Comparing Address Lists Demonstration
Workshop Training
K Lefebvre – Feb 8 2018

Importing/Exporting CSV Files in Excel

CSVs, "comma separated values" files, are especially useful for storing large sets of information in a small text-based format. Tables are stored as plain text, with cells separated or "delimited' by a single non-alphanumeric character, usually commas, but alternately delimiters can be tabs, semicolons, commas, vertical lines, or a number of other symbols. You will need to often work with CSV files, both importing and exporting them, for GUPS purposes. In general it is best to import rather than open CSV files as opening them may cause formatting to drop zeroes from census block codes (e.g. 013 would become 13).

To import a CSV, we'll open a blank workbook, go to the Data tab and click the "From Text" button in the "Get External Data" menu. Navigate to the folder containing the CSV file, and then click Open. In the "Text Import Wizard" window, leave the "Original data type" as the default "Delimited" and, in general, you'll want to start at import row "1" with the default file origin. Click Next> and then select your delimiter if it's listed, or otherwise enter it in the "Other" input box. If you do this, make sure any other delimiters are not checked by accident. You'll be able to see if your import will be a success in the data preview. Click Next>. In Step 3 you can select each of your columns and change their types to General, Text, Date, or set them to be skipped during the import. In most cases you will likely want to leave them as General.

After clicking Finish, the Import Data window will prompt whether you want to add this to an existing worksheet or create a new one. In this case we'll use the default value for the existing worksheet, as our workbook is blank.

Now that our CSV is imported, we can save it with any formatting changes like highlighting, font, etc.; features that would otherwise be lost as a CSV file.

Concatenating and De-concatenating Spreadsheet Fields

Even if you have no prior experience, concatenating addresses is very easy in Excel. To do this, all you need to do is select a blank column for your formula, and then type in =CONCATENATE(
This initiates the concatenate formula, once you've typed in the end parenthesis, hold CTRL and click on any number of cells to put them together. To add a space between cells you will have to enter it in manually between their names, with two quotation marks. So for example if I wanted to concatenate a street name with a municipality with a space in between them, I would enter =CONCATENATE(N2, " ", P2).

If the =CONCATENATE() function doesn't work, double check that you've used commas between cells rather than pluses (+) or any other operative symbols. Additionally, make sure you have "quotations" on both sides of your spaces or other text that you'd like to add. If the function is displaying rather than the expected value (i.e. you've entered the formula but it remains in the cell even after pressing return), right click on the column in question and select "Format Cells". Then change the category from "Text" to "General". Click okay, and then after double clicking on the cell with the formula and hitting enter, it should return the correct value.

To separate spaced information like addresses, we could use Text to Columns, however one limitation of this tool is that it will cause problems where streets have multiple words in their names
e.g.

Address	NewField1	NewField2	NewField3	NewField4
99 WATER ST	99	WATER	ST	
99 MIDDLE WATER ST	99	MIDDLE	WATER	ST

One alternative for this is the LEFT() and RIGHT() functions to place the first and last number of characters into another cell. Again this has limitations as irregular addresses beginning and ending with words may become unrecognizable with this tool. E.g. whereas 99 WATER STREET yields 99 and ST, SHERI LANE ESTATES gives us SH and ES.

Nested Formulas

An Excel template is provided with this handout which includes these formulas-

First word/numbers of address-

=TRIM(LEFT(SUBSTITUTE(A2," ",REPT(" ",100)),100))

Middle words/name of street-

=MID(A2,FIND(" ",A2)+1,FIND(CHAR(1),SUBSTITUTE(A2," ",CHAR(1),LEN(A2)-LEN(SUBSTITUTE(A2," ","")))))-(FIND(" ",A2)+1))

Last word/street type-

=TRIM(RIGHT(SUBSTITUTE(A2," ",REPT(" ",100)),100))

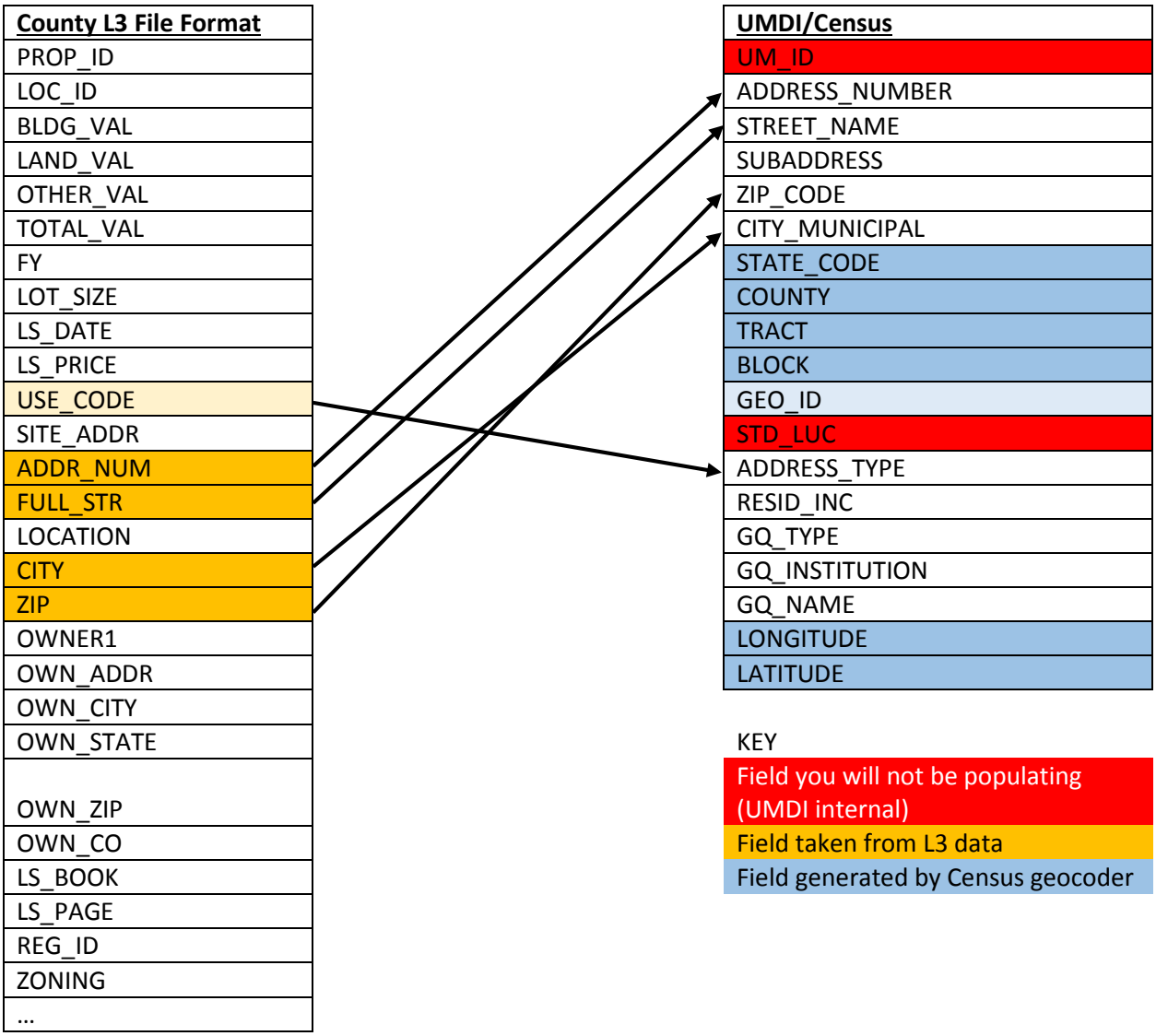
Using the combination of TRIM SUBSTITUTE formulas provided above will allow you to extract the entire first, middle, and last words, including any irregular street names or suffixes. Note the text in green represents the cells that you would change out depending on the location of your address information. After processing data with these operations, we can filter for unusual addresses by selecting our column, going to the data tab and clicking the filter button. A column with a filter is indicated by an arrow button on the right side of the first cell in the column in question.

Some common problems- if your function is displaying on its own rather than returning the correct value, this could be because the cells it is entered in are formatted as text. To change this simply right click on the column, row, or cells in question and select Format Cells. There you will be able to change the format from text to “General”, after doing this, double click on the cells in question and press enter to reactivate the formula, which should allow the function to return a value.

Standardizing Data Fields

If your municipality is working with GUPS, the Census Bureau's Geographic Update Partnership Software (GUPS), you'll want to standardize your L3 assessor's data into a format which mimics that of the Census's. In creating Dashboard 2, which shows counts of group quarters (GQ) and individual housing units, UMDI has attempted to make formatting that is compatible with this.

To convert L3 assessors files to this format, you'll need to select your USE_CODE's and convert them using your assessor's/municipal key (often a local key based Property Type Classification Codes by the Massachusetts Dept. of Revenue) to populate the GUPS ADDRESS_TYPE field. Address numbers, street names, zip codes, and the city names will all easily be reorganized, however some fields like latitude/longitude, and codes for the state, county, tract, and census block, will need to be obtained by using the Census geocoder.



Adding Census Data

Latitude/Longitude, State Code, County, Tract, and Block

Part of this will require finding the Latitude/Longitude coordinates of the entries. We can do this using the U.S. Census Geocoder, which accepts CSV files with the fields-
[Entry Number] | [Number and Address] | [Town/City] | [State] | [ZIP]
but without any headers included in the file.

Copy over numbers and addresses from your file, and concatenate them into one field if they're in separate columns. Then copy the town/city name, and if your data is missing a "state" field, create a new column, enter "MA" for that field and double click the black box in the lower right-hand corner of that cell to populate all subsequent rows thereon.

Formatting Note

When copying over ZIP codes, make sure that your records are showing all 5 digits. If your ZIP code starts with zero, oftentimes Excel will drop that preceding digit. If this happens, it can be fixed by right clicking on your column header, selecting Format Cells, and changing the column type to Text. After doing this, filter for the ZIP code you are trying to fix (if you are using a sheet with multiple towns), type in the correct ZIP code with the preceding zero, and double click the black box or otherwise click and drag it to replace all preceding incorrect entries.

Our first field, the Entry Number, is a dummy variable only needed by the geocoder. It will simply be an integer generated by Excel by entering 1 and 2 in the first two respective rows, selecting both, and double clicking on the lower right corner box to auto-fill integers for every record.

Again, **the Census Geocoder cannot function with table headers in your CSV** (e.g. Entry, Address, City, etc.) so if your file contains field names in the first row, delete said row before proceeding to upload this file to the geocoder. To access the tool, go to https://geocoding.geo.census.gov/ and then select Address Batch from the "Find Geographies Using…" menu on the left hand side. Select the CSV file we just created with our addresses, and leave the Benchmark as Public_AR_Current, which should be the default. Click "Get Results" and the geocoded addresses should download. In extreme cases this process may take up to an hour, so please do not refresh the page or close out of your browser; you can however minimize this window while the geocoder is processing.

If you open the resulting GeocodeResults.csv file, you'll find that the latitude and longitude have been placed in a single field. To separate them, first create a new column to the right of your Lat/Long column, we'll select the column for this field and in the "Data" menu select "Text to Columns". In this tool we'll selected "Delimited", click Next>, set the delimiter to just commas, click Next>, leave the Column data format to "General" and click Finish. If prompted to replace existing data, click yes, as this will replace the combined Lat/Long field into two separate fields.

Geo_ID

The GEO_ID is a unique identifier for all addresses based on census fields for State_Code, County, Tract and Block. Using the =CONCATENATE() function, combine these fields as follows [State_Code][County][Tract][Block]. Do not place any spaces between them, for example-

	G	H	I	J	K
1	[STATE_CODE]	[COUNTY]	[TRACT]	[BLOCK]	[GEO_ID]
2	25	013	812902	2015	250138129022015

The GEO_ID in K2 would be generated by entering the following formula-

=CONCATENATE(G2,H2,I2,J2,K2)

Repeat this for all entries by double clicking the lower-right, or otherwise clicking and dragging this square down the column of other entries

Finding Duplicate Records

If you have the time and patience, sorting and filtering can be useful tools for looking through data for duplicate entries. However if you want to find all duplicate addresses quickly, you can do this with a combination of concatenations and pivot tables.

First we'll make a new "dummy field" in which we'll concatenate our address data so we have the street number, name, subaddress (units, etc.), and town in one column. No spaces between fields will be necessary. Once you've done this, double click on the box in the lower right hand corner to apply the formula to the entire column. If this does not work (which may be the case with larger sets of data), you can also apply this formula to all rows by selecting the column, and then holding CTRL, click on the cell with your concatenation formula; click your cursor to the formula textbox above the table and press CTRL-Enter. After a moment, the spreadsheet should fill the entire column with the formula.

After generating this field, select the entire column and in the Insert menu, select PivotTable. Leave the table/range as default, and choose the PivotTable to be placed on a New Worksheet. In the PivotTable Fields menu, if it's not already populated, click and drag set the one field you generated to the Rows and Values sections, leaving the latter as the default "Count of [FieldName]". After the PivotTable has generated, select both its columns and copy them. Make another worksheet and then right-click to Paste Options>Values (Ctrl-V will not do this). Select the column that displays the count and in the Home menu, click the "Sort & Filter" button and select Sort from Largest to Smallest. If prompted to expand your selection, keep the default option and click Sort. You can filter out those addresses which have a count of 1, and scroll to the top of the pasted PivotTable values and you'll find addresses which have multiple entries.