Hall, Charles
SSCI 593
Final Project Data Report

1) What is the theme and contents of the data?

   a. The theme of my data set is walkable areas and points of interest in downtown Los Angeles. I have decided that the boundaries will be a half mile buffer around Metro Rail stations (Gold Line: China Town, Union Station, Little Tokyo. Red / Purple Line: Union Station, Civic Center, Pershing Square, 7th Metro Station. Expo / Blue Line: 7th Metro, Pico Station.) This is, in my opinion, a decent approximation of the downtown area served by Metro Rail.

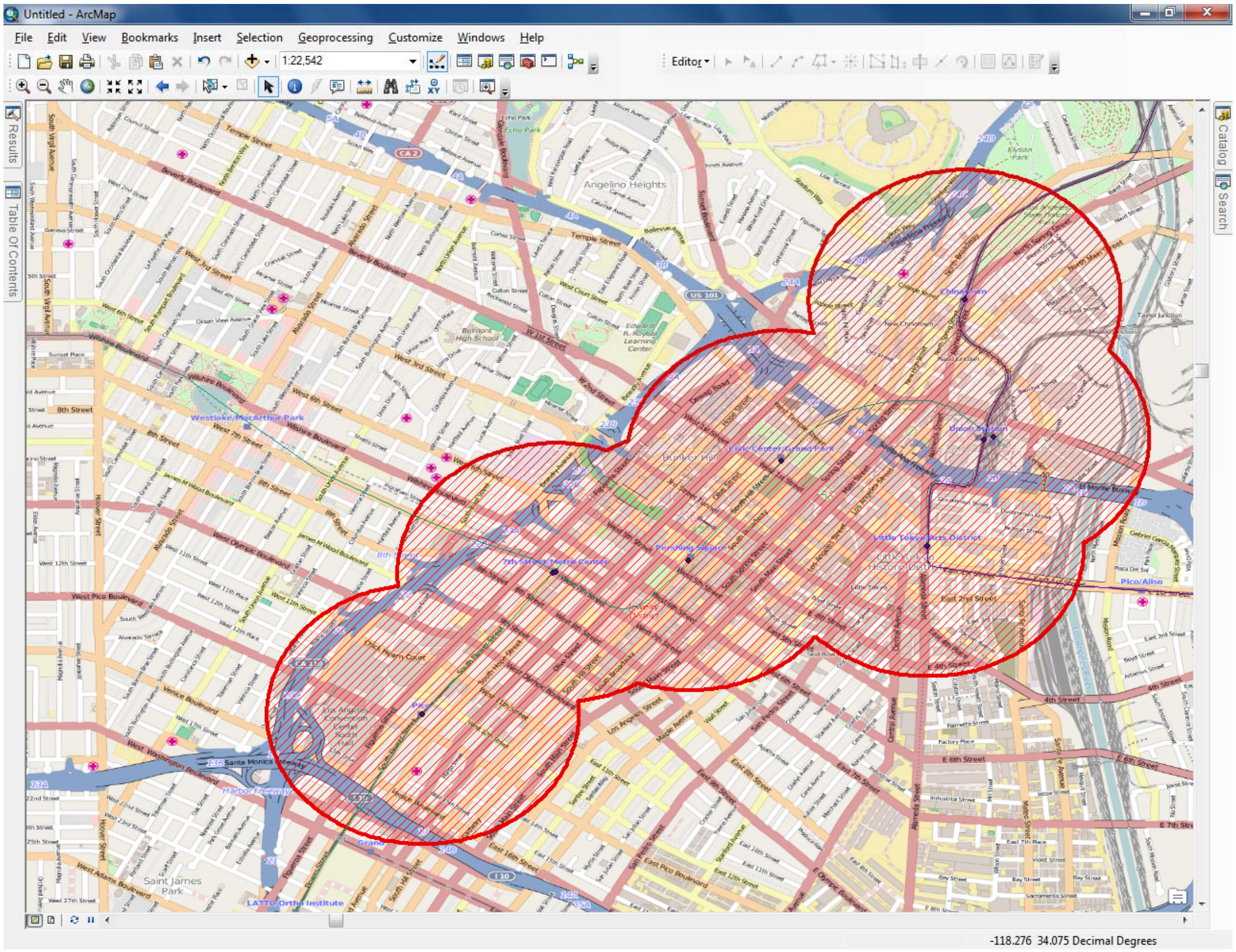


Figure 1 - Proposed Study Area

2) Does the data cover my study geographic area?

   a. In order to answer this question I first had to define my geographic area. I decided to use a half mile buffer around selected Metro Rail stations to define the downtown area. It is an arbitrary definition, but it is good enough for this project. An interesting way to define the downtown area would be to use a kernel density plot of geotagged posts such as tweets or photos that include labels like downtown LA, DTLA, and downtown Los Angeles. Time permitting I will try to incorporate a thematic layer that uses this idea.

b. I selected zip codes that are within my buffers as search criteria in order to limit the number of businesses returned by the business database search. More than one hundred thousand businesses are otherwise returned for the city of Los Angeles.

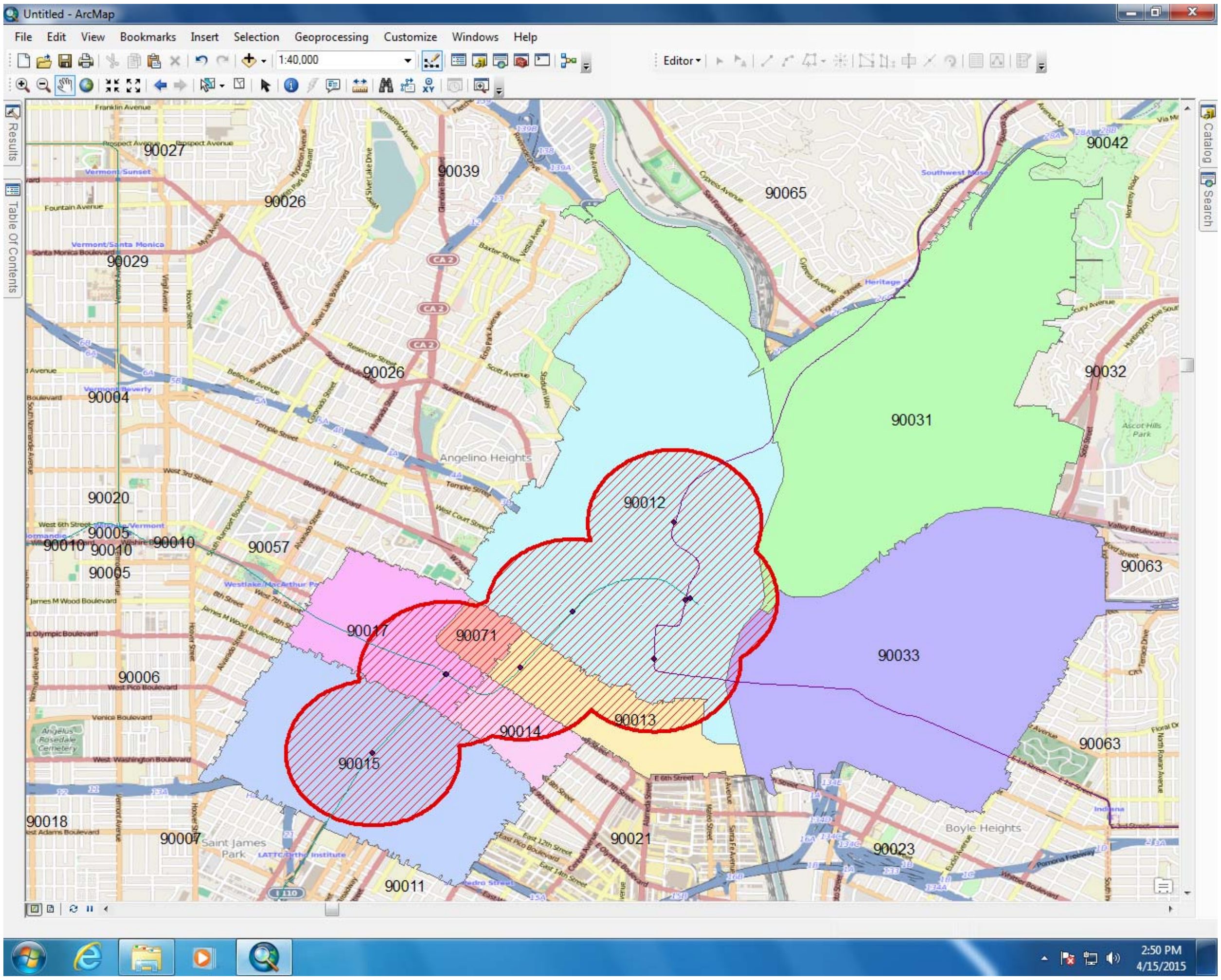


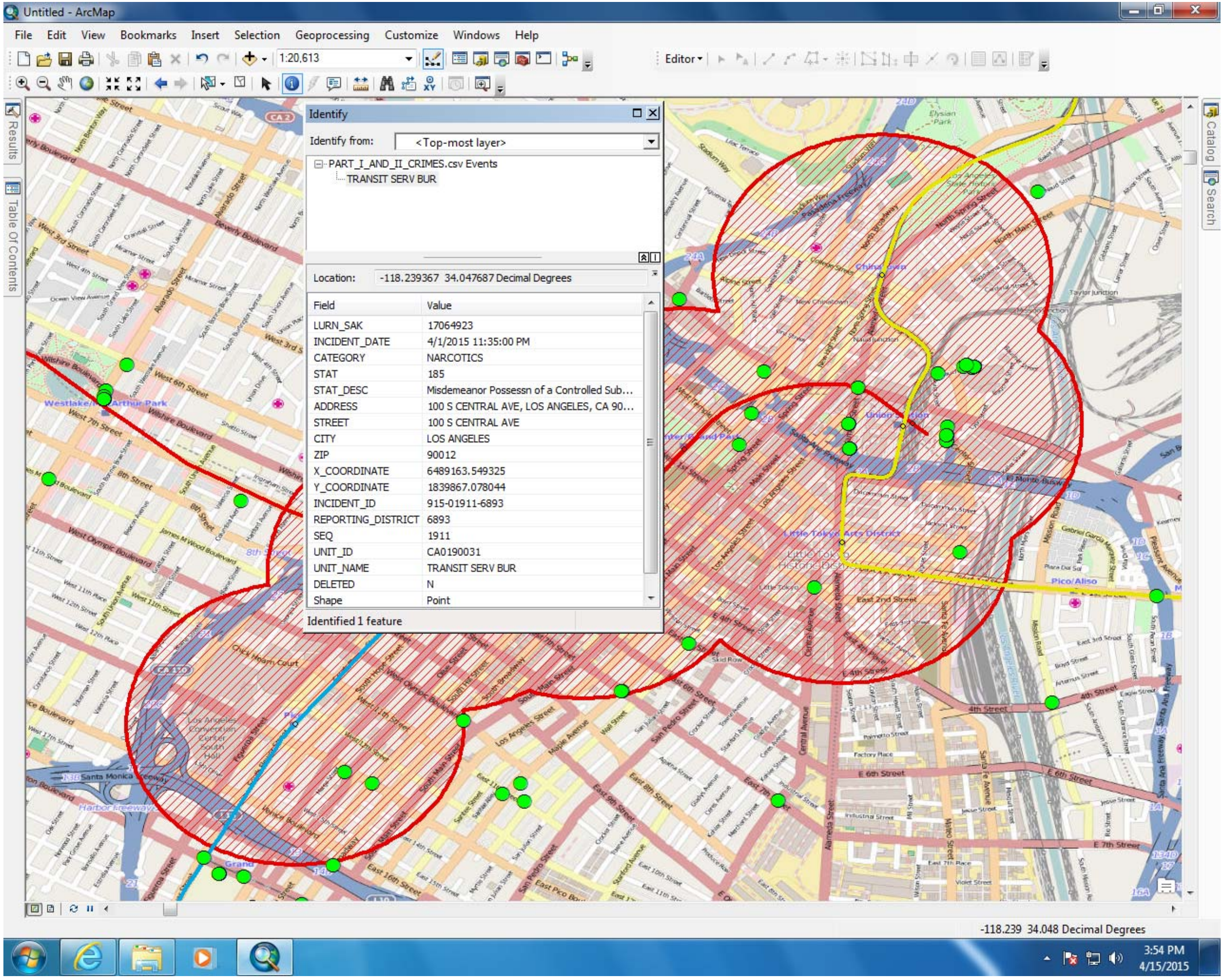
Figure 2 – Zip codes

3) Are the data authoritative data? Were the data developed by a local, state or federal institute? Or, were the data developed through a VGI or crowdsourcing project?
   a. Much of the data that I am using is authoritative data provided by the LA Metro, County of Los Angeles, and referenceUSA. I will also be using VGI from OpenStreetMap as my base map.
   b. If I am able to pull data from API's for services like Flickr or Instagram, those data would be VGI and subject to all of the potential problems presented by VGI that we have discussed to date.
4) Are the data free of charge?
   a. All of the data that I am using for this project is free of charge; however, I must note that some of the data is free of charge because of my affiliation with USC. Data provided

by LA Metro, LA County, LAPD, and via API from sites such as Flickr or Instagram are free of charge; however, they each have their own licensing use requirements.

5) Are the data downloadable? Are the downloadable data map tiles (i.e., only georeferenced images) or features (i.e., the data usable for most spatial analysis)?
   a. That data sets that I am using are mostly downloadable shapefiles and databases. LAPD crime data that I have found available for download is in a text file with no column headings, which makes it rather difficult to work with. The LA County Sheriff provides data in a CSV that includes approximate address as well as XY coordinates in state plane 5 format.
   b. The referenceUSA business data can be downloaded as a CSV in twenty five business blocks. I have not yet found a way to download larger datasets from this service. In addition, the data includes street address, but no spatial data. Consequently, the data would have to be geocoded before it could be used.
6) Are the metadata provided?
   a. Metadata are provided for some of the data sets, but in various ways.
   b. LA Metro does not provide any specific metadata, though they do provide a website with information about when data is updated, update frequency, and who to contact if assistance is needed.
   c. The LA County GIS Portal aggregates data from numerous agencies and there is some variation in the quality of metadata provided. That said, the parcel and zipcode files that I am currently using do contain basic metadata including a point of contact. However, lineage and consistency are not addressed in the metadata.
   d. The LASD crime data does include a disclaimer that locations have been obscured slightly "in order to de-identify specific locations."[1]
7) Is additional data processing required for use in a GIS environment?
   a. Yes, absolutely. Some of the data sets are only used to provide a constraint while querying another data set. Zip codes, for example, help to refine the business search, but provide no additional use in the application. Nevertheless, work must be done to join the zip code data with the buffers in order to determine which zip codes should be used in the business selection criteria.
   b. The rail stations for each line are provided in individual shapefiles. The desired stations must first be selected from each shapefile. The resulting selections must then be merged in order to simplify the process of providing one large buffer that represents the study area. I should note that some attribute names had to be edited or shortened in order for the merge operation to complete successfully.
   c. The LASD data is in a CSV file and must be imported into ArcMap. The coordinate system must be changed once the data is imported.
   d. The LAPD data must be turned into a delimited format and geocoded before it can be used.
   e. The LASD data does not contain object ID's and consequently can't be selected. This creates a very large data set that goes well beyond my research area.

[1] http://egis3.lacounty.gov/dataportal/2012/03/05/crime-data-la-county-sheriff/


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:20,613
Editor
Identify
Table Of Contents
Results
Catalog
Search
-118.239 34.048 Decimal Degrees
3:54 PM
4/15/2015

Identify from: <Top-most layer>

PART_I_AND_II_CRIMES.csv Events
TRANSIT SERV BUR

Location: -118.239367 34.047687 Decimal Degrees

Field	Value
LURN_SAK	17064923
INCIDENT_DATE	4/1/2015 11:35:00 PM
CATEGORY	NARCOTICS
STAT	185
STAT_DESC	Misdemeanor Possessn of a Controlled Sub...
ADDRESS	100 S CENTRAL AVE, LOS ANGELES, CA 90...
STREET	100 S CENTRAL AVE
CITY	LOS ANGELES
ZIP	90012
X_COORDINATE	6489163.549325
Y_COORDINATE	1839867.078044
INCIDENT_ID	915-01911-6893
REPORTING_DISTRICT	6893
SEQ	1911
UNIT_ID	CA0190031
UNIT_NAME	TRANSIT SERV BUR
DELETED	N
Shape	Point

Identified 1 feature

Figure 3 – LASD Data

8) Are the data in a public domain?
   a. None of the data that I have reviewed to date has an explicit public domain tag, a Creative Commons license for example. The data are, however, on public websites and freely available to download, which makes the public domain status slightly murky.
   b. The data from LA Metro and LA County are publicly available and require no credential for access. ReferenceUSA must be accessed via the USC library and has tighter restrictions.
   c. Crime data from both the LAPD and LASD are available for download in CSV format.
9) What kinds of disclaimers are informed?
   a. Each of the data sets includes terms and conditions for use. Some, including those for the LA Metro, are quite extensive and require careful review.
10) What licensing options are available? Are there guidelines for redistribution of the raw data and distribution of your derivative data?
   a. The LA County data explicitly states that there are no usage restrictions.
   b. The LA Metro data states that the data may be displayed in conjunction with other data. The license goes on to state that the data should not be used to the detriment of Metro

or other parties, must not be tampered with, must not be used in advertisements, must be cited as provided by Metro.

c. The referenceUSA data has tighter restrictions, and only public knowledge about a business can be published. Other, private data must remain private.

11) Are there any legal or ethical issues with respect to the data?

a. To my knowledge there are no legal or ethical issues with respect to the data sets that I have chosen. They are all made available to the public with explicit terms and conditions. Consequently, legal and ethical issues will not arise if the data is used according to the terms and conditions.

12) What organization or project developed the data?

a. I currently have data from five sources

i. LA County
ii. LA Metro
iii. referenceUSA
iv. LAPD
v. LASD

13) For what purpose was the data developed?

a. The LA County and LA Metro data were developed for both internal use and an open data initiative.

b. referenceUSA data was aggregated and delivered by a commercial company for profit. Consequently, there are limits on the amount of data that can be retrieved from their database.

c. The LASD and LAPD data are made available to the

14) When was the data created or released?

a. LA Metro Data – December 2014
b. LA County Zipcodes – 2010
c. LA County Parcels – 2011
d. referenceUSA – Unknown
e. LAPD – Unknown
f. LASD – 2015 (Past 30 days)

i. There are historical files for 2005 – 2014
ii. There is no way to get previous months in 2015

15) Which data model (raster or vector) is used?

a. All of the data is in vector format so far.

b. I am investigating transforming the parcel data into a binary raster, but have not yet done this.

16) What is the file format?

a. Files are provided as shapefiles or CSV files.

17) What is the spatial resolution of the data? What is the original cell size if the data model is raster? What is the original scale if the data model is vector?

a. With the exception of rail lines, all of the data is point data. Consequently, resolution is not stated. That said, it is noted that the LASD data has been altered slightly in order to protect specific locations.

b. The OSM basemap data can be scaled to several resolutions.

18) What is the coordinate system of the data? Which datum and map projection are used?
    a. LA Metro data uses GCS_WGS_1984
    b. LASD and LA County zip code data is in NAD_1983_StatePlane_California_V_FIPS_0405_Feet
    c. LAPD and referenceUSA data require geocoding and consequently have no coordinate system or datum natively associated with them.

19) Which attribute item(s) can be used? What is the unit(s) of the item(s)?
    a. LA Metro Line and Station Names
    b. Crime Category, Location, and Description
    c. Business Name, Type, and Location

Data Quality Assessment

LA Metro Stations:

1) Lineage
    a. This data is provided directly by LA Metro and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap
    b. Stations do appear to align with where they would be located on OSM
    c. Stations are represented as points, consequently, different lines that connect in the same station have slightly different locations
3) Attribute Accuracy
    a. Station names are correct
4) Logical Consistency
    a. Names are not always consistent between data sets for different lines
        i. eg. 7th Metro Station vs. Metro Station
5) Completeness
    a. The data set includes all stations for each line as of December 2014
    b. This could be an issue when new stations are being opened eg. Expo Line
    c. All stops for my study area are available
6) Temporal Accuracy
    a. This data is current as of December 2014

LA Metro Lines:

1) Lineage
    a. This data is provided directly by LA Metro and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap
    b. Lines that are above ground (light rail) appear to follow the track on OSM
3) Attribute Accuracy
    a. Line names are labeled correctly
4) Logical Consistency
    a. The Expo and Blue lines have separate files, however, Red and Purple do not. It is not clear why since both share a track at some point, and both go to different locations.
5) Completeness
    a. The lines are all fully represented on the map
6) Temporal Accuracy
    a. This data is current as of December 2014

LA County Zipcodes:

1) Lineage
    a. This data is provided directly by LA County and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap
    b. I can't say with authority where the zip code boundaries are; however, the zip codes are consistent with my own local knowledge
3) Attribute Accuracy
    a. Zip code labels are correct
4) Logical Consistency
    a. Each polygon is consistently labeled
5) Completeness
    a. All zip codes for my study area are available
6) Temporal Accuracy
    a. Current as of 2010
    b. Zip codes don't change often

LA County Parcels:

1) Lineage
    a. This data is provided directly by LA County and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap
    b. The polygons do appear to align with the OSM polygons
3) Attribute Accuracy
    a. Attributes appear to be correctly labeled
4) Logical Consistency
    a. It is not clear what scheme is used to define labels. A large swatch of my study area, multiple polygons, are simply labeled "Civic Center"
5) Completeness
    a. There is very little data in this data set for my study area
6) Temporal Accuracy
    a. This data is from 2011, and the area changes rapidly. Consequently

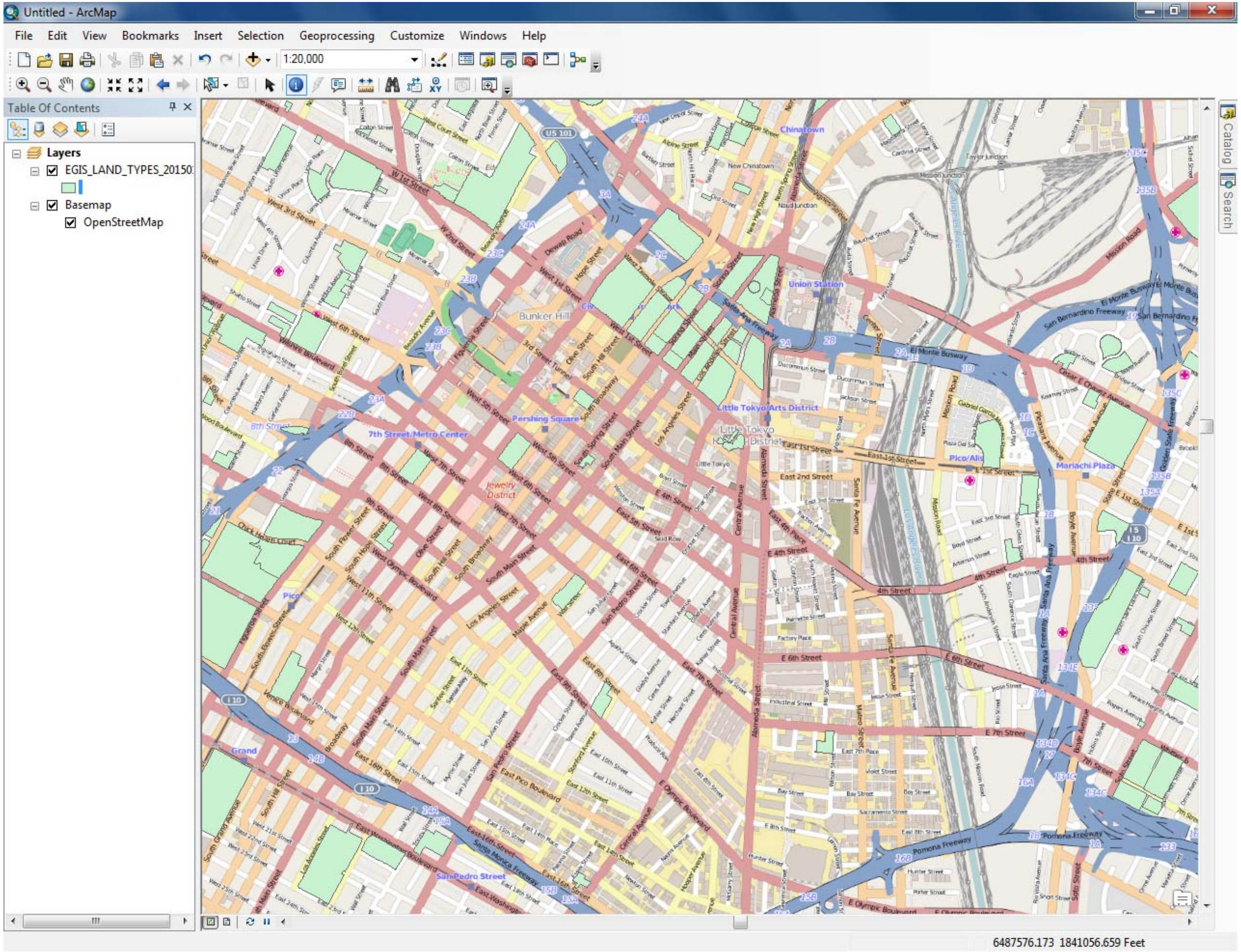

Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:20,000
Table Of Contents
Layers
EGIS_LAND_TYPES_20150
Basemap
OpenStreetMap
Catalog
Search
6487576.173 1841056.659 Feet

Figure 4 – Incomplete Parcel Data

LAPD Crime Data:

1) Lineage
    a. This data is provided directly by LAPD and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap
    b. This data needs to be geocoded based on the address provided
3) Attribute Accuracy
    a. The attribute labels are not clear
4) Logical Consistency
    a. Attributes are not clearly labeled
    b. It is difficult to determine the quality of this data set
5) Completeness
    a. I don't know how to define completeness for crime data
6) Temporal Accuracy

LASD Crime Data:

1) Lineage
    a. This data is provided directly by LASD and is authoritative
2) Positional Accuracy
    a. I should note that I am using OpenStreetMap as my basemap.
    b. XY Coordinates roughly correlate with the addresses provided in the report
    c. Location data has been intentionally obscured
3) Attribute Accuracy
    a. Points are clearly classified and labeled
4) Logical Consistency
    a. Labels are consistent from point to point
    b. Clear classification system in place
5) Completeness
    a. Again, I don’t know how to define completeness for crime data
6) Temporal Accuracy
    a. This data represents crimes that occurred in the last 30 days
    b. Historic data sets are available
    c. Each point has a timestamp

referenceUSA:

1) Lineage
    a. This data is provided directly by referenceUSA
2) Positional Accuracy
    a. Data has to be geocoded using addresses
    b. Some records have a lat/long in the description
        i. I missed this initially because it is in the demographic pane and not the location pane.
3) Attribute Accuracy
    a. Clear labeling schema
    b. There is a verified label attached to verified entries
4) Logical Consistency
    a. Clear categories and organization
    b. The same categories and labels are used throughout
5) Completeness
    a. There is a significant amount of data for each point
6) Temporal Accuracy
    a. There is no timestamp that I can find that defines when the data was collected
    b. The only measure of time is “years in database”

Challenges

I have encountered several challenges in this project to data. The biggest challenge is obtaining complete business data for my study area. I would really like to be able to identify as many points of interest as possible in order to provide the best possible value to the end user. I would like to be able to sort or select points of interest by category such as type of location e.g. museum, restaurant, etc. It would also be nice to provide time windows that various attractions are open

I would also like to incorporate recent crime data to determine whether or not some areas have a higher concentration of crimes. This is partially achieved through LASD data; however, the LAPD data has proven to be difficult to work with.

Finally, determining walkability by land use has proven difficult because of the lack of completeness of the parcel data. I will have to check the LA County portal for a different, more complete, dataset.