Cateloging ~ G-Drive

The G-Drive Contents

Authors: Kirsten Riley and Jorge Valdebenito

Copy and Paste this short cut to find the folder on https://rds.syr.edu/rdweb/webclient/ in the search bar.

"G:/MAX-Filer/Collab/Labs-kbuzard-S18" (you may have to change to backslashes)

Opening the files:

File type	How to Open
.shp	open ArcGIS, open a blank map, add a layer, and open the file in the program
python files	open using Spyder (Anaconda3)
.xml	open using excel

The ramosRivera Folder

bg06_d00_shp Folder

has three documents - all of them make up the map of California used in paper.

- **bd06_d00** - shows the map of California broken down by zipcodes (opened in ArcMap)
- **bg06_d00** - 9 observation: Area, Perimeter, BG06_D00, BG06_D00_I, State, County, Tract, BLKGROUP, and NAME. These make up the information needed to recreate the California map (opened in excel)
- **bg06_d00.shx** - shaping file for the California map (when opened in excel, it does not make much sense)

DART_IRL Scans Folder

Has two Pdfs copies of information on R&D labs and their location

- **1979IRL16** - Industrial Research Laboratories of the US, 16th Edition 1979
  - Original document containing information on the Industrial Research Laboratories of the US. Has information on 9,907 R&D facilities belonging to 6,323 organizations in 1979.
- **1989DART23** - Directory of American Research and Technology 1989, 23rd Edition
  - Original document containing information on organizations active in product development for business in American. Content includes information on 11,275 organizations in alphabetical order.

Summer2021_Dylan Folder

Has eight documents all pertaining to the 1979 and 1989 data from IRL and DART pdfs

- **1979_Digitized** - digitized version of the 1979 IRL pdf
- **1979IRL16** - copy of the 1979IRL16 pdf in DART_IRL Scans folder

- **1989_Digitized** - digitized version of the 1989 DART pdf
- **1989DART23** - copy of the 1989DART23 pdf in DART_IRL Scans folder
- **corr_cattLabs97_Wgeocode 1-6200** - excel file with lines 1-6200 corrected by Dylan
- **corr_cattLabs97_Wgeocode** - original excel file before Dylan and Kelly worked on it
- **corr_cattLabs97_Wgeocode_Line 6200 to Line 12765** - excel file with lines 6200-12765 corrected by Kelly
- **OCR_Result_NO_user** - Antonio's intial OCR scan.
  - This data was input into the corr_cattLabs97_Wgeocode excel sheets

Summer2021_Kelly Folder

Task From Antonio 1 folder

Original Material Folder

- **corr_cattLabs97_Wgeocode** - original excel file before Dylan and Kelly worked on it (duplicate)
- **letter_I_cattell** - digitized version of the research labs starting with the letter "I"
- **letter_O_cattell** - digitized version of the research labs starting with the letter "O"
- **letter_S_cattell** - digitized version of the research labs starting with the letter "S"

  this separation by letter section was done to make digitization process faster.
- **OCR_Result_NO_user** - Antonio's intial OCR scan. (duplicate)
  - This data was input into the corr_cattLabs97_Wgeocode excel sheets
- **corr_cattLabs97_Wgeocode_Line 6200 to Line 12765** - excel file with lines 6200-12765 corrected by Kelly (duplicate)

Task From Antonio 2 Folder

- **1979_Digitized** - digitized version of the 1979 IRL pdf (duplicate)

Task From Antonio 3 Folder

- **1989_Digitized** - digitized version of the 1989 DART pdf (duplicate)
- **1989_OCR_Digitized** - digitized version of the 1989 DART pdf from the OCR machine (unedited)
- **corr_cattLabs97_Wgeocode_Line 6200 to Line 12765** - excel file with lines 6200-12765 corrected by Kelly (duplicate)
- **OneDrive_2021-08-27** - zip drive that leads to the original material folder in Task From Antonio 1 folder (duplicate)

T-Burk Folder

It has 11 folders:

ArcMap Folder

- **Converted_Graphics (.cpg, .dbf, .prj, .shp, .shx)** - it only shows a greeen rectangle
- **Textile Labs (.cpg, .dbf, .prj, .sbn, .sbx, .shp, .shx) -**

- **ZCTAs (.cpg, .dbf, .prj, .sbn, .sbx, .shp, .shx)** - Openning the files in ArcMap it shows California in the ZCTAs areas and the location of the labs (dots).

BlockData Folder

- **nhgis0003_shapefiles_tl2000_560_block_2000**
- **AK_block_2000 (.dbf, .prj, .sbn, .sbx, .shp, .shx)** - one of these files for each state. I can open all the files, but hard to visualize (I am not familiar with ArcMap)
- **DC_block10_2000 (.dbf, .prj, .sbn, .sbx, .shp, .shx)** - Able to open. Shape of DC in zip blocks (I assume)
- **MA_block10_2000 (.dbf, .prj, .sbn, .sbx, .shp, .shx)** - Able to open.
- **ak_wac_S000_JT00_2002.csv.gz** - (This are 7.zip files. There is one for each state. The year vary for some states. I can open the cvs file in Excel. It is geocode)
- **USA_block (.cpg, .dbf, .prj, .shp, .shx)** - input and output for shapstich. Input file for USA_block_emp
- **usa_blockEmp (.cpg, .dbf, .prj, .shp, .shx)** - I do not know what this is showing. Output file for USA_block_emp.

k-function_local_results Folder

- **Manufa_Emp_C000_0.5_Buffers_2 (.cpg, .dbf, .shp, .shx)** -
- **Manufa_Emp_C000_0.25_Buffers_2 (.cpg, .dbf, .shp, .shx)** - input file for start_calc
- **Manufa_Emp_C000_0.75_Buffers_2 (.cpg, .dbf, .shp, .shx)** -
- **Manufa_Emp_C000_1_Buffers_2 (.cpg, .dbf, .shp, .shx)** -
- **Manufa_Emp_C000_2_Buffers_2 (.cpg, .dbf, .shp, .shx)** - input file for start_calc
- **Manufa_Emp_C000_5_Buffers_2 (.cpg, .dbf, .shp, .shx)** -
- **Manufa_Emp_C000_10_Buffers_2 (.cpg, .dbf, .shp, .shx)** - input file for start_calc

All the previous files have missing spatial reference information. The data can be drawn in ArcMap , but not projected. ArcMap doesn't show anything

- **Manufa_Emp_C000_Points_2 (.cpg, .dbf, .shp, .shx)** - input file for start_calc
- **Manufa_Emp_C000_local.txt** - This is a log file with the date (04/05/2021) and time slot of some code running.

The files show the location of Manufacturing employment clusters I belive in California. They should correspond to Figure 1 and 2 of the draft.

LabData Folder

- **cal_lab_fields** – A folder for 34 different indutries i.e. AERO –AERO (.cpg, .dbf, .prj, .shp, .shx). Output file for field_org.
- **comb_emp_C000_local** -
- **Manufa_Emp_C000_0.5_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_0.25_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_0.75_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_1_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –

- **Manufa_Emp_C000_2_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_5_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_10_Buffers_cal0 (.cpg, .dbf, .shp, .shx)** -

  All the previous files have missing spatial reference information. The data can be drawn in ArcMap , but not projected. ArcMap doesn't show anything
- **Manufa_Emp_C000_Points_cal0 (.cpg, .dbf, .shp, .shx)** –
- **Manufa_Emp_C000_local.txt** - This is a log file with the date (04/18/2021) and time slot of some code running.
- **USA_labs_2000 (.cpg, .dbf, .prj, .shp, .shx)** - output file form prep_Labs

PatentData This is probably used to replicate Buzard 2017.

- **.RData** -
- **Rhistory** -
- **CA Control_1_ALT_ramos** (SAS Program)
- **cite_same** (Excel) - input file for start_calc
- **cite76_06** (SAS Data set)
- **clustpatents** (SAS Data set)
- **columnsEFI_CAbaseline** -
- **columnsEFI_NEbaseline** -
- **LA5A_ALT (.cpg, .dbf, .shp, .shx)** –
- **LA5B_ALT (.cpg, .dbf, .shp, .shx)** –
- **LA5C_ALT (.cpg, .dbf, .shp, .shx)** –
- **LA10A_ALT (.cpg, .dbf, .shp, .shx)** –
- **LA10B_ALT (.cpg, .dbf, .shp, .shx)** –
- **list_of_matches_CAbaseline_ramos** –
- **originating** (SAS Data set) –
- **pat76_06** (SAS Data set) –
- **replications_CAbaseline** (Excel) –
- **SASclustpatentsCA** (Excel) –
- **SASoriginatingCA** (Excel) –
- **SASpossiblenclassCA** (Excel) –
- **SB5_ALT (.cpg, .dbf, .shp, .shx)** –
- **SB10_ALT (.cpg, .dbf, .shp, .shx)** –
- **SD5A_ALT (.cpg, .dbf, .shp, .shx)** –
- **SD5B_ALT (.cpg, .dbf, .shp, .shx)** –
- **SD10_ALT (.cpg, .dbf, .shp, .shx)** –
- **SF5A_ALT (.cpg, .dbf, .shp, .shx)** –

- **SF5B_ALT (.cpg, .dbf, .shp, .shx)**–
- **SF10_ALT (.cpg, .dbf, .shp, .shx)** –
- **tables** (word) –Table 2a is Table 3 is the draft –Table 2b is Table 4 in the draft –Table 3b is Table 5 in the draft. The draft only uses 5 and 10 miles ratio

PngData Folder

- **OCR_Output_1998** –
- **letter_I_cattell** (text) –
- **letter_O_cattell** (text) –
- **letter_S_cattell** (text) –
- **OCR_Result** (text) – input for Address_ID, output for OCR
- **OCR_Result_NO_user** (text) All this files looks like the registry of labs.
- **1979 Digitized** (text): Registry of labs
- **1989_OCR_Digitized** (text): Registry of labs
- **calLabs97** (Excel): File with company name, facility name, state, ID and address for 1997
- **cattell_1997_raw** (STATA) - input file for state_code_rep
- **Cattell_corr_list** (STATA) - input file for pngwork.
- **cattell-all** (STATA)
- **cattLabs97** (Excel) - input file for Address_ID, Geo_coder, and state_code_rep. Output file for pngwork
- **CattwithBuzID** (Excel)
- **corr_cattLabs97** (Excel) - input file for field_org, firm_struc, and start_calc. Output files for state_code_rep.
- **corr_cattLabs97_Wgeocode** (Excel): This one has a column counting the observations. Only difference with the file below.
- **corr_cattLabs97_Wgeocode** (Excel)
- **field** (STATA) - input file for field_org, pngwork, start_calc
- **field_lab_counts** (EXCEL): count by sector. There are no differences with the file below
- **field_lab_counts2** (EXCEL): count by sector
- **field-master** (STATA) - input file for field_org
- **geocoded_facilities (EXCEL)**: has 8,737 observations. Input file for prep_Labs. Input file for shapify. Output file for Geo_coder
- **geocoded_facilities_cal** (EXCEL): has 1,728 observations
- **geocoded_facilities_I** (EXCEL): has 394 observations
- **geocoded_facilities_O** (EXCEL): has 198 observations
- **geocoded_facilities_S** (EXCEL): has 886 observations
- **id_dataString** (EXCEL): has the id, the full address and the buzzID
- **matched_data** (EXCEL): has 8,941 obs. input file for Geo_coder.

- **matched_data_I** (EXCEL): has 199 obs. Not sure what is matching or with which file.
- **matched_data_O** (EXCEL): has 199 obs. Not sure what is matching or with which file.
- **matched_data_S** (EXCEL): has 890 obs. Not sure what is matching or with which file.
- **newData** (EXCEL): has 28,515 obs. 39 variables. information from the entire US (by loking at the states)
- **pngbuzz** (EXCEL): has 2,951 obs. 39 variables. information from the entire US (by loking at the states)
- **pngCatIDList** (EXCEL): has 11,313 obs. 5 variables. Output for Address_ID.
- **single_lab_firm** (EXCEL): has 7,430 obs. 21 variables. Input file for firm_struc. Out file for pngwork.

Next step is to go to png website and see which files are downloaded from there and which ones were created by Antonio.

Python Scripts Folder

- **.pylint.d** -
- **stat_calc1.stats** (STATS):
- **Address_ID** – Preparing and cleaning addresses
- **clust_pat_maker** – Python Script to read in patent data and conduct a spatial join with clusters then keep the patents that fall into those clusters as geodataframes and export them
- **countSim_speedUP** – Point Count Simulation Computation. It turns the dictionary back into a dataframe.
- **countSim_tester** – Same as "countSim_speedUpr:" but measure the time for each individual loop and the entire system (time it takes to preform the simulation).
- **field_org** – Read in 1997 cattell lab data as well as my geocoded data and combine the two to produce geodataframes (gdf) for each technology field in the cattell directory which then get saved as shapefile currently the program is set in such a way so as to produce gdfs for the country wide data and gdfs for california and the NE corridor the script will then take the dictionary containing the california labs by field and save each gdf as a shapefile.
- **firm_struc** – Reads in the 1997 Catteell directory data produced by Ivan Png 2016. It takes the data in this data set of American R&D Labs and organizes it based on firm structure. The final product is an Exceell File that gets outputted.
- **GeoCode_OCR**
- **GeoCoder** – Iterate through the address data and geocode each input address.
- **multiprocess_test2** – processing time information
- **multiprocessing_tester** –
- **OCR** – imports images, edit them to use with "tesseract"
- **Pdf2Jpg.py** –
- **pngwork** – create dataframe for firms that have at most 2 establishments, this will become the dataframe for firms with only one research establishment.
- **Prep_Labs** – python script which reads in two csv of geocoded labs and joins them resulting in a pandas dataframe. Then it takes the coordinates for the labs in the dataframe and creates a geometry column to turn the df into a geodataframe it then saves the resulting geodataframe as a shapefile.

- **Prep_ZBP** – python script that uses pandas and geopandas packages to read in census manufacturing employment data at the ZCTA level and shapefile of all ZCTA bounderies in the contiguous US. The employment data is prepared and the merged into the shapefiles data table resulting in a geopandas geodataframe which gets saved as a shapefile.
- **shapeStich** – create the new geodataframe by appending all state level gdfs.
- **shapify** – Reads in a csv file that contains point data in latitude and longitude form and converts them into geopandas geodataframe the result is then exported as a shapefile.
- **stat_calc** – Read in the point file and associated cluster files produced after running the 3Stage_Local program and calculates various statistics from it. Different sections of this program produce different stats and have been partitioned and commented accordingly.
- **state_code_rep** – Fix the state_code column of the cattell png data for 1997
- **usa_block_emp** - reads in a list of csv files containing employment data at the block level. It then reads in a shape file of all US census block boundaries and merges the employment data into the shapefiles data table and saves the resulting geodataframe as a shapefile.

Tables folder excel tables used in paper

- **5_mile_LDS** - Shows Originating Patents, Citing Patents, From Same Cluster, Percent (C/B), Treatment Patents, Treatment Citing For Same Cluster, Percent (F/E), Control Patents, Control Citing From Same Cluster, Percent (I/H), Location Differential (G/J), and P-values for 5-mile cluster in California. (excel) Output file for stat_calc
- **10_mile_LDS** - Shows Originating Patents, Citing Patents, From Same Cluster, Percent (C/B), Treatment Patents, Treatment Citing For Same Cluster, Percent (F/E), Control Patents, Control Citing From Same Cluster, Percent (I/H), Location Differential (G/J), and P-values for 10-mile cluster in California. (excel) Output file for stat_calc.
- **Spatial_LDS** - Table that compares the 5 and 10 mile clusters (excel) Output file for stat_calc.

ZipData folder first folder is a dublicate folder of the "nhgis0005_csv" folder found below

tl_2010_us_ZCTA500 Folder

- **tl_2010_us_zcta500** (DBF, PRJ, Adobe, XML, SHX) - input file for prep_ZBP

Not in folder

- **0SF3_geo_header** - Data dictionary, explains U.S. Abbreviations, Geographic Area Codes by region, divisions, state (census, state (FIPS), county size code, FIPS County Subdivisions Class Code, Place Size Code, etc. (Word document)
- **Employment** - SAS Graph document created to collect ZIP code employment data for California
- **USA_ZCTA_emp(CPG, DBF, PRJ, SBN, SHX)** - Map of the US separated by ZCTA or zipcodes, output for prep_ZBP.

nhgis0005_shape Folder Files will not open because the folders are compressed.

nhgis0004_shapefile_tl2000_330_block_2000 (zipped Folder)

- **NH_block_2000** - (DBF, PRJ, SHX, SBN, SBX, XML) -

nhgis0004_shapefile_tl2010_110_block_2000 (zipped Folder)

- **DC_block10_2000 (DBF, PRJ, SHX, SBN, SBX, XML)** -

nhgis0004_shpaefile_tl2010_250_block_2000 (zipped Folder)

- **MA_block10_2000 (DBF, PRJ, SHX, SBN, SBX, XML)** -

nhgis0005_csv Folder NHGIS data from 2000

- **nhgis0005_ds151_2000_zcta** - excel files with 55 variables but only 32 variables have observations. Contains GISJOIN from the year 2000. Input file for prep_ZBP.
- **nhgis0005_ds151_2000_zcta_codebook** - describes the variable labels in the nhgis0005_ds151_2000_zcta excel file and where the data was ciphered from.
  - ex: GISJOIN: GIS Join Match Code
  - It also contains what the NHGIS codes are (ex: GMH001: Male » Agriculture, forestry, fishing and hunting, and mining)

tl_2010_06_zcta500 Folder

Has five documents in different formats, builds map of California by census block.

- **tl_2010_06_zcta500.dbf** - 11 observations: STATEFP00, ZCTA5CE00, GEOID00, CLASSFP00, MTFCC00, FUNCSTAT00, ALAND00, AWATER00, INTPTLAT00, INTPTLON00, PARTFLG00 (opened in excel)
- **tl_2010_06_zcta500 (.prj, .shp, .shx, .xml)** Map of California by census block (opened with arcGIS), U.S Department of Commerce, U.S. Census Bureau, Geography Division 2010. (xml File)
  - Vector digital data from http://www.census.gov/geo/www/tiger

~ **Not in a Folder** ~

- **cattell-all** - 18 variables: Parent ID (new) *(referring to parent facility)*, year, parent ID (Cattell original) *(referring to pdf scans parent facility ID)*, Parent name, Facility name, Facility ID (Cattell original) *(referring to pdf scans facility ID)*, Facility ID (new), zipcode *(the zipcode the facility is in)*, Facility level, user, prof, doct, tech, parent name (alternative 2), parent name (alternative 3), parent name (alternative 4), state. (Stata file)
  - figure out what new vs cattell original is?
- **Dylan & Kelly notes from Summer 2021** - (5/30/2022) Dylan and Kelly's documentation on their work
- **Dylan & Kelly notes from Summer 2021** - (6/3/2022) Dylan and Kelly's documentation on their work with notes from Prof. Buzard
- **field** - Shows Stata data on the Cattell ID for R&D fields and R&D sub-fields, the year the data was on, and the facilities (Stata file).
- **pngwork** - python script that uses the cattel-all.dta, field.dta, and a file called "GoodLabs.shp" for points*(has not been found)*

ramosRivera - Backup062122 Folder

This folder is a duplicate of the ramosRivera Folder created on June 21, 2022 as a backup to the original ramosRivera folder.