$, Crop$, P, County_area	*417.42*	314.96	538.70	3406.89	2505.87

† Diesel$ = diesel price of the previous year; Fert$ = fertilizer price of the previous year; Crop$ = crop price of the previous year; CRP = crop area set aside under the Conservation Reserve Program during the current year; County_area = area of the county; P = precipitation of the previous year.

Table 4. Percentage relative errors [i.e., (observed − predicted) × 100/observed] for predicted values of crop area from the mixed-effect model runs compared with predicted values from the default method (i.e., mean of all the observed values).

Crop	Relative error with the mixed model	Relative error with the default option
Alfalfa hay	−2.6	−10.4
Barley	−10.5	−41.8
Corn for grain	−7.3	−26.8
Corn for silage	−8.2	−21.1
Oat	−10.0	−41.2
Other hay	−3.4	−16.1
Sorghum	−9.9	−52.7
Soybean	−6.3	−33.5
Wheat	−8.9	−30.3

The gaps in the few counties that had outliers were filled using the default method (mean of the observed values). Overall, 83% of the total gaps in the crop areas reported by NASS (and 98.5% of the remaining gaps in NASSus database) were filled with the linear mixed effect model approach and only 1% of crop area gaps in NASS were filled using the default method. Thus, by filling the missing annual data in those counties where the crops are present (Fig. 8), we were able to create complete yield and area data for all the major crops concerned (Fig. 8). Table 5 provides a summary of the gaps filled in the crop area of each crop.

While significant at the national level, the consequence of gap filling of crop areas was even greater for certain crops and local areas. The impact was mostly obvious for alfalfa hay, other hay, and corn for silage. The gap filling increased the county-level total alfalfa hay area accounted for by 26 to 46% for most years; similarly, the total area of other hay increased by >50% and that of corn for silage increased by up to 26% across the different years. In comparing the final crop area with the cropland area reported by NRI points at state level, we found that out of the 48 states in the USA, 13 states had total crop area (all the crops combined together) that marginally exceeded the NRI cropland area. How-

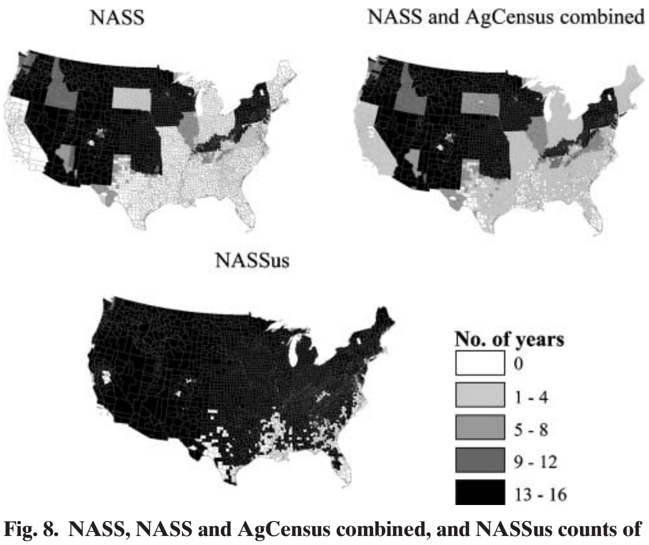

Fig. 8. NASS, NASS and AgCensus combined, and NASSus counts of years with data on alfalfa hay yields and crop area.

Table 5. Percentage gaps† filled by the different methods.

Crop	Gaps in NASS	% gaps filled by combining NASS + AgCensus databases	% gaps in NASS filled by mixed models	% gaps in NASS filled by the default option‡
Alfalfa hay	23017	20.8	77.8	1.4
Barley	7049	3.2	95.0	1.9
Corn	7822	10.9	87.8	1.3
Corn for silage	12417	18.6	81.0	0.5
Oat	9626	7.5	89.9	2.7
Other hay	27554	24.8	75.1	0.0
Sorghum	8938	2.5	94.6	2.9
Soybean	5854	7.8	89.0	3.2
Wheat	7371	10.5	89.4	0.1

† Gaps correspond to the number of missing data in different years during the 16-yr period in all the counties where each crop is grown.

‡ Only the outliers of the predicted values from the mixed models were filled with the default option (i.e., means of the observed values).

ever, the final total cropland area aggregated at state level was very close to the state-level cropland area according to the NRI, with an R^2 exceeding 99%. This was observed for 1982, 1992, and 1997, during years which NRI had also reported data (Fig. 9).

DISCUSSION

Out of the 3111 counties in the conterminous USA, only 67 counties had no crops reported. An initial evaluation of the existing discrepancies between the two main crop statistical databases, NASS and AgCensus, showed that most of the data were very close and comparable (Table 1). However, the difference in the survey methods yielded occasional, extremely different values in NASS compared with AgCensus. Certain differences in the county-level crop statistics could also be attributed to the way NASS and AgCensus reports the farm (or farmer) information; if a farmland extends over several counties, NASS surveys use the location of the household as the location of the farmland. However, for AgCensus, the county in which the operator earns most of his income is reported. This discrepancy becomes visible in those counties with little agriculture (R. Korkosh, 2004, personal communication). For some crops, NASS does not report county-level data for certain states where the crop is present, and overall, NASS has a significant number of missing data at county level. NASS reporting is also restricted by Title 13 of the U.S. code that stipulates that data are not to be published if it would disclose the operations of a single farm within a county, but it is permitted to release the number-of-farms information observed for a county (Griffith, 1999). This is another reason for county-level missing data in NASS. Initially, we filled the gaps in NASS using the data from AgCensus; however, <20% of the total gaps in NASS data could be filled using the AgCensus information.

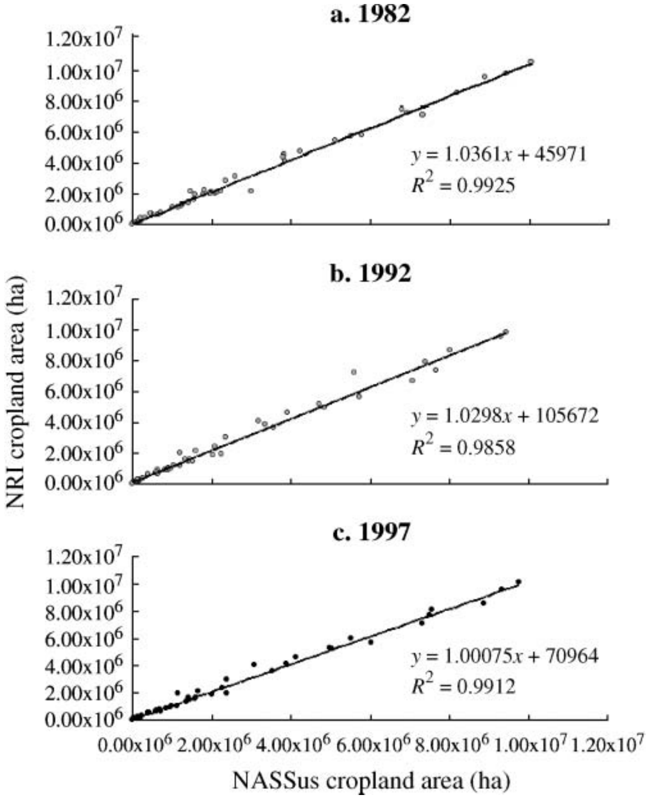

Fig. 9. Final cropland area (NASSus) of all the major crops for 1982 (top), 1992 (middle), and 1997 (lower), aggregated at state-level, plotted against the state-level U.S. total cropland area according to the National Resources Inventory (NRI).

Using linear mixed-effect models with environmental, management, and economic variables to impute missing data yielded lower relative errors compared with a default method of simply using the mean of the observed values for a county. Overall, the linear mixed effect model approach filled >80% of the total gaps in NASS data. In a few instances, where county data were very sparse, models needed to be modified by dropping either autoregression or random effects to meet the convergence criteria. Less than 1% of the missing or imputed values were filled using the county-level means as a default method. Availability of a very low number of observations has been problematic in certain other studies as well. According to Tao et al. (2005), the CASA model overestimated yields in areas with few observations, while it performed better in areas with dense crop coverage.

Using the AR(1) covariance structure yielded predicted values that showed a good time correlation or trend in the predicted yields (Fig. 8). Overall, the mixed model approach seemed to perform well, and only a handful of outliers had to be replaced with the means of the observed values. Compared with other complex simulation models, our approach was simpler, with fewer parameters. It also incorporated essential environmental and economic factors, in addition to spatial and temporal autocorrelation effects.

Incorporation of the county area was essential in the models for predicting the crop areas. In our study, we incorporated log-transformed county area as a predictor variable, while Griffith (1999) had considered the density of area (by dividing by the county area) to incorporate any effect from the size of a county. Griffith (1999) had taken the relationship between an agricultural commodity and the number of farms producing that commodity, along with the spatial autocorrelation in the statistical models used in small area estimation in

Michigan and Tennessee. With the model options having log-transformed county area, we got better relative errors compared with having the area density as the dependent variable.

Since no other ground-based database is available (except for NASS and AgCensus) to compare the final results at county-level, we aggregated the predicted crop area at state-level, and compared those with the state-level cropland area based on the NRI. The final crop area for different crops at state-level was very close, although there were slight differences in certain crops at state level. This could be mostly due to differences in the reporting by NRI and NASS or AgCensus, especially in terms of the differences of small grain crops (due to differences in sampling time), and differences in reporting hay crop categories. The total cropland from all the major crops according to our final crop area (after filling all gaps) were very close to the cropland area from NRI estimates ($R^2 = 0.99$).

CONCLUSIONS

Overall, the methodology we used in this study enabled us to reach the goal of creating complete county-level yield and acreage datasets for major crops in the USA. The effect of gap-filling was greater for certain counties and certain crops, especially for hay crops in states where NASS does not report county-level data, and certain counties with small crop areas. The use of environmental, economic, and management variables in linear mixed models, while taking the spatial and temporal correlation into account, allowed filling the largest proportion of the data gaps; regression analyses with AgCensus also helped fill a significant portion of the gaps during 1982, 1987, 1992, and 1997.

The new NASSus database provides an improved ground-based estimate of cropland productivity that can be used to compare with remote-sensing based estimates and/or to assess temporal and spatial trends in primary productivity and carbon cycling in U.S. cropland.

ACKNOWLEDGMENTS

We are grateful for the assistance and support given by Mark Easter, John Norman, and S.R. Koontz of Colorado State University, Roy Korkosh of the NASS-USDA, Hari Eswaran of NRCS/USDA, Paul Hesse of National Energy Information Center, Sian Mooney of University of Wyoming, Shawn Bucholtz and Alexander Barbarika of the FSA-USDA, and Wen Huang, Tim Parker, and Daniel Hellerstein of the ERS-USDA. Support for this work was provided by the Consortium for Agricultural Soil Mitigation of Greenhouse Gases funded by Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement No. 2001-38700-1109.

REFERENCES

Allen, R., G. Hanuschak, and M. Craig. 2002. History of remote sensing for crop area in USDA's National Agricultural Statistics. Available at www.usda.gov/nass/nassinfo/remotehistory.htm [accessed 25 Jan. 2006; verified 1 Feb. 2007]. USDA-NASS, Washington, DC.

Barkley, M. 2002. Pricing corn silage. Penn State Cooperative Extension Office in Bedford County, Bedford, PA.

Bauer, M.E., M.M. Hixson, B.J. Davis, and J.B. Etheridge. 1978. Area estimation of crops by digital analysis of Landsat data. Photogramm. Eng. Remote. Sens. 44:1033–1043.

Berka, L.M.S., B.F.T. Rudorff, and Y.E. Shimabukuro. 2003. Soybean yield estimation by an agrometeorological model in a GIS. Sci. Agric. (Piracicaba, Braz.) 60:433–440.

Cavero, J., E. Playan, N. Zapata, and J.M. Faci. 2001. Simulation of maize grain yield variability within a surface-irrigated field. Agron. J. 93:773–782.

Csornai, G., C. Wirnhardt, Z. Suba, P. Somogyi, G. Nádor, L. Martinovich, L. Tikász, A. Kocsis, B. Tarcsai, and G. Zelei. 2002. Remote sensing based crop monitoring in Hungary. Available at www.fomi.hu/internet/magyar/Projektek/remsensmonit.htm [accessed 10 Jan. 2005; verified 1 Feb. 2007]. Inst. of Geodesy, Cartography, and Remote Sensing, Budapest.

Daly, C., R.P. Neilson, and D.L. Phillips. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J. Appl. Meteorol. 33:140–158.

Doraiswamy, P.C., J.L. Hartfield, T.J. Jackson, B. Akhmedov, J. Prueger, and A. Stern. 2004. Crop condition and yield simulations using Landsat and MODIS. Remote Sens. Environ. 92:548–559.

Doraiswamy, P.C., S. Moulin, P.W. Cook, and A. Stern. 2003. Crop yield assessment from remote sensing. Photogramm. Eng. Remote Sens. 69(6):665–674.

Doraiswamy, P.C., T.R. Sinclair, S. Hollinger, B. Akhmedov, A. Stem, and J. Prueger. 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sens. Environ. 92:192–202.

Economic Research Service. 2005. Conservation Reserve Program Database. Available at http://usda.mannlib.cornell.edu/ [accessed 15 May, 2005; verified 1 Feb. 2007]. USDA, Washington, DC.

Gonzalez-Alonso, F., J.M. Cuevas, R. Arbiol, and X. Baulies. 1997. Remote sensing and agricultural statistics: Crop area estimation in north-eastern Spain through diachronic Landsat TM and ground sample data. Int. J. Remote Sens. 18(2):467–470.

Griffith, D.A. 1999. A methodology for small area estimation with special reference to one-number agricultural census and confidentiality: Results for selected major crops and states. RD Res. Rep. RD-99–04. National Agricultural Statistics Service, USDA, Washington, DC.

Hixson, M.M., B.J. Davis, and M.E. Bauer. 1981. Sampling Landsat classifications for crop area estimation. Photogramm. Eng. Remote Sens. 47:1343–1348.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS system for mixed models. SAS Inst., Cary, NC.

Lobell, D.B., J.A. Hicke, G.P. Asner, C.B. Field, C.J. Tucker, and O. Los. 2002. Satellite estimates of productivity and light use efficiency in United States agriculture, 1982–98. Glob. Change Biol. 8:722–735.

Lokupitiya, R.S., E. Lokupitiya, and K. Paustian. 2006. Comparison of missing value imputation methods for crop yield data. Environmetrics 17:339–349.

MacDonald, R.B., and F.G. Hall. 1980. Global crop forecasting. Science (Washington, DC) 208:670–679.

Pawel, D., and R. Fecso. 1988. On the use of correlations in crop yields. Proc. of the Survey Research Methods Section. Am. Statistical Assoc., Alexandria, VA.

Prince, S.D., J. Haskett, M. Steninger, H. Strand, and R. Wright. 2001. Net primary production of US Midwest croplands from agricultural harvest yield data. Ecol. Appl. 11:1194–1205.

Reynolds, C.A., M. Yitayew, D.C. Slack, C.F. Hutchinson, A. Huete, and M.S. Petersen. 2000. Estimating crop yields and production by integrating the FAO crop specific water balance model with real-time satellite data and ground-based ancillary data. Int. J. Remote Sens. 21(18):3487–3508.

Rudorff, B.F.T., and G.T. Batista. 1990. Spectral response of wheat and its relationship to agronomic variables in the tropical region. Remote Sens. Environ. 31(1):53–63.

Rudorff, B.F.T., and G.T. Batista. 1991. Wheat yield estimation at the farm level using TM-Landsat and agrometeorological data. Int. J. Remote Sens. 12(12):2477–2484.

Smith, R.C.G., J. Adams, D.J. Stephens, and P.T. Hick. 1995. Forecasting wheat yield in a Mediterranean-type environment from the NOAA satellite. Aust. J. Agric. Res. 46:113–125.

Tan, G., and R. Shibasaki. 2003. Global estimation of crop productivity and the impacts of global warming by GIS and EPIC integration. Ecol. Modell. 168:357–370.

Tao, F., M. Yokozawa, Z. Zhang, Y. Xu, and Y. Hayashi. 2005. Remote sensing of crop production in China by production efficiency models: Model comparisons, estimates and uncertainties. Ecol. Modell. 183:385–396.

Thornthwaite, C.W. 1948. An approach toward a rational classification of climate. Geogr. Rev. 38(1):55–94.

USDA. 1981. Land resource regions and major land resource areas of the United States. Agriculture Handbook 296. USDA, Washington, DC.

USDA. 1982–1997. Agricultural prices. Annual summary 1981–Annual summary 1996. Available at http://usda.mannlib.cornell.edu/usda/nass/AgriPricSu/ [accessed 05 May, 2005; verified 12 Feb. 2007]. USDA, Washington, DC.

USDA. 1998. USDA's National Agricultural Statistics Service: The fact finders of agriculture. USDA, NASS, Washington, DC.

Yang, P., G.X. Tan, Y. Zha, and R. Shibasaki. 2004. Integrating remotely sensed data with an ecosystem model to estimate crop yield in North China. Paper presented at the Geo-Imagery Bridging Continents XXth ISPRS Congr., Istanbul, Turkey. 12–23 July 2004. www.isprs.org/istanbul2004/comm7/papers/29.pdf [accessed 25 Jan. 2006; verified 20 Feb. 2006]. Int. Soc. for Photogrammetry and Remote Sensing, Rockville, MD.