The Case of the Missing Data: Short Samples and Multiple Imputation in Rating Crop Insurance*

Kyoungin Choe

August 12, 2025

Abstract

The federal crop insurance program plays a central role in US farm income support but often faces challenges in rating coverage due to missing county-level yield data. Traditional approaches, such as spatial aggregation, may fail to capture true risk by treating proxy values as observed data. This study applies a multiple imputation (MI) framework to address missing yield data using information from related crops, neighboring counties, and temporal indicators. We construct four imputation scenarios using corn and soybean yield data from Adair County, Iowa, varying the type and scope of supplemental inputs. Results show that MI-based approaches outperform direct proxies in capturing uncertainty and producing more stable parameter estimates. While cross-crop data alone introduce variability, combining spatial, cross-crop, and temporal indicators yields the most robust estimates. These findings highlight the importance of accounting for uncertainty in proxy yields and suggest that MI can improve the actuarial soundness of premium rate determination in data-sparse settings.

Keywords: multiple imputation, crop insurance, missing data
JEL Classification: C81, G22, Q14

*Department of Agricultural Economics, Texas A&M University, College Station, TX. Email: kyoungin.choe@ag.tamu.edu

1 Introduction

The federal crop insurance program has become the cornerstone of farm income support and risk management policy in the United States. The program is administered by the Risk Management Agency (RMA) under the US Department of Agriculture (USDA), which is responsible for establishing insurance premium rates and terms of coverage for all insured crops. According to the RMA, the program currently insures over 490 million acres and covers more than $180 billion in total liability. The Congressional Budget Office (CBO) estimated that the program cost taxpayers $16.66 billion in Fiscal Year 2023, with average annual costs projected at $12 billion over the next decade. Participation in the program is nearly universal, with more than 90% of planted acreage insured for most major crops.

The universal nature of the program—insuring nearly all crops—means that the agency often encounters situations where data are thin or missing. The RMA dataset contains substantial gaps, with approximately 20% of irrigated and 26% of nonirrigated county-level corn yield records missing (Park et al., 2022). The issue is even more pronounced for soybeans, with over 36% of irrigated and nearly 40% of nonirrigated yield data absent. Despite such gaps, the RMA must still establish insurance guarantees, premium rates, and other terms of coverage. To address this, various methods are used to proxy missing data when determining contract parameters. The most common approach is spatial aggregation, whereby yield data are aggregated across increasingly larger geographic areas to estimate expected yields and yield histories.

The shortcoming of this method is that the proxy measures are subsequently treated as if the data were actually observed at a more disaggregated level—typically the county. Chen et al. (2020) find that roughly 40% of counties exhibit some degree of misrating, highlighting the limitations of relying solely on spatial aggregation. However, little attention has been given to measuring the uncertainty and variability of such proxy measures. This approach often overlooks relevant information that could better inform proxies for missing data. For example, consider a corn-producing county that begins

producing a new crop, such as grain sorghum. While little may be known about the yield risks associated with grain sorghum, a considerable yield history may exist for corn. We propose using extensive yield histories from related crops and/or surrounding regions to construct proxy measures for crops with limited or missing data. To this end, we apply multiple imputation methods to generate plausible yield values and quantify the uncertainty introduced by imputation.

Multiple imputation replaces each missing value with a set of plausible alternatives that reflect uncertainty about the true value. Rather than relying on a single proxy, this method uses all available information to generate multiple completed datasets. A reasonably large number of replications of the complete datasets is generated, and predicted yields and associated premium rates are computed from each. This produces a distribution of estimated yields and premium rates, which allows us to quantify how uncertainty from proxy methods affects key insurance parameters.

We apply our approach to county-level corn and soybean yield data collected from the USDA RMA's Information Reporting System, with a particular focus on counties in Iowa.[1] To assess imputation performance, we artificially introduce missingness into the observed data and compare the imputed values against the known yields. As noted, our framework leverages auxiliary information—such as yields from related crops and neighboring counties—to guide the imputation process in data-sparse settings. Based on the imputed yields, we estimate yield distributions, compute loss probabilities, and calculate corresponding premium rates for various coverage levels. This approach allows us to explore how imputation-based methods may enhance premium determination in the presence of incomplete yield data.

This study contributes to the literature on agricultural risk management and crop insurance pricing by addressing a critical yet underexplored issue: how to construct proxy yield measures and quantify their uncertainty when historical data are missing or sparse. While prior research has documented the prevalence of data gaps and the

[1]Data are from the USDA RMA's Information Reporting System, available online at https://webapp.rma.usda.gov/apps/RIRS/RMACountyYieldsReport.aspx.

potential for misrating in premium determination (e.g., Chen et al., 2020; Goodwin et al., 2024), few studies offer statistically rigorous methods for generating proxies that reflect the uncertainty introduced by missing data. Previous work has emphasized the use of supplemental data—such as historical weather—to improve the accuracy of actuarially fair premium rates (e.g., Liu and Ramsey, 2023). Researchers have also proposed spatial smoothing techniques, including Bayesian kriging, to enhance yield estimation in data-scarce regions (e.g., Park et al., 2019). In agricultural contexts, multiple imputation has mostly been applied to survey-based settings to address nonresponse or item-level missingness (e.g., Robbins and White, 2011; Zhong et al., 2018).

To our knowledge, no prior study has addressed the specific challenges posed by missing or thinned yield data for crop insurance rating by employing statistical methods that generate multiple proxy yields using relevant information. We propose the use of multiple imputation as a flexible framework for constructing yield proxies that incorporate yield information from related crops or neighboring counties, while also accounting for the uncertainty introduced by the imputation process. In doing so, our approach improves the accuracy of expected yield estimates and enables the quantification of premium rate variability resulting from data uncertainty. This is particularly important for enhancing actuarial soundness in data-scarce regions or for crops with limited historical records.

The subsequent sections are structured as follows: the next section discusses the empirical setting of our analysis. The third section presents the data and empirical results from the multiple imputation approach and evaluates their performance across scenarios. In particular, we examine the implications of our proposed method for generating crop insurance rate loads that incorporate uncertainty associated with missing yield data. Finally, we summarize the main findings and suggest directions for future research in the concluding section.

2 Empirical Approach

This study employs a multiple imputation (MI) approach to address missing county-level yield data—a common challenge in crop insurance rating. Our methodology explicitly accounts for the uncertainty inherent in missing data and incorporates auxiliary information, including spatial context and yields from related crops. Specifically, we examine how the inclusion of surrounding counties, other crop types, and temporal indicators influences the quality of proxy yield measures and the resulting insurance estimates.

A Markov Chain Monte Carlo (MCMC)-based MI approach is well suited in this context, as it can incorporate auxiliary variables and generate draws from the posterior distribution, reflecting parameter and data uncertainty. We assume the data are missing at random (MAR), meaning that the probability of missingness depends only on observed values (Rubin, 1976).[2] Let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ denote the complete data vector, where Y_{obs} and Y_{mis} represent the observed and missing components, respectively. Missing values are imputed by drawing from the posterior predictive distribution:

$$P(Y_{\text{mis}} \mid Y_{\text{obs}}) = \int P(Y_{\text{mis}} \mid \theta, Y_{\text{obs}}) P(\theta \mid Y_{\text{obs}}) d\theta. \tag{1}$$

The imputation is implemented using an MCMC method, with initial parameter values obtained from the Expectation–Maximization (EM) algorithm. We generate $m = 100$ completed datasets, each representing a plausible realization of the missing values. Each dataset is analyzed independently, producing parameter estimates $\hat{\theta}^{(1)}, \hat{\theta}^{(2)} \ldots, \hat{\theta}^{(m)}$ and their associated variances.

Following Rubin (1976) and Yuan (2011), the total variance of the combined point estimate is:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B, \tag{2}$$

where $\bar{U} = \frac{1}{m} \sum_{l=1}^{m} \widehat{\text{Var}}(\hat{\theta}^{(l)})$ denotes the average within-imputation variance, $B =$

[2] Although we assume a missing-at-random (MAR) mechanism here, the MCMC algorithm in `PROC MI` can accommodate arbitrary missing patterns (Schafer, 1997).

$\frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\theta}^{(l)}-\bar{\theta}\right)^2$ is the between-imputation variance, and $\bar{\theta}=\frac{1}{m}\sum_{l=1}^{m}\hat{\theta}^{(l)}$ is the overall estimate.

To examine the contribution of spatial, cross-crop, and temporal auxiliary information to the imputation of missing yield data, we construct four scenarios: (1) the mean of corn yields from surrounding counties as a direct proxy, without applying MI; (2) an MI approach based solely on the soybean yield from the central county and year; (3) an MI approach using corn yields from neighboring counties, the central county's soybean yield, and year; and (4) an MI approach using corn and soybean yields from surrounding counties, along with year.[3] These scenarios allow us to assess the marginal contribution of each type of auxiliary information to imputation accuracy and evaluate its implications for actuarial estimates in crop insurance.

To evaluate the effect of imputation on crop insurance rating, we fit a Weibull distribution to the imputed county-level yield data in each completed dataset:

$$f(y \mid \alpha, \beta) = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{y}{\beta}\right)^{\alpha}\right], \quad y > 0, \tag{3}$$

where α and β are the shape and scale parameters. Parameter estimates from each imputed dataset are combined using Rubin's rules.

Simulated yields from the fitted Weibull distributions are then used to calculate loss probabilities and premium rates at three common US crop insurance coverage levels—65%, 75%, and 85%. This allows us to evaluate how different imputation strategies affect rating accuracy, risk load estimation, and premium variability.

3 Empirical Results

Our empirical analysis focuses on corn yield data in Adair County, Iowa. Using historical yield data from 1993 to 2023, we artificially set corn yield values in Adair County

[3]In our analysis, Adair County, Iowa is referred to as the central county, and neighboring counties are the seven adjacent counties—Adams, Audubon, Cass, Dallas, Guthrie, Madison, and Union—as illustrated in Figure A1 in the Appendix.

to missing for all years after 2000. County-level corn and soybean yield data were collected from the USDA RMA's Information Reporting System. Adair County has seven neighboring counties—Adams, Audubon, Cass, Dallas, Guthrie, Madison, and Union—which serve as the basis for incorporating spatial auxiliary data.

Table 1 presents summary statistics for county-level corn and soybean yields. Adair County shows moderate mean yields and standard deviations compared to its neighboring counties. Notably, we generally observe consistent patterns across the two crops in yield levels and variability. For example, Audubon County has the highest mean yields for both crops, along with the highest standard deviation for corn and relatively high variability for soybeans. Dallas County shows the lowest standard deviations for both crops, indicating more stable yield patterns over time. Union County has the lowest mean yields for both corn and soybeans. These patterns support the inclusion of cross-crop information in the imputation framework. While yield dynamics differ across crops, similarities in mean yields and variability suggest that soybean yields can better inform proxy measures for missing corn yields.

To evaluate the quality of the imputation process, we begin by examining MCMC diagnostic tests. As shown in Figure 1, the trace plots demonstrate good mixing and the autocorrelation functions decrease rapidly, indicating a low dependence on the lag. These diagnostics provide evidence of convergence and support the validity of the imputation results. Table 2 presents the estimated Weibull parameters, loss probabilities, and premium rates across the four imputation scenarios. In Scenario 1, which uses the mean corn yield from surrounding counties as a direct proxy without applying MI, the shape and scale parameters are relatively high ($\alpha = 5.62$, $\beta = 169.71$). The resulting loss probabilities and premium rates are substantially lower than those in all MI-based scenarios. Specifically, the probability of loss at 75% coverage is only 12.1%, with a corresponding premium rate of 1.85%. These findings may understate actual risk, as the proxy yields do not capture underlying variability.

Scenario 2 yields a lower shape parameter ($\alpha = 3.67$) and a higher scale parameter

($\beta = 217.05$), both with relatively large standard deviations—indicating greater uncertainty and dispersion in the imputed yield distribution. The probability of loss at the 75% coverage level is 43.6%, with a corresponding premium rate of 12.25%. While these values are similar to those in Scenarios 3 and 4, the greater variability in the parameter estimates suggests less precision in the underlying imputation model. This may reflect the limitations of relying solely on cross-crop information, which could be less reliable due to differing yield dynamics between corn and soybeans.

Scenario 3 incorporates corn yields from neighboring counties, along with soybean yields and year indicators. The estimated parameters ($\alpha = 3.73$, $\beta = 143.23$) exhibit much lower standard deviations, reflecting greater stability. Loss probabilities and premium rates remain nearly identical to those in Scenario 2, suggesting that while spatial information improves the precision of parameter estimates, it has limited influence on overall risk predictions.

Scenario 4 uses corn and soybean yields from all surrounding counties, along with year indicators. It yields the highest shape parameter ($\alpha = 7.27$) and the lowest standard deviations for both shape and scale parameters, indicating the most stable estimates across all scenarios. The loss probabilities and premium rates remain similar to those in Scenarios 2 and 3. Thus, while the point estimates of risk are consistent with the other scenarios that incorporate MI, the broader auxiliary information in Scenario 4 enhances the reliability and robustness of the underlying model.

Overall, these findings highlight the importance of auxiliary data when addressing missing yield values. Direct proxies, such as simple spatial averages, may substantially underestimate risk, and cross-crop information alone can introduce bias. Notably, combining spatial, cross-crop, and temporal indicators yields the most reliable estimates for crop insurance rating.

4 Conclusion

The federal crop insurance program often faces challenges in rating coverage in counties with limited historical yield data. Traditional approaches—such as spatial aggregation—may overlook the uncertainty inherent in proxy estimates, potentially resulting in mispricing. To address these limitations, this study introduces a multiple imputation framework that leverages auxiliary information from related crops, neighboring counties, and temporal indicators to construct more reliable yield estimates. We evaluate four imputation scenarios that vary in the type and scope of supplemental data—ranging from simple spatial averages to fully integrated spatial, cross-crop, and temporal inputs.

Using corn and soybean yield data from Adair County, Iowa, our empirical analysis reveals several key insights. Direct proxies without multiple imputation substantially underestimate risk and may result in unrealistically low premium rates. In contrast, MI-based approaches explicitly incorporate uncertainty and produce more stable parameter estimates, particularly when spatial and temporal information is included. While the choice of auxiliary inputs affects model precision, the resulting loss probabilities and premium rates remain broadly consistent across MI scenarios, suggesting that accounting for uncertainty enhances reliability without significantly altering expected outcomes.

The proposed framework offers a practical tool for the Risk Management Agency to improve premium rate accuracy when historical yield data are limited. Compared to conventional approaches, our framework yields more stable and actuarially sound estimates by directly modeling uncertainty and drawing on readily available auxiliary data. Although this study focuses on corn yields in Adair County, the methodology is broadly applicable to other crops—such as soybeans and wheat—and to a broader set of counties and states facing similar data challenges. Future research could evaluate performance under alternative missing data mechanisms, integrate additional auxiliary sources (e.g., weather, soil, or remote sensing), or apply the framework across other crop

insurance programs. These extensions would further strengthen the role of data-driven risk assessment in enhancing the accuracy of crop insurance pricing.

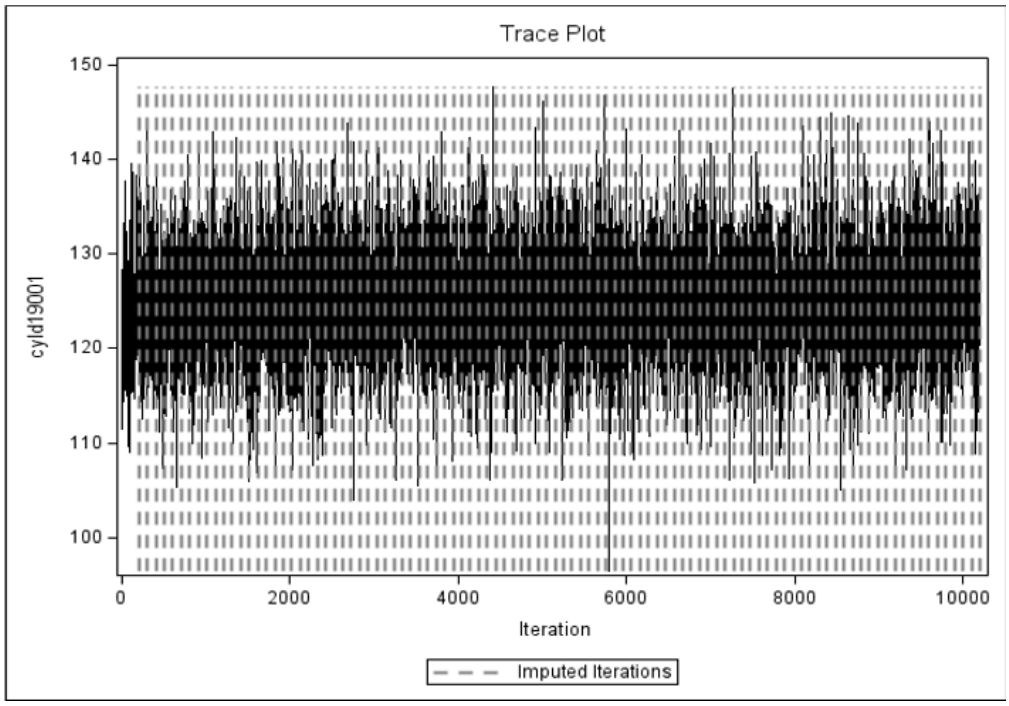
Table 1: Summary Statistics for County-Level Corn and Soybean Yields

County	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**
		Corn		
Adair	156.91	32.56	76.30	223.37
Adams	151.14	30.74	82.40	213.50
Audubon	168.94	38.05	78.60	247.90
Cass	162.24	34.98	75.20	230.60
Dallas	165.99	30.53	95.60	226.60
Guthrie	154.88	37.26	83.00	234.00
Madison	149.93	31.26	79.60	207.90
Union	142.69	34.03	71.70	217.90
		Soybeans		
Adair	47.25	7.98	28.70	63.20
Adams	46.95	7.73	29.50	63.70
Audubon	51.46	8.55	32.00	69.30
Cass	49.33	8.30	31.50	62.10
Dallas	49.07	7.51	32.20	65.00
Guthrie	47.12	9.14	27.60	67.10
Madison	47.20	8.05	27.40	60.40
Union	45.40	9.07	24.70	65.50

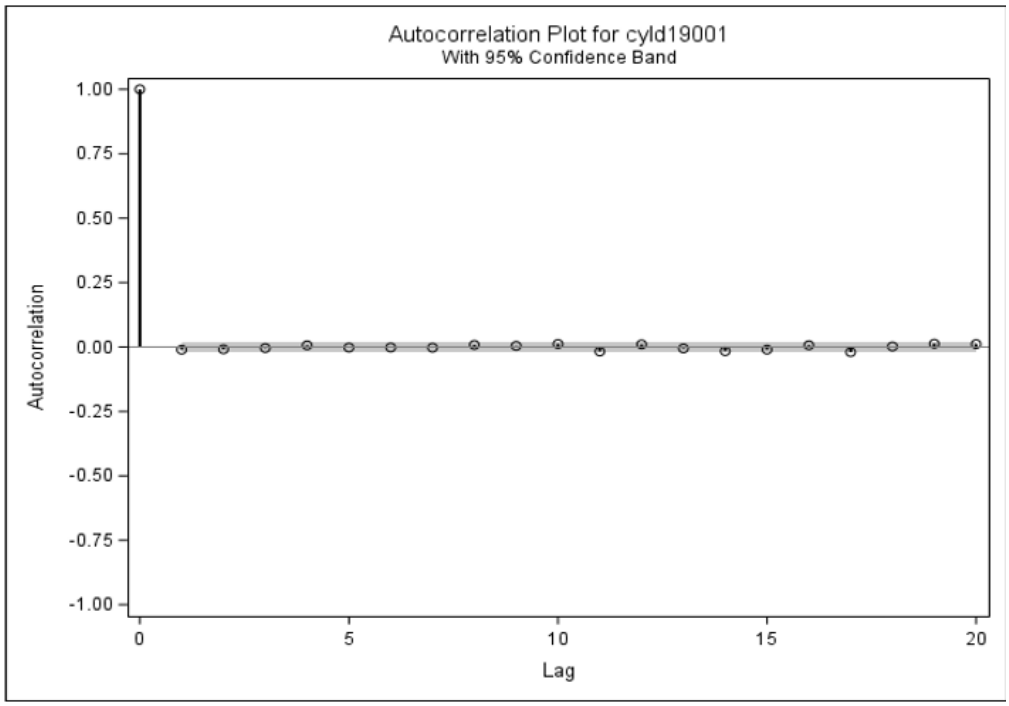
Note: Yields are measured in bushels per acre. Each series contains 31 annual observations from 1993 to 2023. Adair County is the central county in our analysis, and the remaining counties are its neighbors.

Table 2: Imputation Results Across Scenarios: Corn

Parameter	**Scenario**	**Mean**	**Std. Dev.**
Weibull alpha	Using mean of surrounding counties	5.6244	—
Weibull beta	Using mean of surrounding counties	169.71	—
Pr(Loss) at 65% level	Using mean of surrounding counties	0.0547	—
Pr(Loss) at 75% level	Using mean of surrounding counties	0.1211	—
Pr(Loss) at 85% level	Using mean of surrounding counties	0.2325	—
Rate at 65% level	Using mean of surrounding counties	0.0082	—
Rate at 75% level	Using mean of surrounding counties	0.0185	—
Rate at 85% level	Using mean of surrounding counties	0.0366	—
Weibull alpha	MI with soybean yields and year	3.6694	0.3126
Weibull beta	MI with soybean yields and year	217.0512	13.4320
Pr(Loss) at 65% level	MI with soybean yields and year	0.3110	0.0044
Pr(Loss) at 75% level	MI with soybean yields and year	0.4359	0.0052
Pr(Loss) at 85% level	MI with soybean yields and year	0.5653	0.0049
Rate at 65% level	MI with soybean yields and year	0.0840	0.0016
Rate at 75% level	MI with soybean yields and year	0.1225	0.0018
Rate at 85% level	MI with soybean yields and year	0.1670	0.0020
Weibull alpha	MI with surrounding corn, soy, and year	3.7294	0.0068
Weibull beta	MI with surrounding corn, soy, and year	143.2315	0.0061
Pr(Loss) at 65% level	MI with surrounding corn, soy, and year	0.3100	0.0044
Pr(Loss) at 75% level	MI with surrounding corn, soy, and year	0.4346	0.0048
Pr(Loss) at 85% level	MI with surrounding corn, soy, and year	0.5639	0.0048
Rate at 65% level	MI with surrounding corn, soy, and year	0.0839	0.0015
Rate at 75% level	MI with surrounding corn, soy, and year	0.1223	0.0018
Rate at 85% level	MI with surrounding corn, soy, and year	0.1666	0.0019
Weibull alpha	MI with surrounding corn and soy, and year	7.2650	0.0001
Weibull beta	MI with surrounding corn and soy, and year	133.5616	0.0001
Pr(Loss) at 65% level	MI with surrounding corn and soy, and year	0.3115	0.0042
Pr(Loss) at 75% level	MI with surrounding corn and soy, and year	0.4363	0.0049
Pr(Loss) at 85% level	MI with surrounding corn and soy, and year	0.5659	0.0052
Rate at 65% level	MI with surrounding corn and soy, and year	0.0842	0.0015
Rate at 75% level	MI with surrounding corn and soy, and year	0.1227	0.0017
Rate at 85% level	MI with surrounding corn and soy, and year	0.1672	0.0019



(a) Trace Plot



(b) Autocorrelation Plot

Figure 1: MCMC Diagnostic Tests

References

Chen, Z., S. Dall'Erba, and B. J. Sherrick (2020). Premium misrating in federal crop insurance programs: Scale, geography, and fiscal impacts. *Agricultural Finance Review 80*(5), 693–713.

Goodwin, B. K., R. Rejesus, A. Harri, and K. Coble (2024). Directional aspects of crop yield correlation: Implications for insurance rating and contract design. *Journal of the Agricultural and Applied Economics Association 3*(4), 604–617.

Liu, Y. and A. F. Ramsey (2023). Incorporating historical weather information in crop insurance rating. *American Journal of Agricultural Economics 105*(2), 546–575.

Park, E., B. W. Brorsen, and A. Harri (2019). Using bayesian kriging for spatial smoothing in crop insurance rating. *American Journal of Agricultural Economics 101*(1), 330–351.

Park, E., A. Harri, and K. H. Coble (2022). Estimating crop yield densities for counties with missing data. *Journal of Agricultural and Resource Economics 47*(3), 634–655.

Robbins, M. W. and T. K. White (2011). Farm commodity payments and imputation in the agricultural resource management survey. *American Journal of Agricultural Economics 93*(2), 606–612.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Yuan, Y. (2011). Multiple imputation using SAS software. *Journal of Statistical Software 45*, 1–25.

Zhong, H., W. Hu, and J. M. Penn (2018). Application of multiple imputation in dealing with missing data in agricultural surveys: The case of BMP adoption. *Journal of Agricultural and Resource Economics*, 78–102.

Appendix

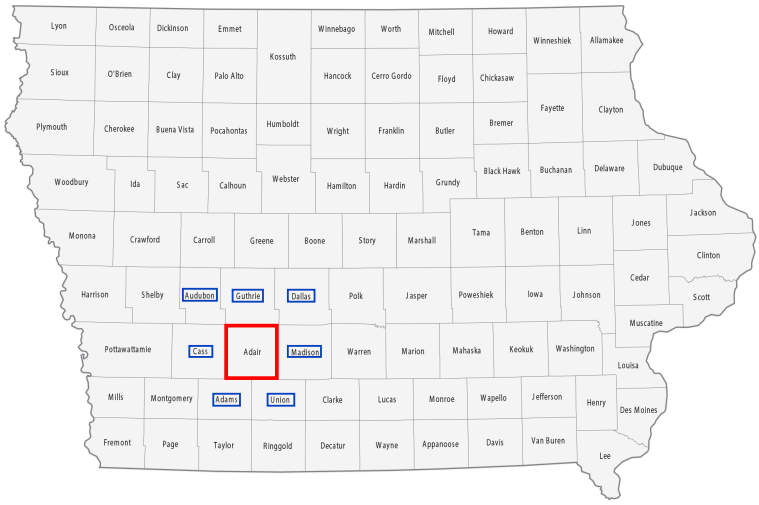


Figure A1: Map of Adair County and neighboring counties in Iowa