Data Cleaning and Construction of the Final Dataset

Silencing the Rails: A Study of the Noise-Safety Trade-off in Railway Quiet Zones

Emtiaz Hossain Hritan

Cleaning of Highway-rail Crossing Accidents Data: The data cleaning process for this study required integrating three datasets to examine the effects of quiet zones on railway crossing accidents from 1994 to 2023. The first dataset, obtained from the Federal Railroad Administration (FRA) under a Freedom of Information Act (FOIA) request, provided details on quiet zones, including crossing status, establishment dates, and geographic coordinates. After standardizing variable names, this data was saved as quietzone.dta. Next, I imported accident data from the FRA's Office of Railroad Safety, which records daily occurrences of highway-rail grade crossing accidents with variables such as vehicle speed, weather conditions, and user demographics. Using a unique crossing identifier, I merged the accident data with quiet zone records, adding a quiet zone indicator for each crossing.

To incorporate geographic context, I processed the FRA railroad crossing inventory dataset, which includes latitude, longitude, and location-specific identifiers. After selecting relevant fields, I merged this inventory data with the combined accident and quiet zone data, assigning each crossing to its respective county and state. Accident dates were then reformatted to Stata date format, from which year and month were extracted for annual aggregation. Non-essential variables, such as those related to hazardous materials, crossing warnings, and maintenance details, were removed to streamline the final dataset.

Cleaning of Demographic Dataset: The data cleaning process for the Census Bureau's intercensal data from 1990 to 2019 followed several essential steps to prepare and standardize the dataset for analysis. First, raw data files for each period (1990–2000, 2000–2009, and 2010–2019) were imported from CSV and Excel formats and converted to Stata-compatible .dta files. A unique county identifier, $county_fips$, was generated by merging state and county codes. Year values were adjusted to reflect actual calendar years, and interim population estimates were removed to retain a single observation per year. For age-specific data, descriptive labels were applied, and the dataset was reshaped to create individual columns for each age range, while retaining aggregate demographic variables, such as male and female populations, for broader analyses. Data cleaning for each period was customized: for 2010–2019, age-specific population data was cleaned and reshaped; for 2000–2009, the data was transformed from wide to long format, demographic variables by sex, origin, and race were generated, and these subsets were then sequentially merged. For the 1990–2000 period, yearly files were appended, age groups were aggregated, and summary demographic variables were created. The data from all three periods was then merged into a single, comprehensive dataset, $county_1990_2019.dta$, with consistent variable names and removal of redundant fields. To integrate railway crossing accident data, the demographic dataset was first spatially joined with geographic location data for railway crossings to assign each crossing to a specific county and state. This combined data was then merged with $county_1990_2019.dta$ using $county_fips$ and year identifiers, resulting in a standardized dataset organized by county, year, age group, and demographic category.

Cleaning of Residential Property Sales Data: To compile the final dataset for the analysis of noise pollution's impact on property values, I utilized six distinct datasets. First, I employed the "Assessor - Parcel Sales" dataset, encompassing sales records for real property in Cook County from 1999 to the present year, 2023. This dataset contains essential information such as Parcel Identification Number (PIN), Property class, sale date, sale price, and an indicator variable denoting whether

a parcel was sold individually or as part of a larger group of PINs.

The second dataset utilized is the "Assessor - Parcel Universe," which provides a comprehensive historical database of Cook County parcels, complete with attached geographic, governmental, and spatial data. The geographic location data is represented in terms of latitude and longitude.

The third dataset, "Assessor - Residential Condominium Unit Characteristics," contributes data regarding condominium properties, including details such as total square footage, bedroom count, land square footage, parking/garage spaces, storage unit indicators, common area indicators, and information on whether the condominium building contains non-residential units.

The fourth dataset, "Assessor - Single and Multi-Family Improvement Characteristics," is a valuable resource featuring information about property improvements, encompassing details like basements, attics, garages, porches, fireplaces, and central heating and air conditioning systems. All of these datasets are available in .csv format upon download.

To prepare the data for analysis, I imported these datasets individually into Stata 17. In Stata, I eliminated unnecessary variables, renamed important ones, and saved the datasets in .dta format. Subsequently, I establish linkages within the data by merging the sales data with unit characteristics, then with single and multi-family improvement characteristics, and finally with location data, employing the "PIN" and "year" variables as shared identifiers. The merged data were then exported in .csv format and imported into ArcGIS Pro.

In ArcGIS Pro, I first displayed the data in x-y format using the latitude and longitude coordinates. I then projected the dataset by converting the coordinate system from WGS 1984 to NAD 1983 (2011). Furthermore, I incorporated railway crossing data, which was obtained from the "Crossing Inventory Data (Form 71) - Current," provided by the U.S. Department of Transportation and published by the Federal Railway Administration (FRA). This dataset includes critical information about grade crossings, including crossing ID, crossing type, the number of passenger trains per day, and daily movements. To ensure compatibility, I projected this data to NAD 1983 (2011), aligning it with our property dataset's coordinate system.

In the next step, I performed a "Spatial Join" using the geoprocessing tool, with the property data as the target feature and the railway crossing data as the join feature. I selected the "contains" and "one to many" joining options to establish the relationships between these datasets. The resulting joined data were then converted into .csv format and imported back into Stata, where I merged it with the quiet zone data.

The quiet zone data, provided by the Federal Railroad Administration (FRA) following a Freedom of Information Act (FOIA) request (FOIA-23-00084), contains essential information, including railroad crossing IDs, quiet zone IDs, dates of whistle ban establishment, and latitude and longitude coordinates to pinpoint the locations of these quiet zones, as well as information regarding whether the crossing is open or closed.