China Data Lab

Following the objectives and work plans in our last report, the China Data Lab project is proceeding smoothly in the second project year, including customization and enhancement of the CDL platform, integrating new datasets, developing case studies, organizing training workshops, and building global collaborations.

1. Spatial Data Lab Platform

China Data Lab released an enhanced version of the Spatial Data Lab platform deployed at Harvard University. The platform integrates sharable research data, tools and workflow-based case studies for reproducible, replicable, and expandable research, which include datasets from China Data Institute, patent data from IncoPat Inc., COVID-19 datasets collected by China Data Lab, and some other collections of socioeconomic and environmental data, as well as more than 40 workflow-based case studies on regional development, innovation studies, public health, environment, and the COVID-19 pandemic. More information is available on the following websites.

- The Cloud Platform at Harvard University: http://chinadatalab.org/

2. Integrated New Datasets

This Spatial Data Lab platform added China patent datasets from 2008 to 2018 provided by IncoPat Inc. and new COVID-19 datasets which are updated weekly on Harvard Dataverse, including human mobility datasets from Google, Apple, Safegraph, and Foursquare, and US/Global vaccination data. Additionally, the project team collaborated with Geoinformation and Big Data Research Laboratory (GIBD) at the University of South Carolina, extracting ODT (Origin-Destination-Time) flows data derived from Geotagged Tweets and Safegraph. To make them easily accessible datasets in different platforms, such as database, open repository—Harvard Dataverse, and RESTful web APIs, we developed easy-to-use workflows, with which users can comfortably and efficiently connect any datasets with straightforward configurations. More information is available in the shared document.

- Datasets introductions and user guide: https://docs.google.com/document/d/1MV0C7ro8n7bFum7q6DYISD0YkKkV1vBMlEMuC-ievmk

3. Workflow-Based Case Studies

The project team has been actively developing workflow-based case studies to help collaborating researchers easily and quickly conduct data analysis, develop methodologies, produce publications, and deliver education. More than 40 case studies have been built covering the topics of public health, innovation, environment, and economics. Each case study includes background and motivations, data source, methodologies, flowchart,

workflow implementation, results and findings, and running steps for replication. More information is available on the following websites:

- Paper Replication List: https://docs.google.com/spreadsheets/d/1uZqXzLBHUOtqMrNRi5igMVPqp83wGTM-09jKvDuz0TI/edit#gid=622313122

4. Research Grants Support

The COVID-19 data collection and platform development help external grant applications and provide direct support for the following projects:

- "Building a Spatiotemporal Platform for Rapid Response to COVID-19", funded by NSF Covid-19 rapid response grant;
- "Cities's COVID Mapping and Mitigation (C2M2)", funded by the Department of State;
- "Coronavirus Visualization Project", sponsored by Harvard CVT (Coronavirus Visualization Team) students group.

5. Research Capacity

The China Data Lab team has been collaborating with global scholars from different institutions, including scholars and students from Harvard University, George Mason University, University of Washington, University of South Carolina, Tulane University, the University of Queensland, Colorado State University, Ball State University, Georgia State University, TAMU, CUNY, Tsinghua University, Wuhan University, and many others, on COVID-19 data collection, analysis, and publications. Those teams (mobility, modeling, innovation, and visualization) have published more than 10 peer-reviewed articles on COVID-19 topics. More information is available on the following websites.

- Publications: https://projects.iq.harvard.edu/chinadatalab/publications-0
- Presentations: https://projects.iq.harvard.edu/chinadatalab/presentations-0

6. Meetings and Webinars

The CDL project is an initiative in the NSF I/UCRC Spatiotemporal Innovation Center. This Center held its 14th semi-annual Industrial Advisory Board meetings in November of 2020. The CDL team presented the project progress and received high scores and positive feedback from the Industry Advisory Board members.

Besides, the China Data Lab team jointly organized three series of webinars on different COVID-19 research topics: (1) Resources for COVID-19 Study; (2) Modeling COVID-19 Pandemic: Resources, Methodology and Applications; and (3) COVID-19 Impact Analysis. In addition, the team jointly hosts two monthly training webinars on (1) "Research Data:

Sources, Tools, and Applications", and (2) "Workflow-based Data Analysis", starting from September 2020 to August 2021. More information is available on the following websites.

- The 14th IAB meeting reports: https://drive.google.com/drive/folders/1KVwgnmnnos1NQiAw5rgqJdg6bGTkuskx
- Webinars and recorded video: https://projects.iq.harvard.edu/chinadatalab/conferences and https://dataverse.harvard.edu/dataverse/cdl_training

7. Global Impact

The Spatial Data Lab platform supported global researchers to work on COVID-19 studies and nearly 120 scholars have applied and used the platform, including the University of Oxford, University of Maryland, University of Illinois at Chicago, University of Queensland, et al. The COVID-19 research webinars also involved 75 scholars (speakers, discussants and chairs) from different institutions and attracted about 12,000 participants globally.

The COVID-19 data collections shared on Harvard Dataverse have attracted users from over 150 countries with a record of more than 400,000 data file downloads as of March 10, 2021. The datasets are also cited by researchers and institutions, such as UCGIS, Emory University Libraries, the World Bank/IMF library, George Washington University Library, Domino Data Lab, and Nanyang Technological University (NTU) library in Singapore.

8. Plans for the next phase

- Expand data collections for China study and spatial study.
- Enhance the Spatial Data Lab platform for global collaborations on China studies and spatial studies.
- Continue the development of workflows based case studies for reproducible, replicable and expandable research on spatial data science.
- Organize training workshops on spatial data science, including the on-going webinar series on "Spatiotemporal Study of Urban Dynamics" from February 2021 to May 2021, jointly with CPGIS and some other institutions.

9. Appendix

[1] China Data Lab project website: https://projects.iq.harvard.edu/chinadatalab/home
[2] China Data Lab project annual reports:
https://projects.iq.harvard.edu/chinadatalab/project-reports?admin_panel=1
[3] Spatial Data Lab Platform at Harvard University:
http://chinadatalab.org/

[4] COVID-19 data achieving on Harvard Dataverse:
https://dataverse.harvard.edu/dataverse/2019ncov
[5] CDL webinar achieving on Harvard Dataverse:
https://dataverse.harvard.edu/dataverse/cdl_training
[6] Spatial Data Lab platform datasets introductions and user guide:
https://docs.google.com/document/d/1MV0C7ro8n7bFum7q6DYISD0YkKkV1vBMlEMuC-ievmk
[7] Publications: https://projects.iq.harvard.edu/chinadatalab/publications-0
[8] Presentations: https://projects.iq.harvard.edu/chinadatalab/presentations-0
[9] The 14th IAB meeting reports:
https://drive.google.com/drive/folders/1KVwgnmnnos1NQiAw5rgqJdg6bGTkuskx
[10] China Data Lab Webinars: https://projects.iq.harvard.edu/chinadatalab/conferences