AJAE Appendix for "Agricultural Payments and Land Concentration: A Semi-parametric Spatial Regression Analysis"

Michael J. Roberts and Nigel Key[♦]

November 20, 2007

Note: The material contained herein is supplementary to the article named in the title and published in the American Journal of Agricultural Economics (AJAE).

[♦] The authors are economists at the USDA Economic Research Service. Views expressed are the authors and not necessarily those of the USDA.

This appendix reports documentation of our methods and reports summary statistics and results from supplementary analyses excluded from the main AJAE paper due to space limitations. Summary statistics for the different years and payment quintiles are given in Table A1. Figures A1 and A2 show features of the estimated GAM model reported in the main paper. Tables A2, A3, and A4 report results from supplementary analyses.

We first document some technical details of the generalized additive model (GAM) and present figures showing fitted non-parametric components of a key model reported in the main paper. We then report results from a linear model with state fixed effects for comparison to the GAM and a series of regional models that show our main finding—that concentration growth has occurred much more rapidly in areas with higher per-acre payments—is both extensive and robust: we find the relationship both across and within regions and with a set of non-parametric control functions that are even more flexible than the pooled model reported in the main paper.

The Generalized Additive Model

As specified in the main paper, our GAM model is:

$$\Delta c_i = \mathbf{X}_i\boldsymbol{\beta} + f(x_i,y_i) + g_c(c_{0i}) + g_a(a_{0i}) + g_s(s_{0i}) + \varepsilon_i \qquad (2)$$

where Δc is the percent change in concentration, $(c_1 - c_0)/ \frac{1}{2}(c_1 + c_0)$, $f(x,y)$ is a smooth function of zip code centroids (x, y), and $g_c(c_0)$, $g_a(a_0)$, and $g_s(s_0)$ are smooth functions of initial concentration (c_0), ratio of cropland to zip code area (a_0), and sales-per-acre (s_0), respectively. These control variables were chosen because they are likely correlated with land quality and land-quality is the most plausible source of a spurious correlation between payments and concentration growth (land quality is linked to payment levels because of the way agricultural program are designed and land quality may also be tied to concentration growth via technological or demographic channels).

Payment effects are estimated with the parametric component of the model, $\mathbf{X}_i\boldsymbol{\beta}$, to facilitate ease of interpretation. The matrix **X** includes indicator variables detonating zip codes with zero payments and each of five payment quintiles, and $\boldsymbol{\beta}$ is a vector of payment-category effects. Although the payment effects are parametric, because we have divided observations into six discrete ordered payment groups, the form of the relationship is flexible. The additive separability of the non-parametric partial effects constrains the functional form, but much less than a standard linear model.[1]

One may think of the spatial surface $f(x, y)$ as 'smoothed' location fixed effects. Using state fixed effects rather than the smooth spatial surface creates false discontinuities near state borders, which reduces efficiency and may induce bias. The smooth non-parametric surface eliminates these sharp discontinuities. The smooth functions of the other control variables allow for non-linearities and capture effects of high-leverage points (those far from the mean).

We estimate the smooth functions using "loess", short for "local polynomial regression," which fits the smooth functions by repeatedly estimating weighted linear regressions using only points local to each fitted point. Locations of the fitted points are selected such that divide the covariate space into sections with similar numbers of observations using a "k-d tree."[2] Each fitted point then is estimated with a separate simple weighted linear regression using observed points local to each fitted point. Among the observed points considered local, the weights are higher for observed points near the fitted point as compared to observed points further away. This process is repeated for all fitted points. Points between the fitted points are estimated by interpolation of the fitted points. Loess or any other standard non-parametric procedure (such as cubic splines) can by used to estimate each non-parametric component of the model, although loess is more robust to outliers due to a re-weighting scheme. The smooth functions are

[1] Two-dimensional smooth functions (like $f(x,y)$ in our model) are now easily estimable with standard non-parametric techniques, even for large data sets. This was more difficult with large data sets a decade ago. Smooth functions of three or more dimensions are extremely expensive and will not be feasible for large data sets for many years. High dimensional smooth estimates also require a lot of data to adequately fill the volume of the space.

[2] Loosely speaking, a k-d tree divides the covariate space into sections using quantiles of the covariate joint distribution so that roughly equal numbers of observed points will lie in spaces between fitted points. In multiple dimensions, this can be somewhat complicated. See Cleveland and Grosse (1991).

estimated simultaneously with payment effects parameters (β) using a Gauss-Seidel backfitting method, as described and implemented by Hastie. See this reference for more technical details about the procedure as originally conceived and Wood (2006) for modern implementation using R software.[3]

The key modeling decision concerns the share of points considered local to each fitted point on the smooth functions. For the models reported in the main paper, each fitted point on the smooth spatial surface was estimated using the nearest 5 percent of zip codes, which is the smallest share that was computationally feasible for the two-dimensional spatial surface covering all sample zip codes in the U.S. For consistency, we used the same share for the one-dimensional concentration function.[4] The other key modeling decisions include the number of fitted points used to estimate each curve or surface and the kernel used to weight observations according to their distance from each fitted point. These decisions generally have little influence on results. For these we used the default values in our software package.[5]

With respect to the control variables (location, initial concentration, sales-per-acre, and the ratio of cropland are to zip code area) the generalized additive model is very flexible. A potential shortcoming to using such a flexible model is that the many degrees of freedom can limit statistical power or prevent identification of the model altogether. This is not a problem in the current application because the sample is large (approximately 21,500 zip codes in each panel). Because the purpose of using non-parametric controls is to check the robustness of our estimates, making the controls as flexible as possible lends greater credibility to the estimated effects of payments. Moreover, our focus is on payment effects, not effects of the controls, so

[3] Briefly, the backfitting algorithm first fits the parametric components of the model and then uses the residuals to estimate the first additively separable non-parametric function; the residuals from non-parametric estimates are then used to estimate the second non-parametric function; the parametric components are then re-estimated by subtracting the fitted values of the two non-parametric function from the dependent variables; and so on, iterating until estimated functions on successive iterations converge.

[4] Depending on the application, many model-selection criteria have been developed to aid model selection, including cross-validation and methods based on the Akaike information criterion (AIC) or unbiased risk estimator (UBRE) (Hurvich and Simonoff, 1998; Wood, 2004). Our choice of 5% for the single-dimension smoothes appear to over-fit. Given the large number of observations and a principal focus on payment effects rather that control effects, the resulting loss in degrees of freedom comes at a low cost.

[5] The software package used was the public domain package 'R' with the 'gam' package by Hastie (see www.r-project.org).

tangible interpretation of the non-parametric components is less important than for the parametric components.

In the main paper we reported estimated payment effects for models relating two concentration growth measures (for cropland and farmland) for each 2-year panel and a long-run panel examining concentration changes from 1987 to 2002. We also reported summaries of the overall fit and statistical significance of each non-parametric function. Here we present figures showing the smooth functions for the long-run analysis of cropland. Figure A1 shows estimates of the single dimension functions, $g_c(c_0)$, $g_a(a_0)$, and $g_s(s_0)$, and figure A2 shows a contour plot of the smooth function of location, $f(x,y)$. Figure A1 shows a strong downward slope of $g(c_0)$, indicating that zip codes with the higher initial concentration generally have less growth in concentration. A logical interpretation of this relationship is regression toward the mean, or trend. Particularly at the zip code level, changes in concentration growth are not entirely permanent, perhaps reflecting, in part, transitory events or even response errors, so extreme values are likely to moderate over time. The negative relationship is not surprising given the large standard deviation of concentration growth rates across zip codes (about 81 and 72 percentage points for long panels of cropland and farmland concentration growth, respectively). Figure A1 also shows the fitted smooth curves of initial crop sales per-acre of cropland (s_0) and the initial ratio of cropland area to zip code area (a_0), which are less statistically significant than initial concentration. Figure A2 displays the contour plot of the fitted spatial surface $f(x, y)$ for the long panel of cropland concentration growth. The fitted surface uses between 70.7 and 87.3 non-parametric degrees of freedom. In comparison to state fixed effects, which use 48 degrees of freedom, the surface may be viewed as somewhat more location specific.

Alternative Model Specifications

For comparison to the GAM results, tables A2 and A3 report results from linear models with state fixed effects. Specifically, for each panel and concentration measure, the tables report estimates of the model:

(3) $$\Delta c_i = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{b}_g \log(c_{0i}) + \mathrm{b}_a \log(a_{0i}) + \mathrm{b}_s \log(s_{0i}) + v_s + \varepsilon_i$$

Where $\mathrm{b_g}$, $\mathrm{b_a}$, and $\mathrm{b_s}$ are coefficients on the logs of initial concentration, ratios of cropland area to zip code area, and sales-per-acre, v_s is the state fixed effect, and ε is the error. We take logs of the control variables due their highly skewed distributions. The estimated payment effects are similar to those estimated by the GAM model, but the goodness of fit (R^2) of the models are approximately one-third greater in the GAM models.

Finally, in Table A4 we report estimated concentration growth, adjusted for controls, for GAM models estimated separately for each of the nine USDA-ERS resource regions. The map in figure A3 shows the nine regions. Because payments per acre can vary markedly between regions, we redefine the payment quintiles so they are specific to each region. That is, in each region, each payment quintile includes 20% of the zip codes within that region. Since each GAM regression includes only a fraction of the total number of observations, each fitted point in the smooth loess-estimated functions uses 15% of the local observations in each region rather than 5% of all observations as in the pooled model in the main paper. Thus, despite the higher percent of observations used to estimate each fitted point, the number of points used to fit each point is smaller for the regional regressions than for the pooled regression, and the overall goodness of fit is higher. Estimates of payment effects are not sensitive to the degree of smoothing for shares between 10% and 50%. For shares less than 10%, estimates sometimes do not converge.

The reported percentages in table A4 give predictions for the area-weighted average growth in concentration across all zip codes in each region with all payment levels set the quintile delineated by each the row. The overall estimated payment effects are similar to the pooled model. The estimates also show how concentration growth increases systematically both within and between regions, with low per-acre payment areas (e.g., The Fruitful Rim) having generally lower concentration growth rates than high per-acre payment areas (e.g., The Heartland).

References

Cleveland, W. S. and E. Grosse (1991). “Computational methods for local regression.” *Statistics and Computing* 1 (1): 47-62.

Hurvich, C. M., and Simonoff, J. S. (1998), “Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion” *Journal of the Royal Statistical Society Bulletin* 60, 271-293.

Wood, S.N. 2004. “Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models.” *Journal of the American Statistics Association* 99:673-686.

Wood S.N. (2006) *Generalized Additive Models: An Introduction with R*. CRC Press

Table A1. Distribution of Zip Codes, Farms, and Land by Payments-Per-Acre Category

Panel Years	Payments per Acre of Cropland/Farmland in Beginning Year					
	No Payments	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Cropland						
1987-1992						
Payments per acre	0	0.01-5.75	5.76-15.74	15.74-27.26	27.26-41.81	>41.81
% of zip codes	9.2	18.2	18.2	18.2	18.2	18.2
% of farms	2.3	15.2	21.0	20.8	20.7	20.0
% of cropland	0.6	4.9	9.6	21.9	31.0	32.0
1992-1997						
Payments per acre	0	0.01-3.27	3.28-7.16	7.17-11.28	11.28-16.97	>16.97
% of zip codes	10.5	17.9	17.9	17.9	17.9	17.9
% of farms	2.7	18.0	21.4	21.1	20.2	16.5
% of cropland	0.6	6.1	12.2	23.4	31.1	26.5
1997-2002						
Payments per acre	0	0.01-2.99	3.00-6.82	6.83-10.43	10.44-14.70	>14.70
% of zip codes	9.7	18.1	18.1	18.1	18.1	18.1
% of farms	2.5	16.7	20.3	21.7	21.7	17.2
% of cropland	0.5	5.4	12.5	25.0	31.1	25.5
Long panel						
Payments per acre	0	0.01-5.75	5.76-15.74	15.74-27.26	27.26-41.81	>41.81
% of zip codes	9.2	18.2	18.2	18.2	18.2	18.2
% of farms	2.3	15.2	21.0	20.8	20.7	20.0
% of cropland	0.6	4.9	9.6	21.9	31.0	32.0
Farmland						
1987-1992						
Payments per acre	0	0.01-1.06	1.07-4.18	4.19-10.90	10.91-22.41	>22.41
% of zip codes	9.4	18.1	18.1	18.1	18.1	18.1
% of farms	2.3	15.8	19.7	20.0	21.1	21.1
% of farmland	3.5	21.7	17.7	19.5	20.2	17.4
1992-1997						
Payments per acre	0	0.01-0.65	0.66-2.12	2.13-4.87	4.88-9.29	>9.29
% of zip codes	10.7	17.9	17.9	17.9	17.9	17.9
% of farms	2.8	18.1	20.4	19.5	20.6	18.7
% of farmland	2.1	22.6	18.4	18.6	20.0	18.3
1997-2002						
Payments per acre	0	0.01-0.58	0.59-1.98	1.99-4.69	4.70-9.11	>9.11
% of zip codes	9.8	18.0	18.0	18.0	18.0	18.0
% of farms	2.5	16.5	20.2	20.1	21.1	19.7
% of farmland	2.7	24.1	17.5	19.4	19.3	17.0
Long panel						
Payments per acre	0	0.01-1.06	1.07-4.18	4.19-10.90	10.91-22.41	>22.41
% of zip codes	9.4	18.1	18.1	18.1	18.1	18.1
% of farms	2.3	15.8	19.7	20.0	21.1	21.1
% of farmland	3.5	21.7	17.7	19.5	20.2	17.4

Note: All payments converted to 1997 dollars using the consumer price index.

Figure A1. Estimated Non-Parametric Controls for 1987-2002 Change in Cropland Concentration

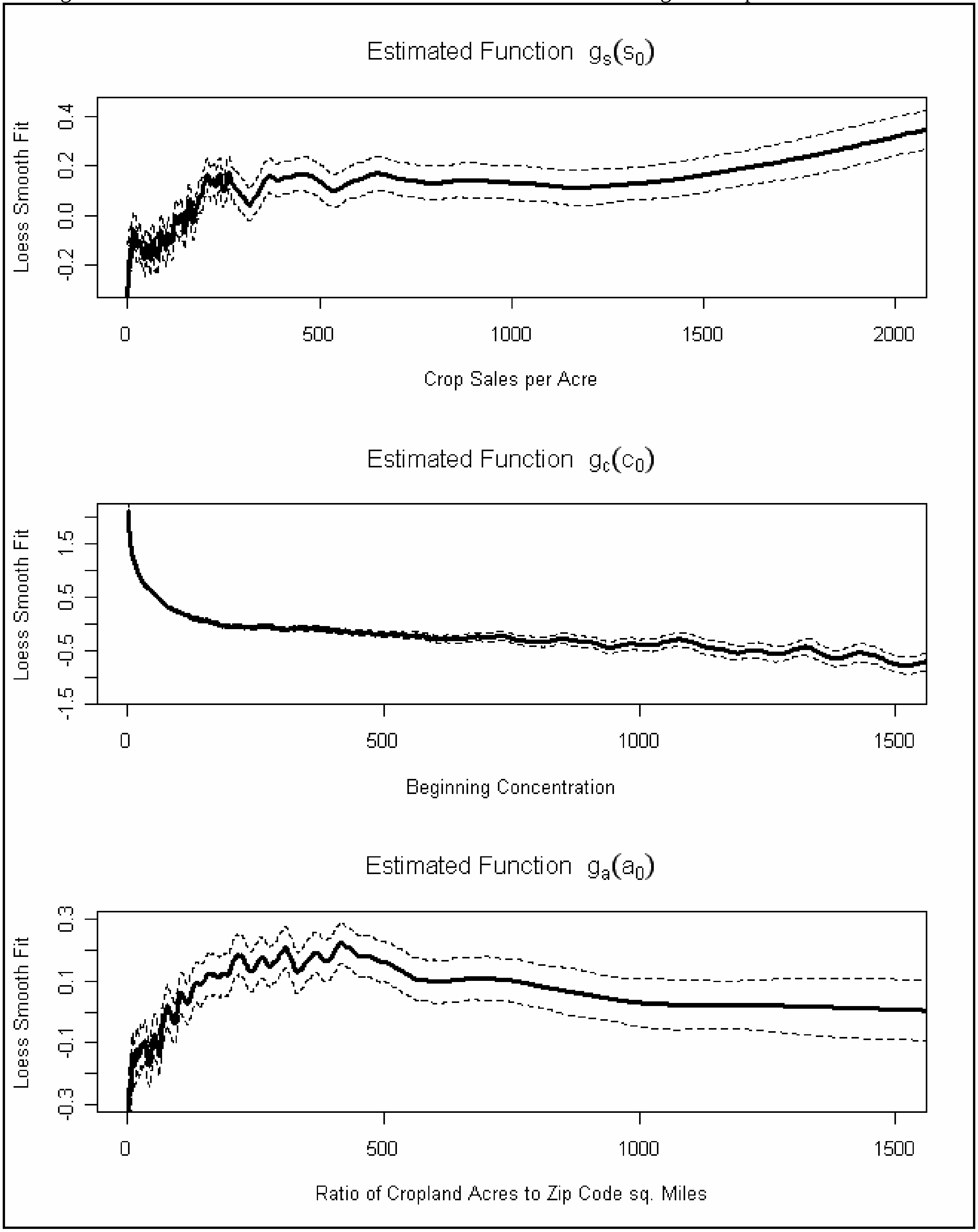
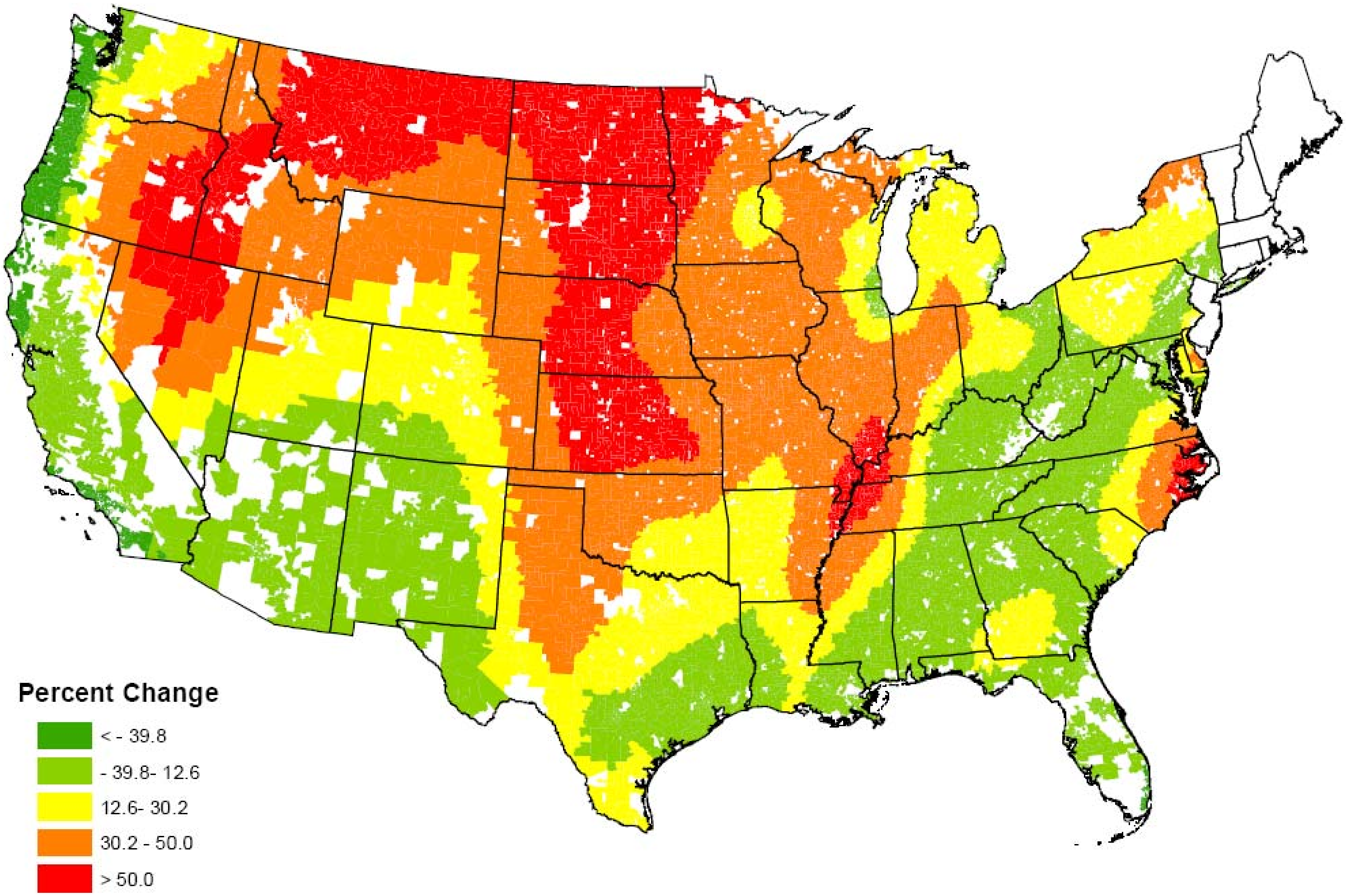


Figure A2. Estimated Spatial Surface, $f(x,y)$ for pooled model of 1987-2002 Change in Cropland Concentration.



Notes: White indicates zip codes dropped from the analysis due to missing observations or extreme outliers. All other variables fixed at population medians.

Table A2. State Fixed Effects Models for Cropland Concentration Growth

Model	*Variable*	*Estimate*	*Std. Error*	*t*
Panel #1	Payment Quintile #1	0.018	0.016	1.12
	Payment Quintile #2	0.084	0.017	5.10
Dependent	Payment Quintile #3	0.187	0.017	10.73
Variable	Payment Quintile #4	0.275	0.018	15.10
	Payment Quintile #5	0.276	0.018	14.97
% Change in Concentration,	Log(1987 Crop Sales/Acre)	0.022	0.004	5.85
1987 - 1992	Log(1987 Concentration)	-0.213	0.005	-46.08
R^2: 0.115 R^2-Adj: 0.113	Log(1987 Cropland Area / Zip Code Area)	0.061	0.004	17.14
Panel #2	Payment Quintile #1	0.022	0.015	1.45
	Payment Quintile #2	0.086	0.016	5.39
Dependent	Payment Quintile #3	0.165	0.017	9.89
Variable	Payment Quintile #4	0.212	0.017	12.39
	Payment Quintile #5	0.237	0.017	14.23
% Change in Concentration,	Log(1992 Crop Sales/Acre)	0.037	0.004	10.13
1992 - 1997	Log(1992 Concentration)	-0.194	0.005	-43.03
R^2: 0.096 R^2-Adj: 0.094	Log(1992 Cropland Area / Zip Code Area)	0.051	0.004	13.73
Panel #3	Payment Quintile #1	0.107	0.018	6.05
	Payment Quintile #2	0.169	0.018	9.22
Dependent	Payment Quintile #3	0.264	0.019	13.61
Variable	Payment Quintile #4	0.318	0.020	15.86
	Payment Quintile #5	0.250	0.020	12.76
% Change in Concentration,	Log(1997 Crop Sales/Acre)	0.042	0.004	10.34
1997 - 2002	Log(1997 Concentration	-0.210	0.005	-43.66
R^2: 0.105 R^2-Adj: 0.103	Log(1997 Cropland Area / Zip Code Area)	0.058	0.004	14.03
Long Panel	Payment Quintile #1	0.193	0.033	5.89
	Payment Quintile #2	0.239	0.033	7.18
Dependent	Payment Quintile #3	0.391	0.034	11.39
Variable	Payment Quintile #4	0.532	0.035	15.16
	Payment Quintile #5	0.569	0.035	16.12
Sum of % Change in Concentration,	Log(1987 Crop Sales/Acre)	0.053	0.006	8.99
1987 – 1992, 1992 – 1997, 1997 – 2002	Log(1987 Concentration)	-0.276	0.006	-43.90
R^2: 0.146 R^2-Adj: 0.143	Log(1987 Cropland Area / Zip Code Area)	0.091	0.005	18.87

Notes: This table summarizes results from ordinary least squares regressions with state fixed effects. The estimated effects of the payment quintiles are relative to zip codes with no payments.

Table A3. State Fixed Effects Models for Farmland Concentration Growth

Model	*Variable*	*Estimate*	*Std. Error*	*t*
Panel #1	Payment Quintile #1	0.041	0.014	2.85
	Payment Quintile #2	0.036	0.015	2.42
Dependent	Payment Quintile #3	0.069	0.016	4.33
Variable	Payment Quintile #4	0.113	0.017	6.69
	Payment Quintile #5	0.163	0.018	8.90
% Change in Farmland	Log(1987 Crop Sales/Acre)	0.005	0.003	1.39
Concentration, 1987 – 1992	Log(1987 Concentration)	-0.174	0.003	-52.85
R^2: 0.128	Log(1987 Cropland Area / Zip Code Area)	0.001	0.003	0.18
R^2-Adj: 0.126				
Panel #2	Payment Quintile #1	0.046	0.014	3.26
	Payment Quintile #2	0.035	0.014	2.48
Dependent	Payment Quintile #3	0.099	0.015	6.59
Variable	Payment Quintile #4	0.149	0.016	9.22
	Payment Quintile #5	0.213	0.017	12.58
% Change in Farmland	Log(1992 Crop Sales/Acre)	0.013	0.003	3.91
Concentration, 1992 - 1997	Log(1992 Concentration)	-0.175	0.003	-51.45
R^2: 0.133	Log(1992 Cropland Area / Zip Code Area)	-0.008	0.003	-2.50
R^2-Adj: 0.130				
Panel #3	Payment Quintile #1	0.080	0.015	5.17
	Payment Quintile #2	0.057	0.016	3.57
Dependent	Payment Quintile #3	0.126	0.017	7.43
Variable	Payment Quintile #4	0.200	0.018	11.03
	Payment Quintile #5	0.266	0.020	13.60
% Change in Farmland	Log(1997 Crop Sales/Acre)	-0.002	0.004	-0.52
Concentration, 1997 – 2002	Log(1997 Concentration	-0.181	0.004	-50.00
R^2: 0.124	Log(1997 Cropland Area / Zip Code Area)	-0.006	0.003	-1.89
R^2-Adj: 0.121				
Long Panel	Payment Quintile #1	0.186	0.028	6.59
	Payment Quintile #2	0.149	0.029	5.18
Dependent	Payment Quintile #3	0.210	0.030	7.11
Variable	Payment Quintile #4	0.360	0.031	11.77
	Payment Quintile #5	0.513	0.032	16.03
Sum of % Change in Farmland	Log(1987 Crop Sales/Acre)	0.007	0.005	1.35
Concentration, 1987 – 1992, 1992 – 1997,	Log(1987 Concentration)	-0.263	0.004	-61.83
1997 – 2002	Log(1987 Cropland Area / Zip Code Area)	-0.003	0.004	-0.70
R^2: 0.204				
R^2-Adj: 0.202				

Notes: This table summarizes results from ordinary least squares regressions with state fixed effects. The estimated effects of the payment quintiles are the proportional change in concentration relative to zip codes with no payments.

Table A4. Regional Analyses of Concentration Growth by Payment Quintile Adjusted with Non-Parametric Controls

Cropland

Within-region payments per acre	*Heartland*	*Northern Crescent*	*Northern Great Plains*	*Prairie Gateway*	*Eastern Uplands*	*Southern Seaboard*	*Fruitful Rim*	*Basin and Range*	*Mississipp i Portal*	*Average weighted by farmland area*
	(Percent change in weighted median, adjusted for controls, 1987-2002)									
No Payments	26.6	10.4	23.6	4.6	6.5	6.3	-18.5	-23.4	27.7	13.7
Quintile #1	37.1	30.1	28.9	20.7	11.6	11.7	-9.3	5.9	26.4	24.7
Quintile #2	56.1	31.9	41.5	30.7	14.3	18.3	-11.3	11.8	38.4	35.9
Quintile #3	59.9	38.8	52.8	40.6	21.6	18.8	-4.6	16.5	41.8	42.3
Quintile #4	59.7	44.3	56.7	45.5	20.9	28.3	4.9	7.6	40.2	45.0
Quintile #5	60.8	49.2	58.1	46.6	32.5	35.6	12.2	21.2	51.1	48.7

Farmland

Within-region payments per acre	*Heartland*	*Northern Crescent*	*Northern Great Plains*	*Prairie Gateway*	*Eastern Uplands*	*Southern Seaboard*	*Fruitful Rim*	*Basin and Range*	*Mississipp i Portal*	*Average weighted by cropland area*
	(Percent change in weighted median, adjusted for controls, 1987-2002)									
No Payments	22.4	-1.5	24.3	-3.9	3.1	0.8	-22.7	-18.7	-2.4	2.4
Quintile #1	10.8	7.6	11.4	-11.5	9.4	3.4	-41.8	-25.5	-0.8	-5.6
Quintile #2	26.9	13.1	17.1	-1.2	10.5	7.9	-31.6	-9.9	15	4.7
Quintile #3	32.1	14.8	26.9	18.4	10.5	10.8	-18.8	-11.9	17.7	13.6
Quintile #4	37.9	20.9	37.4	30.6	12.6	18.7	8.4	9.8	32.7	26.2
Quintile #5	41	24.7	46.5	40.8	25.5	25.6	31.4	31.5	45.3	37.3

Notes: The table reports 1987-2002 concentration growth for each payment quintile of each region, adjusted for region-specific non-parametric controls for location and 1987 levels of concentration, sales-per acre, and the ratio of agricultural area to zip code area. The regions are those defined by USDA's Economic Research Service and are displayed in Figure A3. Where estimates reported in the main paper were based on regressions using the pooled data of all regions, these estimates were derived from estimating a separate GAM regression model for each region. Quintiles of payment levels are also region-specific, so that an equal number of zip codes are in quintile within each region, but the same quintile may represent different payment levels in across regions. Since each regression includes only a fraction of the total number of observations, each fitted point in the smooth loess-estimated functions uses 15% of the local observations in each region rather than 5% of all observations as in the pooled model. Thus, despite the higher percent, the bandwidth is effectively smaller for the regional regressions than it is for the pooled regression, and the overall goodness of fit is higher. Estimates of payment effects are not sensitive to the degree of smoothing for shares between 10% and 50%. For shares less than 10%, estimates sometimes do not converge. See the main paper for a description of the GAM regression model. The reported percentages give predictions for the area-weighted average growth in concentration across zip codes in each region and are scaled so that the estimated average percentage for the third quintile equals the actual area-weighted average growth for the third quintile—that is, non-payment factors are held constant at each region's third-quintile average.

Figure A3. ERS Farm Resource Regions Used for Separate Regional Analyses.