Railroads of the Raj: Estimating the Impact of Transportation Infrastructure ONLINE APPENDIX

Dave Donaldson
MIT and NBER

A Data Appendix

This appendix provides information (supplementary to that in Section 2) on the data used in this paper.

Sample of Districts:

The data cover the areas of modern-day India, Pakistan and Bangladesh, most of the area known as British India. I work with a panel of 235 geographic units of analysis that I refer to as districts, for as much of the period 1870 to 1930 as possible.[1]

Trade Cost Proxy Variables:

I construct trade cost proxy variables using a newly constructed GIS database on the Indian transportation network, from 1851 to 1930. The database covers four modes of transportation: railroads, roads, rivers and coastal shipping. To construct this database, I begin with a GIS database that contains the locations of contemporary railroad, river and coast lines from the *Digital Chart of the World*. Each segment (approximately 20 km long) of the railroad

[1]The majority of British India was under direct British control, and was divided into nine large, administrative units known as provinces. Due to poor data availability, the provinces of Assam and Sindh are not in my main sample directly (though they appear as trade partners in the trade data descried below.) Each province was further sub-divided into a total of 223 districts (as of 1891, the district definition that I use throughout). Areas not under direct British control were known as 'Princely States'. For administrative purposes these were grouped into divisions similar to the provinces and districts described above, so in princely state areas I use the lower administrative units as my units of analysis and refer to them as districts, following the *Indian Administrative Atlas* (Singh and Banthia, 2004). There were 251 of these districts in my sample area, but data collection in the princely states was extremely incomplete and I include only 24 districts from the princely state regions in my final sample.

network is coded according to the year in which it was opened.[2] For river transport I keep only those rivers that are reported in Schwartzberg (1978) or Bourne (1849) as navigable in 1857. The final component of the colonial India GIS database that I construct is the location of each district and salt source. To calculate district locations I digitize a map of the district borders in India (as they existed in 1891) based on maps in the *Indian Administrative Atlas* and *Constable's Hand Atlas of India* (Bartholomew, 1893). I use this to calculate district centroids, which I take to be the 'location' of each district. Finally, I obtain the location of each salt source from contemporary maps.

I then convert the GIS database of transportation lines and district/salt source locations into a graph of nodes and arcs, as is common in the transportation literature (Black, 2003). I work with a simplified graph representation of the Indian transportation network, where the number of nodes and the sparsity of arcs is low enough for network algorithms to be feasibly operated on it using a desktop computer (the resulting network has 7651 nodes). To do this, I use the 'simplify' command in ArcGIS. A line in ArcGIS is a series of vertices connected by straight lines. The 'simplify' command removes vertices in such a way as to minimize the sum of squared distances between the original line and the simplified line. The original *Digital Chart of the World* railroad layer, for example, consists of approximately 33,000 vertices; I simplify the railroad layer to one of only 5616 vertices.

Because the density of informal roads was extremely high (Deloche, 1994), I allow road transport to occur along the straight line between any two nodes on the network, but only if the two nodes either represent districts or salt sources, or the two nodes are within 1000 km of each other.[3] The result is a network with 7651 nodes, 5616 of which represent the railroad network, 660 of which represent the navigable river network, 890 of which represent coastal shipping routes, 477 of which represent the centroids of the 477 districts in India (in 1891 borders), and 8 of which represent the locations of the sources of 8 different types of salt. Because the railroad arcs are coded with a year of opening indicator, this network can be restricted to represent the transportation network for any year from 1851 to 1930 by simply turning these arcs on or off.

Finally, I use this network representation of the Indian transportation system to calculate

[2]To do this I use the publication *History of Indian Railways, Constructed and in Progress* (1918 and 1966 volumes), the 1966 volume of which refers to railroad lines in modern-day India only. To obtain years of opening for line segments in modern-day Pakistan and Bangladesh from 1919 to 1930 I use the annual *Railway Reports* published by the Railways Department, which list all line section openings in each year.

[3]Allowing straight-line road travel between any two nodes would yield a network with over 58 million arcs. The shortest path between each of the nodes on such a dense network cannot be calculated using a desktop computer, so I restrict many of these arcs to be non-existent; the result is that the 7651-by-7651 matrix representing the network can be stored as a sparse matrix, and analyzed using sparse matrix routines (that increase computation speed dramatically) in Matlab.

the variable $LCRED(\boldsymbol{\alpha}, \mathbf{R}_t)$, described in Section 4. This variable is a measure of the cost of traveling between any two points (where a point is either a district or a salt source) in a year using the lowest-cost route along the network (available in that year). The lowest-cost route depends on the value of the relative per unit distance costs of using each mode (rail, river, coast, or road), $\boldsymbol{\alpha}$, and the available transportation network, $\mathbf{R}_t$, whose construction was described above. Conditional on values of $\boldsymbol{\alpha}$, I use a standard algorithm from graph theory and transportation science (Dijkstra's algorithm, implemented using the Boost Graph Library for Matlab) to calculate the shortest path between every pair of points, along the transportation network available in each year from 1870 to 1930. The resulting measure, $LCRED_{odt}(\boldsymbol{\alpha}, \mathbf{R}_t)$, is in units of railroad-equivalent kilometers due to the normalization of $\alpha^{rail} = 1$.[4]

Bilateral Trade Flows:

Data on rail and river trade (along the Brahmaputra, Ganges and Indus river systems) within India were published separately for each province.[5] The geographic unit of analysis in these records is the 'trade block', which typically spans between three and five districts.[6] Four of these trade blocks represented the four major ports of colonial India (Bombay, Calcutta, Karachi and Madras). When a port was represented in this publication its imports included the sum of imports from other regions of India destined for export out of the port city by sea (for either international export or export by coasting trade to another port within India), or destined for consumption/absorption within the port city; an analogous situation held for exports. I therefore treat these four port city trade blocks as four economic units whose trade demands and supplies represent the sum of both intra-city and international demands and supplies for and of goods. (This treatment is described in detail in Appendix C below.) The trade flow data represents final shipments between two regions (even if a shipment changed railroad companies).[7] Only if a shipment was taken off the railroad system and re-shipped

[4]For an estimate of intra-district trade costs (not needed for estimation, but necessary when solving for the full model equilibrium) I follow the CEPII GeoDist database (Mayer and Zignago, 2011) and proxy for the average distance among points within a district with a simple formula that involves the district's area, and apply the road-based distance cost to these distances.

[5]Data on intra-Indian sea trade (between the main ports of each province) were also collected and published but without data on port city net absorption such data cannot be used to improve upon the procedure in Appendix C.

[6]Trade blocks split into smaller blocks over time, but I aggregate over these splits to maintain constant geographic units. The trade blocks were typically drawn so as to include whole numbers of districts.

[7]All bilateral block-to-block intra-provincial trade flows were published, except that from a block to itself (which was always unreported). Inter-provincial trade flows were published from each internal block to each external province (and vice versa), but not by trade block within the external province. I therefore create a full set of inter-provincial block-to-block flows by assigning a province's trade block's imports from each of another province's trade blocks in proportion to the exporting blocks' stated exports to the entire importing

onwards would it be counted as two separate shipments. I collect these data from various annual, provincial publications from 1882 onwards.[8]

Trade data on the physical quantities shipped were published disaggregated by commodities.[9] In order to compare commodities across these different levels of aggregation, I aggregate all data to the level of the 17 commodities (listed below) for which agricultural output and price data are available, plus salt. This was not possible for three grain crops (barley, maize, and ragi) for which the trade categories were usually insufficiently disaggregated; in addition two crops (bajra and jowar) were usually combined so I work with those as a composite crop in the trade data. This data is available (as an unbalanced panel) from 1882 to 1920 only. This generates the variable X^k_{odt} in the text.

Rainfall Data:

A thick network of 3614 rain gauges at meteorological stations scattered throughout colonial India recorded daily rainfall amounts from 1891-1930. From 1901 onwards, these records have been digitized by the Global Historical Climatology Network (Daily) project; the GHCN dataset also provides the latitude and longitude of each station. For the years 1891-1900, I collect the data from the publication, *Daily Rainfall for India in the year....* In the years 1870 to 1890, very little daily rainfall data were published in colonial India, but monthly data from 365 stations (spread throughout India) were published by each province.[10] I convert monthly station-level data to daily station-level data using a modeling procedure that is common in the meteorological statistics literature (eg, Ngo-Duc, Polcher and Laval (2005)).[11] I convert

province (and vice versa for exports). Some bilateral flow reports were sufficiently minor that they were reported for a combined group of partner regions, so I drop these flows.

[8]The titles of these publications changed over time, from *Returns of the Rail [and River-borne] Trade of [Province]* to *Report on the trade carried by rail [and river] in [Province]* to *Report on Inland Trade of [Province].* In the province of Madras, these statistics were only published from 1909 onwards. Railroad trade statistics were not published by the princely states themselves, but each province's external trade to/from the large princely states (Central India Agency, Hyderabad, Mysore, Rajputana and Travancore) were published. I therefore treat these as a single trade block.

[9]Data on trade values are available but these were reported in original sources by converting quantities into values on the basis of average prices within a commodity across partners, a correction that is irrelevant given the fixed-effects that I use here.

[10]These publications included the *Administration Reports* for each province, described in the agricultural price data section below. I use additional data (to increase the number of stations) that were published in selected provinces' *Sanitary Reports.*

[11]Specifically, using daily data from 1891 to 1930, I estimate the district-specific relationship between the pattern of monthly rainfall in a year and the rainfall on any day of that year; I then use these estimated relationships to predict the rainfall on any day in a given district and year from 1870 to 1890, conditional on the pattern of monthly rainfall actually observed in that district and year. While these daily rainfall predictions are likely to be imprecise, much of the imprecision is averaged over when I construct crop-specific rainfall shocks, which are measures of the total rainfall in a given period (a length ranging from 55 to 123 days.)

station-level data to district-level data by simply averaging over the many stations in each district (using spatial interpolation in the case of missing observations). Finally, as described in Section 5, I use the *Indian Crop Calendar* (Directorate of Economics and Statistics, 1967) to compute the total amount of rain that fell during each crop's growing season (or over the total of multiple seasons, where relevant), as defined in the *Crop Calendar*), in each district and year.[12] This generates the variable $RAIN_{ot}^{k}$ used in the text.

Prices of Salt and Agricultural Commodities:

I use data on six different types of salt[13] for each of the six provinces in Northern India as well as data on 17 agricultural commodities[14] from all of India. I collect this price data from various annual, provincial publications.[15] Prices reported in these publication were an average of observations taken by district officers once per fortnight at each of typically 10-15 leading retail markets per district.

Real Agricultural Income:

I use data that present the area under each of the 17 crops, and the yield per acre for each of these crops, in each district and year.[16] I take the product of each area and yield pair

[12]This crop calendar covers the regions of colonial India that are in modern-day India, but not in Bangladesh or Pakistan. For districts in Bangladesh and Pakistan I assign growing seasons (to each crop) that reflect an average of the start and end dates (of the reported growing season) in proximate Indian districts. Likewise, I spatially interpolate in cases where the *Crop Calendar* reports a missing value for a given crop-district, and use the annual total rainfall amount for crops not covered at all.

[13]These six salt types are those from: the Bombay sea salt sources near the city of Bombay, salt from the UK distributed via Calcutta, the Didwana salt source in Punjab, the Kohat mines in Punjab (principally the Jatta mine, according to Watt (1889)), the Salt Range mines in Punjab (principally the Mayo mine, according to Watt (1889)), and the Sambhar Salt Lake in Rajputana, Two other types (Mandi and Sultanpur) were widely discussed in salt reports of the time but featured in price statistics too infrequently to be relevant conditional on the fixed effects used.

[14]These crops are: bajra, barley, cotton, gram, indigo, jowar, jute, linseed, maize, opium, ragi, rice, sesamum, sugarcane, tea, tobacco, and wheat. District-level retail price data are not available for the cash crops indigo, opium, and tea. I therefore use unit values from India's export statistics as the nation-wide price of these three crops.

[15]These publications are: *Prices and Wages in India*; *Administration Reports* from all provinces; the *Salt Report of Northern India*; the *Statistical Atlas of Andhra State* with agricultural price data (for the Madras Presidency); the *Season and Crop Reports* from various provinces with agricultural price data; and the *Sanitary Reports* from various provinces with data on prices of food grains. When two sources report on the same price I take the average. In total, salt price data are available from 133 districts and agricultural price data from 235 districts.

[16]These data were published in *Agricultural Statistics of India* from 1884 to 1930. For the years 1870-1883 I use data on crop areas and yields in the provincial *Administration Reports*, as described in the agricultural prices data section above. Data on agricultural output were published in each province's *Administration Report* except for Punjab and Bengal. For supplementary data I use each province's *Season and Crops Report* between 1904 and 1930. These various sources use varying crop classification schemes, so I work with 17 major crops that minimize the need for aggregation and concordance. When yield or price observations are not available for a given crop-district-year (that nevertheless has a non-missing area observation) I

to create a measure of real output for each crop, district and year. I then evaluate this bundle of 17 real output measures at the retail prices prevailing for these crops (from the agricultural price data described above), in each district and year, to create a measure of total nominal agricultural output for each district and year. Finally, I divide nominal output by a consumer price index (the Törnqvist index) to create a measure of real income.[17]

B Proof of Result 3

Consider the simplified version of the model, as in Section 3.2 (the simplified version of which was only assumed to obtain Result 3). That is, there are only three regions o (X, Y and Z), one commodity, and the regions are initially symmetric ($L_o = 1$, $A_o = A\lambda_1^\theta$, $T_{od} = T$ for all $o \neq d$, and $T_{od} = 1$ for all $o = d$). Now consider a symmetric change in trade costs between regions X and Y only (ie $dT_{XY} = dT_{YX} \neq 0$) and solve for the change in region X's real income ($dW_X = dr_X - dp_X$). Let $r_X = 1$ at all times (ie $dr_X = 0$) by choice of the numeraire; by symmetry the same holds true for r_Y. Solving for the change in real income is then simply a matter of solving for dp_X, since $dW_X = -dp_X$.

Totally differentiating the price equation for p_X (equation (4)) and evaluating this around the symmetric initial equilibrium we obtain

$$p^{-(\theta+1)}dp_X = AT^{-(\theta+1)}dT_{YX} + AT^{-\theta}dr_Z, \tag{1}$$

where $p_o = p = A^{-1/\theta}(1 + 2T^{-\theta})^{-1/\theta}$. To obtain an expression for dr_Z, totally differentiate region Z's land market clearing condition (equation (6)) around the symmetric equilibrium to obtain

$$[(1+\theta)A^{-1} - p_Z^\theta]dr_Z = 2\theta p^{\theta-1}T^{-\theta}dp_X + \theta p^{\theta-1}dp_Z, \tag{2}$$

where this step uses the fact that, because of symmetry, $dp_X = dp_Y$. Finally, note from the price equation for region Z (ie equation (4)), total differentiation around the symmetric

interpolate first over time within a district between non-missing observations, and then fill in any remaining missing values via spatial interpolation. In total, output data are available for 192 districts but base spatial interpolation of price data on the full dataset of 235 districts. While Blyn (1966), Heston (1973), and Dewey (1979) have discussed the potential for measurement error in these sources, these authors have not been concerned with mechanisms through which measurement error might be correlated (conditional on the fixed effects in place) with the regressors I use in this paper.

[17] In order to compute this consumer price index I use district and year specific consumption weights for each crop, computing consumption as output minus net exports (assigning net exports, within each commodity, proportionally across districts within each trade block) or zero if apparent net exports exceed output (as is possible due to the proportional assignment procedure). For years post-1920 (when trade data was not available) and pre-1890 (when trade data was incomplete) I assign the 1920 and 1890 values, respectively. Likewise, for the one native state for which agricultural output data was available, Mysore, I assign the trade weights (not available for Mysore) from Madras, a neighboring province.

equilibrium again implies

$$dp_Z = Ap^{\theta+1}dr_Z. \tag{3}$$

Substituting equations (2) and (3) into equation (1) we obtain

$$\left[1 - \frac{2\theta T^{-2\theta}}{(1+\theta)A^{-2}p^{-2\theta} - \theta - A^{-1}p^{-\theta}}\right] dp_X = p^{\theta+1}AT^{-(\theta+1)}dT_{XY}. \tag{4}$$

Noting that $A^{-1}p^{-\theta} = (1 + 2T^{-\theta})$, this simplifies to

$$\left[1 - \frac{2\theta T^{-2\theta}}{4(1+\theta)T^{-2\theta} + (4\theta+2)T^{-\theta}}\right] dp_X = p^{\theta+1}AT^{-(\theta+1)}dT_{XY}. \tag{5}$$

Since the expression in square brackets is positive (for $\theta > 0$, as maintained throughout the paper) the change in region X's prices (dp_X) is of the same sign as the change in trade costs ($dT_{XY} = dT_{YX}$). That is, real income in region X rises as trade costs between region X and another region (here, region Y) fall, which is Result 3. This concludes the proof.

C Including Port Cities

The four main port cities of colonial India (Bombay, Calcutta, Karachi and Madras) present a number of circumstances that require them to be handled differently from other regions of India. I describe here how these cities entered my analysis, section by section.

Section 4:

The 133 districts of Northern India that are included in my analysis of salt price data include the relevant Northern port city of Calcutta (Karachi is a Northern port city but it is in Sindh, a province not covered, as described above, in my district-level data).

Section 5:

The 47 trade blocks included in my analysis of bilateral trade patterns include 4 blocks that refer to the four main port cities. These blocks included, as described above, trade data reporting each Indian region's total trade with the given port city with no distinction between whether that trade was with inhabitants of the city or with the wider world (or with the rest of India via so-called coasting sea trade) via the particular port in question. Because of this, the four trade blocks should be interpreted as representing the composite economic activity of two sub-economies: (i) the port city in question, and (ii) the segments of the rest of the

world that choose to trade with India via the port in question.[18] This decision has no bearing on the estimation of θ_k in equation (14) due to the inclusion of importer and exporter fixed effects (separately for each commodity and year). However, when estimating κ in equation (15) the four port city blocks are not included. This is because I lack data on r_{ot} (nominal agricultural output per acre) for cases in which o is a port city block, so I cannot equate the estimate, $\ln \widehat{\beta}^k_{ot} + \theta \ln r_{ot}$, to $\ln A^k_{ot}$. Further, I do not observe the appropriate international and within-port equivalent of $RAIN^k_{ot}$ required to estimate equation (15) for these port city blocks.

Section 6:

The port cities are not included in the regressions in Sections 6 and 7 because, as described above, my analysis is focused on the determinants of real agricultural income, and these four port cities did not report agricultural output data (because so little agriculture was taking place in these city-districts). An exception is the port city of Bombay which was not enumerated as a separate district.

Section 7:

A final challenge presented by the four port cities (which, recall, are treated as a composite of domestic and international economic regions) is that their demand for and supply of goods is likely to have had important effects on the 192 non-port Indian districts in whom my interest lies, and these effects will have changed as interior districts became connected by railroads to the port cities and thence the wider world. For this reason, I compute the equilibrium to the model with 196 separate regions: 192 Indian districts plus four port cities (which include the rest of the world). Doing this requires data on the key exogenous variables—the effective productivities (A^k_{ot}) and land areas (L_o)—from the four port city regions, but such data are unavailable. These exogenous variables are required to compute equation (6) (which depends explicitly on L_o and implicitly, through the definition of π^k_{od}, on A^k_o). To circumvent this data shortage I use the structure of the model to proceed as follows. First, the estimates of $e^{\widehat{\beta}^k_{ot}}$ obtained in Section 5 for the port city regions are equal (up to a scale constant that cannot be identified from the regression in equation (14) that I set equal to one in what follows) to $A^k_{ot} r_{ot}^{-\theta_k}$.[19] Since A^k_{ot} appears in equation (6) only when multiplied by $r_{ot}^{-\theta_k}$ I substitute the estimates of $e^{\widehat{\beta}^k_{ot}}$ (for the four port city blocks only) from Section 5 into equation (6).

[18] In the case of Karachi the block in question actually includes all of the province of Sindh as well as its main port city, Karachi, due to the lack of intra-Sindh trade data.

[19] For years with missing trade data I interpolate $e^{\widehat{\beta}^k_{ot}}$ linearly between non-missing years, and set it equal to the first available value for a block for years prior to that (and to the last available value for years after that).

Second, as an estimate of L_o for port regions I use the fifth percentile of the non-port district size distribution.

Appendix References

Bartholomew, J. G. 1893. *Constable's Hand Atlas of India.* London: A. Constable.

Black, William R. 2003. *Transportation: A Geographical Analysis.* Guilford: The Guilford Press.

Blyn, George. 1966. *Agricultural Trends in India, 1891-1947: Output, Availability, and Productivity.* Philadelphia: University of Pennsylvania Press.

Bourne, John. 1849. *Indian River Navigation.* London: W. H. Allen.

Deloche, Jean. 1994. *Transport and Communications in India Prior to Steam Locomotion: Volume I: Land Transport.* Oxford: Oxford University Press.

Dewey, Clive. 1979. "Patwari and Chaukidar: Subordinate Officials and the Reliability of India's Agricultural Statistics." In *The Imperial Impact in Africa and South Asia.* , ed. Clive Dewey and A.G. Hopkins. London.

Directorate of Economics and Statistics. 1967. *Indian Crop Calendar.* Delhi: Government of India Press.

Heston, Alan. 1973. "Official Yields per Acre in India, 1886-1947: Some Questions of Interpretation." *Indian Economic and Social History Review*, 10(40): 303–334.

Mayer, Thierry, and Soledad Zignago. 2011. "Notes on CEPII's Distances Measures: The GeoDist Database." CEPII Working Paper 2011-25.

Ngo-Duc, T., J. Polcher, and K. Laval. 2005. "A 53-year Forcing Data Set for Land Surface Models." *Journal of Geophysical Research*, 110: 1–13.

Schwartzberg, Joseph E. 1978. *A Historical Atlas of South Asia.* Chicago: University of Chicago Press.

Singh, R.P., and Jayant Banthia. 2004. *India administrative atlas, 1872-2001.* New Delhi: Conroller of Publications.

Watt, George. 1889. *A dictionary of the products of India.* London: J. Murray.