# The view from space: Theory-based time-varying distances in the gravity model


Julian Hinz
- Paris School of Economics -



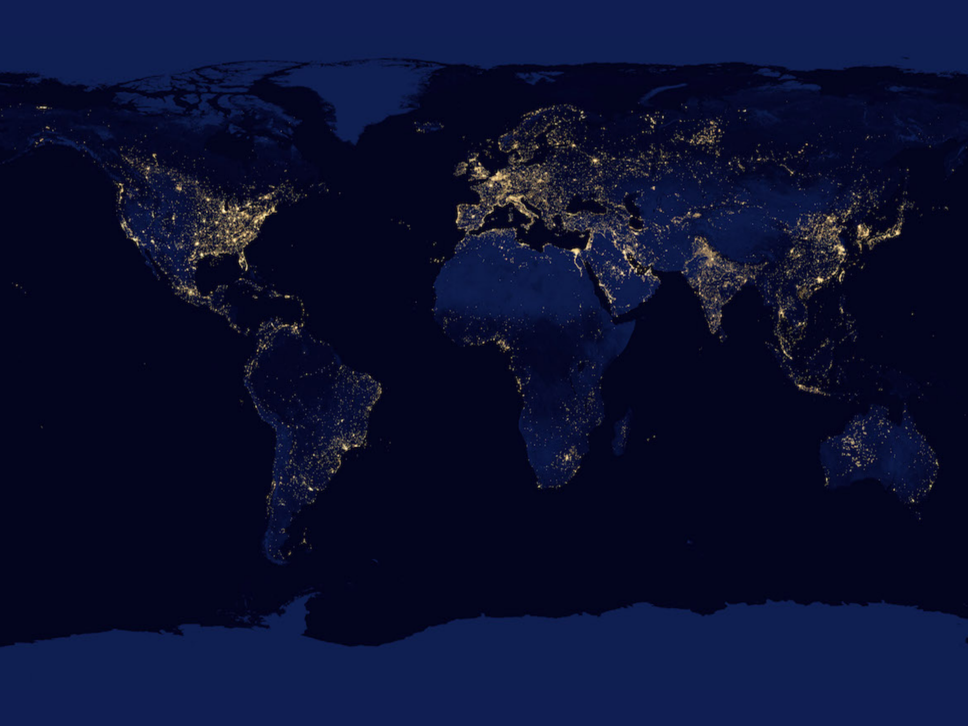
Presentation at the 8th FIW Research Conference
4 December 2015

# Idea

- Trade data mostly aggregated to some geographic entity
  $\rightarrow$ need for aggregated trade costs
- People move, countries develop, borders change
  $\rightarrow$ variation of geography over time
- Nightlight data: shows where people live, produce, trade
  $\rightarrow$ good proxy for economic activity, high resolution

# Contributions

- Theory-based trade cost aggregation
  $\rightarrow$ agnostic to underlying gravity framework
  $\rightarrow$ trade cost elasticity as parameter in generalized mean
- Computation of aggregate distances using annual nightlight data
  $\rightarrow$ vast improvement in detail and coverage
  $\rightarrow$ variation over time, between and within countries
- Iterative estimation of gravity equation
  $\rightarrow$ distance coef. around $-1$ when estimated in within dim.
  $\rightarrow$ border coef. $18-52\%$ lower with harmonic mean distance
  $\rightarrow$ neighbor coefficient similarly affected




1992


1993


1994


1995


1996


1997


1998


1999


2000


2001


2002


2003


2004


2005


2006


2007


2008


2009


2010


2011


2012


1992

# Roadmap

- Distance (and borders) in the literature
- Theory-based trade cost aggregation
- Aggregate distances with nightlight data
- Iterative estimation of gravity equation
- Effect on common gravity variables

# Distances in the gravity model

- Since Tinbergen *et al.* (1962): distance matters
- "Alive and well": Disdier & Head (2008), Head & Mayer (2014)
- Distance effect in within-dimension: Feyrer (2009) and Hugot & Umana Dajud (2014)
- Distance and borders: Helliwell (2000), Helliwell & Verdier (2001) and Head & Mayer (2009) note importance of internal distance
  $\rightarrow$ (almost) no border puzzle with micro data: Hillberry & Hummels (2008)

# "Effective" distances

Helliwell (2000) and Head & Mayer (2009) show that border puzzle could be driven by "wrong" distances:

- Central point (or capital) distance vs. effective distance
- Internal distance as $c \cdot \sqrt{\text{area}/\pi}$ systematically too low
- Better distances with weighted mean:

$$d_{ij} = \left( \sum_{k \in i} (\text{pop}_k / \text{pop}_i) \sum_{l \in j} (\text{pop}_l / \text{pop}_j)\, d_{kl}^{\theta} \right)^{1/\theta}$$

# GeoDist distances

Mayer & Zignago (2011) provide effective distances

- Up to 25 largest cities, geocoded, weighted by population size
- Data (mostly) from 2004 from world-gazetteer.com (UN data)
- Internal distances for countries with only one city simple distances $\sqrt{\text{area}/\pi}$

# Theory - Initial setup

- Trade cost aggregation à la Head & Mayer (2009)
- Assume generic structural gravity equation:

$$x_{kl} = G s_k m_l \phi_{kl}^{\theta}$$

- $x_{kl}$ is trade flow from a location $k$ to another location $l$
- $s_k$ is exporter-specific term, $m_l$ importer-specific term
- $\phi_{kl}$ is the bilateral resistance term
- $\theta$ is the trade elasticity

Let $k$ now be a location in the geographic entity $i$ and $l$ in $j$. Then

$$\begin{aligned} x_{ij} &= \sum_{k\in i}\sum_{l\in j} x_{kl} \\ &= G\sum_{k\in i} s_k \sum_{l\in j} m_l \phi_{kl}^{\theta} \end{aligned}$$

Calling $m_j = \sum_{l\in j} m_l$ and $s_i = \sum_{k\in i} s_k$

$$x_{ij} = G s_i m_j \sum_{k\in i}\sum_{l\in j} \frac{s_k}{s_i}\frac{m_l}{m_j}\phi_{kl}^{\theta}$$

Repackaging the right term then yields the gravity equation for geographic entities

$$x_{ij} = G s_i m_j \phi_{ij}^{\theta}$$

where trade costs are aggregated as

$$\phi_{ij} = \left( \sum_{k \in i} \sum_{l \in j} \frac{s_k}{s_i} \frac{m_l}{m_j} \phi_{kl}^{\theta} \right)^{1/\theta}$$

# Location- and entity-specific trade costs

Let now $\phi$ be described by the function

$$\phi_{kl} = \psi_{ij}^{\epsilon} \chi_{kl}^{\delta}$$

- *Location-specific* component $\chi_{kl}$
  $\rightarrow$ distance between the two locations
- *Entity-specific* component $\psi_{kl} = \psi_{ij} \; \forall \; k \in i, \; l \in j$
  $\rightarrow$ e.g. common official language of the two entities
- $\delta$ is elasticity of trade costs to location-specific trade costs
- $\epsilon$ is elasticity to entity-specific trade costs

# Entity-level gravity equation

Entity-level gravity equation can then be rewritten as

$$x_{ij} = G s_i m_j \psi_{ij}^{\theta\epsilon} \chi_{ij}^{\theta\delta}$$

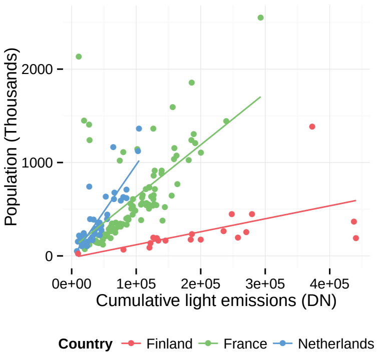
where location-specific trade costs are aggregated as

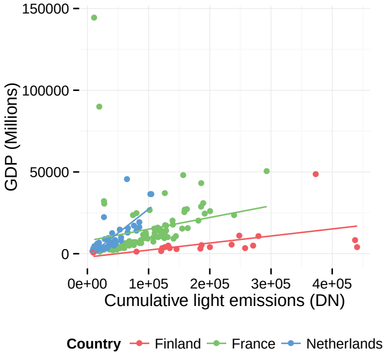
$$\chi_{ij} = \left( \sum_{k \in i} \sum_{l \in j} \frac{s_k}{s_i} \frac{m_l}{m_j} \chi_{kl}^{\theta\delta} \right)^{1/\theta\delta}$$

# Nighttime light data

- Map of human activity (Croft, 1973)
- Used in Economic research before (e.g. Henderson *et al.* (2011), Alesina *et al.* (2012))
- Satellite circling earth about 14 times in 24h, image captured between 8:30pm and 10pm local time
- Since 1972, digitally available since 1992
- Resolution of 30 arc-seconds, about 860m at the equator
  $\rightarrow$ $725,820,001$ pixels, about 60,000,000 illuminated
  $\rightarrow$ not true radiance, DN between 0 and 63

# Light emissions, GDP and population



# Computing aggregate distances

Recall location-specific aggregation:

$$\chi_{ij} = \left( \sum_{k \in i} \sum_{l \in j} \frac{s_k}{s_i} \frac{m_l}{m_j} \chi_{kl}^{\theta\delta} \right)^{1/\theta\delta}$$

Expressed in matrix notation for distances

$$d_{ij} = \left( \mathrm{w}_i^T \mathrm{D}_{ij}^{\theta\delta} \mathrm{w}_j \right)^{1/\theta\delta} \qquad \text{where}$$

$$\mathrm{w}_i = \frac{1}{\sum_{k \in i} \mathrm{DN}_k} \begin{pmatrix} \mathrm{DN}_1 \\ \vdots \\ \mathrm{DN}_k \end{pmatrix} \qquad \text{and } w_j \text{ accordingly.}$$

# Computing aggregate distances

- Calculate great circle distances (about $60,000,000^2$)

$$\mathrm{D}_{ij} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,l} \\ \vdots & \ddots & \vdots \\ d_{k,1} & \cdots & d_{k,l}, \end{pmatrix} \quad k \in i,\ l \in j$$

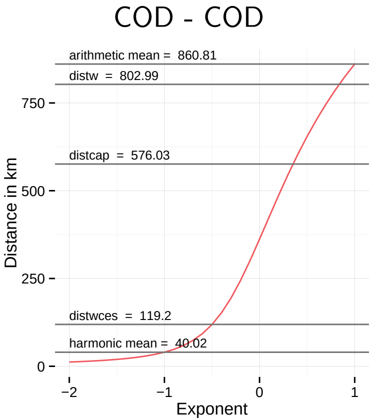
$\rightarrow$ sampling 1 percent and min. 1000 per country, 100 draws

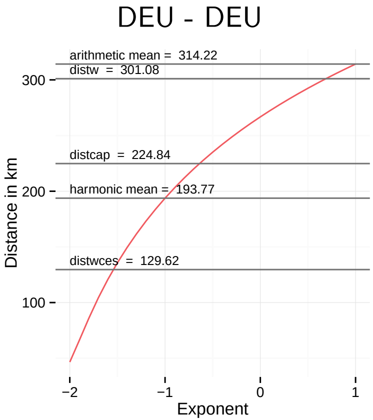
- Exponent $\theta\delta$ a priori unknown:

$$d_{ij} = \left( \mathrm{w}_i^T \mathrm{D}_{ij}^{\theta\delta} \mathrm{w}_j \right)^{1/\theta\delta}$$

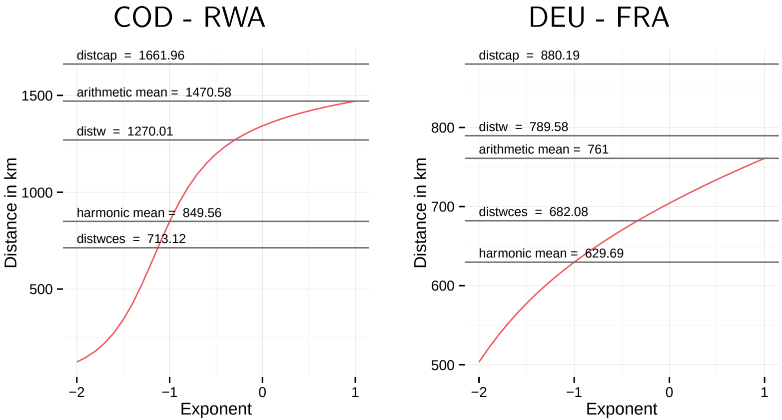
- Very important for aggregate distance
  $\rightarrow$ can be estimated in gravity equation!

# Distance variation by exponent

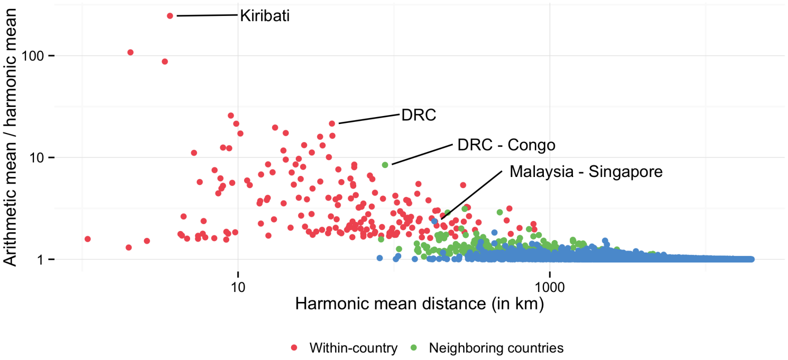


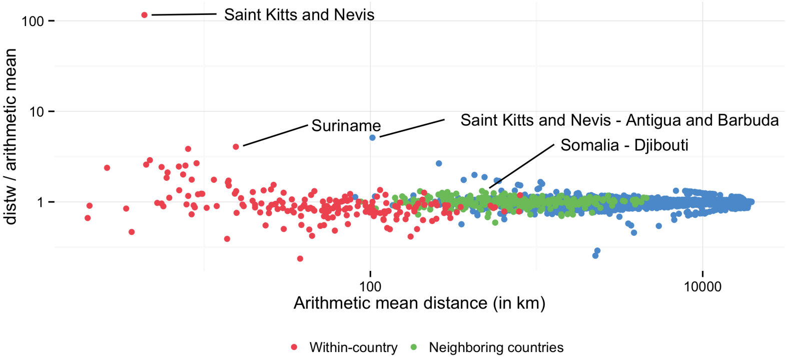
# Distance variation by exponent



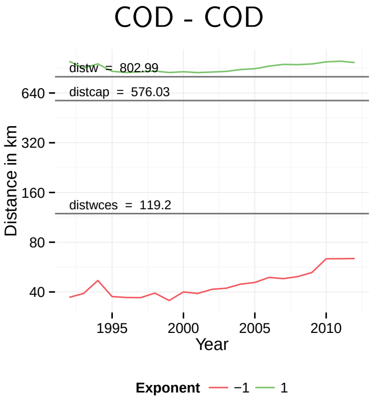
# Bias of arithmetic over harmonic mean

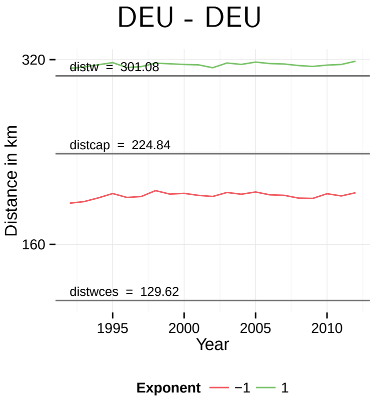


# Measurement error of Mayer & Zignago (2011)'s *distw*

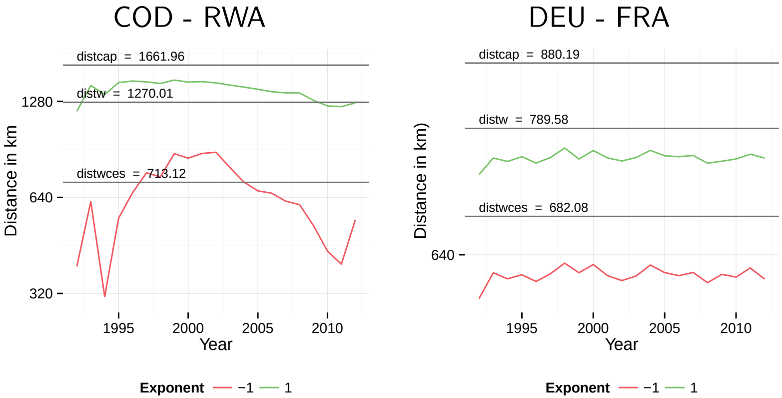


# Distance variation over time



# Distance variation over time



# Estimation of gravity equation

Recall that

$$\chi_{ij} = \left( \sum_{k \in i} \sum_{l \in j} \frac{s_k}{s_i} \frac{m_l}{m_j} \chi_{kl}^{\theta\delta} \right)^{1/\theta\delta}$$

where the exponent $\theta\delta$ is a parameter in the gravity equation

$$x_{ij} = G s_i m_j \psi_{ij}^{\theta\epsilon} \chi_{ij}^{\theta\delta}.$$

# Estimation of gravity equation

Easily estimated with OLS as

$$\log X_{ij} = \alpha_0 + S_i + M_j + \alpha_1 \cdot \mathrm{Controls}_{ij} + \\ \beta_0 \cdot \mathrm{Border}_{ij} + \beta_1 \cdot \log \mathrm{Distance}_{ij} + \epsilon_{ij}$$

or using PPML

$$X_{ij} = \exp\left(\alpha_0 + S_i + M_j + \alpha_1 \cdot \mathrm{Controls}_{ij} + \\ \beta_0 \cdot \mathrm{Border}_{ij} + \beta_1 \cdot \mathrm{Distance}_{ij}\right) + \epsilon_{ij}.$$

$\rightarrow$ Coefficient of interest is $\widehat{\beta}_1 = \widehat{\theta\delta}$

# Iterative estimation

- Iterative estimation: initial value of $\theta\delta$, estimate, update
  $\rightarrow$ until convergence (i.e. no change in 5th digit)
- Time variation of distance allows estimation of distance coefficient in within-dimension of panel
- Multiple samples and data sources: IMF DOTS, UN Comtrade, TradeProd (De Sousa *et al.* , 2012)
  $\rightarrow$ TradeProd has internal flows
- Gravity controls from CEPII and De Sousa (2012)

# Distance coefficient

| | Dependent variable: *log*(flow) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *log*(Distance) | $-1.282^{***}$ | $-1.264^{***}$ | $-1.260^{***}$ | $-0.407^{***}$ | $-0.950^{***}$ | $-0.927^{***}$ |
| | (0.007) | (0.006) | (0.006) | (0.125) | (0.100) | (0.100) |
| Distance | arithmetic | harmonic | iterate | arithmetic | harmonic | iterate |
| Pair FE | No | No | No | Yes | Yes | Yes |
| No. of Iterations | - | - | 4 | - | - | 12 |
| Observations | 177,996 | 177,996 | 177,996 | 177,996 | 177,996 | 177,996 |
| $R^2$ | 0.785 | 0.787 | 0.787 | 0.925 | 0.925 | 0.925 |
| Adjusted $R^2$ | 0.776 | 0.778 | 0.778 | 0.918 | 0.918 | 0.918 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | |

# Distance coefficient: Results

- Distance coefficient close to $-1$ in preferred specification
  $\rightarrow$ suggests use harmonic mean distances
- Now: estimate effect of using harmonic mean distance on other variables in cross section
  $\rightarrow$ captured by fixed effect in previous estimation
- Hypotheses:
  - strong effect on variables correlated with distance: border, neighbor
  - little effect on other variables

# Border effect in 2000

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | log(flow) | | flow | |
| | (1) | (2) | (3) | (4) |
| log(distance) | $-1.530^{***}$ | $-1.464^{***}$ | $-0.886^{***}$ | $-0.813^{***}$ |
| | (0.033) | (0.031) | (0.025) | (0.018) |
| border | $1.959^{***}$ | $0.956^{***}$ | $2.091^{***}$ | $1.728^{***}$ |
| | (0.202) | (0.212) | (0.053) | (0.050) |
| Estimator | OLS | OLS | PPML | PPML |
| Distance | arithmetic | harmonic | arithmetic | harmonic |
| Observations | 4,220 | 4,220 | 4,220 | 4,220 |
| $R^2$ | 0.856 | 0.856 | | |
| Adjusted $R^2$ | 0.848 | 0.849 | | |
| *Note:* | | | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ | |

# Effect on other variables in 2000

| Dependent variable: | log(flow) | | flow | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| neighbor | 0.266** | 0.099 | 0.372*** | 0.355*** |
| | (0.103) | (0.105) | (0.024) | (0.024) |
| rta | 0.539*** | 0.550*** | 0.871*** | 0.903*** |
| | (0.063) | (0.063) | (0.033) | (0.033) |
| comcur | −0.061 | −0.083 | −0.088*** | −0.119*** |
| | (0.138) | (0.138) | (0.030) | (0.030) |
| colony | 0.808*** | 0.805*** | −0.018 | 0.004 |
| | (0.093) | (0.093) | (0.027) | (0.027) |
| comlang off | 0.485*** | 0.488*** | 0.143*** | 0.115*** |
| | (0.054) | (0.054) | (0.027) | (0.027) |
| Estimator | OLS | OLS | PPML | PPML |
| Distance | arithmetic | harmonic | arithmetic | harmonic |
| Observations | 8,811 | 8,811 | 8,811 | 8,811 |
| $R^2$ | 0.805 | 0.804 | | |
| Adjusted $R^2$ | 0.797 | 0.797 | | |

*Note:* *p<0.1; **p<0.05; ***p<0.01

# Effects with simulated data

- Suppose that

$$\phi_{ij} = \exp(-\ln \mathrm{Dist}_{ij})$$

where $\mathrm{Dist}_{ij}$ is *harmonic mean* distance

- Generate data with simple structural model à la Head & Mayer (2014)

$$X_{ij} = \frac{Y_i}{\Omega_i} \cdot \frac{X_j}{\Phi_j} \cdot \phi_{ij}$$

where

$$\Omega_i = \sum_k \frac{X_k \phi_{ik}}{\Phi_k} \quad \text{and} \quad \Phi_j = \sum_k \frac{Y_k \phi_{ik}}{\Omega_k}$$

# Effects with simulated data

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | log(trade_flow) | | | | trade_flow | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| log(distance_harm) | $-1.000^{***}$ | | | | $-1.000^{***}$ | |
| | (0.000) | | | | (0.000) | |
| log(distance_arith) | | $-1.075^{***}$ | $-1.028^{***}$ | $-1.015^{***}$ | | $-1.077^{***}$ |
| | | (0.001) | (0.001) | (0.001) | | (0.001) |
| border | $0.000^{***}$ | | $1.136^{***}$ | $1.185^{***}$ | $-0.000^{**}$ | $0.541^{***}$ |
| | (0.000) | | (0.006) | (0.006) | (0.000) | (0.003) |
| neighbor | $0.000^{***}$ | | | $0.171^{***}$ | 0.000 | $0.108^{***}$ |
| | (0.000) | | | (0.003) | (0.000) | (0.002) |
| Estimator | OLS | OLS | OLS | OLS | PPML | PPML |
| Observations | 32,041 | 32,041 | 32,041 | 32,041 | 32,041 | 32,041 |
| $R^2$ | 1.000 | 0.999 | 1.000 | 1.000 | | |
| Adjusted $R^2$ | 1.000 | 0.999 | 1.000 | 1.000 | | |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

# Conclusion

- Theory-based trade cost aggregation calls for general mean with trade elasticity as exponent
- Distances computed using annual nightlight data allow estimation of distance coefficient in within-dimension $\rightarrow$ coefficient of $-1$ suggest harmonic mean distances
- Border and neighbor effect reduced by 18 to 52% when using harmonic mean distances
- Magnitude of effects is supported by regression on simulated data
- Next: evaluate results on lower levels of aggregation

Thank you for your attention!

Julian.Hinz@univ-paris1.fr

Raw image (Source: NOAA)

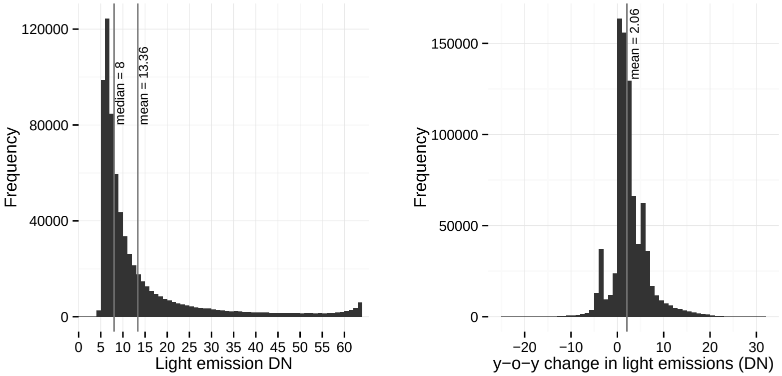
# Processing the data

- Remove artifacts (gas flares), boats, etc.
  $\rightarrow$ fires already removed, only non-cloud days
- Detect borders with data from Weidmann *et al.* (2010)
- Cross-year intercalibration (from NOAA/Elvidge *et al.* (2014)):

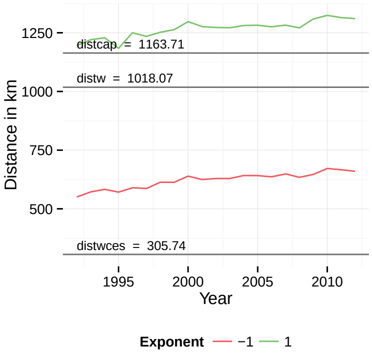
$$\mathrm{DN}^{'} = c_0 + c_1\mathrm{DN} + c_2\mathrm{DN}^2$$

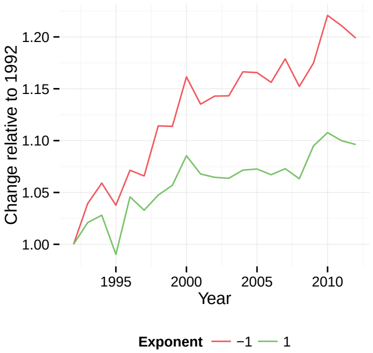
- Average for multi-satellite years

# Light distribution and y-o-y changes



# Internal distance of China

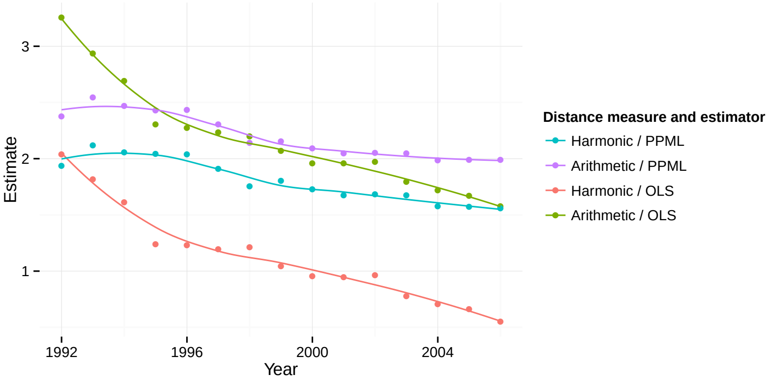


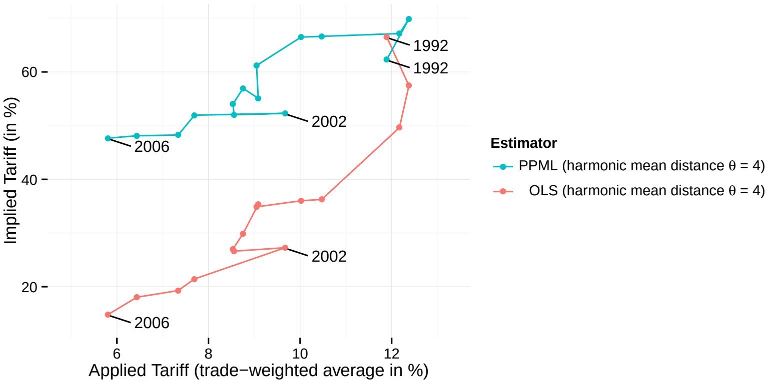
# Distance coefficient: Robustness

| | Dependent variable: *log*(flow) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *log*(Distance) | $-1.019^{***}$ | $-1.177^{***}$ | $-0.615^{*}$ | $-0.860^{***}$ | $-1.399^{***}$ | $-0.771^{**}$ |
| | (0.098) | (0.208) | (0.322) | (0.282) | (0.288) | (0.323) |
| Distance | iterate | iterate | iterate | iterate | iterate | iterate |
| Pair FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Dataset | DOTS | DOTS | DOTS | DOTS | COMTRADE | TradeProd |
| Sample | Neighbors | External | High inc. | Low inc. | all | all |
| No. of Iterations | 14 | 6 | 21 | 13 | 6 | 27 |
| Observations | 30,429 | 175,140 | 31,395 | 2,646 | 87,969 | 132,795 |
| $R^2$ | 0.971 | 0.919 | 0.959 | 0.967 | 0.929 | 0.927 |
| Adjusted $R^2$ | 0.961 | 0.911 | 0.954 | 0.934 | 0.921 | 0.918 |
| *Note:* | | | | | | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ |

# Border effect over time



# Border effect over time: Implied vs applied tariffs



# Neighbor effect over time