Supplementary Materials

The cost of joining forces: collective organisations and the economic performance of dairy farms in South West England

Paolo Nota[1], Daniele Curzi[1], Mauro Vigani[2,3]

1. Department of Environmental Science and Policy, University of Milan, Italy
2. European Commission, Directorate-General for Agriculture and Rural Development (DG AGRI), Brussels, Belgium*
3. Countryside and Community Research Institute, University of Gloucestershire, Cheltenham, UK

Table S1. Summary statistics for outcome, treatment, and control variables

	N	Mean	SD	Min	Max
Collective sale (dummy)	169	0.438	0.498	0	1
Average selling price (€/litre)	169	0.248	0.032	0.18	0.41
Production costs (€/litre)	169	0.205	0.050	0.053	0.324
Markup (%)	169	0.355	0.731	-0.231	3.348
Age (4 categories)	169	2.71	0.876	1	4
Education (4 categories)	169	2.793	0.723	1	4
Hectares for milk production (ha)	169	141.189	152.642	4	1011
Total milk production (litre)	169	1851650	2683897	20000	23000000
Herd size (number of cows)	169	234.142	214.657	17	1600
Yield (litre/cow)	169	7960.009	1886.203	999	15000
Full-time workers (number of)	169	3.698	3.512	0	32

Note: Authors' calculation on survey data. 'Age' is divided into four categories (1-4): less or equal to 40 years, between 41 and 50, between 51 and 65, and more than 65 years. 'Education' is divided into four categories (1-4): primary, lower secondary, higher secondary, and university. The variable 'hectares for milk production' refers to the total area cultivated for milk production (including land used for growing fodder crops).

* **Disclaimer**: The authors are solely responsible for the content of the paper. The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Table S2. Covariates imbalances before (A) and after (B) Coarsened Exact Matching

Panel A

Multivariate L1 distance: 0.865

Univariate imbalance:

	L1	mean	min	25%	50%	75%	max
Age	0.1778	0.2033	0	1	0	0	0
Education	0.0514	0.0078	1	0	0	0	0
Full-time workers	0.1932	0.2964	-1	0	0	0	8
Herd size	0.1316	24.749	6	-10	27	5	200
Hectares	0.0176	17.741	9	-7	-10	0	364
Production	0.0815	4.9e+05	55000	-1.2e+05	0	0	8.0e+06
Yield	0.1737	-289.73	-2801	-700	-500	0	3000

Panel B

Multivariate L1 distance: 0.637

Univariate imbalance:

	L1	mean	min	25%	50%	75%	max
Age	0.0648	-0.0648	0	0	0	0	0
Education	0.0833	0.0401	0	0	0	0	0
Full-time workers	0.0432	0.1049	0	0	0	0	8
Herd size	0.0741	4.292	23	-10	15	-5	200
Hectares	0.1512	2.394	9	2	-5	-11	364
Production	0.1142	5.9e+05	-18000	41000	0	-2.0e+05	8.0e+06
Yield	0.0694	91.88	-300	-3	0	0	3000

Note: Panel A shows the multivariate and univariate imbalance of covariates without matching observations. Panel B shows the multivariate and univariate imbalance after implementing the CEM algorithm. The cutpoints for the covariates are: all four categories for ‘age’; category 1 for ‘education’; 4 workers for the variable ‘Full-time workers’; 200 cows for ‘herd size’; 141.2 hectares for ‘hectares for milk production’; 800,000 litres for ‘production’; 6666, 8000, and 9000 litre/cow for ‘yield’.

Table S3. Average covariates imbalance before and after CEM and t-test on the difference

	Non-matched			Matched		
	Mean		Difference	Mean		Difference
	Individual	Collective		Individual	Collective	
Age (4 categories)	2.62	2.82	-0.20 (0.14) [0.13]	2.60	2.63	-0.03 (0.15) [0.87]
Education (4 categories)	2.79	2.80	-0.01 (0.11) [0.94]	2.77	2.81	-0.04 (0.13) [0.75]
Full-time workers	3.57	3.86	-0.30 (0.55) [0.59]	3.55	3.83	-0.29 (0.81) [0.72]
Ha for milk production	223	248	-25 (33) [0.46]	226	237	-11 (48) [0.82]
Herd size (n. of cows)	133	151	-18 (24) [0.46]	123	133	-11 (30) [0.73]
Milk production (litre)	1637568	2126487	-488918 (415658) [0.24]	1748755	2340593	-591838 (636183) [0.35]
Yield (litre/cow)	8087	7797	290 (292) [0.32]	8273	8267	5 (357) [0.99]

Note: 'Individual' and 'Collective' represent farms that sell milk exclusively through individual businesses with bilateral contracts (e.g., retailers) or through collective organisations (e.g., cooperatives). Values in parentheses indicate standard errors from a two-sample t-test, while values in square brackets represent p-values for the alternative hypothesis that the difference is not equal to zero.

Table S4. Alternative results of selling through collective organisations on farmer's outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Unmatched OLS	Unmatched OLS	Unmatched OLS	Unmatched OLS	CEM	PSM
Dependent variable:						
Average selling price (€/litre)	0.002 (0.005)	0.00001 (0.005)	0.004 (0.005)	0.001 (0.005)	0.003 (0.006)	0.003 (0.005)
Production costs (€/litre)	0.010 (0.007)	0.009 (0.008)	0.012 (0.007)	0.011 (0.008)	0.015 (0.009)	0.012* (0.007)
Markup (%)	-0.236* (0.132)	-0.251* (0.140)	-0.183* (0.107)	-0.204* (0.107)	-0.229* (0.126)	-0.186* (0.108)
Observations	196	196	169	169	107	169
Exclusive sale only			x	x	x	x
Covariates		x		x	x	

Note: the treatment is the variable 'collective' that is a binary variable equal to 1 if the sale is through a collective organisation, 0 otherwise. 'Exclusive sale only' indicates that farmers sell only through a single agreement. Standard errors (in parentheses) are adjusted for heteroskedasticity in the case of OLS and CEM, while produced by a bootstrap procedure (100 times) in the case of Propensity Score Matching (PSM) to account for the fact that the propensity score is estimated. Significance levels: * p<0.1, ** p<0.05, *** p<0.01

Table S5. LASSO and Elastic Net standardised coefficients of selected variables via Cross-Validation (10-folds).

		(1)	(2)	(3)	(4)
		LASSO	Elastic Net	LASSO	Elastic Net
		Full sample	Full sample	Matched	Matched
Payment entirely after delivery	(QC26)	0.788	0.759	0.471	0.453
Price varies with the share of profits	(QC21)	0.545	0.531	0.379	0.368
Logistic costs	(QC31)	0.243	0.239	0.239	0.235
Marketing costs	(QC32)	0.098	0.101		
Price varies with production costs	(QC17)	-0.165	-0.161		
Quality standards	(QC36)	0.068	0.064		
Animal welfare standards	(QC39)	-0.030	-0.028		

Note: the dependent (dummy) variable is 'collective'. The definition of each explanatory (dummy) variable is provided in the text below.

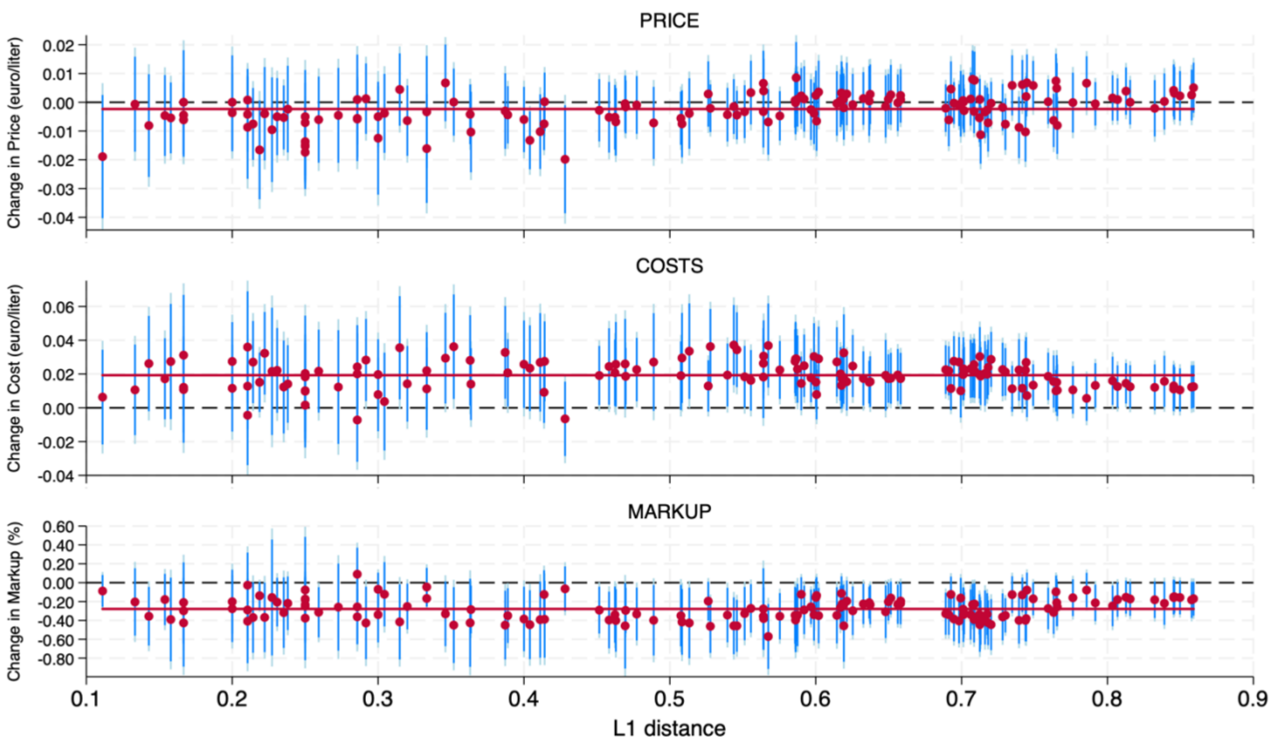

Figure S1. Estimated coefficients after different CEM combinations. The horizontal axis represents the corresponding L1 statistics. Red dots represent point estimates, while blue bars are the 90% and 95% confidence intervals. The red lines are the average of the estimated coefficients (-0.002 for price, 0.019 for costs, and -0.278 for markup).

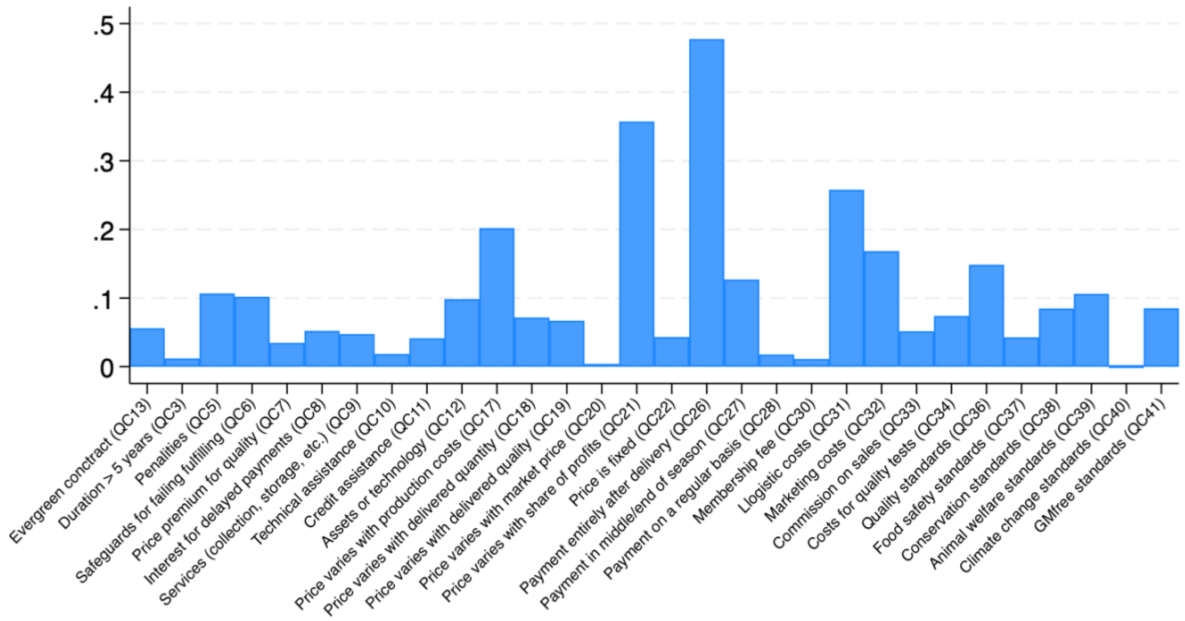

Figure S2. Standardised absolute values of the coefficients from Ridge estimation

Definition of the variables selected by the LASSO and Elastic Net (Table 2 and Table S5)

Variable: “Payment entirely after delivery”

“According to the agreement of this sale/membership rules of a collective organisation, when do you get paid? Please select one option:”

QC.26	Entirely after the delivery of products	Yes	No

Variables: “Price varies with the share of profits”; “Price varies with production costs”

“On what basis is the price of milk determined by the agreement?”

QC.21	Variable price based on share of organisation’s profit	Yes	No
QC.17	Variable price based on production costs	Yes	No

Variables: “Logistic costs”; “Marketing costs”

“In this sale agreement, which of the following costs do you incur?”

QC.31	Collection, storage, transport, handling, etc	Yes	No

QC.32	Promotional and marketing costs	Yes	No

Variables: “Quality standards”; “Animal welfare standards”

“Does the buyer/collective organisation require specific production/quality standards that you have to comply with?”

QC.36	Standards on the quality of the final product(s) (taste, colour, shape, nutritional content, chemical composition, etc…)	Yes	No
QC.39	Standards on animal welfare	Yes	No

Definition of the variables presented in Table 3

“On a scale from 1 (strongly disagree) to 5 (strongly agree) how much do you agree with the following statements regarding your satisfaction with respect to this sale agreement?”

QC.44	I do not have any alternative options to sell my products
QC.45	This sale agreement provides higher prices than alternative buyers
QC.46	This sale agreement provides more stable prices from year to year than alternative buyers
QC.47	This sale agreement provides more possibilities for negotiating prices
QC.48	There are delays in the payments
QC.49	The costs associated with this sale agreement are too high (e.g. storage, marketing and promotion, commission on sales)
QX.15	I am willing to take more risks than my colleagues with respect to dairy production (e.g. adopt new milking regimes, new fodder, new breed of cow)
QX.16	I am willing to take more risks than my colleagues with respect to marketing and pricing risks (e.g. experiment with new sales channels, and new markets)
QX.17	I am willing to take more risks than my colleagues with respect to finance (e.g. loans for production facilities, new investments)

Description of the CEM implementation and cutpoints selection

To investigate the impact of coarsening cutpoints in the implementation of Coarsened Exact Matching (CEM), we run the algorithm multiple times using different sets of cutpoints. Specifically, we explore various combinations of cutpoints as follows.

First, consider the histograms in Figure A3. We observe that most variable distributions are right-skewed, and using too many thresholds would result in many empty or sparsely populated strata. To avoid this, we use the mean, median, mode, and quartiles of each variable's distribution as candidate cutpoints. For categorical variables such as ‘age’ and ‘education’, which take discrete values from 1 to 4, we use the actual values instead of quartiles.

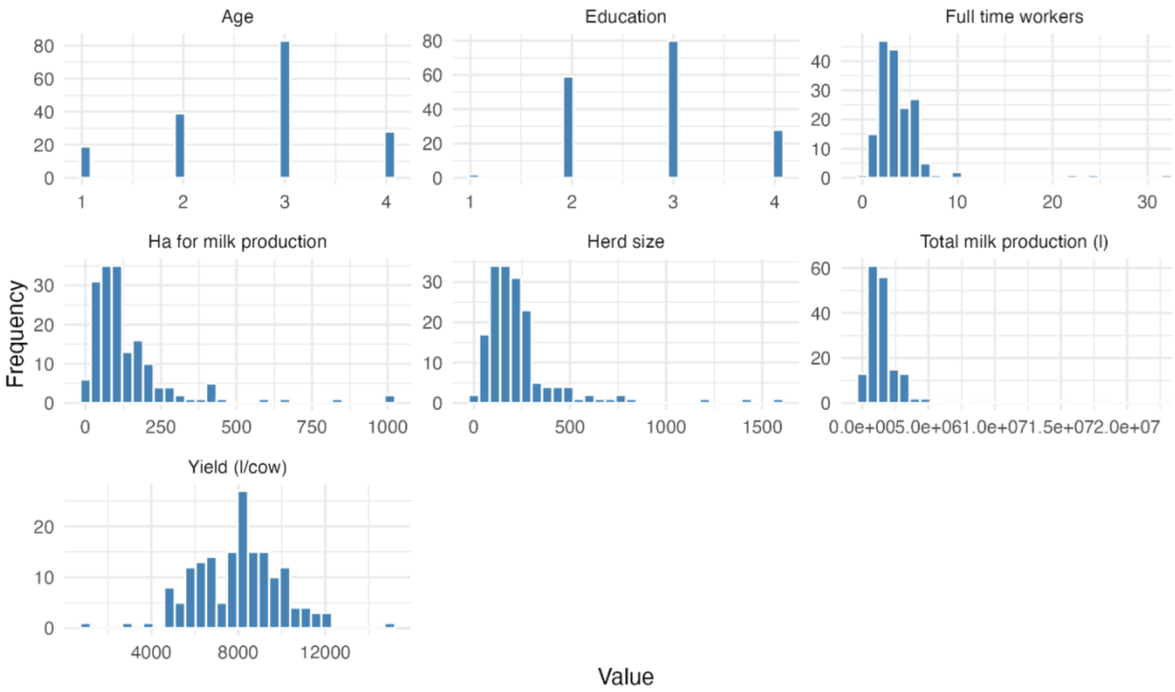

FIGURE S3. Histograms of the variables used in the CEM.

In total, we evaluate all 75000 unique combinations of these cutpoints and report the corresponding L1 statistics in Figure A4, panel A. For each group of matched observations (e.g., those resulting in 50 matched units), we retain only the combination that yields the lowest L1 statistic. We discard combinations with L1 values greater than 0.86, which corresponds to the L1 value in the unmatched case. This filtering process reduces the number of relevant combinations from 75000 to 151, shown in panel B.

As illustrated in both panels of Figure A4, there is a clear positive relationship between the L1 statistic and the number of matched observations: lower L1 values tend to correspond to smaller matched samples, and vice versa. This tradeoff has implications for the statistical power and internal validity of our estimates. A smaller number of matched observations typically implies lower statistical power—that is, a reduced probability of rejecting the null hypothesis ($\beta = 0$) when the alternative hypothesis ($\beta \neq 0$) is true. Moreover, matching on a small subset of the original sample may compromise the internal validity of the estimates, as they are based on a less representative subset.

These considerations suggest that relying solely on the CEM combination that minimises the L1 statistic is not optimal in this context. Instead, we select our preferred combination by jointly considering the L1 statistic, the number of matched observations, and the statistical power of the resulting estimate.

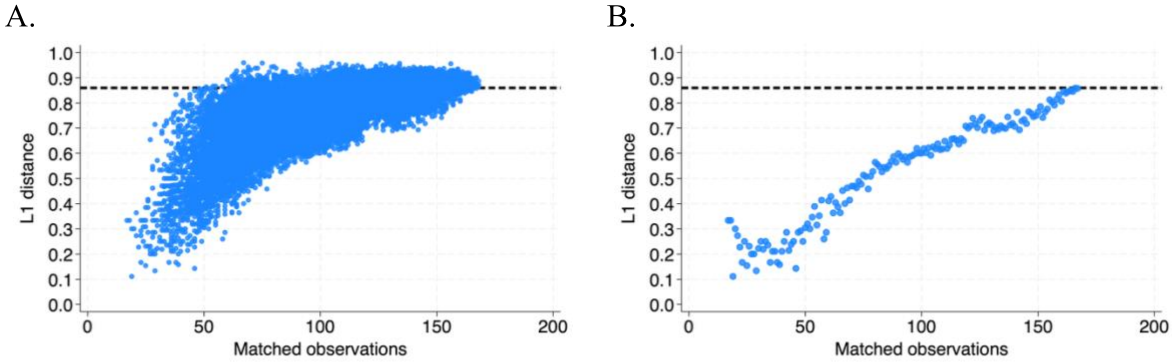

FIGURE S4. Scatter plots showing the relationship between the L1 distance statistic and the number of matched observations after applying CEM. Panel A displays all 75000 combinations of cutpoints, while Panel B shows the subset of combinations that minimize the L1 statistic for each fixed number of matched observations. The dashed line represents the reference L1 value (0.86) obtained without applying any matching.

Figure A5 illustrates this selection process. Panel A displays the L1 statistic, the estimated coefficient (with markup as the dependent variable), and the number of matched observations (represented by bubble size). All but one of the 151 combinations yield negative coefficients; however, those with the lowest L1 values typically correspond to small sample sizes. We select the combination highlighted with a red circle, which substantially reduces the L1 statistic (from 0.865 to 0.637) without excessively reducing the number of matched observations. Notably, the estimated coefficient from this combination is close to that obtained from the unmatched OLS (red cross), suggesting consistency with the full-sample estimate.

In panel B, we demonstrate that this selected estimate has sufficient statistical power (above 80%, indicated by the dashed line). It also outperforms alternatives with fewer observations or coefficients that diverge significantly from the full-sample estimate. Importantly, choosing similar combinations yields comparable results (see Figure A1), indicating that our findings are robust to reasonable variations in the choice of cutpoints.

A. B.

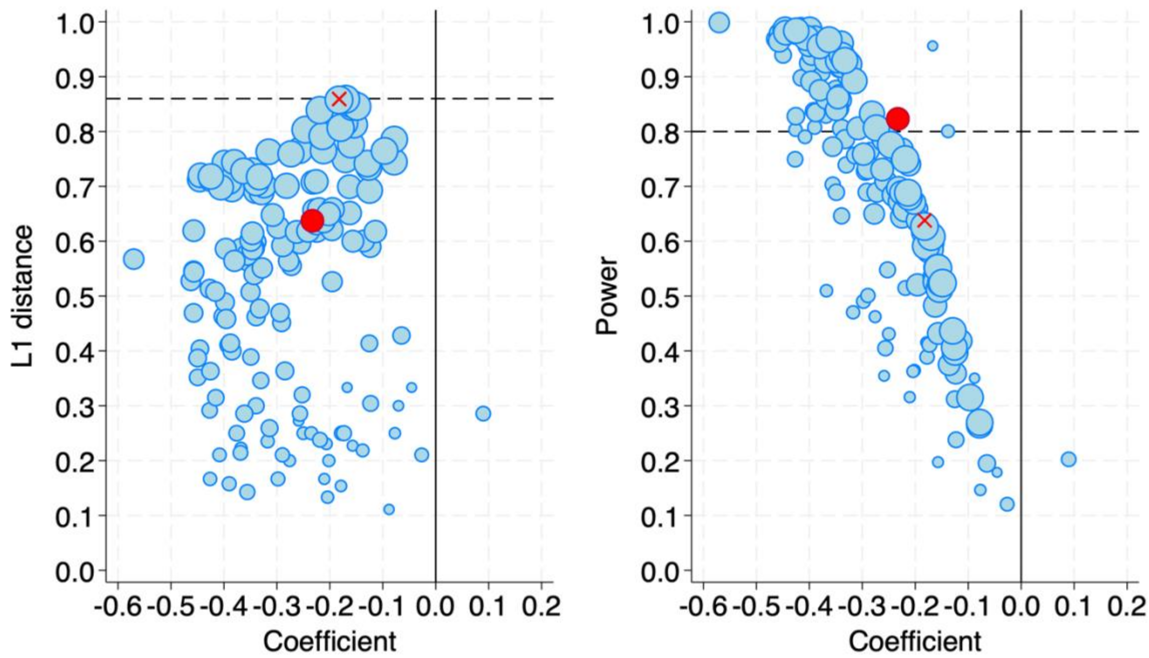

FIGURE S5. Scatter plots of the estimated markup coefficient against the L1 distance statistic (Panel A) and the statistical power (Panel B). Bubble size is proportional to the number of matched observations. The red circle marks the preferred CEM combination, while the red cross represents the unmatched OLS estimate. In Panel A, the dashed line indicates the reference L1 value (0.865) obtained without applying cutpoints. In Panel B, the dashed line marks the commonly used threshold for sufficient statistical power (80%). The L1 statistics are obtained directly from the *cem* command in Stata, while power statistics are calculated using the *power oneslope* command after each regression.