China Data Lab

HARVARD UNIVERSITY

Workflow Data Analysis with KNIME

Presenter: Tao Hu and Lu Chen, and Shuming Bao

Date: 2019-11-21

https://chinadatalab.net

Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Challenges for Data Research and Teaching

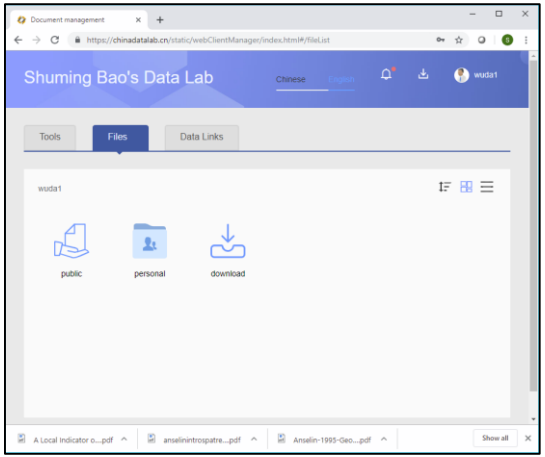
- **Data Sharing**
  - Licensed data
  - Restricted data
  - Sensitive data
  - Large size data
  - Research data generated from different projects
- **Tool Sharing**
  - Licensed and free tools
  - Integrated environment for tools for data
  - Maintenance and updates
- **Research Results Sharing**
  - Research (**reproducible, replicable, generalizable**)
  - Teaching (students with different interests and skills)
  - Decision support (efficient, effective, and expandable)

An Integrated Platform for Research and Teaching

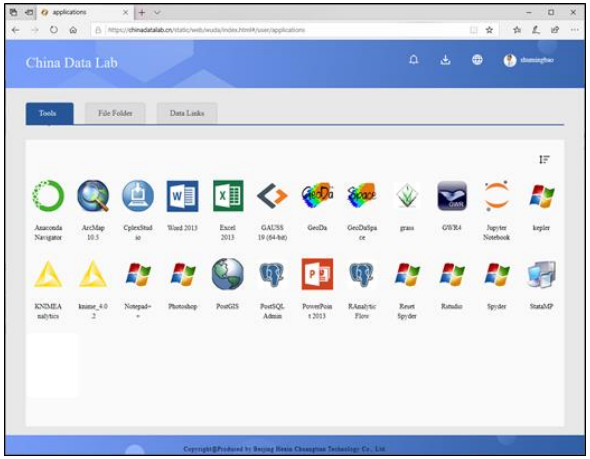
Main Features:

- ❑ Data available only on the cloud
- ❑ Tools available on the cloud
- ❑ All computation are on the cloud
- ❑ No maintenance required for end users

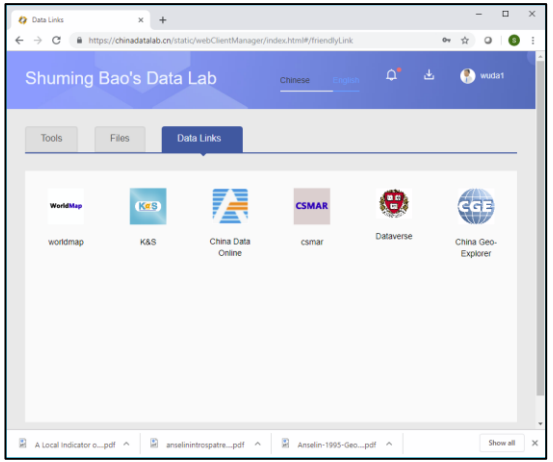
Data Center



Tool Center



Sharing Center



A Comparison of Workflow Tools (I)

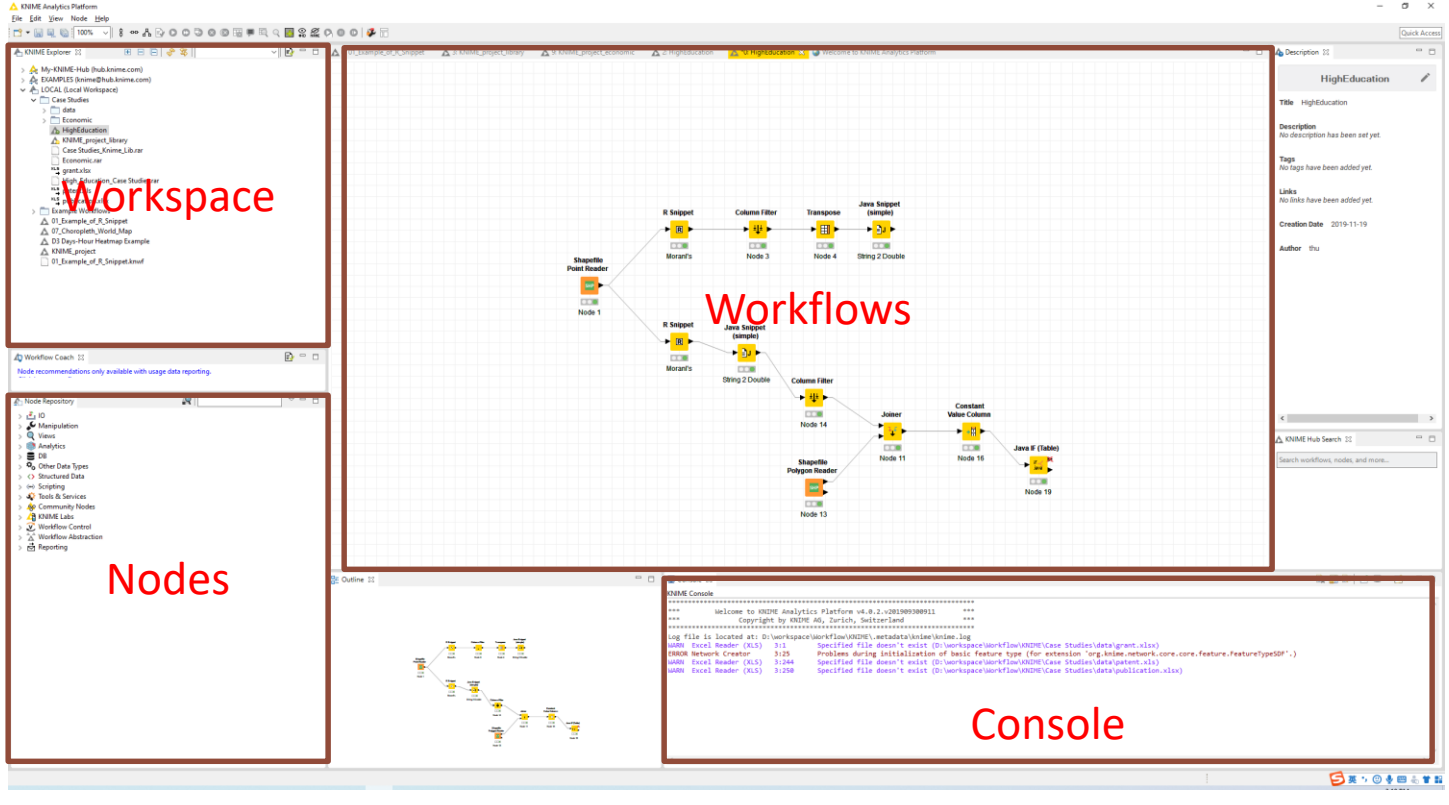
Name	Category	Free	Open Source	Education Licence	Popularity	Open source community	Latest version	Website
KNIME analytics	***Data analytics***	✔	✔	✔	✔✔	✔✔	***4.0.2 (2019)***	***https://www.knime.com***
Talend Open Studio	Data analytics	✔	✔	✔	✔✔✔	✔✔	7.3 (2019)	https://www.talend.com/products/talend-open-studio/
Pentaho Kettle	ETL	✔	✔	✔	✔✔	✔✔✔	8.3 (2019)	https://community.hitachivantara.com/s/article/data-integration-kettle
GeoKettle	Spatial ETL	✔	✔	✔			2.5 (2015)	http://www.spatialytics.org/
RapidMiner Studio	Data analytics	Limited		✔	✔✔✔		9.4 (2019)	https://rapidminer.com/
Tableau Prep	ETL			✔	✔✔✔		2019.3.2 (2019)	https://www.tableau.com/products/prep
Alteryx Designer	Data analytics			✔	✔✔✔		2019.3 (2019)	https://www.alteryx.com/
Dataiku DSS	Data analytics	Limited		✔	✔✔		5.1(2019)	https://www.dataiku.com/

A Comparison of Workflow Tools (II)

Name	Data Input/Output	Data Manipulation	Spatial Processing	Regression Analysis	Machine Learning	R	Python	Charts	Reporting
KNIME analytics	***Tabular data, shapefile***	✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔	✔✔	✔✔
Talend Open Studio	Tabular data, shapefile	✔✔✔	✔	✔	✔✔✔	✔	✔	✔✔	✔✔
Pentaho Kettle	Tabular data, shapefile	✔✔✔	✔	✔	✔	✔	✔	✔✔	✔✔
GeoKettle	Tabular data, shapefile	✔	✔✔✔					✔	✔
RapidMiner Studio	Tabular data, shapefile	✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔	✔✔✔	✔✔
Tableau Prep	Tabular data	✔✔						✔	
Alteryx Designer	Tabular data, shapefile	✔✔✔	✔✔✔	✔✔	✔✔	✔✔	✔✔	✔✔	✔✔
Dataiku DSS	Tabular data, shapefile	✔✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔✔	✔✔✔	✔✔

An Introduction to KNIME

https://www.knime.com



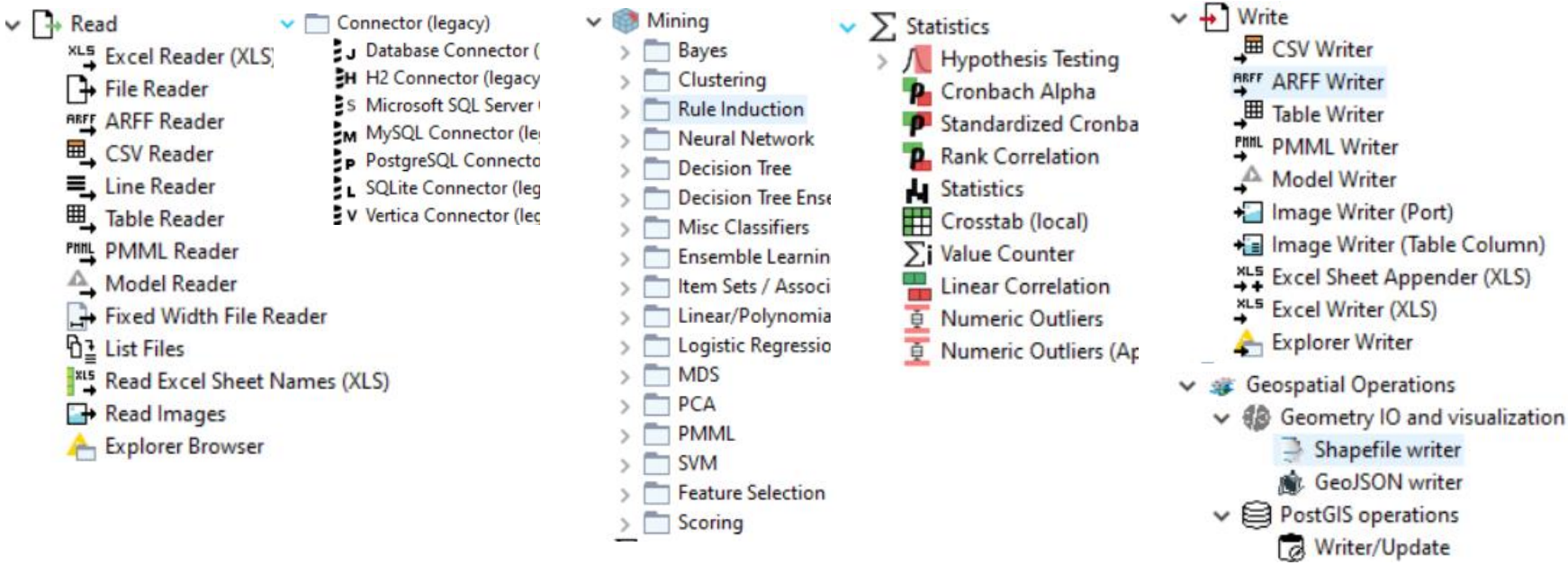
Input and Output

Input/Ouput	Category	Data Format
❑ Input Data	• Local File Reader	Excel; CSV; Table; PMML; Images; File; etc.
	• Database Connector	MySQL; SQLite; PostgreSQL; H2; SQL Server; etc.
	• Big Data Connector	HDFS; webHDFS; HttpFS; HDFS; Hive; etc.
	• Spatial Data Reader	Shapefile Point; Shapefile Polygon; GeoJSON; WFS Connector;
❑ Output Data	• Local File Writer	Excel; CSV; Table; PMML; Images; File; etc.
	• Database Connector	MySQL; SQLite; PostgreSQL; H2; SQL Server; etc.
	• Big Data Connector	HDFS; webHDFS; HttpFS; HDFS; Hive; etc.
	• Spatial Data Writer	Shapefile; GeoJSON; Map Viewer

Knime Core Modules (I)

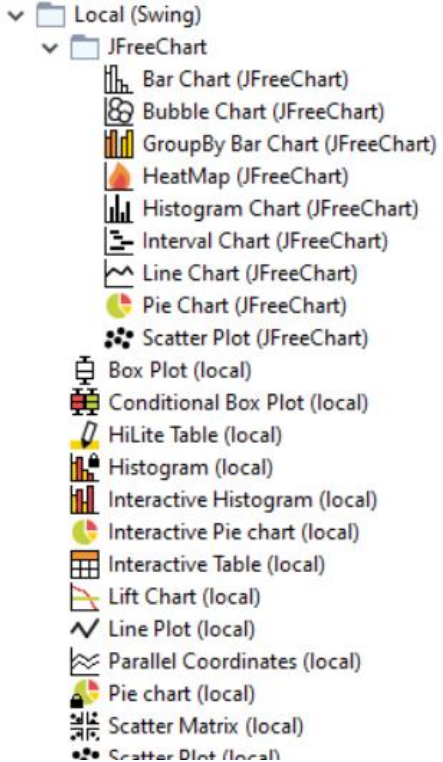
- **Node Repository**

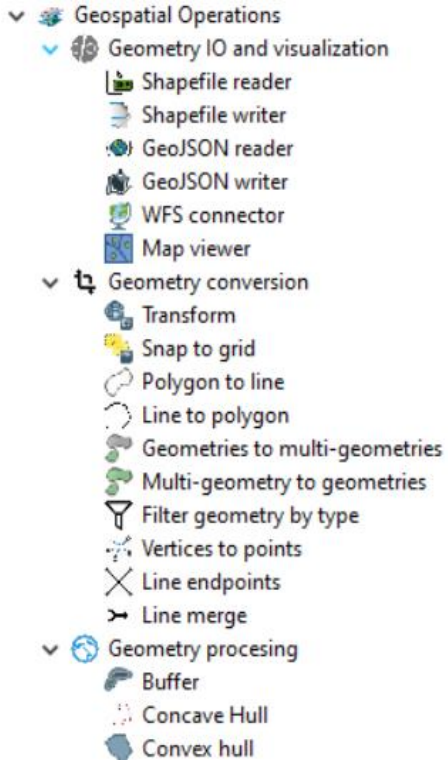
Input	Analysis	Output

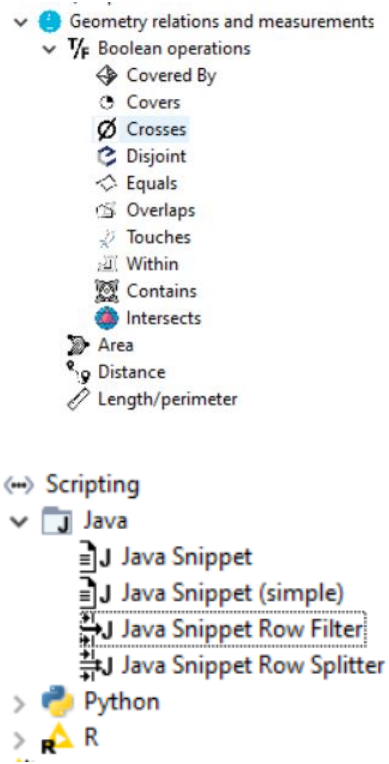


Knime Core Modules (II)

❑ Other Nodes

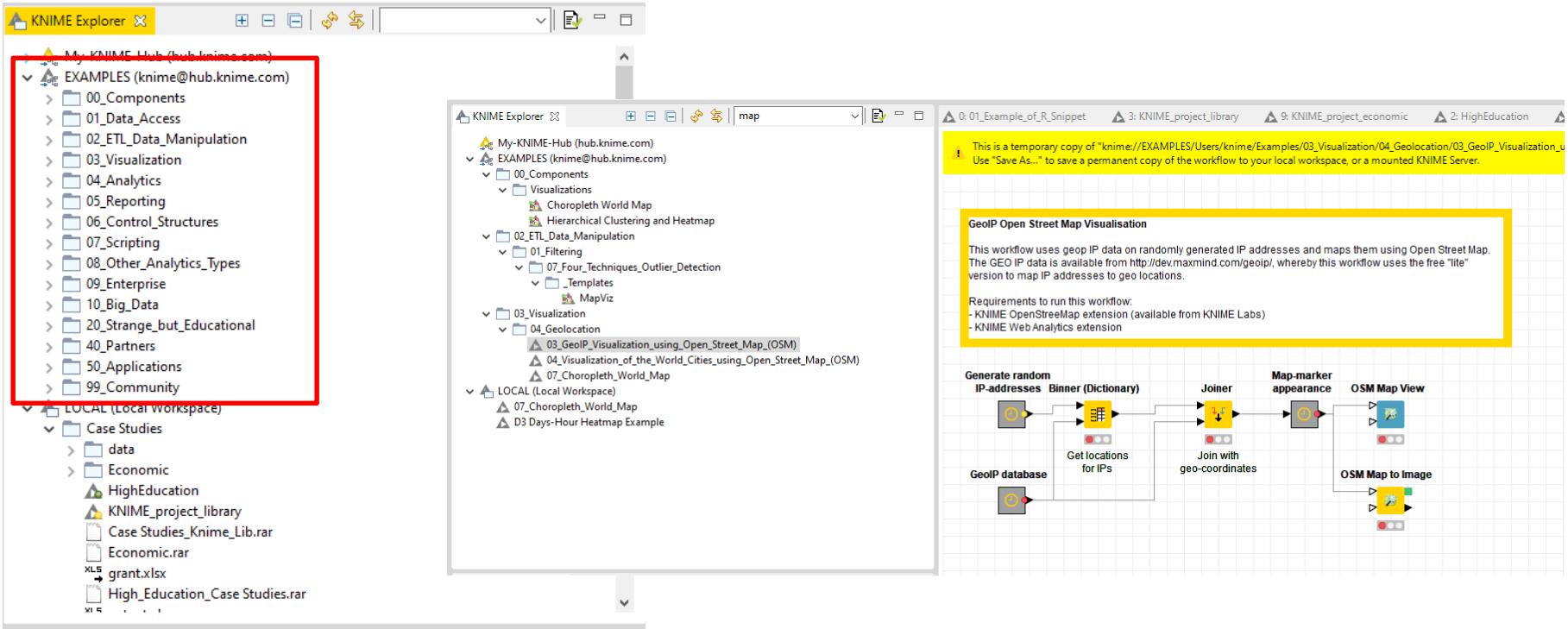




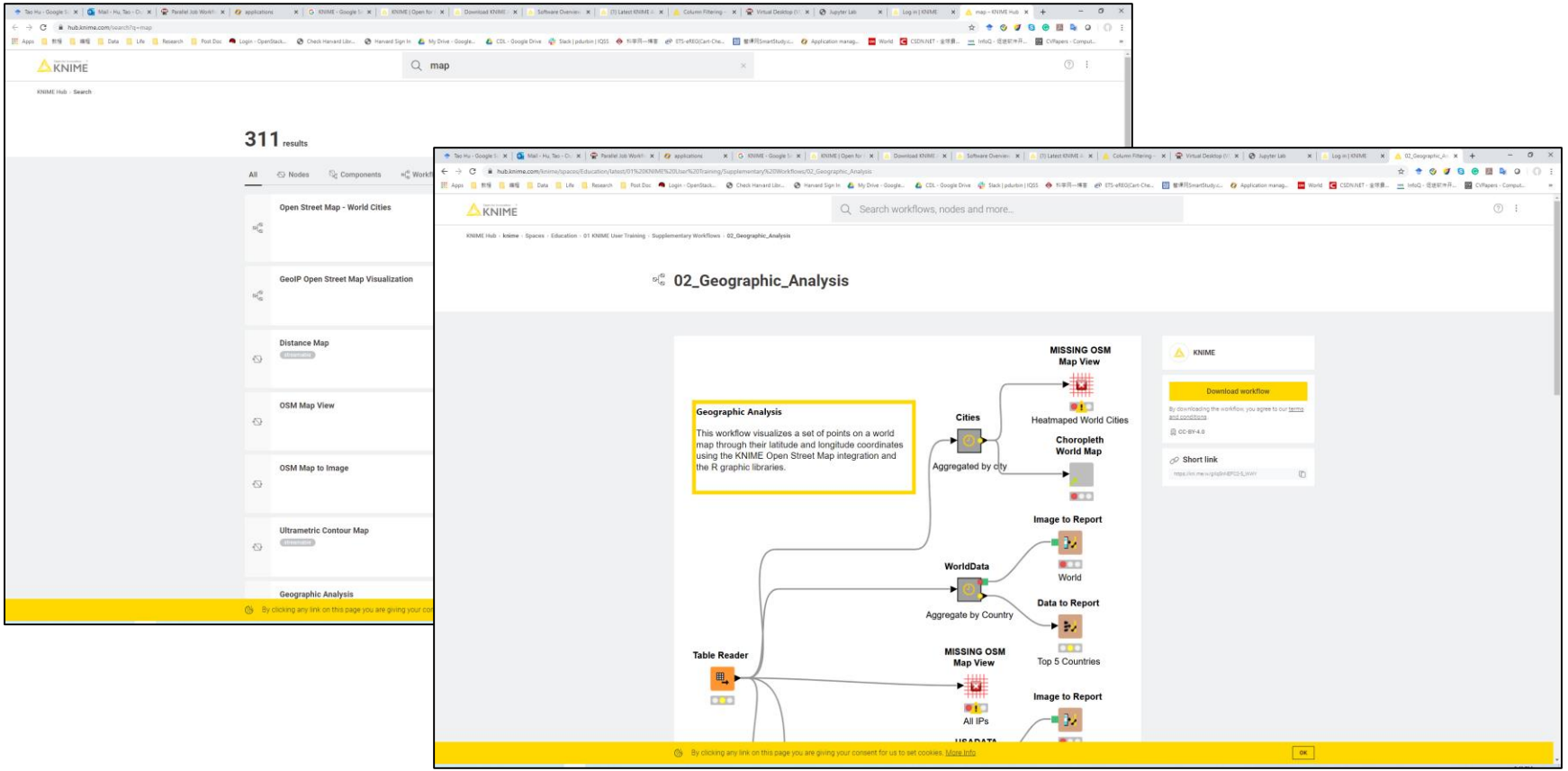


Knime Examples

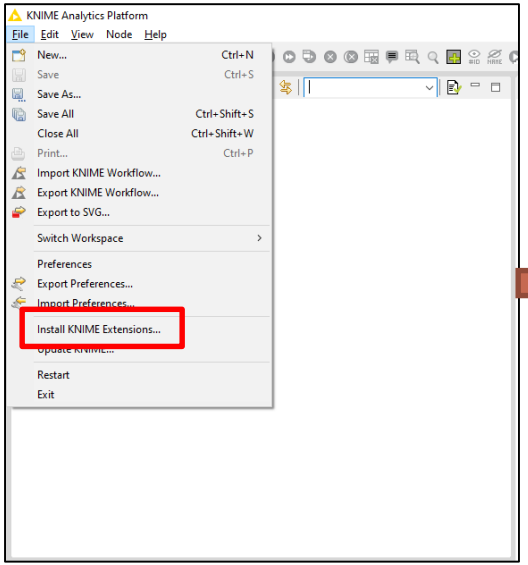
- **Workflow examples**

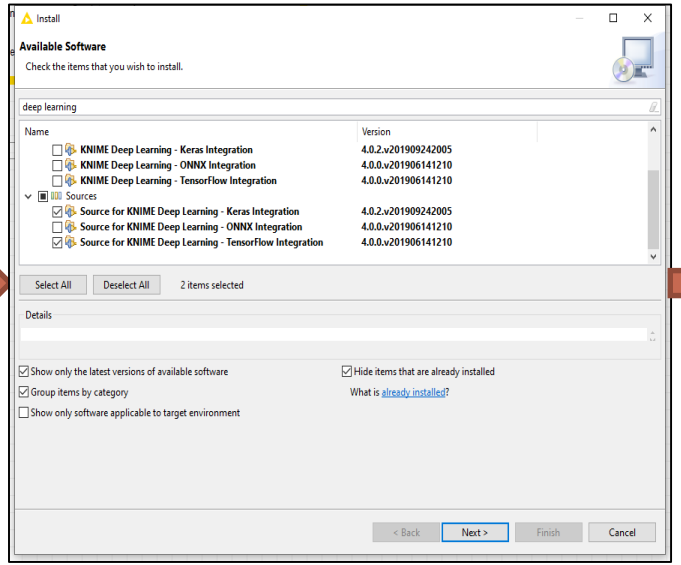


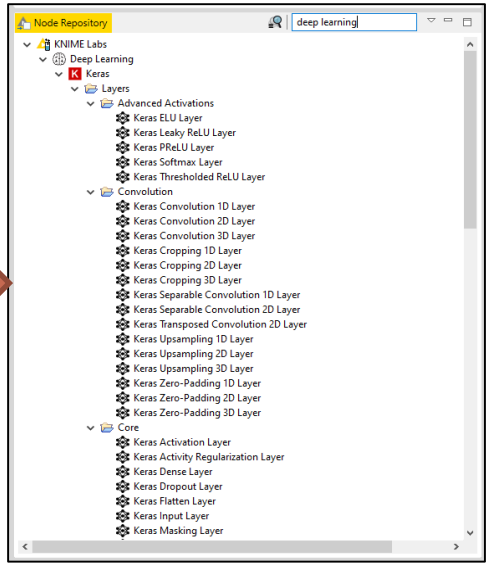
Knime Exchange Hub



Knime Extensions

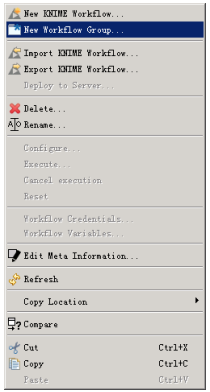


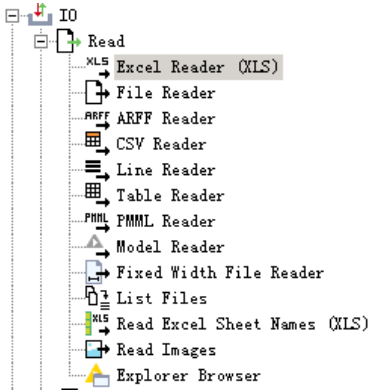


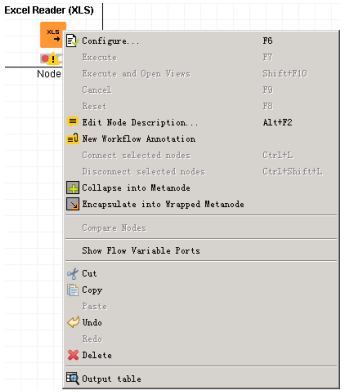


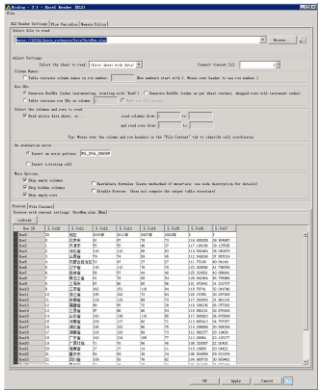
Create a Workflow

① New workflow → ② Select nodes → ③ Configure nodes →

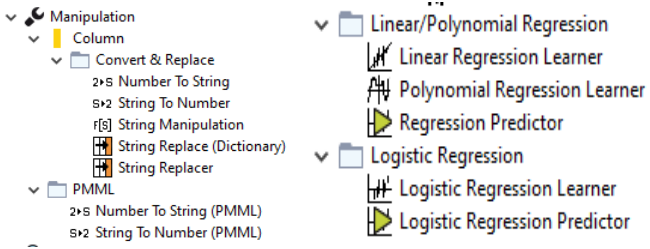


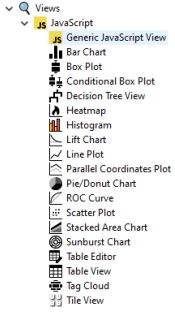


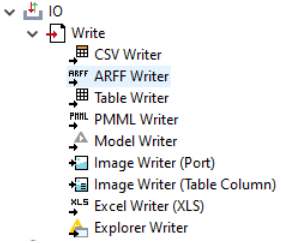


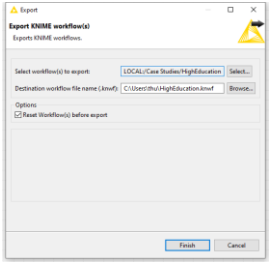


④ Data analysis → ⑤ Data visualization → ⑥ Data output → ⑦ Export workflow







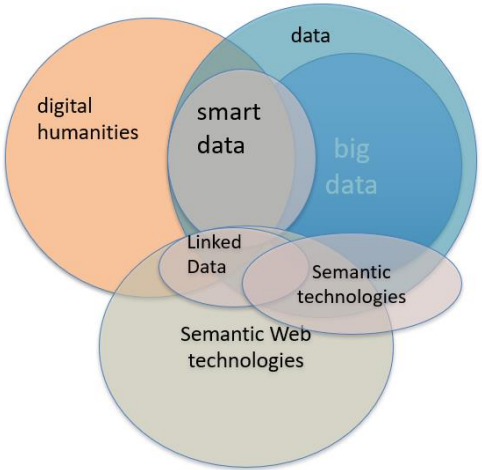


Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Network Analysis on Text Data

- **Goal**: develop and demonstrate a network framework of the historical Innovation and Invention at the Liquid Crystal Institute, Kent State University (PI: Marcia Lei Zeng, et al.)



References:

- Li, H., Zeng, M., Zhang, Y., Ye, X., & Hu, T. (2017). Tackling Innovation Networks with Smart Data: A Case Study of the Liquid Crystal Institute at Kent State University. In DH.
- Zeng, M. L., Zhang, Y., Li, H., & Polyakov, S. (2015). Exploring Smart Data Approaches to the History of Innovation and Invention at Liquid Crystal Institute at Kent State University. In Digital Libraries: Providing Quality Information: The 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings (Vol. 9469, p. 346). Springer.

Objectives

- Replicate data analysis procedures using previous scientific literature data based on workflow;
- Expanded data analysis based on publication, patent, and NSF grant data;
- Applications of workflow for research and teaching related to network analysis based on publication, patent, grant data, as well as other data.

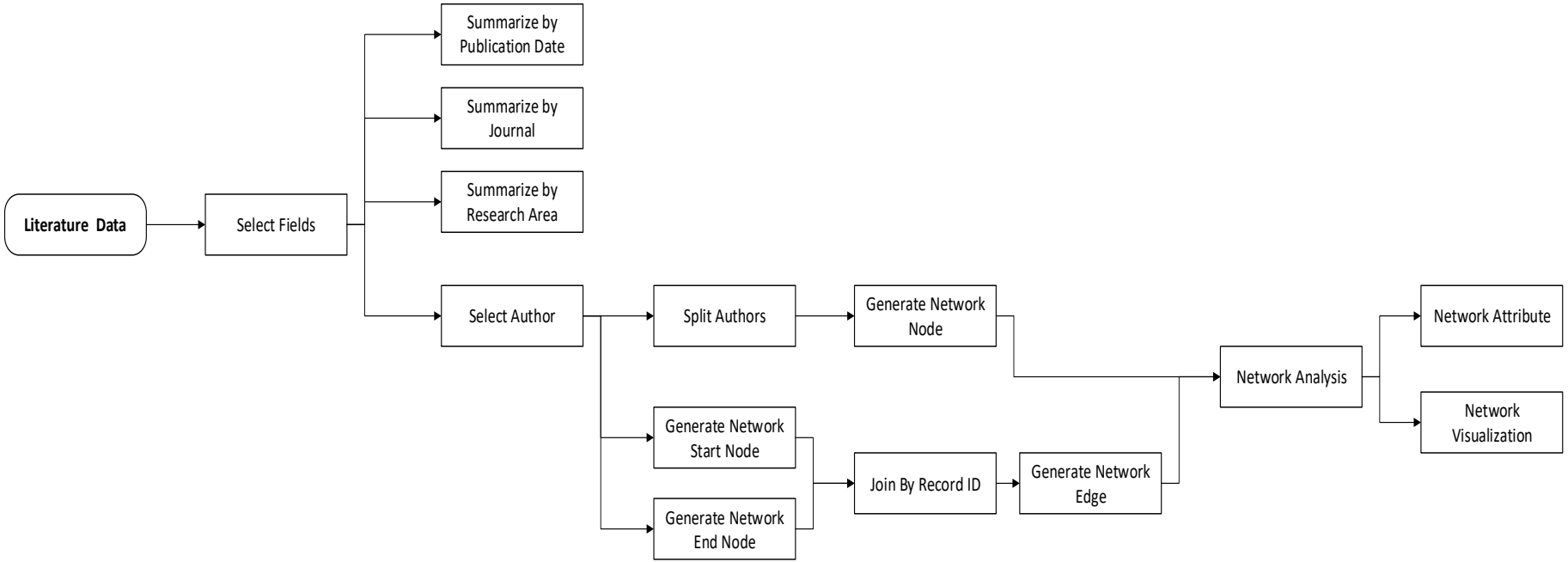
Data Sources

- **Publication Data**
  - Title
  - Author
  - Affiliation
  - Key words
  - Abstract
  - Publication Date
  - Journal
  - Volume
  - Issue
  - …
- **Patent Data**
  - Title
  - Inventor
  - Inventor Location
  - Publication Date
  - Assignee
  - Assignee Location
  - CPC
  - IPC
  - USPC
  - Abstract
  - …
- **Awarded Grants**
  - Title
  - PI
  - Co-PI
  - Email Address
  - Institution
  - NSF Organization
  - Start Date
  - Expiration Date
  - Awarded Amount
  - NSF Directorate

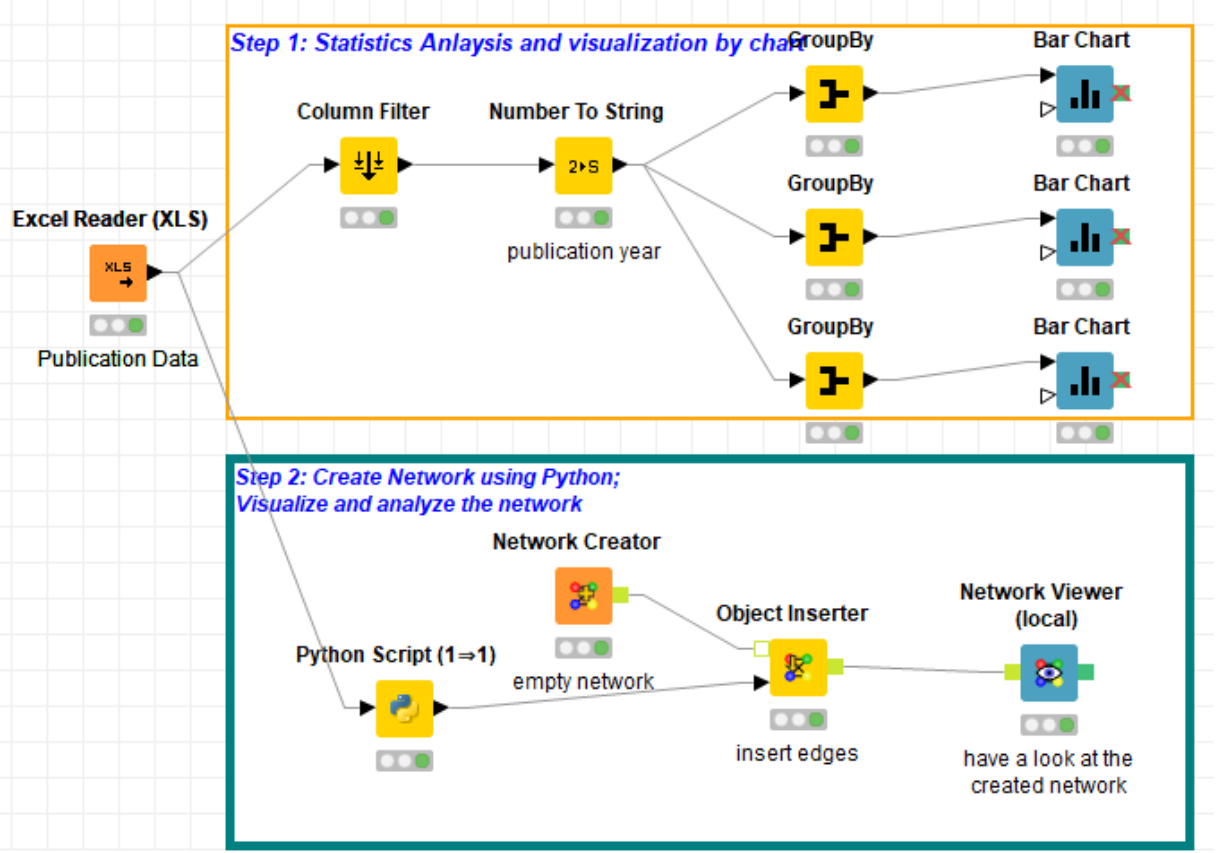
Data Input

File Name	Format	Data Sources
publication.xls	.xls	Web of Science
patent.xls	.xls	ProQuest
grant.xls	.xls	NSF website

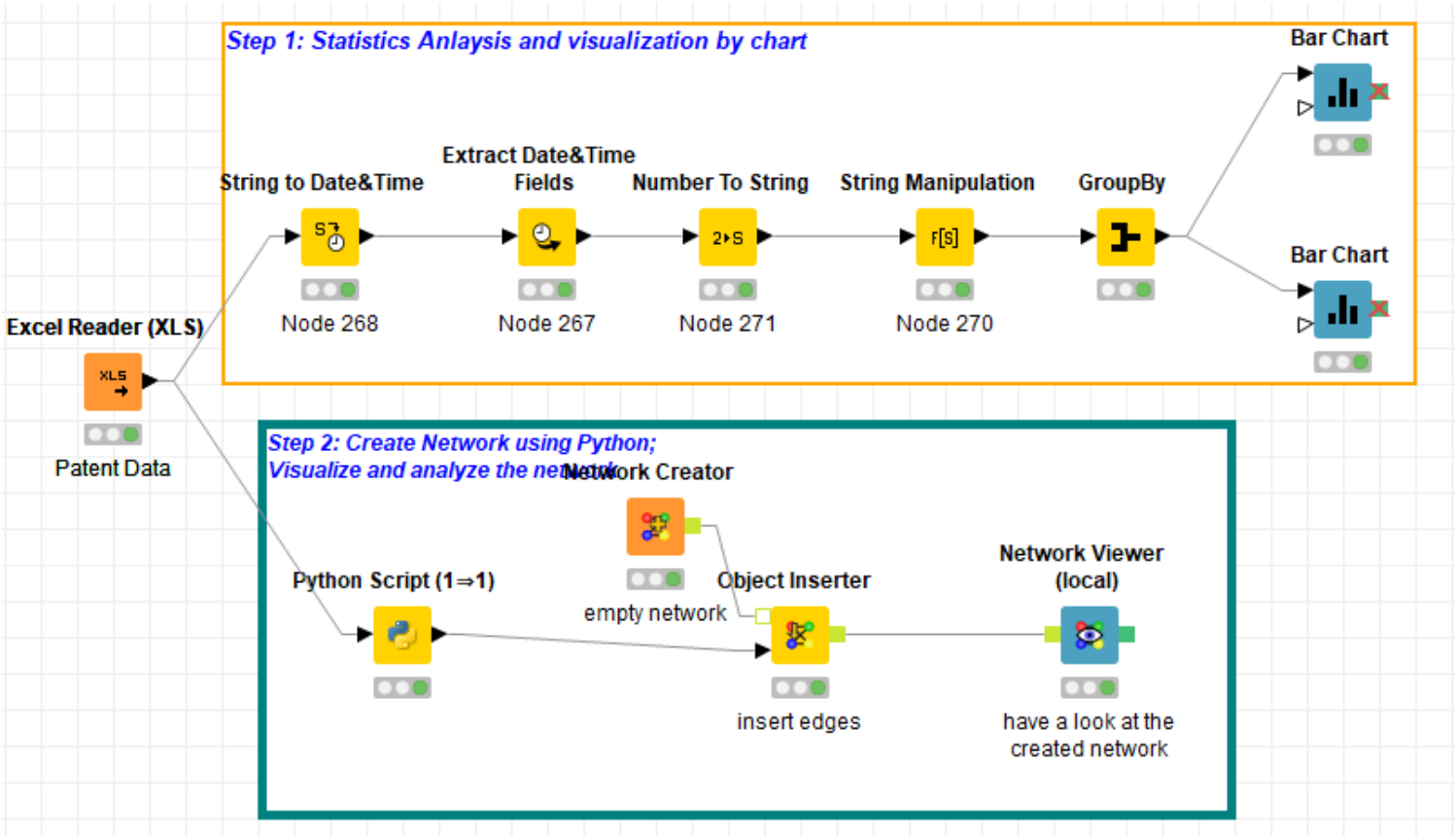
The Flowchart for Data Analysis



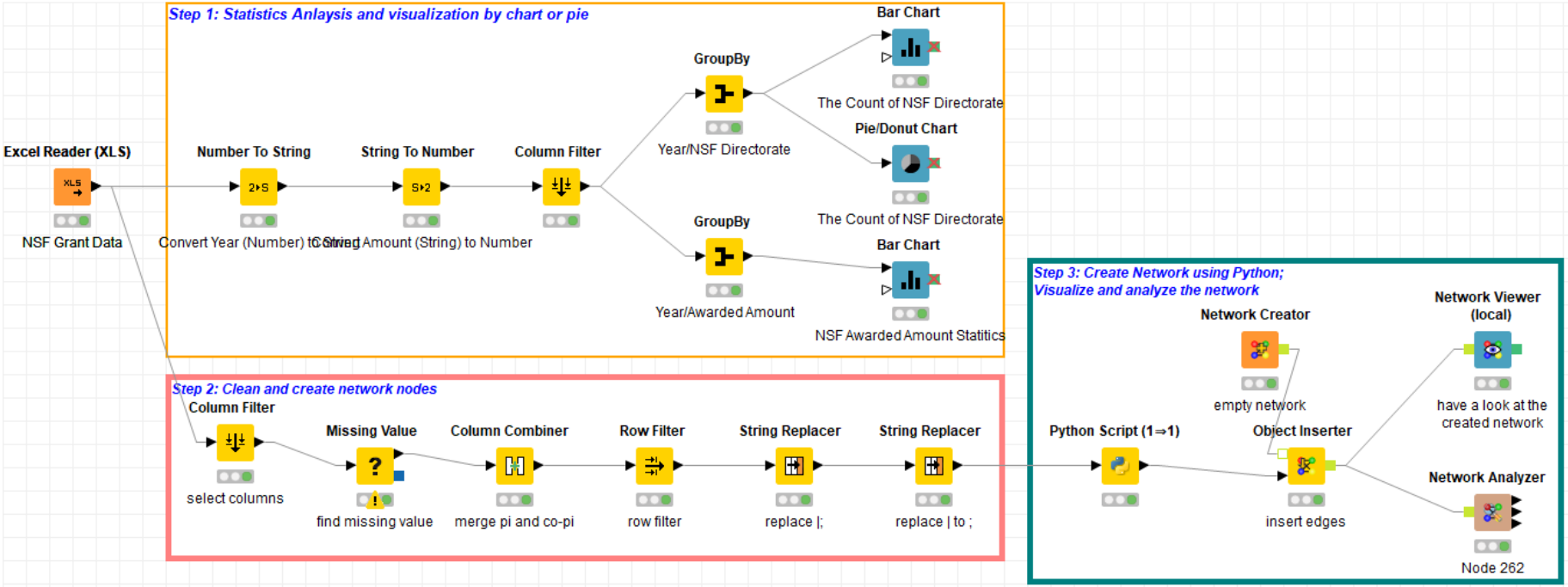
Knime Workflow for Publication Analysis



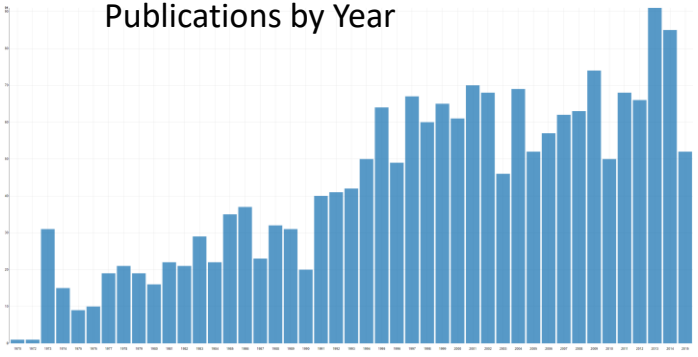
Knime Workflow for Patent Analysis

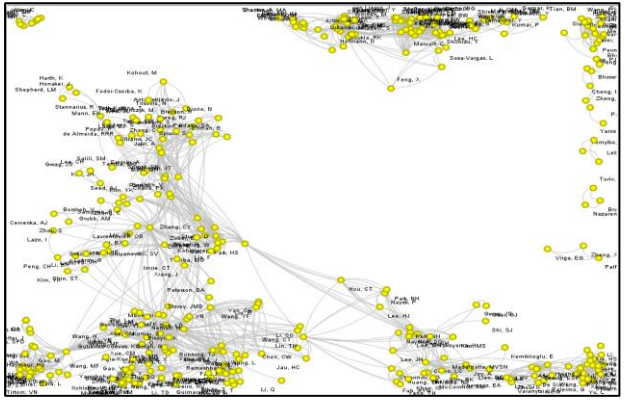


Knime Workflow for Grant Analysis



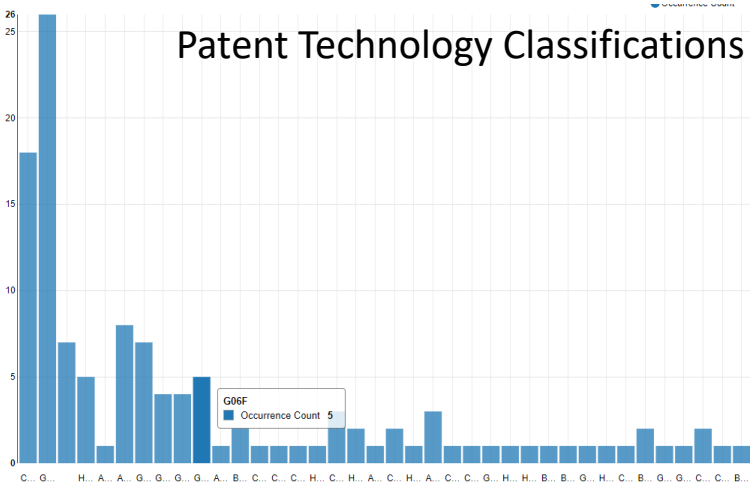
Results from Publication Analysis



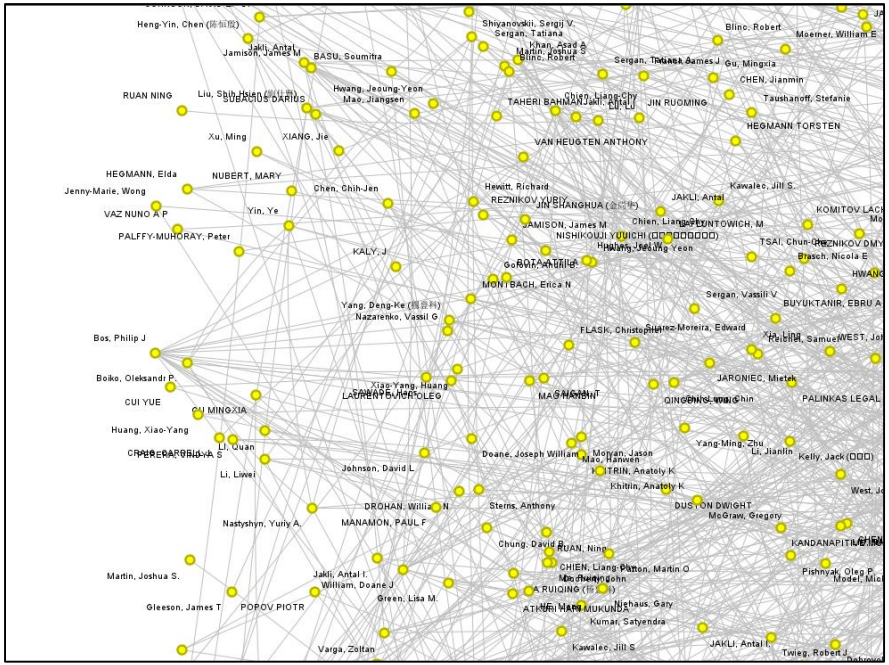


Object id	Node d...	Node d...	In degree	In degr...	Out de...	Out de...	Closen...	Node w...	Avg. n...	Clusteri...	Hub score	Authori...	Baryce...
Feng, X	5	0.326	5	0.326	5	0.326	0.483	5	1	1	0.069	0.069	0.008
Wang, F	5	0.326	5	0.326	5	0.326	0.621	5	1	1	0	0	0.04
Sharma, A	22	1.433	22	1.433	22	1.433	0.589	22	1	0.706	0.801	0.801	0.009
Yao, WH	8	0.521	8	0.521	8	0.521	0.292	8	1	1	0	0	0.001
Wang, MF	5	0.326	5	0.326	5	0.326	0.32	5	1	0.9	0	0	0.001
Lu, W	9	0.586	9	0.586	9	0.586	0.49	9	1	1	0.091	0.091	0.007
Wang, H	4	0.261	4	0.261	4	0.261	0.32	4	1	1	0	0	0.001
Antanasijev...	7	0.456	7	0.456	7	0.456	0.287	7	1	1	0	0	0.001
Reich, R	8	0.521	8	0.521	8	0.521	0.923	8	1	1	0	0	0.1
Lu, L	8	0.521	8	0.521	8	0.521	0.917	8	1	1	0	0	0.1
Lebovka, N	2	0.13	2	0.13	2	0.13	1	2	1	1	0	0	0.5
Ma, J	21	1.368	21	1.368	21	1.368	0.258	21	1	0.367	0	0	0.001
Malgras, V	7	0.456	7	0.456	7	0.456	0.375	7	1	1	0.01	0.01	0.006
Park, HS	23	1.498	23	1.498	23	1.498	0.405	23	1	0.526	0	0	0.002
Beltrano, G	18	1.173	18	1.173	18	1.173	0.517	18	1	1	0.776	0.776	0.008
Kohlmeier, A	15	0.977	15	0.977	15	0.977	0.363	15	1	1	0	0	0.002
Sampson, P	5	0.326	5	0.326	5	0.326	0.327	5	1	1	0	0	0.001
Umadevi, S	8	0.521	8	0.521	8	0.521	0.497	8	1	0.571	0.079	0.079	0.008
Moheghi, A	2	0.13	2	0.13	2	0.13	0.286	2	1	1	0	0	0.001

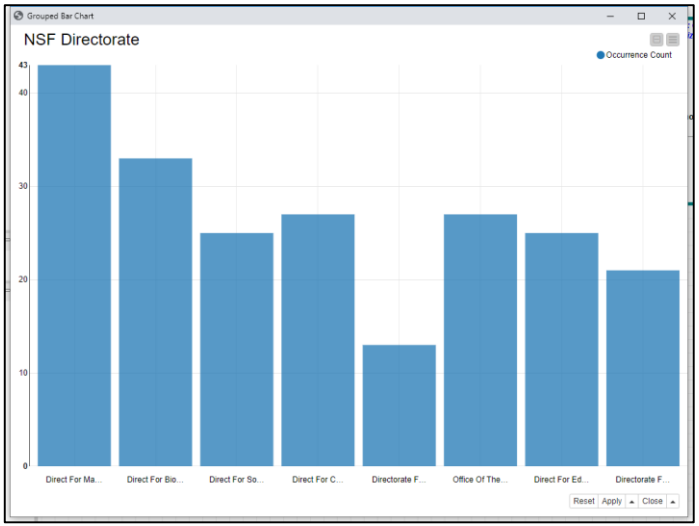
Results from Patent Analysis

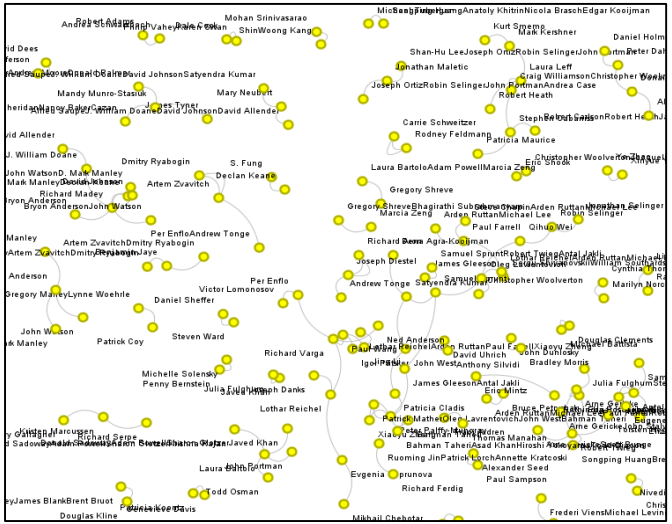


Object id	Node d...	Node d...	In degree	In degr...	Out de...	Out de...	Node w...	Avg. n...
Chao-Chiun, Liang	9	0.771	9	0.771	9	0.771	9	1
SHOKOUHIMEH...	5	0.428	5	0.428	5	0.428	5	1
Nemati, Hossein ...	9	0.771	9	0.771	9	0.771	9	1
Hwang, Jeoung ...	3	0.257	3	0.257	3	0.257	3	1
QIAN LIANGQI (...	2	0.171	2	0.171	2	0.171	2	1
Soehnlen, Eric S...	8	0.686	8	0.686	8	0.686	8	1
Chen, Cheng	4	0.343	4	0.343	4	0.343	4	1
Osher, Lawrence	9	0.771	9	0.771	9	0.771	9	1
Nastyshyn, Yuri...	6	0.514	6	0.514	6	0.514	6	1
Bhowmik, Achint...	2	0.171	2	0.171	2	0.171	2	1
Gleeson, James T	1	0.086	1	0.086	1	0.086	1	1
Dobrovolskyy, A...	3	0.257	3	0.257	3	0.257	3	1
GLEESON, Jame...	3	0.257	3	0.257	3	0.257	3	1
Kelly, Jack (□□□)	4	0.343	4	0.343	4	0.343	4	1
Li, Liwei	5	0.428	5	0.428	5	0.428	5	1
Palffy-Muhoray,...	7	0.6	7	0.6	7	0.6	7	1
Tsai, Chen Chu (...	9	0.771	9	0.771	9	0.771	9	1



Results from Grant Analysis

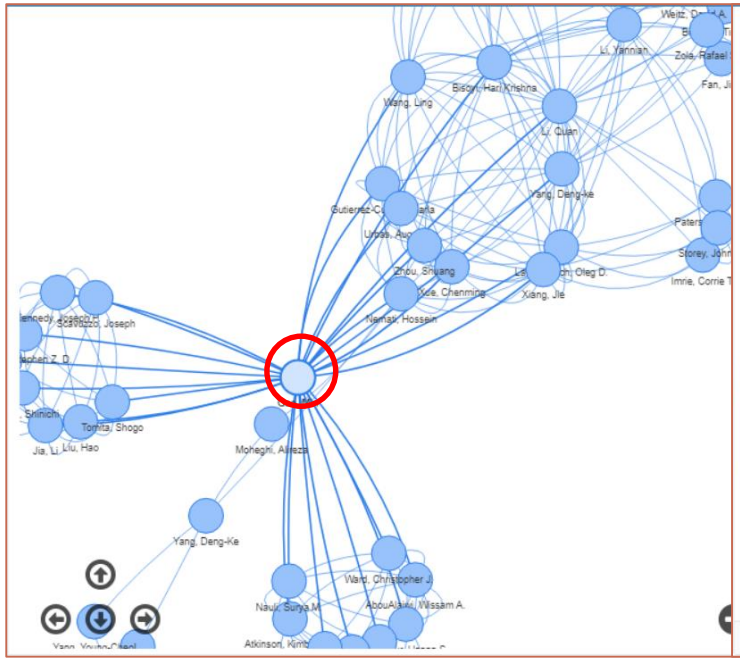




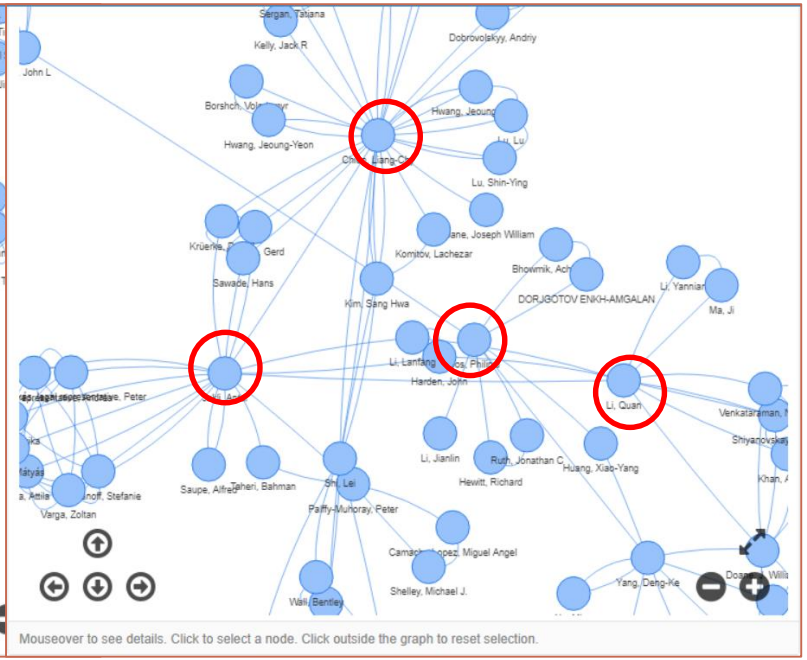
Object id	Node d...	Node degree %	In degree	In degree %	Out degree	Out degree %	Node weigh...	Avg. n...
Genevieve Davis	1	0.787	1	0.787	1	0.787	1	1
Gerassimos Pe...	1	0.787	1	0.787	1	0.787	1	1
Xiaoyu Zheng	2	1.575	2	1.575	2	1.575	2	1
Noah FriedkinE...	1	0.787	1	0.787	1	0.787	1	1
William Kalkhoff	1	0.787	1	0.787	1	0.787	1	1
Robin Selinger	2	1.575	2	1.575	2	1.575	2	1
Paul Farrell	3	2.362	3	2.362	3	2.362	3	1
Joseph OrtizD...	1	0.787	1	0.787	1	0.787	1	1
Ben FinneyMar...	1	0.787	1	0.787	1	0.787	1	1

Network Analysis of Scholars

Authors of Publications



Participants of Grants



Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Financial Analysis with Firm Data

Going public in China: Reverse mergers versus IPOs

Aim: This study examines the decision to go public in China through an initial public offering (IPO) versus a reverse merger (RM) transaction.

ELSEVIER

Journal of Corporate Finance
Volume 58, October 2019, Pages 92-111

Going public in China: Reverse mergers versus IPOs ☆

Charles M.C. Lee [a], Yuanyu Qu [b], Tao Shen [c]

Show more

https://doi.org/10.1016/j.jcorpfin.2019.04.003

Get rights and content

Highlights

- We study firms' choice to go public through reverse mergers (RMs) versus initial public offerings (IPOs) in China.
- Pre-listing RM firms are larger, more profitable, and less politically-connected than pre-listing IPO firms.
- RM firms also have superior post-listing performance, both in terms of operations and stock returns.
- These results are in sharp contrast to the evidence on RMs from developed countries.

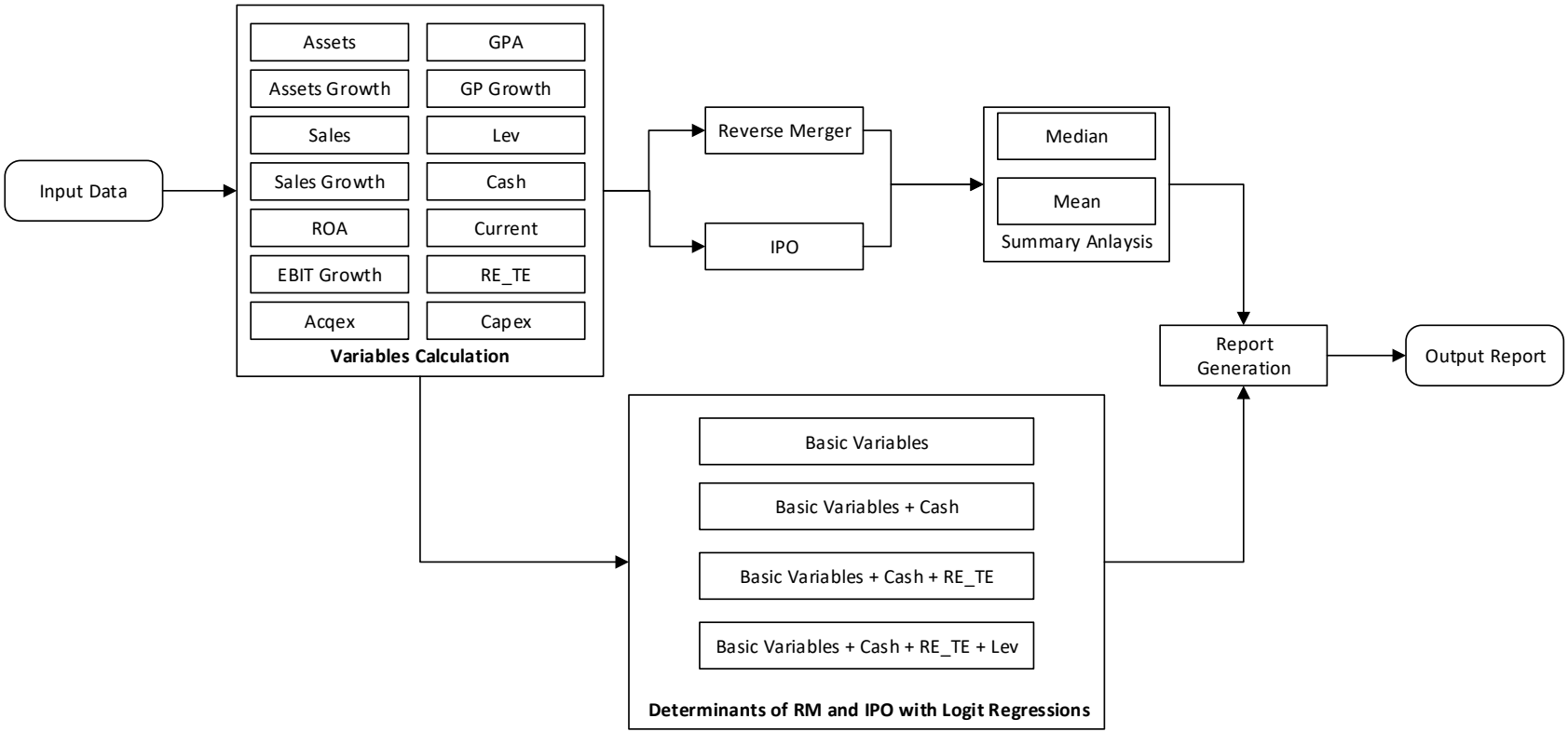
Abstract

We study firms that go public through reverse mergers (RMs) versus initial public offerings (IPOs) in China. Using a manually assembled data set, we show that pre-listing RM firms are larger, more profitable, and less politically connected than pre-listing IPO firms. Chinese RM firms also have superior post-listing performance, in terms of both operations and stock returns, compared to IPOs matched on industry and size. Unlike IPOs, RM firms do not underperform the market in the long run. These results are in sharp contrast to the evidence on RMs from developed countries. We trace these differences to China's stringent and potentially biased IPO policies, which appear to preclude even high-quality firms from accessing public markets.

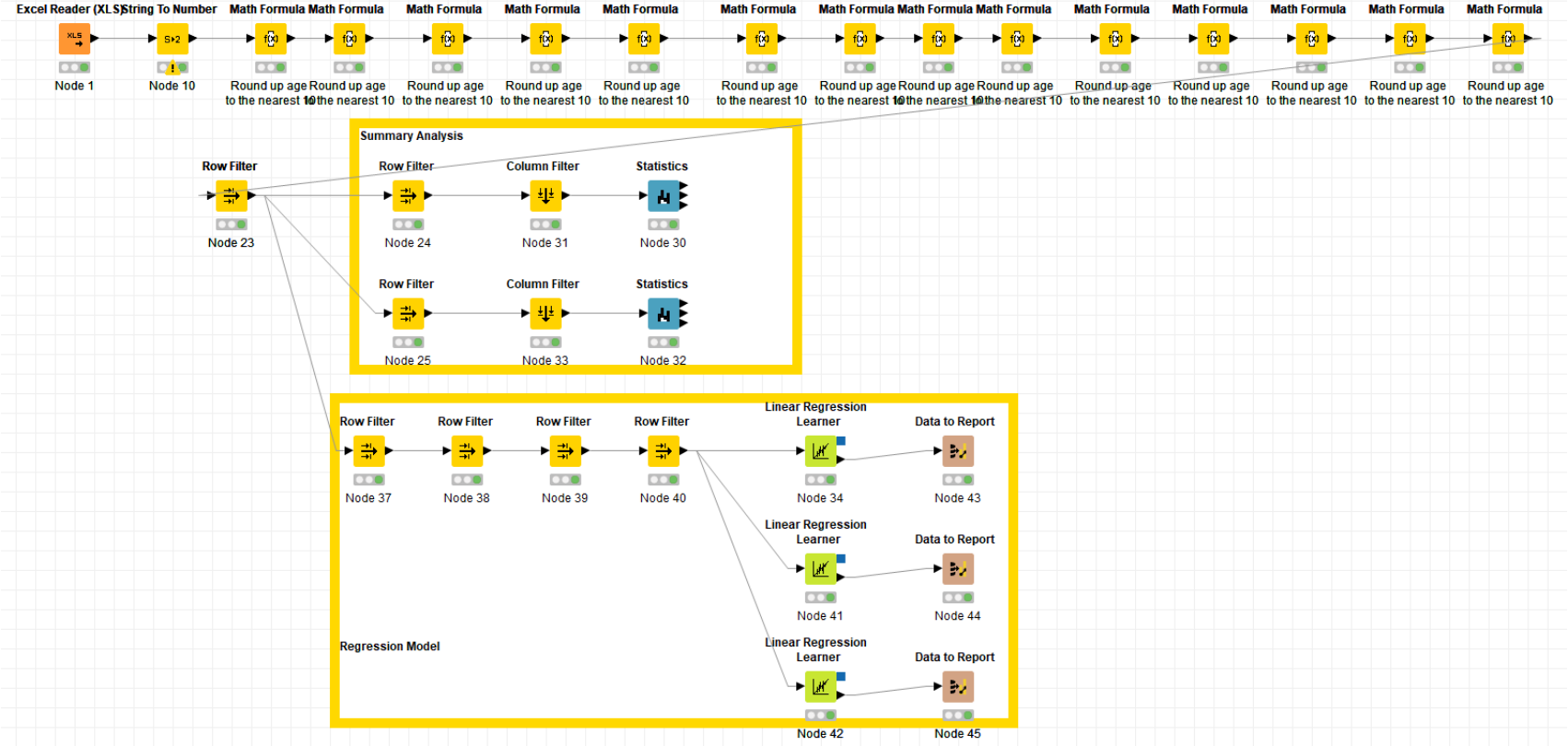
Data Sources

Name	Format	Description
IPO.xlsx	xlsx	The financial and stock returns data of listed firms are from the China Stock Market and Accounting Research (CSMAR) Database
RM.xlsx	xlsx	The data is from the iFinD database provided by Tong Hua Shun (THS), a major financial data service company in China
Firm.xlsx	xlsx	The financial information on each RM proposal from www.cninfo.com.cn, a CSRC-authorized website that archives documents and filings for listed firms

Flow Chart for Data Analysis



KNIME Workflow for Data Analysis



Results from Data Analysis

Summary Statistics of RM and IPO

Row ID	Column	Min	Max	Mean	Std. deviation	Variance	Overall ...	Histogram
GPA	GPA	-0.007	0.638	0.162	0.141	0.02	17.024	
ROA	ROA	-0.157	0.42	0.085	0.099	0.01	9.094	
GP Growth	GP Growth	-627....	56,679....	1,260.835	7,203.23	51,886,524....	132,387.667	
assets	assets	309,7...	359,80...	18,987,...	43,036,002,985....	1,852,097,5...	2,088,583,6...	
assets gr...	assets gr...	-0.299	376.438	4.996	36.275	1,315.889	549.613	
sales	sales	12,65...	57,482,...	5,055,5...	8,239,117,930.447	67,883,064,...	530,827,55...	
sales growth	sales gro...	-0.918	87.484	2.938	12.784	163.419	323.195	
EBIT	EBIT	-476,...	10,705,...	805,421...	1,450,585,376.419	2,104,197,9...	86,180,152,...	
lev	lev	0.045	0.956	0.492	0.235	0.055	54.115	

Results from Logit Regressions

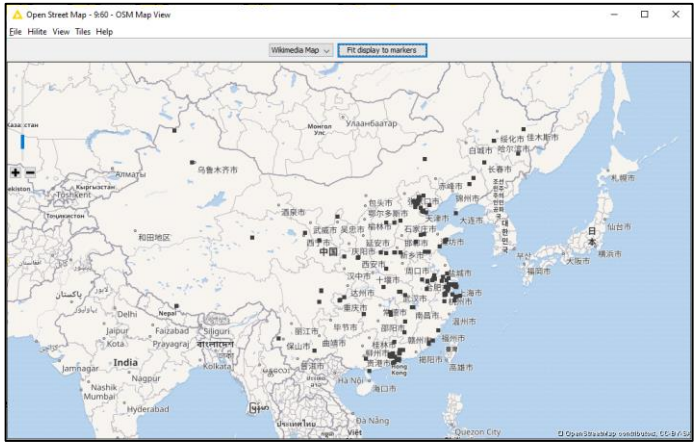
Variable	Coeff.	Std. Err.	t-value	P>|t|
GPA	-0.197	0.117	-1.687	0.092
ROA	0.624	0.19	3.287	0.001
assets	0	0	1.458	0.145
assets growth	-0	0.001	-0.078	0.938
sales	0	0	0.12	0.905
sales growth	0.012	0.002	5.151	0
EBIT	-0	0	-0.792	0.428
current	-0.029	0.085	-0.336	0.737
re_te	0.018	0.005	3.762	0
acqex	-0	0	-0.604	0.546
capex	0	0	2.998	0.003
net payout	-0	0	-1.294	0.196
Intercept	0.08	0.02	4.044	0

Variable	Coeff.	Std. Err.	t-value	P>|t|
GPA	-0.164	0.117	-1.403	0.161
ROA	0.67	0.19	3.519	0
assets	0	0	0.868	0.385
assets growth	0	0.001	0.056	0.955
sales	-0	0	-0.176	0.86
sales growth	0.012	0.002	4.927	0
EBIT	-0	0	-0.767	0.443
lev	0.129	0.052	2.477	0.013
current	0.014	0.087	0.159	0.873
re_te	0.018	0.005	3.707	0
acqex	-0	0	-0.445	0.657
capex	0	0	2.677	0.008
net payout	-0	0	-0.635	0.526
AO	0	0	0.186	0.852
Intercept	0.011	0.034	0.31	0.757

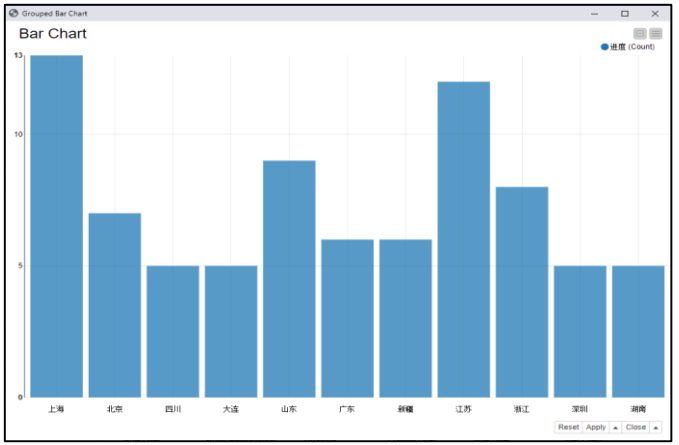
*保留三位小数

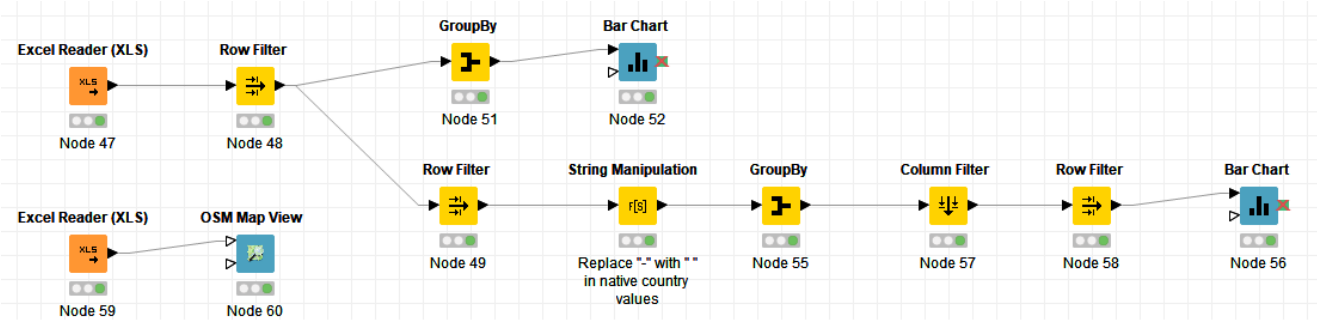
Results from the Expanded Analysis

- **Spatial Distribution of RM Firms**



- **Number of RM Firms by Province**



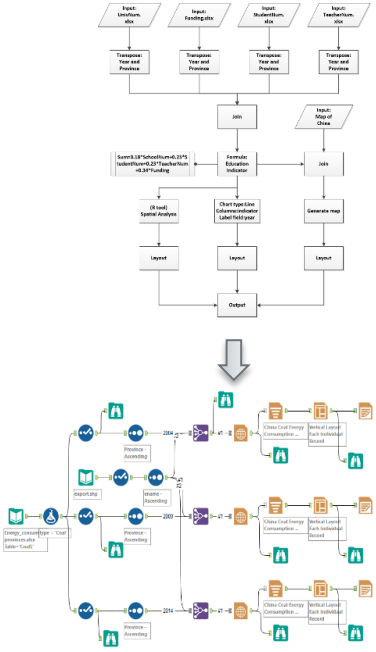


Output Files 输出文件

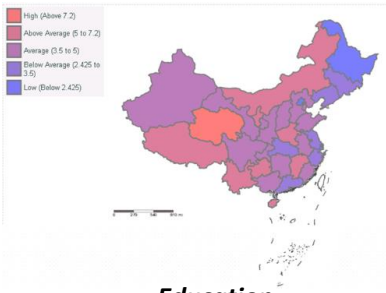
Name	Format	Description
RM_vs_IPO	.yxmd	The workflow file generated by Alteryx
RM_vs_IPO.	.knime	The workflow file generate by KNIME
workflow	. vsdx	The workflow for this study
RM_vs_IPO	.pdf	The results of summary analysis and regression analysis
Case_Study_Economics	.pptx	The description of case study

Replicable, Reproducible and Expandable Research

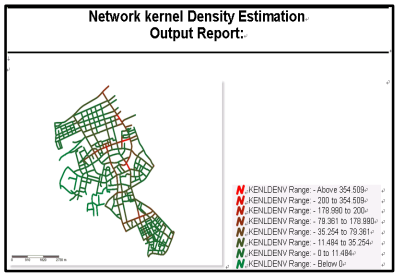
❑CDL Platform for Workflow Data Analysis



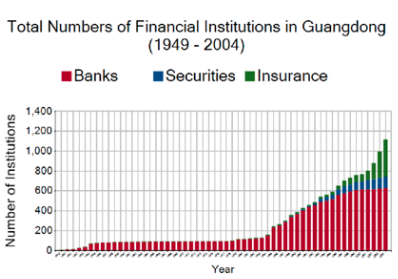
Environment



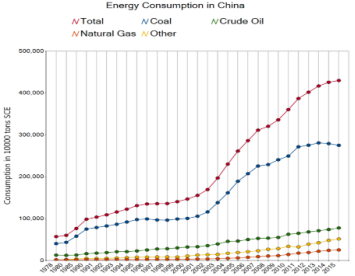
Education



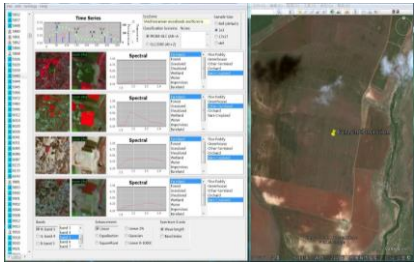
Transportation



Economics



Energy



Land Use

Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Background

Aim: Study spatiotemporal distribution patterns and its changes of China's higher education resources by visualizing the education indicator and applying spatial autocorrelation analysis.

Reference: Qin, Kun, Ping Luo. 2018. Spatio-temporal analysis of high education resources in China
秦昆,罗萍,王彩勤,卢宾宾,林曾,罗教讲. 高等教育资源的空间统计与时空格局分析. 华中师范大学学报(自然科学版), 2018,教学与研究卷.第1期: 167-172

Related work from previous studies:

- ➢Mainly focus on single indicator to analysis
- ➢Basically use mathematical statistics methods

Expansion by this study：

- ✓Multiple indicators involved
- ✓Spatial statistics analysis

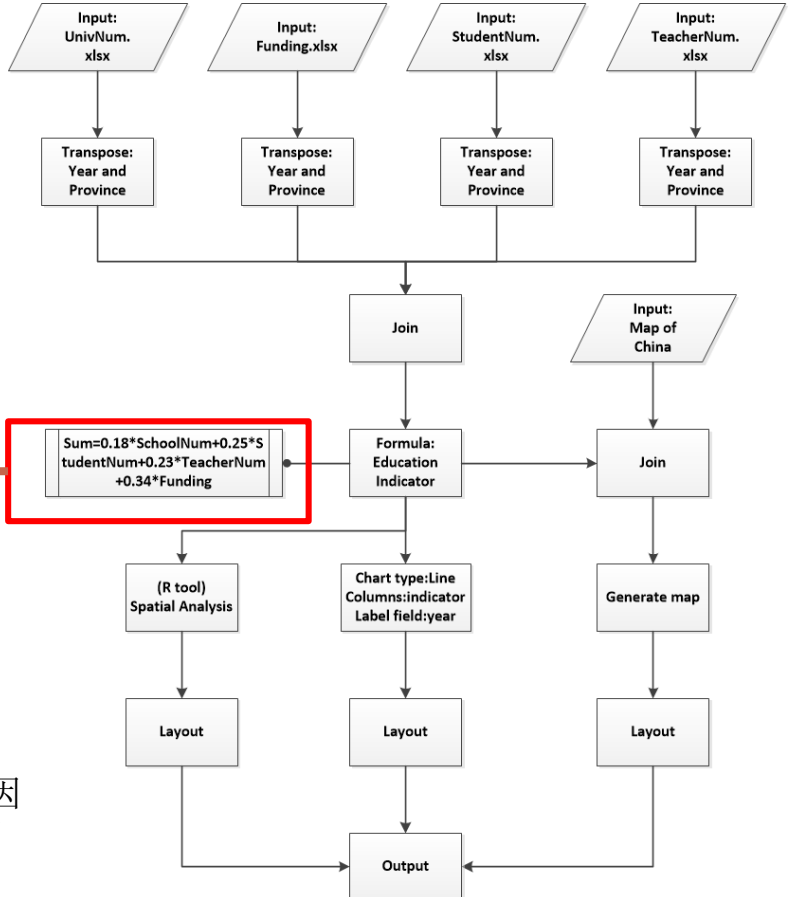
Study Area:

- ✓All universities in China.

Data Sources

Data Category	Data Description		Data Source	Format
High education resources of each province	Number of universities	2003、2007、2011、2015	China statistical yearbook	Excel
	Education funding			
	Number of students			
	Number of full-time teachers			
Map of China	With province boundary		Public data	Shape file

Flow Chart for Data Analysis

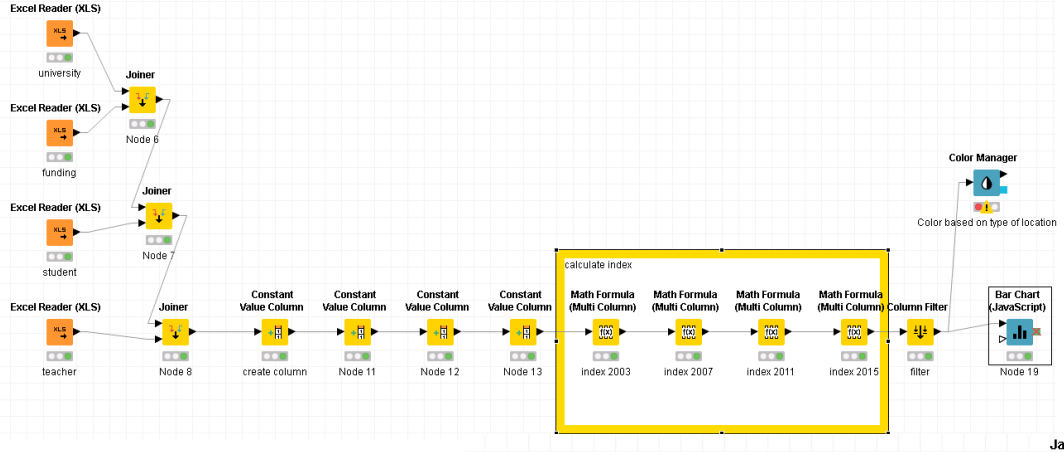


To calculate the Education Indicator：

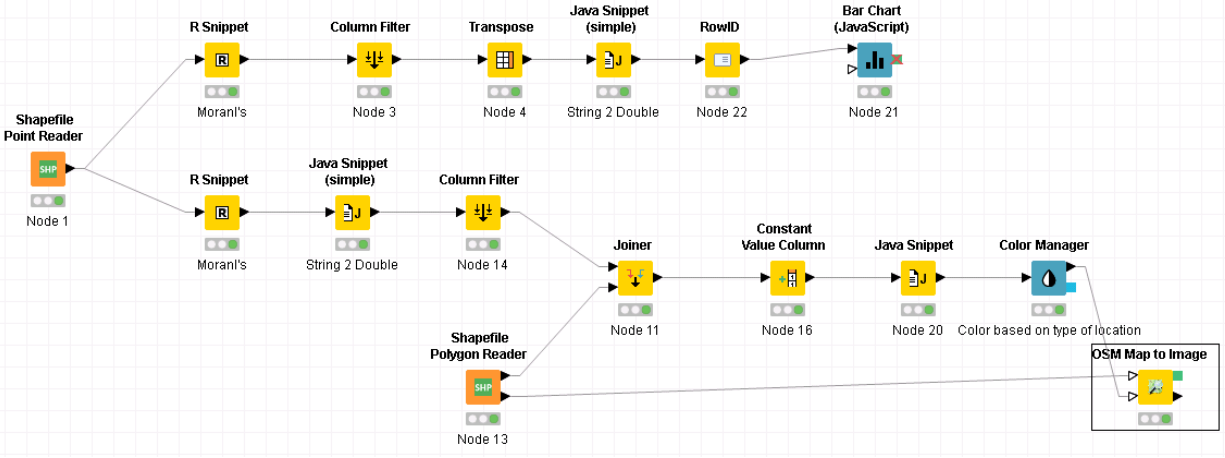
According to paper[1], we get weights of 4 single indicators of 4 years. The average value of weights in 4 years will be the final weight for each single indicator.

[1]邹志红，孙靖南，任广平．模糊评价因子的熵权法赋权及其在水质评价中的应用

Knime Workflow for Data Analysis

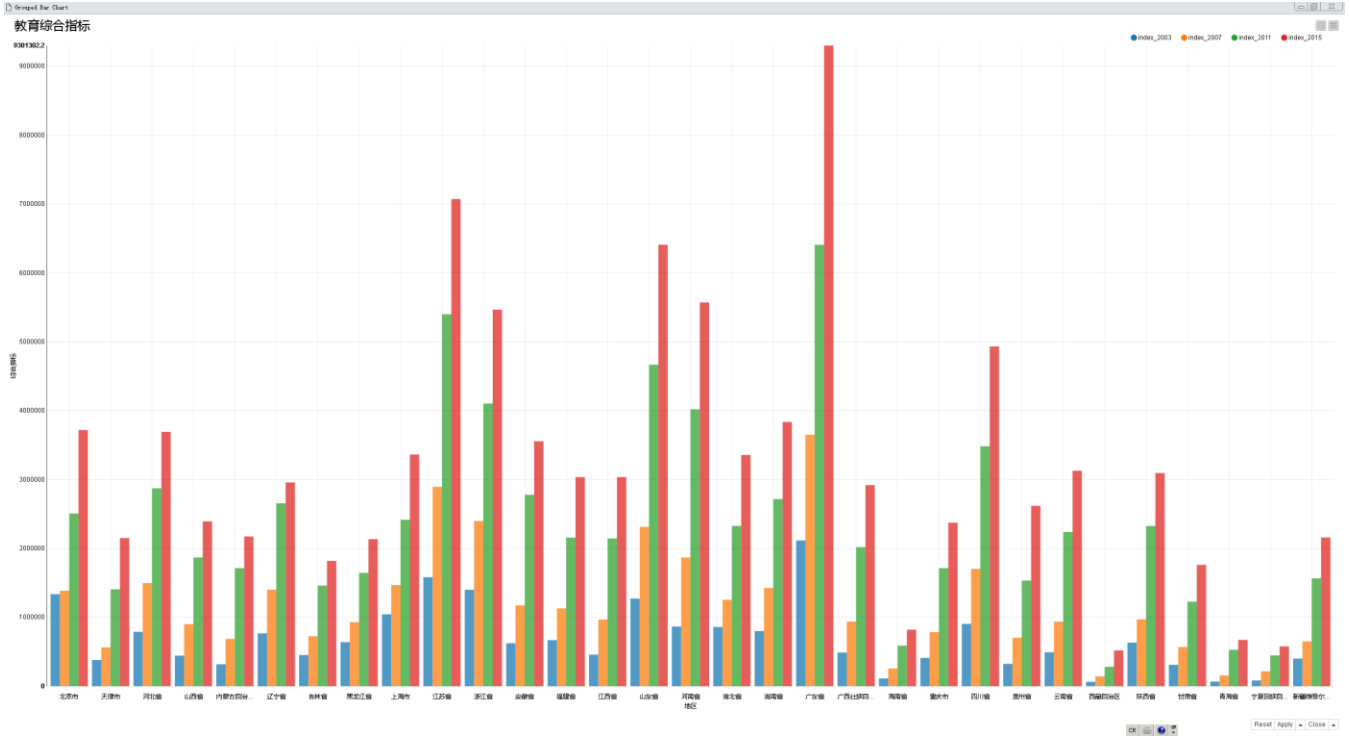


Index Calculation



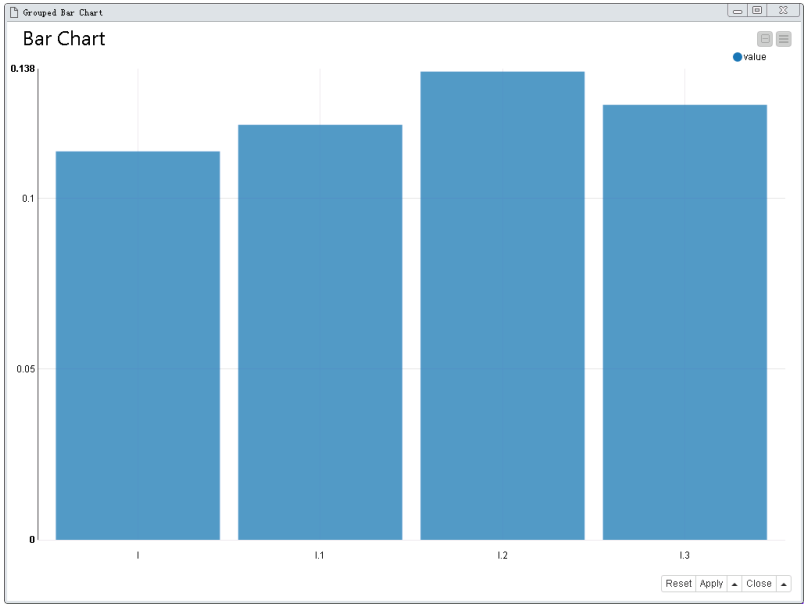
Moran I and G Index

Results from Analysis

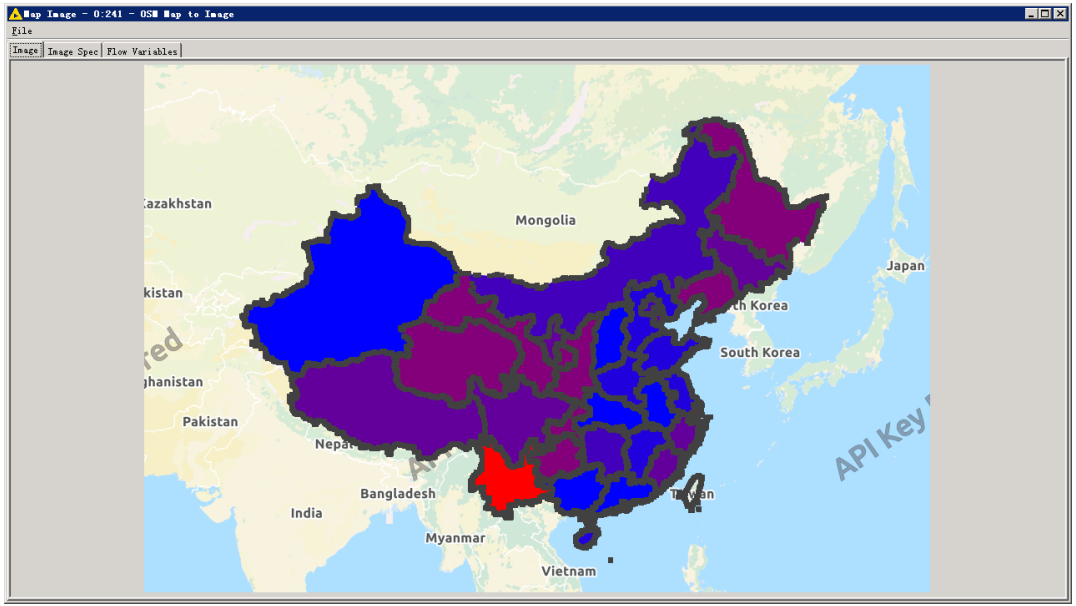


Results from Spatial Analysis

Moran's I of 4 years



Local Geary of 2015



Output Files

Documents	Description	File
Case study reference	PDF file	HighEducation_paper.pdf
Case study presentation	Presentation file	HighEducation_case.pptx
KNIME module	KNIME workflow	HighEduCase Spatial autocorrelation
Technical output	Output file generated by KNIME workflow module	output file

Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Objectives

- **Background**

  As an important form of industrial spatial distribution, it is an urgent and challenging problem to study the measurement and influencing factors of industrial co-agglomeration, especially at the level of urban cluster.

- **Aims**

  Accurate measurement of industrial co-agglomeration index, especially the direction of industrial co-agglomeration.

Definitions

- **Definition of Industrial Co-agglomeration**

  Industry locate near one another OR agglomeration across industries as a whole(Ellison et al.,2010)

- **The peculiarity of industrial co-agglomeration: Directionality Based on Ecology Theory**
- **Commensalism**: a long-term biological interaction (symbiosis) in which members of one species gain benefits while those of the other species neither benefit nor are harmed.
- **Protocooperation**: a form of mutualism, but the cooperating species do not depend on each other for survival.
- **Mutualism**: the ecological interaction between two or more species where each species has a net benefit.

Data Sources

- 2013 China industrial enterprise database
- Use the Baidu API to get enterprise geographic latitude and longitude
- Classified according to Chinese urban clusters

Name	Format	Description
Enterprise data	csv	Enterprise longitude Enterprise latitude Industry code Cluster code
Selected industry list	csv	The selected industries code

$\hat{k}$

Methdologies

Step1: Calculate the Wasserstein Distance of spatial distribution between two industries

$$W_{j,k} = \left(\inf_{\gamma \in \Gamma(f_j, f_k)} \int_{\mathbb{R}^2 \times \mathbb{R}^2} d(x,y)^2 d\gamma(x,y) \right)^{1/2}$$

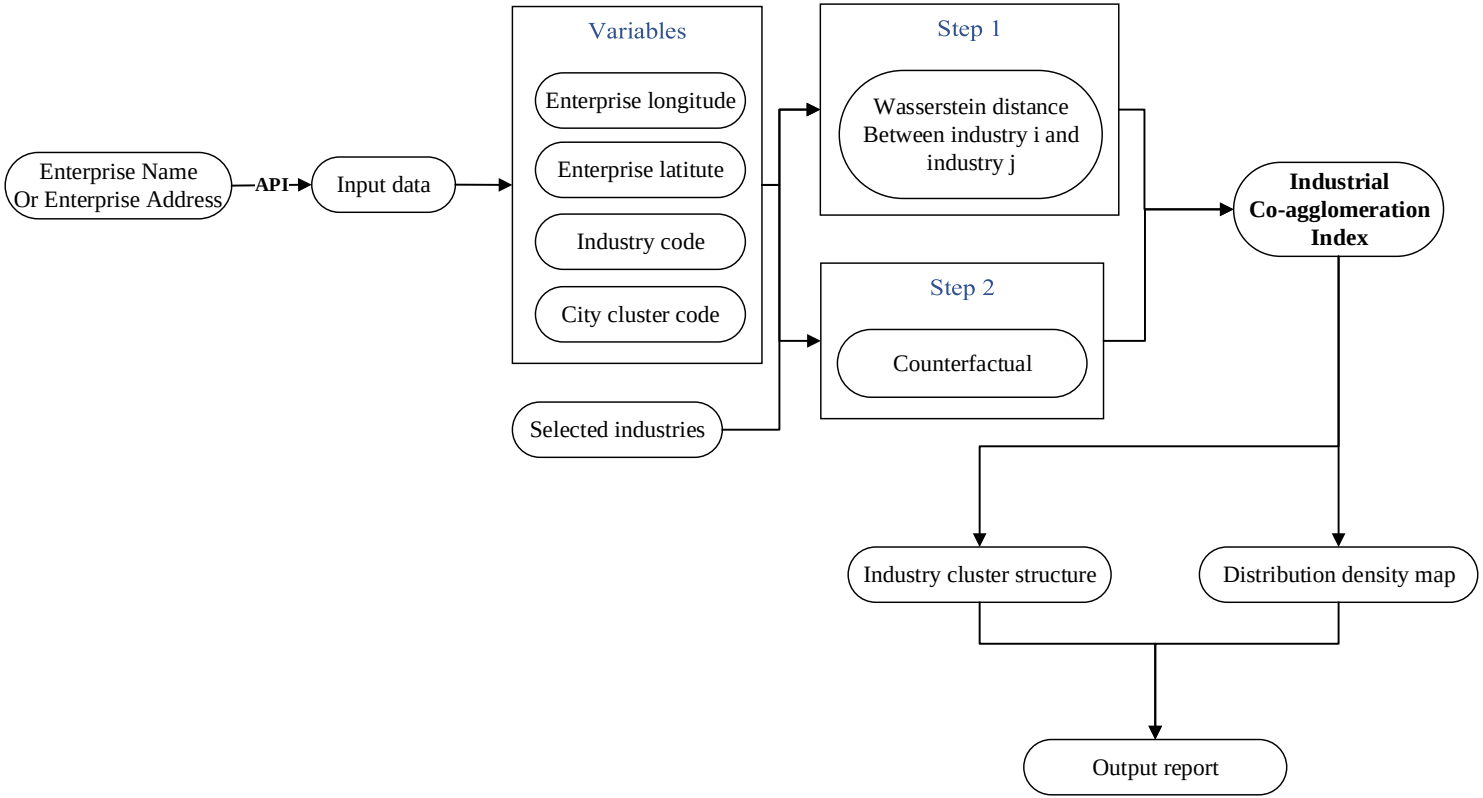
Step2: construct our counterfactual of randomly located (pseudo) industries

Null hypothesis: no spatial similarity between industry j and industry k conditional on the spatial density of industry j

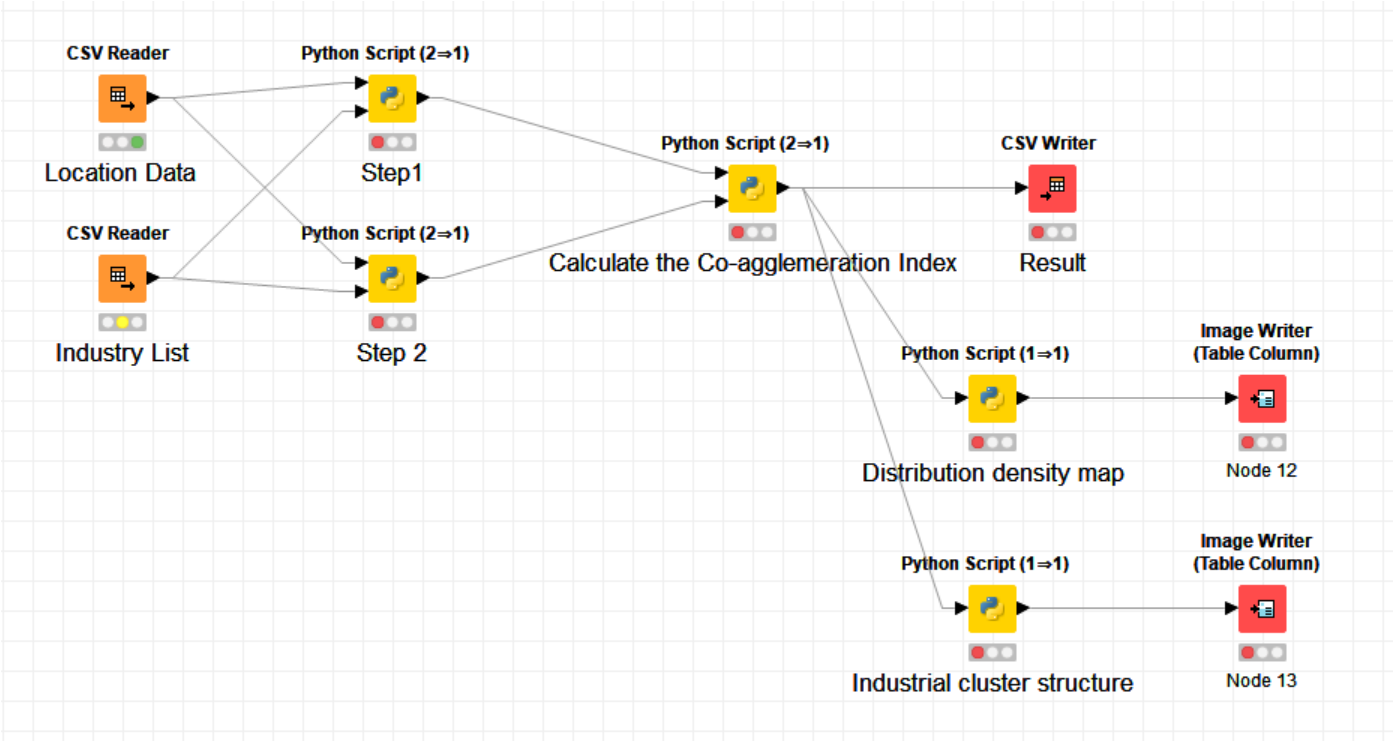
1000 pseudo industries to generate an empirical null distribution for colocalization of industry j to pseudo industry . $\hat{k}$

We construct our colocalization index by determining the number of pseudo industries for which $W_{j,k}$ is less than $W_{j,\hat{k}}$

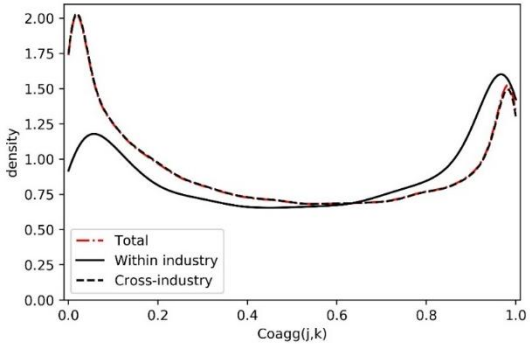
Flow Chart for Data Analysis



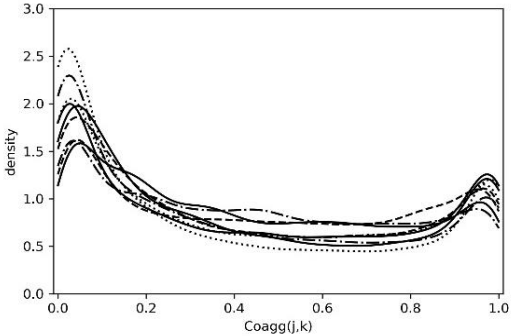
Kinme Workflow for Data Analysis



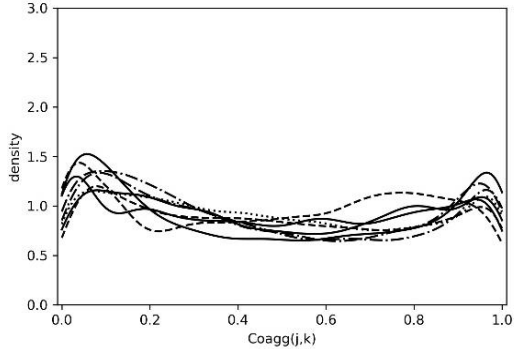
Results from Analysis



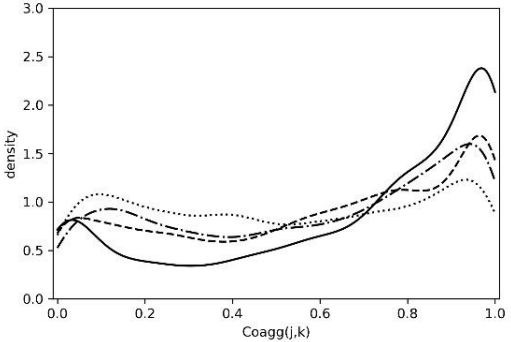
General Pattern



Pattern One



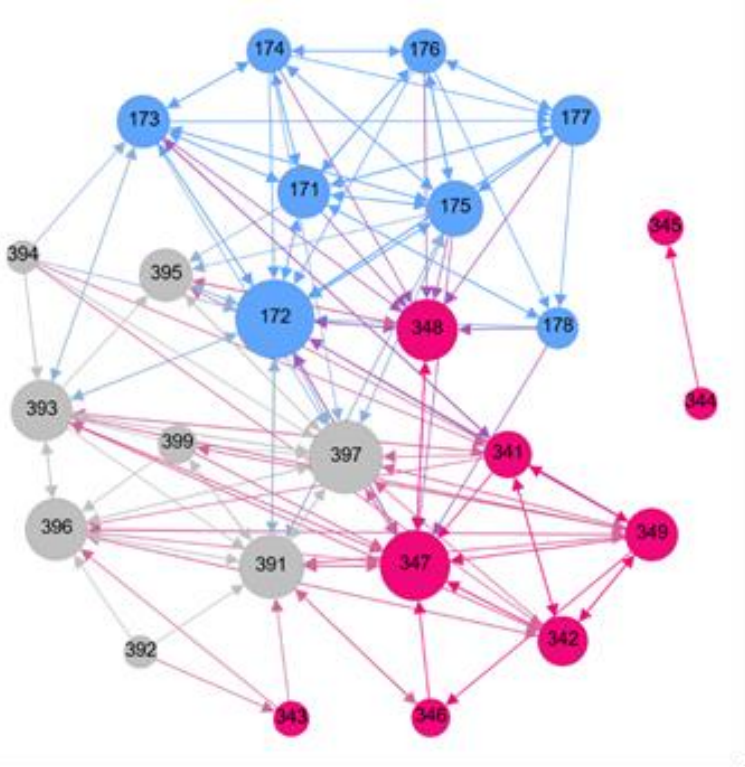
Pattern Two



Pattern Three

Cluster Analysis

- Map: Part of the Industrial cluster structure of Yangtze river delta city cluster



171 棉纺织及印染精加工
172 毛纺织及染整精加工
173 麻纺织及染整精加工
174 丝绢纺织及印染精加工
175 化纤织造及印染精加工
176 针织或钩针编织物及其制品制造
177 家用纺织制成品制造
178 产业用纺织制成品制造
341 锅炉及原动设备制造
342 金属加工机械制造
343 物料搬运设备制造
344 泵、阀门、压缩机及类似机械制造
345 轴承、齿轮和传动部件制造
346 烘炉、风机、包装等设备制造
347 文化、办公用机械制造
348 通用零部件制造
349 其他通用设备制造业
391 计算机制造
392 通信设备制造
393 广播电视设备制造
394 雷达及配套设备制造
395 非专业视听设备制造
396 智能消费设备制造
397 电子器件制造
399 其他电子设备制造

Output Files

ID	Name	Format
1.	The industrial co-agglomeration index between selected industries in 2013	csv table
2.	Map of industrial co-agglomeration index distribution density map	Chart
3.	the Industrial cluster structure of the selected industries within selected city cluster	Map

The Guideline
for
Workflow Data Analysis

Outline

- An introduction to KNIME
- Network analysis on text data
- Financial data analysis on firm level data
- Space-time analysis on educational statistics
- Machine learning based industrial co-agglomeration analysis
- Guide for workflow data analysis

Objectives for Case Studies

- ❑ Cultivate professional data analysis
- ❑ Improve the efficiency of data analysis
- ❑ Promote knowledge and information sharing
- ❑ Promote research collaborations
- ❑ Promote data applications in research and education

Data Sources

- Government Statistics （Province, City, County)
- Population Census
- Economic Census
- Survey Data
- Establishments
- Geography and Environment (Land Use, Night-Time)
- Base Maps (Census Maps, Yearly Administrative Maps)
- GRID Data
- Remote Sensed Data
- Local databases
- Special databases
- Big Data Sources (Twitter, Weibo, QQ, ….)
- Research References

Case Study Report

- Title
- Key words
- Abstract
- Background (hypothesis, debates from literature)
- Data
- Methodology
- Results of the Analysis
- Conclusions and Discussions
- Acknowledgements
- References
- Diagram of the workflow for data analysis

Tool Selection

1. ArcGIS
2. GeoDa
3. GAUSS
4. Jupyter
5. KNIME
6. R AnalyticFlow
7. Alteryx

Sample：Workflow Based Case Study

Religious Diversity and Regional Development

Get Access Export

China Economic Review

Volume 46, December 2017, Pages 1-9

ELSEVIER

Religious diversity and regional development in China

Zheng Ying [a, b], Shibao Liu [c], Shuming Bao [d, *], Jianbo Zhou [a]

Show more

https://doi.org/10.1016/j.chieco.2017.08.003 Get rights and content

Highlights

- Identify possible impacts of religious diversity on regional development at province level under the same political regime
- Apply fragmentation and polarization indices as measures of religious diversity
- Solve the endogeneity issue with the fixed-effect two-stage regression

Abstract

This paper investigates possible impacts of religious diversity on regional development in China. We developed the religious fragmentation and polarization indices as measurements for the religious diversity by using the religious site data from the 2004 economic Census data in China. The results from the panel data regressions within national coverage suggest that the religious diversity has positive and significant impacts on regional development in general. The results from the panel data regressions within regional coverage also suggest that religious fragmentation has a positive and significant association with the economic development in the eastern region of China while religious polarization has a positive and significant association with the economic development in the central and western regions of China.

Aim: Test the impacts of religious diversity on regional development in China

Findings from previous studies: Religious diversity may have positive or negative impacts on economic development

Background of this study: Test the impacts of religious diversity on regional development under the same political regime at sub-national level (province) in a peaceful environment

Data Sources

Name	Format	Description
Data_religion.xlsx	xlsx	The number of religion sites of different types at different provinces from 1987 to 2004
Data_statistics.xlsx	xlsx	variables included in the regression analysis, including real income per capita, average education attainment, investment in physical capital, policy variable, Growth rate of labor & tech, dep.
province_boundary.shp	shp	the boundaries of provinces in mainland China

Methodology

Measurements of religious diversity: fragmentation and polarization indices

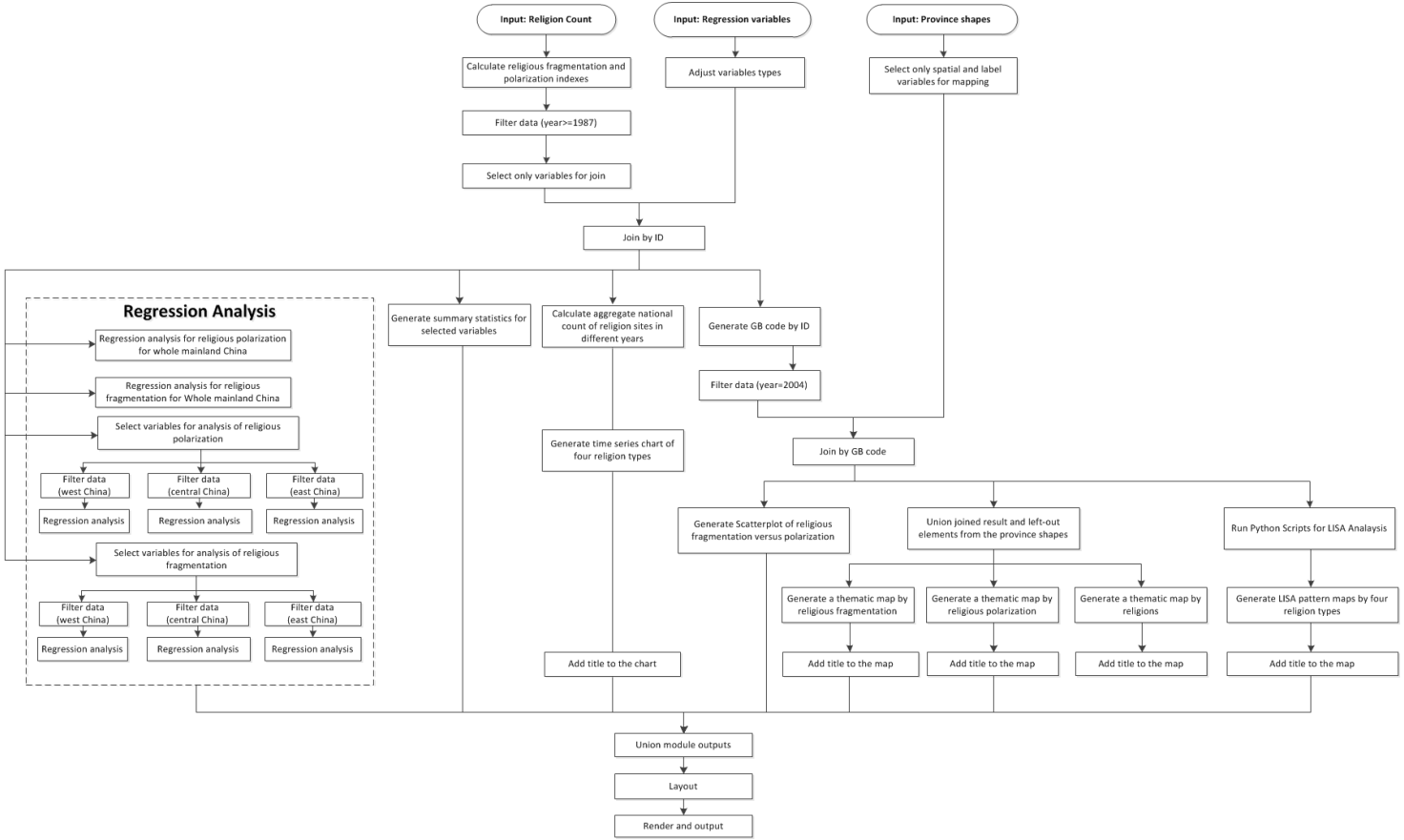
Fragmentation: $FRAG_i = 1 - \sum_{j=1}^{J} \left(\frac{n_{ij}}{N_i}\right)^2$

Polarization: $POL_i = 1 - \sum_{j=1}^{J} \left(\frac{0.5 - \pi_{ij}}{0.5}\right)^2 \pi_{ij}$

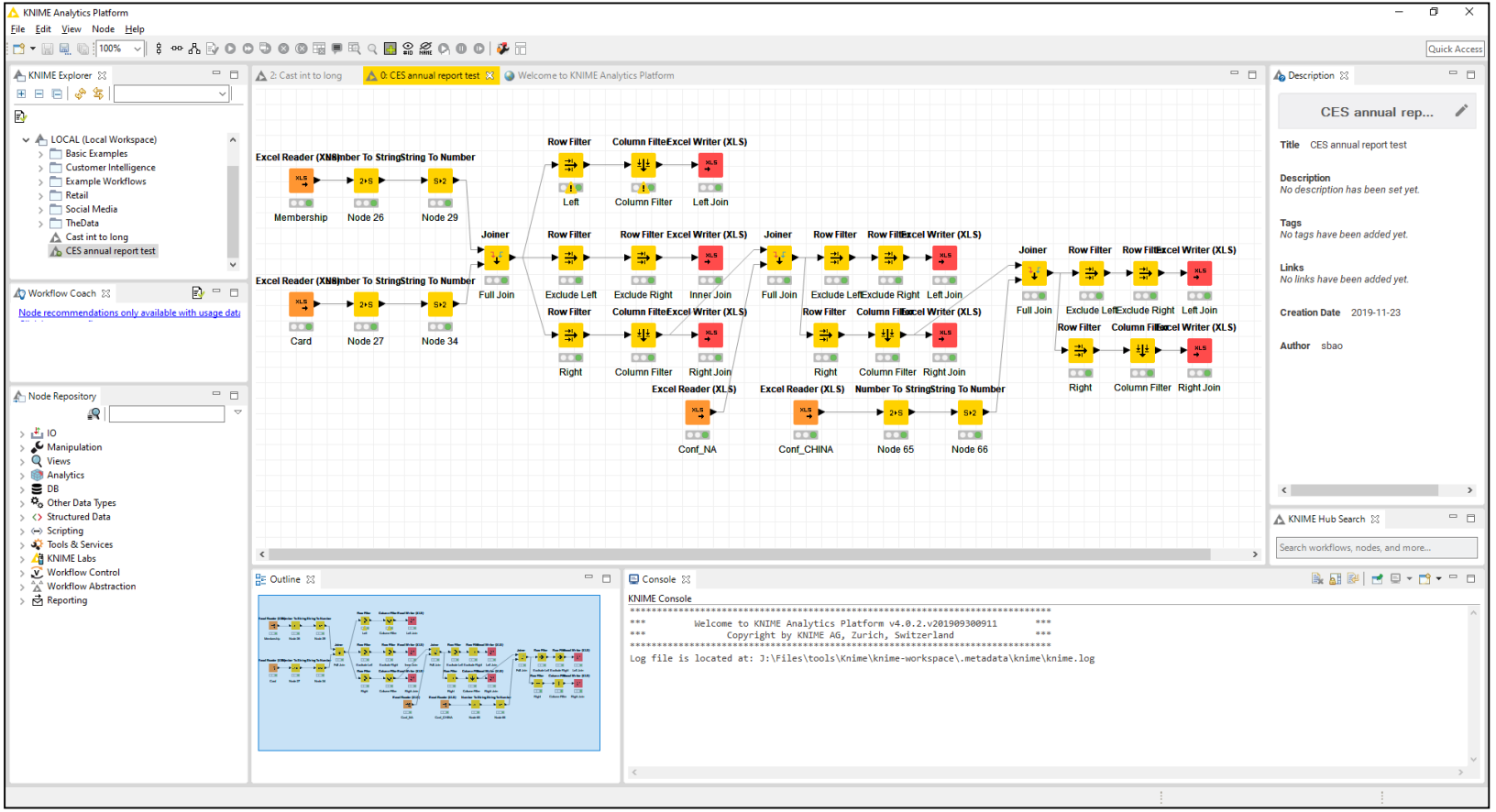
Regression: $ln\frac{Y(t)}{L(t)} = \beta_0 + \beta_1 lns_k + \beta_2 lns_h + \beta_3(n + g + \delta) + u$

Moran I test for spatial autocorrelation: $I(d) = \frac{\sum_{i}^{n}\sum_{j\neq i}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i}^{n}\sum_{j\neq i}^{n} w_{ij}}$

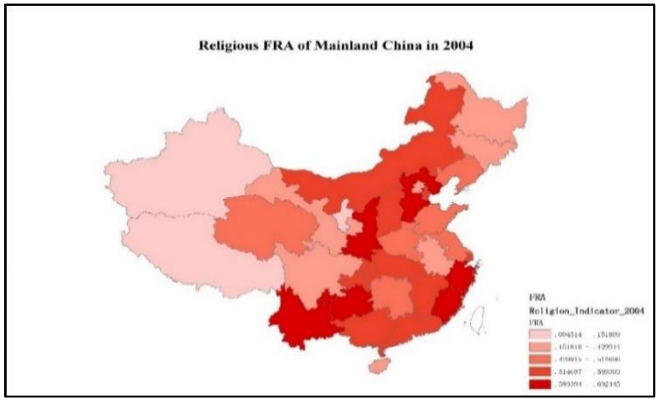
Flowchart for Data Analysis

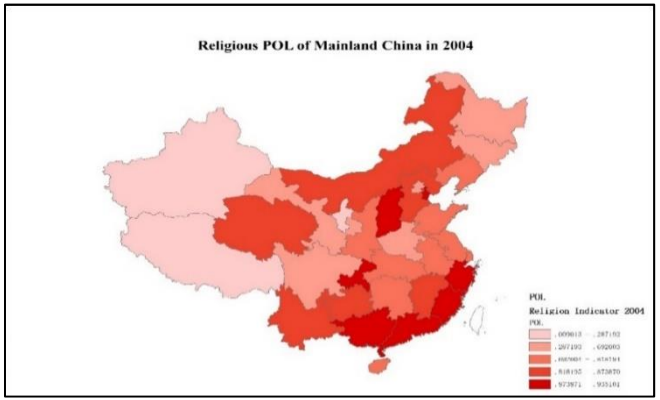


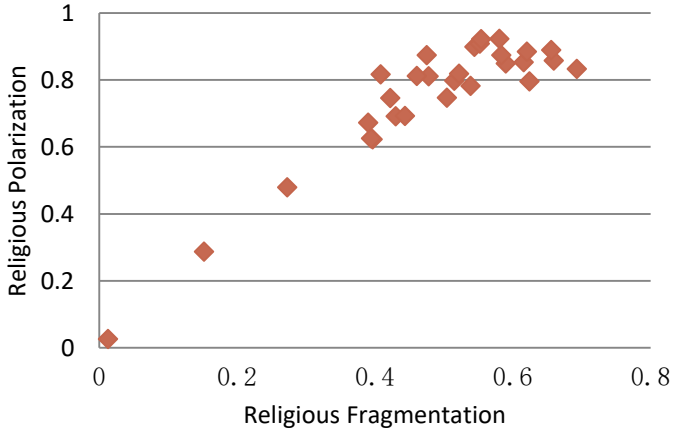
Workflow for Data Analysis



Results from Analysis







	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Whole	East	Central	West	Whole	East	Central	West
frag	0.482***	0.550***	0.0433	-0.243				
	(4.31)	(3.89)	(0.44)	(-0.88)				
(n+g+δ)	-0.00281	-0.00185	-0.0169*	-0.0484***	-0.00286	-0.00271	-0.0178*	-0.0458***
	(-0.54)	(-0.36)	(-1.71)	(-3.58)	(-0.54)	(-0.51)	(-1.81)	(-3.37)
lnedu	-0.415***	0.221	0.213	-0.565***	-0.402***	0.202	0.203	-0.580***
	(-2.91)	(0.89)	(1.08)	(-3.63)	(-2.77)	(0.77)	(1.03)	(-3.77)
lncap	0.0198	0.227***	0.0638**	0.0990**	0.0275	0.235***	0.0734**	0.0919**
	(0.67)	(5.30)	(1.99)	(2.22)	(0.91)	(5.21)	(2.27)	(2.08)
lnpub	-0.119***	-0.281***	0.0840	-0.128**	-0.131***	-0.308***	0.0880	-0.108*
	(-2.84)	(-5.49)	(1.53)	(-2.00)	(-3.08)	(-5.81)	(1.60)	(-1.69)
pol					0.141	0.165	-0.0408	0.417*
					(1.52)	(1.29)	(-0.50)	(1.69)
_cons	6.967***	6.089***	6.469***	7.690***	7.067***	6.233***	6.570***	7.333***
	(24.91)	(13.18)	(15.20)	(23.02)	(24.19)	(12.04)	(15.32)	(20.27)
N	504	180	144	180	504	180	144	180
R^2	0.972	0.987	0.994	0.982	0.971	0.986	0.994	0.982
A-R^2	0.969	0.984	0.992	0.978	0.968	0.983	0.992	0.978
F Sta.	727.2	503.6	794.4	363.0	701.5	461.6	794.8	368.3

Conclusions and Discussions

- The results from the panel data regressions within national coverage suggest that the religious diversity has positive and significant impacts on regional development in general.
- The results from the panel data regressions within regional coverage suggest that religious fragmentation has a positive and significant association with the economic development in the eastern region of China while religious polarization has a positive and significant association with the economic development in the central and western regions of China.

References

- Ying, Zheng, Shibao Liu, Shuming Bao and Jianbo Zhou, 2017. Religious Diversity and Regional Development. China. China Economic Review (Forthcoming).
- Alesina, A., & Zhuravskaya, E. (2011). Segregation and the quality of government in a cross section of countries. American Economic Review, 101 (5), 1872–1911.
- Alesina, A., Harnoss, J., & Rapoport, H. (2016). Birthplace diversity and economic prosperity. Journal of Economic Growth, 21 (2), 1-38.
- Knack, S., & Barro, R. (1998). The Determinants of Economic Growth. Determinants of economic growth: MIT.
- Kodila-Tedika, O. (2012). Determinants of Peace : A Cross-Country Analysis. (Vol.2, pp.180-200). University Library of Munich, Germany.
- Lu H, Page S E. (1998).Diversity and Optimality. Research in Economics.
- Mankiw, N. G., Weil, D. N., & Romer, D. (1992). A contribution to economic growth. , 107(2), 407-437.
- Montalvo, J.G. and M. Reynal-Querol, 2003. Religious polarization and economic development. Economics Letters, 80 (2), 201-210.
- Montalvo, J.G., Reynal-Querol, M. (2000). “The effect of ethnic and religious conflict on growth”, IVIE WP-EC 2000-04. An updated version can be found in http: / /www.wcfio.harvard.edu/programs/ prpes.
- Okten, C., & Osili, U. O. (2004). Contributions in heterogeneous communities: evidence from indonesia. Journal of Population Economics,17(4), 603-626.
- Ottaviano, Gianmarco, and Giovanni Peri. (2003）. “The Economic Value of Cultural Diversity: Evidence from US Cities.” University of California, Davis. Unpublished.

List of Documents for Case Study

Documents	Description	File
Alteryx workfows	Alteryx workflow files	*.yxmd
Flowcharts	The flowchart of workflows	*.doc
Case study report	Word or PDF file	Religion_study.docx
Case study presentation	Presentation file	Religion_case.pptx
Data	Religious data table in Excel	Data_religion.xlsx
	Statistical table in Excel	Data_statistics.xslx
	Province map in Shape file	Province_boundary.shp
Output files	Output files	Output_religion.pdf

Output Files

Description	Format
Map of Religious Fragmentation by Province in 2004	Map
Map of Religious Polarization by Province in 2004	Map
Scatterplot of Religious fragmentation versus Religious polarization in 2004	Chart
Summary statistics of regression variables	Text
Regression analysis (Residual table, coefficient table, ANOVA Analysis, Basic Diagnostic Plots)	Text

The Cloud Based Platform for Research and Teaching

❑Research ❑Tool Development ❑Data Case Studies ❑Training

Participated by

- China Data Institute (chinadatacenter.net)
- All China Marketing Research (www.acmr.com.cn)
- Center for Geographic Analysis, Harvard University (gis.Harvard.edu)
- China Consortium for Finance and Economics Education (knowledgeatshare.cn)
- Geo-computation Center for Social Sciences, Wuhan University (www.lmars.whu.edu.cn/gcss)
- Business School, East China University of Sciences and Technology (https://bs.ecust.edu.cn）
- Spatiotemporal Innovation Center (www.iucrc.org/center/spatiotemporal-thinking-computing-and-

Learning Steps for Workflows

- Step 1: Get familiar with the workflow tool
- Step 2: Practice with the sample work of case studies
- Step 3: Find a published paper and define the flowchart of data analysis
- Step 4: Develop the workflow with tool
- Step 5: Prepare the PPT report
- Step 6: Make the presentation
- Step 7: Finalize the case study

Related Web Sites

China Data Lab

http://chinadatalab.net

China Data Lab on the Cloud

http://chinadatalab.cn

Contacts:

- Dr. Tao Hu
- taohu@fas.harvard.edu
- (330) 990-8943

- Dr. Shuming Bao
- office@chinadatacenter.net
- (734) 680-3943