Anthony Hakim, Marc Loeb, Sasha Filippova, Yifu Hou

Deliverables #2: Project Check-in

Please provide one paragraph description of the goals of your project. You can list the same description from the previous deliverable or provide new details about aspects that have changed since Week 4.

Our goals have remained largely consistent with those of Week 4. We intend to explore the factors that influence urban growth/urban decay in the city of Chicago, and attempt to predict which parts of the city will experience urban change in future given those factors. The unit for our regression is Chicago's 77 community areas. The dependent variable of our analysis is the relative number of new construction to building demolitions in each community area. We still intend to run regressions with data on race, income, and other metrics of socioeconomic status as independent variables.

However, for now we have set aside our ambition to display our results with an *interactive map* that allows users to select different regressors and neighborhoods. This functionality will be more command-line oriented, and will return static maps to the user.

For each source of data that you expect to use, please list the source of data, who will be responsible for collecting data from that source, and a date by which you expect the work of gathering the data from that source to be complete.

Below are some potential datasets we found for the regression model (subject to change). We access each dataset through API, and test on some variables within each dataset to find out the best dependent and independent variables for the regression model. Yifu and Sasha already wrote a function that allows extraction of data from the City of Chicago portal through API. Yifu & Sasha are expected to complete all data gathering by March 5 except the Building Permits dataset (Marc already accomplished the task).

SOURCES:

Community Area Boundaries:

The 77 community areas of Chicago serve as our unit of analysis. This dataset consists of the spatial polygons of each Chicago Community Area.
Link:
https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

Building Permits, Chicago Open Data Portal:
Link:
https://data.cityofchicago.org/Buildings/Building-Permits/ydr8-5enu

This database of the building permits issued by the city of Chicago since 2006 will serve as the basis of our dependent variable. The permits are organized into several categories, two of which are permits for new construction projects, and permits for building demolitions.

This dataset required substantial pre-processing. Many entries were missing their community area codes, or their geographic coordinates entirely. Luckily, almost all had a complete street address. Marc used the geocoding package GeoPy to convert these street addresses into lon and lat coordinates. These could in turn be used to spatially join the building permits to the community area polygons. With this process complete, Marc used groupby and count to aggregate the permits, and find the total number and value of new construction and demolition by community area.

The work of gathering data from this source is complete.

2020 Census Supplement: Chicago Community Areas:
Time: 2020
Geo-information: Community area name
Description: 2020 Chicago census data, total population count and counts by race, by community areas. This will allow us to normalize the data for building permits, crimes, and other count-based datasets.
Link:
https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data/resource/0916f1de-ae37-4476-bf4e-6485ba08c975

Crimes:
Time: 2001-present:
Geo-information: Longitude and Latitude

Description:
This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present.

Link: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

Socioeconomic and Health Indicators:
Time: 2008 – 2012
Geo-information: Community area name

Description:
This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," by Chicago community area, for the years 2008 – 2012.

Link:
https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Hardship Index:
Time: 2006 – 2010
Geo-information: Community area name

Description: This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," by Chicago community area, for the years 2006 – 2010.

Link: https://data.cityofchicago.org/Health-Human-Services/hardship-index/792q-4jtu

Red Light Camera Violations:
Time: 2014 – current
Geo-information: Longitude and Latitude

Description: This dataset reflects the daily volume of violations created by the City of Chicago Red Light Program for each camera.

Link: https://data.cityofchicago.org/Transportation/Red-Light-Camera-Violations/spqx-js37

Affordable Rental Housing Development:
Time: 2013 - current
Geo-information: Community area number

Description: The rental housing developments listed below are among the thousands of affordable units that are supported by City of Chicago programs to maintain affordability in local neighborhoods.

Link:
https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi

Vacant and Abandoned Buildings:
Time: 2011 - current
Geo-information: Street address

Description: Vacant and abandoned building violations issued on properties owned by financial institutions since January 1, 2011.
Link:
https://data.cityofchicago.org/Buildings/Vacant-and-Abandoned-Buildings-Violations/kc9i-wq85

Please give a brief sketch of the work that needs to be done to complete your project (other than data collection), include a description of which team member(s) will be responsible for completing this work and the expected timeline for completion

Marc:
Mapping of spatial data. Grouping and normalization of count based datasets. Any additional geocoding or geo joining (March 1). Command line program to allow user to final package to test the geocoding of permit dataset, which is too computationally intensive to be done "on the fly" (March 5).

Yifu:
Complete function to programmatically gather data with API, process DataFrame to provide a joint dataset for regression (March 5).

Sasha:
Data cleaning and explore possible solutions to build a predictive module (March 5).

Anthony:
Use SKLearn package to find regressors and design regression function to access DataFrame (March 6).

All:
Build the interactive part of the project(March 12)

Include any additional information you wish to provide about your project.

None

If you need additional feedback from me then please make sure to include that in your report.

- How can we test the accuracy of our model to predict development/decay?
- Could you recommend specific packages to run a regression model?
- Our most pressing concern is how to handle interactivity. Initially we considered giving users the ability to select their own regressors, but we have agreed to set that aside. Instead, we are considering allow users to specify a community area to get more information about. And perhaps run a sub-regression between building permits and crime within that community area, to see how the regression coefficients shift across the city as a whole.