

Online Supplement to “Feeding the City: How Rural Productivity Shapes Urban Structural Change”

Ida Grigoryeva

This supplement reports data definitions, coverage and geographic construction, additional identification and inference checks, employment and food-price results, and urban-emergence evidence. Service-flow measurement and model reference material follow the empirical results. Table, figure and section numbers beginning with S refer to this supplement; other references identify the paper and its main appendix.

Contents

S1 Data, Outcomes, and Coverage	3
S1.1 Statistical Sources and Survey Years	3
S1.2 Earlier Household-Survey Outcomes	10
S1.3 Coverage by Source and Sample	10
S1.4 Crop Units and Missing Observations	10
S2 Geographic Units and Exposure Measurement	11
S2.1 Census Links Across Administrative Boundaries	11
S2.2 Spatial Inputs and Analysis Units	11
S2.3 Historical Satellite Coverage	15
S2.4 Allocation Across Changed Boundaries	17
S2.5 Village-Level Links Across Survey Years	18
S2.6 Individual-Level Linkage Across Census Rounds	19
S2.7 Settlement Dates and Village Agro-Climatic Similarity	19
S3 Additional Identification and Inference	20
S3.1 Common City Specification	20
S3.2 Whole-Match and Origin-Composition Reassignments	20
S3.3 Assignment-Based Inference	23
S3.4 Birthplace Composition and the Baseline Census	25
S3.5 Exposure Restricted to Post-Baseline Settlements	25
S3.6 Settler-Retention Proxy Definitions	26

S3.7 Settler-Count Controls in City Regressions	27
S3.8 Settler-Count Controls in Rural Public Provision	29
S4 Additional Urban Employment Results	30
S4.1 Service Subsectors	30
S4.2 Manufacturing and Services	31
S4.3 Robustness of the Migration Bound and the Agricultural-Share Decomposition	31
S5 Food Prices, Consumption Composition, and Trade Openness	36
S5.1 Food-Price Indices and Rice Unit Values	36
S5.2 Price Changes on the Common-City Sample	36
S5.3 Rice Quality and Expenditure Composition	37
S5.4 Trade Openness by Location Type	38
S6 Urban Emergence and Endline Sample Selection	39
S6.1 Lattice Sample and Historical Satellite Inputs	40
S6.2 Built-Up Growth and Urban Founding	40
S6.3 Sensitivity to the City Inclusion Threshold	41
S7 Service Flows and Travel-Time Gradients	46
S7.1 Survey Coverage and Located Service Records	46
S7.2 Comparisons Across Services	46
S8 Urban-Center-Level Empirical Facts and Model Coverage	48
S9 Theories of Structural Transformation	49
S9.1 The closed-economy prediction	50
S9.2 The open-economy reversal	50
S9.3 Limits of the classical setups	51
S9.4 Spatial frameworks and their open questions	51
S9.5 Three rival explanations of the same shock	52
S9.6 Requirements for an empirical test	52
S10 Employment Comparisons on the Observed-Price Samples	55

S1 Data, Outcomes, and Coverage

S1.1 Statistical Sources and Survey Years

The empirical analysis combines three BPS statistical collections including reference years 1980, 1996, 2002, 2003, 2010, and 2011, together with the Transmigration Census replication data of [Bazzi et al. \(2016\)](#), which generate the exposure measure, and the geospatial travel-time and raster inputs cataloged in [Table S2](#). The Population Census provides 2010 outcomes (population, employment by sector, migration) and 1980 baseline controls. The Susenas household survey provides 1996 food shares, consumption quantities, calories, and unit values, 2002 consumption quantities by food item, paired with PODES 2003 production for surplus measures, and 2011 prices, wages, food shares, and housing rent, paired with the 2010 census outcome year. The PODES village survey provides 2003 production quantities by crop. Keeping consumption near the production year gives a cleaner rice surplus = production – consumption measure, while keeping prices and wages near the outcome year supports the structural-transformation interpretation. [Table S1](#) provides a per-outcome summary including the wave / year of measurement, the analysis-level sample size (out of 187 urban centers and 2,025 subdistricts, and within the main city sample), and the construction notes. The analysis sample has 187 urban centers, every one of which has a travel time to a program village ([Appendix A.1](#)), and the main city sample is the 134 centers inside the 300-minute travel-time window whose 1980 baseline the census recorded directly and whose 2010 outcomes are measured.

Table S1: Outcome Variables and Baseline Controls: Source Wave, Sample Size, and Construction. *Counts of units holding each variable, as urban centers / subdistricts, in the universe and in the main city sample.*

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
(A) Population Census: BPS Sensus Penduduk				
Log total population and total employment (2010)	Census 2010	186 / 2,020	134 / 405	Headline urbanization outcomes; constructed from individual-level Sensus Penduduk 2010 records aggregated to the urban-center and subdistrict level.
Log sectoral employment in agriculture, manufacturing, and services (2010)	Census 2010	186 / 2,020	134 / 405	Sectoral employment counts from the 1-digit industry classification on the individual record, aggregated to the geographic unit.
Sectoral employment shares (0–1): agriculture, manufacturing, and services (2010)	Census 2010	186 / 2,020	134 / 405	Sectoral employment divided by total employment. Shares are stored as 0–1 fractions.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
Unemployment rate (2010)	Census 2010	186 / 2,020	134 / 405	Unemployed individuals divided by labor-force participants, stored as a 0–1 fraction.
Birthplace decomposition (4 exhaustive categories): share of 2010 residents born in same district / different district same province / different outer-island province / Java-Bali origin	Census 2010	186 / 2,020	134 / 405	Categories are mutually exclusive and exhaustive; they sum to 1. The Java-Bali origin category captures the channel through which the Transmigration Program is hypothesized to redistribute people.
1980 baseline controls: log total population, log total employment, urban-population share, log agricultural employment	Census 1980	183 / 1,897	134 / 384	Conditioning controls in the headline reduced-form specification. The 1980 aggregates apply the 1980 census (SP1980) per-record inflation weights at the individual level.
(B) Susenas 2011: prices, wages, food share, housing rent (aligned with the 2010 outcome year)				
Household food share of total consumption (2011)	Susenas 2011	186 / 993	134 / 344	Household-weighted mean of food expenditures divided by total consumption expenditures (0–1). The share is computed from food and total consumption expenditures over quarters Q1–Q3. Used as a structural moment in the spatial-equilibrium model.
Median monthly wage: all sectors, agricultural workers, non-agricultural workers (2011)	Susenas 2011	186 / 969	134 / 332	Constructed at the subdistrict level from the Susenas 2011 individual labor module by taking the 50th percentile of monthly wage income for currently employed workers. The wage item is net monthly earnings from the main job, in money and goods, and is asked of everyone who worked, so the earner base includes the self-employed and own-account farmers; unpaid family workers, who do not report earnings, drop out of the median.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
Log nominal wage and log real (price-deflated) wage (2011)	Susenas 2011	186 / 993	134 / 344	Log transforms of the subdistrict-level median wage. The real (price-deflated) version is constructed for the all-sector wage only, by dividing it by a household-expenditure-weighted price index (the agricultural-goods price index weighted by the food-expenditure share, plus the housing price index weighted by the non-food share).
Composite agricultural goods price index (broad), crops-only price index, and rice-only price index	Susenas 2011	174 / 993	129 / 344	Built from household item-level unit values (expenditure divided by quantity), winsorized at the 5th/95th percentiles within item×province×urban, taken to the survey-weighted median over the city's households and divided by the item's median over all urban centers' households, then aggregated across the raw agricultural food items into three index variants: Laspeyres (national expenditure weights), Paasche (the city's own weights), and an equal-weight geometric mean; the crops index is the geometric mean over the rice, corn and cassava items; the rice unit value is the survey-weighted value of rice consumed over the kilograms consumed, over the local and higher-quality rice items. As unit values, these indices reflect price × quality (see Appendix D.2).
Housing price index and median monthly rent (2011)	Susenas 2011	186 / 993	134 / 344	Imputed monthly rent for owner-occupied dwellings combined with median actual rent for renters; it is the urban land-scarcity / non-tradables price proxy in the model.
Non-food price indices: fuels and utilities (2011)	Susenas 2011	169 / 855	125 / 300	Cost-of-living controls; constructed parallel to the agricultural goods price index from household-level unit values.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
(C) Susenas 2002: consumption quantities and values (aligned with PODES 2003 production)				
Rice consumption per capita per year: quantity (kg), value (Rp), and calories (kcal)	Susenas 2002	142 / 877	108 / 282	Built from the Susenas 2002 food-consumption module weekly recall and scaled to annual.
Total dietary calories per person per year (kcal)	Susenas 2002	144 / —	109 / —	Sum of the calorie content of every item the Susenas 2002 food-consumption module records, so it measures the whole diet. Built by pooling households within an urban center, a median of 48 households per center, and excluding household-quarters that record fewer than ten food items, which are partly-recorded modules. A genuine diet of fewer than ten items is implausible. Constructed at the urban-center level only, which is why the subdistrict columns are empty.
Subdistrict-total annual rice consumption (kg and tons), winsorized	Susenas 2002	144 / 877	109 / 282	The sum over the subdistrict's surveyed households of annual rice kilograms times the household's survey weight; a copy winsorized at 0.5% is stored beside it. Its ratio to the weighted population of the same households is the rice consumption per person that the city rice imports and surplus apply to the 2003 population of the villages with a rice output figure.
Staples consumption per capita per year (kg)	Susenas 2002	142 / 877	108 / 282	Sum of per-capita consumption (kg) across all staple food items: rice, corn, cassava, sweet potato, taro, and other tubers. The aggregate excludes wheat and sago (not produced in measurable quantities in the PODES sample).
Legumes consumption per capita per year (kg)	Susenas 2002	142 / 877	108 / 282	Sum of per-capita consumption (kg) across peanut, soybean, and mungbean items.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
Coffee & tea consumption value per capita per year (Rp)	Susenas 2002	142 / 877	108 / 282	Coffee and tea consumption are reported in monetary value because the underlying items use distinct physical units (whole beans, ground beans, leaf).
Spices consumption value per capita per year (Rp)	Susenas 2002	142 / 877	108 / 282	Sum of pepper, nutmeg, and clove monetary consumption. Units differ across spices, so values are aggregated.
(D) Susenas 2011 consumption by food category (constructed from the quarterly food-module raw data)				
Staples consumption per capita per year (kg), Susenas 2011	Susenas 2011	181 / 1,954	132 / 397	Constructed analogously to the Susenas 2002 staples aggregate using the same item set. Quarters Q1, Q2, and Q3 are averaged then scaled to annual; Q4 was never released by BPS, so the result represents the Q1–Q3-average annualized consumption, with no fourth quarter imputed.
Legumes consumption per capita per year (kg), Susenas 2011	Susenas 2011	181 / 1,954	132 / 397	Same items as the Susenas 2002 legumes aggregate.
Coffee & tea consumption value per capita per year (Rp), Susenas 2011	Susenas 2011	181 / 1,954	132 / 397	Companion to the Susenas 2002 coffee & tea measure, aligned with the 2010 outcome year.
Spices consumption value per capita per year (Rp), Susenas 2011	Susenas 2011	181 / 1,954	132 / 397	Companion to the Susenas 2002 spices measure.
(E) PODES 2003: agricultural production (aligned with the Susenas 2002 consumption year)				
Rice production per capita (kg / person)	PODES 2003	182 / 1,913	132 / 373	Total rice production divided by the population of the villages with measured rice output in the same 2003 village census. Headline PODES outcome; available across both village-overlap and non-overlap subdistricts.
Log subdistrict-level rice production (tons)	PODES 2003	182 / 1,913	132 / 373	Log of absolute subdistrict-level rice production in tons; primary moment used in the model calibration.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
Rice output per planted hectare (tons / ha) and planted area (ha)	PODES 2003	176 / 1,777	129 / 325	Rice output divided by planted area over the preceding year, summed over the villages with measured rice output. Each planting is counted, so repeated cultivation of a plot can contribute more than its physical area.
Staples planted area (ha)	PODES 2003	185 / 1,955	133 / 388	Aggregate of planted area across the staples set (rice, corn, cassava, sweet potato, taro, other tubers, sorghum); mirrors the staples production aggregate. Intermediate land measure between rice-only and total arable.
Total planted crop area (ha)	PODES 2003	184 / 1,951	133 / 389	Sum of planted area across all 26 PODES crops: staples (rice, corn, cassava, sweet potato, taro, other tubers, sorghum), legumes (peanut, soybean, mungbean), coffee & tea, spices (pepper, nutmeg, clove), and the remaining plantation, fiber, and tree-crop items (rubber, palm oil, cocoa, sugar palm, sugarcane, cashew, cinnamon, cotton, vanilla, tobacco, jute). Equivalent to the PODES native “total ag area” aggregate; the broadest available land measure.
Staples production (tons)	PODES 2003	185 / 1,945	133 / 382	Aggregate of subdistrict-level production in tons for rice, corn, cassava, sweet potato, taro, other tubers, and sorghum.
Legumes production (tons)	PODES 2003	167 / 1,722	117 / 320	Aggregate of peanut, soybean, and mungbean production in tons.
Coffee & tea production (tons)	PODES 2003	133 / 1,383	100 / 261	Aggregate of coffee and tea production in tons.
Spices production (tons)	PODES 2003	102 / 1,259	71 / 225	Aggregate of pepper, nutmeg, and clove production in tons.

Continued on next page

Table S1, continued from previous page

Outcome / variable description	Source / wave	Universe	Main city sample	Construction notes
(F) Derived surplus outcomes (PODES 2003 production minus Susenas 2002 / 2011 consumption)				
Rice surplus (tons): subdistrict rice production minus subdistrict rice consumption	PODES 2003 – Sus 2002	143 / 833	108 / 259	The surplus subtracts, from production in milled weight, the consumption per person of the surveyed subdistricts times the 2003 population of the villages with a rice output figure. The rice-import figure uses consumption minus production and restricts both terms to surveyed subdistricts with reported rice output. Bilateral rice shipments are not observed.
Staples surplus (tons)	PODES 2003 – Sus 2002	143 / 833	108 / 259	PODES 2003 production of rice, corn, cassava, sweet potato, taro, other tubers, and sorghum minus Susenas 2002 consumption of rice, corn, cassava, sweet potato, taro, and other tubers. The consumption aggregate does not include a separate sorghum item.
Legumes surplus (tons)	PODES 2003 – Sus 2002	130 / 759	95 / 222	Production minus consumption for legumes (peanut, soybean, mungbean).
Staples surplus using 2011 consumption (tons) [robustness]	PODES 2003 – Sus 2011	182 / 1,849	132 / 366	Robustness variant: pairs PODES 2003 production with Susenas 2011 consumption, where the main rows above use Susenas 2002. Trades the consumption–production year-gap (now 8 years) for higher consumption coverage and alignment with the 2010 outcome year.
Legumes surplus using 2011 consumption (tons) [robustness]	PODES 2003 – Sus 2011	167 / 1,666	117 / 313	Analog of the staples-2011 surplus for legumes.

Notes: Each count is the number of units holding every measure a row names, as urban centers / subdistricts. The universe is the 187 urban centers of the analysis universe and the 2,025 outer-island subdistricts in year-2000 boundaries; the main city sample is its 134 cities and the 405 subdistricts that make them up. A city enters the 1980 row when the census recorded its 1980 baseline directly, and a share or a rate counts where its numerator and a positive denominator are present. The rice-only price index is built for urban centers only, so the subdistrict counts of the price-index row cover the broad and crops-only indices; a dash marks a measure built for urban centers only.

S1.2 Earlier Household-Survey Outcomes

The 1996 Susenas wave supplies food expenditure shares, quantities, calories, and unit values. Food indices are constructed from household item-level expenditure and quantity; the rice comparison distinguishes local and higher-quality rice. The food-price, rice-quality, and price-change exhibits report their own samples and transformations. The price-change comparison uses cities observed in both 1996 and 2002. The coverage comparison in [Table A6](#) distinguishes survey-wave availability from the number of observations in each outcome regression.

S1.3 Coverage by Source and Sample

Susenas 2011 outer-island subdistrict-level coverage is 993 of 2,025 subdistricts (49.0% by count) but 66.9% of the outer-island population, with a strong urban-bias by sampling design: subdistricts in the largest population quintile are 5.0 times more likely to be sampled than those in the smallest, and subdistricts with $\geq 52\%$ urban population are essentially fully covered while those with $< 15\%$ urban are only 18.7% covered. The 2011 sample is appropriate for the structural-transformation analysis but describes the urban / kotamadya subset of the subdistrict panel. The 993-subdistrict figure describes the prices-and-wages module; the consumption-by-food-category aggregates in block (D) of [Table S1](#) come from the quarterly food-consumption module, whose coverage is much wider (1,954 of 2,025 subdistricts), which is why that block's sample counts differ from the coverage stated here. Susenas 2002 outer-island subdistrict-level coverage is 877 of 2,025 (43.3% by count, 61.6% by population). Aceh has zero Susenas 2002 coverage, a pre-existing structural coverage gap in the survey-frame design.

S1.4 Crop Units and Missing Observations

For the PODES 2003 crop totals, a total is set to missing when at least one village observation is absent and the observed village values sum to zero. Otherwise the observed sum is retained with a count of absent village observations; the same rule applies when subdistrict totals are summed to cities. An observed zero is retained when all component observations are present. Rice production used in the production and consumption comparison is expressed in milled weight. Susenas 2002 subdistrict-level rice consumption is winsorized at the 0.5% level for the surplus calculation. The Susenas 2011 consumption-by-food-category aggregates are not winsorized at the variable-construction stage.

Coffee & tea and spices surplus measures not constructed. PODES 2003 production for coffee, tea, and spices is recorded in tons; Susenas 2002 and 2011 consumption for the same items is recorded in monetary value (Rupiah), because the underlying items use distinct physical units across coffee, tea, pepper, nutmeg, and clove. There is no clean tonnage conversion without auxiliary price imputation, so the corresponding surplus measures are not constructed. Log production and log consumption value are reported as separate outcomes instead.

Land measure for the agricultural production function. Several subdistrict-level land measures are available from PODES 2003: rice planted area, staples planted area, and total planted crop area,

plus additional sub-aggregates for legumes (peanut, soybean, mungbean), plantation crops (rubber, palm oil, coffee, tea, cocoa), coffee & tea on their own, and spices (pepper, nutmeg, clove), each summed from the individual-crop planted-area variables. Individual planted-area variables are also retained separately for each of the 26 PODES crops (the staples and legumes items listed above, plus coffee, tea, pepper, nutmeg, clove, rubber, palm oil, cocoa, sugar palm, sugarcane, cashew, cinnamon, cotton, vanilla, tobacco, jute). The land endowment that enters the model is the satellite-derived built-up surface $\bar{T}_j$ (GHS-BUILT-S, fixed at the model-unit level; see the model section). It is measured consistently across rural and urban units and does not depend on crop-mix reporting. The PODES harvested-area measures remain available as agricultural-land alternatives for robustness. The rice-only measure sharply understates productive land in plantation-dominant outer-island districts (especially Sumatra and Kalimantan), the total arable measure is the broadest, and staples is an intermediate option.

S2 Geographic Units and Exposure Measurement

S2.1 Census Links Across Administrative Boundaries

Both the 2000 and the 2010 censuses record each person’s village, so the link between the two vintages is made at the village level and then aggregated. I use the BPS village-history file (*Histori Master File Desa*, 1998-2013), which has one row per village with that village’s administrative code in every vintage, so a 2010 census village and its year-2000 predecessor appear on the same row. Each 2010 village is counted in the year-2000 subdistrict of its year-2000 predecessor.

Three rules govern the matching. First, only rows that are a village in both vintages are used, since the file also includes subdistrict and district heading rows. Second, a 2010 village whose year-2000 predecessors lay in two different subdistricts is left unresolved; this affects five villages nationally. Third, where the file does not list a village, the village keeps the assignment from the boundary overlay, so no village is dropped.

The file covers 96.3% of the 77,480 year-2010 villages, and places 45,993 of them on the year-2000 outer-island map. Two checks support the construction. First, the number of 2010 villages each year-2000 subdistrict receives is compared with the number the 2000 census recorded there: 41.7% of outer-island subdistricts are within one village of that count and 87.1% are within a factor of 0.8 to 2.0, against 22.1% and 71.5% for a geometric assignment of the same villages. Second, on the Inner Islands, which this paper does not analyze and on which the construction was not chosen, the same comparison is 95.9% against 56.4%, so the improvement is not an artifact of tuning on the estimation sample. The total 2010 population is unchanged: a village changes subdistrict and is never created or dropped.

S2.2 Spatial Inputs and Analysis Units

The empirical analysis aggregates village- and individual-level data to two spatial units: 2000-vintage subdistricts and outer-island urban centers (UCs). The mapping between the original survey-year

administrative codes and the analysis units relies on a suite of crosswalks plus globally harmonized raster products. [Table S2](#) summarizes the spatial and crosswalk inputs.

Table S2: Spatial Boundary Definitions, Administrative Crosswalks, and Gridded Raster Inputs

Spatial input	Source	Description / use in this paper
(A) Spatial boundary definitions		
Transmigration village sites and polygons	Bazzi et al. (2016); replication archive	The exposure sample is the complete set of 814 program villages in the Transmigration Census analysis sample, settled during Indonesia’s Transmigration Program (1979-1988). Each village has the program-recorded year of settlement, the count of initially placed settlers, and the agro-climatic similarity of the assigned settlers; the village-population summary statistics use the 716 sites with GHS population measured on the earlier mapped footprints, while host classification matches all 814 to year-2000 census-village polygons by administrative code. Travel times are computed from the replication village locations for all 814. This set of villages is the spatial reference for constructing the agro-similarity treatment.
Urban-center (UC) delineation, outer-island Indonesia	GHS Urban Center Database 2025, GHS-UCDB R2024A V1.1 (European Commission, Joint Research Centre), fixed-boundary general-characteristics layer	Outer-island urban centers aggregated from 2000-vintage subdistricts, derived from the GHS-UCDB urban-center polygons; the reclassified analysis sample has 187 urban centers, every one of them with a travel time to a program village (see Appendix A.1). UCs are the headline analysis unit.
Functional urban area (FUA) delineation, outer-island Indonesia	GHS-FUA R2019A (European Commission, Joint Research Centre)	Wider commuting-zone polygons around the urban centers, shown in the descriptive maps.
2000-vintage subdistrict administrative boundaries	Badan Pusat Statistik (BPS), Indonesia’s central statistics agency: subdistrict boundary shapefiles	2000-vintage subdistrict polygons used as the village-aggregation target and the subdistrict-level analysis unit.
(B) Administrative-code crosswalks across census vintages		
Subdistrict 1976 / 1980 to 2000 / 2002 crosswalk (“Olken crosswalk”)	B. Olken, courtesy of replication / supplementary materials	Hand-built crosswalk linking 1980 Census subdistrict codes to the 2000/2002 administrative vintage. The published file excludes Aceh, East Timor, and the Malukus by construction, the regions where 1976–80 boundary data were not available. The build reads an edited version of that file, which restores subdistricts the published one omits, Aceh among them, under dated correction flags held in the file itself; a 1980 code with no crosswalk row is mapped to the year-2000 code with the same digits. The 1980-to-2000 subdistrict mapping is the workhorse for assembling the 1980 baseline-control panel.

Continued on next page

Table S2, continued from previous page

Spatial input	Source	Description / use in this paper
Village 2000 to 2010 crosswalk	BPS <i>Histori Master File Desa</i> (1998–2014) village-history file; boundary overlay of 2000 and 2010 village polygons as the fallback	The primary link between the two census rounds, and the one the city outcomes use. Each 2010 village is placed in the year-2000 subdistrict of its year-2000 predecessor, which the history file names on the same row; it covers 96.3% of the 77,480 year-2010 villages. A village the file does not list keeps the boundary-overlay assignment, and a village descending from predecessors in two different subdistricts is left unresolved. See Section S2 .
Subdistrict 2000 to 2010 correspondence	Derived by aggregating the village-level link above	The 2000-vintage subdistrict totals are built by summing the villages assigned above, so the subdistrict correspondence is an output of the village lineage and includes its fallback assignments. Direct polygon intersection of 2000 and 2010 subdistrict boundaries is used only where no village-level assignment exists.
Host-subdistrict classification	Program-village administrative codes and year-2000 census-village and subdistrict polygons	A subdistrict hosts a program village when the village records it as its administrative subdistrict or more than 0.1% of the village polygon lies inside it. The classification includes all 814 program villages, including the 98 without population measured on the earlier mapped footprints. The rural comparisons use entire subdistricts with and without a settlement.
(C) Travel-time and great-circle distance matrices		
Cost-distance (travel-time) matrix between UCs and transmigration villages	Constructed for this paper from an inherited general land-friction surface augmented with maritime connections; geographic alignment, pairwise calculations, and source limitations are documented in Appendix B.1	Pairwise UC×village travel-time matrix (minutes) used to construct the agro-similarity treatment as the unweighted mean over villages within the headline 300-minute window (weighted variants are robustness). The corresponding subdistrict-level matrix is built analogously for the surrounding-villages extension.
Great-circle (haversine) distance matrix between UCs and transmigration villages	Constructed for this paper from the centroid coordinates of UC and village polygons	Robustness companion to the cost-distance matrix; a free-space variant of the design-based recentering exercise in Appendix C.1 .
(D) Gridded raster products		

Continued on next page

Table S2, continued from previous page

Spatial input	Source	Description / use in this paper
GHS-POP gridded population estimates (1975, 1980, 1985, ..., 2020)	GHS-POP R2023A (European Commission, Joint Research Centre), 100m resolution	5-year epoch population grids calibrated to the BPS Sensus Penduduk totals at the year-2000 subdistrict level. Used as a check on the census population series. Satellite-derived recovery of 1980 population for subdistricts missing from the Olken crosswalk was scoped but not implemented, and no analysis variable in the paper draws on it; those subdistricts enter through the crosswalk corrections and the same-digits mapping described in panel (B).
GHS-SMOD gridded settlement-model classification (1975, 1985, 2015, 2020)	GHS-SMOD, R2022A for the 1975, 1985, and 2015 epochs and R2023A for 2020 (European Commission, Joint Research Centre), 1km resolution	Cell-level settlement-density classification (Urban Center / Urban Cluster / rural). Used as an alternative urban-share measure for robustness. The Urban Center boundaries themselves come from the GHS-UCDB polygon layer in panel (A).
GHS-BUILT-S built-up-surface raster (5-year epochs, 1975-2020)	GHS-BUILT-S R2023A (European Commission, Joint Research Centre), 100m and 1000m resolution	Built-up-surface fraction per grid cell. It is the land endowment $\bar{T}_j$ in the spatial-equilibrium model (Section 6); the satellite-baseline-recovery exercise it was also scoped for was not implemented.

The census as the 1980 baseline. Every 1980 quantity in this paper comes from the population census, for three reasons. First, there is no observed satellite record for 1980. The GHSL built-up series has Landsat input epochs at 1975, 1990, 2000 and 2014, and every other epoch is interpolated between them; the 1980 population grid is close to a fixed one-third blend of the 1975 and 1990 layers, with an implied interpolation weight of 0.28 to 0.33 in every island group. The satellite 1980 baseline is interpolated between published epochs. Second, the gridded population product is not independent of the census in any case, since its counts come from census-based population estimates and the built-up layer only redistributes them within an input unit. At the subdistrict level it tracks the 1980 enumeration at a correlation of 0.57 in levels and 0.62 in logs, with the grid-to-census ratio running from 0.23 at the tenth percentile to 1.87 at the ninetieth. Third, the observed comparison epoch after the program began is the 1990 one, and it already includes the program’s own settlement: within a subdistrict that split between census rounds, a successor’s share of the gridded population rises between 1975 and 1990 by more when that successor hosts transmigrant villages. The paper uses individual records from a 5% sample of the 1980 population census, expanded by census weights for the baseline controls.

S2.3 Historical Satellite Coverage

Civelli et al. (2023) report sparse 1975 Landsat coverage over much of the archipelago, Sumatra included, in an earlier GHSL release. That gap persists in the release used here, and its extent can be measured: GHSL publishes an input-count layer, GHS-SDATA_LDS-QUANT, giving the number of Landsat measurements behind each cell of each epoch (European Commission, Joint Research Centre 2024). On

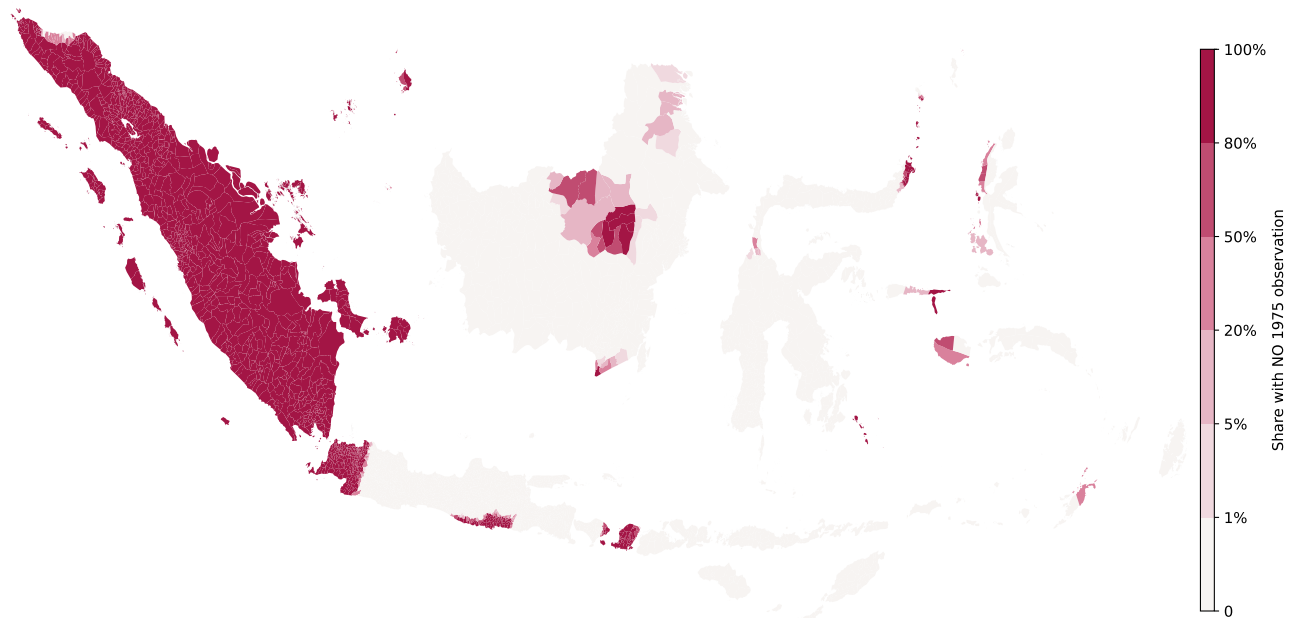
that layer **99.3% of Sumatra’s land cells have no 1975 Landsat observation at all**, against 0.7% in 1990, and 98.2% of the island’s 1975 built-up surface and 97.7% of its 1975 gridded population sit on cells with none. Across the outer islands as a whole, 60.7% of 1975 gridded population is unobserved. Where imagery is absent the producers solve the built-up prediction by spatio-temporal interpolation whose dynamic component is drawn from the epochs that *were* observed, which here begin in 1990, after the program. The satellite baseline therefore includes interpolated values informed by settlement recorded after the program. The census does not have this problem. It measures 1980 directly.

Table S3: Landsat Observation Coverage for the 1975 Epoch

Island group	Land cells	% cells with no 1975 observation	% of 1975 built-up	% of 1975 population
Sumatra	47,743,508	99.3	98.2	97.7
Java and Bali	13,919,709	17.7	26.5	20.4
Nusa Tenggara	6,727,677	6.5	28.6	34.8
Kalimantan	53,835,817	5.5	1.1	0.9
Sulawesi	18,731,243	1.5	3.8	3.6
Maluku	7,822,973	8.1	16.9	13.9
Outer islands, the estimation region		38.4	53.5	60.7

Notes: Computed from the Global Human Settlement Layer’s record of how many Landsat measurements support each 100-meter cell of its 1975 layer, 2023 release ([European Commission, Joint Research Centre 2024](#)), summed over the land cells of year-2000 subdistricts. A cell with none has 1975 built-up and population assigned by spatio-temporal interpolation. The last two columns give the share sitting on such cells, and the outer islands are every group except Java and Bali. Nationally the mean is 0.83 measurements per land cell for 1975 against 1.19 for 1990, 1.26 for 2000 and 1.43 for 2014.

Figure S1: Spatial Coverage of Landsat Observations for the 1975 Epoch



Notes: Year-2000 subdistrict boundaries, shaded by the share of the unit’s land cells with no Landsat measurement behind the 1975 epoch. 889 of 921 Sumatran subdistricts are entirely unobserved and 909 of 921 have under 80% of their land cells observed, the threshold [Civelli et al. \(2023\)](#) use to mark a village as missing. Source: the Global Human Settlement Layer, 2023 release ([European Commission, Joint Research Centre 2024](#)).

The gap follows the coverage of the Landsat archive itself. It is not scattered cloud cover: 61.7% of the variance in the subdistrict-level unobserved share sits between island groups and only 5.8% within a district. Sumatra, where the program placed most settlers, is the region with almost no record at all, and Indonesia had no Landsat receiving station of its own during the Multispectral Scanner era (Table S3 and Figure S1).

A gridded product reports a value for every cell in every published epoch, and nothing in the raster distinguishes a measurement from a model output. The grid is complete even where nothing was measured. So a 1980 satellite baseline for Indonesia cannot support a level comparison across places, cannot support a place-specific pre-period control, and cannot be used to validate the census, because over the treated region it is not an independent record of anything. The historical satellite archive limits these baseline uses.

A narrower use survives. Contemporaneous satellite measures, from epochs that were densely observed, remain sound: the urban-center definition used throughout this paper is a 2020-vintage product, and the built-up land measure enters only at a contemporaneous epoch. For the 1980 baseline itself the alternatives are the population census, used here through its weighted 5% sample, and the village census, PODES, which is the only other source with observations in 1980 at a fine spatial unit. The paper uses those two.

Travel-time construction. Appendix B.1 describes the inherited land-friction surface, the maritime connections added for this paper, the pairwise travel calculations, and their source-date limitations. It also presents the comparison of straight-line distance and travel-time connectivity in Figure B1.

S2.4 Allocation Across Changed Boundaries

Administrative boundaries change between census rounds, so all spatial objects are kept on the 2000-vintage subdistrict geography, which is the only vintage digitized consistently across the sources this paper uses. Census rounds on other vintages are mapped onto it by administrative code. Where a unit split between rounds, its counts are allocated across the successor units, so that recombining a split unit’s own successors recovers its enumerated total exactly. The allocation divides a split unit’s counts equally among its successors. That is weaker than it first appears: an equal split reproduces a subdistrict, district, urban-center or rural-subdistrict total only when the successors of every split unit inside it stay inside it, and where they fall in different reporting areas the equal division stands in for a true split this paper does not observe. Three quarters of split units have all their successors inside the same reporting aggregate, so the approximation is confined to the remaining quarter. Weighting instead by a gridded population layer would import into the 1980 baseline a quantity that is itself correlated with where settlers were placed, and those baselines enter the estimating equation as controls. The weighting choice is close to immaterial in any case: weighting instead by the 1975 gridded population moves no coefficient in Table 4 by more than a third of a standard error.

S2.5 Village-Level Links Across Survey Years

No village-level link, administrative or spatial, spans the years between 1980 and 1996, which sets how far back village-level work can reach. Three facts set the year and the unit. Four village crosswalk families exist, and they start in three different places. Cassidy and BPS's Master File Desa panel both begin at the 1998 administrative vintage; a direct 1996 map covers 88.9% of the 1996 wave across 26 provinces, Aceh, Maluku and Irian Jaya included; and a 1980 map covers West, Central and East Java, so it holds none of the outer-island sample. No village file has a vintage between 1980 and 1996, and spatial matching does not open those years either: the earliest village map is drawn on the 2000 census, and no map of the three is complete even against the village list of its own year, holding 95.2%, 99.2% and 95.5% of the SP2000, SP2010 and PODES 2021 village lists. The subdistrict is therefore the finest unit that reaches back to 1980 everywhere the paper estimates, through the Olken crosswalk at 82 to 91% per wave, which is the level at which this paper works. [Table S4](#) measures every route on the drives against every PODES wave.

Table S4: Village-Survey Coverage in Administrative Links and Boundary Maps

Wave	Village, crosswalk		Subdistrict, Olken		Village, spatial	
	Best file and column	Covered	Best column	Covered	Best layer	Covered
1976	Not matched	–	Not matched	–	Not matched	–
1980	PODES 80 Java ID1980	31.3	id1980	90.9	Not matched	–
1983	Not matched	–	id1983	88.9	Not matched	–
1986	Not matched	–	id1983	87.8	Not matched	–
1990	Not matched	–	id1990	89.5	Not matched	–
1993	Not matched	–	id1993	88.2	Not matched	–
1996	PODES 96-03 iddesa_1996	88.9	id1997	89.8	Not matched	–
2000	PODES 96-03 iddesa_2000	92.3	id1999	82.0	2000 census	79.9
2003	BPS MFD id2002	95.0	id2002	86.2	2000 census	77.8
2005	Cassidy ID_SE06	94.0	Not matched	–	2010 census	80.6
2008	Cassidy IDPODES	98.7	Not matched	–	2010 census	94.9
2011	Cassidy ID2011_1	98.8	Not matched	–	2010 census	96.3
2014	Cassidy IDPODES14	99.9	Not matched	–	2020 BPS	96.9
2018	Cassidy ID2015_1	96.3	Not matched	–	2020 BPS	96.3

Wave	Village, crosswalk		Subdistrict, Olken		Village, spatial	
	Best file and column	Covered	Best column	Covered	Best layer	Covered
2021	Cassidy ID2015_1	95.6	Not matched	–	2020 BPS	95.5
2024	Cassidy ID2015_1	90.7	Not matched	–	2020 BPS	91.5

Notes: Each row is a wave of the village census (PODES), and each “Covered” column is the share, in %, of the distinct codes of the wave found in the best-matching column of the village crosswalk named, of the subdistrict crosswalk of Olken, or of the best-matching village boundary map. The village pair reports whichever of four crosswalk families reaches furthest into the wave: Cassidy’s crosswalks and the village-history panel of the national statistics agency (Master File Desa), both from the 1998 vintage; a direct 1996 to 2000 to 2003 village map over 26 provinces, the only route into the 1996 wave; and a 1980 to 2000 village map of West, Central and East Java only, which reaches 90.1% of the 1980 wave inside Java and none of the outer islands. “Not matched” means that no file of the route spans the wave (no village crosswalk has a vintage between 1980 and 1996, the Olken crosswalk ends at the 2002 vintage, and the earliest village boundary map is drawn on the 2000 census) or, in the 1976 subdistrict cell, that under 10% of the codes match, because the 1976 codes have no province part; matches outside the span of a file are suppressed. The “Village, spatial” pair matches on the code against three village boundary maps: the 2000 and 2010 census maps (66,280 and 77,480 distinct codes) and the 2020 boundary file of the national statistics agency (81,912), which hold 95.2, 99.2 and 95.5% of the village lists of the 2000 census, the 2010 census and the 2021 village census. Shading of the coverage cells steps up at 80, 90 and 95%, and cells below 80% share the shade of “Not matched”. *Source:* the village census waves, the crosswalks and the boundary maps, each measured from the file itself. In the cells, BPS MFD is the Master File Desa panel of the national statistics agency (BPS), PODES 96-03 the 1996 to 2000 to 2003 village map, PODES 80 Java the 1980 to 2000 Java map, 2020 BPS the 2020 boundary file of the agency, and the entry in typewriter type is the column of the crosswalk that holds the codes of the wave.

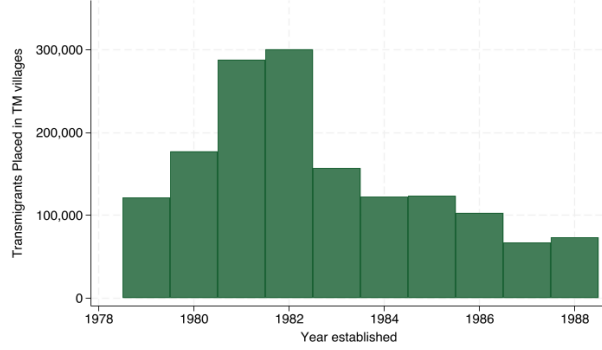
S2.6 Individual-Level Linkage Across Census Rounds

The 2000-2010 crosswalks above are administrative-boundary mappings and support unit-level aggregation only. Person- or household-level linkage of the 2000 and 2010 census microdata is not feasible: the 2000 census records do not contain digitized respondent names, and a full-sample probabilistic matching experiment linked only 0.08–0.43% of households at a placebo-calibrated 1% false-discovery rate. No result in this paper therefore relies on following individuals or households across census rounds; all 2000-2010 comparisons are aggregates within crosswalked spatial units.

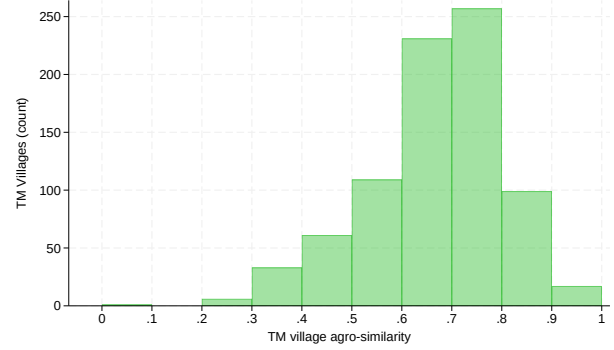
S2.7 Settlement Dates and Village Agro-Climatic Similarity

Figure S2: Transmigration Villages: Settlement by Year and Agro-Climatic Similarity

(a) Transmigrants Placed, by Settlement Year



(b) Village-Level Agro-Similarity (TM Villages)



Notes: Panel (a) counts the transmigrants placed in each settlement year from 1979 to 1988 across the 814 program villages. Panel (b) counts the same villages by their agro-climatic similarity to their settlers' places of origin, on its 0 to 1 scale. Source: Transmigration Census settlement records from the [Bazzi et al. \(2016\)](#) replication data.

S3 Additional Identification and Inference

S3.1 Common City Specification

The main city sample consists of the 134 outer-island urban centers within 300 travel minutes of a program village, with a directly recorded 1980 census baseline and measured 2010 outcomes. Agro-climatic similarity is the unweighted mean across program villages in that window. The city regressions control for the inverse-travel-time-weighted settler count, in thousands, and 16 destination characteristics averaged across those villages: three slope measures, ruggedness, log elevation, soil organic carbon, exchangeable sodium percentage, soil pH, two soil-texture and two drainage measures, distances to coasts and rivers, and mean rainfall and temperature over 1948-1978. The four city baselines are log population, log employment, the urban population share, and log agricultural employment in 1980. Tables state the transformations, standardization, and uncertainty method. The main city estimates use Conley standard errors allowing spatial dependence within 200 kilometers; village, lattice, and district exercises define their own samples and controls.

S3.2 Whole-Match and Origin-Composition Reassignments

The recentering of [Appendix C.1](#) uses two permutations of the settler match. The whole-match reshuffle permutes each village's agro-climatic similarity across sites within province and Repelita settlement wave; the origin re-draw holds each site's destination agro-climate fixed and re-draws only the settler-origin composition.

Realized and recentered exposure are both correlated with the measured village characteristics. On the main specification's agro-climatic control set, the sixteen village covariates averaged unweighted inside the same 300-minute window the exposure itself uses, the realized exposure has an R^2 of 0.775 on the estimation sample and the province-by-cohort recentered measure 0.556. [Table S5](#) subtracts the

expected exposure from the treatment, reporting $\tilde{z} = z - \mu$, with coefficients expressed per standard deviation of the recentered measure. Their magnitudes are not directly comparable to the headline coefficients. The agricultural share reads -0.016 with a standard error of 0.011 and log modern-sector employment $+0.038$ with a standard error of 0.038 , so of the three compositional outcomes only the modern employment share is statistically significant.

The origin-composition reassignment holds each site’s destination fixed and re-draws only the settler-origin composition, which is the more literal form of the design, since the destination geography is a predetermined exposure component and the shock is which settlers arrived. This is feasible because the similarity index of [Bazzi et al. \(2016\)](#) is linear in the origin composition, $AS_v = \sum_o \omega_{ov} s(o, v)$, where $s(o, v)$ is the origin-to-destination dyadic similarity and ω_{ov} is the share of village v ’s settlers born in origin district o . Their replication data provide the full dyadic grid (125 Java and Bali origins by 814 villages) together with the realized migrant flows that give ω , so I can hold each site’s similarity column $s(\cdot, v)$ fixed and swap in another village’s origin composition $\omega_{\cdot v'}$, permuted within province-by-wave cells; every settlement village is linked to the grid by its rank in the census-code ordering. The composition-weighted grid reproduces the published village index up to an affine rescaling ($r = 1.00000$, slope 0.56), so the rebuilt exposure and its expected value are differenced on their own scale and then standardized, and the swap is the same reassignment [Bazzi et al. \(2016\)](#) perform in their own analysis, holding the destination fixed and re-choosing origins. Holding the destination fixed raises the correlation between realized and expected exposure to 0.96 , deep in the range where a constructed exposure typically collapses under recentering, and [Table S6](#) shows that this one does. Per standard deviation of the origin-recentered exposure, the modern employment-share coefficient is $+0.027$, statistically significant, and the log agricultural-employment coefficient is -0.100 , marginally significant. The agricultural-share coefficient is -0.011 and imprecise; log population and log modern-sector employment have negative, statistically insignificant coefficients of -0.063 and -0.006 . The design-based inference built from the same draws is weaker still, returning $p = 0.545$ for the agricultural share where the analytic p is near 0.006 ([Table S7](#)). One standard deviation of the origin-recentered treatment is about 30% of a standard deviation of realized exposure, so the column retains real variation and its nulls speak to variation in origin composition. These regressions use variation in settler origins conditional on the sites. The headline also uses variation associated with site land after the linear controls; [Appendix C.1](#) reports the components separately and discusses the qualifications from the productivity and 1980 tests.

That correlation of 0.96 is easy to mistake for evidence that the recentering barely changed the treatment, and it means the reverse. Since $\tilde{z} = z - \mu$, the correlation between the raw and the recentered exposure is $[\text{SD}(z) - \text{corr}(z, \mu)\text{SD}(\mu)]/\text{SD}(\tilde{z})$, which falls as $\text{corr}(z, \mu)$ rises: subtracting something closely aligned with the raw measure removes most of it and leaves a remaining component weakly correlated with it. Under the scalar permutation, where $\text{corr}(z, \mu)$ is 0.75 , the recentered and raw measures still correlate at 0.60 . Under the origin permutation they correlate at 0.28 , with a rank correlation of 0.09 ; 49% of cities move from above-average to below-average exposure or the reverse, and the average city shifts 41 places in a ranking of 134. The robustness column is therefore

a different regression, identified off a treatment that orders the sample almost independently of the one the headline uses, and its coefficients are not on a scale the headline can be compared against. Agro-climatic match quality is mostly a property of the land. Once a site's destination is held fixed, the arriving cohort can move that site's similarity only so far, and the destination accounts for roughly 91% of the variance in exposure. The agro-climatic controls explain much of that component. Their R^2 on realized exposure is the 0.775 reported above, so roughly a quarter of the exposure's variance is variation no linear combination of them reproduces. Re-drawing the origin composition strips the destination component out and leaves the variation that is genuinely about which settlers arrived, and only the modern employment share survives on that remainder. Under the coarser province-only stratification the remaining variance share is similar, 10% against 9%, confirming that the correlation comes from the land's dominance under either stratification. Reshuffling the whole village match instead leaves 43% of the variance, so the two exercises bracket the design. The recentered estimates differ from the headline estimates under both reassignments. Because each column standardizes its own regressor, the two recentered coefficients are not on a common scale. One standard deviation of the origin-recentered treatment is about 30% of a standard deviation of realized exposure, against 66% for the whole-match version.

Table S5: Design-Based Robustness: Agro-Similarity Recentered Against Counterfactual Settler Assignments (Borusyak & Hull)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ln pop	ln emp	unemp	ln ag	ag share	ln modern	modern share
Agro-similarity recentered $\tilde{z}^{PW}$	0.011 (0.050)	0.012 (0.049)	-0.008* (0.005)	-0.036 (0.075)	-0.016 (0.011)	0.038 (0.038)	0.021** (0.010)
Settler count	✓	✓	✓	✓	✓	✓	✓
Base ln pop 1980	✓	✓	✓	✓	✓	✓	✓
Base ln empl 1980	✓	✓	✓	✓	✓	✓	✓
Base ln agri empl 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Observations	134	134	134	134	134	134	134

Notes: Each column is one regression of a 2010 population census outcome on recentered agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Check marks show the controls. Recentered agro-similarity is agro-similarity minus its expected value over 2,000 random reassignments of village agro-climatic similarity among the villages settled in the same province and the same five-year plan (1979 to 1983 or 1984 to 1988), standardized. Outcomes are log population, log employment, the unemployment rate, log agricultural employment, the agricultural employment share, log modern-sector (manufacturing and services) employment and the modern-sector employment share; shares and the unemployment rate are fractions. Standard errors are in parentheses. Agro-similarity recentered, written $\tilde{z}^{PW}$, is agro-similarity minus its expected value under the reassignments within province and five-year plan (the μ^{PW} of the appendix's recentering equation), standardized; the settler count weights the settlers of each transmigration village within 300 travel minutes by the inverse of its travel time, in thousands; and the Base ln rows are logs of 1980 population (pop), employment (empl) and agricultural employment (agri empl). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Randomization-inference p-values from the same 2,000 reassignments are in column (2) of [Table S7](#).

Table S6: Origin-Permutation Recentering: Destination Held Fixed, Origin Re-Drawn (UC)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ln pop	ln emp	unemp	ln ag	ag share	ln modern	modern share
Agro-similarity							
origin-recentered $\tilde{z}^{orig}$	-0.063 (0.047)	-0.060 (0.045)	-0.005 (0.004)	-0.100* (0.058)	-0.011 (0.013)	-0.006 (0.056)	0.027*** (0.009)
Settler count	✓	✓	✓	✓	✓	✓	✓
Base ln pop 1980	✓	✓	✓	✓	✓	✓	✓
Base ln empl 1980	✓	✓	✓	✓	✓	✓	✓
Base ln agri empl 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Observations	134	134	134	134	134	134	134

Notes: Each column is one regression on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. The regressor is origin-recentered agro-similarity, standardized: realized exposure minus its mean over 2,000 draws that permute the settlers' origin composition across villages within province and settlement wave, holding each village's similarity to every origin region fixed. The outcomes are 2010 census log population, log employment, unemployment rate, log agricultural employment, agricultural employment share, and log employment and employment share in manufacturing and services. Standard errors are in parentheses; exact randomization-inference p-values from the same draws are in [Table S7](#), column (4). Agro-similarity origin-recentered, written $\tilde{z}^{orig}$, is agro-similarity minus its mean over the draws that re-draw the settlers' origin composition, standardized; ln (the natural log), emp (employment), unemp (the unemployment rate), ag (agricultural) and modern (manufacturing-plus-services employment), each share a fraction of total employment; the settler count weights the settlers of each transmigration village within 300 travel minutes by the inverse of its travel time, in thousands. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S3.3 Assignment-Based Inference

[Table S7](#) reports assignment-based inference for both schemes.

Table S7: Randomization-Inference p -Values by Statistic and Permutation Scheme

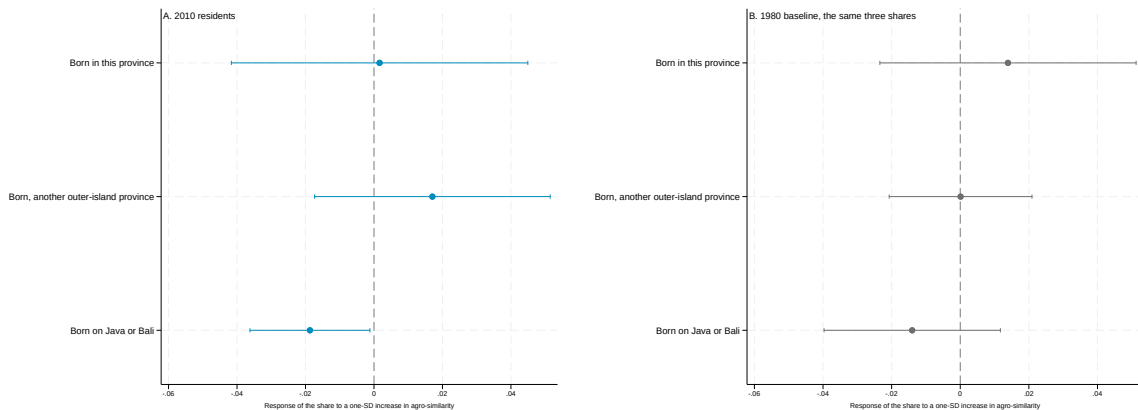
Outcome	Conley 200 km (1)	Whole-match		Origin		Controlling for expected exposure			
		recentered		recentered		t -statistic		coefficient	
		prov.	province	prov.	province	prov.	province	prov.	province
		×wave (2)	(3)	×wave (4)	(5)	×wave (6)	(7)	×wave (8)	(9)
ln population	0.389	0.823	0.881	0.211	0.537	0.663	0.668	0.831	0.837
ln employment	0.674	0.824	0.910	0.304	0.332	0.778	0.763	0.885	0.888
Unemployment rate	0.216	0.095	0.736	0.431	0.647	0.488	0.517	0.738	0.721
ln agricultural employment	0.075	0.630	0.393	0.158	0.244	0.447	0.429	0.677	0.632
Agricultural employment share	0.006	0.336	0.246	0.545	0.582	0.196	0.189	0.483	0.468
ln modern employment	0.011	0.599	0.688	0.938	0.906	0.181	0.192	0.615	0.609
Modern employment share	0.000	0.180	0.134	0.091	0.222	0.050	0.052	0.336	0.317
Observations	134	134	134	134	134	134	134	134	134

Notes: Every entry is a two-sided p -value for the null of no effect of agro-similarity on the outcome, on the main city sample. Each cell is shaded by the strength of the evidence, deepening green from the 10% to the 1% level, with pale pink for a p at or above 0.10. Column 1 is the analytic p -value from Conley SE within 200km, behind the stars of Table 4. Columns 2 to 9 are exact randomization inference over 2,000 counterfactual assignments of settlers to sites: the observed statistic is compared with the distribution of the same statistic when the settler match is re-drawn, stratified by province and Repelita settlement wave (even columns) or by province (odd columns). Columns 2 and 3 test the coefficient on the recentered exposure under the whole-match reshuffle, and columns 4 and 5 under the origin re-draw, which holds each site's destination agro-climate fixed and re-draws only the settler-origin composition. Columns 6 to 9 test the coefficient that controls for expected exposure (Table C4): columns 6 and 7 its t -statistic with a heteroskedasticity-robust (HC1) standard error, columns 8 and 9 the coefficient itself. The t -statistic is the test the appendix reports. With the expected exposures and the controls partialled out, the realized exposure has a residual standard deviation of 0.41 on the scale of the standardized treatment, against 0.20 on average across the reshuffles, and none of the 2,000 reshuffles reaches it under either stratification, so a placebo coefficient is estimated on about half the variation behind the observed one, a difference the t -statistic divides out. At the 10% level, in both columns of a pair, the whole-match recentering rejects for no outcome; the origin recentering rejects for no outcome; the t -statistic test of the controlled coefficient rejects for the modern employment share; the test of the coefficient itself rejects for no outcome. Of the cells whose pair does not reject, 2 clear 10%: the unemployment rate under the whole-match recentering stratified by province and wave (0.095); the modern employment share under the origin recentering stratified by province and wave (0.091).

S3.4 Birthplace Composition and the Baseline Census

I compare the two decompositions side by side at the only resolution both censuses support. SP1980 records place of birth at the province level, so the same-district and same-province cells of the 2010 decomposition cannot be separated at baseline; collapsing them into one same-province share makes the two panels the same object, and it is why the comparison has three categories where Figure 13 has four. The 2010 own-province coefficient is close to zero, and the other-outer-island-province coefficient is positive but imprecise. The Java-or-Bali coefficient is negative in both years; all three 1980 coefficients are imprecise (Figure S3).

Figure S3: Where Residents Were Born, 2010 Against the 1980 Baseline, at the Three-Way Resolution



Notes: Each panel regresses three exhaustive birthplace shares on standardized agro-similarity on the main city sample (outer-island urban centers with a transmigration village within 300 travel minutes and a 1980 census baseline; 134 cities in both panels), in the main specification with the full set of controls and Conley SE within 200km. Markers are the coefficients, in share units, and bars their 90% confidence intervals from the same standard errors; the three coefficients in each panel sum to zero by construction, which both panels satisfy exactly. Panel A covers 2010 census residents born 1955 to 1990, with the same-district and same-province categories of Figure 13 added together; panel B covers all 1980 census residents at the same three categories. The two panels share one horizontal scale. Table C6 adds the 1980 shares of the same birthplace categories to the main specification.

S3.5 Exposure Restricted to Post-Baseline Settlements

Restricting the exposure to the villages settled after the count leaves it nearly unchanged and lowers the settler-count control by roughly a fifth. No city loses its hinterland (Table S8).

Table S8: Exposure From All Villages and From Post-1980 Villages

	All 814 villages	687 post-1980 villages
<i>Exposure (unweighted mean similarity, 300 minutes)</i>		
Mean	0.6800	0.6828
Standard deviation	0.1025	0.1016
Minimum	0.3777	0.3777
Maximum	0.9563	0.9563
Correlation with the other, levels		0.9901
Correlation with the other, ranks		0.9762
<i>Villages within 300 travel minutes</i>		
Mean per city	30.4	25.8
Median per city	12	11
Cities losing exposure entirely		0
<i>Settler count (inverse travel time weights, 300 minutes, thousands)</i>		
Mean	0.466	0.365
Correlation with the other		0.8704
Urban centers (main city sample)		134

Notes: Computed over the main city sample. The exposure is the mean agro-climatic similarity of the transmigration villages within 300 travel minutes, and the settler count sums their settlers weighted by inverse travel time, in thousands. The left column builds both from all 814 villages, as the main specification does; the right column from the 687 villages settled after the 1980 census, dropping the 127 placed by 1980, which hold 19.5% of all settlers. No city loses its hinterland under the restriction.

S3.6 Settler-Retention Proxy Definitions

The settler stay-share featured in the text divides a village’s 2000-census residents born in Java or Bali by the individuals the program placed in the village, under two restrictions that each remove one class of non-settler the census can identify: residents whose recorded residence five years before enumeration was still Java or Bali, and residents born in Java or Bali after their own village was settled. The broader ratio without the two restrictions, which [Figure 14](#) plots, counts anyone who reached the village from the main islands on their own after the program ended, so it proxies for retention. The restrictions remove a median of 10% of the numerator, yet the conditional gradient on agro-climatic similarity moves only from 0.078 to 0.080, the definitions correlate above 0.98, and every version stays indistinguishable from zero under both the wild cluster bootstrap and Conley standard errors ([Table S9](#)). The definitions give similar estimates in this sample. No version of the numerator recovers the settlers’ locally born children, who are outer-island-born and never enter it, so the measure cannot establish the retention of all original settlers and their descendants.

Table S9: Settler-Retention Proxy Under Alternative Numerator Definitions

Numerator definition	Gradient on agro- similarity	Island- clustered std. error	Wild bootstrap p-value	Conley 200 km p-value	Corr. with the plotted measure
All Java/Bali-born residents (plotted measure)	0.078	(0.106)	0.69	0.57	1.000
Excluding arrivals in the previous five years	0.080	(0.093)	0.67	0.52	0.989
Excluding those born after settlement	0.071	(0.091)	0.67	0.54	0.995
Excluding both groups	0.080	(0.086)	0.66	0.49	0.985

Notes: Each row is the slope of one stay-share definition on the agro-climatic similarity of the village itself (an index from 0 to 1), across the same 808 transmigration villages, with the seven controls of Figure 14: the shares of land with slopes of 0-2%, 2-8% and 8-30%, ruggedness, log altitude, distance to the coast, and mean 1948-1978 temperature. The stay share is the residents born in Java or Bali in the 2000 census over the individuals the program placed in the village; rows 2 to 4 remove from that numerator the residents who arrived in the five years before the census, those born after the village was settled, or both. Parentheses hold standard errors clustered by island (four clusters); the wild cluster bootstrap p-value uses Webb weights over the four islands, and the p-value uses Conley SE within 200km over village centroids. Corr. with the plotted measure is the correlation, across the same villages, between the row's stay-share and the one plotted in the text (row 1).

S3.7 Settler-Count Controls in City Regressions

The headline main specification controls for the inverse-distance-weighted count of nearby transmigrant settlers ($SettlerCount_c^{IDW}$, in thousands) in addition to the agro-climatic similarity exposure. Because the agro-similarity measure and the migrant-count control are built over the same villages and the same travel-time window, the two are correlated, and the choice of functional form for the migrant-count control could in principle absorb some of the agro-similarity variation. This appendix reports the headline UC-level regressions under three alternative functional forms for the migrant-count control:

- (i) **Linear count** (thousands of settlers, IDW): the form used in the main specification.
- (ii) **Log count** (log of the IDW-weighted settler count): tests sensitivity to a constant-elasticity functional form.
- (iii) **Count + count-squared** (count and its square, both reported): probes non-linearity in the count effect, motivated by the gap in significance between (i) and (ii).

Findings. The agro-similarity coefficients are essentially invariant to the form of the settler count. Manufacturing employment is +0.249 under the linear count ($p = 0.039$), +0.266 under the log ($p = 0.033$) and +0.274 under count and count squared ($p = 0.023$), and agricultural employment is negative at the 10% level under all three (-0.166 , -0.171 and -0.164). The settler term itself is positive on every outcome under the linear count and under the log, and significant at the 5% level on 5 of the seven levels outcomes under the linear count and 5 under the log (population, total employment, manufacturing, construction and services in both), with agricultural employment and the unemployment rate positive but imprecise: more settlers placed near a city go with a larger city and more workers in every modern sector, and the count control absorbs that. Under count plus count squared the agricultural terms are +0.137 and -0.009 , both imprecise ($p = 0.609$ and 0.875), so these estimates do not establish a non-monotone agricultural association; the concave shape is in manufacturing, where the count enters

at +0.604 and its square at -0.126 , both at the 5% level, and more weakly in services, whose squared term is -0.055 at the 10% level. The agro-similarity coefficients are similar across the count-control specifications examined.

The marginal effect of the settler count, $\partial Y/\partial \text{count} = \alpha_1 + 2\alpha_2 \cdot c$ from specification (iii), traces how the response changes across the empirical range of settler counts, with 90% Conley confidence bands, for population, total employment, agricultural employment, modern-sector employment, the household food-expenditure share, and housing price. For population, total employment and modern-sector employment the marginal effect is positive across the whole plotted range, the 5th to the 95th percentile of the settler count, and largest where settlement is thinnest: the estimated marginal association decreases as the settler count rises, and the curves flatten without turning down inside the range. Agricultural employment does not respond detectably at any count. The household food-expenditure share falls with the count, by less as settlement thickens. In the count and count-squared estimates for the main outcomes (Table S10, below) the squared term is negative on every employment outcome, significant at the 5% level for manufacturing and at the 10% level for services. Figure S4 plots these marginal effects with their bands.

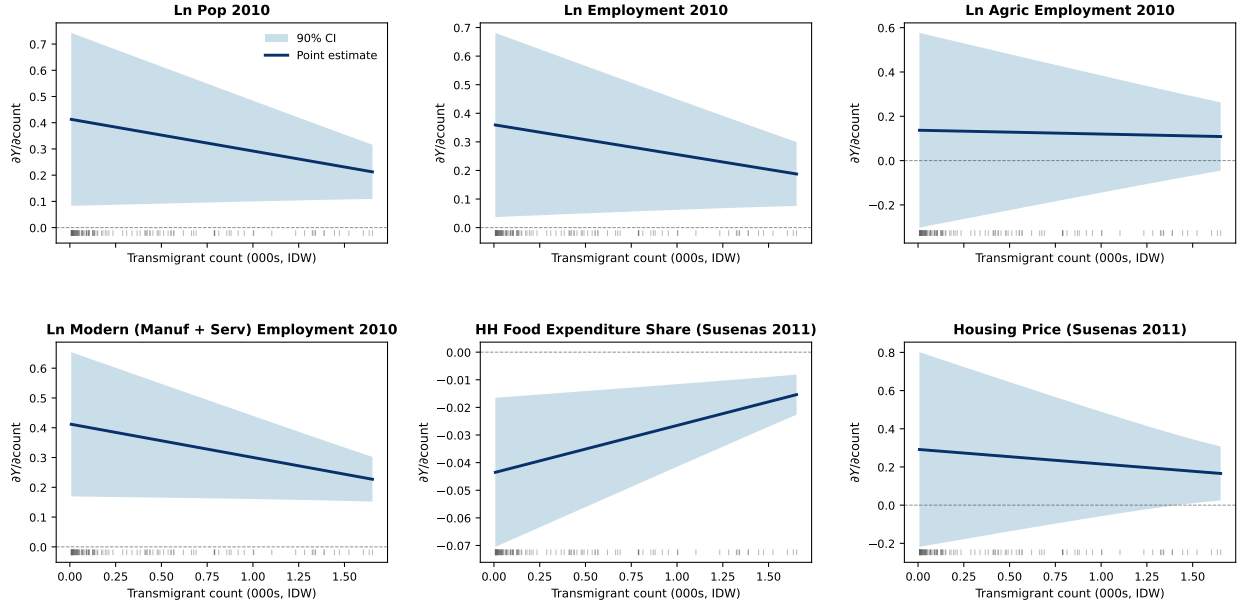
Table S10: Population and Employment in 2010 With Settler-Count Coefficients

	Population + Employment			Log sector employment			
	(1) ln pop	(2) ln employed	(3) unemp rate	(4) L agric	(5) L manuf	(6) L constr	(7) L serv
Agro-similarity	0.063 (0.060)	0.034 (0.058)	0.009 (0.006)	-0.164* (0.093)	0.274** (0.120)	0.058 (0.072)	0.155** (0.068)
Settler count (inverse travel time weighted)	0.414** (0.201)	0.360* (0.196)	0.021 (0.017)	0.137 (0.268)	0.604** (0.251)	0.438*** (0.162)	0.420*** (0.144)
Settler count squared	-0.061 (0.042)	-0.052 (0.040)	-0.005 (0.005)	-0.009 (0.055)	-0.126** (0.056)	-0.045 (0.035)	-0.055* (0.032)
Baseline Pop 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Total Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agri Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Outcome mean	11.842	10.933	0.122	9.822	7.556	8.011	10.136
Outcome SD	0.854	0.840	0.040	0.896	1.356	1.111	0.966
Observations	134	134	134	134	134	134	134

Notes: Each column is a separate urban-center regression on standardized agro-similarity on the main city sample. Every column is estimated in the main specification with the full set of controls and Conley SE within 200km. The settler count within 300 minutes, weighted by inverse travel time and in thousands, enters with its square, and both coefficients are shown. Outcomes are from the 2010 population census (the unemployment rate is a fraction). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Count and count-squared estimates for the main outcomes.

Figure S4: Conditional Marginal Associations With the Settler Count



Notes: Each panel plots the marginal effect of the settler count, $\partial Y / \partial \text{count} = \alpha_1 + 2\alpha_2 c$, from one regression on the main city sample of the outcome on agro-similarity, the settler count within 300 minutes (thousands, weighted by inverse travel time) and its square, 16 agro-climatic covariates within 300 minutes, and four 1980 baselines. For population, employment, and agricultural employment the two count coefficients are those of Table S10. The effect is evaluated at 30 points from the 5th to the 95th percentile of the settler count across the 138 urban centers with a transmigration village within 300 travel minutes, tick marks along the bottom show their counts in that range, and the shading is the 90% confidence interval from Conley SE within 200km. Outcomes are 2010 census log population, log employment, log agricultural employment and log modern (manufacturing plus services) employment, and the 2011 household-survey (Susenas) mean household food expenditure share and housing price index, the weighted median rent of the urban center relative to the national level.

S3.8 Settler-Count Controls in Rural Public Provision

Leaving the settler control (for a subdistrict with a settlement, the number of settlers placed in the transmigration villages that overlap it; for one without, the settlers within 300 travel minutes, weighted by inverse travel time) out of the regressions of public provision in rural subdistricts with and without a settlement, or counting the settlers within 300 travel minutes for the subdistricts with a settlement too, changes the stars of 2 and the sign of 1 of the 56 corresponding estimates on the sensitivity sample (t-ratios between -0.07 and 0.04 where the sign changes), and the largest move of an estimate is 0.16 of the main specification's Conley standard error; the table leaves West Kalimantan out of its 1980 to 2011 presence-share columns, and this check keeps it (Table E26).

S4 Additional Urban Employment Results

The census outcomes behind the headline table are reported in the three-sector levels and shares layouts that the main text consolidates into the modern-sector aggregation of [Table 4](#) ([Table S11](#) and [Table S12](#)).

Table S11: Population and Sectoral Employment Levels in 2010 (aggregated to urban centers)

	Population + Employment			Ln Sectoral Employment			
	(1) ln pop	(2) ln employed	(3) unemp rate	(4) L agric	(5) L manuf	(6) L constr	(7) L serv
Agro-similarity	0.051 (0.059)	0.024 (0.057)	0.008 (0.007)	-0.166* (0.093)	0.249** (0.120)	0.049 (0.067)	0.144** (0.067)
Settler count	✓	✓	✓	✓	✓	✓	✓
Baseline Pop 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Total Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agri Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Outcome mean	11.842	10.933	0.122	9.822	7.556	8.011	10.136
Sector empl share (2010)				0.374	0.047	0.060	0.470
Outcome SD	0.854	0.840	0.040	0.896	1.356	1.111	0.966
Observations	134	134	134	134	134	134	134

Notes: Each column is one regression of the outcome on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Standard errors are in parentheses. Outcomes are 2010 census log population, log employment, the unemployment rate (a fraction), and log employment by sector; the sector employment share row is the share of each sector in 2010 employment. SD is the standard deviation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S12: Population and Sectoral Employment Shares in 2010 (aggregated to urban centers)

	Population + Employment			Sectoral Employment Shares			
	(1) ln pop	(2) ln employed	(3) unemp rate	(4) agric %	(5) manuf %	(6) constr %	(7) serv %
Agro-similarity	0.051 (0.059)	0.024 (0.057)	0.008 (0.007)	-0.051*** (0.018)	0.010** (0.005)	-0.002 (0.004)	0.056*** (0.018)
Settler count	✓	✓	✓	✓	✓	✓	✓
Baseline Pop 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Total Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agri Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Outcome mean	11.842	10.933	0.122	0.374	0.047	0.060	0.470
Outcome SD	0.854	0.840	0.040	0.183	0.042	0.027	0.145
Observations	134	134	134	134	134	134	134

Notes: Each column is one regression of the outcome on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Standard errors are in parentheses. Outcomes are 2010 census log population, log employment, the unemployment rate, and the share of each sector in total employment (agriculture, manufacturing, construction and services); the rate and the shares are fractions between 0 and 1. Baseline rows are 1980 logs of population, total employment and agricultural employment, and Urban share 1980 is the urban share of the 1980 population; SD is the standard deviation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S4.1 Service Subsectors

[Table S13](#) reports employment in trade, transport, and the service group comprising information, finance, education, and health. These categories do not exhaust total service employment; their shares use total employment as the denominator.

Table S13: Urban Employment by Service Type 2010 (aggregated to urban centers)

	All (Ln)	Service Types (Ln)			All (%)	Service Types (% of All Empl)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Services	Trade	Transport	Skilled	Services	Trade	Transport	Skilled
Agro-similarity	0.144** (0.067)	0.114 (0.077)	0.135*** (0.049)	0.105 (0.068)	0.056*** (0.018)	0.013*** (0.005)	0.006** (0.003)	0.009* (0.005)
Settler count	✓	✓	✓	✓	✓	✓	✓	✓
Baseline Pop 1980	✓	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓	✓
Total Empl 1980	✓	✓	✓	✓	✓	✓	✓	✓
Agri Empl 1980	✓	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓	✓
Outcome mean	10.136	9.137	7.889	8.442	0.470	0.172	0.053	0.086
Sector empl share (2010)	0.470	0.172	0.053	0.086				
Outcome SD	0.966	0.999	1.057	0.915	0.145	0.051	0.024	0.028
Observations	134	134	134	134	134	134	134	134

Notes: Each column is one regression of the outcome on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Standard errors are in parentheses. Columns 1 to 4 are 2010 census log employment in services and in three selected service groups; columns 5 to 8 are their shares of total employment, as fractions between 0 and 1. Trade is industry 11; transport is transportation and warehousing (13); the group labeled skilled services combines information and communication, finance and insurance, education, and health (14 to 17). These groups exclude hotels and restaurants (12) and social, government, and other services (18), so they do not exhaust services; the grouping does not identify private ownership. The sector employment share row is the share of 2010 employment in the service type of each column. Baseline rows are 1980 logs of population, total employment and agricultural employment, and Urban share 1980 is the urban share of the 1980 population; SD is the standard deviation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S4.2 Manufacturing and Services

The headline estimates aggregate manufacturing and services into the modern sector. The manufacturing and services columns of [Table S11](#) report log employment responses of +0.249 and +0.144, respectively. Their employment-share responses are +0.010 and +0.056 in [Table S12](#). The log coefficients describe proportional differences within each sector.

S4.3 Robustness of the Migration Bound and the Agricultural-Share Decomposition

These exhibits belong to the migration bound of [Appendix E.3](#) and the decomposition of [Appendix E.4](#).

Robustness. I repeat the test under the assumptions that could overturn it. Two are substantive. Exposure may shift the arrivals' own sector mix, which would lower a and with it the requirement; the test rejects at the high-exposure city's own value. Exposure is also imperfectly balanced on the 1980 agricultural share, so conditioning on that share turns Δ into a 1980-to-2010 change in the share; the test still rejects ([Table S14](#)).

The required composition of arrivals

The question can be pushed some way without measuring the arrivals' sector at all. Suppose every worker the exposure adds to a city works outside agriculture, the most generous assumption available to the migration account. With the incumbent workforce fixed and net employment growth equated to

Table S14: The Migration Bound Under Alternative Assumptions

	Required inflow	One-sided p
Headline: 2010 base, lifetime migrant stock, $a = 0.269$	93.9%	0.034
Five-year movers instead, $a = 0.210$	44.7%	0.025
a at the high-exposure city's own value, 0.228	53.5%	0.026
1980 agricultural share added as a control	95.4%	0.001
Both corrections together	54.1%	0.001
Conservative 1980 base, pre-treatment $a = 0.402$	47.6%	0.025
Conservative 1980 base, 2010 $a = 0.210$	17.0%	0.065

Notes: Log employment and the agricultural employment share are each regressed on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. The required inflow is the gross inflow, as a percentage of the workforce, that [Equation 21](#) needs to produce the estimated fall in the agricultural share per standard deviation of agro-similarity when arrivals farm at share a and every incumbent stays in its sector; the base is the mean agricultural share of the sample in 2010, or in 1980 in the conservative rows. a is the agricultural share of 2010 workers born outside the district (the lifetime migrant stock) or who moved district within five years (five-year movers), or of 1980 workers born outside the province; the high-exposure city is one standard deviation higher in agro-similarity, and its a adds the estimated effect of that rise. The one-sided p , from the same standard errors, tests the null that the log employment response is at least as large as the requirement; its cell takes one shade at 0.10 or above and deepening shades below 0.10, 0.05 and 0.01.

arrivals, a log employment response β_L implies an arrivals share $1 - \exp(-\beta_L)$ of realized employment, and at a zero agricultural share among them the city share falls by that fraction times the baseline share. At the estimated $\beta_L = 0.023$ and a sample mean agricultural share of 0.374 this ceiling is 0.9 percentage points, 17.1% of the estimated 5.1-point fall. An alternative version replaces the employment response with the estimated 3.5 percentage-point rise in the share of workers born outside the district and multiplies by the locally born agricultural share of 0.404, giving 1.4 points, or 28.1%.

Both point-estimate ceilings sit below the fall: under the stated assumptions, arrivals account for 28.1% of it at most. [Figure S5a](#) reports the two ceilings, the ceiling at the top of the employment response's 90% range, and the measured composition component.

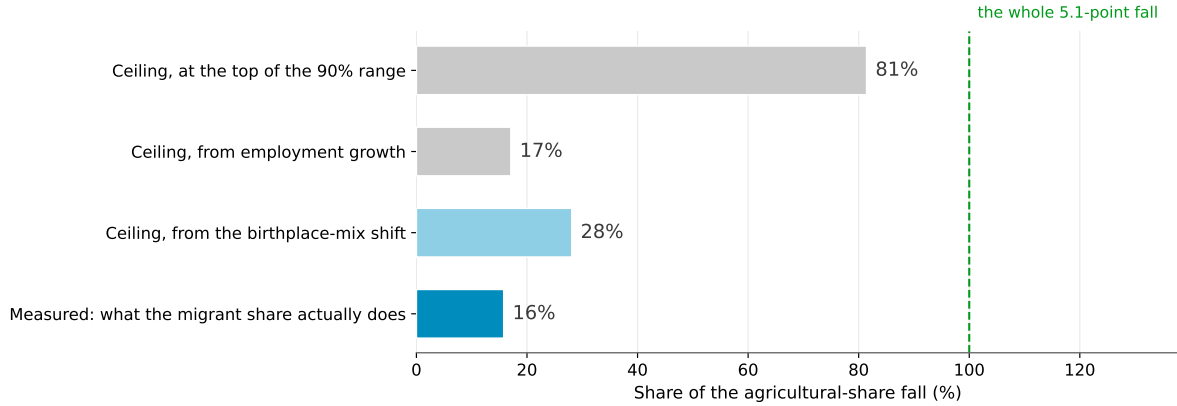
The composition channel is present but small. Migrants are markedly less agricultural than the locally born, 20 to 30% against 40%, which makes a migrant-composition account plausible before it is measured. It stays small because the heavily agricultural group is two-thirds of the workforce and exposure raises the migrant share by only a few percentage points, which cannot produce the estimated 5.1-point fall ([Figure S5b](#)).

Estimates and robustness

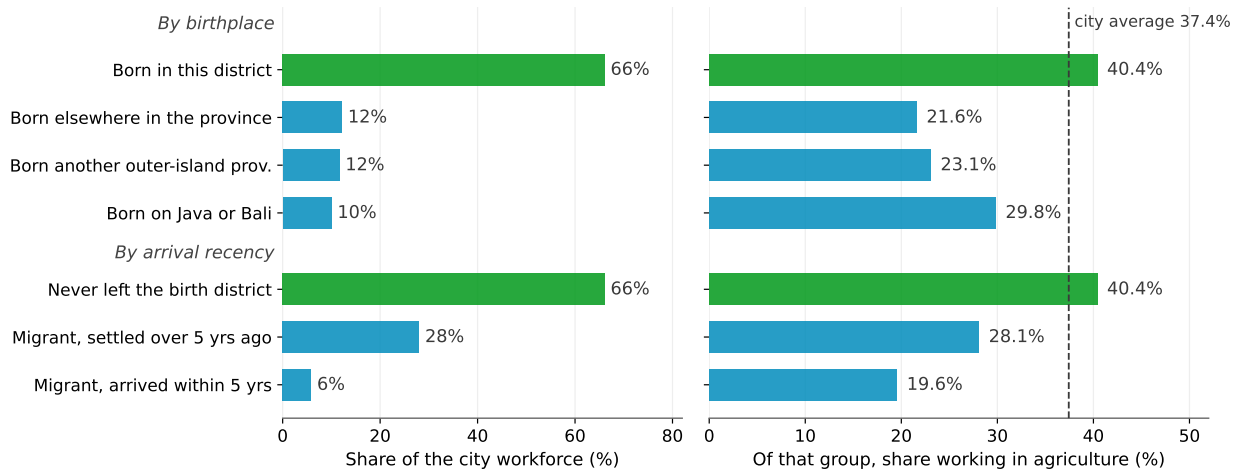
Dropping the 5 cities that draw part of their workforce from a subdistrict shared with a neighbor ([Appendix E.4](#)) gives a within share of 83.7% on the remaining 129 cities, against 83.7% on the full sample. [Table S15](#) reports every component under both partitions, and [Table S16](#) collects the responses, baselines, ceilings, and measured components in one place.

Figure S5: The Migration Ceilings and the Size and Agricultural Share of Each Origin Group

(a) The Migration Ceilings Against the Measured Composition Component



(b) Origin Groups: Share of the Workforce and Agricultural Share



Notes: Both panels use the main city sample and the 2010 population census; the estimates are regressions on standardized agro-similarity in the main specification with the full set of controls and Conley SE within 200km. Panel (a): the three ceilings ask how much of the estimated agricultural-share fall migration could produce if every arrival worked outside agriculture, so they are maxima under that assumption. Each bar is a percentage of the estimated fall in the agricultural employment share per standard deviation of agro-similarity, and the dashed line at 100% is the whole fall. The employment-growth ceiling multiplies the inflow implied by the estimated log employment response by the mean 2010 agricultural share; the ceiling at the top of the 90% range uses the upper end of that response's 90% confidence interval from the same standard errors; the birthplace-mix ceiling multiplies the estimated rise in the share of workers born outside the district by the agricultural share of locally born workers. The bottom bar is the measured composition component of Figure E5. Panel (b): unweighted means across cities of each origin group's share of the workforce (left) and of its own agricultural employment share (right), with the unweighted mean city agricultural share as the dashed line. The two blocks are two partitions of the same workforce, so the shares within each block sum to one. Every entry is a mean across cities, so the group shares and the group agricultural shares do not multiply out to the city average; the gap is the covariance across cities between a group's share of the workforce and its agricultural share, summed over the groups.

Table S15: Decomposition of the Fall in the Urban Agricultural Employment Share, by Birthplace and Arrival Recency

	By birthplace				By arrival recency		
	(1) Total	(2) Migrant share	(3) Within origin	(4) Interaction	(5) Migrant share	(6) Within origin	(7) Interaction
Agro-similarity	-0.0509*** (0.0184)	-0.0080** (0.0040)	-0.0425*** (0.0165)	-0.0003 (0.0055)	-0.0057 (0.0035)	-0.0445*** (0.0171)	-0.0007 (0.0044)
Settler count	✓	✓	✓	✓	✓	✓	✓
Baseline Pop 1980	✓	✓	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓	✓	✓
Total Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agri Empl 1980	✓	✓	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓	✓	✓
Observations	134	134	134	134	134	134	134

Notes: Each column is one regression on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Standard errors are in parentheses. Column 1 is each city's 2010 agricultural employment share in the 2010 population census. Columns 2 to 4 split it exactly over four birthplace groups (born in the district, elsewhere in the province, in another outer-island province, and on Java or Bali): Migrant share, the composition component, lets only the birthplace makeup of the workforce vary; Within origin, the within component, lets only each group's own agricultural share vary; Interaction is the remainder. Columns 5 to 7 repeat the split over arrival recency: workers who never left their birth district, migrants present five years before the census, and migrants who arrived within five years. Because the census is a repeated cross-section, the within-origin component includes cohort replacement. Agro-climatic controls are the 16 village covariates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S16: Migrant Composition as a Bound on the Agricultural-Share Fall

	Estimate	Share of the fall
<i>Panel A. Estimated responses per +1 SD of agro-similarity</i>		
Agricultural employment share	-0.0509*** (0.0184)	
Conley 200 km p-value	0.006	
Log employment	0.0235 (0.0570)	
Conley 200 km p-value	0.680	
Share of workers born outside the district	0.0354 (0.0257)	
Conley 200 km p-value	0.170	
<i>Panel B. Baseline agricultural shares, sample means</i>		
All workers, 2010	0.374	
All workers, 1980	0.560	
Workers born in the district, 2010	0.404	
Workers who arrived within 5 years, 2010	0.196	
<i>Panel C. Maximum composition effect, arrivals assumed entirely non-agricultural</i>		
Inflow based, 2010 baseline	0.0087	17.1%
at the 90% upper end	0.0414	81.4%
on the 1980 baseline	0.0130	25.6%
Composition based, 2010 baseline	0.0143	28.1%
at the 90% upper end	0.0314	61.8%
<i>Panel D. Measured, replacing the assumption with the observed arrival mix</i>		
Migrant share, by birthplace	-0.0080**	15.8%
Within origin, by birthplace	-0.0425***	83.7%
Migrant share, by arrival recency	-0.0057	11.1%
Within origin, by arrival recency	-0.0445***	87.5%

Notes: Panels A and D report regressions on standardized agro-similarity on the main city sample, in the main specification with the full set of controls and Conley SE within 200km. Standard errors are in parentheses, with the p-value beneath each, shaded by significance level. Workers are counted in the 2010 population census, and in the 1980 census for the rows marked 1980; Panel B gives sample means. Panel C bounds the composition effect by assuming every arrival works outside agriculture: the inflow-based bound multiplies the implied employment inflow, one minus the exponential of minus the log employment response, by the baseline agricultural share; the composition-based bound multiplies the estimated rise in the share of workers born outside the district by the agricultural share of locally born workers; the upper-end rows use the upper end of the response's 90% confidence interval; the larger bound, here the composition based one, is the ceiling. Panel D replaces the assumption with the arrivals' measured agricultural share, from [Table S15](#). Share of the fall divides each Panel C bound by the Panel A fall in the agricultural employment share, and each Panel D component by the total it decomposes. SD is one standard deviation of agro-similarity, and Log employment is the log of the city's 2010 employment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S5 Food Prices, Consumption Composition, and Trade Openness

S5.1 Food-Price Indices and Rice Unit Values

The 1996 food-price basket estimates are statistically insignificant under each weighting: the raw-food index is -0.012 (s.e. 0.025) with city-specific expenditure weights, -0.012 with pooled urban-center weights, and -0.013 equal-weighted, and the all-food index including prepared food is $+0.007$. The rice unit-value estimate is also statistically insignificant. Both the 1996 and the 2002 waves predate the 2004 rice-import ban, whose openness-differential price effects (Bazzi 2017) run through the 2011 wave, and the 1996 wave includes more of the main city sample than the 2002 wave (Table A6). Table S17 reports the full set of these food-price estimates.

Table S17: 1996 Food-Price Responses to Agro-Similarity

	Raw-ag Paasche	Raw-ag Laspeyres	Raw-ag eq-wt	All-food	Ln rice price
Agro-similarity	-0.012 (0.025)	-0.012 (0.024)	-0.013 (0.025)	0.007 (0.025)	-0.009 (0.029)
Settler count	✓	✓	✓	✓	✓
Base ln pop 1980	✓	✓	✓	✓	✓
Base ln empl 1980	✓	✓	✓	✓	✓
Base ln agri empl 1980	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓
Outcome mean	1.058	1.041	1.040	1.070	6.840
Outcome SD	0.130	0.115	0.135	0.130	0.136
N (cities)	112	112	112	112	112

Notes: Each column is a separate urban-center regression of a 1996 food price measure on standardized agro-similarity across the 112 urban centers of the main city sample with food prices in the 1996 national socioeconomic household survey (Susenas), in the main specification with the full set of controls and Conley SE within 200km. Check marks show the controls. Standard errors are in parentheses. Columns 1 to 3 are a price index of raw agricultural foods (cereals to spices): the median unit value of each item in the urban center relative to its median over the households of all urban centers, averaged with the spending weights of the urban center, with spending weights pooled over all urban centers, or as an equally weighted geometric mean. Column 4 is the same index over every food from cereals through prepared food, with the spending weights of the urban center; column 5 is the log rice unit value, rice spending over kilograms consumed. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S5.2 Price Changes on the Common-City Sample

Both the 1996 and the 2002 waves include the agricultural price index, so on the 102 urban centers that appear in both it can be differenced, and exposure can be asked whether it moved the price over the intervening period as well as where the price stood at a point in time. The estimates for the change and both cross-sections are statistically insignificant. Differencing removes a time-invariant city price component. The interval ends before the 2004 rice-import restrictions, and the index is a unit value, which moves with the quality and the composition of the basket as well as with the price of a fixed good (Table S18).

Table S18: Agricultural Price-Index Responses to Agro-Similarity in Levels and Changes, 1996 to 2002

	Log price 1996	Log price 2002	Change 1996 to 2002
Agro-similarity	-0.004 (0.013)	0.013 (0.011)	0.017 (0.011)
<i>p</i> -value	0.752	0.226	0.131
Settler count	✓	✓	✓
Base ln pop 1980	✓	✓	✓
Base ln empl 1980	✓	✓	✓
Urban share 1980	✓	✓	✓
Base ln agric empl 1980	✓	✓	✓
Agro-climatic controls	✓	✓	✓
Observations	102	102	102

Notes: Each column is one regression on standardized agro-similarity on the 102 urban centers of the main city sample with an agricultural price index in both the 1996 and 2002 household-survey (Susenas) waves, in the main specification with the full set of controls and Conley SE within 200km. Coefficients are in log points of one plus the index; standard errors are in parentheses, and the *p*-values come from the same inference. The outcome is the log of one plus the raw-food Paasche price index, in 1996 (column 1), in 2002 (column 2), and as the 2002 log minus the 1996 log (column 3); column 1 of [Table S17](#) gives the same index in levels on every city of the main city sample with a 1996 price. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S5.3 Rice Quality and Expenditure Composition

The 1996 module splits rice into a local and a high-quality variety, so the pooled unit value can be separated into a within-variety price effect and a shift of spending across varieties. Households shift toward the high-quality variety: its share of rice spending is 19.3 percentage points higher per standard deviation of exposure, from a mean of 26.8%, and the share of rice buyers who buy any of it rises with it, both statistically significant. The local-variety unit-value estimate is positive and marginally statistically significant; the high-quality and pooled estimates are statistically insignificant. In 1996, then, more households buy higher-quality rice with a statistically insignificant pooled unit-value response ([Table S19](#)).

Table S19: The 1996 Rice Unit Value: Within-Variety Price vs Quality Composition

	Ln rice (pooled)	Ln local rice	Ln high-qual rice	HQ spend share	HQ buyer share
Agro-similarity	-0.009 (0.029)	0.031* (0.017)	0.054 (0.069)	0.193*** (0.061)	0.182*** (0.063)
Settler count	✓	✓	✓	✓	✓
Base ln pop 1980	✓	✓	✓	✓	✓
Base ln empl 1980	✓	✓	✓	✓	✓
Base ln agri empl 1980	✓	✓	✓	✓	✓
Urban share 1980	✓	✓	✓	✓	✓
Agro-climatic controls	✓	✓	✓	✓	✓
Outcome mean	6.840	6.845	6.833	0.268	0.262
Outcome SD	0.136	0.130	0.150	0.297	0.289
N (cities)	112	97	49	112	112

Notes: Each column is one regression of its outcome on standardized agro-similarity, on the urban centers of the main city sample where the 1996 household survey (Susenas) measures it, in the main specification with the full set of controls and Conley SE within 200km; the last row gives the number of urban centers. Standard errors are in parentheses. Column 1 is the log unit value of rice (expenditure over quantity, in rupiah per kilogram), local and high-quality varieties pooled; columns 2 and 3 are the log unit values of local and of high-quality rice; column 4 is the high-quality share of rice expenditure and column 5 the share of rice-buying households that buy any high-quality rice, both fractions. Ln is the natural log and HQ is high-quality; the Base ln rows are 1980 baselines in logs (pop is population, empl employment, agri empl agricultural employment); SD is the standard deviation; and N (cities) is the number of urban centers in each regression. *** 1%, ** 5%, * 10%.

S5.4 Trade Openness by Location Type

The paper's figure of the response by location type split on trade openness ([Figure 18](#)) shows a star beside each gap between the more closed and the more open group of a location type whose p -value under the figure's Conley standard errors is below 10%. [Table S20](#) reports, for each of its rows, how many units of the location's sample the outcome covers, the estimation rows in each group, the response of the more open group less that of the more closed group with its standard error, and the p -value of that gap under the Bartlett kernel at the same cutoff and with clusters by province. Population, employment and the unemployment rate are from the 2010 census; the real wage, the food spending share and spending per person are from the 2011 wave of the household survey (Susenas).

Table S20: The Openness Split of the Response by Location Type: Coverage and the Test of the Difference

Outcome	Location	Measured of sample	Observations		Open minus closed	<i>p</i> -value of the difference		
			Closed	Open		Conley	Bartlett	Province
<i>Panel A. Population and employment</i>								
Ln population	Cities	134 / 134	66	68	−0.150 (0.103)	0.144	0.185	0.159
	No settlement	596 / 596	302	294	0.111 (0.101)	0.270	0.222	0.363
	With settlement	191 / 191	96	95	−0.234 (0.208)	0.261	0.200	0.237
Ln employment	Cities	134 / 134	66	68	−0.149 (0.116)	0.198	0.216	0.178
	No settlement	596 / 596	302	294	0.156* (0.092)	0.092	0.072	0.184
	With settlement	191 / 191	96	95	−0.305 (0.224)	0.173	0.122	0.171
Ln agricultural employment	Cities	134 / 134	66	68	0.078 (0.170)	0.645	0.611	0.632
	No settlement	596 / 596	302	294	0.091 (0.097)	0.351	0.342	0.389
	With settlement	191 / 191	96	95	−0.331* (0.196)	0.091	0.053	0.097
Ln modern employment	Cities	134 / 134	66	68	−0.249** (0.112)	0.027	0.033	0.061
	No settlement	596 / 596	302	294	0.117 (0.107)	0.274	0.273	0.345
	With settlement	191 / 191	96	95	−0.178 (0.211)	0.399	0.355	0.390
<i>Panel B. Wages, unemployment and spending</i>								
Ln real wage	Cities	134 / 134	66	68	0.073 (0.152)	0.633	0.695	0.709
	No settlement	574 / 596	292	282	−0.177 (0.108)	0.100	0.188	0.211
	With settlement	188 / 191	95	93	0.008 (0.222)	0.970	0.969	0.970
Unemployment rate	Cities	134 / 134	66	68	−0.240 (0.289)	0.406	0.376	0.460
	No settlement	596 / 596	302	294	−0.032 (0.210)	0.879	0.863	0.880
	With settlement	191 / 191	96	95	0.208 (0.144)	0.151	0.167	0.252
Food spending share	Cities	134 / 134	66	68	0.140 (0.161)	0.385	0.387	0.534
	No settlement	576 / 596	293	283	−0.217 (0.168)	0.197	0.163	0.299
	With settlement	188 / 191	95	93	0.569*** (0.209)	0.006	0.011	0.025
Ln spending per person	Cities	134 / 134	66	68	−0.040 (0.120)	0.739	0.801	0.856
	No settlement	576 / 596	293	283	0.274** (0.108)	0.011	0.017	0.031
	With settlement	188 / 191	95	93	−0.354** (0.158)	0.025	0.037	0.031

Notes: The rows of Figure 18: one regression per outcome and location type, on the units of that location's sample that report the outcome, with each location type split at the median route distance to the Thai and Vietnamese rice hubs. Open minus closed: the response of the more open group less that of the more closed group, in standard deviations of the outcome per standard deviation of agro-similarity, with its Conley SE within 200km (uniform kernel) beneath; its stars are the figure's. The *p*-values test that gap under those Conley SE, the Bartlett kernel at the same cutoff, and province clusters; a cell takes one shade at 0.10 or above and deepening shades below 0.10, 0.05 and 0.01. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

S6 Urban Emergence and Endline Sample Selection

The set of cities that defines the estimation sample uses the GHS Urban Center Database's fixed 2025 boundaries, so the timing defense in Section 4.3 requires knowing whether the shock itself created urban

units that a 2025-defined sample would then select. This appendix reports that test directly, estimating the effect of the same exposure on urban emergence outside the fixed sample.

S6.1 Lattice Sample and Historical Satellite Inputs

The unit is an equal-area 25 km lattice over outer-island land (2,648 cells, excluding Java, Bali, and Papua as in the main sample), with the pre-shock urban baseline taken from the GHS-SMOD 1975 classification, the published 1975 classification. The treatment is the paper’s exposure rebuilt at cell centroids, the unweighted mean of village agro-climatic similarity within 300 route-aware cost-minutes, with the agro-climatic control set and the transmigrant-count control re-derived from the same cell-to-village pairs and Conley standard errors at a 200-kilometer cutoff for the lattice regressions. The estimation sample is the baseline-rural, exposure-defined cells containing no transmigrant settlement point (program settlement points are excluded): 572 cells at 25 km. New-town founding is a rare event (single-digit foundings in the exposed sample by 2010), so the founding analyses use the finer 15 km lattice (2,074 cells), defining a founding as a nightlights cluster disjoint from the 1992 urban extent with a 5 km guard against an existing city’s expanding edge.

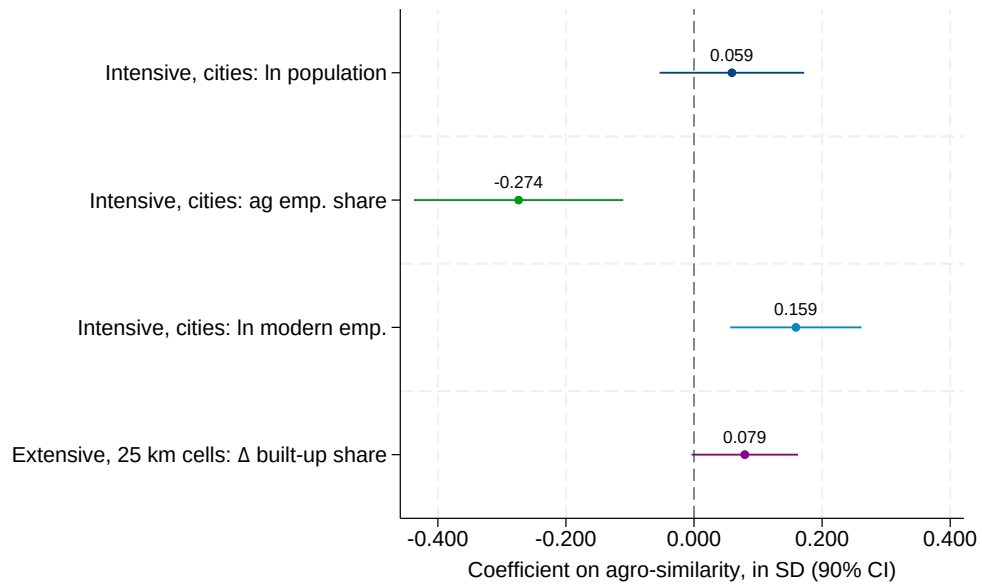
The early classification, built-up share, and gridded population depend partly on interpolated satellite inputs. The historical coverage limitations in [Section S2](#) therefore also apply to the sample definition and baseline controls here. The exercise does not establish the direction or size of the resulting bias.

S6.2 Built-Up Growth and Urban Founding

On one standardized axis, the extensive response is smaller than the intensive responses of the agricultural employment share and modern employment, and marginal ([Figure S6](#)). The intensive trio sits off zero on its two composition outcomes (ln population +0.059, agricultural employment share -0.274 , ln modern employment +0.159), while the extensive response, the 1975-2010 change in built-up areal share, is smaller and marginal at +0.079, its 90% interval crossing zero. [Table S21](#) reports the raw-units version: +0.00011 of built-up areal share per standard deviation of exposure over 1975-2010 (SE 0.00007), stable across the 1975-2000 and 1975-2020 windows. The estimate is sensitive to geographic controls. [Table S22](#) shows the same coefficient conventionally significant within islands (island fixed effects +0.00017, province fixed effects +0.00016, both $p < 0.01$). The additional tables vary the baseline and founding definitions. [Table S23](#) re-runs the headline under all five baseline-rural definitions and [Table S24](#) the founding event under all four founding definitions, positive throughout.

The nearby-exposure comparison examines spatial redistribution. [Table S25](#) adds the mean exposure of cells 30 to 100 km away: the own effect stays positive while the ring enters negative (founding -0.0034 , marginally significant), consistent with local displacement. The ring’s precision is sensitive to the band definition, so it is illustrative; the magnitude statements that make the displacement point are in checkable units. The count of founded towns per district does not rise with exposure ([Table S26](#)), and the population bridge in [Table S27](#) reports a positive but imprecise population association within

Figure S6: Intensive Versus Extensive Margin of the Agro-Similarity Shock



Notes: Each marker is a coefficient, printed above it, in SD of the outcome per SD of agro-similarity; bars are 90% confidence intervals from Conley SE within 200km. The first three rows are 2010 log population, agricultural employment share and log manufacturing and services employment on the main city sample (outer-island urban centers with a transmigration village within 300 travel minutes and a 1980 census baseline; 134 cities), in the main specification with the full set of controls. The fourth is the 1975-2010 change in the built-up share of land, from the Global Human Settlement Layer, across 572 square 25-kilometer cells of outer-island land outside Papua that were rural in that source's 1975 settlement map, lie within 300 travel minutes of a transmigration village and contain none; its controls are the cell's settler count and 16 agro-climatic covariates within 300 minutes, its 1975 built-up share, log 1975 gridded population and log land area.

incumbent footprints and a negative association within newly founded footprints ($-1,050$ per standard deviation, statistically significant).

The built-up, founding, and district-population comparisons use different outcomes and geographic units. Their standardized coefficients do not measure contributions to aggregate urban growth. Sparse founding events, imprecise population estimates, and the early-satellite limitations qualify the interpretation.

S6.3 Sensitivity to the City Inclusion Threshold

One source of selection is left over, and the lattice cannot test it: the estimation sample above is restricted to baseline-rural cells, which excludes precisely the incumbent towns that sat near the inclusion threshold. Membership in the urban-center sample requires 50,000 people in the 2025 population grid, a post-treatment quantity, so an incumbent town near that line enters only if it grew, and exposed marginal towns would select in while unexposed ones select out. Of the 134 sample cities, 18 have a 2025 polygon population within 20% of the threshold, and the sample's smallest is 50,287, close enough to the line that the concern is not hypothetical. Dropping all 18 and re-estimating the headline table does not move the population coefficient in the direction the concern predicts. The agricultural employment share coefficient goes from -0.051 to -0.044 ($p = 0.031$), modern-sector employment from

0.158 to 0.152 ($p = 0.038$), and log population from 0.051 to 0.082 ($p = 0.234$, against $p = 0.389$ on the full sample), all on the remaining 116 cities. The population estimate remains statistically insignificant without the near-threshold cities. This exclusion is a sensitivity check; it does not identify the direction of selection bias or exclude selection in the sectoral-share estimates. Two further sample cities have no GHS polygon population because the reclassification created them as new nodes instead of selecting them on the threshold, so they are not exposed to this selection and are retained.

Table S21: Headline Extensive Margin: Built-Up-Share Change on Agro-Similarity (Conley SE within 200km)

	(1) Δ built-up	(2) Δ built-up	(3) Δ built-up	(4) Δ built-up
TMV agro-similarity	0.00011 (0.00007)	0.00012 (0.00008)	0.00009* (0.00005)	0.00010 (0.00012)
N	572	572	572	906
Window	1975-2010	1975-2020	1975-2000	1975-2010 (NTL base)

Least-squares regressions of the change in the built-up share of each cell (Global Human Settlement Layer) from 1975 to the end year in the Window row on agro-similarity of the transmigration villages within 300 travel minutes, standardized over all cells with a measured exposure. The sample is the square 25-kilometer cells of outer-island land outside Papua with a transmigration village within 300 travel minutes and none inside the cell, and with no urban pixel in the 1975 settlement map of that source (columns 1 to 3) or no urban nighttime-lights pixel in 1992 (column 4). Controls: the settler count and 16 agro-climatic covariates within 300 minutes, and the 1975 built-up share, log 1975 population and log land area of the cell. Conley SE within 200km in parentheses. In the labels, Δ built-up is the change in the built-up share over the years in the Window row, and NTL base is the column 4 sample, rural by 1992 nighttime lights (NTL); TMV agro-similarity is the standardized agro-similarity of the transmigration villages (TMV), and N is the number of cells.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S22: Built-Up Share Changes With Alternative Geographic Controls

	(1) No FE	(2) Island FE	(3) Province FE
TMV agro-similarity	0.00011 (0.00007)	0.00017*** (0.00006)	0.00016*** (0.00006)
N	572	572	572
Fixed effects	None	Island	Province

Each column regresses the 1975 to 2010 change in the built-up share of each cell (Global Human Settlement Layer) on agro-similarity of the transmigration villages within 300 travel minutes, standardized over all cells with a measured exposure. The sample is the square 25-kilometer cells of outer-island land outside Papua with no urban pixel in the 1975 settlement map of that source, a transmigration village within 300 travel minutes and none inside the cell. Controls: the settler count and 16 agro-climatic covariates within 300 minutes, and the 1975 built-up share, log 1975 population and log land area of the cell. Column 2 adds island fixed effects and column 3 province fixed effects. Conley SE within 200km in parentheses. FE is fixed effects, TMV agro-similarity is the standardized agro-similarity of the transmigration villages (TMV), and N is the number of cells.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S23: Built-Up Share Changes Under Alternative Baseline Rural Definitions

	(1) SMOD ≥ 21	(2) SMOD ≥ 23	(3) SMOD ≥ 30	(4) Built-up 1975	(5) NTL 1992
TMV agro-similarity	0.00011 (0.00007)	0.00019* (0.00011)	0.00008 (0.00013)	0.00021 (0.00014)	0.00010 (0.00012)
<i>N</i>	572	727	883	927	906

Each column regresses the 1975 to 2010 change in the built-up share of each cell (Global Human Settlement Layer) on agro-similarity of the transmigration villages within 300 travel minutes, standardized over all cells with a measured exposure. The sample is the square 25-kilometer cells of outer-island land outside Papua that count as rural at baseline and have a transmigration village within 300 travel minutes and none inside the cell. The rural rule differs across columns, and a stricter urban threshold admits more cells: no pixel of the 1975 settlement map of that source at class 21 or above (any urban class) in column 1, at 23 or above (dense urban cluster) in column 2, or at 30 (the urban-center class) in column 3; a 1975 built-up share below 3% in column 4; and no urban nighttime-lights pixel in 1992 in column 5. Controls: the settler count and 16 agro-climatic covariates within 300 minutes, and the 1975 built-up share, log 1975 population and log land area of the cell. Conley SE within 200km in parentheses. In the column labels, SMOD is the settlement map of the Global Human Settlement Layer, the number after it is the class threshold, and NTL is nighttime lights; TMV agro-similarity is the standardized agro-similarity of the transmigration villages (TMV), and N is the number of cells.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S24: New Urban Settlement Under Alternative Definitions

	(1) New cluster 2010	(2) Any urban 2010	(3) New cluster 2020	(4) Any urban 2020
TMV agro-similarity	0.00180 (0.00132)	0.00086 (0.00108)	0.00190 (0.00120)	0.00140 (0.00142)
<i>N</i>	2074	2074	2074	2074
Events	6	8	7	12

Each column is a linear probability model across square 15-kilometer cells of outer-island land outside Papua with no urban pixel in the 1975 Global Human Settlement Layer settlement map, a transmigration village within 300 travel minutes and none inside the cell. In the new-cluster columns the outcome is one when the cell overlaps an urban nighttime-lights cluster of the year in the column heading that does not touch the 1992 urban extent and has its center at least 5 kilometers from it; in the any-urban columns it is one when the cell holds any urban nighttime-lights pixel in that year. The Events row counts sample cells with an outcome of one. The regressor is agro-similarity of the transmigration villages within 300 travel minutes, standardized over all cells with a measured exposure. Controls: the settler count and 16 agro-climatic covariates within 300 minutes, and the 1975 built-up share, log 1975 population and log land area of the cell. Conley SE within 200km in parentheses. TMV agro-similarity is the standardized agro-similarity of the transmigration villages (TMV), and N is the number of cells.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S25: Urban Settlement and Exposure in Surrounding Areas

	(1) New town	(2) Δ urban share
TMV agro-similarity	0.0047*** (0.0015)	0.0001*** (0.0000)
Ring term (30-100 km)	-0.0034* (0.0021)	-0.0001 (0.0001)
<i>N</i>	2074	2074

Each column is a regression across the 2,074 cells, on a 15-kilometer lattice over the outer islands, that had no urban settlement in the 1975 Global Human Settlement Layer, lie within 300 travel minutes of a transmigration village and do not contain one. Column 1 is 1 when the cell overlaps a new town, a 2010 night-light urban cluster disjoint from the 1992 urban extent with its center at least 5 kilometers from it; column 2 is the 1992 to 2010 change in the night-light urban share of the land area of the cell. The regressors are the agro-similarity of the cell, the mean agro-climatic similarity of the transmigration villages within 300 travel minutes, and the ring term, the mean agro-similarity of the cells with a defined exposure whose centers lie 30 to 100 kilometers away; both are standardized over the lattice cells with a defined exposure, and a negative ring coefficient means that exposure nearby suppresses founding. Controls are the settler count and 16 agro-climatic covariates within 300 minutes, and the 1975 built-up share, log 1975 population and log land area of the cell. Conley SE within 200km in parentheses. TMV agro-similarity is the standardized agro-similarity of the transmigration villages (TMV), Δ urban share is the 1992 to 2010 change in the night-light urban share, and *N* is the number of cells.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S26: District-Level New-Cluster Count on Agro-Similarity, with the Log of At-Risk Cells as Offset

	(1) Poisson	(2) Poisson + islFE
Agro-similarity (district)	0.3807 (0.4275)	0.7321 (0.5929)
<i>N</i>	129	129
FE	none	island

Each column is a Poisson regression, across the 129 outer-island districts with at least one at-risk cell, of the number of new towns whose centers fall in the at-risk cells of the district on the district mean of standardized cell agro-similarity, with the log number of at-risk cells as offset; column 2 adds island-group fixed effects (Sumatra, Kalimantan, Sulawesi and Maluku). Standard errors clustered by province in parentheses. At-risk cells are 15-kilometer lattice cells with no urban settlement in the 1975 Global Human Settlement Layer that lie within 300 travel minutes of a transmigration village and do not contain one. A new town is a 2010 night-light urban cluster outside the 1992 urban extent whose center lies at least 5 kilometers from it. Cell agro-similarity is the mean agro-climatic similarity of the transmigration villages within 300 travel minutes of the cell, standardized over the lattice cells with a defined exposure. Agro-similarity (district) is the district mean of standardized cell agro-similarity, FE is fixed effects, islFE is island-group fixed effects, and *N* is the number of districts.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S27: Population Bridge: Persons Added per SD of Agro-Similarity, by Footprint Type

	(1)	(2)	(3)	(4)
	d pers incumbent	d pers new town	incumb 90-20	new 90-20
TMV agro-similarity	1,346 (1,223)	-1,050** (438)	2,269 (2,338)	-1,465** (620)
<i>N</i>	150	150	150	150
Footprint	incumbent	new town	incumbent 90-20	new town 90-20

Each column regresses the change in population, in persons, inside one footprint type of a district on district agro-similarity, standardized over the districts, across the 150 outer-island districts with a 15-kilometer lattice cell within 300 travel minutes of a transmigration village. Population is summed from the Global Human Settlement Layer population grid inside incumbent footprints, the 1992 night-light urban extent (columns 1 and 3), or new-town footprints, the 2020 night-light urban clusters outside that extent with centers at least 5 kilometers from it (columns 2 and 4); the change runs from 1990 to 2010 in columns 1 and 2 and from 1990 to 2020 in columns 3 and 4. District agro-similarity averages, over the lattice cells of the district, the mean agro-climatic similarity of the transmigration villages within 300 travel minutes of each cell. Standard errors clustered by province in parentheses. In the labels, d pers is the change in population in persons, incumb and new are the incumbent and new-town footprints, 90-20 is the change from 1990 to 2020, TMV agro-similarity is the standardized district agro-similarity of the transmigration villages (TMV), and N is the number of districts.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

S7 Service Flows and Travel-Time Gradients

The two-tradable-sector variant has one parameter that no moment in the estimation disciplines: δ_S , the distance decay on the catchment-traded bundle. This appendix measures it directly.

Service trade costs in the model are $\tau_{oj}^M = \max(d_{oj}, 1)^{\delta_S}$, entering CES demand as $\tau^{1-\sigma}$ with $\sigma = 4$, so bilateral service expenditure falls in travel time as $d^{-(\sigma-1)\delta_S}$. A gravity regression of service flows on log travel time therefore recovers $-(\sigma - 1)\delta_S$ and nothing finer: σ and δ_S are not separately identified from flows, but their product is, and it is the product that governs how far a city’s service market reaches.

S7.1 Survey Coverage and Located Service Records

The Indonesia Family Life Survey records the provider’s location and the respondent’s residence for outpatient use in the four weeks before interview. Pooling the 2007 and 2014 waves gives 9,275 records with both locations identified. Of these, 1,925 link different year-2000 subdistricts, across 1,074 positive provider and residence pairs. The descriptive shares in [Table S28](#) instead use the survey’s same-subdistrict indicator and its survey-year boundaries. Estimating

$$F_{oj} = \exp(\gamma_o + \delta_j + \beta \ln d_{oj})$$

by Poisson pseudo-maximum likelihood with origin and destination fixed effects, zeros retained, and standard errors two-way clustered on origin and destination, on the route-aware cost-distance surface in minutes, gives $\hat{\beta} = -3.272$ with a standard error of 0.100, and hence

$$\hat{\delta}_S = -\hat{\beta}/(\sigma - 1) = 1.091 \quad (0.033),$$

with a 90% interval of [1.036, 1.146].

The estimate is similar when flows are weighted by out-of-pocket expenditure (1.131), to excluding traditional practitioners (1.133), to keeping only formal providers (1.117), and to excluding the public health centers that Indonesia sites in every subdistrict by administrative design (1.072). Restricting the sample to the outer islands, which is the estimation footprint of the paper, gives 0.948.

S7.2 Comparisons Across Services

Outpatient visits are one service, and medical care is atypical in being urgent and in having a publicly sited provider network. The survey records the same provider-location block for two further services, providing further comparisons. The additional services are schooling and hospitalization. [Table S28](#) estimates all three identically.

The provider-location shares differ across services, while the estimated distance gradients have overlapping confidence intervals. The share of records with a provider outside the home subdistrict more than doubles from routine medical care to hospitalization, while the estimated decay moves by 0.11 and

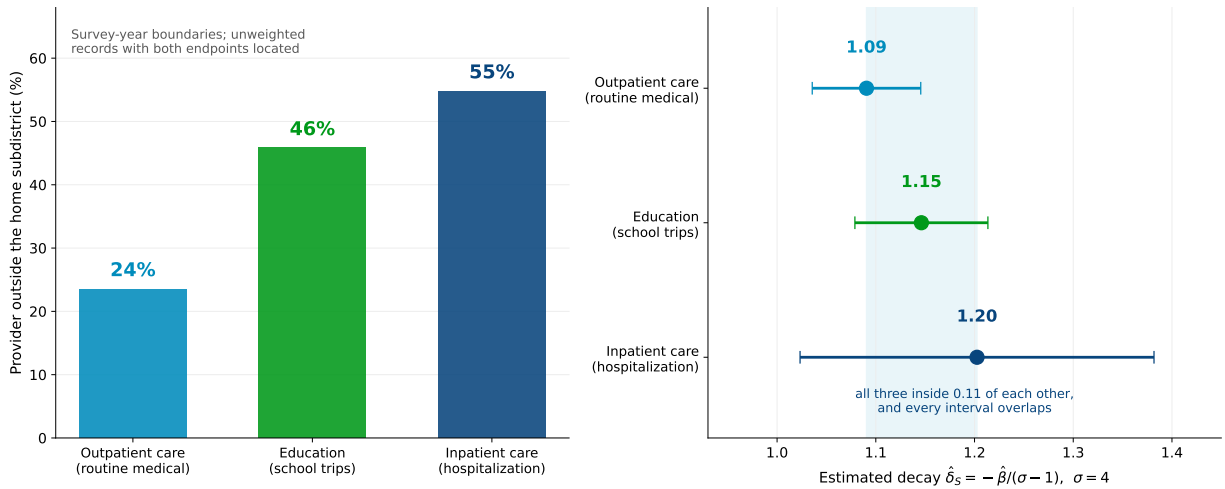
Table S28: Service Locations and Travel-Time Gradients

Service	Provider outside home subdistrict	$\hat{\beta}$	s.e.	$\hat{\delta}_S$	90% interval
Outpatient care (routine medical)	23.5%	-3.272	0.100	1.091	[1.036, 1.146]
Education (school trips)	45.8%	-3.439	0.123	1.146	[1.079, 1.214]
Inpatient care (hospitalization)	54.8%	-3.608	0.327	1.203	[1.023, 1.382]

Notes: The descriptive shares are unweighted, using survey-year subdistrict boundaries among records with both residence and provider located: 9,275 outpatient, 4,719 education, and 699 inpatient records, including own-subdistrict observations. Provider home-visit records are retained, so provider location does not always represent household travel. The gravity regressions use year-2000 boundaries and the product of observed provider and residence locations, retain bilateral zero flows, and exclude own-subdistrict pairs and locations with no observed cross-subdistrict flow. Poisson pseudo-maximum likelihood with origin and destination fixed effects, zeros retained, standard errors two-way clustered on origin and destination, own-subdistrict pairs excluded, estimated on the model's route-aware cost distance in minutes. $\hat{\delta}_S = -\hat{\beta}/(\sigma - 1)$ at $\sigma = 4$. Outpatient care and education pool the 2007 and 2014 waves; inpatient care uses 2007 only, because the 2014 module records the province and subdistrict of the hospital but not the district. Education covers respondents enrolled at their highest level, which in the located sample is predominantly senior secondary and tertiary.

every interval overlaps every other, with a Cochran Q of 1.75 on two degrees of freedom against a 5% critical value of 5.99. These are different objects: the descriptive share records the provider location relative to the residence, while the regression estimates a conditional travel-time gradient. Figure S7 separates the two.

Figure S7: Service Locations and Estimated Travel-Time Gradients



Notes: Panels use the samples of Table S28. The left panel reports providers outside the respondent's home subdistrict under survey-year boundaries; the right panel reports estimated distance decay under year-2000 boundaries. Intervals are 90%.

The recorded provider location does not always describe a journey by the household: records marked as provider home visits are retained. The services covered here also exclude retail and restaurant purchases. The survey covers thirteen provinces, so the gradient is estimated on that geographic coverage and extrapolation beyond it requires an additional assumption.

The calibrated value. The measured value is 1.091 and the calibrated model is solved at $\delta_S = 1.0$. The two differ because the traded-service block does not exist at the higher value: the homotopy that switches the block on converges up to a traded-service expenditure share of 7.8% at $\delta_S = 1.09$ and fails before reaching the 8.2% that the expenditure data require, and the same failure occurs at 1.15 and 1.25. The model is therefore solved at the highest decay below the estimate at which the economy exists, and the gap of 0.09 is reported rather than rounded away. No average of the three services in [Table S28](#) is used. The estimates are not independent draws, sharing households, geography and one cost surface, and every pooled value lies further outside the solvable region than the outpatient estimate does.

S8 Urban-Center-Level Empirical Facts and Model Coverage

Each urban center-level empirical fact maps to an object of the model, and the estimate behind it comes from the main text wherever a main-text table reports it. The population, employment, and sectoral-composition block (F1-F6) is reported here from the main-text [Table 4](#), which consolidates the same census regressions that [Appendix Section E.1](#) reports in a disaggregated layout. Most of the remaining Susenas- and PODES-based rows are likewise reported in the main text alongside the results they support; only the rows without a main-text table of their own (the 1996 household-survey table behind F9, the agricultural-goods outcomes behind F10, and the service-sector composition behind F20) point to [Section E.1](#), where their signs and significance are verified on the same sample. Sample sizes below 134 in those tables follow the narrower Susenas and PODES coverage of the same sample. [Table S29](#) lists the facts and their mapping to the model.

The structural model is calibrated to match the exact standardized (z -on- z) point estimates, reported together with the untargeted over-identifying moments ([Appendix Table B6](#)).

Table S29: Empirical Facts and Model Coverage

#	Fact	Coef (SE)	Source
F1	City population: positive, imprecise	0.051 (0.059)	Table 4 col 1
F2	City employment: positive, imprecise	0.024 (0.057)	Table 4 col 2
F3	Ag employment share ↓	−0.051 (0.018)***	Table 4 col 5
F4	Ln ag employment ↓	−0.166 (0.093)*	Table 4 col 4
F5	Ln modern employment ↑ (serv+manuf)	0.158 (0.062)**	Table 4 col 6
F6	Modern employment share ↑	0.066 (0.015)***	Table 4 col 7
F8	Non-food spending share ↑	Positive, sig.	Table E2 col 1
F9	Housing rent: 1996 positive, imprecise; 2011 flat	Pos., ns (1996); ≈ 0, ns (2011)	Appendix Table E7; Table E2 col 2
F10	Rice surplus ↓, imprecise; own production ↑	−0.141 (1.211); 0.735***	Appendix Table E6 col 5; Table E3 col 1
F12	Non-ag wage: ns	≈ 0, ns	Table E2 col 3
F15	Food spending share ↓	Negative, sig.	Appendix Table B6 row m_7
F17	Village retains ag workers	Positive	Figure 14
F19	Migration from the same province ↑; other outer-isl provinces ns; Java/Bali ↓	+0.042 (0.018), sig. at 5%; +0.018 (0.021), not sig.; −0.018 (0.011), sig. at 10%	Appendix Table C6 Panel B, col 1
F20	Service growth concentrated in transport ↑	Positive, sig.	Table S13 col 3

Notes: Each row is one empirical fact of the paper and the model object it maps to. The coefficient column reports the estimate on standardized agro-similarity from the table and column the Source column names, with the standard error in parentheses and stars as in those tables; rows F1 to F6 read the headline table’s own macros, so a rebuild of that table moves them with it, and a row without a numeric entry states the sign and whether the estimate is statistically significant. F19 reports the three birthplace groups in order: the same province, other outer-island provinces, and Java and Bali.

S9 Theories of Structural Transformation

Two classical accounts explain what higher agricultural productivity does to the rest of the economy, and they make opposite predictions. In the first, cheaper food releases both spending and labor to the modern sector, and the economy transforms. In the second, the same productivity gain strengthens the region’s comparative advantage in farming and pulls labor toward the land. In both cases, the positive productivity shock is within the same economy where outcomes are measured. A city that buys its food from the villages that received the gain is a small economy whose trading partner became more productive at farming, a case Matsuyama (1992) does not state; under his open-economy condition, the buyer’s manufacturing share rises on impact. This appendix sets out both, states the condition that selects between them, and shows what the setups behind them contain. It proceeds in four steps: the closed-economy prediction and the two assumptions it turns on, the open-economy reversal and why the original models treat it as a switch, what the classical setups leave out that this design requires, and the frameworks that added locations, together with the three rival explanations of the same shock that the theory names.

S9.1 The closed-economy prediction

The prediction that a better harvest transforms a nearby city descends from the dual-economy literature and rests on two assumptions. The first is Engel's law: the income elasticity of food demand is below one, so a household that feeds itself more cheaply moves its marginal spending to everything else. The second is that the food market clears at home, so that the extra food has nowhere to go but into local consumption.

[Lewis \(1954\)](#) states the price consequence of a better harvest. His two sectors exchange food against capitalist goods at an internal terms of trade, and because food demand is inelastic, a rise in the subsistence sector's productivity reaches the capitalist sector as a fall in the price of food and a rise in profit. In his model the transfer of labor runs through capitalist accumulation at a subsistence wage. The food-problem tradition states the second. In [Gollin et al. \(2002\)](#) households buy a subsistence quantity of food before valuing anything else, so labor stays on farms until the economy's food requirement is met and low agricultural productivity postpones industrialization ([Gollin et al. 2007](#)); the constraint appears already in [Schultz \(1953\)](#).

[Matsuyama \(1992\)](#) states the mechanism at its sharpest and shows how little it needs, and the subsistence preferences he used, the Stone-Geary form the model in this paper also uses, became the standard non-homotheticity of the structural-transformation literature. Labor is the only factor, preferences include a subsistence food requirement, and the food market clears at home. Food consumption then equals food output, and that single condition is sufficient to pin the manufacturing labor share as an increasing function of agricultural productivity. A better harvest frees workers for the modern sector because the extra food has nowhere else to go. [Glaeser \(2014\)](#) extends the argument to urbanization in poorer countries.

The prediction is about the employment composition, and it is worth separating from the scale prediction because the two are often stated together. Nothing in Matsuyama's closed-economy mechanism requires the city to grow. In Glaeser's closed economy, urbanization increases with farm productivity (his Proposition 1): his urban sector produces traded manufacturing goods, so a better harvest raises the share of workers in the city. Cheaper food raises what a household has left after feeding itself and tilts its marginal spending toward the modern good, and it lowers the return to farming inside the city relative to modern work; both forces move the employment composition without anyone changing location. The classical account that does require the city to grow is the labor one, and it is a different claim.

S9.2 The open-economy reversal

The same model with both goods trading at world prices predicts the opposite. [Matsuyama \(1992\)](#) shows that when the food constraint is lifted, comparative advantage allocates labor instead, and an economy that becomes better at farming specializes in farming: the productivity gain pulls workers into agriculture and slows the transition out of it. The empirical literature has largely worked inside that case. [Bustos et al. \(2016\)](#) model each Brazilian region as a small open economy in which international

prices pin agriculture’s marginal product, so the direction of the local labor response is set by the factor bias of the technical change.

The reversal is a switch in these models because openness is a property of the whole country: both regimes are stated for an economy with one location, so there is no internal geography over which openness could vary. [Matsuyama \(2008\)](#) restates the caveat in the same terms. That is the feature the design changes, and the following subsection takes it up.

S9.3 Limits of the classical setups

I compare the twelve models this paper is positioned against on three facts: whether the setup contains a food market, whether it contains trade of any other kind, and whether it contains locations at a distance from each other. Locations are absent from every classical setup and arrive only with the quantitative spatial frameworks of the last decade, and that absence is why the theory has to be reinterpreted at the regional level before it can be used here. [Table S30](#) records the three facts for each model.

Their absence has a consequence that is easy to miss, because the models do use place names. [Harris & Todaro \(1970\)](#) call one sector rural and the other urban, but the agricultural good sells at a single economy-wide terms of trade and the two sectors sit at no distance from each other. [Gollin et al. \(2016\)](#) do the same. A model with no distance between its locations has one food market by construction, so its food constraint binds for the country as a unit, and it cannot ask whether a particular city can be fed by a particular hinterland. The classical prediction is therefore about national agricultural productivity and national industrialization, and applying it at the level of a city and the villages that feed it changes the model.

I apply that regional interpretation to Indonesia’s outer islands and report the travel-time gradients of district food-price gaps in the main text ([Section 4.1](#)).

S9.4 Spatial frameworks and their open questions

Five recent frameworks contain everything the classical setups leave out, and this paper differs from them in the source of the identifying variation. [Eckert & Peters \(2025\)](#) have discrete locations, costly trade in food, costly migration, and nonhomothetic demand; their frictions are calibrated, their productivities inverted from the static equilibrium, and their question is the incidence of secular structural change. [Nagy \(2023\)](#) is closest in spirit, with farm goods shipping from hinterlands to the cities they feed and city formation endogenous. His cities hold both farm and non-farm population, and he does not report a within-city agricultural response. [Tombe & Zhu \(2019\)](#) is the closest quantitative precedent for separating goods-market from labor-market integration, run on calibrated frictions and aggregate productivity. [Pellegrina & Sotelo \(2025\)](#) formalize the first stage of this paper’s own design, farmers moving with crop-specific knowledge, and follow it into specialization and trade, leaving the urban system outside the model. [Donaldson \(2018\)](#) and [Fajgelbaum & Redding \(2022\)](#) establish how far shipping costs shape where production locates.

Two further theories predict the same direction of reallocation for reasons unrelated to the food market, and the openness heterogeneity separates them. In the price-effect family, structural change follows from sectors growing at different rates when the goods substitute poorly (Ngai & Pissarides 2007; Acemoglu & Guerrieri 2008; Herrendorf et al. 2014); an agricultural productivity gain then moves labor out of agriculture with no reference to who eats the output or where it is sold. In the income-effect family, to which the mechanism here belongs, the reallocation runs through nonhomothetic demand (Kongsamut et al. 2001; Boppart 2014; Comin et al. 2021). Both predict a fall in the agricultural employment share of an exposed city. A city’s connection to outside food markets does not enter the price-effect mechanisms, so the openness contrast separates them from the food-trade account.

S9.5 Three rival explanations of the same shock

The theory also names three ways a rural productivity gain can reach a city, and the paper’s claim is about which one operates.

Labor. The classical account sends workers. A productivity gain means the same output can be grown with less labor, relative returns to farming fall, and workers leave farming (Cao & Birchenall 2013; Bustos et al. 2016); rainfall-driven productivity gains raise local demand and move rural Indian labor into local non-farm work (Emerick 2018). The release is largest in the villages that received the largest gain, which is a prediction about the villages and is testable there.

Capital. Bustos et al. (2020) follow the agricultural surplus out of the region it was earned in, as bank deposits that fund lending elsewhere, so agricultural productivity growth finances industrialization at a distance at a speed set by financial integration. Theirs is the one predecessor that crossed the boundary of the region that received the shock, and it did so through neither goods nor people.

Factor bias. Bustos et al. (2016) show that the direction of the local labor response depends on whether the technical change saves labor or augments land. A land-augmenting gain pulls labor into agriculture in the place that receives it, at a fixed outside price; the food market never enters. That is the same village-side pattern the food-trade account predicts for a different reason, so the factor bias of the shock is a genuine alternative explanation of the village side.

S9.6 Requirements for an empirical test

These mechanisms imply different responses in employment, population, and food prices. Within cities, the food expenditure share falls as spending per person rises (Section 5.2). The city’s composition moves out of agriculture and into the modern sector (Section 5.1). Modern employment grows by more than total employment under the fixed-incumbent-workforce accounting (Section E.3). Net employment growth does not measure gross migration when arrivals replace departing workers or coincide with changes in labor-force participation. The estimate for the city’s rice production less consumption is negative and statistically insignificant (Section 5.2). The village agricultural employment share has a positive, statistically insignificant gradient with agro-similarity. I find no evidence in the measured settler stay-share that settlers leave differentially by productivity. The measure does not follow the

original settlers individually or include their locally born children (Section 5.2 and Section 5.3). The 2011 non-agricultural wage, overall wage and real wage estimates are statistically insignificant (Section 5.2). And the modern-sector response is concentrated in cities weakly connected to outside food markets and damped in cities that can already import cheaply, while the agricultural employment share falls in both groups (Section 5.2). The openness comparison supports the food-market mechanism if the other mechanisms respond similarly in the two groups of cities after controls. Migration, amenities, and rural demand could also vary with trade access.

Table S30: What Each Setup Contains, and Where Its Food Constraint Binds

Model	Food market	Other trade	Locations	Where the food constraint binds, and what the setup leaves open
<i>Panel A. The classical setups: no locations</i>				
Lewis (1954)	•	•	◦	The economy as one unit; the terms of trade between food and capitalist goods move once, nationally. Later cases trade only between whole countries
Harris & Todaro (1970)	•	◦	◦	The country. Rural and urban are place names with no distance and no trade cost, and one national relative price of food. The urban sector has no agriculture, so the transformation can only arrive as migrants
Matsuyama (1992)	•	•	◦	The country when closed, nowhere when open. Openness is a national switch, and with one location there is no migration to compete with the goods channel
Gollin et al. (2002)	•	◦	◦	The country as one unit, which is frictionless internal trade by assumption. The subsistence requirement is the mechanism and cannot be localized
Gollin et al. (2016)	•	•	◦	The country; resource rents are earned at world prices. Their urban sector has no agriculture, so the within-city response is zero by construction
Bustos et al. (2016)	•	•	◦	Nowhere: each region is a small open economy and food sells at the world price whatever its neighbors produce. Distance does not appear, and factor bias sets the sign
Bustos et al. (2020)	◦	•	◦	Nowhere in the model; the geography lives in the bank-branch data. Names the capital channel that the goods and labor accounts have to be separated from
<i>Panel B. The frameworks with locations</i>				
Eckert & Peters (2025)	•	•	•	The location, with food shipped at iceberg costs. Frictions are calibrated and productivities inverted from the static equilibrium, and the focus is the incidence of secular change
Tombe & Zhu (2019)	◦	•	•	The region, through costly goods trade and costly migration. The closest precedent for separating the two integration costs, run on calibrated frictions and aggregate productivity
Pellegrina & Sotelo (2025)	◦	•	•	The region, for agricultural goods. Formalizes the knowledge transfer behind this paper's first stage; the outcomes are specialization and trade, with no urban side
Nagy (2023)	•	•	•	The city and the hinterland that ships to it, which is this paper's unit. His cities hold both farm and non-farm population, and he does not report a within-city agricultural response
Tombe (2015)	•	•	•	The country, with trade costs largest in agriculture. The international statement of the localization argument; the units are whole countries

Notes: • marks an object the setup contains and ◦ one it does not. “Food market” is a market in which the price of food is determined; “other trade” is trade in any other good, including trade with the rest of the world; “locations” means places at a measured distance from each other, so that a shipping or moving cost can be defined. Place names without distance, as in the rural and urban sectors of Harris & Todaro (1970), are not locations in this sense. The setups in Panel A are transcribed from the papers’ own texts; the two Tombe rows and Tombe & Zhu (2019) are stated at the level of their published abstracts. Ordering within each panel is chronological.

S10 Employment Comparisons on the Observed-Price Samples

These tables use the same outcomes, samples, exposure scale, openness groups, and controls as the observed-price comparisons in the paper. Bartlett distance weights provide a sensitivity to the main spatial inference.

Table S31: Employment Responses on the Observed-Price Samples

	Agricultural Share	Log Modern Jobs	Modern Share
<i>Main City Sample and 2011 Prices</i>			
Agro-similarity	-5.086*** (1.861)	0.158*** (0.059)	6.595*** (1.556)
Cities	134	134	134
<i>Cities With Observed Prices in 1996</i>			
Agro-similarity	-3.854 (2.667)	0.071 (0.084)	5.726*** (2.159)
Cities	112	112	112
<i>Cities With Observed Prices in 2002</i>			
Agro-similarity	-2.157 (2.566)	0.097 (0.080)	3.865* (2.230)
Cities	109	109	109

Notes: The first panel uses the main city sample, which is also the observed-price sample in 2011. The other panels retain main-sample cities with the raw-food price index observed in the indicated wave. Exposure and outcome scales remain fixed. These are coverage sensitivities; differences across samples do not identify an effect of the import restrictions. Agro-similarity is standardized over the full city file before any sample restriction. All regressions include the settler count within 300 travel minutes, the 1980 population, employment, urban population share and agricultural employment, and sixteen agro-climatic controls. Agricultural and modern employment shares are in percentage points; modern employment is $\ln(\text{count}+1)$. Modern employment combines manufacturing and services. Sources: population censuses, 1980 and 2010; household survey (Susenas), 1996, 2002 and 2011. Conley standard errors with Bartlett distance weights within 200km are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S32: Employment Responses by Openness on the Observed-Price Samples

	Agricultural Share	Log Modern Jobs	Modern Share
<i>Main City Sample and 2011 Prices</i>			
More closed	-7.069*** (2.569)	0.307*** (0.110)	9.788*** (2.068)
More open	-2.387 (1.914)	0.065 (0.069)	3.529** (1.697)
More open minus more closed	4.683* (2.621)	-0.242** (0.114)	-6.259*** (2.163)
Cities	134	134	134
More-closed	66	66	66
More-open	68	68	68
<i>Cities With Observed Prices in 1996</i>			
More closed	-7.323*** (2.568)	0.233** (0.097)	9.575*** (2.210)
More open	-0.815 (3.543)	-0.091 (0.100)	2.253 (2.801)
More open minus more closed	6.508* (3.361)	-0.324** (0.134)	-7.321*** (2.756)
Cities	112	112	112
More-closed	57	57	57
More-open	55	55	55
<i>Cities With Observed Prices in 2002</i>			
More closed	-4.198 (2.584)	0.197** (0.090)	7.545*** (2.044)
More open	-0.362 (2.956)	0.014 (0.094)	0.818 (2.395)
More open minus more closed	3.836 (2.972)	-0.183* (0.109)	-6.727*** (2.218)
Cities	109	109	109
More-closed	55	55	55
More-open	54	54	54

Notes: The first panel uses the main city sample, which is also the observed-price sample in 2011. The other panels retain main-sample cities with the raw-food price index observed in the indicated wave. Exposure and outcome scales remain fixed. These are coverage sensitivities; differences across samples do not identify an effect of the import restrictions. The more open cities have route hours below the median among all cities within 300 travel minutes of a settlement; this division is fixed before restricting to measured prices. Both group slopes are estimated in one regression with a group indicator and common controls. The difference is more open minus more closed, using their covariance. Agro-similarity is standardized over the full city file before any sample restriction. All regressions include the settler count within 300 travel minutes, the 1980 population, employment, urban population share and agricultural employment, and sixteen agro-climatic controls. Agricultural and modern employment shares are in percentage points; modern employment is $\ln(\text{count}+1)$. Modern employment combines manufacturing and services. Sources: population censuses, 1980 and 2010; household survey (Susenas), 1996, 2002 and 2011. Conley standard errors with Bartlett distance weights within 200km are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

- Acemoglu, D. & Guerrieri, V. (2008). Capital Deepening and Nonbalanced Economic Growth. *Journal of Political Economy*, 116(3), 467–498.
- Bazzi, S. (2017). Wealth Heterogeneity and the Income Elasticity of Migration. *American Economic Journal: Applied Economics*, 9(2), 219–255.
- Bazzi, S., Gaduh, A., Rothenberg, A. D., & Wong, M. (2016). Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia. *American Economic Review*, 106(9), 2658–2698.
- Boppart, T. (2014). Structural Change and the Kaldor Facts in a Growth Model with Relative Price Effects and Non-Gorman Preferences. *Econometrica*, 82(6), 2167–2196.
- Bustos, P., Caprettini, B., & Ponticelli, J. (2016). Agricultural Productivity and Structural Transformation: Evidence from Brazil. *American Economic Review*, 106(6), 1320–1365.
- Bustos, P., Garber, G., & Ponticelli, J. (2020). Capital Accumulation and Structural Transformation. *Quarterly Journal of Economics*, 135(2), 1037–1094.
- Cao, K. H. & Birchenall, J. A. (2013). Agricultural Productivity, Structural Change, and Economic Growth in Post-Reform China. *Journal of Development Economics*, 104, 165–180.
- Civelli, A., Gaduh, A., Rothenberg, A. D., & Wang, Y. (2023). Urban Sprawl and Social Capital: Evidence from Indonesian Cities. *Economic Journal*, 133(654), 2110–2146.
- Comin, D., Lashkari, D., & Mestieri, M. (2021). Structural Change with Long-Run Income and Price Effects. *Econometrica*, 89(1), 311–374.
- Donaldson, D. (2018). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *American Economic Review*, 108(4-5), 899–934.
- Eckert, F. & Peters, M. (2025). *Spatial Structural Change*. NBER Working Paper 30489, National Bureau of Economic Research. September 2022, revised July 2025. Under revision at the Review of Economic Studies.
- Emerick, K. (2018). Agricultural Productivity and the Sectoral Reallocation of Labor in Rural India. *Journal of Development Economics*, 135, 488–503.
- European Commission, Joint Research Centre (2024). *GHSI Data Package 2023*. Technical Report JRC133256, Publications Office of the European Union, Luxembourg.
- Fajgelbaum, P. & Redding, S. J. (2022). Trade, Structural Transformation, and Development: Evidence from Argentina 1869–1914. *Journal of Political Economy*, 130(5), 1249–1318.

- Glaeser, E. L. (2014). A World of Cities: The Causes and Consequences of Urbanization in Poorer Countries. *Journal of the European Economic Association*, 12(5), 1154–1199.
- Gollin, D., Jedwab, R., & Vollrath, D. (2016). Urbanization with and without Industrialization. *Journal of Economic Growth*, 21(1), 35–70.
- Gollin, D., Parente, S., & Rogerson, R. (2002). The Role of Agriculture in Development. *American Economic Review*, 92(2), 160–164.
- Gollin, D., Parente, S. L., & Rogerson, R. (2007). The Food Problem and the Evolution of International Income Levels. *Journal of Monetary Economics*, 54(4), 1230–1255.
- Harris, J. R. & Todaro, M. P. (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *American Economic Review*, 60(1), 126–142.
- Herrendorf, B., Rogerson, R., & Valentinyi, Á. (2014). Growth and Structural Transformation. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of Economic Growth*, volume 2 (pp. 855–941). Amsterdam: Elsevier.
- Kongsamut, P., Rebelo, S., & Xie, D. (2001). Beyond Balanced Growth. *Review of Economic Studies*, 68(4), 869–882.
- Lewis, W. A. (1954). Economic Development with Unlimited Supplies of Labour. *Manchester School*, 22(2), 139–191.
- Matsuyama, K. (1992). Agricultural Productivity, Comparative Advantage, and Economic Growth. *Journal of Economic Theory*, 58(2), 317–334.
- Matsuyama, K. (2008). Structural Change. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan, 2nd edition.
- Nagy, D. K. (2023). Hinterlands, City Formation and Growth: Evidence from the U.S. Westward Expansion. *Review of Economic Studies*, 90(6), 3238–3281.
- Ngai, L. R. & Pissarides, C. A. (2007). Structural Change in a Multisector Model of Growth. *American Economic Review*, 97(1), 429–443.
- Pellegrina, H. S. & Sotelo, S. (2025). Migration, Specialization, and Trade: Evidence from Brazil’s March to the West. *Journal of Political Economy*, 133(12), 3993–4049.
- Schultz, T. W. (1953). *The Economic Organization of Agriculture*. New York: McGraw-Hill.
- Tombe, T. (2015). The Missing Food Problem: Trade, Agriculture, and International Productivity Differences. *American Economic Journal: Macroeconomics*, 7(3), 226–258.

Tombe, T. & Zhu, X. (2019). Trade, Migration, and Productivity: A Quantitative Analysis of China. *American Economic Review*, 109(5), 1843–1872.