Forecasting Innovative Cities with Matrix Completion

<u>Nutarelli F.</u> [1], Edet S.[2], Gnecco G. [1], Riccaboni M. [1]

[1] IMT School for Advanced Studies,
[2] IMF Washington D.C.

SCAN ME

Overview

Which city is more competitive? Why?

(a) New York

(b) Jakarta

Overview

Which city is more competitive? Why?

(a) New York

(b) Jakarta

(c) Tokyo

(d) Johannesburg

Overview

Which city is more competitive? Why?

(a) New York

(b) Jakarta

(c) Tokyo

(d) Johannesburg

(e) Milan

(f) London

Motivation

While it may be simple to pick a city, it is difficult to explain why.

Motivation

While it may be simple to pick a city, it is difficult to explain why.

Understanding the competitiveness of smart cities is challenging [1]

(i) **No clear definition** of urban competitiveness ([5])

(ii) Contemporary urban landscapes are characterized not by singular, monolithic cities, but by diverse, **multi-centered** metropolitan regions.

(iii) There is a shift from traditional industrial chain systems to more dynamic and **unpredictable innovation chain** systems ([6]).

(iv) Urban competitiveness emerges from a **tapestry of complex, high-level social interactions**, encompassing various factors and innovations ([4], [5]).

Motivation

- Unraveling the linkage between competitiveness of cities and their resource allocations and capabilities [11] is crucial to understand why some cities are more competitive, productive, or resilient than others;

Motivation

- Unraveling the linkage between competitiveness of cities and their resource allocations and capabilities [11] is crucial to understand why some cities are more competitive, productive, or resilient than others;
- But why is competitiveness so important?

Motivation

- Unraveling the linkage between competitiveness of cities and their resource allocations and capabilities [11] is crucial to understand why some cities are more competitive, productive, or resilient than others;
- But why is competitiveness so important?

Cross-fertilization of ideas !

Motivation

- Unraveling the linkage between competitiveness of cities and their resource allocations and capabilities [11] is crucial to understand why some cities are more competitive, productive, or resilient than others;
- But why is competitiveness so important?

Cross-fertilization of ideas !

- In this scenario predicting the future competitiveness of global cities in different technological areas is key;

Motivation

- Unraveling the linkage between competitiveness of cities and their resource allocations and capabilities [11] is crucial to understand why some cities are more competitive, productive, or resilient than others;
- But why is competitiveness so important?

Cross-fertilization of ideas !

- In this scenario predicting the future competitiveness of global cities in different technological areas is key;
- Economic complexity and machine-learning literatures provide useful insights when combined.

Contribution

Contribution 1.1: Data Structure

Developing a unique dataset providing insights into the economic complexity and competitiveness of cities across different technological domains.

Contribution 1.2: Forecasting

Developing a forecast of future capabilities of cities taking into account high-order correlations between technologies. We employed concepts from economics complexity (Revealed Technology Advantage, henceforth RTA) and machine-learning (Matrix Completion, henceforth MC).

WHY CHOOSING MC?

Conceptual Reasons

- According to [9] and [10], *"innovation is a linear combination of existing technologies"*;
- Rows or columns of the **RTA** matrix are linearly dependent (low-rank matrix);
- MC's success depends on the fact that the matrix to be reconstructed is low-rank.

WHY CHOOSING MC?

Conceptual Reasons

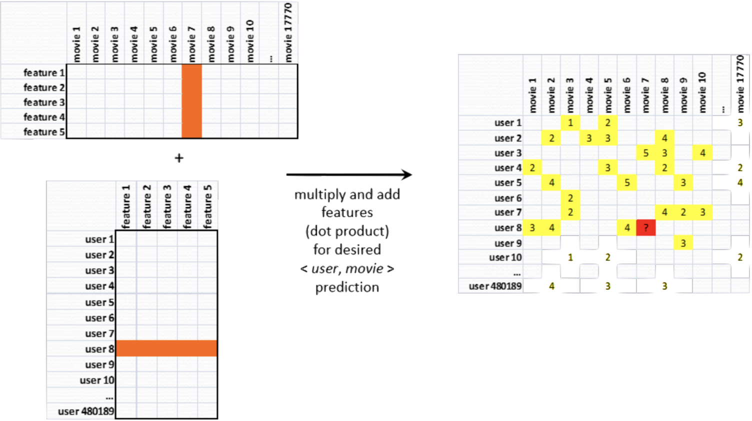
- According to [9] and [10], *"innovation is a linear combination of existing technologies"*;
- Rows or columns of the **RTA** matrix are linearly dependent (low-rank matrix);
- MC's success depends on the fact that the matrix to be reconstructed is low-rank.

Technical Reasons

- Previous methods –focused on complexity– retained only first 2 eigenvalues of a matrix associated with a bipartite network;
- MC retains n singular values, where n is the minimal number to minimize out-of-bag prediction errors!
- Hence MC is better for prediction tasks.

HOW CAN MC HELP UNCOVERING HIGH-ORDER CORRELATIONS?

- MC reconstructs each row of a matrix by a **linear combination** of "latent" factors (e.g. users' preferences) that are extracted by MC **in a nonlinear way**, using the training dataset (i.e. the way the user's preferences are learned from the preferences of other users is nonlinear):



Basic Notation

(i) Element RTA_{ij}^{t} of Revealed Technological Advantage matrix $\mathbf{RTA}^{t}$: *city's i nr. of patents in technology j relative to total market share at time t*;

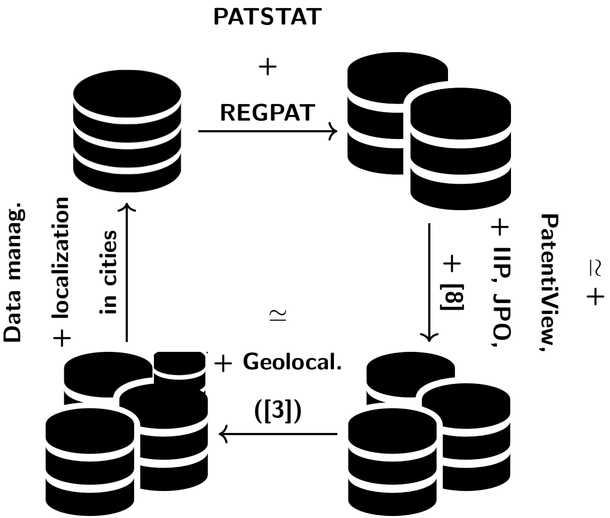
(ii) Competitiveness matrix at time t: $\mathbf{M}^{t}$;

(iii) $\mathbf{M}^{t+5}$: the incidence matrix derived by setting M_{ij}^{t+5} to 1 if $RTA_{ij}^{t+5} \geq 1$, and to 0 otherwise.

AIM: Predicting future competitiveness matrix elements $\mathbf{M}^{t+5}$ using current data $\mathbf{M}^{t}$, with 5-year[1] forecasts based on discretized attributes from the **RTA** matrix $\mathbf{RTA}^{t}$, to reflect long-term investment impacts on urban competitiveness

[1] 5 years is the estimated time needed for investments to significantly impact the competitiveness structure of cities.

Data



PATSTAT
+
REGPAT
+ IIP, JPO,
PatentiView,
+ [8]
≃
+
+ Geolocal.
([3])
≃
Data manag.
+ localization
in cities

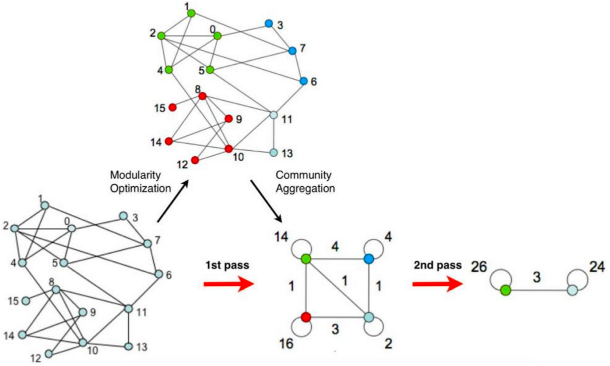
Continent	Number of cities	Time period	Number of patents (thousands)	Average employment (thousands)	Average net migration (thousands)
Africa	6	2000-2004	1.62	1431.78	2.52
		2005-2009	1.76	1632.11	30.78
		2010-2014	2.82	1759.73	38.15
Asia	45	2000-2004	1320.61	3145.15	97.21
		2005-2009	1400.51	3722.14	125.21
		2010-2014	1362.48	4358.46	64.87
Europe	48	2000-2004	230.21	1210.84	10.46
		2005-2009	268.17	1284.38	11.72
		2010-2014	269.01	1310.01	8.36
North America	34	2000-2004	476.99	2069.75	1.83
		2005-2009	504.49	2136.25	2.49
		2010-2014	531.18	2171.25	4.52
Oceania	7	2000-2004	11.53	956.70	11.96
		2005-2009	11.77	1083.13	22.92
		2010-2014	12.58	1186.06	24.11
South America	10	2000-2004	2.13	3121.41	2.86
		2005-2009	3.23	3561.11	5.70
		2010-2014	5.12	4011.93	11.82

Table: Descriptive statistics.

Pre-processing

- $\mathcal{H}$ow? Louvain-clustering
- $\mathcal{W}$hat? Identifying cities that are *similar* to any city i;
- $\mathcal{W}$hy? In order to enhance the prediction power of the models by facilitating their job;
- $\mathcal{W}$here? In the matrix $\mathbf{NRTA}^t := \mathbf{RTA}^t(\mathbf{RTA}^t)' \in \mathbb{R}^{City \times City}$, where City is the number of cities. Number of IPC technological areas in which city i and city j have a competitive technology in common;
- $\mathcal{W}$hen? Before applying the supervised machine-learning models.

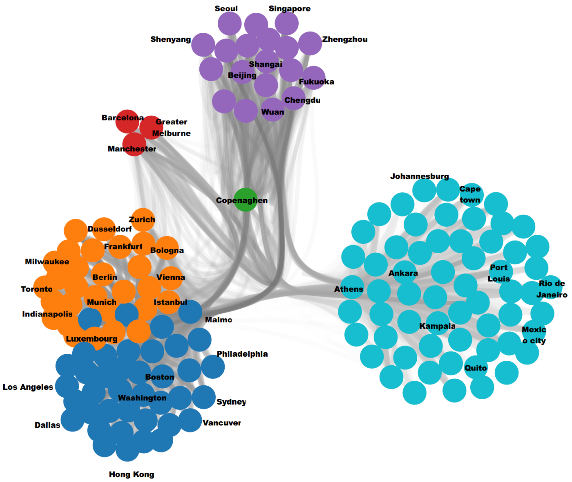
Idea of pre-processing: cluster cities by maximizing modularity $Q^t = \frac{1}{2S^t}\sum_{i,j}\left[Adj^t_{ij} - \frac{k^t_i k^t_j}{2S^t}\right]\delta(c^t_{k(i)}, c^t_{k(j)})$ $\rightarrow$ aggregate the so found clusters $\rightarrow$ repeat until no more modularity gain. Below the representation of a **single** iteration:



Gain in modularity:

$$\Delta Q^t(i,c) = \left[\sum_{j\in c^t} Adj^t_{ij} - \frac{k^t_i \sum_{j\in c^t} k^t_j}{2S^t}\right] - \left[\sum_{j\in c^t_{k(i)}} Adj^t_{ij} - \frac{k^t_i \sum_{j\in c^t_{k(i)}} k^t_j}{2S^t}\right]$$

Example of cluster for $t = 2014$:



Once the optimal partition has been determined for every $t \in \{2000, 2001, \ldots, 2008\}$, we performed a majority voting by counting the number of times that every pair of cities belonged to the same cluster.

Matrix Completion (MC)

MC is used to complete a partially observed matrix. It does so by minimizing the trade off between a data fitting term (in red) and a regularization term (usually nuclear norm in blue).

MC models employed

Soft-impute ([7])

$$\underset{\mathbf{Z}}{\text{minimize}} \; \frac{1}{2}\sum_{(i,j)\in\Omega^{\mathrm{tr}}}(A_{ij}-Z_{ij})^2+\lambda\|\mathbf{Z}\|_*$$

([2]) FE

$$\underset{\mathbf{Z},\mathbf{L},\boldsymbol{\Gamma},\boldsymbol{\Delta}}{\text{minimize}} \; \left(\frac{1}{|\Omega^{\mathrm{tr}}|}\sum(A_{i,j}-Z_{i,j})^2+\lambda\|\mathbf{L}\|_*\right),$$

$$\text{subject to } \mathbf{Z}=\mathbf{L}+\boldsymbol{\Gamma}\mathbf{1}^\top+\mathbf{1}\boldsymbol{\Delta}^\top$$

A is partially observed and reconstructed by **Z**. The second model introduces Fixed Effects (FE) to reduce regularization bias optimally.

Detailed Analysis of MC Model Applications

- **MC Models Variability:** Applied with different $\mathbf{A}$ matrices (one for each choice of city and year), training sets Ω^{tr}, and regularization parameters λ.
- **Tr.set Construction:** For each city and year, Ω^{tr} includes 75% of row of the 50 most similar cities (found in pre-processing). Specifically we generated $R = 500$ unique training sets by randomly choosing the 75% rows. Validation and test were chosen among remaining rows.
- **Optimization:** Identified optimal λ by minimizing Root Mean Square Error (RMSE) on validation set Ω^{val}.
- **Predictive Focus:** Aimed at 5-year predictions, using elements from $\mathbf{RTA}^{t+5}$ as ground truth for minimizing RMSE.
- **Final Testing:** Applied MC for $t = 2009$ with the optimal λ (λ°), which most frequently minimized RMSE.
- **Classifier Construction:** Formed a multi-class classifier from test set predictions; later simplified into a binary classifier.

MC vs RF

M(atrix)**C**(ompletion)

- Input: A portion of a matrix.

R(andom)**F**(orest)

- Input: Feature vectors.

MC vs RF

M(atrix)C(ompletion)

- Input: A portion of a matrix.
- Training Set: Based on specific indices.

R(andom)F(orest)

- Input: Feature vectors.
- Training Set: Based on bootstrap sampling from features.

MC vs RF

M(atrix)**C**(ompletion)

- Input: A portion of a matrix.
- Training Set: Based on specific indices.
- **Classification**: Focused on minimizing RMSE, with the ground truth being values observed 5 years later.

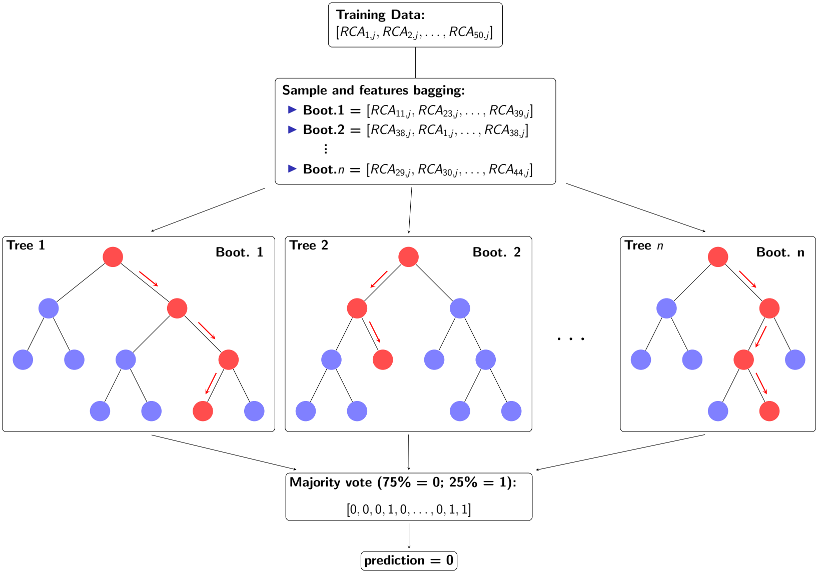
R(andom)**F**(orest)

- Input: Feature vectors.
- Training Set: Based on bootstrap sampling from features.
- **Classification**: Based on majority voting.

Benchmark model: Random Forest (RF)

- **Input Adaptation for MC and RF:** Adapted inputs to ensure comparability between MC and RF models.
- **RF Model Features:**
  - For each city i and IPC j, a 50-dimensional feature vector is constructed.
  - Vector elements: Column j from matrix $\mathbf{A}$, excluding the target element.
  - Target element (at $t+5$) is used as the desired label.
- **RF Hyperparameters:** Tuned number of trees, tree depth, and split quality criteria for optimal performance.
- **Training and Testing:**
  - Trained RF model on similar cities (as in MC) for $t = 2000, 2001, \ldots, 2008$ using cross-validation to tune hyperparameters.
  - Applied optimal hyperparameters for $t = 2009$, predicting for test set at $t + 5 = 2014$.

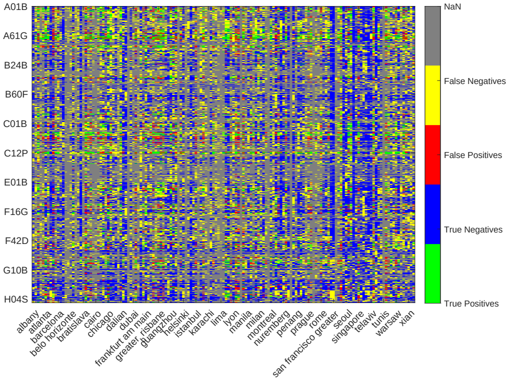
RF in our context (an example)



Results

MC by [2] with pre-processing performs better:

		Random Forest (benchmark)	Matrix Completion (Mazumder et al., 2010)	Matrix Completion (Athey et al., 2021)
Scenario I	**Avg. F1-score**	0.34	0.39	0.42
Scenario II	**F1-score**	0.34	0.67	0.70
	Precision Recall (PR) AUC	0.33	0.65	0.63
	Matthew's coefficient	0.24	0.29	0.31



The figure presents a comparison of the configurations of true positives (green), true negatives (blue), false positives (red), and false negatives (yellow) obtained when utilizing the RTA values of 2014 as the ground truth in the binary classifier derived from MC of [2].

Rank	Predicted competitiveness (MC, Athey et al., 2021)	Actual competitiveness	Predicted ubiquity (MC, Athey et al., 2021)	Actual ubiquity
1^{st}	Shanghai	Chongqing	A61K Preparation for medical, dental or toiletry purposes	A61K Preparation for medical, dental or toiletry purposes
2^{nd}	Chicago	Guangzhou	A61P Specific therapeutic activity of chemical compounds or medicinal preparations	A61P Specific therapeutic activity of chemical compounds or medicinal preparations
3^{rd}	Munich	Dalian	C12Q Measuring or testing processes involving enzymes, nucleic acids or microorganisms	C12Q Measuring or testing processes involving enzymes, nucleic acids or microorganisms
4^{th}	Guangzhou	Chengdu	C07K Peptides	C07K Peptides
5^{th}	Seoul	Chicago	A01N Preservation of bodies of human or animals or plants or parts thereof	H02M Apparatus for converting electrical power, e.g., from DC to AC
6^{th}	Los Angeles Greater	Shanghai	C07D Heterocyclic compounds	C07H Sugars; derivatives thereof; nucleosides; nucleic acids
7^{th}	Paris	Frankfurt	C12N Microorganisms or enzymes; compositions thereof; mutation or genetic engineering	G01N Investigating or analyzing materials by determining their chemical or physical properties
8^{th}	Atlanta	Jinan	G01N Investigating or analyzing materials by determining their chemical or physical properties	A61F Filters implantable into blood vessels; prostheses and similar devices
9^{th}	Frankfurt	Milan	A61J Containers specially adapted for medical or pharmaceutical purposes and similar devices	C12N Microorganisms or enzymes; compositions thereof; mutation or genetic engineering
10^{th}	Tokyo	Shenyang	B01D Separation	A61L Methods or apparatus for sterilising materials or objects in general

Discussion & Conclusions

- **Contribution:** (i) Unique dataset; (ii) Framework for defining competitiveness among urban cities; (iii) Prediction of future competitiveness of global cities across technological areas without major structural assumptions;
- **Approach:** (i) Integration of various data sources; (ii) Adoption of RTA in a complexity framework; (iii) Integration of MC and Louvain community detection;
- **Performance:** Superior prediction accuracy compared to benchmark (Random Forest) under similar pre-processing;
- **Policy Implications:**
  - Design of tailored strategic innovation policies for individual cities;
  - Map of future excess supply (demand) in cities;
  - Tracing the mobility of inventors.

THANK YOU FOR THE ATTENTION

For more visit my website ▸ HERE, *at*

`https://www.federiconutarelliphd.com/`

References I

[1] Francesco Paolo Appio, Marcos Lima, and Sotirios Paroutis. "Understanding Smart Cities: Innovation ecosystems, technological advancements, and societal challenges". In: *Technological Forecasting and Social Change* 142 (2019), pp. 1–14.

[2] Susan Athey et al. "Matrix completion methods for causal panel data models". In: *Journal of the American Statistical Association* 116.536 (2021), pp. 1716–1730.

[3] G. De Rassenfosse, J. Kozak, and F. Seliger. "Geocoding of worldwide patent data". In: *Scientific Data* 6.1 (2019), pp. 1–15.

[4] Nigel Harris. "City Competitiveness". In: *Originally drafted for a World Bank study of competitiveness in four Latin American cities* (2007).

References II

[5] Magdalena Kachniewska, Arkadiusz Michał Kowalski, and Ewelina Szczech-Pietkiewicz. "The Competitiveness of Cities: Components, Meaning and Determinants". In: (2018).

[6] M Kamiya et al. "Global Urban Competitiveness Report (2019–2020)". In: *UN HABITAT: Nairobi, Kenya* (2020).

[7] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. "Spectral regularization algorithms for learning large incomplete matrices". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2287–2322.

[8] G. Morrison, M. Riccaboni, and F. Pammolli. "Disambiguation of patent inventors and assignees using high-resolution geolocation data". In: *Scientific Data* 4 (2017).

[9] Hal R Varian. "Computer mediated transactions". In: *American Economic Review* 100.2 (2010), pp. 1–10.

References III

[10] Youngjin Yoo et al. "Organizing for innovation in the digitized world". In: *Organization Science* 23.5 (2012), pp. 1398–1408.

[11] F. Zhang, Y. Wang, and W. Liu. "Science and technology resource allocation, spatial association, and regional innovation". In: *Sustainability* 12.2 (2020), p. 694.