**Analysis description for “Superfund Cleanups and Children’s Lead Exposure”**

**October 2019**

**Data inputs - non-restricted access data:**

* List of Superfund sites for each study state:
	+ Source: <https://cumulis.epa.gov/supercpad/cursites/srchsites.cfm>
	+ F:\Lead Project\NPL Polygon Tables\Superfund Sites By State with Match (Cumulis Data).xlsx
* Polygon NPL Boundaries (from EPA/OSRTI)
	+ F:\Superfund Data\NationalPrioritiesList.gdb
		- Used files corresponding to EPA regions 1, 4, 5, and 7
		- The National file of Superfund Geospatial Boundaries was created and released in 2014 as part of a Freedom of Information Act (FOIA) Request.  Files from each region containing geospatial layers were combined and loaded to a national schema.  The process of creating this dataset did not include QA/QC of regional data or standardization for inconsistencies.  There has also been no official project to update it with boundaries for sites added to the NPL or to update boundaries which were already identified.
		- Note that the Agency provides geospatial information as a public service and does not vouch for the accuracy, completeness or currency of data. Data provided by external parties is not independently verified by EPA. These data are made available to the public strictly for informational purposes. Data do not represent EPA 's official position, viewpoint or opinion, express or implied. This information is not intended for use in establishing liability or calculating Cost Recovery Statutes of Limitations and cannot be relied upon to create any rights, substantive or procedural, enforceable by any party in litigation with the United States or third parties. EPA reserves the right to change these data at any time without public notice.
* Polygon NPL Boundaries (from ATSDR)
	+ Source: <http://sedac.ciesin.columbia.edu/data/set/superfund-atsdr-hazardous-waste-site-ciesin-mod-v2>
	+ G:\Lead\Shapefile Data\National\Hazardous Waste Sites\ATSDRHazardousWasteSitePolygonDataWithCIESINModificationsVersion2\ATSDRHazardousWasteSitePolygonDataWithCIESINModificationsVersion2.shp
* Polygon NPL Boundaries for Federal Facilities that were missing boundary data in the 2014 FOIA boundary data (from EPA/FFRRO)
	+ Source: Census TIGER/Line Shapefile, 2013, Series Information File for the Military Installation National Shapefile (from 2012 inventory and boundaries of most military installations the U.S. Department of Defense and the U.S. Department of Homeland Security): <https://catalog.data.gov/dataset/tiger-line-shapefile-2013-series-information-file-for-the-military-installation-nati>
		- F:\State Maps\Massachussets\2012\_us\_mil\_MA.shp
		- Source: Department of Defense Military Installations, Ranges, and Training Areas <https://catalog.data.gov/dataset/military-installations-ranges-and-training-areas>
		- F:\State Maps\Massachussets\MIRTA\_MA.shp
* Polygon NPL boundaries – final file used in our analysis (discussed in STEP 1)
	+ G:\Lead\Stata\Multi state\Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.xlsx
* Brownfields points for sites where grantees indicated lead was present
	+ Source: Clean Ups in My Community(CIMC): <https://www.epa.gov/cleanups/cleanups-my-community>
	+ ….CIMC\_export\_06222017.xlsx
	+ F:\State Maps\National\All\_Brownfields\_.shp
* RCRA Corrective Action points for sites where lead was released
	+ Source: An excel file containing XY coordinates was provided to us directly from EPA’s Office of Resource Conservation & Recovery. Points for all RCRA facilities are available in the downloadable GIS module at <https://rcrapublic.epa.gov/rcra-public-export/>.
		- RCRA CA sites in CIMC Dev 07192017.xls
	+ To identify whether lead was released at each RCRA CA facility, data was pulled from EPA’s internal RCRA Info user interface. Publicly available RCRA data is available to download at <https://rcrapublic.epa.gov/rcra-public-export/> and the files below could be compiled using data from the following downloadable modules: Handler, Biennial Report and Corrective Action.
		- …RCRA\_CA\_*state*.dta, where *state*=MA, MI, MO, NC, RI, and WI
		- …LEAD\_CLEANUP\_*state*.dta, where *state*=MA, MI, MO, NC, RI, and WI
* NPL site data from Superfund Enterprise Management System (SEMS) and Agency for Toxic Substances and Disease Control (ATSDR)
	+ SEMS data includes cleanup milestone dates, contaminant types, site categories, and information on removal actions.
	+ G:\Lead\Stata\SEMS\NPL Categories and CC Dates 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\NPL Contaminants 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\NPL P L D Dates 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\NPL Removals 08 25 2017.xlsx
	+ Source: Individual superfund site information from SEMS is available publicly at the links below, however; the batch site date files were pulled from EPA’s internal SEMS database. The SEMS contact at EPA is Jennifer Sutton, Chief, Information Management Branch, OSRTI (Sutton.Jennifer@epa.gov)
		- <https://cumulis.epa.gov/supercpad/cursites/srchsites.cfm>
		- <https://www.epa.gov/enviro/sems-overview>
	+ ATSDR data provides information on NPL sites where lead is present and which may have a completed exposure pathway for lead
	+ G:\Lead\Excel Spreadsheets\ATSDR\_Data.xlsx
		- Source: Annually, ATSDR publishes a Priority List of Hazardous Substances (also known as the Substance Priority List) and a Completed Exposure Pathway (CEP) Site Count Report. These reports and some supporting data can be downloaded from here: <https://www.atsdr.cdc.gov/spl/resources/>. The data file above was obtained from ATSDR. The contact at ATSDR is Mike Fay (rmf4@cdc.gov) or you can fill out a general inquiry form here: <https://wwwn.cdc.gov/dcs/ContactUs/Form?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcdc-info%2Frequestform.html>
* Ambient air concentration data
	+ F:\State Maps\National\NATA\LEAD COMPOUNDS 1999.xlsx
		- Source: <https://archive.epa.gov/airtoxics/nata1999/web/html/tables.html>
		- Scroll to “Pollutant-Specific Database,” select “Lead Compounds” from Group 2 drop-down menu, convert downloaded Access database to Excel
	+ F:\State Maps\National\NATA\LEAD COMPOUNDS 2002.xlsx
		- Source: <https://archive.epa.gov/nata2002/web/html/tables.html>
		- Scroll to “2002 Tract-Level Modeled Ambient Concentrations, Exposures, and Risks,” select “Lead Compounds” from the drop-down menu, convert downloaded Access database to Excel
	+ F:\State Maps\National\NATA\LEAD COMPOUNDS 2005.xlsx
		- Source: <https://www.epa.gov/national-air-toxics-assessment/2005-nata-assessment-results>
		- Scroll to “Tract Level Modeled Ambient Concentrations,” select “Lead Compounds” from the drop-down menu, convert downloaded Access database to Excel
	+ F:\State Maps\National\NATA\LEAD COMPOUNDS 2011 clipped.xlsx
		- Source: <https://www.epa.gov/national-air-toxics-assessment/2011-nata-assessment-results>
		- Scroll to “2011 Modeled Ambient Concentrations, Exposures and Risks,” “Lead Compounds” from Pollutant Specific Results” drop-down menu, convert downloaded Access database to Excel
* Temperature data
	+ Source: https://www.esrl.noaa.gov/psd/data/usclimate/tmp.state.19712000.climo
	+ G:\Lead\Stata\Multi state\State\_Weather.dta
* Road and VMT data:
	+ Source: National Highway Planning Network (NHPN) spatial data from the National Transportation Database: <http://osav-usdot.opendata.arcgis.com/datasets/0ad03e8a10e9445f8d2f14f0955b18cb_0>
	+ F:\State Maps\National\TRANS\National\_Highway\_Planning\_Network.shp
	+ Source: Vehicle Miles Traveled per Lane-Mile from the Bureau of Transportation Statistics: <https://www.bts.gov/content/roadway-vehicle-miles-traveled-vmt-and-vmt-lane-mile-functional-class>
		- C:\XXXXXXXX\Lead\Superfund\BTS\_VTM\_per\_lane\_mile.xlsx
	+ Final file used in analysis: G:\Lead\Stata\Multi state\Natl\_VMT\_Density\_Updated.xlsx
* National Health and Nutrition Examination Survey (NHANES) data on BLL for children age 1-5
	+ Source: <https://wwwn.cdc.gov/nchs/nhanes/Default.aspx>
	+ C:/XXXXXXXX/Lead/Superfund study/July2018.xlsx
* NPL site data from Superfund Enterprise Management System (SEMS) to identify sites used in the benefit transfer:
	+ - Source: Individual superfund site information from SEMS is available publicly at the links below, however; the batch site data files were pulled from EPA’s internal SEMS database. The SEMS contact at EPA is Jennifer Sutton, Chief, Information Management Branch, OSRTI (Sutton.Jennifer@epa.gov)
			* <https://cumulis.epa.gov/supercpad/cursites/srchsites.cfm>
			* <https://www.epa.gov/enviro/sems-overview>
	+ …SEMS\_contaminants\_all.dta
	+ …SEMS\_PLDdates\_all.dta
	+ …SEMS\_categories\_CCdates\_all.dta
* NPL expenditure data from EPA:
	+ Source: Data was obtained from EPA’s Office of Superfund Remediation & Technology Innovation and was pulled from internal databases (Integrated Financial Management System (IFMS) and Compass Financials (Compass).
		- …ExpData\_NPLs\_Lead\_Nominal.xls

**Data inputs - restricted access data:**

* State blood lead screening data received from Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin public health departments:
	+ Michigan: 14 files with the following naming convention: Karen\_EPA\_1of14.csv, Karen\_EPA\_2of14.csv,… Karen\_EPA\_14of14.csv
	+ Missouri: MOdata1.csv, MOdata2.csv
	+ North Carolina: 23 files with the following naming convention: 1992NCDataKlemickEPA.csv, 1993NCDataKlemickEPA.csv,… 2015NCDataKlemickEPA.csv
	+ Rhode Island: qryDrEpaSuperFundBloodTestByAddressFlagLeadcareWithClient.txt
	+ Wisconsin: EPA\_Superfund\_14JUN2017.txt
* Data received from Massachusetts Dept of Public Health
	+ Lead\_2000-2016\_Geo\_Final.txt – individual blood lead data
	+ NEAR\_Tables.xlsx – distances from children in blood lead data to NPL, RCRA, and Brownfields sites
		- Shapefiles for NPL, RCRA, and Brownfields sites provided to MDPH by EPA: Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp**,** All\_Brownfields\_.shp, Updated\_Natl\_RCRA\_Sites\_W\_Lead\_Status.shp
	+ CLPPP\_CTs.xlxs – Census tract IDs for children in blood lead data
	+ 2001-2002.xls – blood lead screening rates by census tract for these years
	+ 2003-2016.xls – blood lead screening rates by census tract for these years
* Individual housing characteristics (for each state except Massachusetts) – source: ZTRAX assessment data
	+ Building.txt
	+ Value.txt
	+ Main.txt
* Geolytics Neighborhood Change Database (Census tract characteristics 1990-2010)
	+ The NCDB provides multi-decadal Census data in terms of 2010 Census tract boundaries, making it easier to track demographic changes over time for a given location.
	+ Source: [http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970-2000,Data,Features,Products.asp](http://www.geolytics.com/USCensus%2CNeighborhood-Change-Database-1970-2000%2CData%2CFeatures%2CProducts.asp)
* Alternately, the Census tract characteristics for 1990-2010 are publicly available from American Fact Finder: <https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml>
* Geolytics Demographic Data (2009-2013 American Community Survey 5-year Estimates)
	+ Source: [http://www.geolytics.com/USCensus,Census2010ACS,Categories.asp](http://www.geolytics.com/USCensus%2CCensus2010ACS%2CCategories.asp)
	+ Alternatively, the 2009-2013 American Community Survey 5-year estimates are publicly available from American Fact Finder: <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> or directly from the Census FTP Server here <https://www.census.gov/data/developers/data-sets/acs-5year.2013.html>
* ENR Construction Cost Index:
	+ Source: <https://www.enr.com/economics>
	+ A subscription to Engineering News-Record (ENR) is required to access the index prior to 1989. Subscribe and then contact them directly.

**Contacts to request restricted-access data:**

* Massachusetts: Robert Knorr, Director, Environmental Epidemiology Program, Massachusetts Department of public health, robert.knorr@state.ma.us, [www.mass.gov/dph/research](http://www.mass.gov/dph/research)
* Michigan: Dan Albright, Michigan Department of Health & Human Services, AlbrightD@michigan.gov
* Missouri: Scott Patterson, Missouri Department of Health & Senior Services, Scott.Patterson@health.mo.gov
* North Carolina: Tena Hand, tena.hand@dhhs.nc.gov, Kim Gaetz, kim.gaetz@dhhs.nc.gov, North Carolina Department of Health & Human Services
* Rhode Island: Margaret Gradie, Margaret.Gradie@health.ri.gov, Anne Cardozo, Anne.Cardoza@health.ri.gov, Rhode Island Department of Health
* Wisconsin: Marjorie Coons, Wisconsin Division of Public Health, Marjorie.Coons@dhs.wisconsin.gov
* Zillow ZTRAX data: ZTRAX@zillowgroup.com , <https://www.zillow.com/research/ztrax/>
* Geolytics: <http://www.geolytics.com/>
* NPL Expenditure Data at EPA: Amy Vandenburg, vandenburg.amy@epa.gov, Office of Superfund Remediation & Technology Innovation, Resources Management Division
* SEMS Data at EPA: Jennifer Sutton, Sutton.Jennifer@epa.gov, Chief, Information Management Branch, Office of Superfund Remediation & Technology Innovation
* ENR Construction Cost Index: Scott Lewis, LewisW@enr.com, <https://www.enr.com/economics>

**Analysis steps**

STEP 1 (in ESRI ArcMap): Create master NPL polygon file

* Note: In all ESRI ArcMap steps, data was converted to a GCS NA 1983 projection after being imported
* Downloaded cartographic boundary shapefiles of each state in our study area from the Census Bureau’s TIGER database: <https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html>
* Created a 5 km buffer surrounding all 6 states in our study, which was used to clip national NPL site polygons
* Imported NationalPrioritiesList.gdb
* Clipped these polygons using the 5 km buffer around the state boundaries
* Conducted a search for all NPL sites in six study states on the Cumulis website (<https://cumulis.epa.gov/supercpad/cursites/srchsites.cfm>)
	+ Specified the state, and then selecting options from the “NPL Status” tab, selected both “Currently on the Final NPL” or “Deleted from the Final NPL”
	+ Saved as “Superfund Sites By State with Match (Cumulis Data).xlsx”
* Identified sites in “Superfund Sites By State with Match (Cumulis Data).xlsx” that did not appear in NationalPrioritiesList.gdb. For those sites we:
	+ Identified each site’s lat/long or street address information (whichever was available from each site’s Cumulis profile) and exported that information into ArcMap using sheets named “[State] Import” in “Superfund Sites By State with Match (Cumulis Data).xlsx”
	+ Plotted tables using lat/long or street address (depending on availability), and combined the two to form one point shapefile.
* To get site boundaries for the sites missing from NationalPrioritiesList.gdb, we used polygons from the following files:
	+ Federal facility site boundaries: 2012\_us\_mil\_MA.shp and MIRTA\_MA.shp
	+ ATSDR polygon data: ATSDRHazardousWasteSitePolygonDataWithCIESINModificationsVersion2.shp
	+ No polygon data were available for 2 NPL sites in Michigan (Veliscol Burn Pits and Ten-Mile Drain - MIN000510389 & MIN000510063). We drew these by hand using the polygon drawing tool in the editor toolbar in ArcMap. We used site maps and documents from the site’s Cumulis profile in combination with satellite base maps in ArcMap to draw the polygons accurately.
* We combined all site polygons together into a single shapefile that contains the full set of sites in our study (Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp). We originally hoped to include Illinois, Iowa, and New Jersey in our analysis, so this file contains polygons for those states. Sites in Illinois and Iowa within 5 km of Wisconsin were included in our analysis.
* Outputs (used in STEPS 4, 8, and 10):
	+ Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp
	+ Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.xlsx(attribute table exported to excel)
* For North Carolina only, we generated a list of zip codes intersecting our study buffer for use in ZTRAX data cleaning in STEP 6.

STEP 2 (in ESRI ArcMap): Create Brownfields and RCRA points files

* The Brownfield data were obtained from the ACRES database. ACRES (<https://www.epa.gov/brownfields/brownfields-grantee-reporting-using-assessment-cleanup-and-redevelopment-exchange-system>) is an online database for Brownfields Grantees to electronically submit data to EPA. Brownfields location data from ACRES can be obtained in a variety of ways:
	+ Clean Ups in My Community(CIMC): <https://www.epa.gov/cleanups/cleanups-my-community>
		- Choose “create a listing of cleanups or grants...”
		- Select “Brownfields Properties”
		- Scroll to the bottom of the page and click “Click for More Brownfields Info”
		- Check the all the categories and you will get an excel spreadsheet that includes lat/long locations along with other property information from ACRES such as contaminants found and cleaned up
		- CIMC\_export\_06222017.xlsx
	+ Data were collapsed by ACRES ID. We only used sites with a “Y” designating a yes for either “Containment Found Lead” or “Containment Cleaned Up Lead”
* Plotted the sites by their XY coordinates for each state.
* Merged all state layers to create a single Brownfields layer that also includes sites in other states bordering the six study states
	+ Output: F:\State Maps\National\All\_Brownfields\_.shp (used in STEP 8)
* RCRA Corrective Action siteinformation was obtained from RCRA Info, EPA’s system for tracking the regulated universe of RCRA hazardou**s** waste handlers
	+ To determine which RCRA CA sites had a release/cleanup of lead each states’ files
	(LEAD\_CLEANUP\_*state* and RCRA\_CA\_*state*) were processed with STATA do-file to create the files RCRA\_CA\_LEAD\_state, where state= *state*=MA, MI, MO, NC, RI, and WI. The pdfs provide definitions of codes in the data and excel file identifies the lead codes we included.
		- …RCRA\_CA\_Lead.do
		- …ca\_event\_code\_definitions.pdf
		- …source\_code\_definitions.pdf
		- …waste\_code\_definitions.pdf
		- …waste\_codes\_lead.xlxs
	+ We only mapped sites where lead was present
	+ The file RCRA CA sites in CIMC Dev 07192017.xls was imported, and RCRA CA sites were plotted by their XY coordinates
	+ Once this was done for all states, we merged each state’s layer together to create a single RCRA CA site layer
	+ Output: F:\State Maps\National\RCRA\ Updated\_Natl\_RCRA\_Sites\_W\_Lead\_Status.shp (used in STEP 8)

STEP 3 (in ESRI ArcMap): Calculate 1980 traffic density for each Census tract

* Imported the NHPN shapefile (F:\State Maps\National\TRANS\National\_Highway\_Planning\_Network.shp)
* Clipped the NHPN shapefile using the TIGER CT boundary file for each study state
* Using the FCLASS variable in the NHPN attribute table, exported the data into several different shapefiles based on road type.
* Joinedeach of those individual shapefiles to the Census tracts shapefile, summing the total length of each respective type of road within the tracts
* Did this by running a spatial join with the following settings:
	+ Target Feature - Census Tracts
	+ Join Features - Road shapefile
	+ Join Operation - One to One
* Deleted every field being imported from the road shapefile except the Shape\_Leng (measured in meters)
* Selected the first bullet point in query 2: MERGE RULE - SUM
* Renamed Shape\_Leng after each merge to reflect the type of road (such as freeway, etc.)
* After every join, we used the field calculator to multiply the road length totals for each road class we generated by the 1980 column from the file BTS\_VMT\_per\_land\_mile.xlsx to get an estimation of the total traffic along the road in 1980
* Using the field calculator again, we summed the VMT estimate from each road class to get total 1980 traffic estimations by Census tract
* We then divided by the total land area of the Census tract (in square meters) from the original TIGER census tract shapefile download to get 1980 VMT density
* Output: G:\Lead\Stata\Multi state\Natl\_VMT\_Density\_Updated.xlsx (used in STEP 12)

STEP 4 (in Stata 14): Merge NPL data sources to create single dataset with NPL site characteristics

* Run Stata do-file “SEMS data AllStates Nov2018.do”
* Data needed:
	+ G:\Lead\Stata\SEMS\NPL Categories and CC Dates 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\NPL Contaminants 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\NPL P L D Dates 08 25 2017.xlsx
	+ G:\Lead\Stata\SEMS\\NPL Removals 08 25 2017.xlsx
	+ G:\Lead\Excel Spreadsheets\ATSDR\_Data.xlsx
* Uses the attribute table from NPL polygon file created in STEP 1 (G:\Lead\Stata\Multi state\Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.xlsx)
* Performs the following tasks:
	+ Imports the SEMS and ATSDR data. Creates variables identifying sites where lead is present and where there is at least one completed exposure pathway for lead.
	+ The four separate SEMS datasets are cleaned and collapsed by site and then merged.
	+ Drops sites from our analysis that were not listed prior to 2015, sites that never changed status from proposed, a site that was cleaned up prior to the study period, and two sites within the boundaries of other sites where lead was not listed as a contaminant of concern. One of the sites in our study was also listed, deleted, and then re-listed. We adjust the dates of those milestones accordingly.
	+ Cleans the NPL polygon attribute table and merges with the SEMS/ATSDR data
* Output: SEMS\_merge\_NPLlist\_Nov2018.dta (used in STEP 12)

STEP 5 (in Stata 14): Process NATA data on lead air concentration for merging with blood lead data

* Run Stata do-file “NATA.do”
* Data needed:
	+ LEAD\_COMPOUNDS 2011 clipped.xlsx
	+ LEAD\_COMPOUNDS 2005.xlsx
	+ LEAD\_COMPOUNDS 2002.xlsx
	+ LEAD\_COMPOUNDS 1999.xlsx
* Performs the following tasks:
	+ Imports and cleans each year of NATA data
	+ Merges the data together using census tract IDs
	+ Saves/exports the data specific to each state for 1999-2005, and then drops the 2011 and saves that dataset as XX\_2000\_Nata .dta files (based on 2000 Census tract ID, “XX” represents each state’s 2-letter abbreviation)
	+ Saves/exports the data specific to each state for 2011 (based on 2010 Census tract ID)
* Outputs: XX\_2000\_Nata.dta, XX\_NATA.dta (used in STEP 12)

STEP 6 (in Stata 14): Process the ZTRAX data for merging with blood lead data in Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin

* Run the Stata do-file “Zillow\_Data\_Import\_AllStates.do”
* Data needed for each state (from ZTRAX):
	+ Building.txt
	+ Value.txt
	+ Main.txt
* Performs the following tasks:
	+ For Michigan, Rhode Island, North Carolina, Missouri, and Wisconsin:
		- Imports data, keeps the needed variables, and generates a dummy variable indicating a lack of listed address
		- Basic cleaning, dropping observations we don’t need (Address= vacant lot, etc.)
		- Generates a cleaned version of the Main.txt Zillow data for each state (containing only an ID variable and spatial variables) and exports for plotting in ArcMap
			* Separate output for each state: Current\_Assesor\_Data\_GIS\_Ready\_LatLong\_RowID\_Sort.txt (used in STEP 9)
		- Merge the three datasets together, keeping relevant variables
		- Standardizing capitalizations in address variables
		- Collapsing data to be unique by full address (street address, city, and zip)
	+ For North Carolina only:
		- North Carolina has a unique problem—there are a large number of duplicate observations by street address/city/zip, which inflates the size of the dataset to over 16 GB, making it slow to run in Stata and ArcMap.
		- To mitigate this issue, we dropped ZTRAX data in all zip codes not intersected by our study buffer (using list of zip codes generated in STEP 1).
* Outputs: XX\_Main\_Building\_Value\_Collapsed.dta (used in STEP 12)

STEP 7 (in Stata 14): Clean the state blood lead screening data from Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin for mapping in ArcMap and further analysis

* Run the following Stata do-files:
	+ Appending MI\_Data\_HK.do
	+ NC\_append\_Dec2017.do
	+ Michigan\_datacleaning\_Oct2017.do
	+ RhodeIsland\_datacleaning\_Oct2017.do
	+ Missouri\_datacleaning\_Jan2018.do
	+ Wisconsin\_datacleaning\_Oct2017.do
	+ NC\_datacleaning\_Dec2017.do
* Data needed from state public health departments:
	+ Michigan: 14 files with the following naming convention: Karen\_EPA\_1of14.csv, Karen\_EPA\_2of14.csv,… Karen\_EPA\_14of14.csv
	+ Missouri: MOdata1.csv, MOdata2.csv
	+ North Carolina: 23 files with the following naming convention: 1992NCDataKlemickEPA.csv, 1993NCDataKlemickEPA.csv,… 2015NCDataKlemickEPA.csv
	+ Rhode Island: qryDrEpaSuperFundBloodTestByAddressFlagLeadcareWithClient.txt
	+ Wisconsin: EPA\_Superfund\_14JUN2017.txt
* Performs the following tasks:
	+ First, two states (Michigan and North Carolina) transferred their data to us in multiple files, so we aggregated the data into one file per state using Appending\_MI\_Data\_HK.do and NC\_append\_Dec2017.doc, respectively.
	+ Then, each do-file named [state]\_datacleaning\_[date].do cleans the blood lead data, including standardizing decimal places, creating testing year and month variables, eliminating data points we can’t use (children outside the age range, addresses that aren’t representative of location like PO boxes), etc.
	+ When such a variable is not already provided by the state, we generate child ID variables that identify unique combinations of factors unlikely to be shared by multiple children. These variables can be seen in the “egen id\_group” commands
		- Outputs for each state: XX\_Cleaned.dta (used in STEP 12)
	+ Then we create files for Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin for use in GIS analysis. We keep only the variables necessary to plot children’s locations and drop duplicates to get a dataset with one observation per unique address.
	+ Exports cleaned address data as csv files for plotting in GIS. In cases where there were too many observations to geocode in ArcMap at once we break down the data into manageable chunks (N=~600,000)
* Outputs (used in STEP 8):
	+ Michigan Sorted\_export1.csv
	+ Michigan Sorted\_export2.csv
	+ MO\_UAs\_Sorted\_Total.csv
	+ NC\_UAs\_Sorted\_1.csv
	+ NC\_UAs\_Sorted\_2.csv
	+ NC\_UAs\_Sorted\_3.csv
	+ Wisconsin\_export.csv
	+ Unique\_Address\_Data.csv (Rhode Island)

STEP 8 (in ESRI ArcMap): Map unique addresses from Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin blood lead data, determine the correct Census tract ID, and calculate distances to contaminated sites

* Imported unique addresses from blood lead data in each state into ArcMap (from STEP 7):
	+ Michigan Sorted\_export1.csv
	+ Michigan Sorted\_export2.csv
	+ MO\_UAs\_Sorted\_Total.csv
	+ NC\_UAs\_Sorted\_1.csv
	+ NC\_UAs\_Sorted\_2.csv
	+ NC\_UAs\_Sorted\_3.csv
	+ Wisconsin\_export.csv
	+ Unique\_Address\_Data.csv
* Geocoded address fields using the OEI\_IGD/Navteq\_USA Address Locator
	+ The address locator uses four fields to plot address: address, city, state, and zip
	+ For states with more than one .csv import, we geocoded them individually. We then used the merge tool to create a single file per state.
* Created a buffer 10 km around the NPL boundary shapefile created in STEP 1 (Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp) using a Geodesic method with no dissolve (Side Type = FULL / End Type = ROUND)
* Clipped the full geocoded set of blood lead test unique addresses in each state by the NPL distance buffer file to identify the set of addresses included in our analysis.
* Did a spatial join with TIGER census tract boundaries from both 2010 and 2000 to assign the relevant FIPS codes to all blood test unique addresses within study area, and exported separate attribute tables for each state
	+ Outputs: XX\_Children\_with\_FID.txt (XX =state abbreviation) (used in STEP 12)
* Conducted a state-by-state evaluation of the geocoding:
	+ The “score” indicates the degree to which the address in the data was a perfect match to the address it was assigned spatially, with 100 being a perfect match. For each state, we determined the acceptable score by comparing geocoded addresses in ArcMap to those same addresses in google maps. We checked multiple addresses (roughly 25 per state) until we identified, for each state, the cutoff at which the addresses in ArcMap differed by more than a few meters from its counterpart in Google maps. We used score thresholds of 87.99-96.15, depending on the state. All addresses with scores below the cutoff are dropped in STEP 12.
* Used the “Generate Near Table” tool to create tables with the distances of each unique address in the study to each NPL boundary within 10 km
	+ Input feature - BLL unique addresses <= 10 km from NPL boundaries
	+ Near feature - Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp
	+ Search radius - 10 km
	+ Uncheck “Find only closest feature”
	+ Method - Geodesic
	+ Outputs: XX\_NPL\_Near\_Table.txt (used in STEP 12)
* Used the “Generate Near Table” tool and same process above to create tables with the distances of each unique address to each Brownfield site within 2 km
	+ Near feature - All\_Brownfields\_.shp (from STEP 2)
	+ Outputs: XX\_Brownfield\_Near\_Table.txt (used in STEP 12)
* Used the “Generate Near Table” tool and same process above to create tables with the distances of each unique address to each RCRA Corrective Action site within 2 km
	+ Near feature - Updated\_Natl\_RCRA\_Sites\_W\_Lead\_Status.shp (from STEP 2)
	+ Outputs: XX\_RCRA\_Near\_Table.txt (used in STEP 12)

STEP 9 (in ESRI ArcMap): Map ZTRAX assessment data and do a spatial join with state blood lead data for Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin

* Imported into ArcMap "Current\_Assesor\_Data\_GIS\_Ready\_LatLong\_RowID\_Sort.txt" files for each state (from STEP 6)
* Converted tables into a dBASE format (which helps avoid plotting errors), plotted Zillow houses by XY coordinates, and exported the resulting “events” file into a shapefile of all Zillow points in the state
* Conducted a NEAR analysis on the unique addresses from each state’s blood lead data (from STEP 8), with the near feature being the shapefile created from the ZTRAX data, and exported attribute tables
	+ This adds the closest ZTRAX point’s FID and relative distance to the attribute table of the blood lead test address data, as long as the Zillow point was within 100 meters
* Outputs: XX\_Zillow\_Data\_With\_FID.txt, XX\_Zillow\_Near\_Data.txt (used in STEP 12)

STEP 10 (in ESRI ArcMap): GIS mapping to allow for calculating blood lead screening rate by Census tract for Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin

* Ran an intersect (under the Analysis section of the ArcMap toolbox) between the 2010 Census tract shapefile and a 10 km buffer around Final\_NPLs\_With\_ATSDR\_and\_Missing\_MI\_Sites.shp
* The resulting shapefile was then joined to the TIGER Census tract boundary shapefile for the state
* The census tracts intersecting the 10 km buffer were exported to their own shapefile
* This shapefile was used to clip the blood lead unique address data plotted in STEP 8
* Conducted a spatial join between the blood lead unique address data and the Census tract shapefile to get each child’s 2010 Census tract ID and exported the attribute tables
* Outputs: (used in STEP 11)
	+ NC\_Screening\_Rate\_Children\_in\_CTs.txt
	+ RI\_Screening\_Rate\_Children\_in\_CTs.txt
	+ WI\_Screening\_Rate\_Children\_in\_CTs.txt
	+ New\_MO\_Screening\_Rate\_Children\_in\_CTs.txt
	+ Updated\_MI\_Screening\_Rate\_Children\_in\_CTs.txt

STEP 11 (in Stata 14): Calculate blood lead screening rate by Census tract for Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin

* Run the following Stata do-files:
	+ Screening\_Rate\_MI.do
	+ Screening\_rate\_RI.do
	+ Screening\_Rate\_MO.do
	+ Screening\_Rate\_WI\_Updated.do
	+ Screening\_Rate\_NC.do
* Data needed (from STEP 10):
	+ NC\_Screening\_Rate\_Children\_in\_CTs.txt
	+ RI\_Screening\_Rate\_Children\_in\_CTs.txt
	+ WI\_Screening\_Rate\_Children\_in\_CTs.txt
	+ New\_MO\_Screening\_Rate\_Children\_in\_CTs.txt
	+ Updated\_MI\_Screening\_Rate\_Children\_in\_CTs.txt
* Data needed (from Geolytics NCDB): Geolytics\_Total\_Children\_0\_4.xlsx
* Performs the following tasks:
	+ Imports the data
	+ Cleans the variable names that were affected by the export from GIS and drops observations that did not accurately plot to a specific street address or point (the address locator in ArcMap assigned them a “loc\_name” of Postal, StreetName, or AdminPlaces; loc\_name further explained in STEP 12)
	+ Drop observations of children above age 0-4 and repeat tests of the same child in the same year
	+ Merge with Geolytics Census data on child population age 0-4
	+ Estimate screening rates by Census tract in each state for separate multi-year periods
		- Determine the number of children screened in each tract during the time period
		- Divide by the number of years in the time period, multiplied by the estimated population of children aged 0-4 in each year. That estimated yearly number is generated by interpolating the populations in 2000 and 2010 Census years
		- Merge the different time periods together using Census tract ID, giving a single file for each state Census with tract screening rates over time.
* Outputs: XX\_Screening\_Rates.dta (used in STEP 12)

STEP 12 (in Stata 14): Merge different datasets generated in Stata and ArcMap together for each state

* Run the following Stata do-files:
	+ MA\_Merging.do
	+ Michigan\_Merge\_Zillow\_Apr2018.do
	+ Missouri\_Merge\_Zillow\_Apr2018.do
	+ NC\_merge\_Zillow\_Apr2018.do
	+ RI\_merge\_zillow\_Apr2018.do
	+ Wisconsin\_merge\_zillow\_Apr2018.do
* Performs the following tasks for each state:
	+ For Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin, imports the following files (from STEP 8):
		- XX\_Children\_with\_FID.txt
		- XX\_NPL\_Near\_Table.txt
		- XX\_Brownfield\_Near\_Table.txt
		- XX\_RCRA\_Near\_Table.txt
	+ Data are renamed, collapsed, merged
	+ Drop addresses from dataset with poor spatial accuracy:
		- Drop if “loc\_name” (which describes how the composite address locator placed certain addresses on the map) was one of the following:
			* Postal (indicates that the locator matched the address only to a postal code and assigned a point within that code)
			* Admin\_Places (indicates the point was assigned within a general administrative boundary, such as a county or city)
			* StreetName (Indicates the point was assigned to the correct street, but along an interpolated point, not an identified house number)
		- Drop if score is less than state-specific numerical cutoff (determined in STEP 8)
	+ Merge GIS data with cleaned blood lead test data (XX\_Cleaned.dta from STEP 7)
	+ For Massachusetts, import and merge data received from MDPH:
		- Lead\_2000-2016\_Geo\_Final.txt
		- NEAR\_Tables.xlsx
		- CLPPP\_CTs.xlxs
		- 2001-2002.xls
		- 2003-2016.xls
	+ Import and merge in data with Geolytics Census data, VMT density, NATA, and screening rate data by Census tract IDs:
		- From Geolytics NCDB:
			* C:\Users\XXXXXXXX\Documents\Blood lead data\Massachusetts\Geolytics\_Data.csv
			* C:\Users\XXXXXXXX\Documents\Blood lead data\Michigan\Geolytics\_Data.csv
			* G:\Lead\Stata\Missouri\Geolytics\_Data.csv
			* G:\Lead\Stata\ North Carolina\Full\_Database\_Work\Geolytics\_Data.csv
			* C:\Users\XXXXXXXX\Documents\Blood lead data\Rhode Island\Geolytics\_Data.csv
			* C:\Users\XXXXXXXX\Documents\Blood lead data\Wisconsin\WI\_Geolytics\_Data.csv
		- From STEP 3: G:\Lead\Stata\Multi state\Natl\_VMT\_Density\_Updated.xlsx
		- From STEP 5: XX\_2000\_Nata.dta, XX\_NATA.dta
	+ Merge in state monthly weather data from NOAA: State\_Weather.dta
	+ Merge in NPL site data from STEP 4: SEMS\_merge\_NPLlist\_Nov2018.dta
	+ Drop observations not used in analysis (children less than six months old, those located near sites with no status change during study period)
	+ Calculate the average number of tests per child
	+ Drop extraneous variables
	+ Output for Massachusetts: MA\_cleaned\_SH.dta (used in STEP 13)
	+ For Michigan, Missouri, North Carolina, Rhode Island, and Wisconsin, merge in Zillow data on year built using two approaches. The first is an exact match on address; the second is based on the spatial join in ArcMap in STEP 9.
		- First approach: Merge XX\_Main\_Building\_Value\_Collapsed.dta based on exact address match (from STEP 6)
		- Second approach: Merge XX\_Zillow\_Near\_Data.txt and XX\_Zillow\_Data\_With\_FID.txt based on GIS location (from STEP 9)
		- For an observation that matches based on both approaches, we give precedence to the first approach.
	+ Outputs: XX\_cleaned\_zillow\_SH.dta (used in STEP 13)

STEP 13 (in Stata 14): Prepare data from six states for pooled regression analysis

* Run the following Stata do-files:
	+ Massachusetts\_analysis\_Dec2018.do
	+ Michigan\_ analysis\_Dec2018.do
	+ Missouri\_ analysis\_Dec2018.do
	+ NorthCarolina\_ analysis\_Dec2018.do
	+ RI\_ analysis\_Dec2018.do
	+ Wisconsin\_ analysis\_Dec2018.do
* Performs the following tasks:
	+ Create variables necessary for regression analysis
	+ Perform coarsened exact matching
	+ Drop extraneous variables, save data for use in pooled regression analysis:
* Outputs: XX\_pooled\_5km\_2018.dta (used in STEP 14)

STEP 14 (in Stata 14): Pooled regression analysis of data from six states and additional analyses

* Run Stata do-file: pooled\_analysis\_Nov2018.do
	+ Append data files from six states: XX\_pooled\_5km\_2018.dta (from STEP 13)
	+ Generate fixed effects and additional variables needed for regressions
	+ Calculate summary statistics
	+ Run regressions
	+ Export regression results to excel
	+ Calculate imbalance statistics comparing full and matched samples
	+ Create figure graphing EBLL over time among children in treatment and control groups
* Run Stata do-file: pooled\_analysis\_figure4\_probEBLL.do
	+ Create figure graphing probability of EBLL at lead-contaminated and non-lead Superfund sites before and after “listing” and “construction complete” cleanup milestones
* Run Stata do-file: pooled\_buffer\_analysis\_Nov2018.do
	+ Run regressions using alternative distance cutoffs besides 2 km to define “close” to a Superfund site

STEP 15 (in Excel and Stata): Estimate national Superfund benefits

* Use Crump et al.’s (2013) estimated relationship between BLL and IQ: *IQ = -3.315\*ln(BLL+1)*
* Use data from the National Health and Nutrition Examination Survey (NHANES) on geometric mean (GM) BLL and percent BLL > 3.3 μg/dL from 1988-2016 for children age 1-5 (NHANES benefit transfer data July2018.xlsx)
	+ In Stata, regress GM BLL on the natural log of percent of BLL above 3.3 μg/dL. We estimated that *GM BLL = -0.054 + 0.753\*ln(EBLL)* (N = 11, R2 = 0.88).
* Determine the number of Superfund sites nationwide where lead was classified as a contaminant of health concern and that had reached construction complete as of September 31, 2017. Run Stata do-file: GenerateBenTransferNPL\_08272018.do; using Stata dta-files.:
	+ - …SEMS\_contaminants\_all.dta
		- …SEMS\_PLDdates\_all.dta
		- …SEMS\_categories\_CCdates\_all.dta
* Estimate number of children age 0-4 living within 2 km of these sites using Superfund site boundary (or point data if boundary data was not available) as of the end of fiscal year 2016 and block group population data from the 2011-2015 American Community Survey. Child counts surrounding the Superfund sites where lead was a contaminant of health concern were calculated by identifying the Census block group centroids that fell within within 2 km of site boundaries (or modeled circular boundaries when only point site location data was available).
* In Excel, calculate change in IQ and lifetime earnings from the change in BLL caused by Superfund cleanup: OneDrive/Lead Superfund Project/Benefits Transfer/Benefits\_Transfer\_10\_15\_2018.xlsx
	+ Use estimate from Salkever (1995) that a 1-point increase in IQ is associated with a 2.38% gain in lifetime earnings on average across males and females.
	+ Use U.S. EPA (2018a) estimate that a 2.38% gain in lifetime earnings is equivalent to $19,359 at a 3% discount rate for a child age 3 in 2016.
* Calculate EPA expenditures at sites where lead was classified as a contaminant of health concern and that reached construction complete as of September 31, 2017.
	+ Data source: EPA’s Office of Superfund Remediation & Technology Innovation data pull from internal databases (Integrated Financial Management System (IFMS) and Compass Financials (Compass).
		- …ExpData\_NPLs\_Lead\_Nominal.xls
		- Expenditures from 1980-2017 were adjusted for inflation using the ENR Construction Cost Index (CCI) (ENR 2018). We used the monthly CCIs to calculate average annual fiscal year CCI and then each fiscal year deflator. All sites did not have expenditure in all 38 years; however, the average total annual expenditure at these 581 sites was calculate dividing the total expenditure at all sites from 1980-2017 by 581 sites, and then dividing by 38 years. An annual average based on the number of years of actual expenditures at each specific site could not be calculated because expenditure data from 1980-1989 were only available as a total for the 10-year period. To adjust expenditure data from 1980-1989, we used the average annual CCI over the 10-year period.