CyberGIS for the virtual co-laboratory for policy analysis

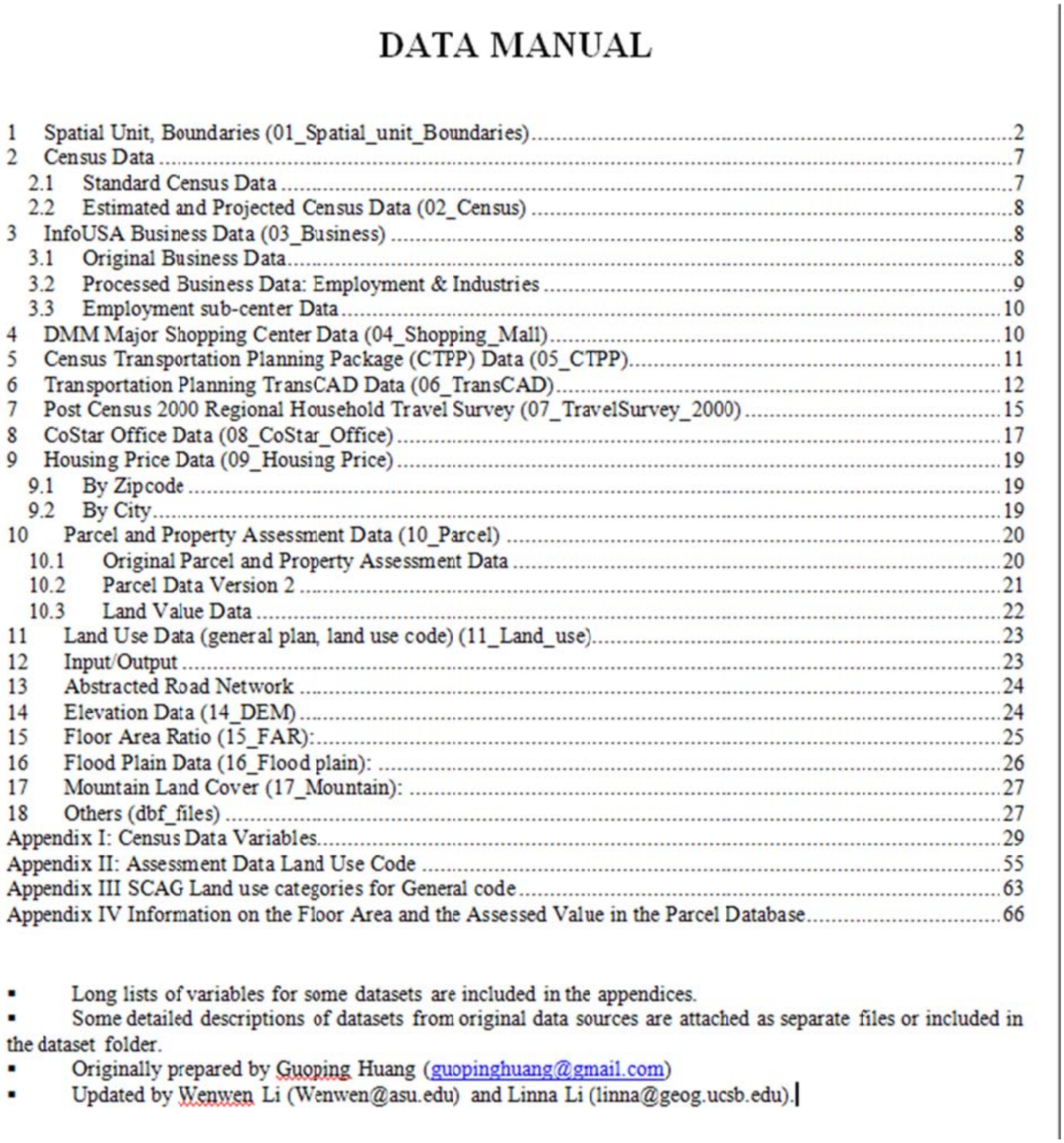
Michael F. Goodchild, Wenwen Li, and Linna Li
Department of Geography
University of California, Santa Barbara
Santa Barbara, CA 93106

Introduction

The goal of the virtual co-laboratory is the use of available cyberinfrastructure and geographic information system (GIS) technologies to support and revolutionize computational economic models for public policy decision making. This co-laboratory is designed to facilitate collaboration of researchers and decision makers with different backgrounds located in different places, including economists, planning professionals, and environmental analysts. Cyberinfrastructure is hardware and software that support data acquisition, data storage, data management, and data visualization through the Internet, especially in collaborative efforts. It integrates information and communication technologies to facilitate knowledge discovery and decision making. GIS, on the other hand, supports the collection, organization, management, and analysis of geographically referenced data, which is important for any economic analysis that contains a spatial component. The integration of the two fields is an emerging field called CyberGIS. This online co-laboratory is based on the technology of CyberGIS and includes two major components. First, relevant spatial and non-spatial data are the basis for further analysis. Second, a graphic user interface (GUI) available over the Internet allows distributed users to access, visualize, share, and download data. This report focuses on two aspects: data cleaning and data organization to ensure easy search and access of data and a GUI that provides an online tool for distributed users to view, visualize, share, and download geographic data for economic analysis.

Data support

A clean and organized data catalog is critical for data search and access. In addition, comprehensive metadata are critical to complement any data. Good data organization is important for navigating through the whole database, finding relevant data and using the data appropriately. This is especially true for a large-scale project LA-Plan that involves research coordinators and potential data users who are from different disciplines and located at different places. Originally, data for this project were collected from various sources under different standards. As the project progresses, many data have been generated: some are very useful for later analysis, such as model zones, while others are intermediate or redundant datasets. Many of them miss metadata and are not organized well. In order to improve the data quality and associated documentation, we went through every line of the data manual, reviewed every data file and attribute, sent emails to data providers regarding missing and ambiguous data information. As a result, we improved metadata for both original and processed data, including data sources, attribute name and metadata, and provenance. Data provenance refers to the processes that have been performed on the data, including the origins of data and its movement from one dataset to another. This type of information is critical for evaluating the fitness and accuracy of data in any analysis. In addition to data documentation, we cleaned all the data folders by establishing correspondence between data files and data manual, removed redundant and wrong data files and attributes that were generated in the intermediate processes. The table of contents of the updated data manual is shown in Figure 1.

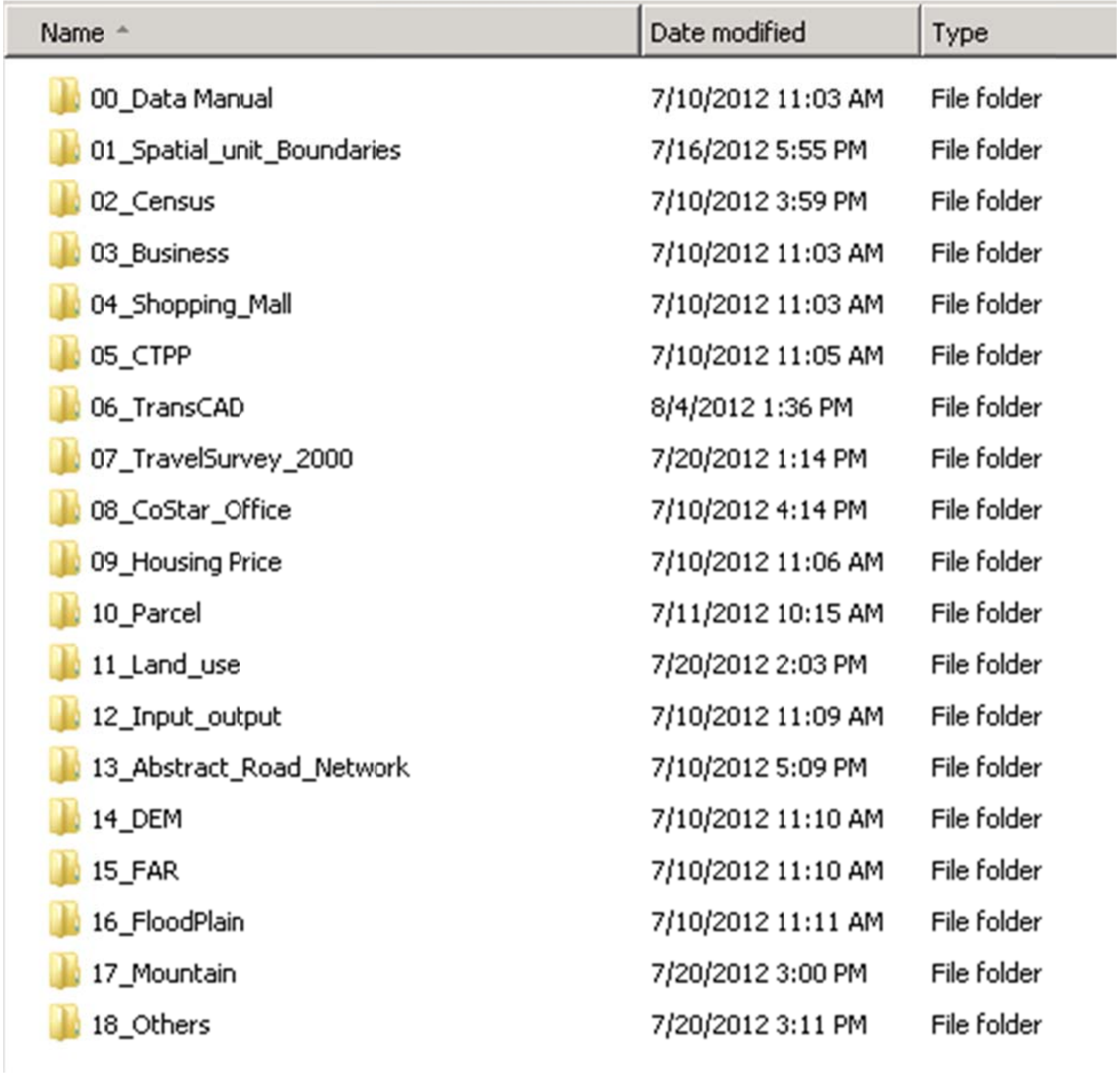


DATA MANUAL

- Long lists of variables for some datasets are included in the appendices.
- Some detailed descriptions of datasets from original data sources are attached as separate files or included in the dataset folder.
- Originally prepared by Guoping Huang (guopinghuang@gmail.com)
- Updated by Wenwen Li (Wenwen@asu.edu) and Linna Li (linna@geog.ucsb.edu).

Figure 1. The table of contents in the updated data manual

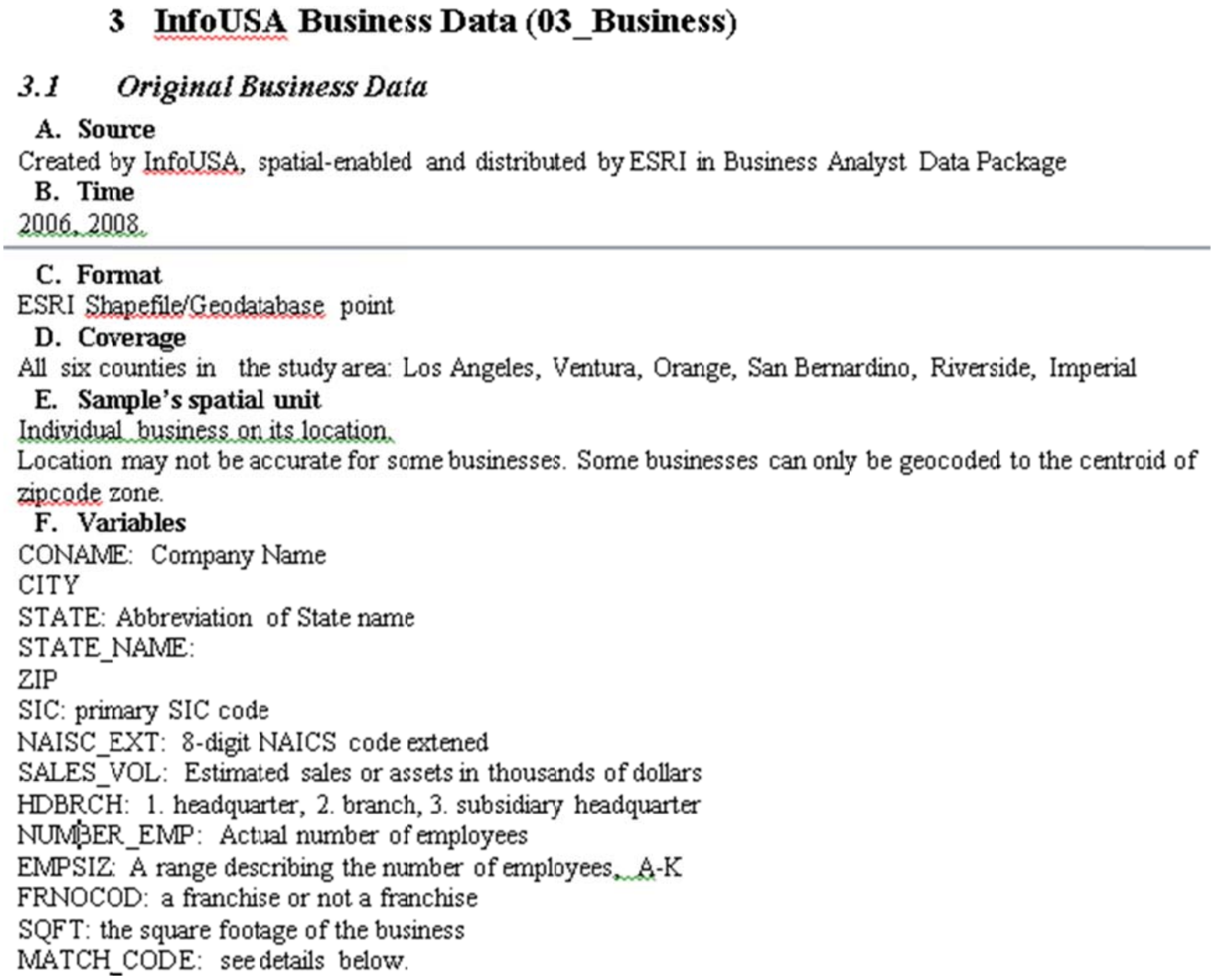
Currently, the data are organized into 18 categories: spatial unit boundaries, census, business, shopping center, CTPP, TransCAD, household travel survey, CoStar office, housing price, parcel, land use, input/output, abstract road network, elevation, floor area ratio, flood plain, mountain land cover, and others (Figure 2). Spatial unit boundaries contain different sets of data that define the spatial analysis units in the study area, including six county boundaries from US Census Bureau, Regional Statistical Areas (RSA) zones from Southern California Association of Governments (SCAG), ZIP code, tract, and block group zones from US Census Bureau, and Transportation Analysis Zones (TAZ) from US Department of Transportation. Census data contain all the attributes that were collected by US Census Bureau at the level of block group, census tract and ZIP code in 1990 and 2000, as well as the estimated and projected Census data in 2006, 2008, 2009, 2011, and 2014 provided by Esri Business Analyst Data Package. The business data were created by InfoUSA and distributed by Esri in Business Analyst Data Package. These data describe basic information of individual businesses at point locations, such as company name, estimated sales, number of employees, etc. Major shopping center data were collected from Directory of Major Malls and also distributed by Esri. Information about major malls includes the size of retail space, total retail sales, distance to nearest competing center, and the number of stores, etc. Census Transportation Planning Package (CTPP) data are derived from the long form in Census 2000, providing information on commute characteristics and socio-economic data. Transportation Planning TransCAD data provided by SCAG contain rich information that is useful in traffic simulation, including hourly capacity of links and traffic flow. The regional household travel survey provided by SCAG contains demographic information about participant households and their individual members, information about the vehicles, visited places, and trips. Office data provided by CoStar contain information about office buildings, such as building address, building name, sale price, and the number of stories, etc. Housing price data were collected from DQNews at the level of ZIP code from 2004 to 2008. They contain information about median price for single family homes, median price per square foot, number of sales, etc. Parcel and property assessment data collected from individual county's assessor's office contain information about the location and property assessment of each parcel in the study area except for Orange county. Land use data contain information about land use code in 1990, 1993, 2001, and 2005. Input/output data describe commodity flows from producers to intermediate and final consumers, which is useful for IMPLAN modeling. Abstracted road networks contain information about nodes and links in the model and their membership to different model zones. Elevation data from USGS provide digital elevation models at the scale of 10m and 30m grid. Floor area ratio data were estimated from Google Earth and Google Street View in Riverside, Orange, and Imperial counties. Flood plain data provided by Federal Emergency Management Agency (FEMA) contain information about flood zones, such as flood hazard zone lines and hydraulic structures. Mountain land cover data are extracted from the SCAG parcel data and represent the areas with mountains. The last category contains dbf files of the six counties.



Name	Date modified	Type
00_Data Manual	7/10/2012 11:03 AM	File folder
01_Spatial_unit_Boundaries	7/16/2012 5:55 PM	File folder
02_Census	7/10/2012 3:59 PM	File folder
03_Business	7/10/2012 11:03 AM	File folder
04_Shopping_Mall	7/10/2012 11:03 AM	File folder
05_CTPP	7/10/2012 11:05 AM	File folder
06_TransCAD	8/4/2012 1:36 PM	File folder
07_TravelSurvey_2000	7/20/2012 1:14 PM	File folder
08_CoStar_Office	7/10/2012 4:14 PM	File folder
09_Housing Price	7/10/2012 11:06 AM	File folder
10_Parcel	7/11/2012 10:15 AM	File folder
11_Land_use	7/20/2012 2:03 PM	File folder
12_Input_output	7/10/2012 11:09 AM	File folder
13_Abstract_Road_Network	7/10/2012 5:09 PM	File folder
14_DEM	7/10/2012 11:10 AM	File folder
15_FAR	7/10/2012 11:10 AM	File folder
16_FloodPlain	7/10/2012 11:11 AM	File folder
17_Mountain	7/20/2012 3:00 PM	File folder
18_Others	7/20/2012 3:11 PM	File folder

Figure 2. Eighteen categories of the data

Each category contains the following information: subject, source, time, format, spatial coverage, spatial unit, variables (attributes), and notes. The subject of the data is indicated in the name of the data files. Source gives the information of the data provider. Time describes when the data were generated to help users decide whether the data are appropriate for the study period. Information about format tells us how to open the file and the data are usually in the form of shapefile or table. Spatial coverage usually contains the six counties in the study area, with some missing counties in several categories of the data. Spatial unit is usually either point location or aggregated data in an area. Variables are all the relevant information about each spatial unit in the spatial coverage. Finally, notes include further explanation of the data and contact information of the data provider and data processor. One sample category of the business data is shown in Figure 3.



3 InfoUSA Business Data (03_Business)

3.1 Original Business Data

A. Source

Created by InfoUSA, spatial-enabled and distributed by ESRI in Business Analyst Data Package

B. Time

2006, 2008.

C. Format

ESRI Shapefile/Geodatabase point

D. Coverage

All six counties in the study area: Los Angeles, Ventura, Orange, San Bernardino, Riverside, Imperial

E. Sample's spatial unit

Individual business on its location.

Location may not be accurate for some businesses. Some businesses can only be geocoded to the centroid of zipcode zone.

F. Variables

CONAME: Company Name

CITY

STATE: Abbreviation of State name

STATE_NAME:

ZIP

SIC: primary SIC code

NAISC_EXT: 8-digit NAICS code extened

SALES_VOL: Estimated sales or assets in thousands of dollars

HDBRCH: 1. headquarter, 2. branch, 3. subsidiary headquarter

NUMBER_EMP: Actual number of employees

EMPSIZ: A range describing the number of employees, A-K

FRNOCOD: a franchise or not a franchise

SQFT: the square footage of the business

MATCH_CODE: see details below.

Figure 3. Business data

GUI development

The goal of the virtual (online) co-laboratory is the use of the available cyberinfrastructure and GIS technologies to support and revolutionize computational economic models for public policy decision making. This co-laboratory will facilitate collaboration of researchers and decision makers with different backgrounds located in different places. There are two major components that will be available directly to potential users, such as economists, planning professionals, and environmental analysts. First, a web mapping interface will be provided for data browsing and data download. Second, an online graphic user interface will also be provided for the execution of the Regional Economy, Land Use and Transportation Model (RELU-TRAN) in the Los Angeles area. In the online mapping interface, users are allowed to interact with the online data catalog, to select data of interest, and to upload and share their own data. It is an online GIS that supports data organization, data sharing, and data visualization (Figure 4).

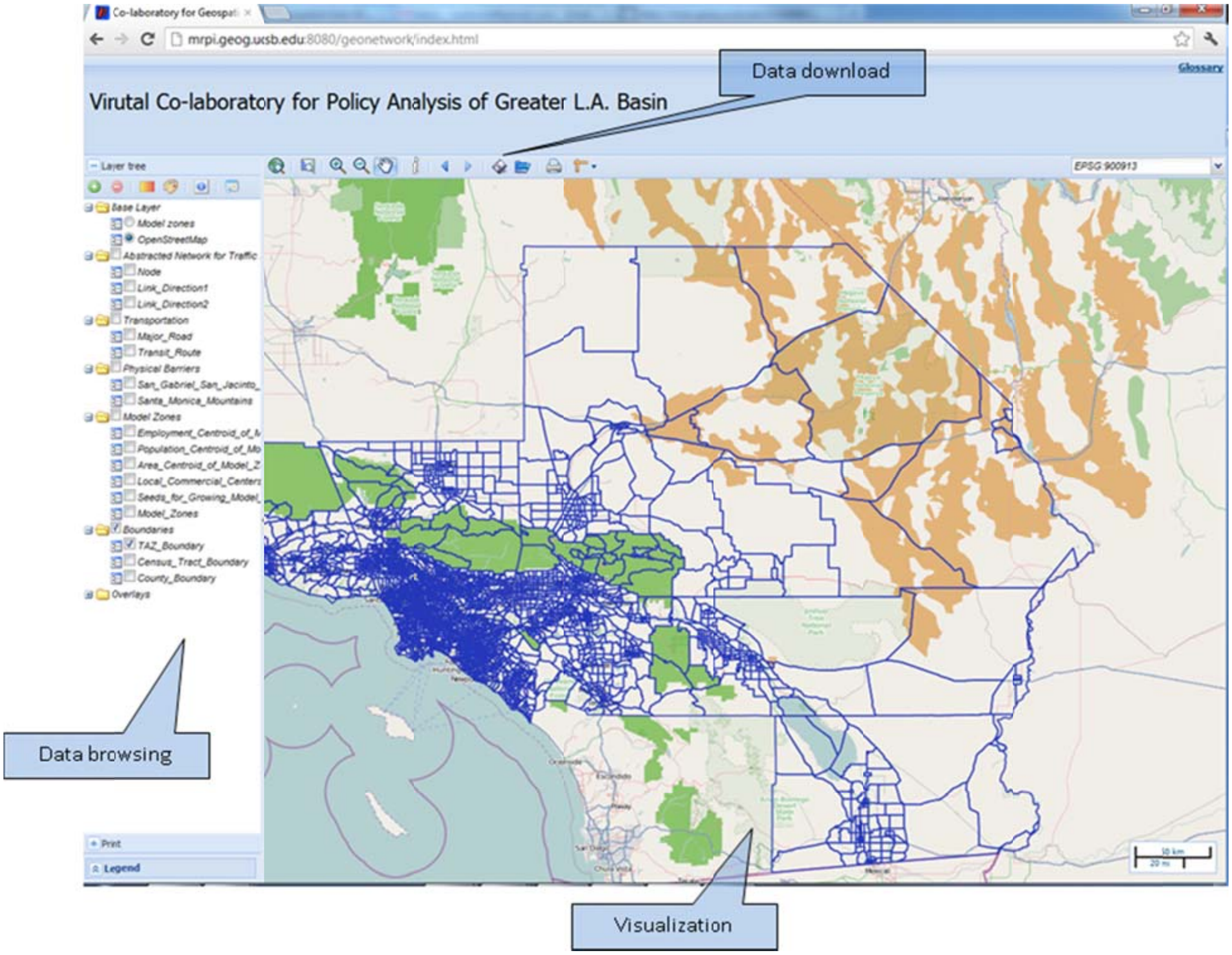
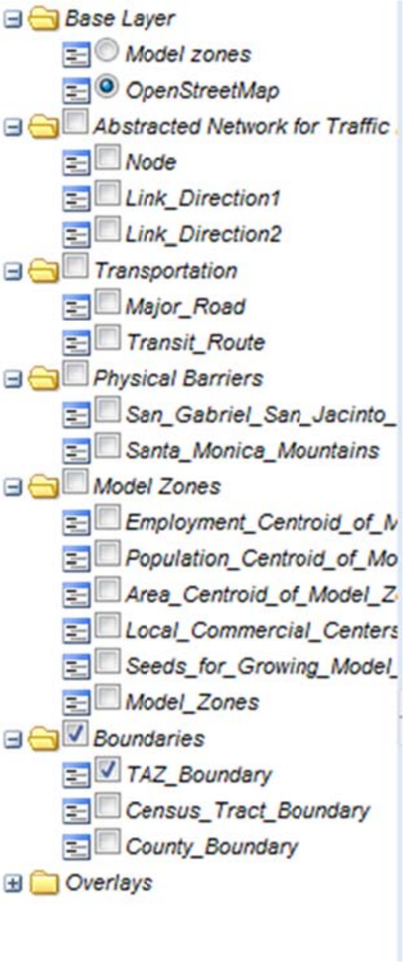


Figure 4. The graphic user interface of the virtual co-laboratory

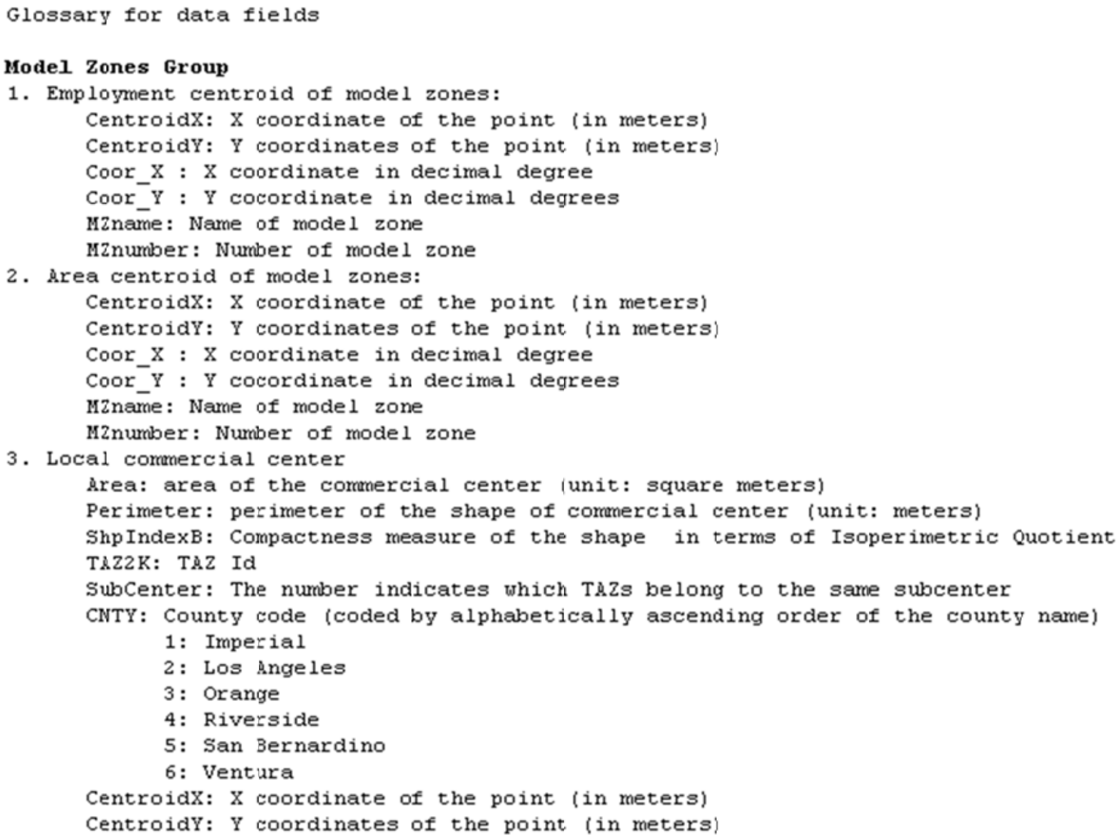
The major functionality of this interface includes data browsing, data download, visualization, and simple analysis. The base map of the mapping interface has two layers: OpenStreetMap and model zones. OpenStreetMap is a free world map that has been increasingly improved by volunteers all over the world. It contains information about basic geographic infrastructures and points of interest. In this interface, it is used as a reference for other data layers. Model zones, another base map, provide the spatial units in the study area of six counties in the Los Angeles region. Five categories of data are currently available in the data panel of the online catalog: abstracted road networks, transportation, physical barriers, model zones, and boundaries (Figure 5). In addition, users can also upload their own data as a Web Map Service (WMS) to share those with other users. This mapping interface provides Internet access to interactive maps and facilitates access to and integrated use of geospatial data.



- ⊟ Base Layer
  - ○ Model zones
  - ● OpenStreetMap
- ⊟ □ Abstracted Network for Traffic
  - □ Node
  - □ Link_Direction1
  - □ Link_Direction2
- ⊟ □ Transportation
  - □ Major_Road
  - □ Transit_Route
- ⊟ □ Physical Barriers
  - □ San_Gabriel_San_Jacinto_
  - □ Santa_Monica_Mountains
- ⊟ □ Model Zones
  - □ Employment_Centroid_of_M
  - □ Population_Centroid_of_Mo
  - □ Area_Centroid_of_Model_Z
  - □ Local_Commercial_Centers
  - □ Seeds_for_Growing_Model_
  - □ Model_Zones
- ⊟ ☑ Boundaries
  - ☑ TAZ_Boundary
  - □ Census_Tract_Boundary
  - □ County_Boundary
- ⊞ Overlays

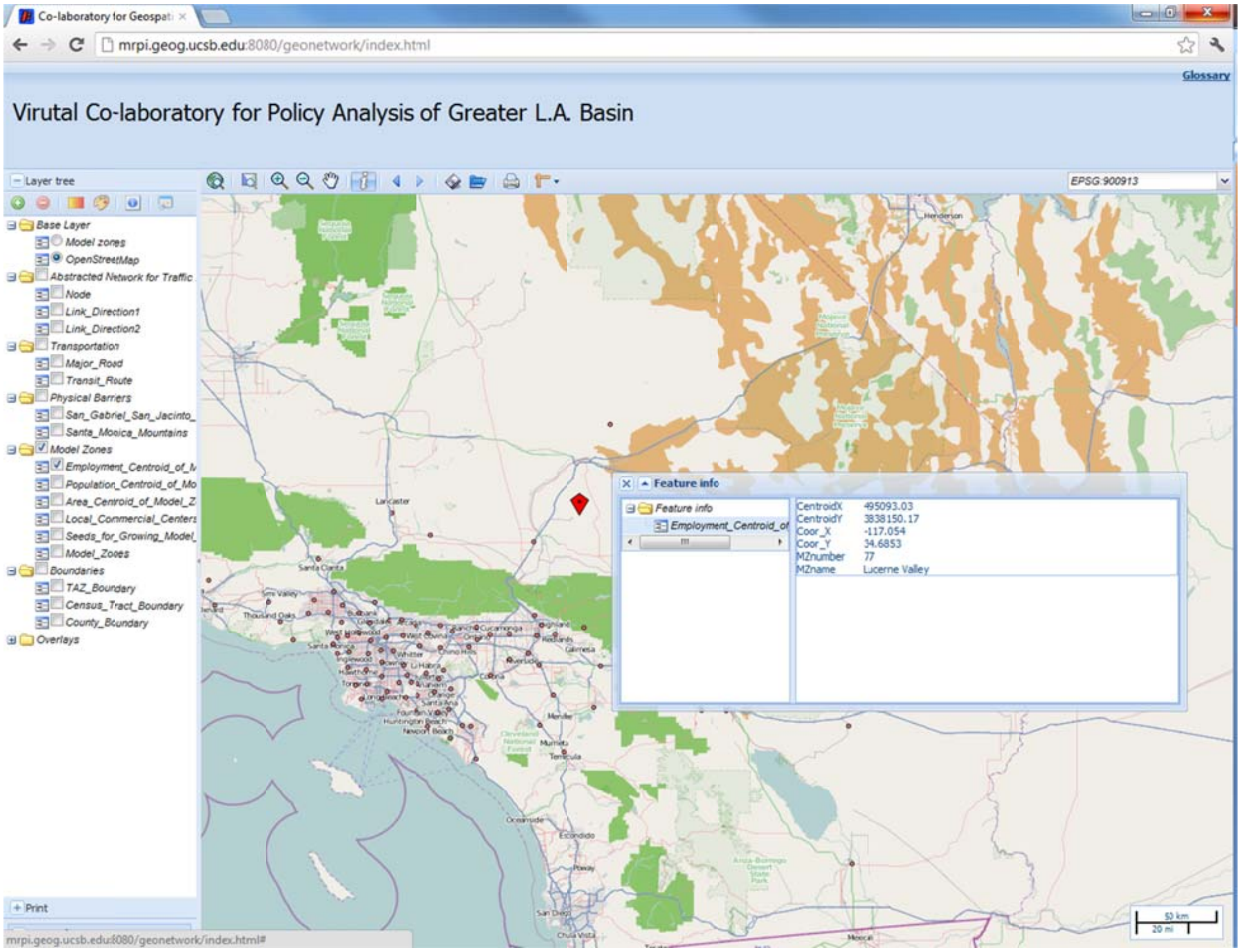
Figure 5. Data panel of the GUI

A glossary is provided to explain each attribute and associated unit (Figure 3). Each attribute name was reviewed and modified for readability when necessary. The number of decimal digits in numeric attributes reflects their accuracy. Several new functions were also added to the graphic user interface. The *i* button used for attribute query is fully supported to allow query of any visible (checked) layers. As shown in Figure 7, when a user clicks the *i* button and then clicks a feature on the interactive map, an attribute window will pop up, giving the attribute names and values associated with this feature. Data download is fully supported. Users can browse the available data on the map and download the data as tables (either in the form of CSV or XLS) if they are interested. Data download is enabled when the floppy disk icon in the tool bar on top of the map is selected (Figure 8). To protect the data copyright, authorization is required for data download. Finally, a scale bar is displayed as part of the interactive map, rather than a representative fraction.



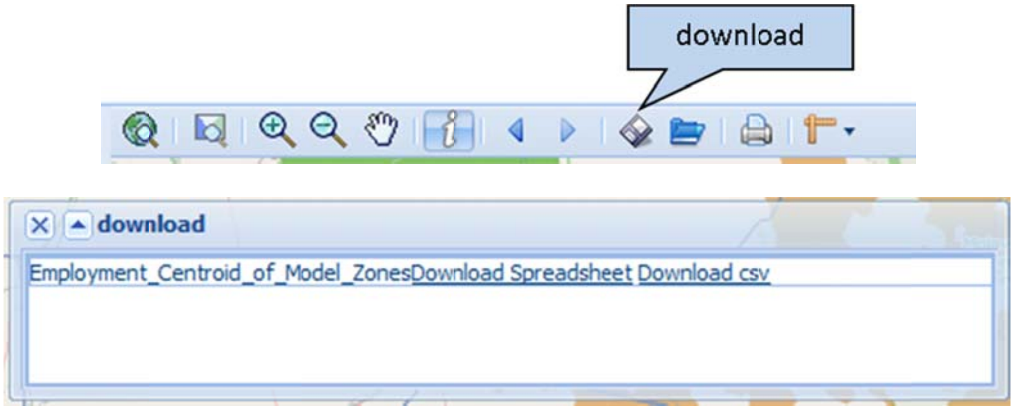
```
Glossary for data fields

Model Zones Group
1. Employment centroid of model zones:
      CentroidX: X coordinate of the point (in meters)
      CentroidY: Y coordinates of the point (in meters)
      Coor_X : X coordinate in decimal degree
      Coor_Y : Y cocordinate in decimal degrees
      MZname: Name of model zone
      MZnumber: Number of model zone
2. Area centroid of model zones:
      CentroidX: X coordinate of the point (in meters)
      CentroidY: Y coordinates of the point (in meters)
      Coor_X : X coordinate in decimal degree
      Coor_Y : Y cocordinate in decimal degrees
      MZname: Name of model zone
      MZnumber: Number of model zone
3. Local commercial center
      Area: area of the commercial center (unit: square meters)
      Perimeter: perimeter of the shape of commercial center (unit: meters)
      ShpIndexB: Compactness measure of the shape  in terms of Isoperimetric Quotient
      TAZ2K: TAZ Id
      SubCenter: The number indicates which TAZs belong to the same subcenter
      CNTY: County code (coded by alphabetically ascending order of the county name)
            1: Imperial
            2: Los Angeles
            3: Orange
            4: Riverside
            5: San Bernardino
            6: Ventura
      CentroidX: X coordinate of the point (in meters)
      CentroidY: Y coordinates of the point (in meters)
```

Figure 6. Glossary for data fields



CentroidX	495093.03
CentroidY	3838150.17
Coor_X	-117.054
Coor_Y	34.6853
MZnumber	77
MZname	Lucerne Valley

Figure 7. Feature query



Employment_Centroid_of_Model_ZonesDownload Spreadsheet Download csv

Figure 8. Data download

This graphic user interface is still in progress. The next step is to complete the development of the GUI. There are two major functions we plan to add. The first one is the query and visualization of interaction data. When a user selects two locations on the map, the interaction between the two locations will be displayed, such as the traffic volume between two model zones. The second one is the implementation of choropleth mapping, using Tobler's interval-free approach. The color of each model zone is based on the selected attribute value of each one. When the GUI is completed, we will do another round of systematic usability test. Here are some example questions we expect to answer.

Does it work in different browsers on different platforms?
Is it easy to navigate through different parts of the GUI?
Are users able to locate key features and functionality in the GUI?
Does every function work properly?
Where do users get lost?
Are there any missing data that are critical for economic modeling?
Are there any important functions that are not available?