Appendix 3: Data analysis

All statistical analyses were done in the R programming environment version 4.0.3 (R Foundation for Statistical Computing 2020) with a Windows 10 Pro version 1909, 64-bit operating system (Microsoft, Redmond, WA, U. S. A.). Data manipulation and descriptive statistics were conducted using the R package "dplyr" (Wickham et al. 2021) and base R. Plots were generated with the R package "ggplot2" (Wickham 2016).

Analysis of survey data

Correlations between ordered categorical variables from the survey were tested using Spearman's rank correlation test.

Analysis of participation in AWM

Four of the independent variables in the regression model (group size, size of the resource system, size of citrus groves, heterogeneity in grove size) were based on information recorded in the database of citrus operations in California maintained by the Citrus Research Board (CRB), hereafter referred to as the *citrus layer*. We obtained access to the June 2020 version of the citrus layer (Rick Dunn, personal communication) and the outlines of each AWM unit in the state of California (Rick Dunn and Robert Johnson, pers. com.). The software ArcGIS Pro (ESRI, Redlands, CA, U. S. A.) was used to overlay the citrus layer and the institutional layer in order to calculate the group size, size of the resource system, size of citrus groves and heterogeneity in grove size in each AWM unit using the "Dissolve" tool. Correlations between numeric independent variables in the regression model were tested using Pearson's correlation test.

- Group size: It was calculated as the number of different PURs within each AWM unit on the CRB citrus layer, which was compared with the number of PURs routinely collected by the grower liaisons and found to be highly correlated (ρ=0.72, P=2E-15).
- Size of the resource system: It was calculated by aggregating all of the citrus properties in each PMA/PCD and calculating the sum of the grove acres. The calculated total citrus acreage under each management unit was highly correlated with data provided by the grower liaisons (ρ=0.97, P<2.2E-16) and with the citrus acreage recorded in the California Statewide Crop Mapping database (ρ=0.98, P<2.2E-16) (Department of Water Resources 2020).
- Size of citrus groves: It was calculated with the "Dissolve" tool from the software ArcGIS Pro by aggregating all of the citrus properties in each PMA/PCD and calculating the mean of the grove acres.

- Heterogeneity in grove size: It was calculated with the "Dissolve" tool from the software ArcGIS Pro by aggregating all of the citrus properties in each PMA/PCD and calculating the standard deviation of the grove acres.

Some preliminary statistical analyses were conducted to guide the hypotheses tested with the zoib regression model.

- Institutional approach (PMA/PCD): there was significantly higher participation in AWM in PCDs than PMAs in every season ($P \leq 0.043$ on t-tests), except the Fall of 2016 ($P = 0.99$).
- Group size: there was a significant negative correlation between the number of pesticide use permits and participation in AWM ($\rho = -0.28$, $P < 2.2E\text{-}16$).
- Size of citrus groves: there was a significant positive correlation between the average size of citrus groves and participation in AWM ($\rho = 0.27$, $P \leq 2.2E\text{-}16$).

Zero-and-one-inflated beta regression models were constructed using the R package "zoib" (Liu and Kong 2015). A zoib model assumes that the dependent variable y (the percentage of citrus acreage in each PMA/PCD treated within the recommended window) follows a piecewise distribution such that

$$f(y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1-p_i)q_i & \text{if } y_i = 1 \\ (1-p_i)(1-q_i)\text{Beta}(\alpha_{i1}, \alpha_{i2}) & \text{if } y_i \in (0,1) \end{cases}$$

where p_i represents the probability $\Pr(y_i=0)$, q_i represents the conditional probability $\Pr(y_i=1 | y_i \neq 0)$, and α_{1i} and α_{2i} represent the shape parameters of the beta distribution for $y_i \in (0,1)$. These distributions are combined to derive the unconditional estimate of the response $E(y_i)$:

$$E(y_i) = (1-p_i)(q_i + (1-q_i)\mu_i^{(0,1)}$$

The zoib regression model estimates the logit [*i.e.*, the log(odds)] of the expected value of the beta distribution, the logit of $P(0)$ and $P(1)$ and the log of the dispersion of the beta distribution as linear functions of fixed and/or random effects. The coefficients of the effects on the mean of the beta regression can be interpreted as the expected change in the logit of participation with a one unit change in the corresponding variable. The coefficients of the effects on P(0) and P(1) are interpreted as the change in the logit of either having Participation=0 or Participation=1 with a one unit change in the corresponding variable. The coefficients of the effects on the dispersion of the beta distribution indicate the change in the log of the dispersion with a one-unit change in the corresponding variable (van Woerden et al. 2019). Based on a Bayesian framework, the coefficients are estimated through a Markov Chain Monte Carlo (MCMC) approach (Liu and Kong 2015). Two independent MCMC chains were run per model, each with 5000 iterations, including 200 iterations for burn-in, and thinned by a factor of 2. We assumed a Normal prior distribution N(0, 0.001) for each regression coefficient.

MCMC convergence was visually checked with trace plots and autocorrelation plots. The potential scale reduction factor (psrf) was calculated for each model parameter and the threshold

psrf≤1.1 was used to determine that convergence had been reached (Gelman et al. 2021). In cases where psrf>1.1, we repeated the MCMC process with three chains, 10000 iterations per chain, 1000 for burn-in and thinned by a factor of 50. Posterior inferences for each parameter are reported as the mean and 95% credible interval (CI). Model selection was based on the deviance information criterion (DIC) (Liu and Kong 2015). Starting with the most complex model including the seven independent variables mentioned in the previous section, we examined the results and iteratively removed variables for which the CI of the posterior estimates was bounded by a negative and a positive value, and therefore comprised zero. Among competing models that fulfilled the previous condition, we chose the one with the lowest DIC (Table A4.1, Table A4.2).

Finally, the participation levels predicted by the zoib regression model were calculated using the pred.zoib function in the R package "zoib" (Liu and Kong 2015). Predictions were based on a new dataset where the independent variable under evaluation was allowed to vary within the range observed in the original dataset and the rest of the independent variables were fixed at their mean value, except in the case of interaction terms, where both variables were allowed to vary within the observed range.

All data sets and R code used in this study will be posted in a repository at the following URL after publication of this manuscript: https://github.com/nmcr01?tab=repositories.

Literature cited

Department of Water Resources. 2020, January 7. 2016 Statewide Crop Mapping GIS Shapefiles. California Natural Resources Agency.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. 2021. *Bayesian Data Analysis*. Third Edition. CRC Press/ Chapman and Hall, Boca Raton, FL.

Liu, F., and Y. Kong. 2015. zoib: An R Package for Bayesian Inference forBeta Regression and Zero/One Inflated Beta Regression. *The R Journal* 7(2):34–51.

R Foundation for Statistical Computing. 2020. *R: A language and environment for statistical computing*. Vienna, Austria.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY.

Wickham, H., R. François, L. Henry, and K. Müller. 2021. *dplyr: A Grammar of Data Manipulation*.

van Woerden, I., D. Hruschka, and M. Bruening. 2019. Food insecurity negatively impacts academic performance. *Journal of Public Affairs* 19(3):e1864.