Estimating response propensity and nonresponse bias for the 2022 Agricultural Production Census

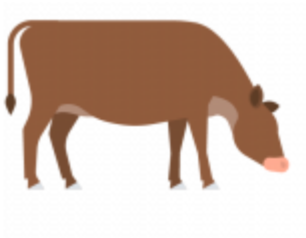
Alba Cervantes Loreto
alba.cervantesloreto@stats.govt.nz

Stats NZ
Tatauranga Aotearoa

The Agricultural Production Census

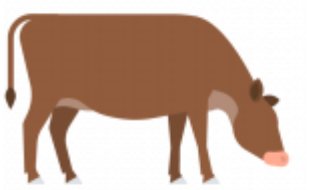
Stats NZ
Tatauranga Aotearoa

- Undertaken every 5 years in partnership with the Ministry of Primary Industries
- Aims to provide a range of summary statistics on the agricultural industry in New Zealand (e.g., total number of cows).
- Target population is all businesses engaged in agricultural production activity during the year ended
- Despite being a survey, we have information on both responding an non responding farms.
- High-level strata consist of a combination of region and farm types
- The final stratification variable is the total land area of each farm measured in hectares.



The Agricultural Production Census

- In 2022 the APC had an atypical low response rate of 69%
- In comparison, in the 2017 census the response rate was 84%
- In addition to the general tendency of low response rates, Groundswell NZ called for all farmers and growers to boycott the APC
- Historically, nonresponse has been handled by donor imputation
- Low response rates can potentially introduce nonresponse bias.



Nonresponse bias

Stats NZ Tatauranga Aotearoa

- Decreasing response rates may not always lead to nonresponse bias. Low response rates are not necessarily “bad” per se.
- Nonresponse bias occurs as a function of how correlated response propensity is to the attributes measured.
- Within the same survey, nonresponse bias can vary across different variables.
- To discern when nonresponse rates lead to nonresponse bias, we must understand how the influences for and against participation are related to the survey measures.

Can nonresponse bias actually be quantified?

Stats NZ
Tatauranga Aotearoa

- Nonresponse bias is notoriously difficult to estimate because we do not know the nonrespondent's values.
- The bias estimation based on the sample respondents will not equal the population bias (i.e., you need the whole population)
- Bias approximations also need Y values for the nonrespondents (which we do not know), or some approximation of them (based on variables that correlate with them).
- Imputation assigns Y values to nonrespondents.
- All the expressions that relate response propensities to nonresponse bias are based on approximations because the estimators are nonlinear.
- While approximations are quite good in many cases, they may be less precise in some situations.

Response propensity

$$\phi_i = \phi(x_i) = \Pr(R_i = 1 | X = x_i)$$

- ϕ_i The response propensity for unit i
- x_i Auxilary data for unit i
- $R_i = 1$ If unit i responds

- Response propensities are unknown, we observe only the binary outcome of response or nonresponse
- We often have auxiliary data available for all sampled units that can be used to understand/adjust for non response
- We assume that $\phi_i > 0$ for all *i*
- Response propensities are often estimated by logistic regression, but probit and non parametric methods can also be used.
- Response propensities are dynamic and likely to vary with the recruitment protocol

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329-353.

Response propensity

- What causes a survey variable to be correlated to the likelihood to respond?

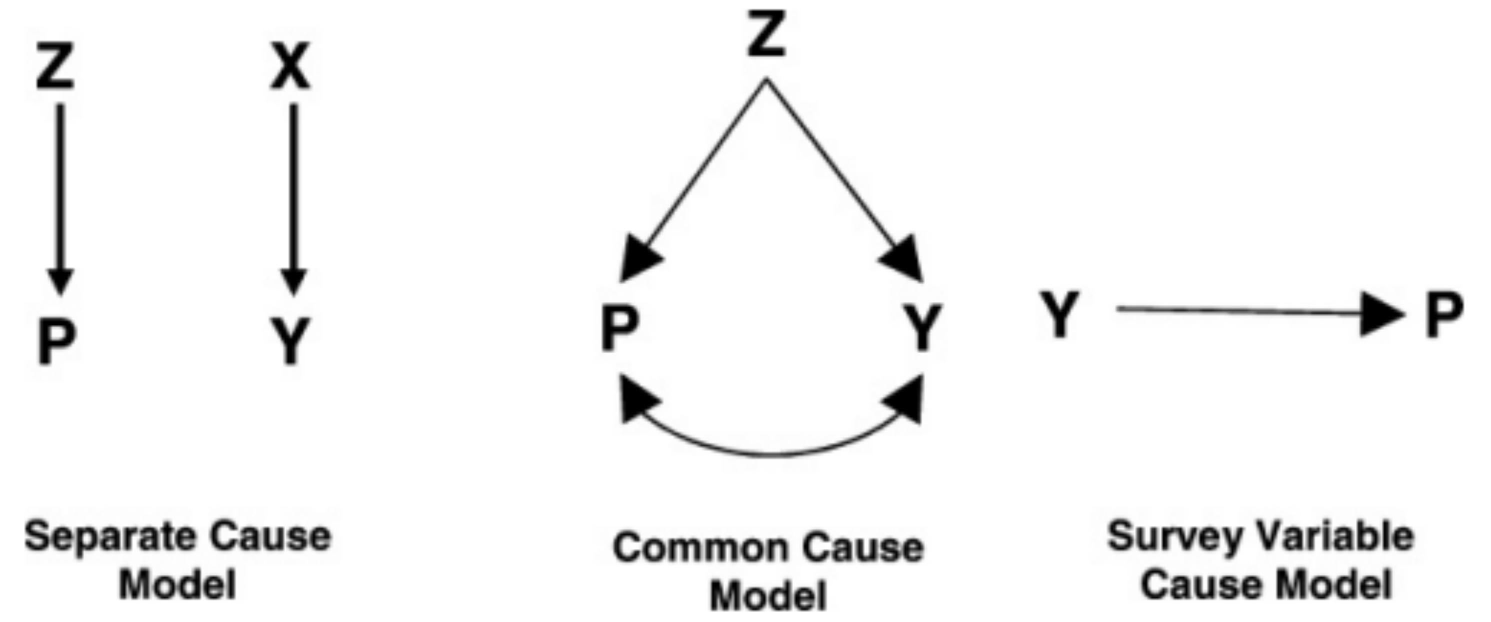


Figure 1. Three Relevant Causal Models Linking Response Propensity with Nonresponse Bias.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, *72*(2), 167-189.

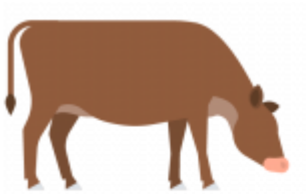
Response propensity weight adjustment

$$\hat{y} = \sum_i d_i \phi_i^{-1} y_i$$

- An method to adjust for nonresponse
- The adjustment factor is the inverse of the estimated propensities of the respondents.
- The idea is to replace the unknown probability of response by an estimate

Research questions

- Which auxiliary variables best correlate with response propensities for the APC?
- Did response propensities change between 2017 to 2022?
- Is there evidence of nonresponse bias in key variables in the Agricultural Production Census?
- Does response propensity weight adjustment give different results to donor imputation?
- Can nonresponse bias decrease with response propensity weight adjustment?



Response propensity model

Stats NZ Tatauranga Aotearoa

$$\text{response}_i \sim \text{Binomial}(1, p_{i,j})$$

$$p_{i,j} = \alpha_{\text{cell}[j]} + \beta_{\text{s,cell}[j]}\text{size}_i + \beta_{\text{r,cell}[j]}\text{past responses}_i$$

$$\begin{bmatrix} \alpha_{\text{cell}} \\ \beta_{\text{s,cell}} \\ \beta_{\text{r,cell}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_s \\ \beta_r \end{bmatrix}, \mathbf{\Sigma} \right)$$

$$\mathbf{\Sigma} = \mathbf{SRS}$$

$$\alpha \sim \text{Normal}(0, 1.5)$$

$$\beta_s \sim \text{Normal}(0, 0.5)$$

$$\beta_r \sim \text{Normal}(0, 0.5)$$

$$\sigma_\alpha \sim \text{Exponential}(1)$$

$$\sigma_{\beta_s} \sim \text{Exponential}(1)$$

$$\sigma_{\beta_r} \sim \text{Exponential}(1)$$

$$\mathbf{R} \sim \text{LKJcorr}(2),$$

- Multilevel Bayesian logistic model
- Imputation cell as a random effect (i.e., each cell had its own intercept and slope)
- Imputation cells are a combination of farm type, region and a range of sizes.
- Predictor variables : size (log transformed) and number of past responses.

Response propensity model

$$\text{response}_i \sim \text{Binomial}(1, p_{i,j})$$

$$p_{i,j} = \alpha_{\text{cell}[j]} + \beta_{\text{s,cell}[j]}\text{size}_i + \beta_{\text{r,cell}[j]}\text{past responses}_i$$

$$\begin{bmatrix} \alpha_{\text{cell}} \\ \beta_{\text{s,cell}} \\ \beta_{\text{r,cell}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta_s \\ \beta_r \end{bmatrix}, \mathbf{\Sigma} \right)$$

$$\mathbf{\Sigma} = \mathbf{SRS}$$

$$\alpha \sim \text{Normal}(0, 1.5)$$

$$\beta_s \sim \text{Normal}(0, 0.5)$$

$$\beta_r \sim \text{Normal}(0, 0.5)$$

$$\sigma_\alpha \sim \text{Exponential}(1)$$

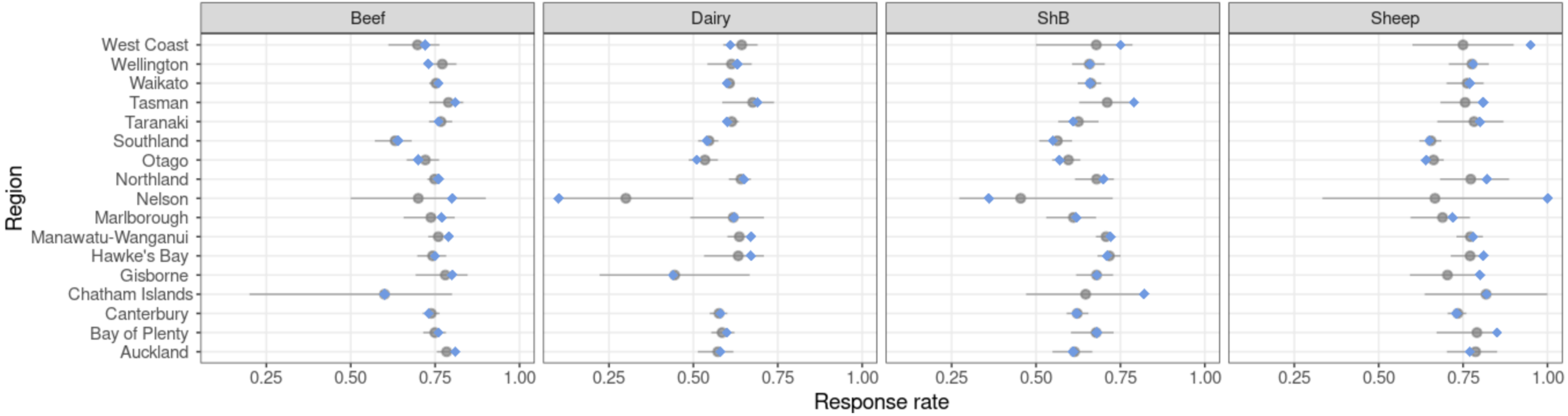
$$\sigma_{\beta_s} \sim \text{Exponential}(1)$$

$$\sigma_{\beta_r} \sim \text{Exponential}(1)$$

$$\mathbf{R} \sim \text{LKJcorr}(2),$$

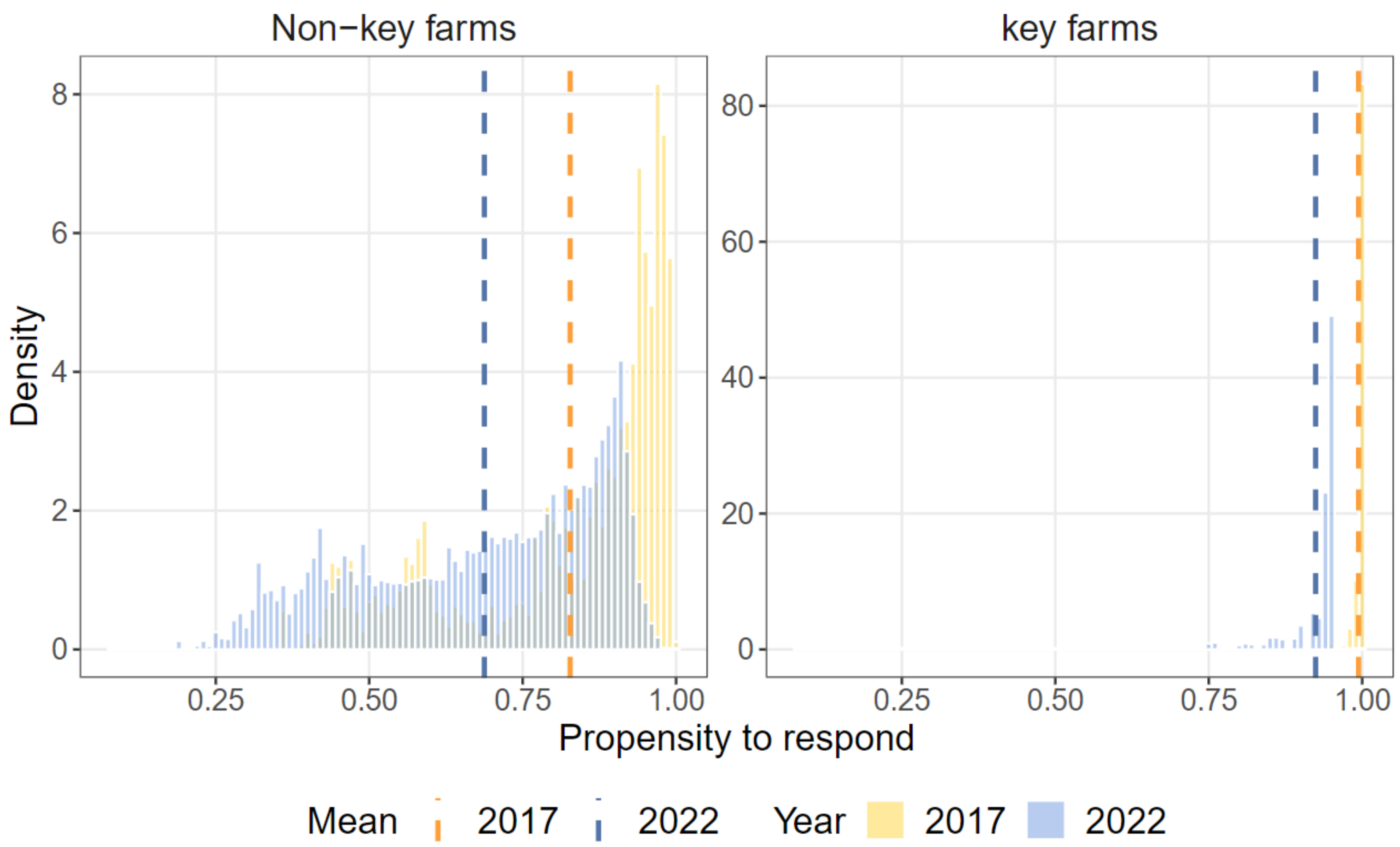
- Adjusted this model to 2017 and 2022 census data.
- Model comparison with LOOIC showed this is the best fit model for both sets of data compared to less complex models (e.g., only intercept)

Response propensity model



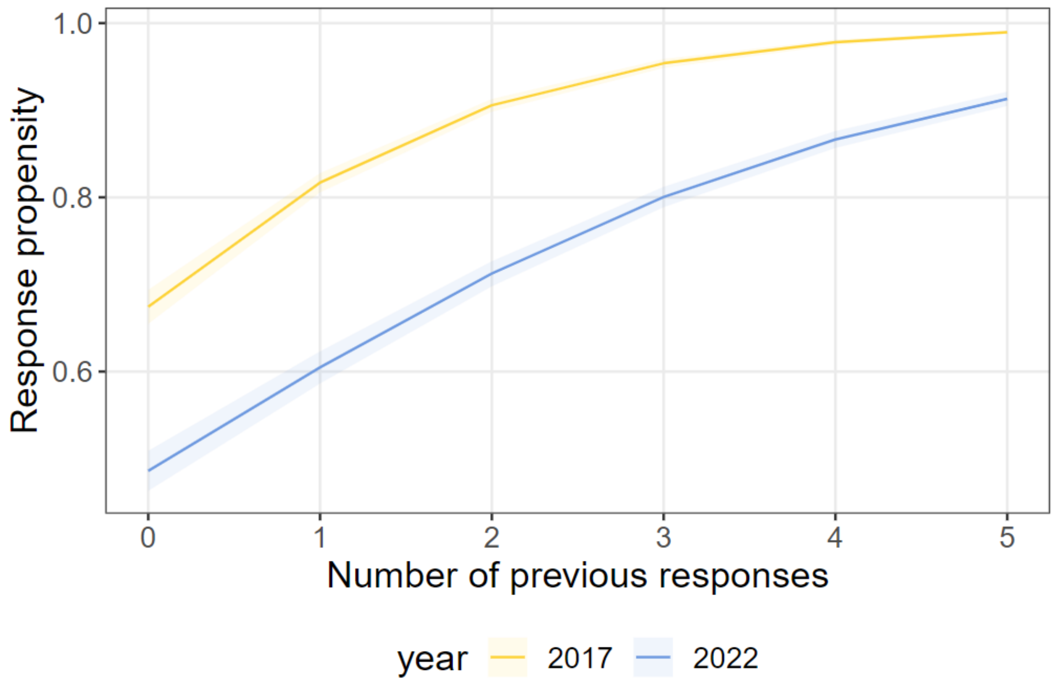
- Posterior predictive checks showed this model is a relatively good fit to the data (figure shows 2022).
- Blue points are observed response rates while grey points and lines are median and 90% of the highest posterior density

Response propensity model

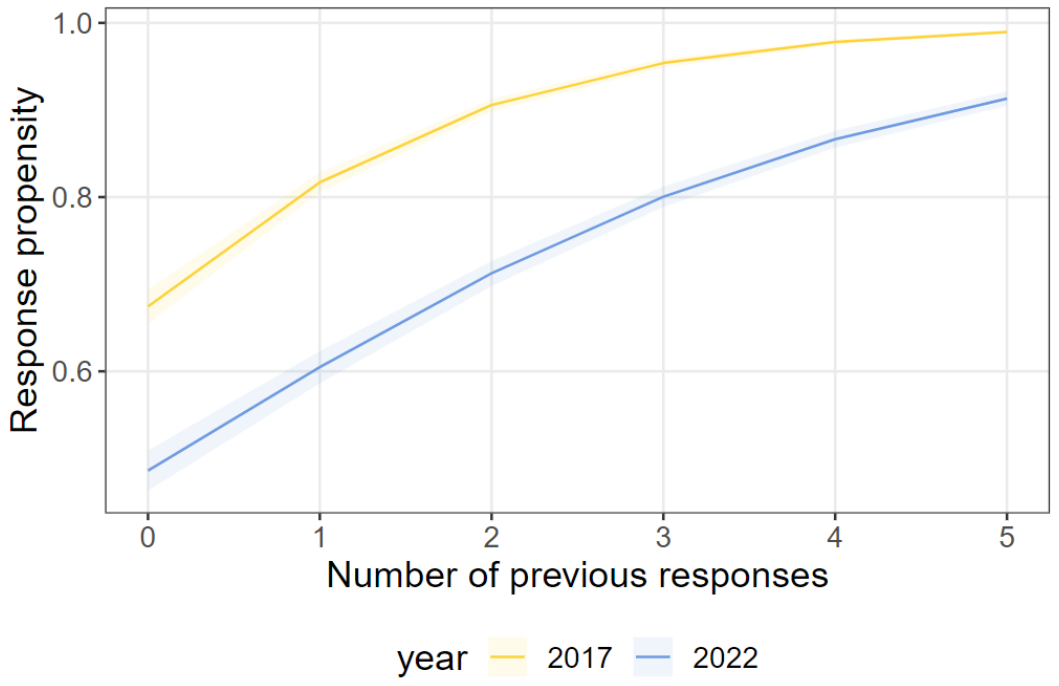


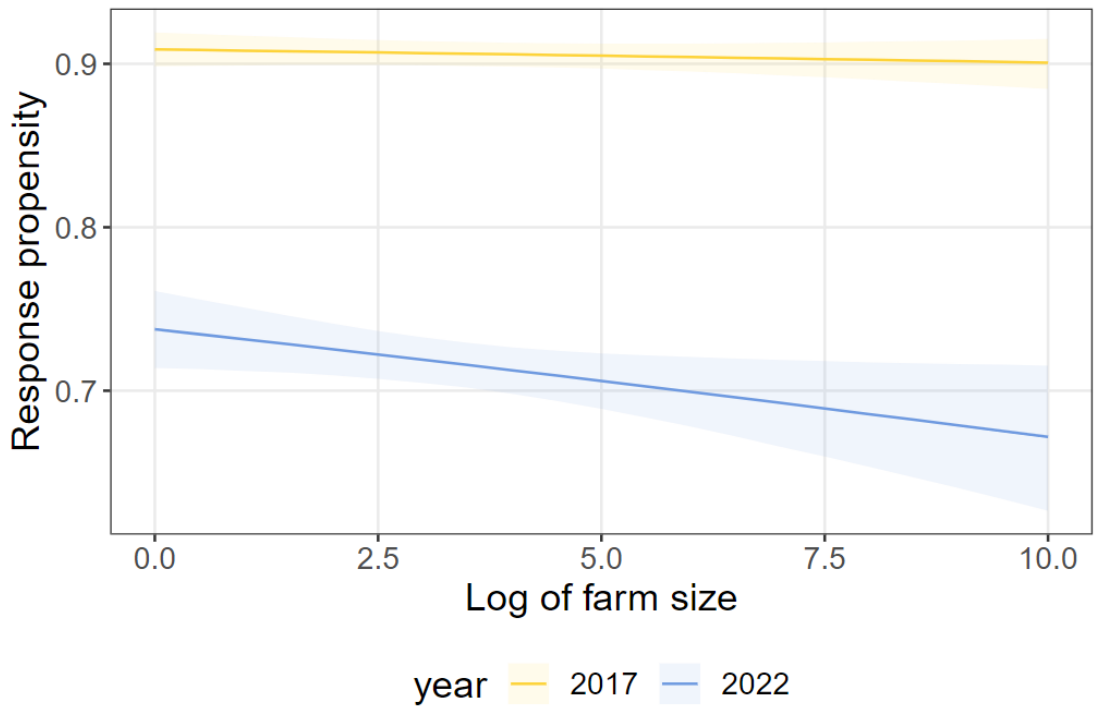
Posterior draws of the expected value of the posterior predictive distribution for every observation in 2017 and in 2022

Response propensity model

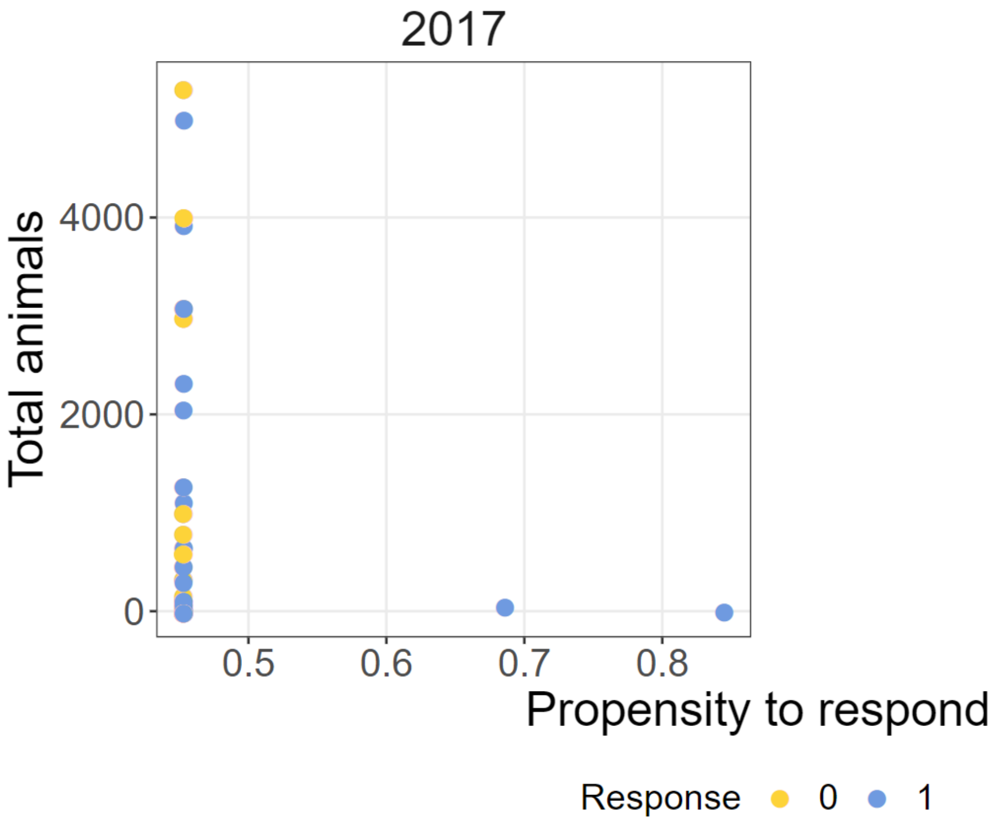


Response propensity model



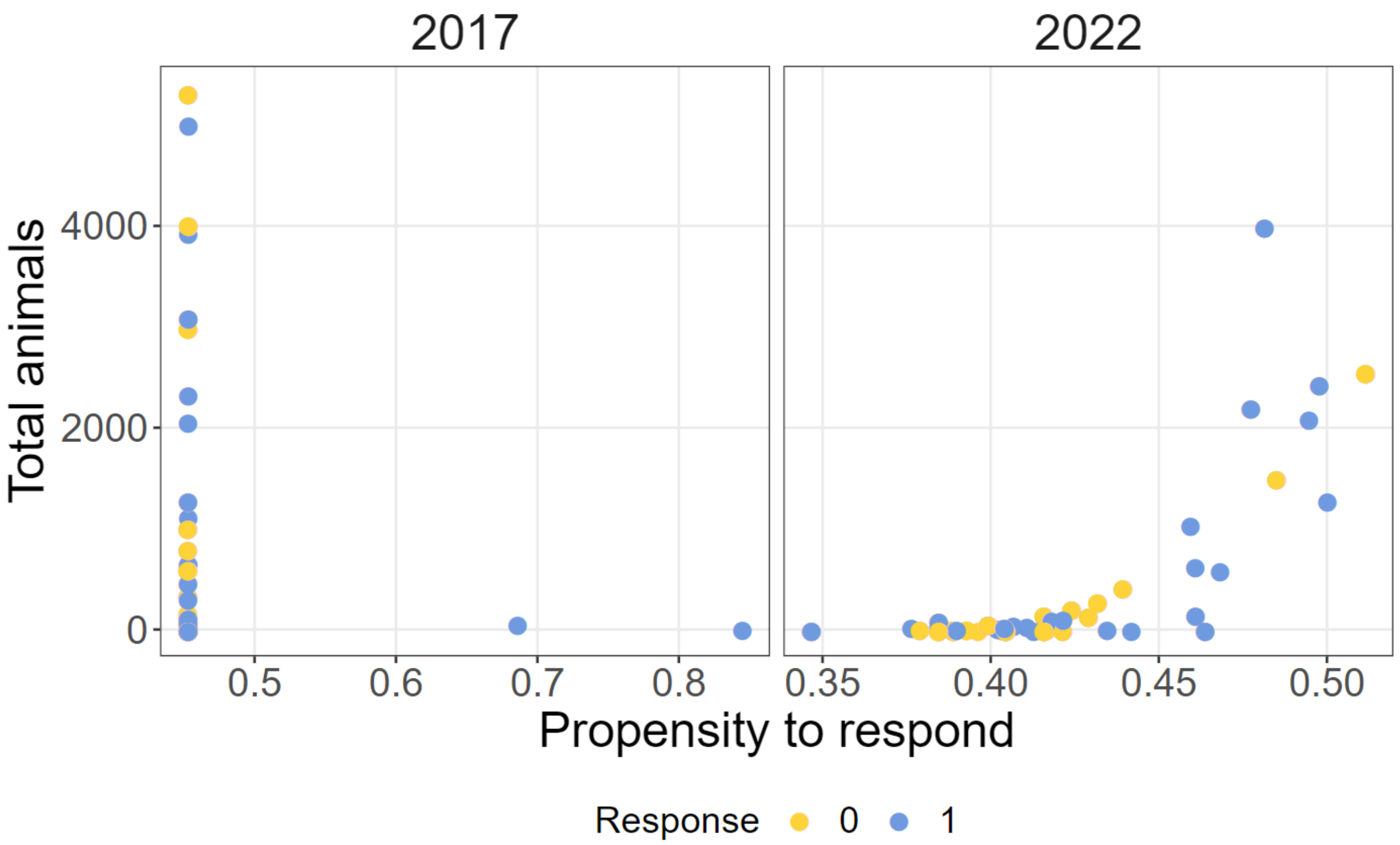


Bias within imputation cells



- Poststratification and imputation will reduce nonresponse bias if response propensities are homogeneous within strata
- And if there is little correlation between response propensities and the response variable

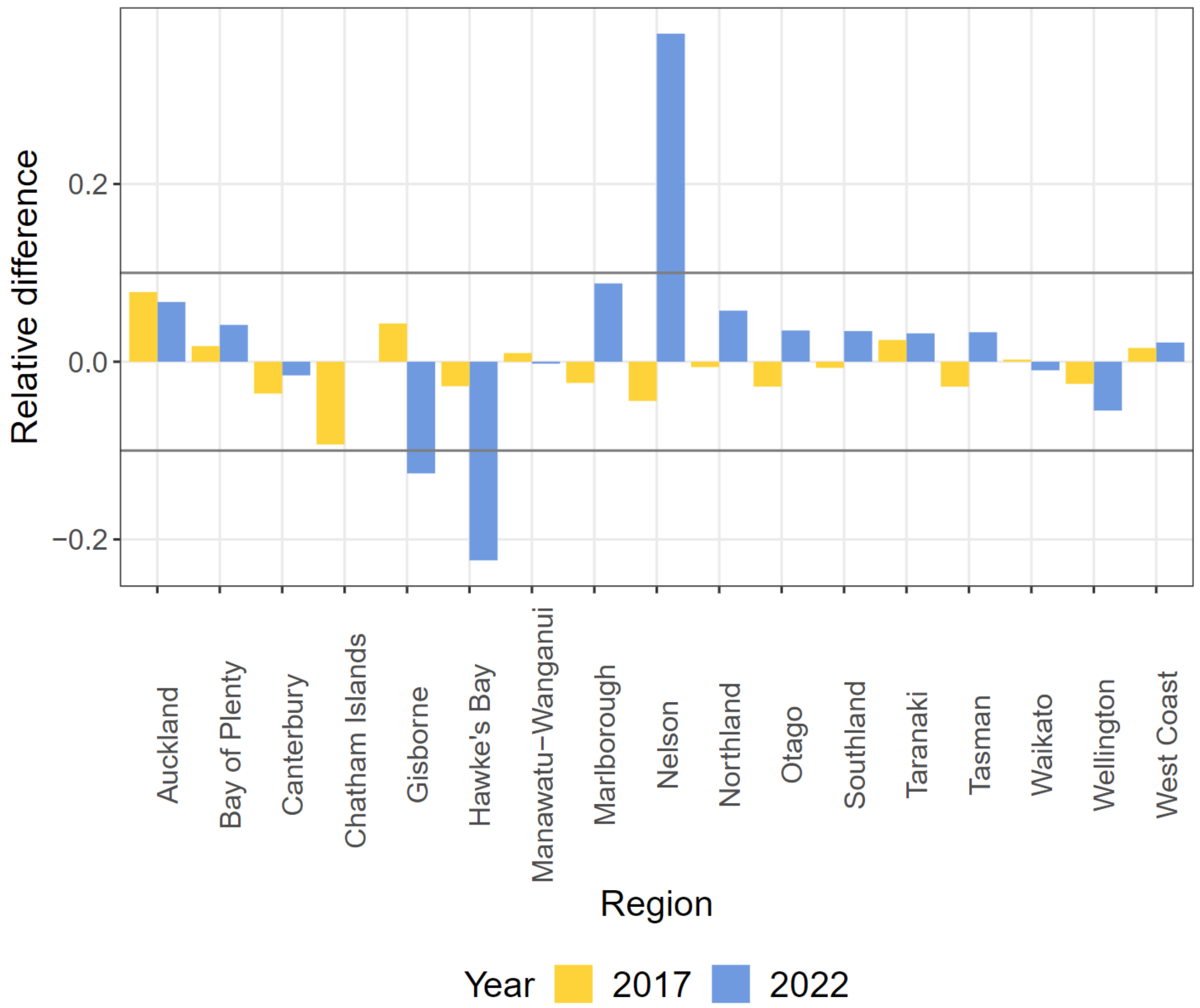
Bias within imputation cells



- Poststratification and imputation will reduce nonresponse bias if response propensities are homogeneous within strata
- And if there is little correlation between response propensities and the response variable

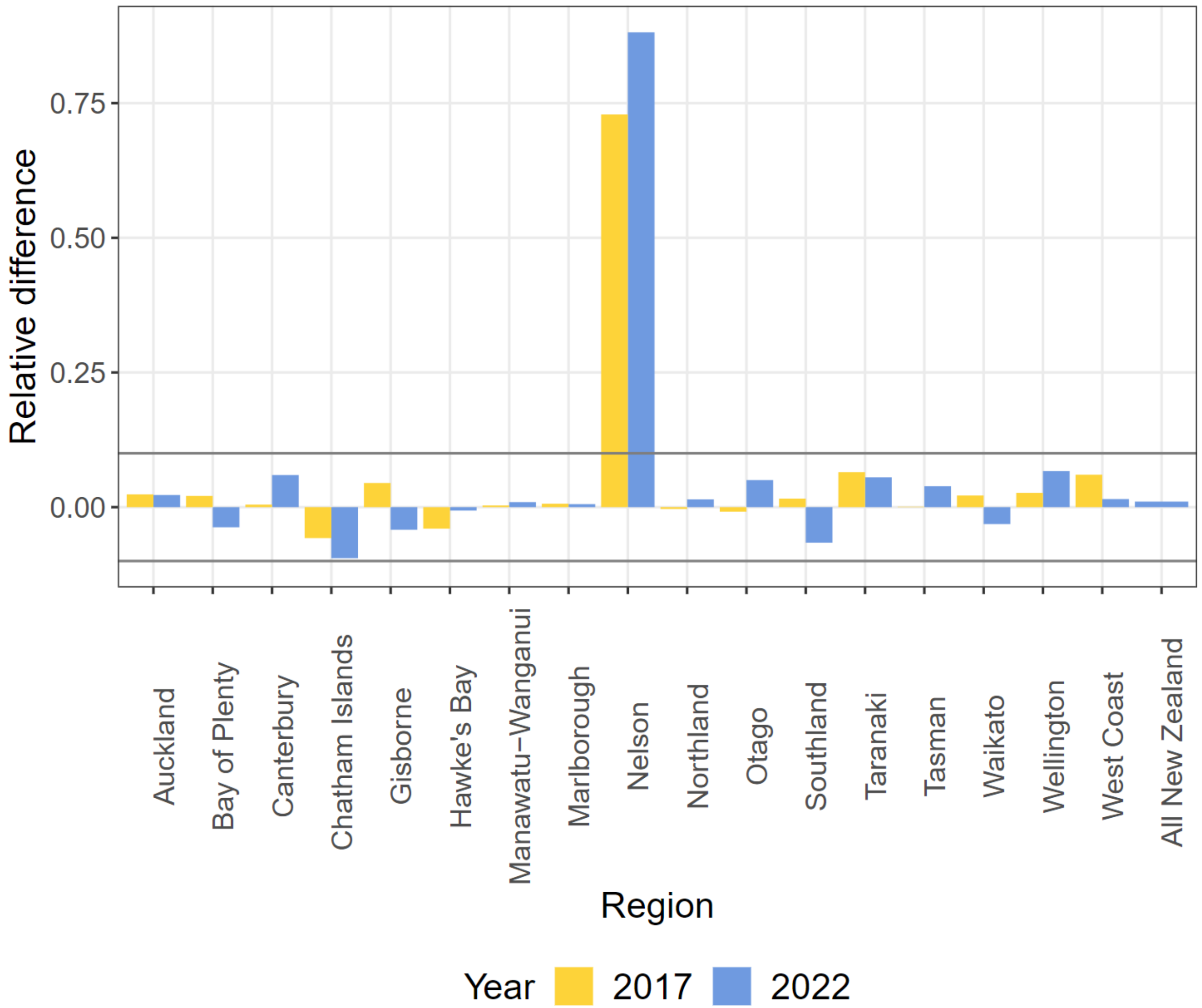
Response propensity weight adjustment

Stats NZ Tatauranga Aotearoa



- Comparison between donor imputation and response propensity weight adjustment
- Dairy cows

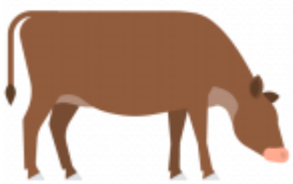
Response propensity weight adjustment



- Beef cows

Main findings

- Main predictors of response propensity were size of the farm and the number of past responses.
- Response propensities decreased from 2017 to 2022 and were more widespread in 2022
- However, the patterns of response propensity remained similar to those observed in 2017
- For some imputation cells in 2022 response propensities were not homogeneous per strata and covaried with the number of animals, which likely introduced nonresponse bias at this level.
- At a regional level the results of the donor imputation and response propensity weighting methods were consistent



Moving forward

- Predicting response propensities is increasingly important to understand nonresponse bias. New covariates and predictive models can and should be explored if we want to accurately quantify nonresponse bias.
- As for the current release:
  - Some cells were divided to make them more homogeneous.
  - Suppression of outputs that derive from a high percentage of biased cells
- For future estimates:
  - New imputation methods can and should be explored under the assumption that response rates might not improve.
  - Multiple imputation
  - Weight calibration methods with response propensity
  - Model assisted estimates.

Thank you

Nonresponse bias

$$bias(\hat{y^{pst}}) \approx N^{-1} \sum \bar{\phi_h}^{-1} \sigma_{\phi_h} \sigma_{Y_h} \rho_{\phi_h, Yh}$$

- $\hat{y}^0$ estimated mean of the postratified estimator
- h denotes stratification classes

- Poststratification will reduce non response bias if the distributions of ϕ or Y are less variable within post strata than across post strata
- Or if their covariance is attenuated within post strata
- A good choice for a post stratification variable would be a variable highly correlated with the response propensities such that response propensities were constant with each level
- The total nonresponse bias is the sum of bias across all strata

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329-353.

Nonresponse bias

$$bias(\hat{\bar{y}}^0) = \bar{\phi}^{-1}\sigma_\phi \sigma_Y \rho_{\phi,Y}$$

- $\hat{\bar{y}}^0$ The unadjusted estimator of the respondents mean
- $\bar{\phi}^{-1}$ The population mean of response propensities
- σ_ϕ The standard deviation of response propensities
- σ_Y The standard deviation of the response variable
- $\rho_{\phi,Y}$ The correlation between response propensity and the response variable

- Stochastic representation of bias.
- It assumes that response is a random variable and the probability of response is like the probability in an additional phase of sampling
- However the probability for every unit in this phase is unknown, thus has to be *estimated.*
- The estimated respondent mean is unbiased if $\rho = 0$.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329-353.

Nonresponse bias of the total

$$bias(\hat{y}^{pst}) \approx \sum_h \bar{\phi}_h^{-1} \sum_i Y_{hi}(\phi_{hi} - \bar{\phi}_h)$$

- Different estimators have different expressions of bias.
- The total bias of the estimate is the sum across all strata
- Imputation allows us to have Y values for respondents and nonrespondents

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329-353.

Can nonresponse bias actually be quantified?

Stats NZ
Tatauranga Aotearoa

- Nonresponse bias is notoriously difficult to estimate because we do not know the nonrespondent's values.
- All equations of nonresponse bias use *Y*, the population values, rather than *y*, sample values
- The bias estimation based on the sample respondents will not equal the population bias (i.e., you need the whole population)
- Bias approximations also need *Y* values for the nonrespondents (which we do not know), or some approximation of them (based on variables that correlate with them).
- Imputation assigns *Y* values to nonrespondents.
- All the expressions that relate response propensities to nonresponse bias are based on approximations because the estimators are nonlinear.
- While approximations are quite good in many cases, they may be less precise in some situations.

Can nonresponse bias actually be quantified?

Stats NZ
Tatauranga Aotearoa

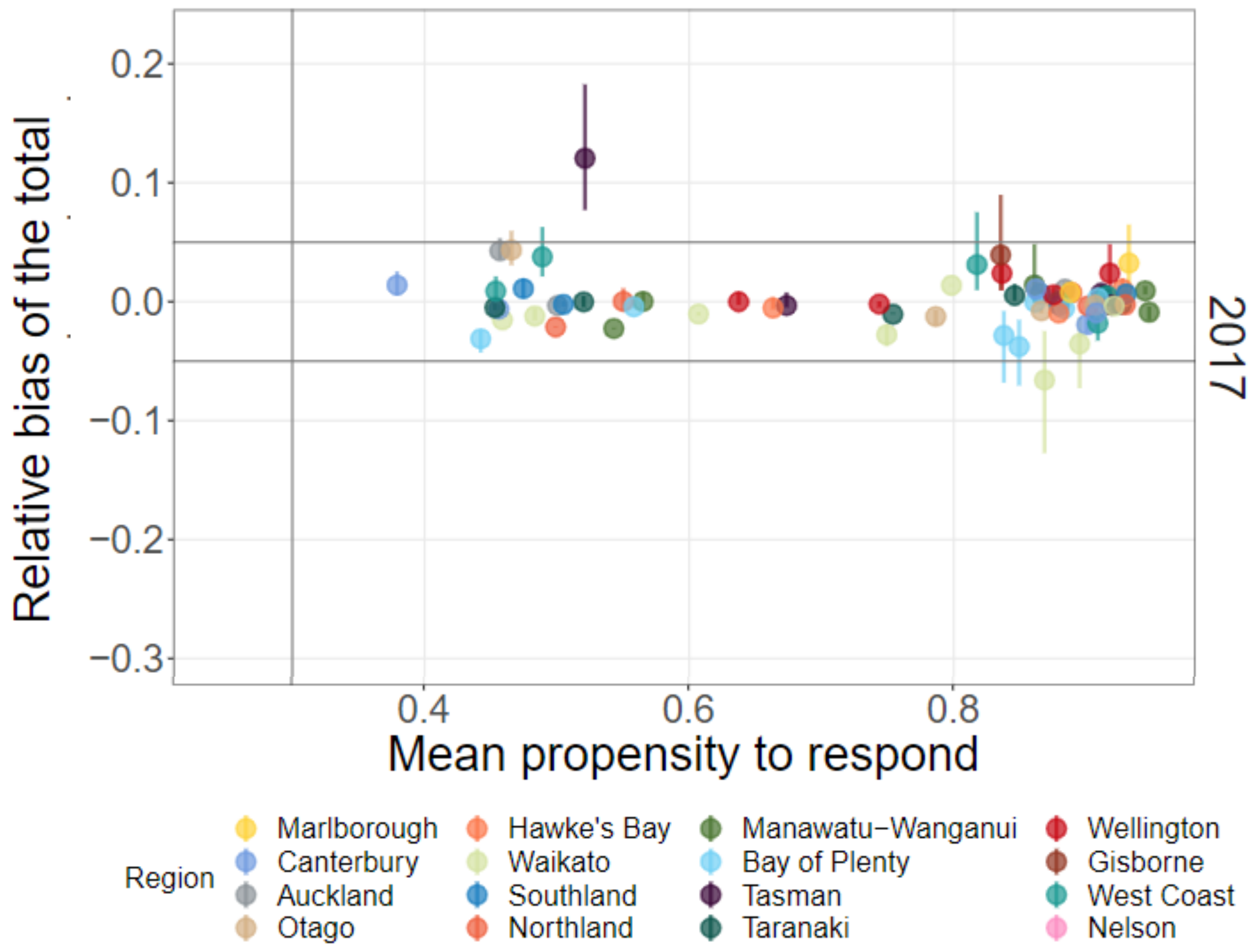
- Types of data that can produce estimates of nonresponse bias
  - Sample frame data (i.e., where records were available both on respondents and nonrespondents)
  - Supplemental data for both respondents and nonrespondents, linked to the sample data.
  - Follow up studies of nonrespondents, comparing the earlier respondent group to those former respondents
  - Reports of intentions to respond to a later survey, comparing those who report agreeing to respond with those who decline to respond
  - Screener interview data

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, *72*(2), 167-189.

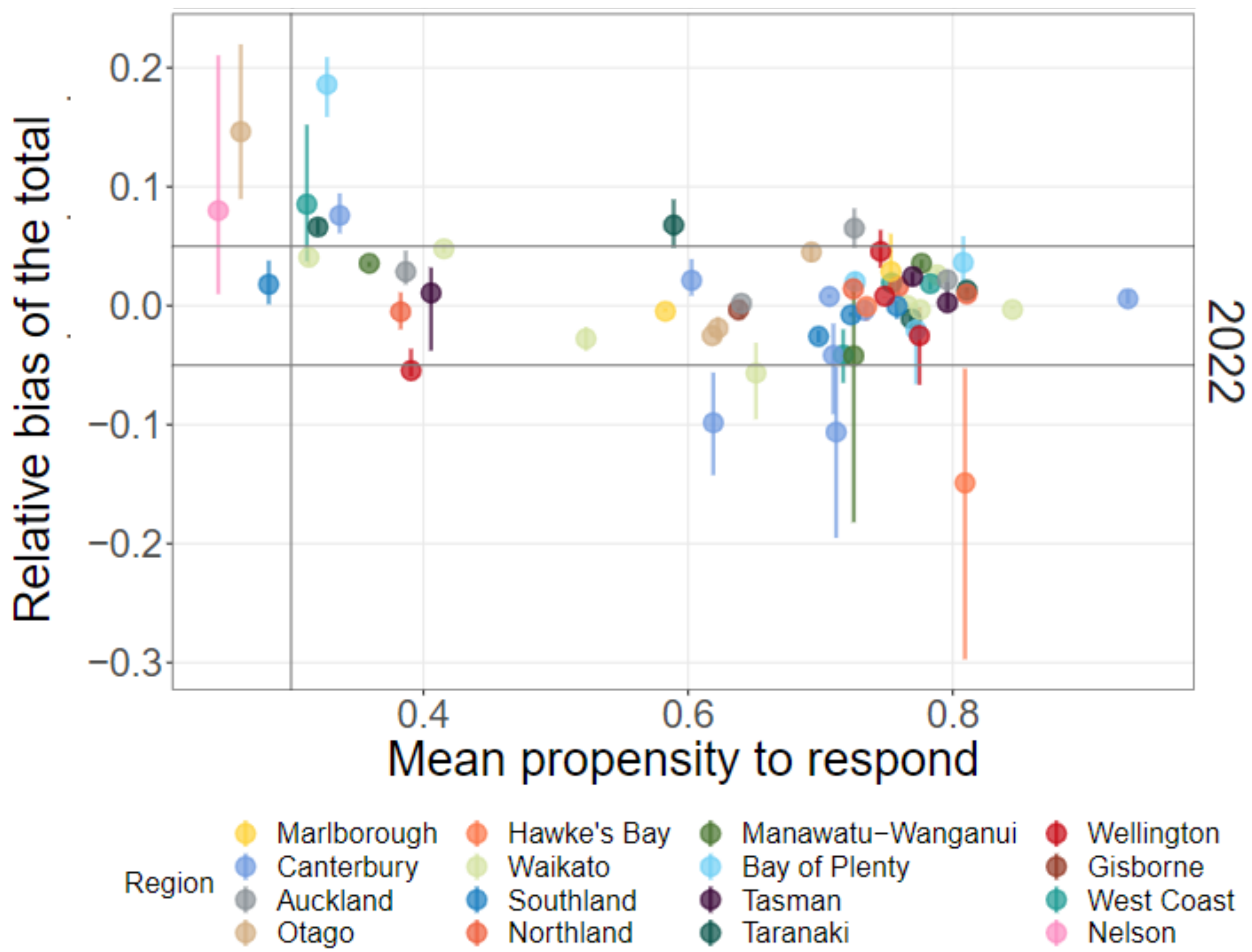
Imputation cells

- The same variables used for forming selection cells are used to form imputation cells (region, farm type and farm size)
- With some minor adjustments for merging small cells
- Farm size is an imputation variable which strongly correlates with key response variables.
- For each nonrespondent the values for all variables to be imputed are copied from the next available donor in the cell
- Each unit can only be used as a donor up to 6 times
- Unlinking may occur
- Key farms are not imputed via donor imputation but by using past information
- If groups are homogeneous imputation will work.

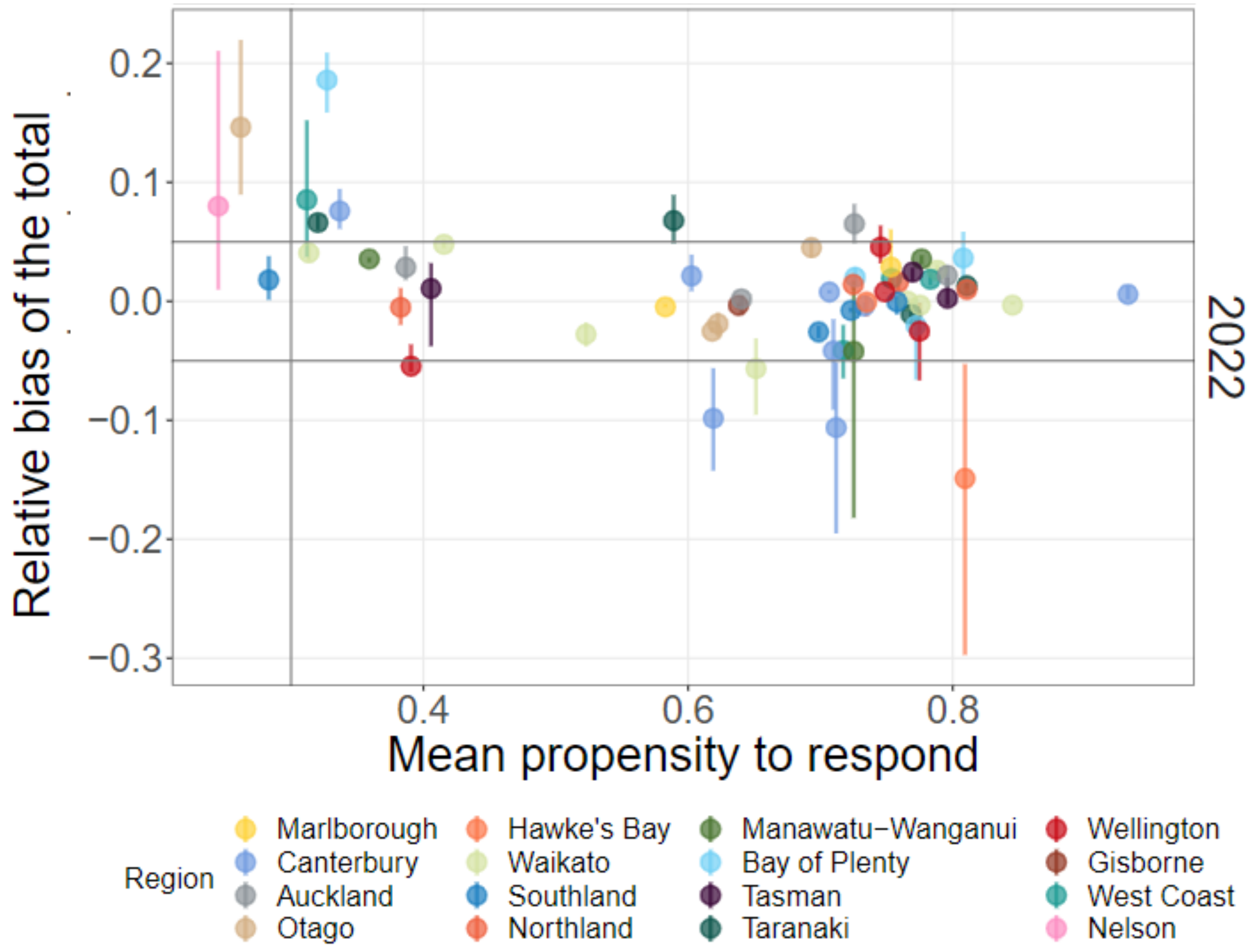
Bias across imputation cells – Dairy cows



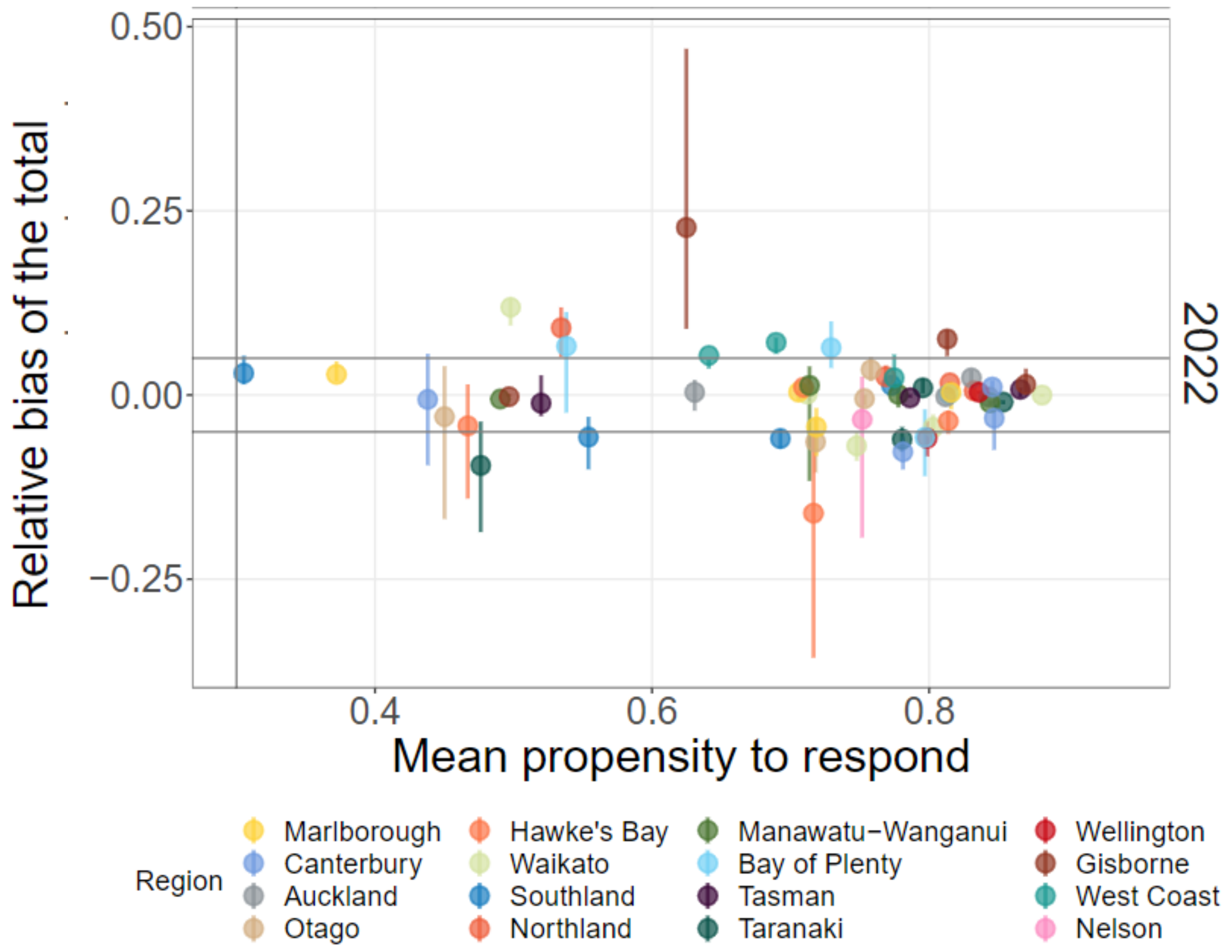
Bias across imputation cells – Dairy cows



Bias across imputation cells – Dairy cows



Bias across imputation cells – Beef cows



Thank you