NOT FOR PUBLICATION: WEB APPENDICES

The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa

by Remi Jedwab and Alexander Moradi

NOT FOR PUBLICATION

DATA DESCRIPTION, VARIABLE CONSTRUCTION AND SAMPLE SELECTION SELECTION FOR GHANA AND AFRICA

This appendix describes in details the data we use in our analysis.

Spatial Units for Ghana:

We assemble data for 2,091 grid cells of 0.1x0.1 degrees (about 11x11 km) from 1891 to 2000. We choose a high resolution grid because we have precise GIS data on railroads and population. Each cell has the same size, except the cells that are coastal or crossed by a border. We create dummies equal to one if the cell is coastal or bordering another country to control for this issue. The 2,091 cells belong to 35 ethnic groups using the Murdock (1959) map that was recreated by Nunn (2011), or 110 districts (2000).

Railway Data for Ghana:

We obtain the layout of railway lines in GIS from *Digital Chart of the World*. We use Gould (1960), Dickson (1968), Tsey (1986), and Luntinen (1996) to reconstruct the history of rail construction. For each line, we know when it was started and finished, and when each station was opened. From the same sources, we know the lines that were planned but not built. Most of those placebo lines became roads later. Using the GIS road network also available from *Digital Chart of the World*, we construct those placebo lines in GIS. We calculate for each cell the Euclidean distance (km) from the cell centroid to each real or placebo line. Lastly, we create a set of cell dummies equal to one if the cell centroid is less than X km away from the line: 0-10, 10-20, 20-30 and 30-40 km. We create a cell dummy equal to one if the cell contains a rail station in 1918. We also know how many tons of cocoa were brought to each station in 1918.[1] Data on railway traffic was obtained from various sources.[2] Lastly, using GIS we construct instruments as the Euclidean distance (km) from each cell centroid to the straight lines Sekondi-Tarkwa-Obuasi-Kumasi and Accra-Kumasi.

Commercial Agriculture Data for Ghana:

A very precise map of cocoa production in 1927 was obtained from the *1927 Yearbook of the Gold Coast* and digitized. This map displays dots for each 100 tons of cocoa production.[3] We then use GIS to reconstruct total cocoa production (tons) for each grid cell in 1927. Cocoa production was nil in 1901. We obtain the international and producer prices from Bateman (1965).

Population and Urban Data for Ghana:

We collect population data from the gazetteers of the *Population and Housing Censuses* 1891, 1901, 1931, 1948, 1960, 1970, 1984, and 2000. They list localities and their population size. Defining as a city any locality with more than 1,000 inhabitants, we obtain a geospatialized sample of 3,057

[1]This information was retrieved from the *Administration Report of Gold Coast Railways 1921*.

[2]These sources are: *Government Gazette Supplement of Accra* 1914, *Administration Report of Gold Coast Railways* 1920, 1921, 1929-30 and 1931-32, Gould (1960), Tsey (1986) and Luntinen (1996).

[3]Aggregating all the dots, we obtain 209,100 tons of cocoa production in 1927, which is very close to the national estimate of 210,600 tons (see Gunnarsson 1978).

different cities for all these years. We used the GeoNet data base to retrieve the geographical coordinates of each city. Using GIS, we recalculate total urban population for each grid cell. We are then able to recreate rural population for each gridcell in 1901, 1931, 1970 and 2000. For the year 1901, from the census gazetteer, we know the population size of each village (locality with less than 1,000 inhabitants). It was impossible to find the geographical coordinates of all of them. Yet, the 1901 census was exhaustively conducted and geospatialized in the South of Ghana (N = 756 cells). We know for each cell the number of *large towns*, *towns* (more than 500 inhabitants), *head chief towns*, *large villages* (100-500 inhabitants) and *villages* (less than 100 inhabitants). Using GIS, we can deduce for each cell the number of villages that are less than 100 inhabitants, the number of villages that have between 100 and 500 inhabitants and the number of villages that have between 500 and 1,000 inhabitants. From the census, we know the average settlement size for each category and we can reconstruct total rural population for each cell in 1901. For the year 1931, we digitised the population map in the 1960 *Ghana Population Atlas*. This map displays at a very fine spatial level settlements that have less than 500 inhabitants and settlements that have between 500 and 1,000 inhabitants. From the census, we know the average settlement size for each category, and we can reconstruct total rural population for each cell in 1931. For the years 1970 and 2000, we use the 2000 Facility Census which reports the population of all localities in 1970 and 2000 and their respective enumeration area. To derive rural population, we used a GIS map of enumeration areas, selected localities below 1,000 inhabitants and calculated population densities by enumeration areas. We then calculated rural population of grid cells in 1970 and 2000 by areal weighting. The boundaries of the enumeration areas were significantly modified between 1960 and 1970, and we do not have a GIS map of them in 1960. Thus, we cannot estimate rural population in 1960, and use rural population in 1970 instead. Another issue is that some locations have a total or urban population of 0 for the early years. Some cells were historically sparsely inhabited tropical forest. Besides, not all cells have a locality above 1,000 inhabitants. When using logs, we add one inhabitant to each cell to ensure they are not dropped from the analysis.

Mining Production Data for Ghana:

Various sources indicate the location of each mine (gold, manganese, and diamonds) in 1901 and 1931. We create a dummy equal to one if the cell contains a mine.[4] The same sources report the value of mineral production (in shillings) for each mine in 1931. We use GIS to reconstruct the standard score of the total value of mineral production for each cell in 1931.

Geographical and Economic Geography Data for Ghana:

Forest data comes from GIS data compiled by Globcover (2009). The data displays those areas with virgin forest or mixed virgin forest/croplands, which were areas with virgin forest before they were cleared for cash crop production. Soil data comes from the *1958 Survey of Ghana Classification Map of Cocoa Soils for Southern Ghana*. The map was digitized, and using GIS we calculated for each cell the share of land that is suitable for cocoa cultivation. We know the respective shares of land which consists of ochrosols (first class, second class, third class, unsuitable), oxysols, and intergrades. We also use the *1971 Survey of Ghana Soil Suitability Map for Export Crops in Ghana*. We know the share of land that is generally unsuitable for cocoa exports. A cell is defined as suitable

[4]Mining data is collected from the following documents: *The Mineral Industry of the British Empire and Foreign Countries 1913-1919*; *Reports of the Mines Department of the Gold Coast* 1931-1958.

for cocoa production if it contains cocoa soils suitable for cocoa exports. It is then highly suitable if more than 50% of its area consists of forest ochrosols, the best cocoa soils. It is very highly suitable if more than 50% of its area consists of class 1 and class 2 ochrosols. Although the map is for 1958, it is a good indication of cocoa suitability in 1901, as soil types are time-invariant. Climate data comes from *Terrestrial Air Temperature and Precipitation: 1900-2007 Gridded Monthly Time Series, Version 1.01*, 2007, University of Delaware. We estimate average annual precipitations (mms) in 1900-1960 for each cell. Topography comes from SRTM3 data. We estimate for each cell the mean and standard deviation of altitude (meters). For each cell, we use GIS to get the Euclidean distances (km) to Accra, the capital city, Kumasi, the largest hinterland city, Aburi, the city where cocoa production originated, a port in 1901 (see Dickson 1968), a navigable river (see Dickson 1968) and the coast.

Other Transportation Networks Data for Ghana:

Transportation networks in 1901 are obtained from Gould (1960) and Dickson (1968), and *Colonial Annual Reports of the Gold Coast* 1903, 1904 and 1907.[5] We use various sources to construct a GIS database of roads in 1901 and 1931: Gould (1960) and *Map of The Gold Coast with Togoland Under British Mandate*, published in 1930. Those maps have a consistent legend showing class 1 roads ("*roads suitable for motor traffic throughout the year*"), class 2 roads ("*roads suitable for motor traffic but occasionally closed*"), and class 3 roads ("*roads suitable for motor traffic in dry season only*"). Other roads are not suitable for motor traffic and are not considered here. We use Michelin paper maps to create the 1965 and 1998 road networks in GIS (which we use as proxies for 1960 and 2000 respectively), distinguishing paved (bitumenized), improved (laterite), and earthen roads. The map of trade routes in 1850 is obtained from Dickson (1968).

Non-Transportation Infrastructure Data for Ghana:

We have data on government and non-government schools (missions) in 1902 (which we use as an approximation for 1901) and 1930-31. The data was compiled from education reports: *Report on the Education Department of the Gold Coast* for the years 1902 and 1930-31. They list all the schools in the country, which we then geocoded. We also have data on hospitals in 1902 and 1930 (which we use as approximations for 1901 and 1931). The data was compiled from health reports: *Report on the Medical and Sanitary Department of the Gold Coast* for the years 1902 and 1929-30. Data on the locations in churches in 1902 and 1929-30 comes from the *Ecclesiastical Returns* of the *Blue Books* of the same years (we use these years as approximations for 1901 and 1931). Besides, we use the 2000 *Facility Census* and the 2000 *Population and Housing Census* to obtain data on infrastructure provision and human capital at the cell level in 2000: shares (%) of inhabitants living less than 5 km from a primary school, junior secondary school (JSS), senior secondary school (SSS), health clinic, hospital, post office, or telephone, shares of inhabitants with access to clean water, in a residence with solid walls, a solid roof, or floor, and shares of adults aged 25 and over that are literate, have ever been to school, and have finished primary school, JSS, or SSS. Since we only have data for 10% of the population census, the most rural cells of our sample do not have enough observations to correctly estimate these shares. Data is missing for them.

[5] We use these sources, Maxwell (1923) and Tsey (1986) to obtain the trade cost associated with each transportation technology in 1901. Cardinall (1932) is our main source for production costs.

Employment and Economic Development Data for Ghana:

Data on urban employment in 1931 comes from the *Population and Housing Census* 1931. The source of the satellite data on night lights is NOOA (2012). We follow the approach of Henderson, Storeygard and Weil (2012) and estimate average light intensity for each cell, for the year 2000-01. Per capita GDP data for the whole country in 1891-1931 comes from Jerven (2011). Data on the value of total exports and cocoa exports in 1891-1931 (constant 1995 USD) comes from Teal (2002). Data on Ghana's total population in 1891-1931 comes from census figures adjusted by Austin (2007).

Spatial Units for Africa:

We assemble data for 194,000 grid cells of 0.1x0.1 degrees (11x11 km) in 39 Sub-Saharan African countries from 1960 to 2000. The 194,000 cells belong to 755 ethnic groups using the Murdock (1959) map recreated by Nunn (2011), or 2,304 districts (2000). Districts are the third-level administrative units for each country, below the "province" or "region" (the second-level administrative unit). The GIS files of districts for each country come from Global GIS (2007).

Railway Data for Sub-Saharan Africa:

We obtain the layout of rail lines in GIS from *Digital Chart of the World*. We use Chaléard, Chanson-Jabeur & Béranger (2006), Wikipedia (2013) and various historical studies available on the internet to recreate the history of each line. We know when each line was built, and the main motivations behind its construction: (i) "military domination" if Wikipedia or the study mentions the fact that a colonial power built the line to mark its territory given the scramble for Africa, or the fact that a colonial power built the line to be able to rapidly dispatch troops in order better control the native population, (ii) "mining" if Wikipedia or the study mentions the fact that a colonial power built the line to connect a mine to the coast or the already existing railroad network, (iii) "cash crops" if Wikipedia or the study mentions the fact that a colonial power built the line to connect agriculturally rich area to the coast or the already existing railroad network, and (iv) "trade" if Wikipedia or the study do not mention any specific motivation for the line except the fact that a colonial power wants to boost trade in general. A same line can obviously be built for several motivations. There are 66,491 km of railway lines in 2000, but 58,716 km were built before independence (57,872 km if we use the year 1960 instead). From Metcalfe (1916) and NGS (1922), we know the lines that were planned but not built. Most of those placebo lines became roads later. Using the GIS road network also available from *Digital Chart of the World*, we recreate those placebo lines in GIS. We calculate for each cell the Euclidean distance (km) from the cell centroid to each real or placebo line. Lastly, we create a set of cell dummies equal to one if the cell centroid is less than X km away from the line: 0-10, 10-20, 20-30 and 30-40 km. Then, we construct the instrument by using GIS to obtain the Euclidean distance (km) from each cell centroid to the Euclidean Minimum Spanning Tree Network based on the initial urban network of the 39 countries in 1900 (the localities above 10,000 inhabitants in 1900, as well as the capital city, the largest city and the second largest city of each country at independence). Lastly, data on colonial budgets in French West Africa and Kenya was compiled by Huillery (2012) and Burgess et al. (2013).

Urban and Satellite Data on Night Lights for Sub-Saharan Africa:

We obtain from various sources a geospatialized database of localities above 10,000 inhabitants for 39 Sub-Saharan African countries in 1890, 1900, 1960, 1970, 1980, 1990, 2000 and 2010. We exclude from the analysis four small island countries and four Southern African countries for which urban data could not be obtained. Web Appendix Table 11 lists the main sources used for each country. For the years 1890 and 1900, we use census reports and colonial handbooks, Wikipedia (2013) and the *Encyclopedia Britannica* 1901 and 1911. For the period 1960-2010, Africapolis (2010) reports the data for 15 West African countries, while Africapolis (2012) reports the data for 18 Central and East African countries. Data for the remaining 6 countries was recreated using the reports of their *Population Censuses*, various official documents listing administrative counts, several demographic studies, *Google Earth* and Wikipedia (2013), for the closest years to 1960, 1970, 1980, 1990, 2000 and 2010: Botswana, Malawi, Mozambique, Namibia, Zambia and Zimbabwe.[6] After obtaining a list of all localities above 10,000 inhabitants, we use GeoNet and Google Earth to retrieve their geographical coordinates (and verify that there are indeed cities today). Using GIS, we estimate total urban population for each grid cell. The source of the satellite data on night lights is NOOA (2012). We follow the approach of Henderson, Storeygard and Weil (2012) and estimate average light intensity for each cell for the year 2010.

Controlling Variables for Sub-Saharan Africa:

Using GIS, we estimate the Euclidean distances (km) from the cell centroid to the coast or a navigable river. We use Michelin paper maps to create the 1998 road networks in GIS (which we use as a proxy for 2000), distinguishing paved (bitumenized) and improved (laterite) roads. We create dummies that take the value of one if the cell is within 10 km from the coast, a navigable river, a paved road or an improved road in 2000. We create dummies indicating if the cell contains the largest city, the second largest city or the capital city (which is not always the largest or second largest city) of the country at independence. We compute the Euclidean distances (km) to the largest city, the second largest city or the capital city. Climate data comes from *Terrestrial Air Temperature and Precipitation: 1900-2007 Gridded Monthly Time Series, Version 1.01*, 2007, University of Delaware. We estimate for each cell average annual precipitations (mms) in 1900-1960. Topography comes from SRTM3 data. We estimate for each cell the mean and standard deviation of altitude (m). We use FAO (2011) to obtain for each cell the shares of of class 1 (the most suitable), class 2, class 3, undetermined, sparsely vegetated and submerged soils. Lastly, we use SNL Metals & Mining (2014) to locate the mines that served as a node for a railroad line built before 1960. We selected all properties that qualified as a mine (Mine Type="Mine"). We included active and inactive mines (Mine Status="open" or "closed") but excluded prospective mines (Mine Status="Feasibility Study", "Exploration" or "Construction"). We only included mines within a 10 km distance to a railroad line. In the end, there are 60 cells with a mine along a colonial railroad line.

APPENDIX REFERENCES

Africapolis. 2010. *Africapolis I: Western Africa*. Paris: SEDET.

[6]We have enough sources to obtain a precise or approximative estimate of the population of each city in each decade. Data was then linearly interpolated or extrapolated across the years used in our analysis.

Africapolis. 2012. *Africapolis II: Central and Eastern Africa*. Paris: SEDET.

Austin, Gareth. 2007. "Labour and Land in Ghana, 1874-1939: A Shifting Ratio and an Institutional Revolution." *Australian Economic History Review* 47(1): 95-120.

Bairoch, Paul. 1988. *Cities and Economic Development: Frow the Dawn of History to the Present*. Chicago: The University of Chicago Press.

Bateman, Merril. 1965. *Cocoa in the Ghanaian Economy*. Ph.D. dissertation, MIT.

Burgess, Robin, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padro-i Miguel. 2013. "The Value of Democracy: Evidence from Road Building in Kenya." Unpublished manuscript, Department of Economics, London School of Economics.

Cardinall, Allan W. 1932. *The Gold Coast, 1931: a Review of Conditions in the Gold Coast in 1931 as Compared with Those of 1921*. Government Printer: Accra.

Chaléard, Jean-Louis, Chantal Chanson-Jabeur, and Chantal Béranger. 2006. *Le chemin de fer en Afrique*. Paris: Karthala Editions.

Dickson, Kwamina. 1968. *A Historical Geography of Ghana.* Cambridge University Press.

FAO. 2001. *Permanent Crop and Arable Land of Sub-Saharan Africa*. Rome: Food and Agricultural Organization.

Frankema, Ewout, and Marlous van Waijenburg. 2011. "Structural Impediments to African Growth? New Evidence from Real Wages in British Africa, 1880-1965." Working Papers 0024, Utrecht University, Centre for Global Economic History.

Global GIS. 2007. *Global GIS*. Reston, Virgina: U.S. Geological Survey.

Globcover. 2009. *Globcover Regional: Africa*. Louvain: European Space Agency Ionia.

Gould, Peter R. 1960. *The Development of the Transportation Pattern in Ghana.* Chicago: Northwestern University Press.

Gunnarsson, Christer. 1978. *The Gold Coast Cocoa Industry 1900-1939: Production, Prices and Structural Change*. Lund: Published Thesis.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102(2): 994-1028.

Huillery, Elise. 2012. "The Black Man's Burden: The Cost of Colonization of French West Africa." Unpublished manuscript, Department of Economics, Sciences Po.

Jerven, Morten. 2011. "Comparing Colonial and Post-Colonial Output: Challenges in Estimating African Economic Change in the Long Run." Working Papers 0010, Utrecht University, Centre for Global Economic History.

Luntinen, Pertti. 1996. *Railway on the Gold Coast: A Meeting of Two Cultures : a Colonial History*. Helsinki: Suomalainen. Tiedeakatemia.

Malanima, Paolo, and Oliver Volckart. 2007. Urbanisation 1700-1780. Chapter prepared for the Third RTN/CEPR Summer Symposium in London: An Economic History of Modern Europe.

Manning, Patrick. 2010. "African Population Estimates, 1850-1960: Appendices." In *The Demographics of Empire: The Colonial Order and The Creation of Knowledge in Africa*, ed. Karl Ittmann, Dennis D. Cordell and Gregory H. Maddox, Ohio University Press.

Maxwell, John. 1923. *Gold Coast Handbook 1923*. Accra: Government Printer.

Metcalfe, Charles. 1916. "Railway Development of Africa, Present and Future." *The Geographical Journal* 47(1): 3-17.

Murdock, George. 1959. *Africa: Its Peoples and their Culture*. McGraw-Hill Inc., US.

NGS. 1922. *Map of Africa and Adjoining Portions of Europe and Asia*. Washington D.C.: National Geographic Society.

NOAA. 2012. "Average Visible, Stable Lights, & Cloud Free Coverages." Version 4 DMSP-OLS Nighttime Lights Time Series. Silver Spring: National Oceanic and Atmospheric Administration

Nunn. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review*, 101 (7): 3221-3252.

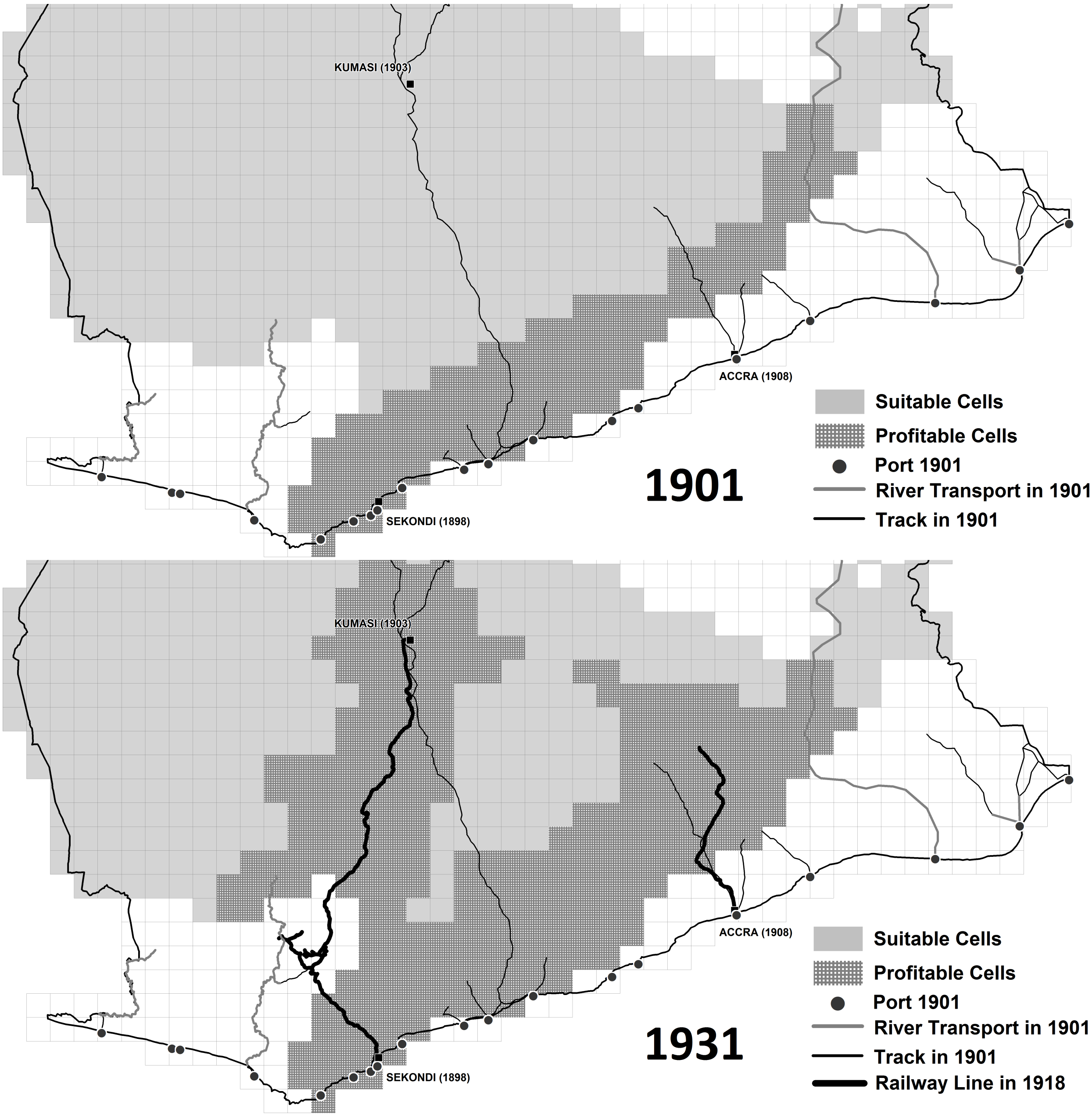
SNL Metals & Mining. 2014. "Mining Property Briefing Books" available online at: http://www.snl.com/Sectors/MetalsMining/MiningProjectBriefingBooks.aspx.

Teal, Francis. 2002. "Export Growth and Trade Policy in Ghana in the Twentieth Century." *The World Economy* 25(9): 1319-1337.

Tsey, Christian E. 1986. *Gold Coast Railways: the Making of a Colonial Economy, 1879-1929*. PhD University of Glasgow, Glasgow.

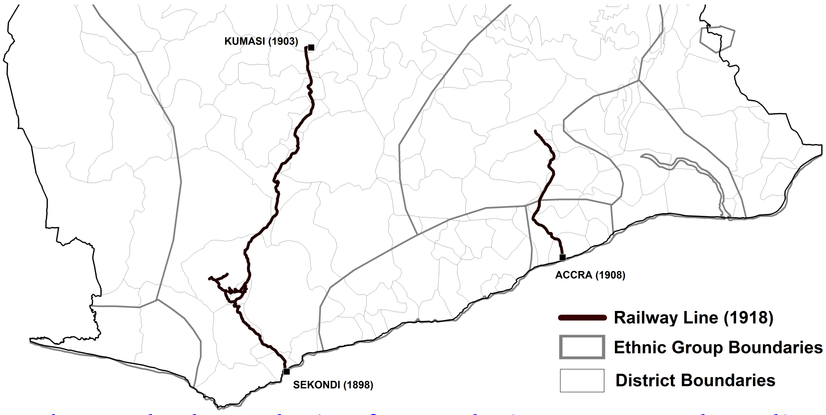
Wikipedia. 2013. *Historical of Rail Transport by Country* and *Lists of Cities by Country*. San Francisco, CA: California.

Web Appendix Figure 1: Transportation Technologies and Area of Profitable Production, 1901-1931



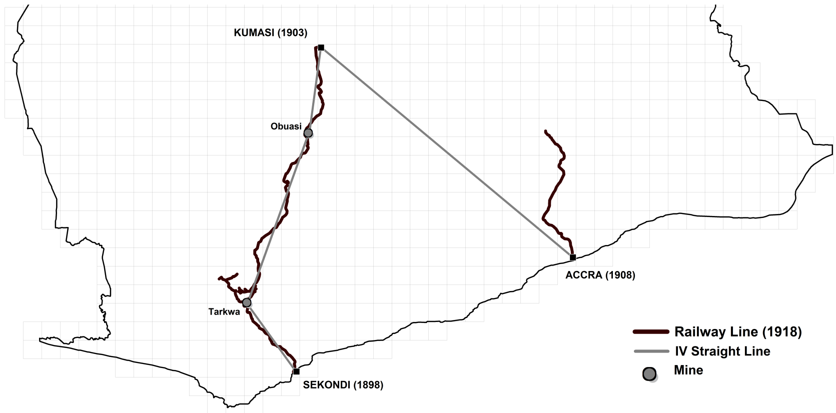
***Notes*: This map shows the cells for which cocoa cultivation would have been profitable in 1927 given the existing transportation methods in 1901: (i) without railroads ("1901" on the map), and (ii) with railroads ("1931" on the map).** 1901: The map shows transportation networks in 1901 (tracks, rivers and ports), the cells suitable for cultivation, and the cells for which cultivation is profitable without the railroad lines. We estimate the following model for each cell (per ton produced): Profit (1920s) = Port price (1920s) - production cost (1920s) - transportation cost (1901). We consider the 1920s, as production in 1927 was determined by economic conditions five years before (1922). We estimated transportation costs by calculating the least cost path from each cell to any trading port in 1901. We used Arcgis "path distance" tool. The cost raster was based on the following costs: head porterage on tracks (5 shillings, but adding a penalty for slopes), head porterage through the forest (8s), cask rolling for the few tracks on which casks could be rolled (1.9s), and canoe or steam launch for the cells along a river (1s-3.5s, depending on the river considered). 1931: The map shows the cells for which production is profitable in 1927 thanks to the 1918 railroads (cost of 0.5s). See *Web Appendix* for data sources.

Web Appendix Figure 2: Ethnic Group and District Boundaries



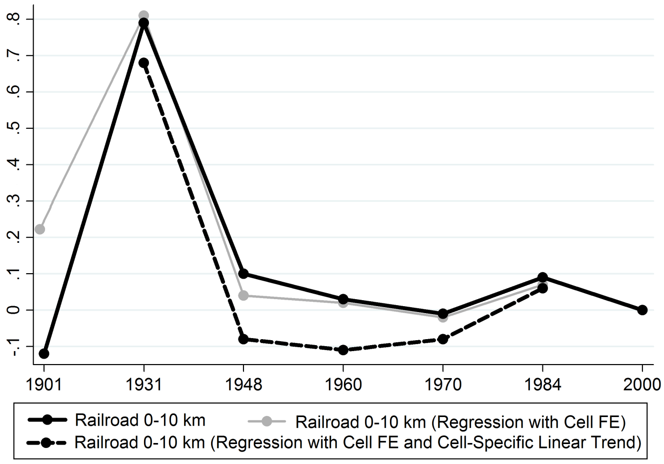
***Notes*: This map shows the boundaries for 9 ethnic groups and 62 districts in Ghana.** The districts are the administrative unit below the province. See *Web Appendix* for data sources.

Web Appendix Figure 3: IV Straight Lines



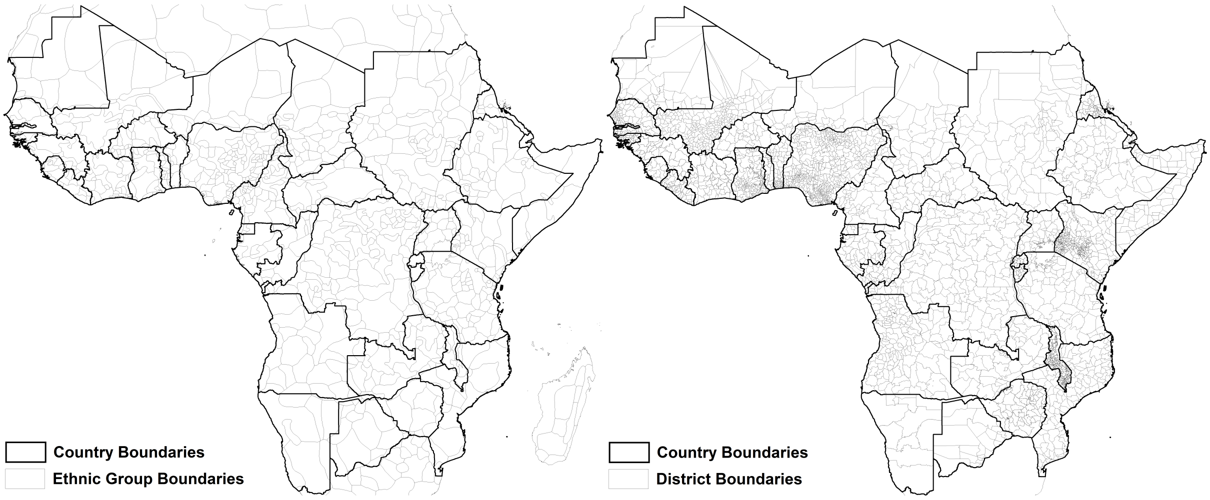
***Notes*: This map shows the IV straight lines Sekondi-Tarkwa-Obuasi-Kumasi and Accra-Kumasi.** For the IV straight line of the *Western Line*, we connect Sekondi, Tarkwa, Obuasi and Kumasi using the shortest Euclidean distance (km) between the four nodes. For the IV straight line of the *Eastern Line*, we connect Accra and Kumasi using the shortest Euclidean distance (km) between the two nodes. We drop the five nodes in the IV regressions. See *Web Appendix* for data sources.

Web Appendix Figure 4: Railroad Effects for Each Period, 1891-2000



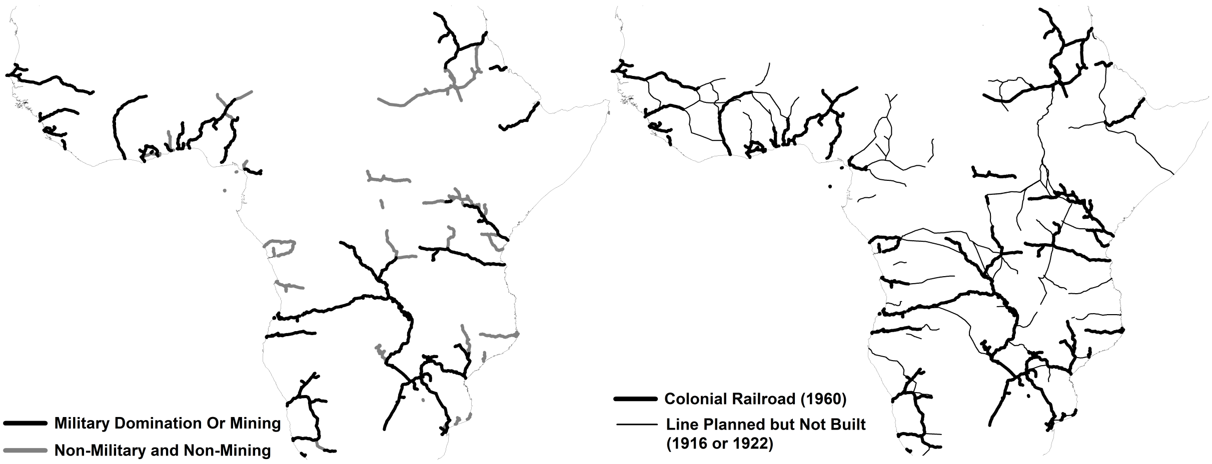
***Notes*: This graph shows that the estimates of the railroad effect (for 0-10 km) for each period in 1891-2000 are unchanged when running panel regressions with: (i) cell and year fixed effects, and (ii) cell and year fixed effects, as well as cell-specific linear trends.** The dependent variable is the *change* in the standard score of the urban population for each cell and each period, using the years = [1891, 1901, 1931, 1960, 1970, 1984, 2000]. The lag of the standard score of the urban population is not included on the right-hand side. We show the results for: (0) the baseline regressions, (i) the panel regression when including cell and year fixed effects (the 1984-2000 period is omitted due to the fixed effects), and (ii) the panel regression when including cell and year fixed effects, as well as cell-specific linear trends (the 1891-1901 period and the 1984-2000 period are omitted due to the fixed effects and the linear trends). We only display the coefficients of the 10 km railroad dummy, as the other coefficients are nil. See *Web Appendix* for data sources.

Web Appendix Figure 5: Ethnic Group and District Boundaries (Africa)



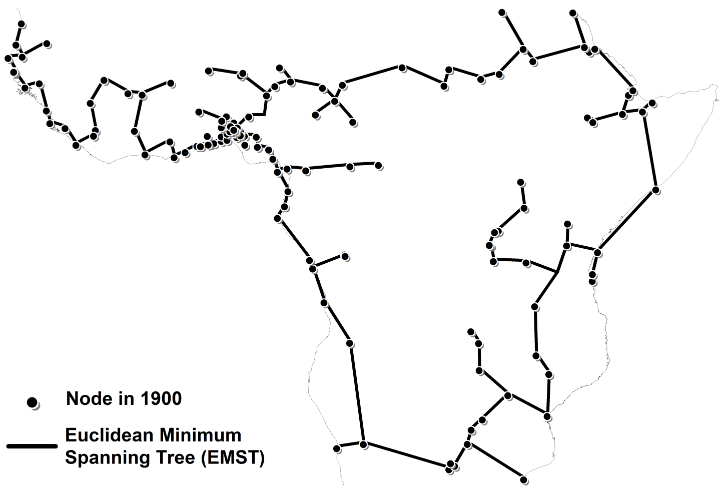
***Notes*: This map shows the boundaries for 755 ethnic groups and 2,304 districts in 39 African countries.** Districts are the third-level administrative unit ("districts", "preferectures", "cercles" or "communes"). The district boundaries are for the year 2000. See *Web Appendix* for data sources.

Web Appendix Figure 6: Military, Mining and Placebo Lines (Africa)



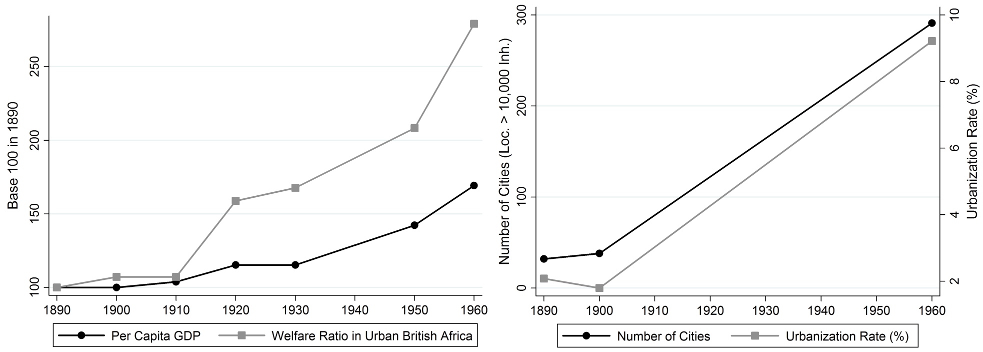
***Notes*: This map shows the lines that were built before 1960 either for military domination or mining and the placebo lines for 39 African countries.** The "placebo" lines are the lines that were planned in 1916 or 1922 but never built. See *Web Appendix* for data sources.

Web Appendix Figure 7: Euclidean Minimum Spanning Tree (Africa)



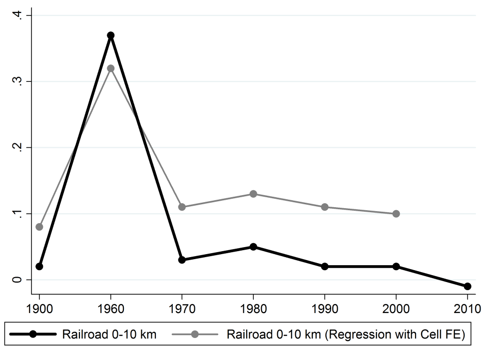
***Notes*: This map shows the Euclidean Minimum Spanning Tree network and the nodes used to compute it for 39 African countries in 1900.** We use the nodes of the initial urban network in 1900 – i.e., all the localities above 10,000 inh. in 1900 as well as the capital, first and second largest cities of each country – to compute the EMST for the whole rail network in 1960. This shows the network that the colonial powers would have built if they had collaborated to optimally connect the initial cities while minimizing construction costs (using the Euclidean distance between each pair of nodes). See *Web Appendix* for data sources.

Web Appendix Figure 8: Income and Urban Patterns for Africa, 1890-1960



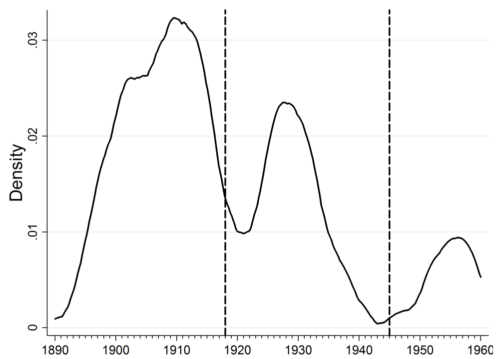
***Notes*: This graph shows that per cap. income, the number of cities and the urbanization rate of the 39 countries have increased in 1900-1960, but not in 1890-1900. The urbanization rate has increased by 8 pct. points. Using the same urban definition, Europe's urbanization rate has increased by 7 pct. points during the Industrial Revolution (from 8% in 1700 to 15% in 1870) (Malanima and Volckart 2007).** The left graph shows the evolution of two measures of income per cap. (base 100 in 1890) in 1890-1960. Per capita GDP data comes from Bairoch (1988). Data on the welfare ratio comes from Frankema and van Waijenburg (2011). The welfare ratio is the fraction by which income exceeds or falls below the subsistence level. Frankema and van Waijenburg (2011) report the welfare ratio for 8 cities of British Africa: Accra, Banjul, Dar es Salaam, Freetown, Kampala, Lagos, Nairobi and Zomba. We use the pop.-weighted average of the welfare ratio of the 8 cities to obtain the welfare ratio for urban British Africa. Data on total pop. for each country comes from Manning (2010). The right graph shows the number of cities and the urbanization rate (%) in 1890, 1900 and 1960. Data on total pop. for each country comes from Manning (2010). See *Web Appendix* for data sources.

Web Appendix Figure 9: Rail Effects for Each Period, Africa 1891-2000



***Notes*: This graph shows that the estimates of the rail effect (for 0-10 km) for each period in 1890-2010 are unchanged when running panel regressions with both cell and year fixed effects.** The dependent variable is the change in the z-score of the urban population for each cell and each period, using the years = [1890, 1900, 1960, 1970, 1980, 1990, 2000, 2010]. The lag of the z-score of the urban population is not included on the right-hand side. We show the results for the baseline regressions, and the panel regression when including cell and year fixed effects (2000-2010 is omitted). We cannot include cell-specific trends because the number of variables is too high for econometric softwares. We only display the coefficients of the 10 km rail dummy. We drop Nigeria when estimating the panel regression as population data is missing for Nigeria's cities in 1890. See *Web Appendix* for data sources.

Web Appendix Figure 10: Kernel Distribution of the Year of Connection for the 0-10 km Railroad Cells, 1890-1960



***Notes*: This graph shows that there were three episodes of rail building in Africa: 1890-1918, 1919-1945 and 1946-1960.** The graph shows the kernel distribution of the year of connection for each 0-10 rail cell for the 39 countries in 1890-1960. 1890 is the first year a cell was reached by a railroad. The two vertical dashed lines represent the last years of World War I (1918) and World War II (1945) respectively. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 1: IDENTIFICATION STRATEGIES, GHANA, 1901-1931

Strategy:	*OLS*	*Ethnic*	*District*	*Long.Lat.*	*Western*	*Placebo*	*C:Placebo*	*IV*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A:	**Dependent Variable = Rural Population (Z-Score) in 1931**							
Rail 1918, 0-30 km	0.46***	0.57***	0.38***	0.40***	0.24***	0.19	0.63***	0.23
(Col.(6): Placebo)	(0.12)	(0.12)	(0.11)	(0.10)	(0.09)	(0.13)	(0.13)	(0.17)
Panel B:	**Dependent Variable = Total Population (Z-Score) in 1931**							
Rail 1918, 0-20 km	0.49***	0.57***	0.38***	0.46***	0.32***	0.02	0.71***	0.54**
(Col.(6): Placebo)	(0.13)	(0.14)	(0.13)	(0.13)	(0.12)	(0.10)	(0.18)	(0.27)
Cell Controls	Y	Y	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the results on rural and total population growth in 1901-1931 are robust when we implement the same identification strategies as in Table 3. For the sake of simplicity, we focus on the 0-30 km dummy for rural population and the 0-20 km dummy for total population.** OLS regressions using data on 554 cells for the years 1901 and 1931. Robust SEs clustered at the district level in parentheses; * $p<0.10$, ** $p<0.05$, *** $p<0.01$. We add the z-scores of the urban and rural populations in 1901 as controls. Table 1 lists the controls. Columns (2)-(3): We add 9 ethnic group FE and 62 district FE. Column (4): We include first, second, third and fourth order polynomials of the longitude and latitude of the cell's centroid. Column (5): The rail dummy is equal to one if the cell is within 10 km from the Western line. Column (6): The rail dummy is equal to one if the cell is within X = [30 and 20] km from a placebo line. Columns (7): The control group (C:) is restricted to all placebo cells. Column (8): We instrument the rail dummy by a dummy equal to one if the cell is within 40 km from the straight lines Sekondi-Tarkwa-Obuasi-Kumasi and Accra-Kumasi (we drop the nodes). In panels A and B, the coefficient (F-stat.) of the instrument in the 1st stage is 0.55 (93) and 0.38*** (39), respectively. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 2: NON-EFFECTS FOR PLACEBO LINES, GHANA, 1901-1931

Type of Placebo Line:	*Planned But Never Built (From West to East)*					*Not Built Yet*	
Placebo Line:	*C.Coast Kumasi 1873*	*Saltpond Kumasi 1893*	*Apam Kumasi 1897*	*Accra Kumasi 1897*	*Accra Kpong 1898*	*Tafo Kumasi 1923*	*H.Valley Kade 1927*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	***Panel A:* Dependent Variable = Cocoa Production (Z-Score) in 1927**						
Placebo 0-40 km	-0.08	0.06	0.12	0.79***	1.20***	0.46**	-0.28*
	(0.21)	(0.21)	(0.20)	(0.19)	(0.33)	(0.22)	(0.16)
	***Panel B:* Dependent Variable = Cocoa Production (Z-Score) in 1927**						
Placebo 0-40 km	-0.39	0.01	-0.06	0.21	0.43	-0.28	0.03
Drop if Rail 0-40 km = 1	(0.29)	(0.18)	(0.15)	(0.13)	(0.44)	(0.17)	(0.14)
	***Panel C:* Dependent Variable = Total Population (Z-Score) in 1931**						
Placebo 0-20 km	-0.05	-0.01	-0.03	0.17	0.18	0.23*	-0.29***
	(0.15)	(0.10)	(0.10)	(0.17)	(0.43)	(0.13)	(0.09)
	***Panel D:* Dependent Variable = Total Population (Z-Score) in 1931**						
Placebo 0-20 km	-0.25**	-0.04	-0.10	-0.13	0.69	0.12	-0.28***
Drop if Rail 0-20 km = 1	(0.11)	(0.09)	(0.08)	(0.11)	(0.52)	(0.12)	(0.10)
	***Panel E:* Dependent Variable = Urban Population (Z-Score) in 1931**						
Placebo 0-10 km	0.30	0.35	0.30	0.46	-1.03	0.45	-0.03
	(0.35)	(0.28)	(0.29)	(0.33)	(1.46)	(0.34)	(0.13)
	***Panel F:* Dependent Variable = Urban Population (Z-Score) in 1931**						
Placebo 0-10 km	0.02	0.13	0.02	0.04	0.55	0.20	0.04
Drop if Rail 0-10 km = 1	(0.16)	(0.16)	(0.15)	(0.16)	(0.75)	(0.19)	(0.17)
Cell Controls	Y	Y	Y	Y	Y	Y	Y
Observations	554	554	554	554	554	554	554

Notes*: This table shows that there are no spurious effects for each placebo line in 1901-1931. Panels B, D and F: We drop the railroad cells to conduct our preferred test, which only compares the placebo cells with the other control cells.** OLS regressions using data on 554 cells for the years 1901 and 1931. Robust SEs clustered at the district level in parentheses; * $p<0.10$, ** $p<0.05$, *** $p<0.01$. ***Panels A-B: The dep.var. is the z-score of cocoa production (tons) in 1931. ***Panels C-D***: It is the z-score of the total population (inh.) in 1931. ***Panels E-F***: It is the z-score of the urban population (inh.) in 1931. For each outcome and each line, we create a dummy equal to one if the cell is less than X km from the line (X = 40, 20 and 10 km respectively). ***Columns (1)-(5)***: We consider the placebo lines that *were planned but not built*. ***Columns (6)-(7):*** We consider lines that *were not built early enough to affect cocoa production in 1927*. ***Panels A, C and E***: We compare the placebo cells with the non-placebo cells. One issue here is that some of the placebo lines overlap with the railroad lines, so that there may be a correlation between the treatment and placebo dummies. Thus, we verify that there are no positive effects for the segments of the placebo lines that do not overlap with the existing lines. ***Panels B, D and F***: We drop the cells within X km from a railroad line (X = 40, 20 and 10 km respectively), in order to compare the placebo cells with the other control cells only. There are no significant positive effects of the placebo lines then. Table 1 lists the controls. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 3: ROBUSTNESS AND SPECIFICATION CHECKS, GHANA, 1901-1931

Panel A: **Dependent Variable = Cocoa Production (Z) in 1927**							
	Baseline (0)	*No Controls* (i)	*Full Sample* (ii)	*No Nodes* (iii)	*No Neighbors* (iv)	*Rail Station* (v)	*Distance* (vi)
Rail 1918, 0-40 km	0.65*** (0.14)	0.77*** (0.22)	1.03*** (0.21)	0.65*** (0.14)	0.65*** (0.14)	0.64*** (0.14)	-0.014*** (0.003)
	Change (vii)	*Panel* (viii)	*Normalization* (ix)	*Norm.: Change* (x)	*Log* (xi)	*Log-Log* (xii)	*Conley SE* (xiii)
Rail 1918, 0-40 km	0.65*** (0.14)	0.66*** (0.19)	1.02*** (0.21)	1.02*** (0.21)	1.42*** (0.29)	-0.97*** (0.17)	0.65*** (0.18)
Panel B: **Dependent Variable = Total Population (Z) in 1931**							
	Baseline (0)	*No Controls* (i)	*Full Sample* (ii)	*No Nodes* (iii)	*No Neighbors* (iv)	*Rail Station* (v)	*Distance* (vi)
Rail 1918, 0-20 km	0.49*** (0.13)	0.48*** (0.17)	0.79*** (0.18)	0.55*** (0.13)	0.44*** (0.14)	0.49*** (0.14)	-0.008*** (0.002)
	Change (vii)	*Panel* (viii)	*Normalization* (ix)	*Norm.: Change* (x)	*Log* (xi)	*Log-Log* (xii)	*Conley SE* (xiii)
Rail 1918, 0-20 km	0.49*** (0.13)	0.53*** (0.18)	0.74*** (0.20)	0.74*** (0.20)	0.75*** (0.25)	-0.60*** (0.13)	0.49** (0.17)
Panel C: **Dependent Variable = Urban Population (Z) in 1931**							
	Baseline (0)	*No Controls* (i)	*Full Sample* (ii)	*No Nodes* (iii)	*No Neighbors* (iv)	*Rail Station* (v)	*Distance* (vi)
Rail 1918, 0-10 km	0.72** (0.28)	0.63** (0.27)	0.87*** (0.32)	0.64** (0.24)	0.60** (0.26)	0.80*** (0.29)	-0.006*** (0.002)
	Change (vii)	*Panel* (viii)	*Normalization* (ix)	*Norm.: Change* (x)	*Log* (xi)	*Log-Log* (xii)	*Conley SE* (xiii)
Rail 1918, 0-10 km	0.72** (0.28)	0.78** (0.39)	1.97** (0.77)	1.97** (0.77)	0.91* (0.51)	-0.41** (0.14)	0.72** (0.24)
Cell Controls	Y	N;Y	Y	Y	Y	Y	Y

Notes*: This table shows that the results in 1901-1931 hold when performing various robustness and specification checks.** OLS regressions using data on 554 cells for the years 1901 and 1931. Robust standard errors clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. ***Column (i): We drop the controls. ***Column (ii):*** We use the full sample. We cannot control for rural population in 1901, because data was not exhaustive for rural population then. ***Column (iii):*** We drop the railroad nodes (Kumasi, Obuasi and Sekondi in our sample of 554 cells). ***Column (iv):*** We drop the nodes and all the cells neighboring a cell containing a node. ***Column (v):*** The rail dummies are generated using the Euclidean distance to railroad stations (km). ***Column (vi):*** The variable of interest is the Euclidean distance (km) of the cell's centroid to the rail. ***Column (vii):*** The dependent variable is the change in the outcome Y between 1901 and 1931 ($ZY_{c,1931} - ZY_{c,1901}$). We do not control for the outcome in 1901. ***Column (viii):*** We use the panel dimension of the data and the dependent variable is for cell c and year $t = [1901;1931]$. The rail dummy is equal to 0 in 1901. We include cell fixed effects (N = 554), year fixed effects (N = 2) and the controls interacted with year dummies. ***Column (ix):*** We use the same normalization as in Black and Henderson (1999). For cell c, year t, the dependent variable is $Y_{c,t}/\overline{Y}_{c,t}$ with $\overline{Y}_{c,t}$ being the mean of outcome Y in the sample. ***Column (x):*** We use the same dependent variable as in Black and Henderson (1999), which is the change in $Y_{c,t}/\overline{Y}_{c,t}$ between 1901 and 1931. We do not control for the outcome in 1901. ***Column (xi):*** We use a log-linear functional form. We use $log(Y_{c,t}+1)$ for the cocoa production and population variables on both the LHS and RHS. We add 1 ton/inhabitant to all cells to make sure we do not drop the cells without production/urban inhabitants. ***Column (xii):*** We use a log-log functional form. We regress the log of the dependent variable ($log(Y_{c,t}+1)$) on the log of the distance (km) of the cell's centroid to the rail. ***Column (xiii):*** Standard errors are corrected for spatial autocorrelation using the approach of Conley (1999), with a distance cut-off of 100 km. The controls are listed in Table 1. See Web Appendix for data sources.

WEB APPENDIX TABLE 4: RAILROADS AND ANTHROPOMETRIC OUTCOMES FOR GHANAIAN SOLDIERS, GHANA, 1867-1937

Dependent Variable:	Soldier Height (cm)
	(1)
Born in Rail 1918, 0-10 km (ζ)	-0.04
	(0.38)
Dummy Born Post-1918 (τ)	1.34***
	(0.35)
Born in Rail 1918, 0-10 km x Dummy Born Post-1918 (β_h)	0.66**
	(0.31)
Cell Controls, World War Dummy, Individual Controls ; Observations	Y ; 5,725

***Notes*: This table shows that the standards of living increased along the lines, using place of birth and anthropometric data of Ghanaian recruits to the British Army.** OLS regressions using data on 5,725 Ghanaian soldiers of the Gold Coast Regiment (GCR) for the years of birth 1867-1937. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The dep.var. is the height (cm) of each soldier. We regress the height on three dummies: a dummy equal to one if the soldier is born in a cell that is within 10 km from a 1918 railroad line, a dummy equal to one if the soldier was born after 1918, and a dummy equal to one if the soldier was born after 1918 in a cell that is within 10 km from a line. The interaction captures the effect of being born along the line. Table 1 lists the controls. We add one dummy equal to one if the soldier was enlisted during World War I (1914-18) or World War II (1939-45), to control for the fact that the selection process of the GCR may have been different during a war. We add the following controls at the individual level: 36 age dummies, 65 ethnic group dummies, and two dummies equal to one if the individual was literate or a farmer before. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 5: COLONIAL RAILROADS, POPULATION GROWTH CONTEMPORARY FACTORS, GHANA, 1901-2000

Dependent Variable in 2000:	Population (Z-Score) Rural	Urban	Total	Urb. Rate (%)	Paved or Impr. Road (%)	Secondary School (%)	Health Clinic (%)	Hospital (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Railroads and contemporary factors in 2000								
Rail 1918, 0-10 km	0.77***	0.87**	0.97***	10.1*	17.2*	11.6***	9.9*	18.6***
	(0.24)	(0.37)	(0.35)	(5.2)	(10.0)	(3.4)	(5.3)	(4.4)
Rail 1918, 10-20 km	0.71***	0.20**	0.32***	2.3	10.2	8.7***	5.5	9.5***
	(0.20)	(0.09)	(0.10)	(5.5)	(8.3)	(2.9)	(4.4)	(3.2)
Rail 1918, 20-30 km	0.68***	0.07	0.19**	-0.6	4.0	3.4	0.2	5.8*
	(0.16)	(0.08)	(0.09)	(4.3)	(8.1)	(2.8)	(3.9)	(3.3)
Rail 1918, 30-40 km	0.49***	0.05	0.14*	-1.9	3.9	2.1	-0.8	4.0
	(0.15)	(0.07)	(0.08)	(4.6)	(7.0)	(3.2)	(4.8)	(3.1)
Panel B: Railroads and contemporary factors in 2000, conditioned on population in 2000								
Rail 1918, 0-10 km	_	_	_	_	5.5	4.7	3.7	7.5*
	_	_	_	_	(3.9)	(5.7)	(3.8)	(10.0)
Rail 1918, 10-20 km	_	_	_	_	3.1	7.1***	5.0	6.2**
	_	_	_	_	(8.2)	(2.6)	(4.2)	(2.8)
Rail 1918, 20-30 km	_	_	_	_	-2.0	2.9	0.8	4.0
	_	_	_	_	(8.2)	(2.8)	(3.9)	(3.0)
Rail 1918, 30-40 km	_	_	_	_	-0.5	1.7	-0.4	2.7
	_	_	_	_	(7.4)	(2.8)	(4.5)	(2.7)
Cell Controls; Mean	Y; 0	Y; 0	Y; 0	Y; 43.3	Y; 44.7	Y; 48.4	Y; 43.3	Y; 14.4

***Notes*: Columns (1)-(4) show that the long-term effects for the rural, urban and total population and the urbanization rate in 2000 are similar to the short-term effects in 1931 (Table 2). Columns (5)-(8) show that railroad cells also have better infrastructure in 2000 (Panel A). These railroad effects on infrastructure today are explained by the higher population densities in the cell in 2000 (Panel B).** OLS regressions using data on 553 cells for the year 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The sample is the same as in Table 2, except we drop Kumasi. Table 1 lists the controls. Panel A: We regress each contemporary outcome on the rail dummies. Panel B: We control for the z-scores of the urban and rural populations in 2000. Column (4): We control for the urbanization rate (%) in 1901. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 6: ROBUSTNESS AND SPECIFICATION CHECKS, GHANA, 1901-2000

Dependent Variable:	Urban Population (Z-Score) in 2000						
	Baseline (0)	*No Controls* (i)	*Full Sample* (ii)	*No Nodes* (iii)	*No Neighbors* (iv)	*Rail Station* (v)	*Distance* (vi)
Rail 1918, 0-10 km	0.71* (0.37)	0.67* (0.37)	1.30** (0.55)	0.94** (0.43)	0.60 (0.36)	0.70** (0.30)	-0.009*** (0.003)
	Change (vii)	*Panel* (viii)	*Normalization* (ix)	*Norm.: Change* (x)	*Log* (xi)	*Log-Log* (xii)	*Conley SE* (xiii)
Rail 1918, 0-10 km	0.80** (0.36)	0.82* (0.49)	3.14* (1.67)	3.14* (1.67)	0.92*** (0.33)	-0.57*** (0.17)	0.71** (0.34)
Cell Controls	Y	N;Y	Y	Y	Y	Y	Y

Notes*: This table shows that the results in 1901-2000 hold when performing various robustness and specification checks.** OLS regressions using data on 553 cells for the years 1901 and 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The sample is the same as in Table 2, except we drop Kumasi. ***Column (i): We drop the controls. ***Column (ii):*** We use the full sample. We cannot control for rural population in 1901, because data was not exhaustive for rural population then. ***Column (iii):*** We drop the railroad nodes (Obuasi and Sekondi in our sample of 553 cells). ***Column (iv):*** We drop the railroad nodes and all the cells neighboring a cell containing a railroad node. ***Column (v):*** The rail dummies are generated using the Euclidean distance to railroad stations (km). ***Column (vi):*** Rail connectivity is measured by the Euclidean distance (km) of the cell's centroid to the rail instead of the four rail dummies. ***Column (vii):*** We use the panel dimension of the data and the dependent variable is for cell c and year $t = [1901;2000]$. The rail dummy is equal to 0 in 1901. We include cell fixed effects (N = 554), year fixed effects (N = 2) and the controls interacted with year dummies. ***Column (viii):*** The dependent variable is the change in outcome Y between 1901 and 2000 ($ZY_{c,2000} - ZY_{c,1901}$). We do not control for the outcome in 1901. ***Column (ix):*** We use the same normalization as in Black and Henderson (1999). For cell c and year t, the dependent variable is $Y_{c,t}/\overline{Y}_{c,t}$ with $\overline{Y}_{c,t}$ being the mean of outcome Y in the sample. ***Column (x):*** We use the same dependent variable as in Black and Henderson (1999), which is the change in $Y_{c,t}/\overline{Y}_{c,t}$ between 1901 and 2000. We do not control for the outcome in 1901. ***Column (xi):*** We use a log-linear functional form. We use $log(Y_{c,t}+1)$ for the population variables on both the LHS and RHS. We add 1 inhabitant to all cells to make sure we do not drop the cells without urban inhabitants. ***Column (xii):*** We use a log-log functional form. We regress the log of the dependent variable ($log(Y_{c,t}+1)$) on the log of the distance (km) of the cell's centroid to the rail. ***Column (xiii):*** Standard errors are corrected for spatial autocorrelation using the approach of Conley (1999), with a distance cut-off of 100 km. The controls are listed in Table 1. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 7: SENSITIVITY OF RESULTS TO VARIOUS CITY THRESHOLDS, GHANA, 1901-2000

Dependent Variable:	Urban Population (Z-Score) in 2000						
Threshold (Inh.) in 2000:	*1,000*	*1,666*	*10,000*	*2,000*	*5,000*	*15,000*	*20,000*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rail 1918, 0-10 km	0.80** (0.36)	0.70** (0.34)	0.76** (0.36)	0.70** (0.34)	0.73** (0.36)	0.75** (0.36)	0.73* (0.38)
Cell Controls	Y	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the results in 1901-2000 are unchanged when using alternative city thresholds for the year 2000.** OLS regressions using data on 553 cells for the years 1901 and 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The dep.var. is the z-score of the urban population (inh.) in 2000. We control for the z-score of the urban population (inh.) in 1901, using 1,000 inh. as the city threshold. ***Column (1): We use 1,000 inh. to define any locality as a city in 2000. ***Columns (2)-(7):*** We use 1,666, 10,000, 2,000, 5,000, 15,000 and 20,000 inh. as the threshold in 2000 respectively. ***Column (2):*** we use 1,666 inh. in 2000, to have a constant relative cut-off between 1901 and 2000. For 1901, for the 66 towns of the 554 cells, the ratio of the minimum to mean city size (2,835 inh.) is 2.84. In 2000 that defines 1,695 towns with a minimum population of 1,666 inh. 1,666 is obtained by dividing the mean city size in 2000 – 4,695 inh. – by 2.84 (see Black & Henderson 1999 for a detailed explanation of this methodology). ***Column (3):*** we use 10,000 inh. in 2000, because the total population of Ghana increased tenfold between 1901 and 2000. The sample is the same as in Table 2, except we drop Kumasi. The controls are listed in Table 1. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 8: SIZE DISTRIBUTION OF CITIES, GHANA, 1960-2000

PANEL A: FOREST:		Actual Distribution in 1960-2000				Steady-State
Cell	Cut-Offs	1960	1970	1984	2000	Distribution
f1	0.32-0.5	0.45	0.45	0.5	0.58	0.58
f2	0.5-1	0.25	0.26	0.23	0.19	0.24
f3	1-2	0.10	0.09	0.09	0.08	0.07
f4	2-4	0.13	0.13	0.12	0.10	0.08
f5	4-max	0.06	0.08	0.06	0.05	0.03
	Threshold	1,000	1,050	1,129	1,341	

PANEL B: GHANA:		Actual Distribution in 1960-2000				Steady-State
Cell	Cut-Offs	1960	1970	1984	2000	Distribution
f1	0.32-0.5	0.49	0.53	0.6	0.63	0.67
f2	0.5-1	0.24	0.2	0.17	0.16	0.18
f3	1-2	0.09	0.08	0.07	0.06	0.05
f4	2-4	0.12	0.12	0.10	0.09	0.06
f5	4-max	0.06	0.07	0.06	0.06	0.04
	Threshold	1,000	1,164	1,276	1,490	

Notes*: This table shows the dynamics and stability in the city size distribution in the Forest and Ghana after 1960. Transition matrices are stable across time, i.e. the dynamics giving rise to the city size distribution is stable; the distribution in 1960 has not yet reached the steady state distribution, whereas the 2000 actual distribution is very close to it.** ***Panel A shows the city size distribution for the Forest (N = 554 cells). ***Panel B*** shows the city size distribution for Ghana (N = 2,091). 5 discrete cells (f1-f5) are used to characterize the city size distribution. Individual city size is expressed relative to mean city size in the respective census. For example, the bottom cell (f1) contains all cities with a population between 0.32 (minimum threshold corresponding to 1,000 inhabitants in the year 1960) and 0.5 of the mean city size in the census year. It is assumed that the city size distribution evolves according to a stationary, first order homogeneous Markov process, with entry. From this a steady state distribution was derived. The methodology was first applied by Black & Henderson (1999). Transition matrices for the post-1960 period are stationary. Existing cities were typically upward mobile, while new cities typically entered in the bottom of the distribution. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 9: IDENTIFICATION STRATEGIES, ROADS, GHANA, 1901-2000

Dependent Variable:	Urban Population (Z-Score) in 2000						
Strategy:	*OLS*	*Ethnic*	*District*	*Long.Lat.*	*IV: Straight Lines for Railroads & Roads 1931*	*Routes 1850*	*1931 & 1850*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rail 1918, 0-10 km	0.77**	0.77**	0.59*	0.81**	1.25***	1.38**	1.23***
	(0.36)	(0.37)	(0.32)	(0.39)	(0.45)	(0.69)	(0.47)
Paved Road 2000, 0-10 km	0.24***	0.24***	0.31***	0.20***	0.53***	0.61*	0.56**
	(0.07)	(0.07)	(0.09)	(0.07)	(0.23)	(0.32)	(0.22)
Improved Road 2000, 0-10 km	0.23**	0.23*	0.12	0.24**	0.28**	0.29**	0.28**
	(0.11)	(0.12)	(0.09)	(0.11)	(0.13)	(0.14)	(0.13)
1st Stage:	*Effects of the instruments on Rail 1918, 0-10 km*						
Straight Line for Rail, 0-40 km					0.17***	0.18***	0.17***
Class 1 Road 1931, 0-10 km					-0.03		-0.03
Class 2 Road 1931, 0-10 km					0.05*		0.05*
Trade Route 1850, 0-10 km						-0.00	-0.00
IV F-Statistic					12.0	16.4	8.0
	Effects of the instruments on Paved Road 2000, 0-10 km						
Straight Line for Rail, 0-40 km					0.00	0.10**	0.01
Class 1 Road 1931, 0-10 km					0.38***		0.34***
Class 2 Road 1931, 0-10 km					0.25***		0.22***
Trade Route 1850, 0-10 km						0.24***	0.17***
IV F-Statistic					18.3	9.7	14.8

Notes*: This table shows that the road effects are lower than the railroad effects by using, when possible, the same identification strategies.** OLS regressions using data on 553 cells for the years 1901 and 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The dep.var. is the z-score of the urban pop. in 2000. The sample is the same as in Table 2, except we drop Kumasi. Table 1 lists the controls. ***Columns (2)-(3): We add 9 ethnic group FE and 62 district FE respectively. ***Column (4):*** We add first, second, third and fourth order polynomials of the longitude and latitude of the cell's centroid. ***Column (5)-(7):*** We instrument the rail and paved road dummies by a dummy equal to one if the cell is within 40 km from the straight lines Sekondi-Tarkwa-Obuasi-Kumasi and Accra-Kumasi (we drop the nodes) and dummies for being within 10 km from existing roads/tracks in 1931 or 1850. We do not instrument the improved road dummy because none of the 1931/1850 road/track variables predict the fact of having an improved road today. The F-stat. are always too low (not shown). ***Column (5):*** The instruments for the roads are two dummies equal to one if the cell was within 10 km from a class 1 road or a class 2 road in 1931 (for the roads). ***Column (6):*** The instrument is a dummy equal to one if the cell was within 10 km from a historical trade route in 1850. ***Column (7):*** The instruments are the three dummies in 1850 and 1931. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 10: RAILROADS, ROADS AND CITIES, GHANA, 1901-2000

Dependent Variable (Z-Score):	Urban Pop. 1931	City 1/0 1931	Urban Pop. 1960	City 1/0 1960	Urban Pop. 2000	City 1/0 2000
	(1)	(2)	(3)	(4)	(5)	(6)
Rail 1918, 0-10 km	0.72**	0.33**	-0.01	-0.11	0.25	-0.10
	(0.29)	(0.14)	(0.11)	(0.13)	(0.15)	(0.13)
Rail 1918, 10-20 km	0.03	0.12	-0.01	0.04	0.04	0.06
	(0.10)	(0.18)	(0.06)	(0.16)	(0.04)	(0.16)
Rail Post-1918, 0-10 km	0.02	0.15	0.04	0.01	-0.08	-0.02
	(0.09)	(0.15)	(0.05)	(0.09)	(0.07)	(0.11)
Class 1 Road 1931, 0-10 km	0.20***	0.45***	0.01	0.34***	0.08	0.31**
	(0.07)	(0.10)	(0.05)	(0.12)	(0.06)	(0.12)
Class 2 Road 1931, 0-10 km	0.09	0.39***	-0.06	0.22*	0.01	0.22*
	(0.11)	(0.13)	(0.06)	(0.12)	(0.04)	(0.12)
Paved Road 1960, 0-10 km			-0.04	0.22**	-0.01	0.16
			(0.08)	(0.10)	(0.07)	(0.14)
Improved Road 1960, 0-10 km			-0.10	0.05	-0.10	-0.06
			(0.07)	(0.13)	(0.14)	(0.15)
Paved Road 2000, 0-10 km					0.01	-0.01
					(0.07)	(0.14)
Improved Road 2000, 0-10 km					0.07	0.16
					(0.12)	(0.14)
Cell Controls	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the effects of the railroads and roads built after 1931 are small. The roads built post-1931 are associated with the creation of new cities (see the effects for City 1/0 in 1960 and 2000), but these cities are too small to modify the urban distribution (see the non-effects for Urban Pop. in 1960 and 2000).** OLS regressions using data on 554 cells for the years 1901, 1931, 1960 and 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. The 20-30 km and 30-40 km rail dummies are added, but the coefficients are not shown. ***Columns (1)-(2): The dep.vars. are the z-scores of the urban pop. and a dummy equal to one if the cell has a city in 1931 respectively. We add the z-scores of the urban and rural pop. in 1901 as controls. ***Columns (3)-(4):*** The dep.vars. are the z-scores of the urban population and a dummy equal to one if the cell has a city in 1960 respectively. We add the z-scores of the urban and rural pop. in 1901 and 1931 as controls. ***Columns (5)-(6):*** The dep.vars. are the z-scores of the urban pop. and a dummy equal to one if the cell has a city in 2000 respectively. We drop Kumasi. We add the z-scores of the urban and rural pop. in 1901, 1931 and 1960 as controls. The variables of interest are six dummies equal to one if the cell is within 10 km from a class 1 road in 1931, a class 2 road in 1931, a paved road in 1960, an improved in 1960, a paved road in 2000 or an improved road in 2000 respectively. Table 1 lists the controls. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 11: ROBUSTNESS FOR PATH DEPENDENCE, GHANA, 1901-2000

Dependent Variable:	Urban Population (Z-Score) in 2000				
	(1)	(2)	(i)	(ii)	(iii)
Rail 1918, 0-10 km	0.32	0.24	0.23	0.21	0.36
	(0.24)	(0.23)	(0.23)	(0.20)	(0.34)
Rail 1918, 10-20 km	0.02	0.07	0.07	0.07	0.13*
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)
Paved Road 2000, 0-10 km	0.04	0.00	0.03	-0.01	0.03
	(0.09)	(0.07)	(0.08)	(0.08)	(0.07)
Improved Road 2000, 0-10 km	-0.02	-0.01	-0.00	0.01	0.04
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Urban Pop. (Z-Score) 1931	0.64***	0.53***	0.54***	0.52***	0.53***
	(0.21)	(0.19)	(0.18)	(0.16)	(0.16)
Rural Pop. (Z-Score) 1931	0.19***	0.17**	0.18**	0.22***	0.25***
	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)
Cell Controls	Y	Y	Y	Y	Y
Historical Factors 1901, 1931	N	Y	Y	Y	Y
Historical Factors 1901, 1931 Squared	N	N	Y	Y	Y
Historical Factors 1901, 1931 Interactions	N	N	N	Y	Y
Cocoa Production 1931	N	N	N	N	Y
Rail Station and Cocoa at Rail Station 1918	N	N	N	N	Y

Notes*: This table shows that the results on the contribution of historical factors to path dependence are unchanged when performing various robustness checks.** OLS regressions using data on 553 cells for the years 1901, 1931, 1960 and 2000. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. We drop Kumasi. The 20-30 and 30-40 km rail dummies are included, but the coefficients are not shown. ***Columns (1)-(iii): We include two dummies equal to one if the cell is within 10 km from a paved road or an improved road in 2000 respectively. ***Columns (2)-(iii):*** We control for historical factors in 1901 and 1931: dummies equal to one if the cell has a school, a hospital, or a church, and the respective numbers of schools, hospitals and churches for each cell, and dummies equal to one if the cell is within 10 km from a class 1 road, class 2 road, or class 3 road respectively. ***Columns (i)-(iii):*** We include the squares of the numbers of schools, hospitals and churches for each cell. ***Columns (ii)-(iii):*** We include two sets of dummies: the interactions of the dummies equal to one if the cell has a school, a hospital, or a church, and the interactions of the dummies equal to one if the cell is within 10 km from a class 1 road, class 2 road, or class 3 road. ***Column (iii):*** We also include the z-score of cocoa production (tons) in 1931, a dummy equal to one if there is a railroad station in the cell in 1918, and the z-score of the amount of cocoa (tons) brought to the station in 1918, as controls for "private capital" in 1931. The controls are listed in Table 1. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 12: CITY DATA SOURCE INFORMATION BY COUNTRY, 1890-2010

Country	Main Sources
Angola	Colonial Population Censuses (Reports), Colonial Statistical Abstracts, Africapolis (2012)
Benin	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Bissau	Colonial Statistical Abstracts, Africapolis (2010)
Botswana	Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)
Burkina-Faso	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Burundi	Colonial Statistical Abstracts, Africapolis (2012)
Cameroon	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
CAR	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
Chad	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
Congo	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
Djibouti	Colonial Statistical Abstracts, Africapolis (2012)
Eq. Guinea	Colonial Handbooks, Africapolis (2012)
Eritrea	Colonial Handbooks, Africapolis (2012)
Ethiopia	Colonial Handbooks, Africapolis (2012)
Gabon	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
Gambia	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2010)
Ghana	Colonial Population Censuses (Reports), Africapolis (2010)
Guinea	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Ivory Coast	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Kenya	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2012)
Liberia	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2010)
Malawi	Colonial Handbooks, Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)
Mali	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Mauritania	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Mozambique	Colonial Handbooks, Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)
Namibia	Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)
Niger	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Nigeria	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2010)
Rwanda	Colonial Statistical Abstracts, Africapolis (2012)
Senegal	Colonial Population Censuses (Reports), Colonial Statistical Abstracts, Africapolis (2010)
Sierra Leone	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2010)
Somalia	Colonial Handbooks, Africapolis (2012)
Sudan	Colonial Handbooks, Africapolis (2012)
Tanzania	Colonial Handbooks, Colonial Population Censuses (Reports), Africapolis (2012)
Togo	Colonial Administrive Counts, Colonial Statistical Abstracts, Africapolis (2010)
Uganda	Colonial Handbooks, Africapolis (2012)
Zaire	Colonial Handbooks, Colonial Statistical Abstracts, Africapolis (2012)
Zambia	Colonial Handbooks, Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)
Zimbabwe	Colonial Population Censuses (Reports), Post-Colonial Population Censuses (Reports)

***Notes*: This table shows the main sources used to reconstruct the city population series for 39 selected Sub-Saharan African countries for the following years = [1890, 1900, 1960, 1970, 1980, 1990, 2000 and 2010].** While some of these sources could be found on the internet, we found many of them at the Census Bureau Library in Washington D.C., the British Library and the libraries of the London School of Economics and SOAS in London, and the National Archives and the Centre Population et Développement in Paris. The other two main sources that we used for all countries for the earlier years (1890 and 1900) are: *Encyclopedia Britannica* 1901 and 1911, and *Wikipedia* (for each city). See *Web Appendix* for data sources.

WEB APPENDIX TABLE 13: COLONIAL RAILROADS AND URBAN GROWTH, IDENTIFICATION AND GENERAL EQUILIBRIUM EFFECTS, AFRICA 1900-1960

Dependent Variable (Z-Score):	Urban Population in 1960 (Column (8): City 1/0 in 1960)							
Strategy:	*Military*	*Mining*	*Non-Mil.Min.*	*Placebo 1916*	*Placebo 1922*	*C:Placebo 1916*	*C:Placebo 1922*	*OLS City1/0*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Rail 1960, 0-10 km	0.32***	0.37***	0.31***	0.04	0.10**	0.29***	0.34***	0.46***
	(0.05)	(0.08)	(0.09)	(0.03)	(0.05)	(0.06)	(0.08)	(0.04)
X km or Pre-Rail Dummy:	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*	*City1900*	*City1890*	*Coastal*
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Col.(17)-(21): Rail, X km	0.01	-0.00	-0.00	0.02	-0.00	17.42***	5.03*	0.02
Col.(22)-(24): Pre-Rail	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(4.76)	(2.86)	(0.02)
Country FE, Cell Controls	Y	Y	Y	Y	Y	Y	Y	Y

Notes*: Columns (1)-(7) show that the results are robust if we focus on the lines built for military domination or mining, when considered separately, or the placebo lines that were proposed in 1916 or in 1922 but never built, when considered separately. Column (8) shows that many cities were created along the lines (the dependent variable is the z-score of a dummy equal to one if there is a city in the cell in 1960). Columns (9)-(13) show that there is no negative effect for the adjacent cells (we only consider the cells until 60 km). Columns (14)-(15) show that there is no negative effect for other competing cells (the cells with a city in 1900 or a city in 1890, or the cells along the coast).** OLS regressions using data on 193,923 cells for 39 countries for the years 1900 and 1960. Robust SEs clustered at the district level (N = 2,304) in parentheses; * p<0.10, ** p<0.05, *** p<0.01. We drop the capital, first and second largest cities of each colony. ***Columns (1)-(3): The rail dummy is equal to one if the cell is within 10 km from a line built for military domination, mining (we drop the 60 mining nodes), or another reason. ***Columns (4)-(5):*** We test that there are no effects for the placebo lines (1916 or 1922). ***Columns (6)-(7):*** The control group (C:) are the placebo cells. ***Columns (9)-(13):*** The variable of interest is a dummy equal to one if the cell is within X km from a line (we drop the cells below X km). ***Columns (14)-(15):*** Pre-Rail is a dummy equal to one if the cell has a city in 1900 or 1890 respectively (we drop Nigeria for 1890, as we do not have data on its cities). ***Column (16):*** It is a dummy equal to one if the cell is coastal. All regressions include country FE and the following controls: Two dummies equal to one if the cell is within 10 km from the coast or a navigable river, Euclidean distances (km) to the coast, a navigable river, the largest city, the second largest city or the capital city respectively, mean and standard deviation of altitude (m), average annual rainfall (mm) in 1900-1960 and the shares (%) of class 1, class 2, class 3, undetermined, sparsely vegetated and submerged soils in the cell. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 14: ROBUSTNESS AND SPECIFICATION CHECKS, AFRICA, 1900-1960

Dependent Variable:	Urban Population (Z) in 1960					
	Baseline (1)	*No Controls* (2)	*No Neighbors* (3)	*Distance* (4)	*Panel* (5)	*Change* (6)
Rail 1960, 0-10 km	0.37***	0.37***	0.38***	-0.005***	0.31***	0.31***
	(0.05)	(0.05)	(0.05)	(0.001)	(0.07)	(0.05)
	Normalization (7)	*Norm.: Change* (8)	*Log* (9)	*Log-Log* (10)	*Country SE* (11)	*Conley SE* (12)
Rail 1960, 0-10 km	13.4***	7.5***	0.19***	-0.03***	0.37***	0.37***
	(1.8)	(2.5)	(0.02)	(0.00)	(0.11)	(0.07)
Country FE, Cell Controls	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the effects for Africa in 1900-1960 hold when performing various robustness checks.** OLS regressions using data on 193,923 cells for 39 countries for the years 1900 and 1960. Robust SEs clustered at the district level reported in parentheses; * p<0.10, ** p<0.05, *** p<0.01. We control for the z-score of the urban population in 1900. ***Column (2): We drop the controls. ***Column (3):*** We drop the cells neighboring a cell containing a node (the largest, second largest and capital cities of each country). ***Column (4):*** The variable of interest is the Euclidean distance (100 km) of the cell's centroid to the rail. Column (5): we use the panel dimension of the data and the dependent variable is for cell c and year $t = [1900;1960]$. The rail dummy is equal to 0 in 1900. We include cell and year fixed effects, country-year fixed effects and the controls interacted with year dummies. ***Column (6):*** The dependent variable is the change in the outcome between 1900 and 1960 ($ZY_{c,1960} - ZY_{c,1900}$). We do not control for the outcome in 1900. ***Column (7):*** We use the normalization of Black and Henderson (1999). For cell c and year t, the dependent variable is $Y_{c,t}/\overline{Y}_{c,t}$ with $\overline{Y}_{c,t}$ being the mean of the outcome. ***Column (8):*** We use the same dependent variable as Black and Henderson, i.e. the change in $Y_{c,t}/\overline{Y}_{c,t}$ between 1900 and 1960. We do not control for the outcome in 1900. ***Column (9):*** We use a log-linear functional form. We use $log(Y_{c,t}+1)$, adding 1 to all cells to avoid dropping the cells without urban residents in 1901 and 1931. ***Column (10):*** We use a log-log functional form. We regress the log of the dependent variable ($log(Y_{c,t}+1)$) on the log of the distance (km) of the cell's centroid to the rail. ***Column (11):*** SEs are clustered at the country level. ***Column (12):*** SEs are corrected for spatial autocorrelation using the approach of Conley (1999, cut-off of 100 km). The controls are listed in the footnote of Web Appendix Table 12. See Web Appendix for data sources.

WEB APPENDIX TABLE 15: USE OF VARIOUS CITY THRESHOLDS, AFRICA 1900-2010

	Panel A: Dependent Variable: Urban Population (Z-Score) in 1960					
Threshold (Inh.) in 1960:	*10,000*	*17,322*	*22,000*	*15,000*	*20,000*	*50,000*
	(1)	(2)	(3)	(4)	(5)	(6)
Rail 1918, 0-10 km	0.37*** (0.05)	0.33*** (0.05)	0.30*** (0.05)	0.34*** (0.05)	0.31*** (0.05)	0.22*** (0.05)
	Panel B: Dependent Variable: Urban Population (Z-Score) in 2010					
Threshold (Inh.) in 2010:	*10,000*	*31,763*	*83,000*	*15,000*	*20,000*	*50,000*
	(1)	(2)	(3)	(4)	(5)	(6)
Rail 1918, 0-10 km	0.34*** (0.04)	0.30*** (0.04)	0.25*** (0.04)	0.33*** (0.04)	0.32*** (0.04)	0.28*** (0.04)
Country FE, Cell Controls	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the results in 1900-1960-2010 are unchanged when using alternative city thresholds for the years 1960 and 2010.** OLS regressions using data on 193,923 cells for 39 countries for the years 1900, 1960 and 2010. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. In ***Panel A, the dep.var. is the z-score of the urban population (inh.) in 1960. We control for the z-score of the urban population (inh.) in 1900, using 10,000 inh. as the city threshold. ***Column (1):*** We use 10,000 inh. to define any locality as a city in 1960. ***Columns (2)-(6):*** We use 17,322, 22,000, 15,000, 20,000 and 50,000 inh. in 1960 respectively. ***Column (2):*** We use 17,322 inh. in 1960, to have a constant relative cut-off over in 1900-1960. For 1900, for the 73 cities with a pop. above 10,000, the ratio of the minimum to mean city size (24,973 inh.) is 2.50. In 1960 that defines 248 cities with a minimum pop. of 17,322 inh. 17,322 is obtained by dividing the mean city size in 1960 – 43,305 inh. – by 2.50 (see Black & Henderson 1999 for details of this methodology). ***Column (3):*** We use 22,000 inh. in 1960, because the population of Sub-Saharan Africa (SSA) was multiplied by 2.2 between 1900 and 1960. In ***Panel B***, the dep.var. is the z-score of the urban population (inh.) in 2010. We control for the z-score of the urban population (inh.) in 1900, using 10,000 inh. as the threshold. ***Column (1):*** We use 10,000 inh. to define any locality as a city in 2010. ***Columns (2)-(6):*** We use 31,763, 83,000, 15,000, 20,000 and 50,000 inh. in 2010 respectively. ***Column (2):*** We use 31,763 inh. in 2010, to have a constant relative cut-off over in 1900-2010. We use the same ratio of the minimum to mean city size (2.50) as for 1900-1960. In 2010 that defines 990 cities with a minimum pop. of 31,763 inh. 31,763 is obtained by dividing the mean city size in 2010 – 79,408 inh. – by 2.50. ***Column (3):*** We use 83,000 inh. in 2010, because SSA's population was multiplied by 8.3 between 1900 and 2010. The controls are listed in the footnote of Table 8. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 16: IDENTIFICATION STRATEGIES, AFRICA, 1900-2010

Dependent Variable in 2010 (Z-Score):	Urban Population								City 1/0
Strategy:	*Baseline*	*Ethnic*	*District*	*Long.Lat*	*Mil.Min*	*Placebo*	*C:Placebo*	*IV*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Rail 1960, 0-10 km (Col.(6): Placebo)	0.34*** (0.04)	0.32*** (0.04)	0.27*** (0.04)	0.34*** (0.04)	0.29*** (0.05)	0.08*** (0.05)	0.23*** (0.06)	0.43*** (0.12)	0.55*** (0.03)
Country FE, Cell Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y

Notes*: This table shows that the long-term effects are causal in Africa in 1900-2010, using the same identification strategies as in 1900-1960.** OLS regressions using data on 193,923 cells for 39 countries for the years 1900 and 2010. Robust SEs clustered at the district level in parentheses; * p<0.10, ** p<0.05, *** p<0.01. We control for the z-score of the urban population in 1900. ***Columns (2)-(3): We add 755 ethnic group FE and 2,304 district FE respectively. ***Column (4):*** We add country-specific first, second, third and fourth order polynomials of the longitude and latitude of the cell's centroid. ***Column (5):*** The rail dummy is equal to one if the cell is within 10 km from a line built for military domination or mining (we drop the 60 mining nodes). ***Column (6):*** We test that there are no effects when using the placebo lines (1916-22) as an alternative. ***Column (7):*** The control group (C:) is restricted to all placebo cells. ***Column (8):*** We instrument the rail dummy by a dummy equal to one if the cell is within 40 km from a straight line of the EMST network based on the initial urban network in 1900 (all the cities as well as the capital, first and second largest cities of each colony). The coefficient (F-stat) of the instrument in the 1st stage is 0.08*** (92). ***Column (9):*** The dep.var. is the z-score of a dummy equal to one if the cell has a city in 2010. The controls are listed in the footnote of Table 8. See *Web Appendix* for data sources.

WEB APPENDIX TABLE 17: SIZE DISTRIBUTION OF CITIES, AFRICA, 1960-2010

		Actual Distribution in 1960-2010						Steady-State
Cell	Cut-Offs	1960	1970	1980	1990	2000	2010	Distribution
f1	0.32-0.5	0.54	0.6	0.62	0.65	0.67	0.69	0.70
f2	0.5-1	0.14	0.13	0.12	0.12	0.10	0.10	0.12
f3	1-2	0.07	0.07	0.07	0.05	0.07	0.05	0.06
f4	2-4	0.15	0.11	0.09	0.08	0.07	0.08	0.06
f5	4-max	0.10	0.10	0.10	0.10	0.09	0.08	0.05
	Threshold	13,833	16,277	18,721	21,872	23,477	26,062	

***Notes*: This table shows the dynamics and stability in the city size distribution in Africa after 1960. Transition matrices are stable across time, i.e. the dynamics giving rise to the city size distribution is stable; the distribution in 1960 has not yet reached the steady state distribution, whereas the 2010 actual distribution is close to it.** The table shows the city size distribution for the 39 African countries as a whole (N = 194,000 cells). 5 discrete cells (f1-f5) are used to characterize the city size distribution. Individual city size is expressed relative to mean city size in the respective census. For example, the bottom cell (f1) contains all cities with a pop. between 0.32 (minimum threshold corresponding to 10,000 inh. in the year 1960) and 0.5 of the mean city size in the census year. It is assumed that the city size distribution evolves according to a stationary, first order homogeneous Markov process, with entry. From this a steady state distribution was derived. The methodology was first applied by Black & Henderson (1999). Transition matrices for the post-1960 period are stationary. Existing cities were typically upward mobile, while new cities typically entered in the bottom of the distribution. See *Web Appendix* for data sources.