State variables and dynamics of economic complexity, and their connection to economic performance

Workshop on Economic Complexity and Development
December 9-10, 2019
Curitiba, Brazil

Andres Gomez-Lievano
Center for International Development,
Harvard University

Contact info:
Email: andres_gomez@hks.harvard.edu
Twitter: @GomezLievano

*"I can calculate the motion of heavenly bodies, but not **the madness of people**"*

SHŌSAI IKKEI 昇斎 景 (active ca. 1870)
Kaika Injun: The New Battles the Old, 1873, triptych, 33.75" x 19.25"

Overarching questions

- Do socio-economic systems have **state variables**?
- If ‘yes’, do economic state variables have **predictable dynamics**?
- If ‘yes’, how are state variables and their dynamics related to observable metrics of **socio-economic performance**?

Organizing ideas

- The performance of an economy is a function of the *size* of the accumulated *collective knowhow*

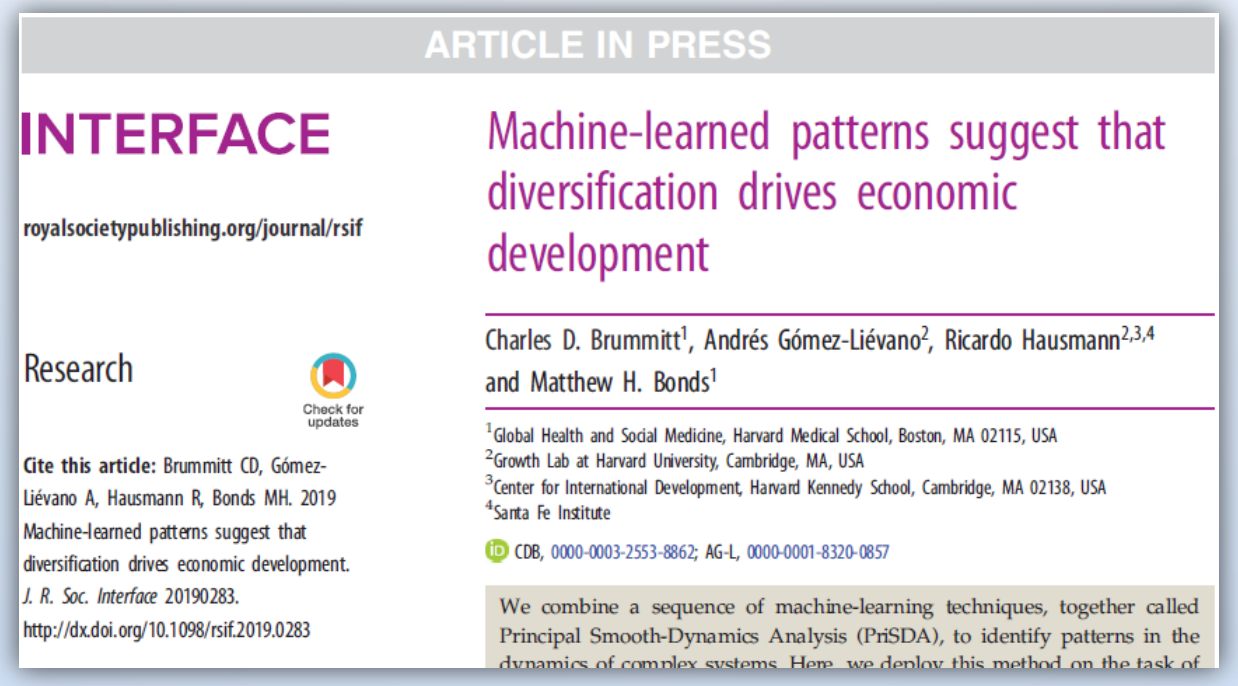
- The development of an economy is given by the dynamics of *collective learning*

Overview

- Paper # 1: (20 mins)
  "Machine-learned patterns suggest that diversification drives economic development"
- Paper # 2: (20 mins)
  "The drivers of urban economic complexity and their connection to urban economic performance"

- Looking forward: (5 mins)
  General lessons and intriguing questions

PAPER # 1



ARTICLE IN PRESS

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Check for updates

Cite this article: Brummitt CD, Gómez-Liévano A, Hausmann R, Bonds MH. 2019 Machine-learned patterns suggest that diversification drives economic development. *J. R. Soc. Interface* 20190283.
http://dx.doi.org/10.1098/rsif.2019.0283

Machine-learned patterns suggest that diversification drives economic development

Charles D. Brummitt[1], Andrés Gómez-Liévano[2], Ricardo Hausmann[2,3,4] and Matthew H. Bonds[1]

[1]Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA
[2]Growth Lab at Harvard University, Cambridge, MA, USA
[3]Center for International Development, Harvard Kennedy School, Cambridge, MA 02138, USA
[4]Santa Fe Institute

CDB, 0000-0003-2553-8862; AG-L, 0000-0001-8320-0857

We combine a sequence of machine-learning techniques, together called Principal Smooth-Dynamics Analysis (PriSDA), to identify patterns in the dynamics of complex systems. Here, we deploy this method on the task of

Questions

- Can a “machine” discover dynamical laws in economic data?

- Do “paths/sequences/progressions” in economic development literally exist?

Motivating literature # 1

Sir Isaac

nature COMMUNICATIONS

Article | OPEN | Published: 21 August 2015

Automated adaptive inference of phenomenological dynamical models

Bryan C. Daniels & Ilya Nemenman

Nature Communications **6**, Article number: 8133 (2015) | Download Citation

Journal of Machine Learning Research 19 (2018) 1-24 Submitted 1/18; Revised 7/18; Published 7/18

Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations

Maziar Raissi MAZIAR_RAISSI@BROWN.EDU
Division of Applied Mathematics
Brown University
Providence, RI, 02912, USA

Editor: Manfred Opper

Abstract

We put forth a deep learning approach for discovering nonlinear partial differential equations from scattered and potentially noisy observations in space and time. Specifically, we approximate the unknown solution as well as the nonlinear dynamics by two deep neural networks. The first network acts as a prior on the unknown solution and essentially enables us to avoid numerical differentiations which are inherently ill-conditioned and unstable. The second network represents the nonlinear dynamics and helps us distill the mechanisms that govern the evolution of a given spatiotemporal data-set. We test the effectiveness of our approach for several benchmark problems spanning a number of scientific domains and demonstrate how the proposed framework can help us accurately learn the underlying dynamics and forecast future states of the system. In particular, we study the Burgers', Korteweg-de Vries (KdV), Kuramoto-Sivashinsky, nonlinear Schrödinger, and Navier-Stokes equations.

Keywords: Systems Identification, Data-driven Scientific Discovery, Physics Informed Machine Learning, Predictive Modeling, Nonlinear Dynamics, Big Data

SINDy

Discovering governing equations from data by sparse identification of nonlinear dynamical systems

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz

PNAS April 12, 2016 113 (15) 3932-3937; published ahead of print March 28, 2016 https://doi-org.ezp-prod1.hul.harvard.edu/10.1073/pnas.1517384113

Edited by William Bialek, Princeton University, Princeton, NJ, and approved March 1, 2016 (received for review August 31, 2015)

RESEARCH ARTICLE | APPLIED MATHEMATICS

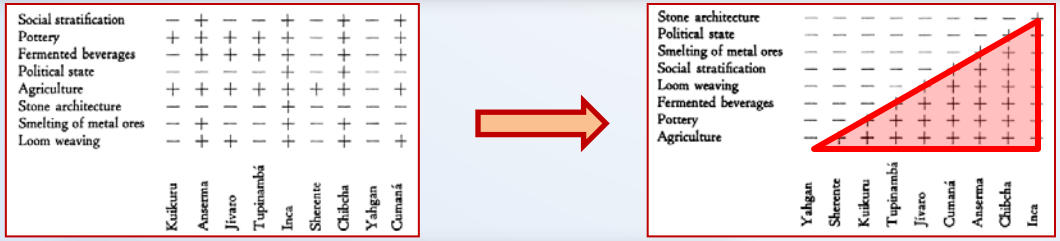
Data-driven discovery of partial differential equations

Samuel H. Rudy[1,*], Steven L. Brunton[2], Joshua L. Proctor[3] and J. Nathan Kutz[1]
+ See all authors and affiliations

Science Advances 2 Apr 2017:
Vol. 3, no. 4, e1602614
DOI: 10.1126/sciadv.1602614

Motivating literature # 2

Scale Analysis as an Instrument for the Study of Cultural Evolution
Author(s): Robert L. Carneiro
Source: *Southwestern Journal of Anthropology*, Vol. 18, No. 2 (Summer, 1962), pp. 149-169
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/3629014
Accessed: 06-04-2016 22:50 UTC



Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization

Peter Turchin, Thomas E. Currie, Harvey Whitehouse, Pieter François, Kevin Feeney, Daniel Mullins, Daniel Hoyer, Christina Collins, Stephanie Grohmann, Patrick Savage, Gavin Mendel-Gleason, Edward Turner, Agathe Dupeyron, Enrico Cioni, Jenny Reddish, Jill Levine, Greine Jordan, Eva Brandl, Alice Williams, Rudolf Cesaretti, Marta Krueger, Alessandro Ceccarelli, Joe Figliulo-Rosswurm, Po-Ju Tuan, Peter Peregrine, Arkadiusz Marciniak, Johannes Preiser-Kapeller, Nikolay Kradin, Andrey Korotayev, Alessio Palmisano, David Baker, Julye Bidmead, Peter Bol, David Christian, Connie Cook, Alan Covey, Gary Feinman, Árni Daníel Júlíusson, Axel Kristinsson, John Miksic, Ruth Mostern, Cameron Petrie, Peter Rudiak-Gould, Barend ter Haar, Vesna Wallace, Victor Mair, Liye Xie, John Baines, Elizabeth Bridges, Joseph Manning, Bruce Lockhart, Amy Bogaard, and Charles Spencer

PNAS January 9, 2018 115 (2) E144-E151; first published December 21, 2017 https://doi.org/10.1073/pnas.1708800115

Contributed by Charles Spencer, November 16, 2017 (sent for review May 26, 2017; reviewed by Simon A. Levin and Charles Stanish)

Human societies across the globe and over 10,000 years differ mainly in their "social complexity"

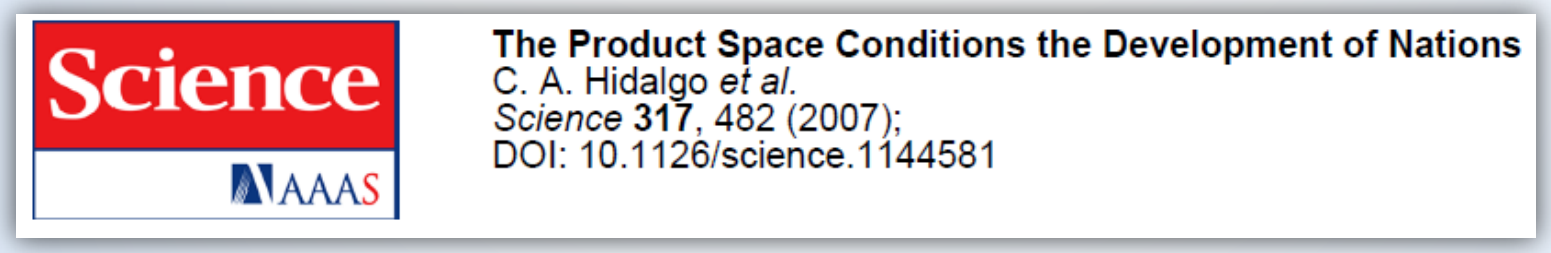
Material wealth in 3D: Mapping multiple paths to prosperity in low- and middle- income countries

Daniel J. Hruschka, Craig Hadley, Joseph Hackman

Published: September 8, 2017 • https://doi.org/10.1371/journal.pone.0184616

Development is multidimensional

Motivating literature # 3



Economic diversification is constrained

"Economic complexity" can be summarized in a single quantity

Economic development differs across locations in a phase space

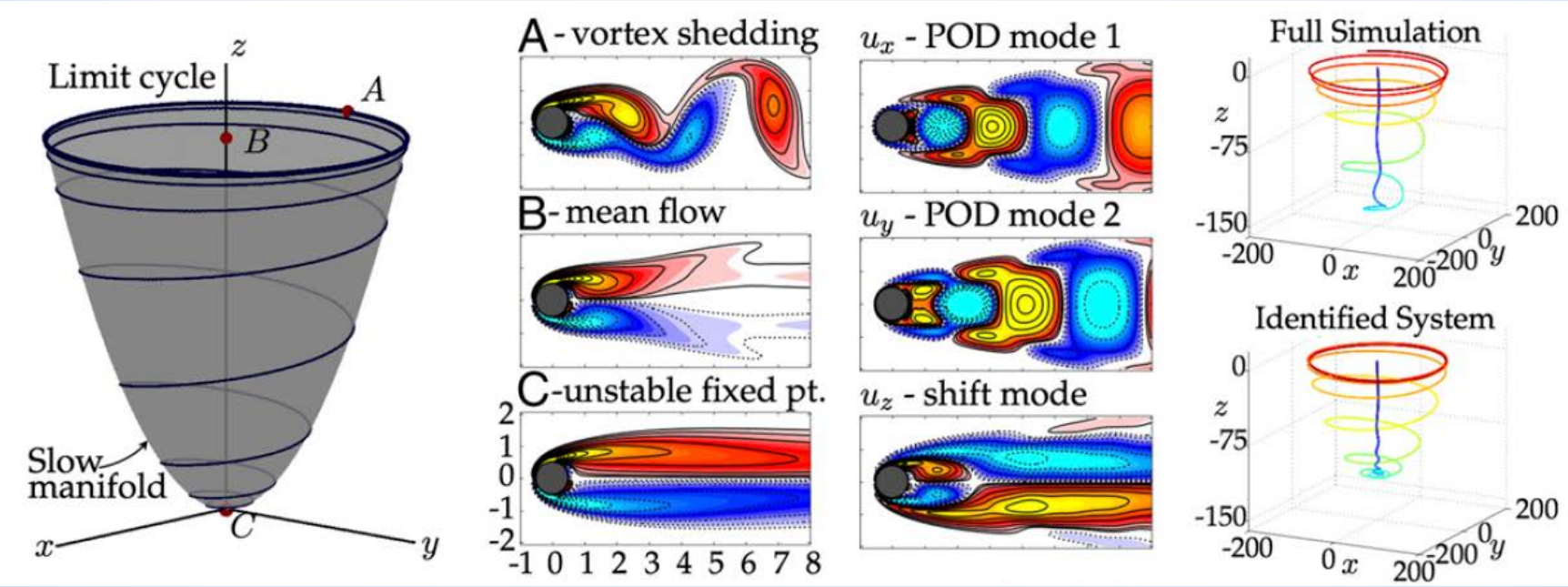
Summary of the paper

- Spirit of the paper:
  - Agnostic of traditions of economic theory, **let the data speak**
  - **Let machine ("PriSDA")** ***teach*** **us** what describes economic change
  - Use machine intelligence to *automate* the construction of relevant quantities and learn their dynamics
- Data:
  - C = 123 countries; P = 59 products (2-SITC); Y = 55 years (1962-2016)

- What PriSDA found:
  - What most distinguishes countries in time and space is **export basket diversity**.
  - Export basket diversity appears to drive *per capita* income, not the reverse
  - **Countries are not split** in clear communities (e.g., manufacturing vs agricultural), and there is a tendency to **converge** on the same basket.

"Law-discovery" methods have been applied successfully to very difficult problems, e.g., high-dimensional chaotic systems.

- **High dimensional problem**. In a fluid, the state of the whole fluid in 2D space is a single point in phase space.
- E.g., Singular Value Decomposition (~ Principal Component Analysis) reduces the dimensionality of the problem. After that, they apply SINDy.



! "[SINDy discovered] the underlying dynamics of a system that took experts in the community nearly 30 years to resolve"

- Our situation:

  Model the export basket vector *x(t)* evolving in time

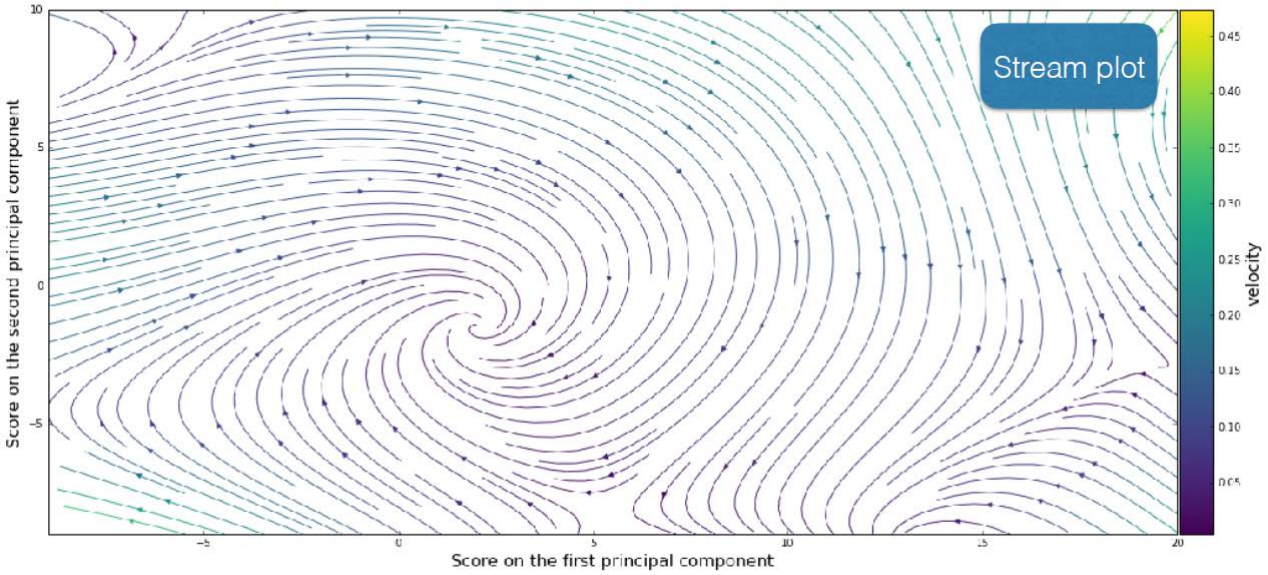
$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t))$$

Take world trade data from 1962-2016, and infer (plausible) underlying dynamical regularities using machine learning.

The promise

Reduce dimensions of the data using PCA on $\frac{\mathrm{RCA}-1}{\mathrm{RCA}+1}+1$

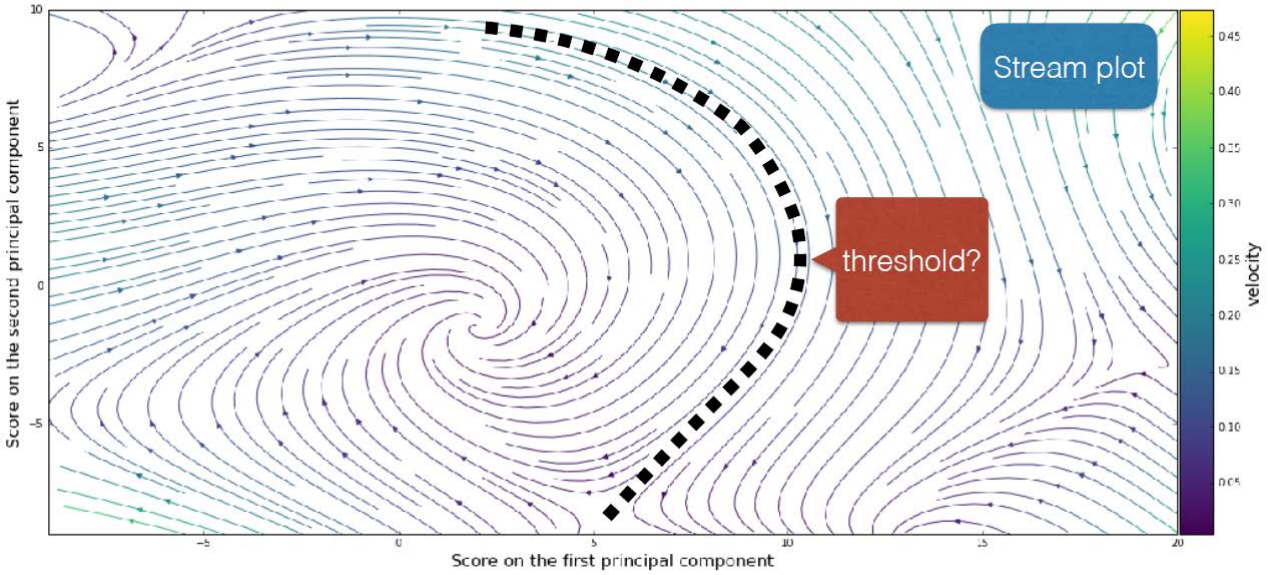
Model $\boldsymbol{x}(t+1) - \boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ with SINDy (polynomials)



The promise

Reduce dimensions of the data using PCA on $\frac{\mathrm{RCA}-1}{\mathrm{RCA}+1}+1$

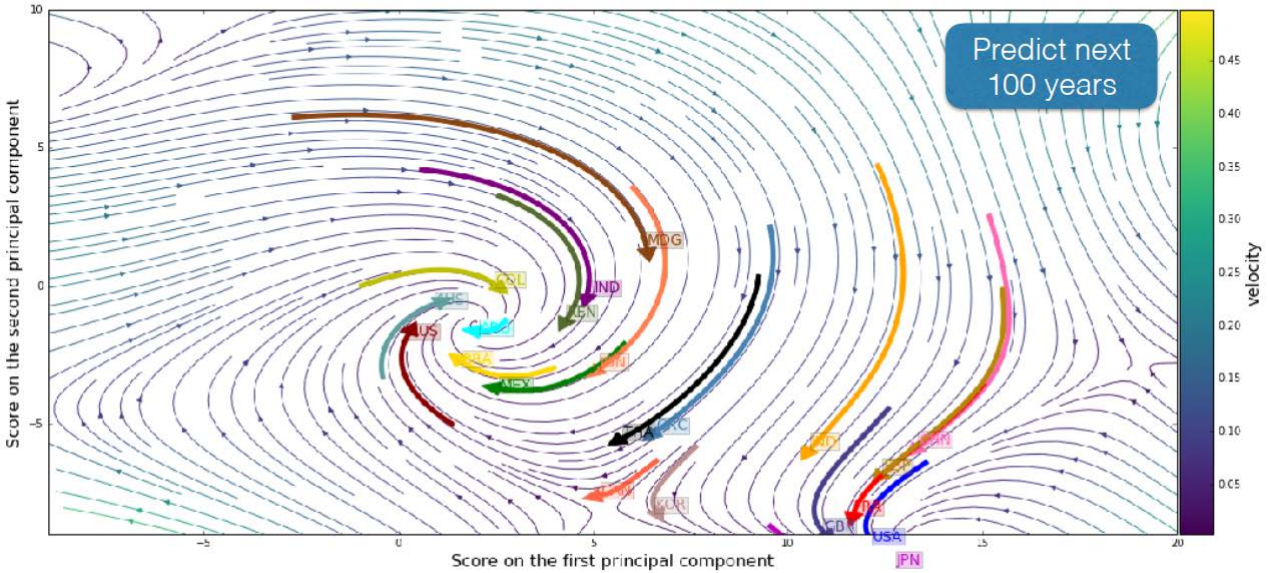
Model $\boldsymbol{x}(t+1) - \boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ with SINDy (polynomials)



The promise

Reduce dimensions of the data using PCA on $\frac{\mathrm{RCA}-1}{\mathrm{RCA}+1}+1$

Model $\boldsymbol{x}(t+1) - \boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ with SINDy (polynomials)

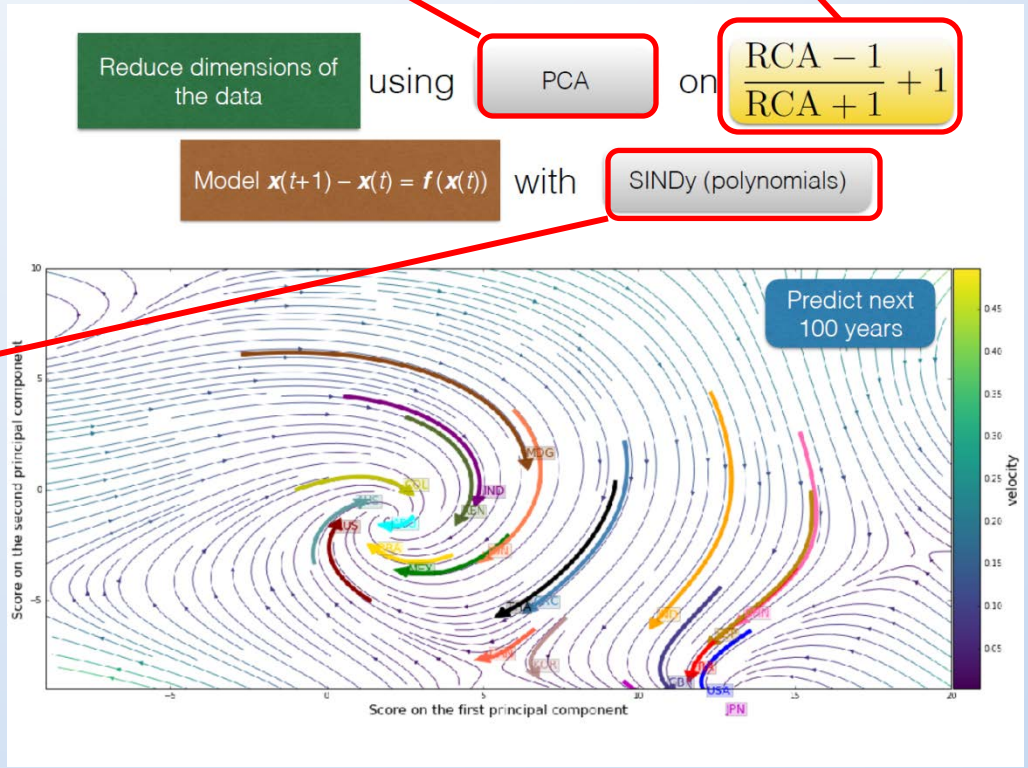


Three elements of decision

1. What is the appropriate transformation?

2. What is the appropriate dimensionality reduction?

3. What is the appropriate model?

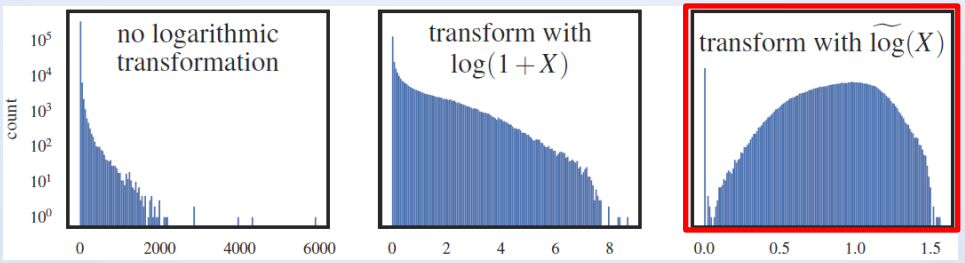


1. Transformations: What is a well-behaved measure of competitiveness?

$$\mathcal{R}_{cpt} := \frac{X_{cpt} / \mathbb{E}\left[X_{cpt} | P_{ct}\right]}{\sum_c X_{cpt} / \mathbb{E}\left[\sum_c X_{cpt} \middle| \sum_c P_{ct}\right]}$$

"Absolute Advantage":

$$\mathcal{R}_{cpt} = \frac{X_{cpt} / \left(\alpha_p \left(P_{ct}\right)^{\beta_p}\right)}{\sum_c X_{cpt} / \left(\gamma_p \left(\sum_c P_{ct}\right)^{\delta_p}\right)}$$



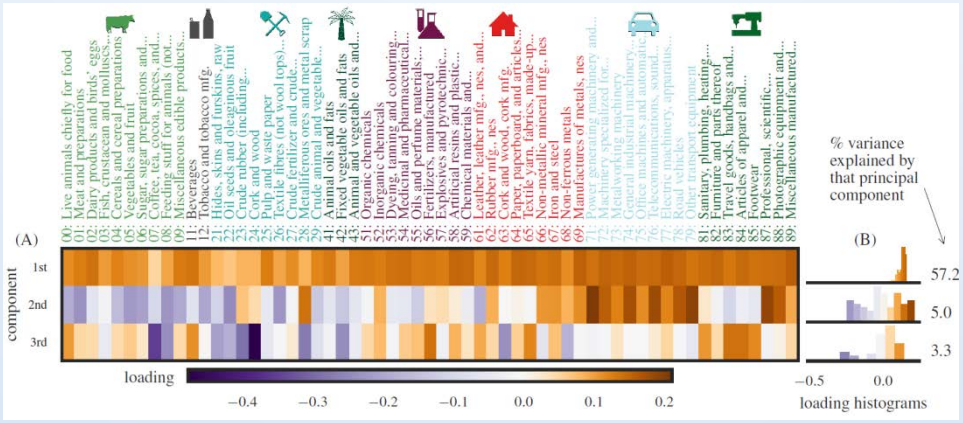
$$\widetilde{\log}(x) \equiv \begin{cases} 1 + s \log(x) & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases}$$

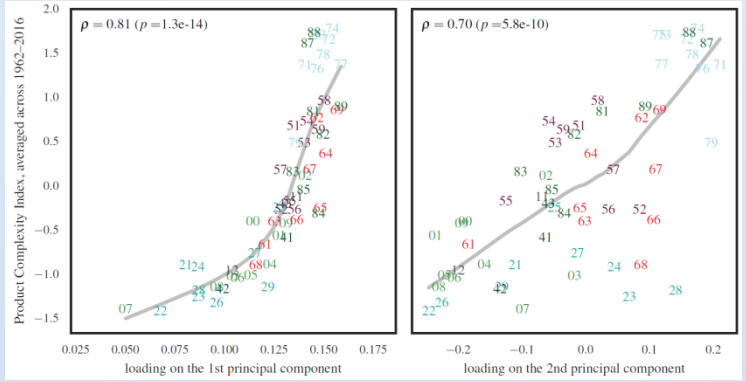
$$\widetilde{\log}(x_m) = x_m,$$
$$\widetilde{\log}(1) = 1,$$

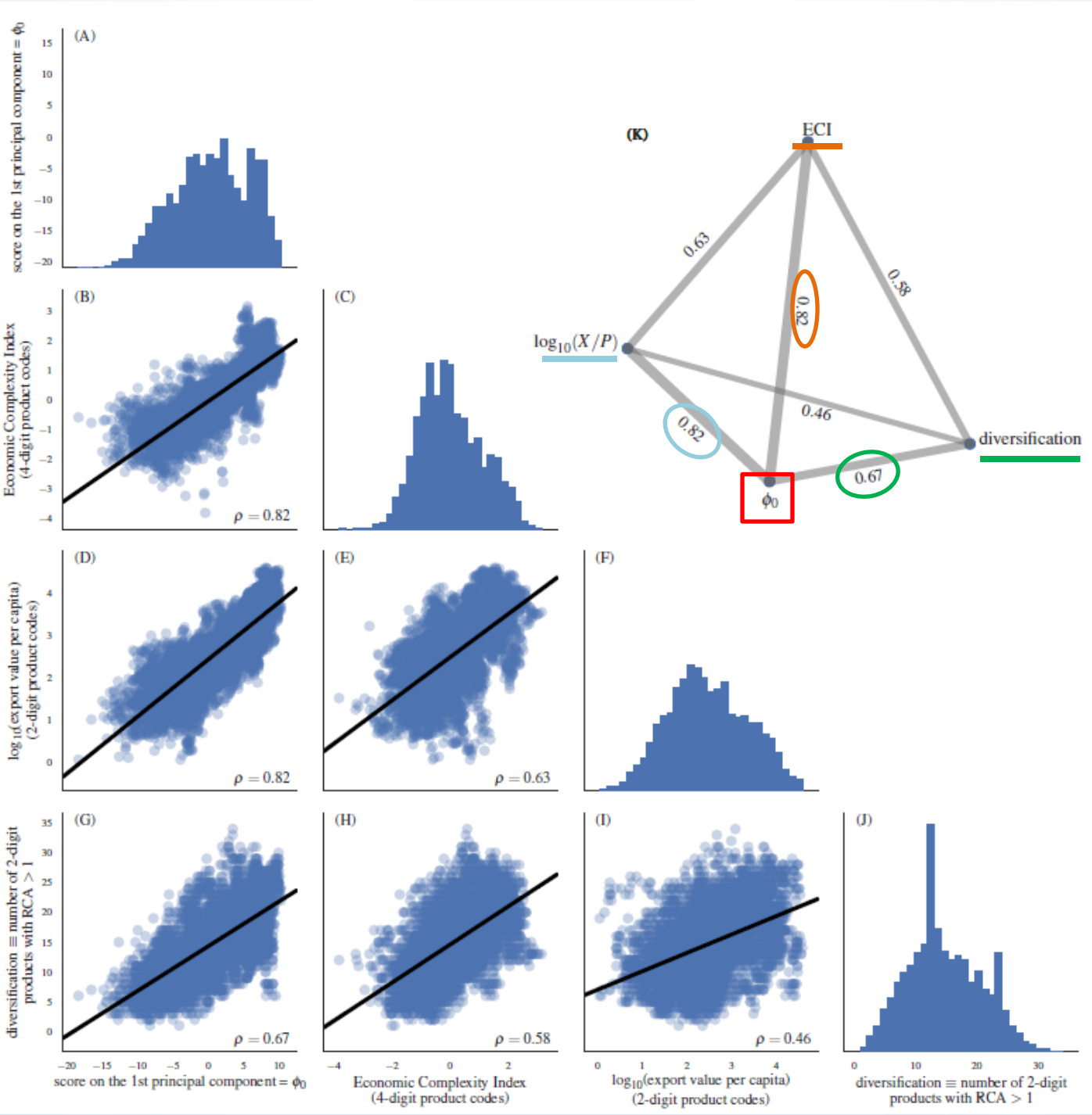
2. Dimensionality Reduction: Stable and invertible

Principal Component Analysis:

- Disadvantage: only linear structures.
- Advantage: deterministic and invertible.






(A)
(B)
(C)
(D)
(E)
(F)
(G)
(H)
(I)
(J)
(K)
score on the 1st principal component = ϕ_0
Economic Complexity Index (4-digit product codes)
$\log_{10}$(export value per capita) (2-digit product codes)
diversification ≡ number of 2-digit products with RCA > 1
$\rho = 0.82$
$\rho = 0.82$
$\rho = 0.63$
$\rho = 0.67$
$\rho = 0.58$
$\rho = 0.46$
ECI
$\log_{10}(X/P)$
diversification
ϕ_0
0.63
0.58
0.82
0.46
0.82
0.67
score on the 1st principal component = ϕ_0
Economic Complexity Index (4-digit product codes)
$\log_{10}$(export value per capita) (2-digit product codes)
diversification ≡ number of 2-digit products with RCA > 1

3. Machine Learning model: Predictive but also interpretable

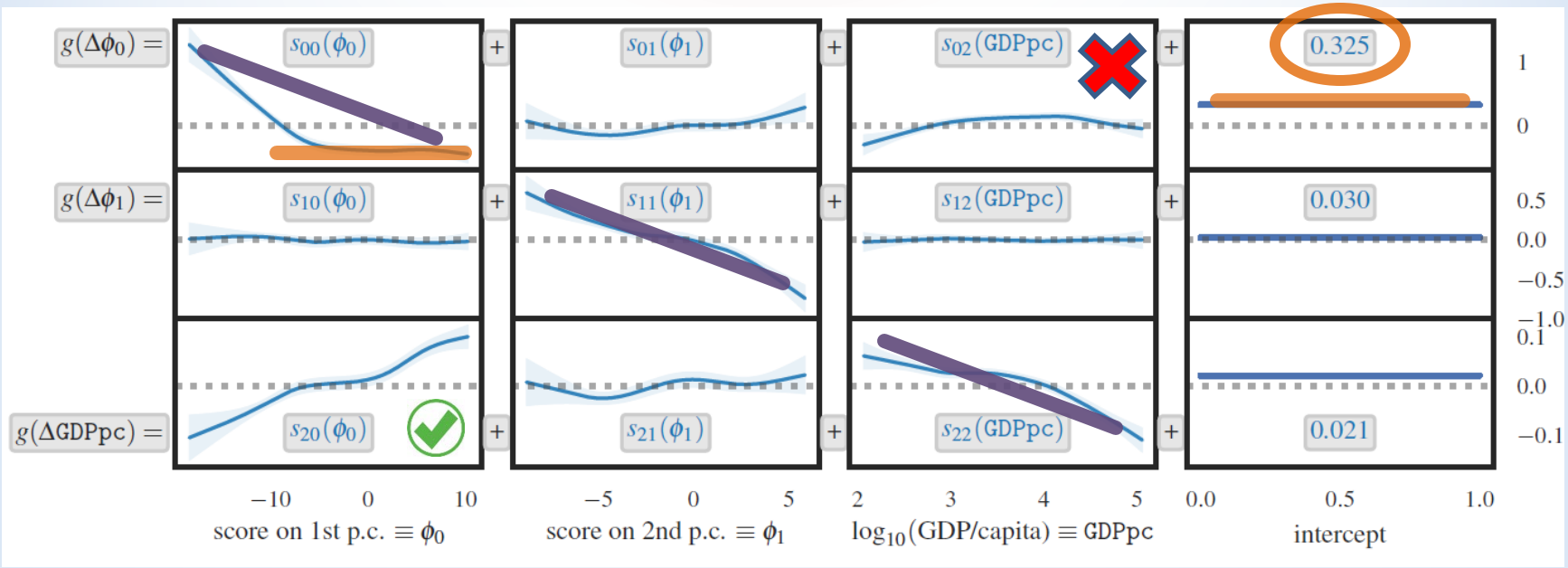
Generalized Additive Models
-> cubic smoothing splines:

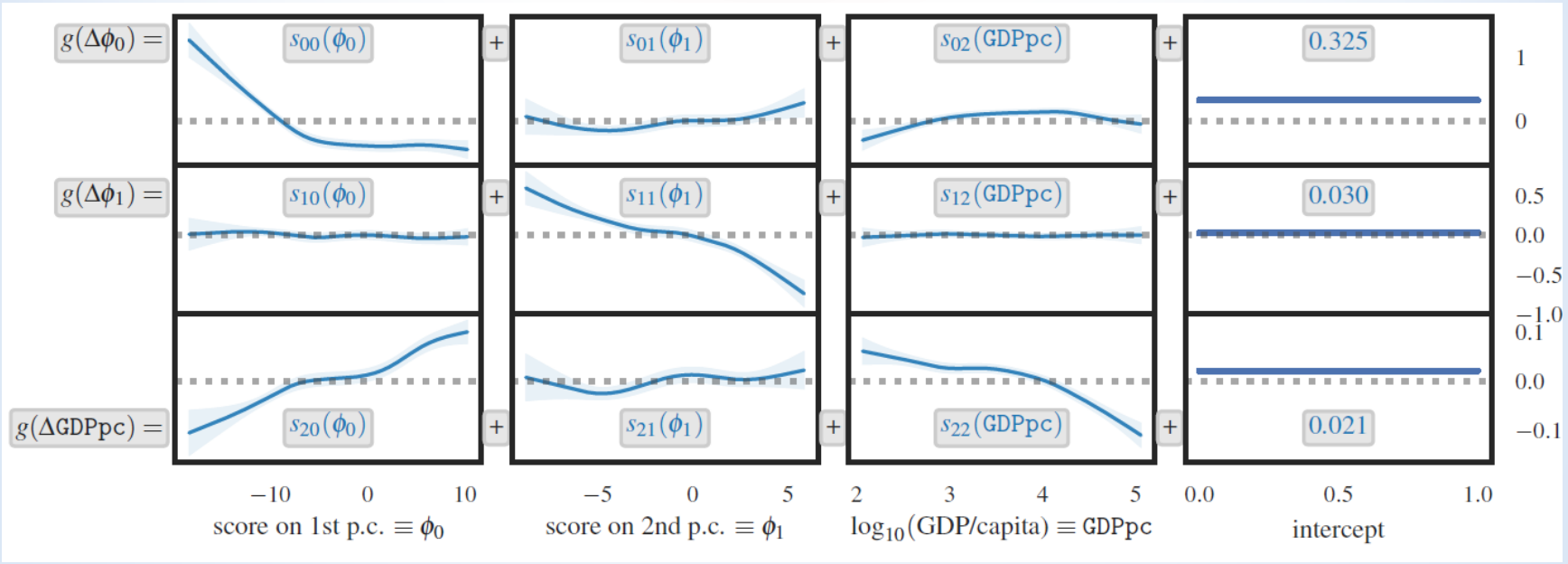
- Disadvantage: no explicit equation.
- Advantage: flexible, efficient and non-parametric, best generalization error.

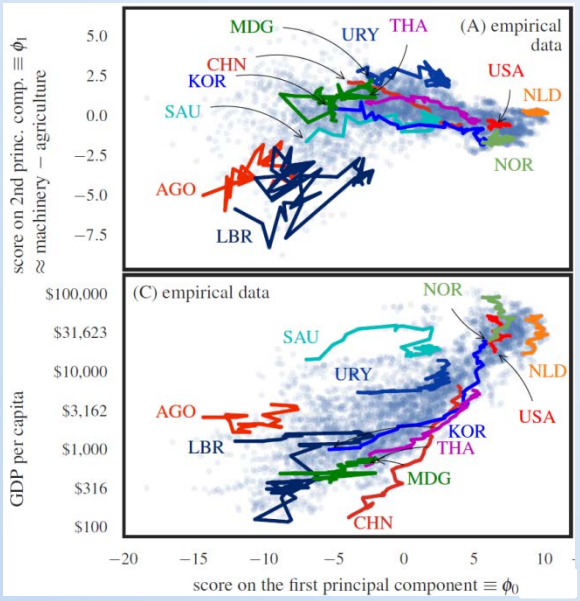
$$g(\Delta\phi_0(t)) = c_0 + s_{00}(\phi_0(t)) + s_{01}(\phi_1(t)) + s_{02}(\mathrm{GDPpc}(t)) \quad (1.2a)$$

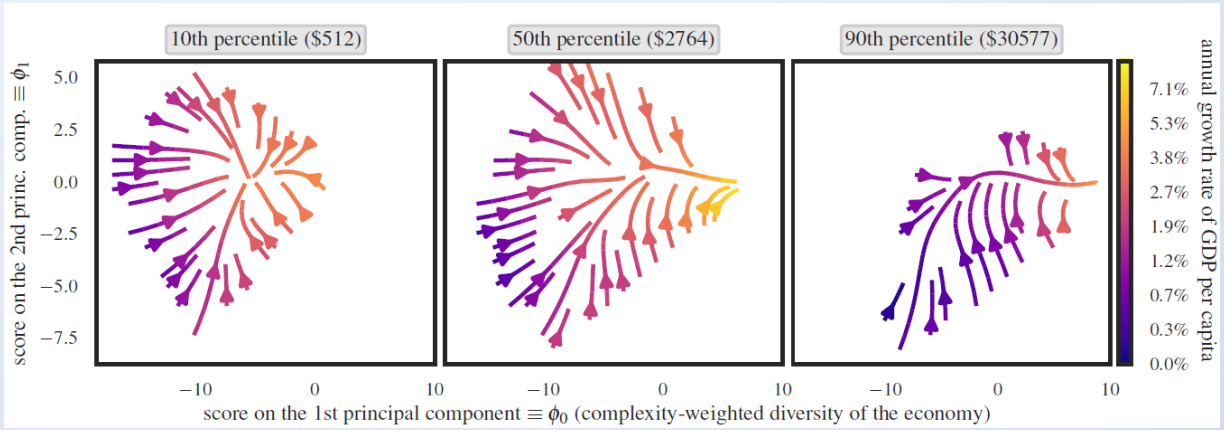
$$g(\Delta\phi_1(t)) = c_1 + s_{10}(\phi_0(t)) + s_{11}(\phi_1(t)) + s_{12}(\mathrm{GDPpc}(t)) \quad (1.2b)$$

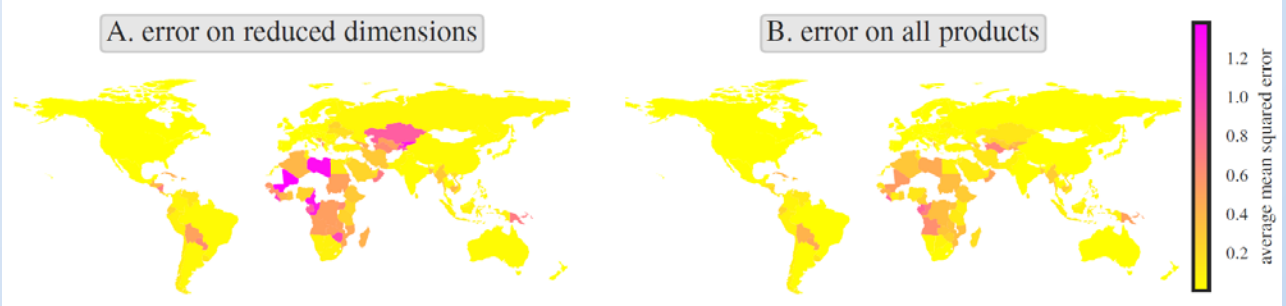
$$\text{and} \quad g(\Delta\mathrm{GDPpc}(t)) = c_2 + s_{20}(\phi_0(t)) + s_{21}(\phi_1(t)) + s_{22}(\mathrm{GDPpc}(t)) \quad (1.2c)$$


$g(\Delta\phi_0) =$
$g(\Delta\phi_1) =$
$g(\Delta\text{GDPpc}) =$
$s_{00}(\phi_0)$
$s_{01}(\phi_1)$
$s_{02}(\text{GDPpc})$
0.325
$s_{10}(\phi_0)$
$s_{11}(\phi_1)$
$s_{12}(\text{GDPpc})$
0.030
$s_{20}(\phi_0)$
$s_{21}(\phi_1)$
$s_{22}(\text{GDPpc})$
0.021
+
1
0
0.5
0.0
−0.5
−1.0
0.1
0.0
−0.1
−10
0
10
−5
0
5
2
3
4
5
0.0
0.5
1.0
score on 1st p.c. $\equiv \phi_0$
score on 2nd p.c. $\equiv \phi_1$
$\log_{10}(\text{GDP/capita}) \equiv \text{GDPpc}$
intercept

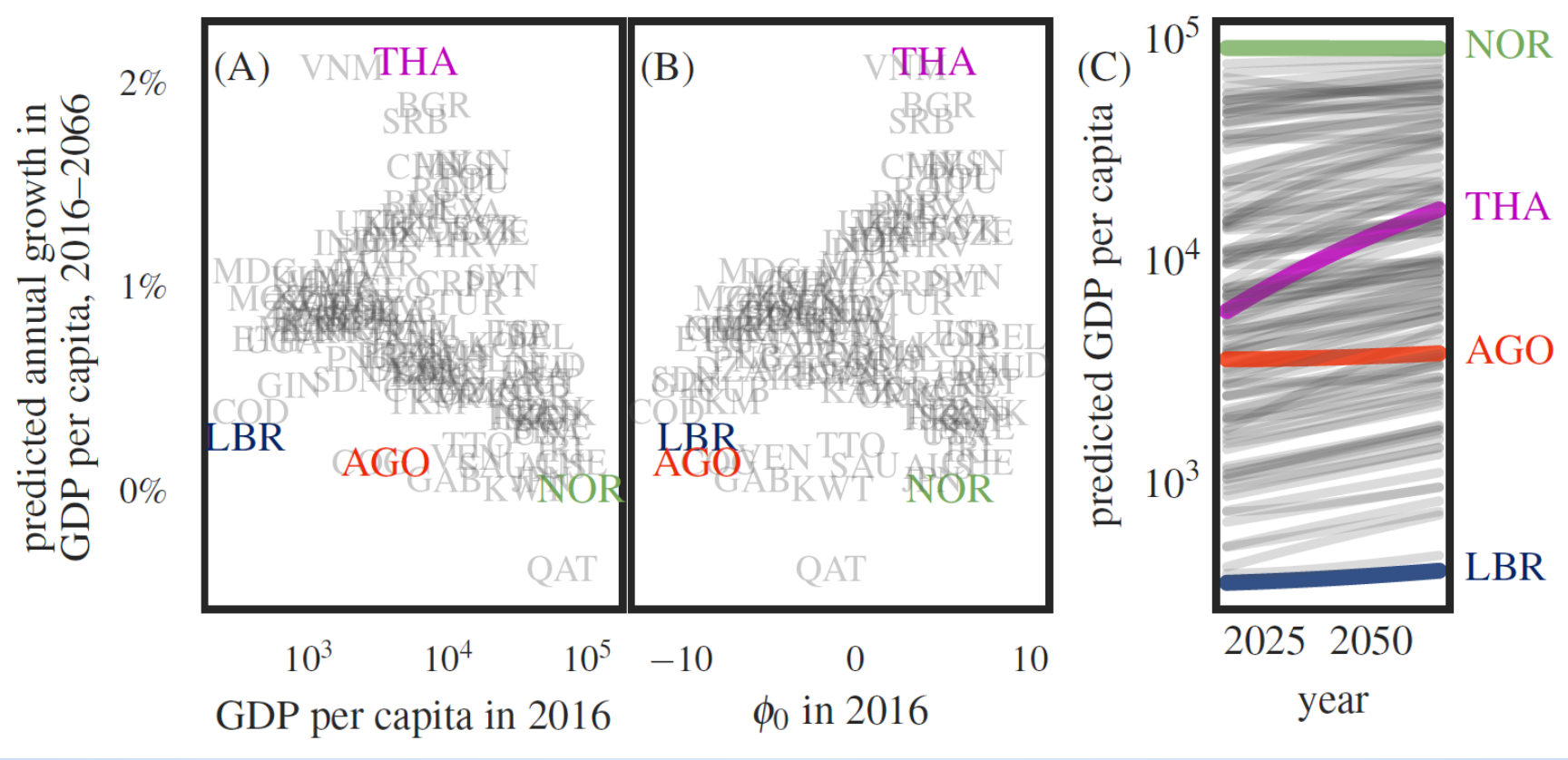

$g(\Delta\phi_0) =$
$s_{00}(\phi_0)$
$s_{01}(\phi_1)$
$s_{02}(\mathrm{GDPpc})$
0.325
$g(\Delta\phi_1) =$
$s_{10}(\phi_0)$
$s_{11}(\phi_1)$
$s_{12}(\mathrm{GDPpc})$
0.030
$g(\Delta\mathrm{GDPpc}) =$
$s_{20}(\phi_0)$
$s_{21}(\phi_1)$
$s_{22}(\mathrm{GDPpc})$
0.021
score on 1st p.c. $\equiv \phi_0$
score on 2nd p.c. $\equiv \phi_1$
$\log_{10}(\mathrm{GDP/capita}) \equiv \mathrm{GDPpc}$
intercept


(A) empirical data
MDG
URY
THA
CHN
KOR
USA
NLD
SAU
NOR
AGO
LBR
score on 2nd princ. comp. $\equiv \phi_1$ ≈ machinery − agriculture
(C) empirical data
GDP per capita
score on the first principal component $\equiv \phi_0$


10th percentile ($512)
50th percentile ($2764)
90th percentile ($30577)
score on the 2nd princ. comp. $\equiv \phi_1$
score on the 1st principal component $\equiv \phi_0$ (complexity-weighted diversity of the economy)
annual growth rate of GDP per capita


A. error on reduced dimensions
B. error on all products
average mean squared error

50-yr forecast



Summary of negative results

- Predictability was low
- No “archetypical” paths
  - However, there are too many angles, and we didn’t explore them all.
- Not clear how many dimensions has the state space.

Summary of positive results

1. We found that the movement of countries occurs in a **low dimensional** "space".
   - Dimension 1: Weighted diversity.
   - Dimension 2: Proficiency in machinery relative to agriculture.
2. High **diversity precedes GDPpc growth**, but high GDPpc does not precede diversification.
3. Evidence of **convergence** of export baskets.

Takeaways

- Economic activity *can (should?)* be represented in low dimensional space
  - Which dimensions?
- Low dimensions *can (should?)* be linked to economic performance
  - What mechanisms?

PAPER # 2

Elsevier Editorial System(tm) for Research Policy

Manuscript Draft

Manuscript Number: RESPOL-D-19-00231

Title: The drivers of urban economic complexity and their connection to urban economic performance

Article Type: VSI: Complexity

Keywords: urban diversification; collective knowhow; industrial structure; employment distribution; economic complexity

Corresponding Author: Dr. Andres Gomez-Lievano, Ph.D.

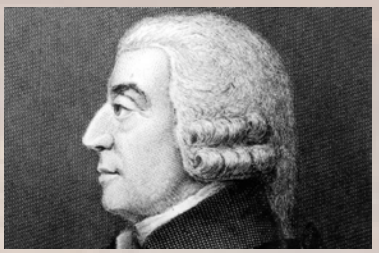
Corresponding Author's Institution: Harvard University

First Author: Andres Gomez-Lievano, Ph.D.

Order of Authors: Andres Gomez-Lievano, Ph.D.; Oscar Patterson-Lomba, PhD

Questions

- Can we motivate from *first principles* which "dimensions" summarize the complexity of economic activity?
- Can first principles suggest the mechanisms?



A theory about countries?

United States

New York City, NY

Holmes County, MS

Cross-city differences are **twice as large** as cross-country differences

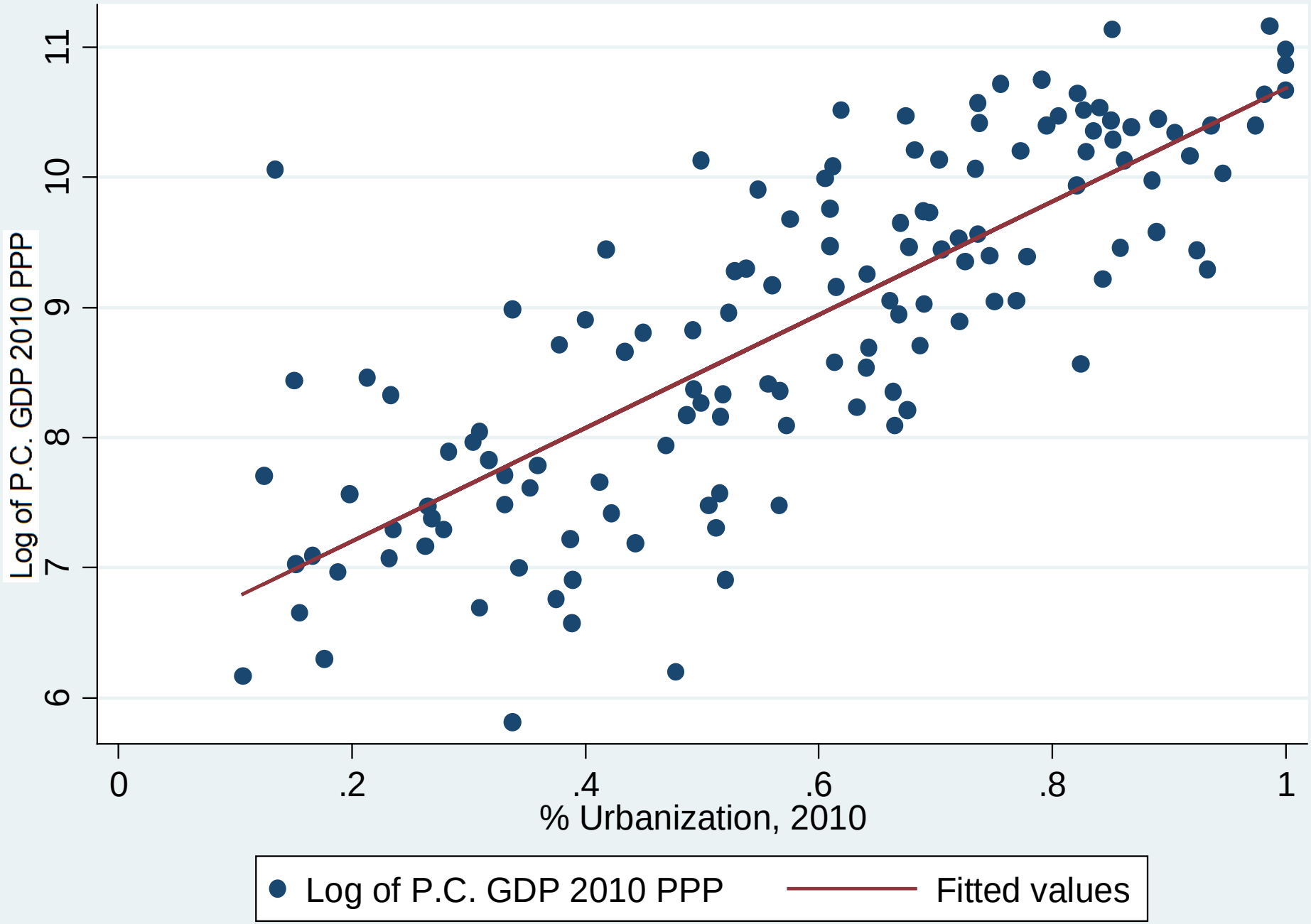
Colombia

Bogota D.C.

Vigía del fuerte, Antioquia

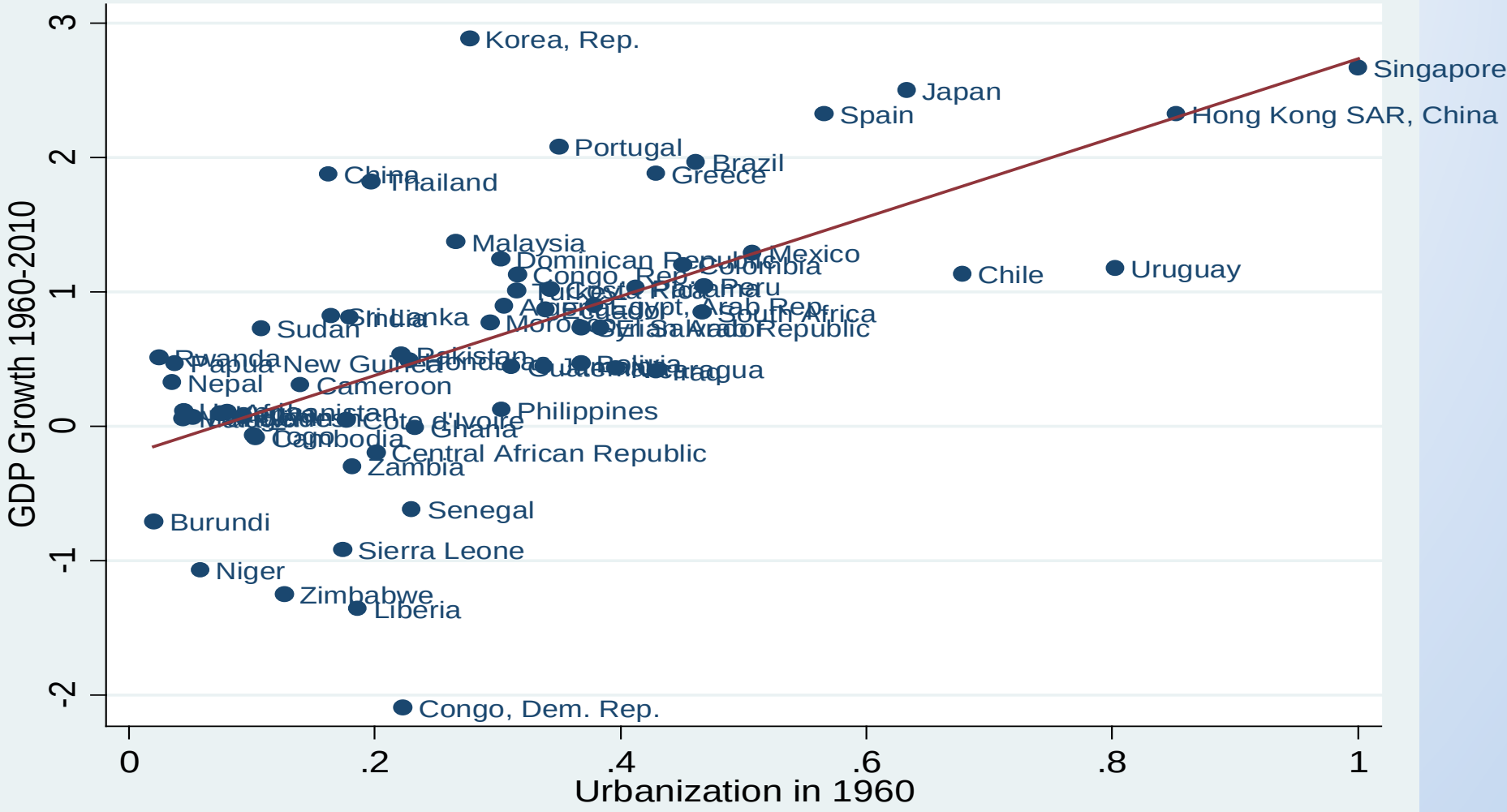
?

Bottom-right photo from : borrowed from S. Fajardo.



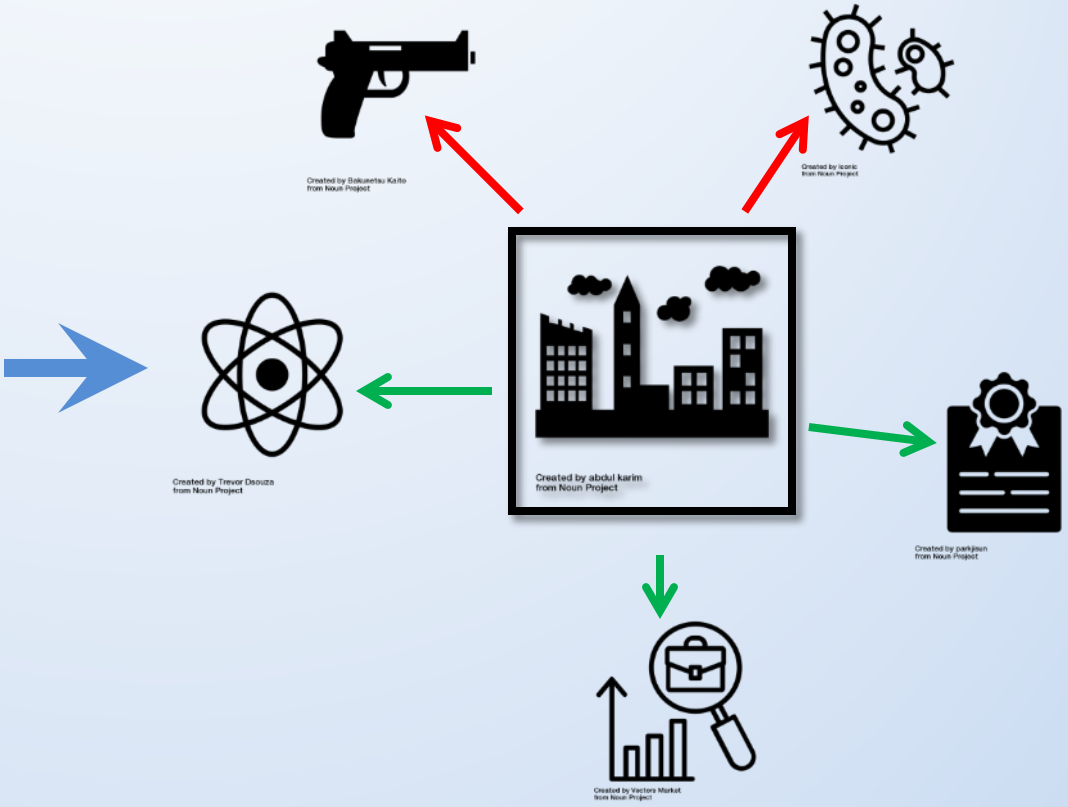
Graph by Prof. Edward Glaeser

Per Capita GDP Growth 1960-2010 (Poor Countries<$5000 PC GDP)



Graph by Prof. Edward Glaeser


KAIKAINIJŪN


Created by Bokunetsu Kaito from Noun Project
Created by Iconic from Noun Project
Created by Trevor Dsouza from Noun Project
Created by abdul karim from Noun Project
Created by parkjisun from Noun Project
Created by Vectors Market from Noun Project

Growing literature on a "new science of cities"

- Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci. USA 104, 7301–7306* (2007).
- Samaniego, H. & Moses, M. E. Cities as organisms: allometric scaling of urban road networks. *J. Transp. Land Use 1, 21–39* (2008).
- Arbesman, S., Kleinberg, J. M. & Strogatz, S. H. Superlinear scaling for innovation in cities. *Phys. Rev. E 79, 016115* (2009).
- Gomez-Lievano, A., Youn, H. & Bettencourt, L. M. A. The statistics of urban scaling and their connection to Zipf 's Law. *PLoS ONE 7, e40393* (2012).
- Bettencourt, L. M. A., Lobo, J. & Youn, H. The hypothesis of urban scaling: formalization, implications and challenges. *Preprint at http://arxiv.org/abs/1301.5919v1* (2013).
- Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nat. Commun. 4, 1961* (2013).
- Bettencourt, L. M. A. The origins of scaling in cities. *Science 340, 1438* (2013).
- Yakubo, K., Saijo, Y. & Korošak, D. Superlinear and sublinear urban scaling in geographical networks modeling cities. *Phys. Rev. E 90, 022803* (2014).
- Bettencourt, L. M., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Sci. Rep. 4, 5393* (2014).
- Patterson-Lomba, O., Goldstein, E., Gómez-Liévano, A., Castillo-Chavez, C., & Towers, S. Per capita incidence of sexually transmitted infections increases systematically with urban population size: a cross-sectional study. *Sex Transm Infect*, *91*:8, 610-614 (2015).
- Youn, H. et al. Scaling and universality in urban economic diversification. *J. R. Soc. Interf. 13, http://dx.doi.org/10.1098/rsif.2015.0937* (2016).
- Gomez-Lievano, A., Patterson-Lomba, O., & Hausmann, R. Explaining the prevalence, scaling and variance of urban phenomena. *Nature Human Behaviour*, 1, 0012 (2016).

The theory in brief:

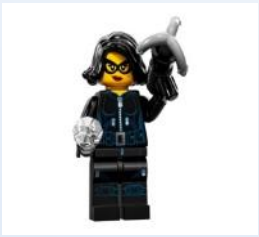
Gomez-Lievano, A., Patterson-Lomba, O., & Hausmann, R. (2016). Explaining the prevalence, scaling and variance of urban phenomena. Nature Human Behaviour, 1, 0012.

1. Most of urban phenomena are the conjunction of complementary factors. The fewer the factors, the less "complex" the phenomenon.

2. Cities accumulate these factors through a stochastic process of accumulation.

3. And each person in the city is different in the factors they bring to the city. The exposure of people to the city is what generates the outcomes.

Gomez-Lievano, A., Patterson-Lomba, O., & Hausmann, R. (2016). Explaining the prevalence, scaling and variance of urban phenomena. Nature Human Behaviour, 1, 0012.



Model inspired on:
Hidalgo, C. A. and Hausmann, R. (2009).
The building blocks of **economic complexity**. PNAS,
106(25):10570-10575.

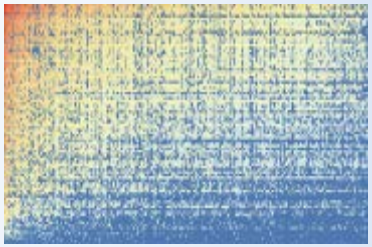
Hausmann, R. and Hidalgo, C.A. (2011),
The network structure of economic output. J Econ Growth,
16:309-342.

After some algebraic manipulations…

$$P\left(X_{icf}=1\right)\cong e^{-M_f(1-s_i)(1-r_c)}$$

Food services and drinking places
Administrative and support services
Ambulatory health care services
Professional and technical services
Hospitals
Specialty trade contractors
Merchant wholesalers, durable goods
Educational services
Food and beverage stores
General merchandise stores
Social assistance
Nursing and residential care facilities
Credit intermediation and related activities
Merchant wholesalers, nondurable goods
Insurance carriers and related activities
Motor vehicle and parts dealers
Accommodation
Management of companies and enterprises
Real estate
Amusements, gambling, and recreation
Personal and laundry services
Fabricated metal product manufacturing
Clothing and clothing accessories stores
Membership associations and organizations
Transportation equipment manufacturing
Truck transportation
Repair and maintenance
Construction of buildings
Building material and garden supply stores
Computer and electronic product manufacturing
Food manufacturing
Health and personal care stores
Machinery manufacturing
Securities, commodity contracts, investments
Miscellaneous store retailers
Chemical manufacturing
Gasoline stations
Telecommunications
Publishing industries, except internet
Heavy and civil engineering construction
Warehousing and storage
Sporting goods, hobby, book and music stores
Support activities for transportation
Couriers and messengers
Plastics and rubber products manufacturing
Electronics and appliance stores
Electronic markets and agents and brokers
Crop production
Transit and ground passenger transportation
Miscellaneous manufacturing
Printing and related support activities
Air transportation
Performing arts and spectator sports
Furniture and home furnishings stores
Motion picture and sound recording industries
Nonstore retailers
Utilities
Agriculture and forestry support activities
Rental and leasing services
Waste management and remediation services
Nonmetallic mineral product manufacturing
Broadcasting, except internet
Primary metal manufacturing
Furniture and related product manufacturing
Paper manufacturing
Private households
Support activities for mining
Data processing, hosting, and related services
Electrical equipment and appliance mfg
Beverage and tobacco product manufacturing
Wood product manufacturing
Other information services
Oil and gas extraction
Apparel manufacturing
Unclassified
Museums, historical sites, zoos, and parks
Animal production and aquaculture
Petroleum and coal products manufacturing
Textile product mills
Mining, except oil and gas
Textile mills
Water transportation
Pipeline transportation
Scenic and sightseeing transportation
Forestry and logging
Leather and allied product manufacturing
Lessors of nonfinancial intangible assets
Fishing, hunting and trapping
Postal service
Funds, trusts, and other financial vehicles
Rail transportation

Data



- API of Bureau of Economic Analysis
  - **Employment** across 3-digit industries, across Metropolitan Statistical Areas (MSAs), 1990—2016
  - **Population size** and **Total employment** by MSA, 1969—2016
- American Community Survey: Educational Attainment ("S1501")
  - Average years of schooling for "**Population with 25 years or older**" 2009—2016

$$\Pr\{X_{i,c,f} = 1\} = e^{-M_f(1-s_i)(1-r_c)}$$

- Can we estimate directly s_i, M_f and r_c?
- Is our model better than other alternatives?
- What can we learn?

What is the model telling us?

$$\Pr\{X_{i,c,f} = 1\} = e^{-M_f(1-s_i)(1-r_c)}$$

- $M_f\,(1\text{-}s_i)\,(1\text{-}r_c)$ = ln(1/Prob) = "Net complexity"

- Net complexity is decomposable:
  - M_f : Phenomenon-specific *complexity*
    = "recipe" of activity
  - s_i : Person-specific *susceptibility*
    = "individual knowhow"
  - r_c : City-specific *suitability*
    = "cultural diversity"
    = "collective knowhow"

Claims / Hypotheses

1. These “drivers” of urban outcomes can be measured

2. Knowing these “drivers” has measurable consequences

Claims / Hypotheses

1. These "drivers" of urban outcomes can be measured:

$$-\ln\left(-\ln\left(\Pr\{X_{icf}=1\}\right)\right) = -\ln(M_f) - \ln(1-s_i) - \ln(1-r_c)$$

$$\ln\left(-\ln\left(Y_{c,f}/N_c\right)\right)$$

linear regression with activity-specific and city-specific **fixed effects**

2. Knowing these "drivers" has measurable consequences

Claims / Hypotheses

1. These “drivers” of urban outcomes can be measured:

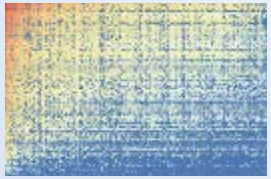
$$-\ln(-\ln(Y_{c,f}/N_c)) = -\ln(p_c) - \ln(M_f)$$

$$\boxed{-\ln\left(-\ln\left(y_{c,f}\right)\right) = \gamma_c + \delta_f + \varepsilon_{c,f}}$$

2. Knowing these “drivers” has measurable consequences

Claims / Hypotheses

1. These "drivers" of urban outcomes can be measured:



Share of employment
of industry *f* in city *c*

$$-\ln\left(-\ln\left(y_{c,f}\right)\right) = \gamma_c + \delta_f + \varepsilon_{c,f}$$

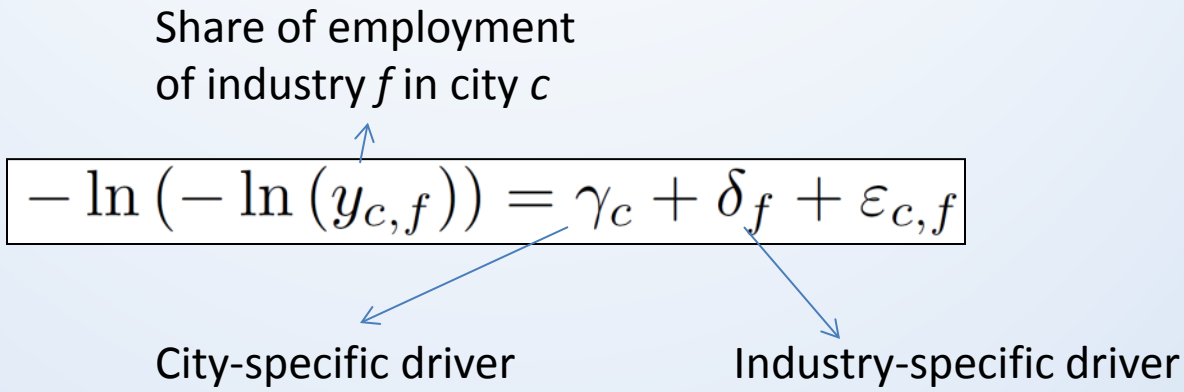
City-specific driver

Industry-specific driver

2. Knowing these "drivers" has measurable consequences

Claims / Hypotheses

1. These "drivers" of urban outcomes can be measured:



2. Knowing these "drivers" has measurable consequences:
   - Higher **predictive** power
   - Understanding urban economic performance:
     average firm size and **average wages**.

$$RMSE \equiv \sqrt{\frac{1}{|\mathcal{S}|}\sum_{(c,f)\in\mathcal{S}}(y_{c,f}-\widehat{y_{c,f}})^2}$$

$$MAE \equiv \frac{1}{|\mathcal{S}|}\sum_{(c,f)\in\mathcal{S}}|y_{c,f}-\widehat{y_{c,f}}|$$

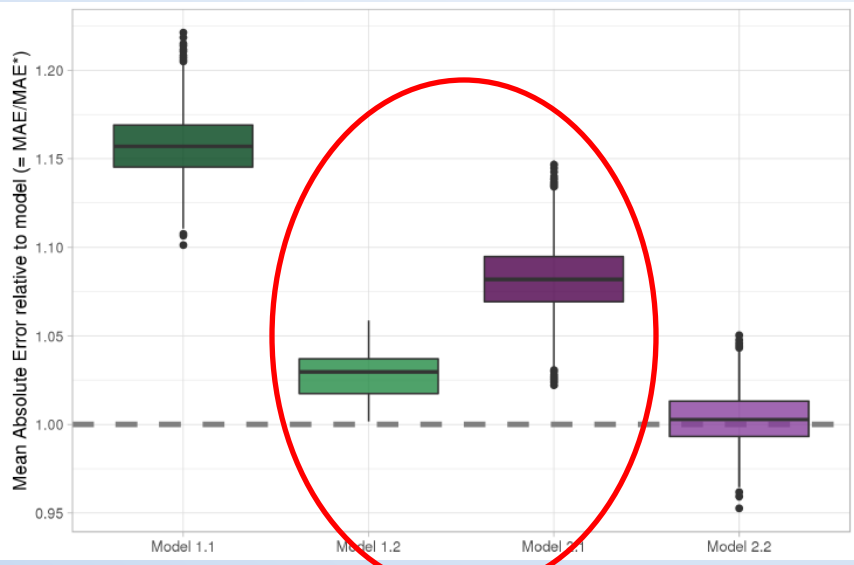
Bootstrap cross-validation:

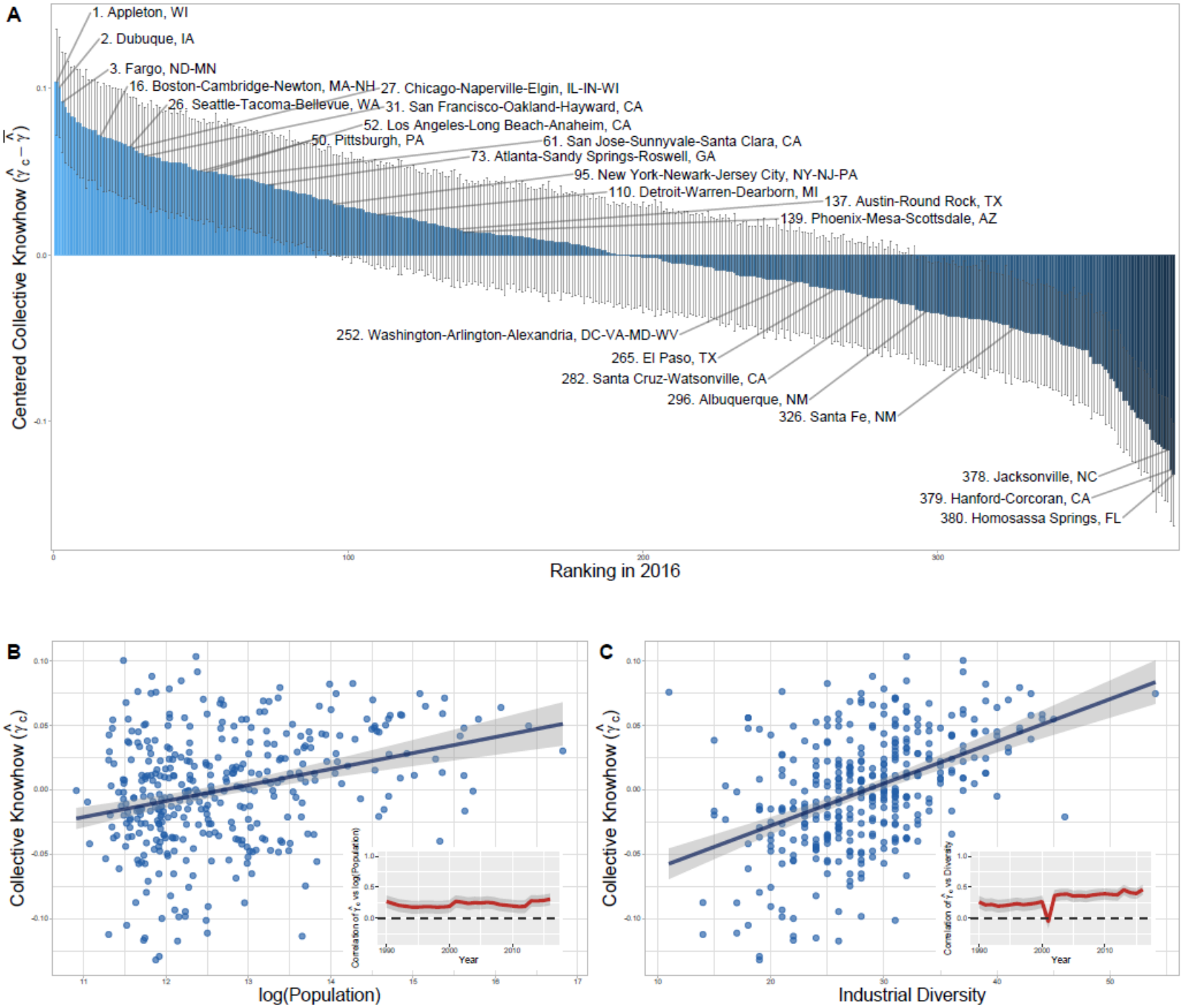
Fit on train → Predict on test

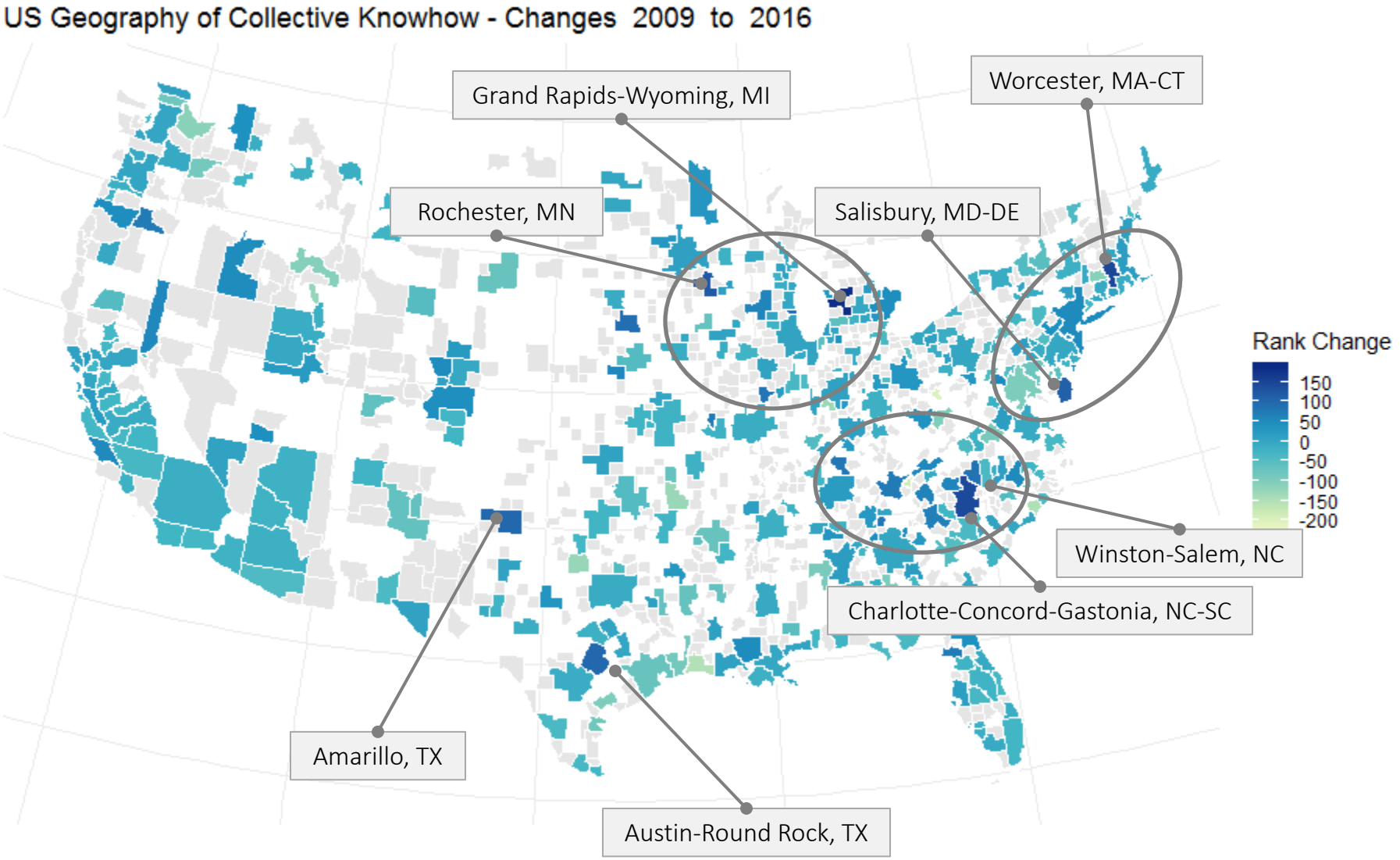
→ Evaluate $\mathrm{MAE}_{(\mathrm{model}\,k)}$ and $\mathrm{RMSE}_{(\mathrm{model}\,k)}$ for each model *k*

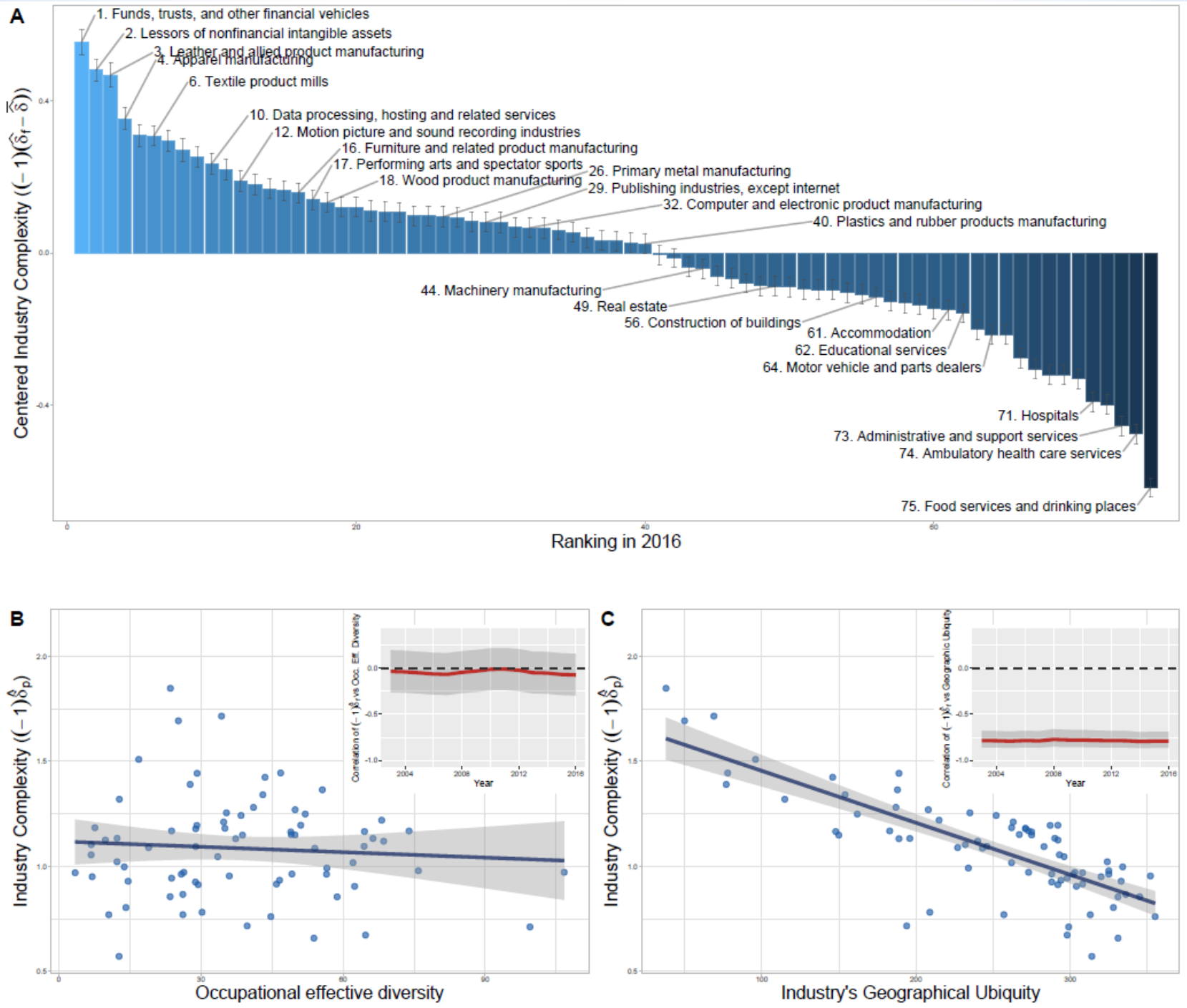
→ store $\mathrm{MAE}_{(\mathrm{model}\,k)}/\mathrm{MAE}_{(base)}$

$$\begin{aligned}
\text{Base (our model):} \quad & -\ln(-\ln(y_{c,f})) = \gamma_c + \delta_f + \varepsilon_{c,f} \\
\text{Model 1.1 (reference):} \quad & y_{c,f} = \gamma_c + \delta_f + \varepsilon_{c,f} \\
\text{Model 1.2 (substitutable factors):} \quad & \ln(y_{c,f}) = \gamma_c + \delta_f + \varepsilon_{c,f} \\
\text{Model 2.1 (universal exponent):} \quad & \ln(y_{c,f}) = \alpha_f + 0.16\ln(N_c) + \varepsilon_{c,f} \\
\text{Model 2.2 (EC + CE):} \quad & \ln(y_{c,f}) = \alpha_f + \beta_f \ln(N_c) + \varepsilon_{c,f}
\end{aligned}$$



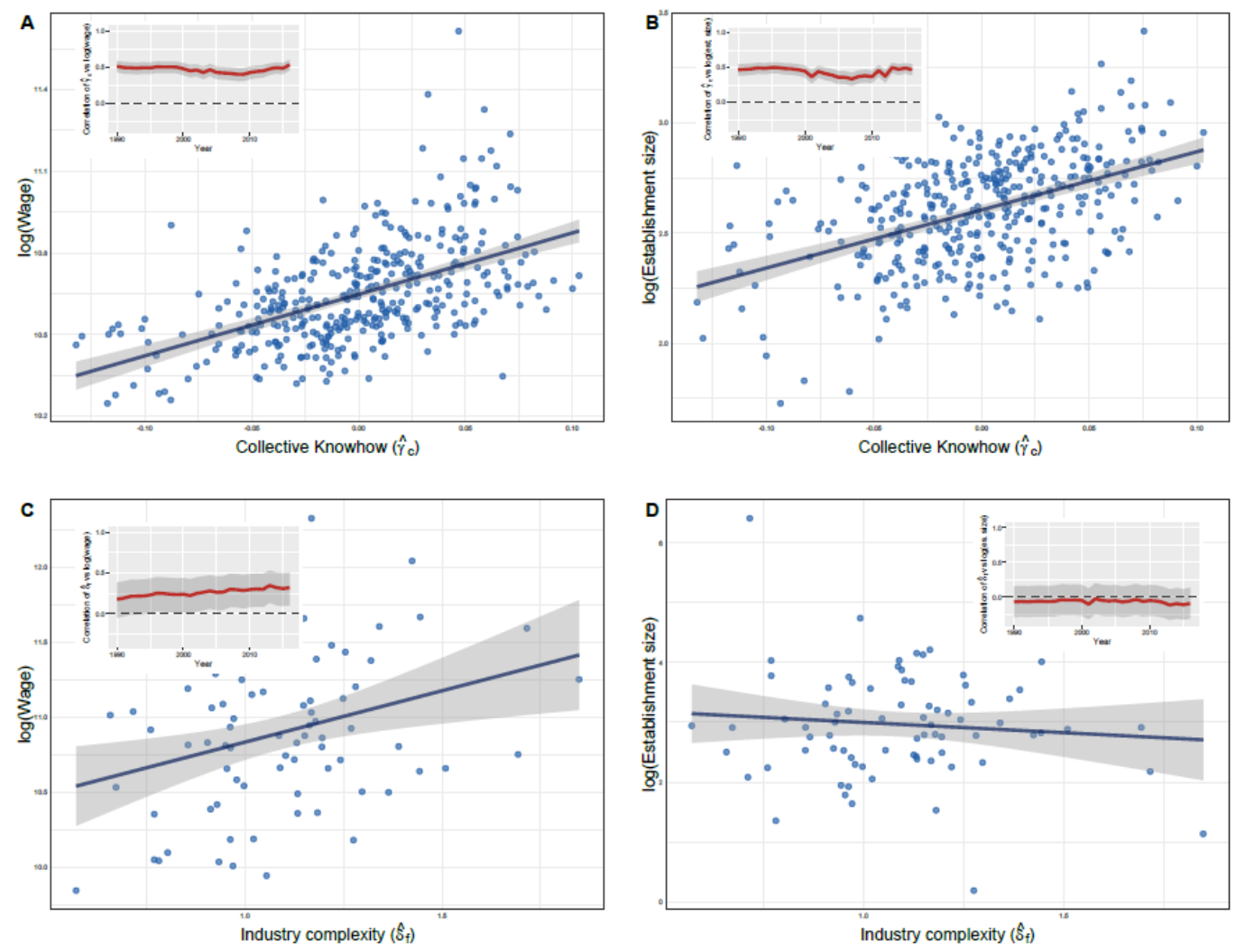

A
1. Appleton, WI
2. Dubuque, IA
3. Fargo, ND-MN
16. Boston-Cambridge-Newton, MA-NH
26. Seattle-Tacoma-Bellevue, WA
27. Chicago-Naperville-Elgin, IL-IN-WI
31. San Francisco-Oakland-Hayward, CA
52. Los Angeles-Long Beach-Anaheim, CA
50. Pittsburgh, PA
61. San Jose-Sunnyvale-Santa Clara, CA
73. Atlanta-Sandy Springs-Roswell, GA
95. New York-Newark-Jersey City, NY-NJ-PA
110. Detroit-Warren-Dearborn, MI
137. Austin-Round Rock, TX
139. Phoenix-Mesa-Scottsdale, AZ
252. Washington-Arlington-Alexandria, DC-VA-MD-WV
265. El Paso, TX
282. Santa Cruz-Watsonville, CA
296. Albuquerque, NM
326. Santa Fe, NM
378. Jacksonville, NC
379. Hanford-Corcoran, CA
380. Homosassa Springs, FL
Centered Collective Knowhow ($|\hat{\gamma}_c - \bar{\gamma}|$)
Ranking in 2016
B
Collective Knowhow ($\hat{\gamma}_c$)
log(Population)
Correlation of $\hat{\gamma}_c$ vs log(Population)
Year
C
Collective Knowhow ($\hat{\gamma}_c$)
Industrial Diversity
Correlation of $\hat{\gamma}_c$ vs Diversity
Year


US Geography of Collective Knowhow - Changes 2009 to 2016
Grand Rapids-Wyoming, MI
Worcester, MA-CT
Rochester, MN
Salisbury, MD-DE
Rank Change
150
100
50
0
-50
-100
-150
-200
Winston-Salem, NC
Charlotte-Concord-Gastonia, NC-SC
Amarillo, TX
Austin-Round Rock, TX


A
1. Funds, trusts, and other financial vehicles
2. Lessors of nonfinancial intangible assets
3. Leather and allied product manufacturing
4. Apparel manufacturing
6. Textile product mills
10. Data processing, hosting and related services
12. Motion picture and sound recording industries
16. Furniture and related product manufacturing
17. Performing arts and spectator sports
18. Wood product manufacturing
26. Primary metal manufacturing
29. Publishing industries, except internet
32. Computer and electronic product manufacturing
40. Plastics and rubber products manufacturing
44. Machinery manufacturing
49. Real estate
56. Construction of buildings
61. Accommodation
62. Educational services
64. Motor vehicle and parts dealers
71. Hospitals
73. Administrative and support services
74. Ambulatory health care services
75. Food services and drinking places
Centered Industry Complexity ((−1)(δ̂_I − δ̄))
Ranking in 2016
B
Industry Complexity ((−1)δ̂_p)
Occupational effective diversity
Correlation of (−1)δ̂_I vs Occ. Eff. Diversity
Year
C
Industry Complexity ((−1)δ̂_p)
Industry's Geographical Ubiquity
Correlation of (−1)δ̂_I vs Geographic Ubiquity
Year

These city- and industry-fixed effects are predictive of average individual wages and average size of establishments!

(see details and controls in paper)



Summary

1. "**Urban phenomena**" are the result of three quantities interacting:
   i. city "diversity"
   ii. phenomenon-specific "complexity"
   iii. and individual "susceptibility"

2. We can, in principle, **estimate** all three with the log-log of per-capita metrics.

3. Knowing them explains socio-economic **outcomes**, and could be useful for prediction and counter-factual analysis.

Considerations and open questions

- The model is based on an assumption about independent Bernoulli trials: $P(X_{icf} = 1|D_c) = s_i^{M_f - D_c}$ (factors are likely to come in "chunks")
- The model is static
- Cities and phenomena are described by a single quantity each. Worth exploring techniques for Matrix Factorization of the matrix – $\ln(-\ln(y_{cf}))$
- The proposed estimation strategy for individual-level data requires individuals to be observed in several phenomena and across several cities

Questions remaining:

- How to account for *Growth* → **Dynamical** models from first principles?
- Can we use individual-level data and a **Bayesian Framework** to simultaneously estimate s_i, M_f and r_c?

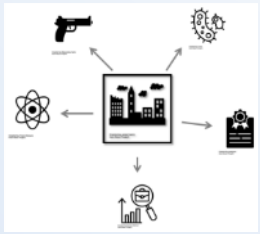
LOOKING FORWARD

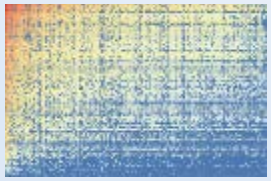
General lessons and intriguing questions

Transformations, dimensionality reduction, and dynamics

1. **Theory and mathematical models** that suggest the appropriate (non-trivial) **transformations** of the raw data
   - E.g., $-\log(-\log\left(\frac{X_{cp}}{X_c}\right))$ vs Revealed Competitive Advantage vs Absolute Advantage
2. **Theory and mathematical models** that suggest the appropriate **dimensionality reduction** technique
   - PCA vs NNMF vs t-SNE vs UMAP vs PHATE vs Isomap vs Diffusion Maps vs Monocle vs …
3. **Theory and mathematical models** that suggest what are the rules of **collective learning**
   - Do we need to change the paradigm from dynamical equations to rule-based modeling in order to account for open-ended growth?

THANK YOU





Contact info:
Email: andres_gomez@hks.harvard.edu
Twitter: @GomezLievano

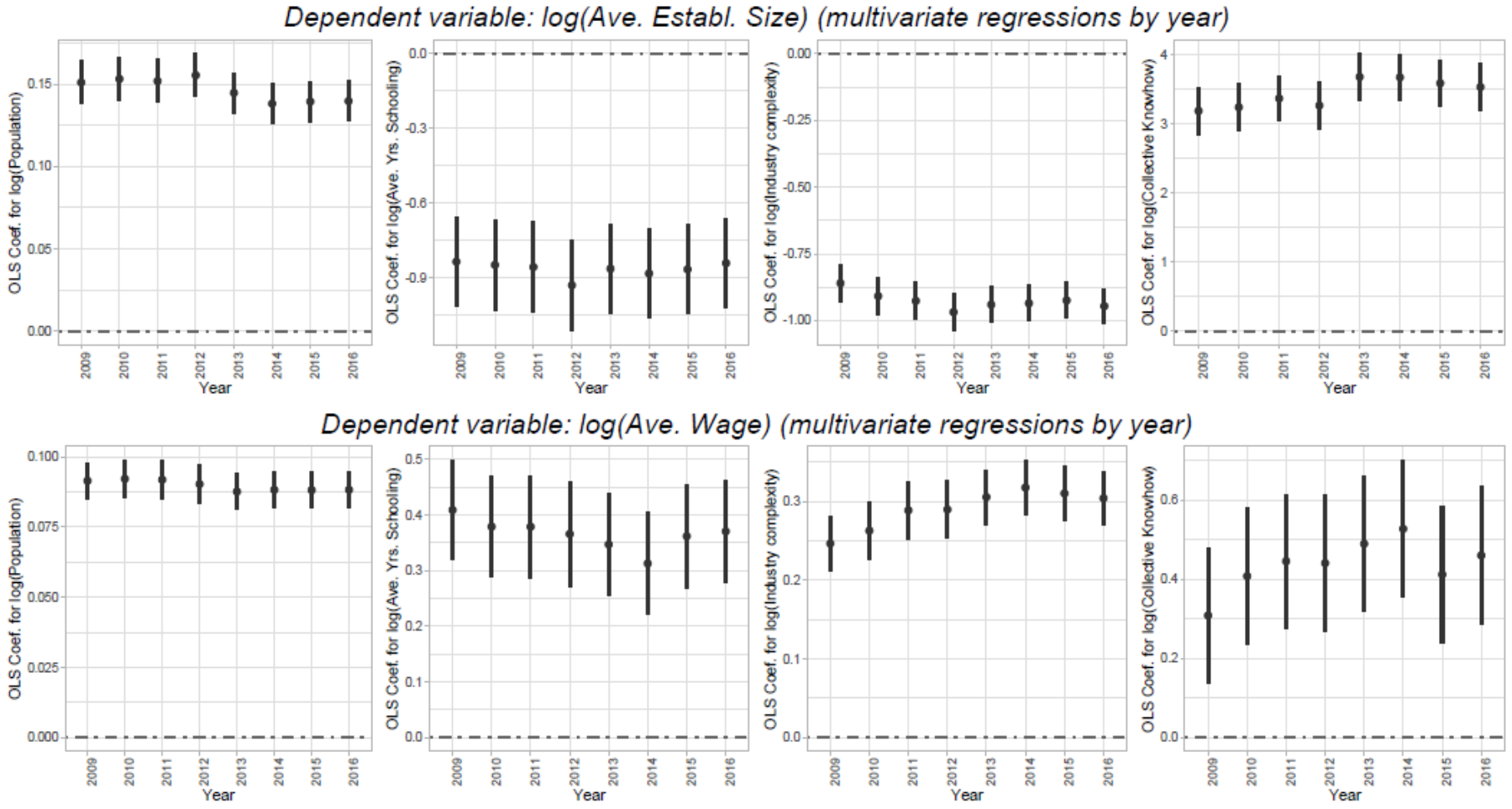


Fig 13. Partial elasticities of firm sizes (above) and wages (below) with respect to regressors. For each year we carry out a multivariate regression including population, average years of schooling, complexity and collective knowhow (column 6 in Tables 1 and 2), and plot the point estimate of the coefficients of such regressions with their corresponding standard error bars.

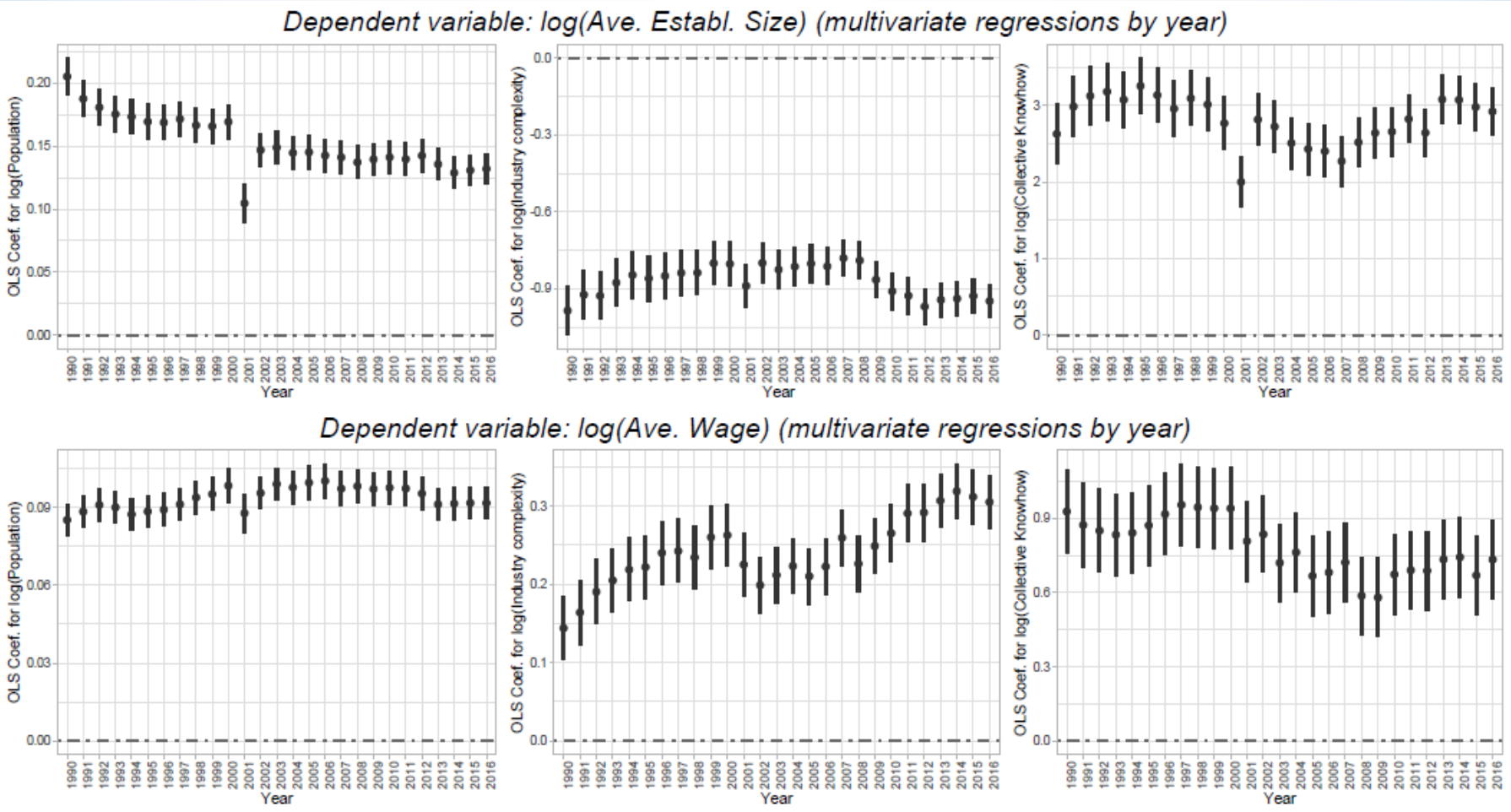


Fig 14. Partial elasticities of firm sizes (above) and wages (below) with respect to regressors. For each year we carry out a multivariate regression including population industry complexity and collective knowhow, and plot the point estimate of the coefficients of such regressions with their corresponding standard error bars.

Model

$$P(X_{icp} = 1) = \sum_{D} s_i^{M_p - D} \binom{M_p}{D} r_c^D (1 - r_c^D)^{M_p - D}$$

Using the Binomial theorem,

$$(a + b)^n = \sum \binom{n}{k} a^k b^{n-k}$$

we get

$$P(X_{icp} = 1) = (r_c + s_i(1 - r_c))^{M_p}$$

Approximation and rearrangements

$$P(X_{icf} = 1) = (r_c + s_i(1 - r_c))^{M_f}$$

$$(r_c + s_i(1 - r_c))^{M_f} = \left(-(1 - r_c) + 1 + s_i(1 - r_c)\right)^{M_f}$$
$$= \left(1 - (1 - s_i)(1 - r_c)\right)^{M_f}$$
$$e^{\ln(.)} = e^{M_f \ln(1-(1-s_i)(1-r_c))}$$

The exponent contains a logarithm $\ln(1 - x)$, which we expand in its Maclaurin series,

$$\ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$$

But $0 < x \ll 1$ (product of two probabilities), so we approximate the series only to first order, and neglect all terms $O(x^2)$:

$$\boldsymbol{P(X_{icf} = 1) \cong e^{-M_f(1-s_i)(1-r_c)}}$$