Progress Report for MPhil Thesis

Tilman Graff – University of Oxford

November 28, 2017

Open Questions – What I need to discuss

Topics to be raised in my next supervisor meeting.

Below starts my latest progress report.

---

1 Research Question

Which factors influence the global distribution of trade network optimality?

In this thesis, I aim to create a (potentially) global – in any case African – dataset of trade network efficiency. Taking the spatial distribution of current economic activity and population as given, I use a model from a recent working paper to determine the optimal trade network for each country. I then compare each country's current road network to its optimal one and derive a measure of how far a country is currently away from its ideal self.

In a second step, I will then investigate the origins of this global distribution. Which factors led to the heterogeneities among countries today? Specifically, I will look at:

- Do networks with large colonial infrastructure investments do better or worse today?
- Does tribal favoritism explain why some areas are lacking lucrative investment?

2 Research steps

In order to conduct this research, I will need to follow a series of steps and transfer data between multiple programming softwares. Here is a step-by-step guide, with progress as of November 28, 2017:

- ✓ Find global raster data on population, night-lights, ruggedness, and colonial infrastructure investments.
- ✓ Grid the world on 50x50km squares and aggregate finer-resolution data into those grids.
- ~~Locate the maximum population point within each grid and call this the (population) centroid of the grid.~~ I decided against this additional step. Mostly because I cannot figure out how to do it in QGIS.

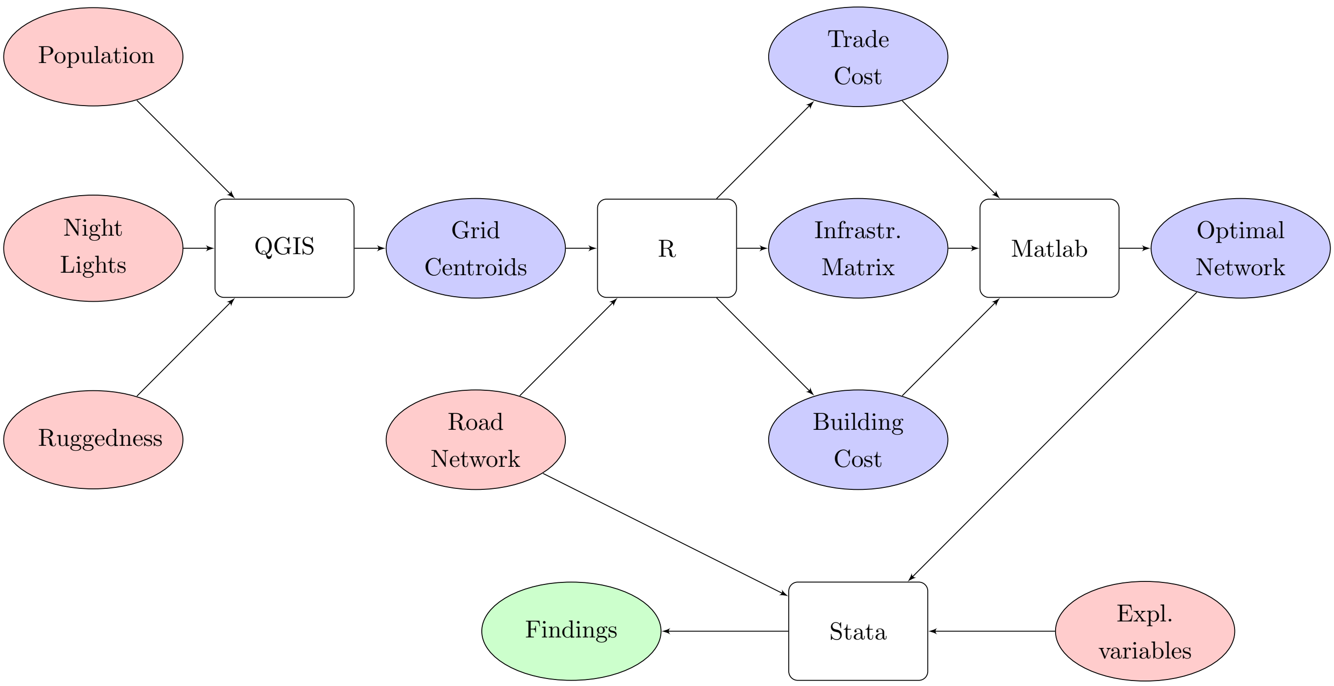
(And it's probably not that important?). Instead, I investigate travel times and distances between unweighted geometric centroids.

- ✓ Use OpenStreetMap to find distance and average speed between neighboring gridcells.
- Use distance and ruggedness to calculate Infrastructure Building Cost Matrix $\delta^I_{i,k}$ for every country.
- Use average speed to calculate current Infrastructure Matrix $I_{i,k}$ for every country.
- Use distance to calculate (iceberg) Trade Constant Matrix $\delta^\tau_{i,k}$ for every country.
- Use $\delta^I_{i,k}$, $\delta^\tau_{i,k}$, and $I_{i,k}$ to find the optimal trade network $I^*_{i,k}$ and optimal tradeflows $Q^*_{i,k}$ for every country. This directly follows the Fajgelbaum and Schaal (2017) Working Paper. I know how to do this, I am however afraid that this might take too long.
- Compare $I_{i,k}$ and $I^*_{i,k}$ for every country to obtain a measure of network optimality ζ_c for every country c.
- Investigate heterogeneity in ζ_c

The following flowchart visualises the process. It shows the path input data (red circles) take through various programming languages (white rectangles) and intermediate datasets (blue circles) into eventual findings (green circle).

3 Notes on individual steps

In this section, I provide a more detailed account of what I did in each of the above-mentioned steps.



3.1 Find global raster data

I use two datasources to create a global raster dataset of relevant variables:

- Data on night lights, ruggedness, land suitability, altitude, malaria index, and precipitation comes from Henderson et al. (2018). These are available on 25km x 25km grids. Most of the data are for the year 2010.
- For census-level spatial population data, I use the Gridded Population of the World (GPW) database from the Socioeconomic Data and Applications Center (2016). These are available on much finer resolution. This database is for the year 2015.

Using GIS, I aggregated these datasets to a 1-by-1 degree global grid (roughly 50km x 50km), following Fajgelbaum and Schaal (2017). In doing so, I take spatial sums of lights and population, and spatial averages of ruggedness, malaria, and weather data.

3.2 Construct Centroids

For each of these gridcells, I calculate the geometric centroid. I do not weigh by population in constructing this centroid, as I have been unable to figure out how to do this. I will need to discuss whether this is ok.

I then crop all global centroids that are not over land. This leaves me with 59,059 gridcells. Doing this, I am aware that I will lose information: I cut gridcells that might be partially over land, as soon as their centroid is not over land. This approach will, however, create equidistant centroid locations, so I'm really just playing one disadvantage against the other.

I then attribute the centroids with the lights, population, and other data from the underlying gridcell. Thus, I act as if all people and all economic activity of a given cell were concentrated on the single centroid point. This is because Fajgelbaum and Schaal's model calculates trade between nodes, not areas. I find this to be a reasonable model simplification.

3.3 Construct Global Road database

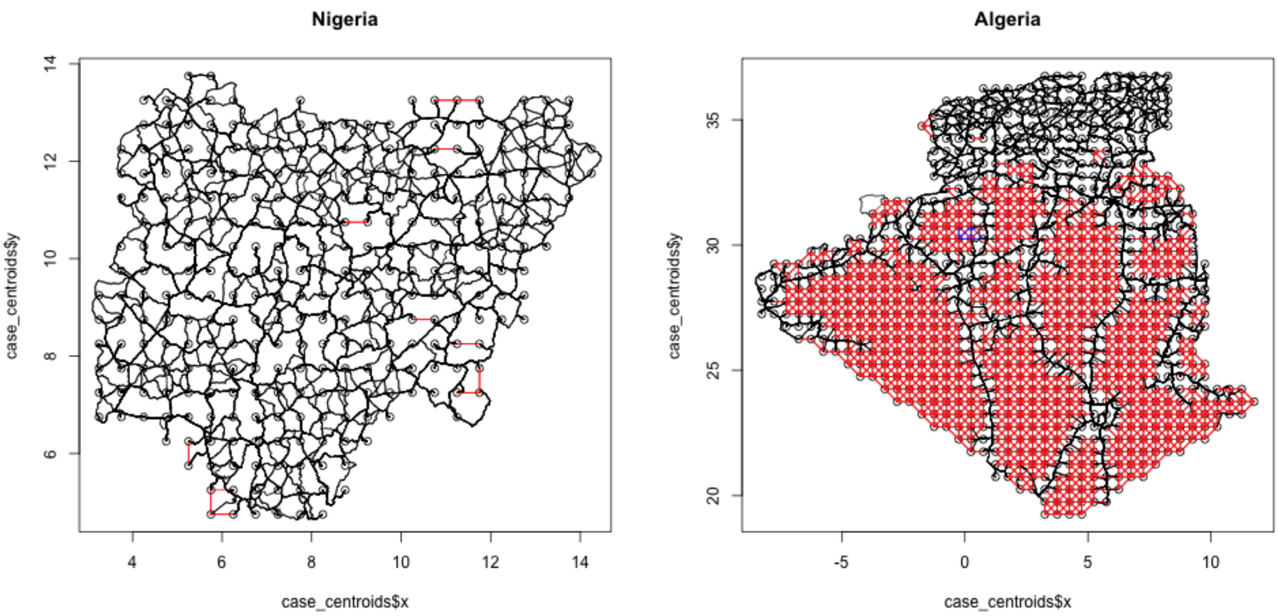
I calculate the optimal route between any centroid and each of its eight surrounding neigbours. In doing so, I rely on the open source online project Open Street Maps (OSM), which is comparable to Google Maps, but allows for unlimited use of its API. I scrape OSM with an R-Package called `osrmRoute`. This package takes start and destination locations, sends them to OSM, and comes back with the optimal route, distance, and speed in virtually no time. I am amazed by how fast this is.

Two problems present itself:

1. OSM's data supply is user generated and hence biased towards more prosperous areas. However, since I mostly care about big highways, I doubt there are all too glaring ommisions. Also, I care about relative inefficiencies within a country, not cross-section differences between countries.
2. I tell OSM to find the optimal route for a car. However, in many remote areas, cars don't get you from A to B. `osrmRoute` then desparately tries to locate the user onto the nearest street (which could be far away!). I hence scrape the entire loc-by-loc route, and calculate the walking distance to the nearest street (and the walking distance from the end of the supplied route to the actual endpoint). I do so by taking direct paths and imply a walking speed of 4 km/h. I then calculate whether walking the

entire distance from A to B is faster (not shorter, but faster!) than the route supplied by `osrmRoute`. If so, I replace the route with the walking route.

Figure 1: Road Networks for different countries as scraped off OSM



(a) Nigeria **(b)** Algeria

Figure 1a displays how this would look like for the country of Nigeria. In black are optimal routes from all 289 centroids to their up to eight closest neighbours. If walking were the preferred option, this direct route is plotted in red. In Nigeria, this is mostly the case in some areas of the (swampy) South and East and some observations in the desert ridden North.

This procedure seems to well capture notions of remoteness: Consider the same graph but for Algeria (Figure 1b). For much of the Saharan parts of the country, walking straight lines through the sand is the best available option. I am actually not too worried about these areas. They will presumably have almost no population or night lights and hence I do not expect the later optimisation to yield unrealistic trans-Saharan highways.

4 Calibration of $I_{i,k}$, $\delta^{\tau}_{i,k}$, $\delta^{I}_{i,k}$

Next up, I use the obtained matrices to construct underlying graph characteristics needed for the model to work. As rough guideline, I follow Fajgelbaum and Schaal (2017).

4.1 Current Infrastructure Network $I_{i,k}$

The matrix $I_{i,k}$ discretises the road network obtained by `osrmRoute`. Basically, it attaches a value to each link between nodes, indicating "How much infrastructure" has been built into the edge. Fajgelbaum and Schaal (2017) do this by calculating the mean number of lanes a road has plus adding a dummy for national roads. I cannot do this, as I don't have comprehensive data on road lanes. However, I believe that the derived speed matrix from `osrmRoute` actually serves as the much more significant statistic for what Fajgelbaum and Schaal want to capture. I hence propose

$$I_{i,k} = \text{Average Speed}_{i,k} \tag{1}$$

This is obviously a very simple calibration and I am open to discuss more flexible approaches. But the general logic is clear: if a certain edge allows a car to drive faster on it, chances are that more investments had been made into it. In a sense, this captures the same notion Fajgelbaum and Schaal with their calibration based on lanes and national roads. Figure 2 shows a distribution of the average speeds (on travelled routes, in km/h) for Namibia. As can be seen, the distribution follows a nicely behaved shape, with an additional hump at 4 km/h for purely walked routes (see above).

4.2 Infrastructure Building Cost Matrix $\delta^I_{i,k}$

I am still working on this one. The entries to this matrix represent the cost to building new infrastructure (think: one additional km/h to stay in the logic from above) onto a given edge from i to k. Fajgelbaum and Schaal use

$$\ln(\frac{\delta^I_{i,k}}{dist_{i,k}}) = \ln(\delta^I_0) - 0.11 * (dist_{i,k} > 50km) + 0.12 * \ln(\text{ruggedness}_{i,k}) \tag{2}$$

which comes from Collier et al. (2015). δ^I_0 is individual for every country. I have all the data to replicate this, but was thinking of including measures on extra low soil quality (which is congruent to Saharan soil) and (maybe) the Malaria index in there (or is that too much a colonial, *"white guys walk in, build a road, and get Malaria"*, kind of thinking?)

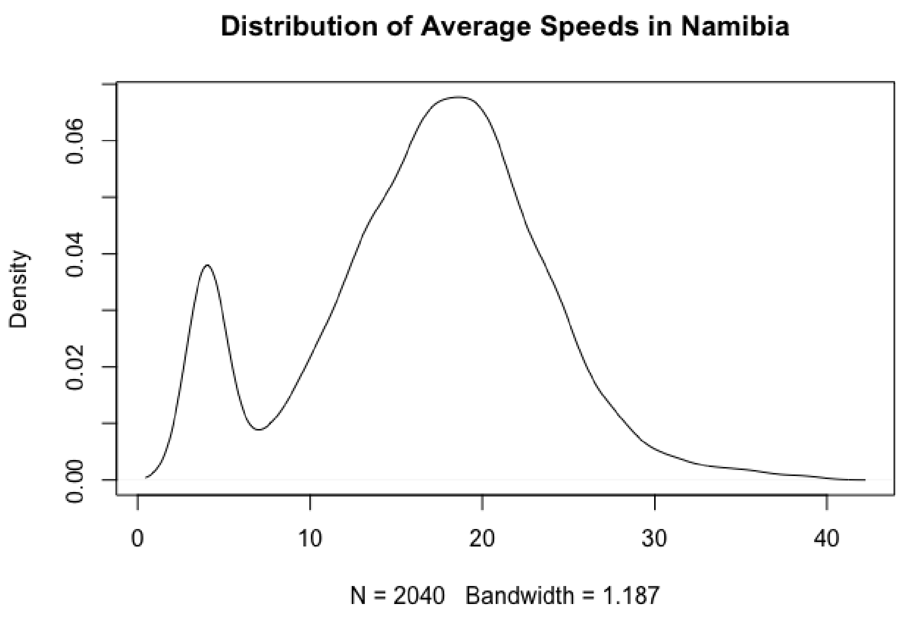


Figure 2: Average Speed for Namibia, in km/h, 288 centroid locations

4.3 Iceberg Trade Constant Matrix $\delta^{\tau}_{i,k}$

I follow Fajgelbaum and Schaal in assuming

$$\delta^{\tau}_{i,k} = \delta^{\tau}_{0} * dist_{i,k}^{\delta^{\tau}_{1}} \tag{3}$$

where δ^{τ}_{0} comes from a calibration using Spanish data and is reported – and $\delta^{\tau}_{1} = 0$. I assume I can go ahead and use this replication.

The actual iceberg trade costs are then computed as

$$\tau_{i,k}(Q_{i,k}, I_{i,k}) = \delta^{\tau}_{j,k} * \frac{Q^{\beta}_{i,k}}{I^{\gamma}_{i,k}} \tag{4}$$

where $Q_{i,k}$ represents trade flows between nodes (to capture notion of congestion), $I_{i,k}$ is the existing Infrastructure Network (derived above), $\beta = 1$ and $\gamma = 1$ in the easiest calibration.

5 Next steps

1. Decide on a stance regarding the questions raised above (1-2 days, possibly longer depending on how much to change)
2. Calculate $I_{i,k}$, $\delta^{\tau}_{i,k}$, $\delta^{I}_{i,k}$ (2 hours)
3. Optimise Matlab Code for better performance, or find better performance elsewhere (1 week)
4. Perform Network optimisations (1-2 weeks)

Immediate goal: Have optimal networks calculated by January 1, 2018.

Past supervisor meetings; topics raised and answered

November 28, 2018

- Is it ok to use geometric centroids as opposed to population-weighted centroids? *Yes, it's fine. Ferdinand did this in Maurer et al. (2017).*
- Night Lights: do I have to translate them into GDP first? Fajgelbaum and Schaal (2017) use G-Econ, but that's only available for the year 1990 and in much coarser resolution. *Tough. But probably go for it just because it's easiest and no direct alternative comes to mind. Keep it at its easiest level. Kocornik-Mina et al. (2015) do use pure lights and cite Henderson et al. (2012) who find a Lights-GDP elasticity of 1.*
- Can I proceed to use Open Street Maps even though it has disadvantages as described below? *Yes.*

- Computational challenges looming. As soon as I'll start the Matlab bit, I will need either patience or better computing power. Does the department / chair grant access to remote desktops etc.? *Probably not.*
- Average speed a good proxy for Infrastructure Matrix $I_{i,k}$? *Not really many other options.*
- The paper by Fajgelbaum and Schaal (2017) has been criticised for its heavy dependence on the congestion assumption (in words, that iceberg trade costs rely on current tradeflows, or that $\tau_{i,k}(Q_{i,k}, I_{i,k})$ is a function of $Q_{i,k}$); namely by Allen and Arkolakis (2016). I think this criticism is overblown. Still worth keeping in mind. *Definitely cite and discuss this. But should be fine.*

References

Allen, Treb and Costas Arkolakis (2016). The Welfare Effects of Transportation Infrastructure Improvements. Working paper, Dartmouth and Yale

Collier, Paul, Martina Kirchberger, and Mans Söderbom (2015). The Cost of Road Infrastructure in Low-and Middle-Income Countries. *The World bank economic review* 30(3), pp. 522–548

Fajgelbaum, Pablo D. and Edouard Schaal (2017). Optimal Transport Networks in Spatial Equilibrium. Working paper, National Bureau of Economic Research

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

Henderson, Vernon, Tim Squires, Adam Storeygard, and David N Weil (2018). The Global Spatial Distribution of Economic Activity: Nature, History, and the Role of Trade. *The Quarterly Journal of Economics* forthcoming

Kocornik-Mina, Adriana, Thomas KJ McDermott, Guy Michaels, and Ferdinand Rauch (2015). Flooded Cities. Working paper

Maurer, Stephan, Jörn-Steffen Pischke, and Ferdinand Rauch (2017). Of Mice and Merchants: Trade and Growth in the Iron Age. Working paper

Socioeconomic Data and Applications Center (2016). Gridded Population of the World (GPW), v4 — SEDAC. `http://sedac.ciesin.columbia.edu/data/collection/gpw-v4`