Online Appendix

Appendix A Additional Tables

Table A1: Summary of Statistics and Correlation Matrix (Matched Panel)

Panel A: Summary of statistics of the matched panel data

Variable	N	Mean	SD	Min	Max
(1) Sales revenue	3,897	6.40	2.00	1.29	11.16
(2) Necessity-driven entrepreneur	3,897	0.22	0.42	0	1
(3) Other SOE entrepreneur	3,897	0.45	0.50	0	1
(4) Crisis	3,897	0.25	0.43	0	1
(5) Management experience	3,897	0.39	0.49	0	1
(6) Technical experience	3,897	0.15	0.36	0	1
(7) Business experience	3,897	0.10	0.30	0	1
(8) Age	3,897	43.68	7.53	23.00	72.00
(9) Female	3,897	0.94	0.24	0	1
(10) Education	3,897	13.13	2.82	6.00	19.00
(11) CPC membershipp	3,897	0.30	0.46	0	1
(12) ln(Registered capital)	3,897	4.42	1.72	0.00	10.33
(13) ln(Total assets)	3,897	5.63	1.85	0.00	12.91
(14) Firm's age	3,897	7.98	4.69	0.00	31.00

Panel B: Correlation matrix of the matched panel data

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Sales revenue	--													
(2) Necessity-driven entrepreneur	0.09	--												
(3) Other SOE entrepreneur	-0.04	-0.48	--											
(4) Crisis	0.18	0.08	-0.12	--										
(5) Management experience	0.12	0.06	0.05	0.17	--									
(6) Technical experience	0.02	0.04	0.08	0.03	-0.34	--								
(7) Business experience	-0.00	0.03	-0.03	0.05	-0.26	-0.14	--							
(8) Age	0.19	0.08	-0.01	0.26	0.18	-0.01	-0.01	--						
(9) Female	0.03	-0.02	0.01	-0.03	0.06	-0.02	-0.01	-0.02	--					
(10) Education	0.31	0.14	0.06	0.22	0.12	0.16	-0.01	-0.00	-0.05	--				
(11) CPC membershipp	0.13	0.10	-0.03	0.17	0.19	-0.02	-0.03	0.20	0.03	0.14	--			
(12) ln(Registered capital)	0.51	0.12	-0.10	0.25	0.10	0.01	0.01	0.16	0.01	0.31	0.13	--		
(13) ln(Total assets)	0.73	0.07	-0.03	0.26	0.14	0.01	0.00	0.25	0.03	0.33	0.14	0.64	--	
(14) Firm's age	0.24	-0.15	0.09	0.41	0.06	0.04	0.02	0.42	0.00	0.13	0.06	0.10	0.31	--

Notes: Numbers in parentheses are variable numbers. Summary statistics and correlations are computed within each panel's stated core sample.

Table A2: Robustness for Construct Validity: Full Controls

Model	logistic		multinomial logistic	
Dependent var.	Push/necessity-driven	Pull/opportunity-driven	Push factor	Mixed
	(1)	(2)	(3)	(4)
necessity-driven entrep.	0.852**	-0.520**	0.917**	0.359
	(0.412)	(0.260)	(0.412)	(0.313)
other SOE entrep.	0.030	0.378*	-0.034	-0.500**
	(0.340)	(0.206)	(0.344)	(0.235)
age	0.004	-0.017**	0.007	0.020**
	(0.017)	(0.008)	(0.017)	(0.010)
female	0.739***	-0.301*	0.754***	0.087
	(0.279)	(0.181)	(0.278)	(0.244)
education	-0.009	0.032	-0.015	-0.039
	(0.035)	(0.021)	(0.034)	(0.026)
CPC membership	0.416	-0.020	0.401	-0.103
	(0.261)	(0.120)	(0.255)	(0.164)
registered capital	-0.132	0.036	-0.134	-0.015
	(0.092)	(0.038)	(0.091)	(0.049)
total asset	-0.283***	0.060	-0.280***	0.026
	(0.058)	(0.039)	(0.058)	(0.046)
coastal provinces	-0.676***	0.219	-0.683***	-0.039
	(0.252)	(0.202)	(0.264)	(0.220)
types of industries	0.009	-0.061	0.022	0.087
	(0.118)	(0.099)	(0.121)	(0.123)
Wald test	5.90**	13.80***	8.17***	
Observations	1,543	1,543	1,543	1,543
Pseudo R-squared	0.077	0.018	0.034	0.034

Notes: Standard errors in parentheses, clustered at province level. ***, **, * represent significance levels of 1%, 5%, and 10%, respectively.

Table A3: Robustness: Including Entrepreneurs with Government Background

Templates	Governance		Welfare		Connection	
Dependent var.	CPC branch	Worker's congress	Labor protection	Labor insurance	PR rank	Public relation
	(1)	(2)	(3)	(4)	(5)	(6)
necessity-driven entrep.	0.066***	0.037**	0.099**	0.183***	0.221**	0.178**
	(0.015)	(0.017)	(0.041)	(0.059)	(0.083)	(0.083)
other SOE entrep.	0.005	0.010	-0.021	-0.003	0.161***	0.009
	(0.013)	(0.012)	(0.040)	(0.054)	(0.056)	(0.079)
gov background entrep.	0.015	0.005	0.021	0.046	-0.112*	0.101
	(0.009)	(0.013)	(0.042)	(0.044)	(0.063)	(0.088)
manager exp.	0.049***	0.045***	0.156***	0.148***	-0.178***	0.107
	(0.010)	(0.012)	(0.037)	(0.043)	(0.039)	(0.091)
tech exp.	-0.003	-0.014	0.125*	0.145**	-0.150**	0.224**
	(0.016)	(0.016)	(0.066)	(0.066)	(0.061)	(0.109)
business exp.	-0.005	-0.001	-0.099*	-0.069	-0.213***	-0.099
	(0.018)	(0.019)	(0.058)	(0.063)	(0.057)	(0.190)
age	0.002***	0.002***	0.002	0.009***	0.008**	-0.002
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)
female	-0.014	-0.010	-0.055	-0.102**	0.021	-0.074
	(0.011)	(0.014)	(0.042)	(0.049)	(0.068)	(0.097)
education	0.013***	0.009***	0.035***	0.069***	0.054***	0.105***
	(0.002)	(0.002)	(0.006)	(0.009)	(0.007)	(0.018)
CPC membership	0.180***	0.057***	0.140***	0.290***	-0.231***	0.129**
	(0.017)	(0.012)	(0.032)	(0.035)	(0.041)	(0.063)
registered capital	0.034***	0.031***	0.118***	0.144***	0.028**	0.135***
	(0.005)	(0.006)	(0.015)	(0.016)	(0.012)	(0.032)
total asset	0.053***	0.027***	0.233***	0.244***	0.055***	0.353***
	(0.005)	(0.005)	(0.021)	(0.024)	(0.014)	(0.037)
firm's age	0.008***	0.006***	0.027***	0.027***	0.007	0.027**
	(0.002)	(0.001)	(0.005)	(0.005)	(0.005)	(0.012)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,090	8,090	8,090	8,090	1,713	6,811
R-squared	0.301	0.136	0.323	0.401	0.214	0.609

Notes: Standard errors in parentheses, clustered at province level. ***, **, * represent significance levels of 1%, 5%, and 10%, respectively.

Appendix B Construction of the Matched CPES Firm-Year Panel

The CPES is a repeated cross-sectional survey, and successive waves do not carry a common firm identifier. To track the same enterprise over time, we build a matched firm-year panel by linking observations that share time-invariant founder and firm information across waves. We follow the stable-characteristic matching logic of Marquis and Qiao (2020) and extend it into a deterministic entity-resolution procedure that enforces longitudinal consistency at every stage. The procedure runs over all eleven available CPES waves from 1993 to 2014, which together contain about 39,943 firm-year records. This wide window strengthens identification, because a firm observed in the study period can be cross-checked against its appearances in earlier and later waves. For the performance analysis we then restrict the linked data to the 1997 to 2010 window.

The first stage harmonizes the survey waves. The CPES revised variable names, coding rules, value labels, and measurement conventions over its history, so we standardize a compact set of founding and founder characteristics that are fixed or slow-moving over a firm's life. These cover core identity information, founding conditions, and selected indicators of the founder's career or political history. We retain a few additional attributes whose evolution follows a clear temporal logic and therefore serve as validation signals. Some founder characteristics stay constant across waves, and some institutional statuses can advance over time but should not reverse. This design lets the procedure judge whether a candidate link forms a coherent founder-firm history across the full sequence of waves.

Candidate pairs are generated through strict blocking rules that substantially reduce the search space before any match decision is made. Observations are compared only when they share a set of highly stable identity fingerprints, and only across different survey years. Other founding characteristics and scale-related information are then used to distinguish stronger from weaker candidates. Missing information is treated as uncertainty rather than automatic

rejection, but ambiguous candidates must receive additional support from other consistent attributes. The adopted procedure generates 161,701 candidate pairs, of which 4,669 are classified as high-confidence candidates and 157,032 as more ambiguous candidates requiring further adjudication.

The next stage evaluates each candidate pair through a hierarchical decision framework. The framework first rejects links that violate hard longitudinal or institutional constraints. It then accepts high-confidence links when multiple stable features and rare historical signals point to the same underlying entity. Standard links require broader consistency across founding characteristics, founder attributes, and noncontradictory institutional histories. For early survey waves with sparser information, the procedure allows limited recovery of plausible links, but only when there is no direct contradiction and only under stricter downstream controls. This design reduces false positives while avoiding the mechanical loss of early observations. In total, the procedure accepts 33,342 matched pairs and rejects 128,359 candidate pairs.

Accepted pairwise links are then converted into firm entities using graph-based clustering. Each observation is treated as a node and each accepted match as an edge. Edges are processed according to confidence, and the algorithm prevents a constructed firm entity from containing more than one observation from the same survey year. After clustering, each firm entity is subjected to posterior validation. This validation checks whether the cluster satisfies basic identity invariance, founding-history consistency, institutional-status monotonicity, and other longitudinal restrictions. When a connected component contains both valid and invalid links, the procedure does not either accept the entire contaminated chain or discard the whole component. Instead, it performs posterior splitting, removing inconsistent links while preserving valid subclusters whenever possible. This step addresses the problem of inconsistent linkage chains in repeated cross-sectional data.

The final output is a long-format matched panel that preserves all 39,943 original observations while distinguishing multi-year entities from singletons. It contains 31,289 firm entities, including 5,497 multi-year firms and 25,792 singletons. The multi-year portion contains 14,151 firm-year observations. Among multi-year firms, the mean panel length is 2.57 waves and the maximum panel length is 9 waves. The final validation shows that all key hard constraints pass with no status-year conflicts, no rare-trait conflicts, and no institutional-status monotonicity or timing violations. All 33,342 accepted matched pairs have zero basic information difference.

This panel is intentionally more conservative than a match-maximizing panel. In a repeated cross-sectional survey with low re-sampling probability, very long average panels may reflect over-matching rather than more accurate recovery firm histories. This is especially relevant in the CPES setting, where many private firms founded in the same province and period may share similar founding characteristics. Our comparison experiments confirm this trade-off. When longitudinal constraints are relaxed, the same data can easily generate panels with an average length above three waves and a maximum length of ten waves. However, this comes at the cost of more than 20 percent founding-year violations and numerous cases in which institutional status moves backward over time. These comparison results will be documented in the online replication archive. We therefore choose a shorter but higher-fidelity panel.

This procedure contributes to the use of repeated cross-sectional surveys for longitudinal entrepreneurship research in three ways. First, it extends the stable-characteristic matching logic in Marquis and Qiao (2020) by complementing longitudinal consistency constraints. Second, it introduces graph-based clustering and posterior splitting, which allows the procedure to preserve valid subpanels while removing inconsistent linkage chains. Third, it treats early-wave data sparsity as a separate identification problem, allowing bounded

recovery of plausible early observations without permitting weak evidence to create artificial long panels. This approach is an entity-resolution procedure and better suited to archival management and entrepreneurship research than a black-box matching algorithm.