About the TWDB Flood Planning Building Dataset

To assist with Flood Exposure analysis, the Texas Water Development Board (TWDB) will be providing statewide (~10 million) building data to the Regional Flood Planning Groups (RFPGs). In addition to footprints, improvement value, and land use, each building has an associated population estimate and Social Vulnerability Index (SVI). Improvement value and use come from StratMap 2019 Land Parcel . Population estimates are based on 2019 LandScan USA data, a 90m raster dataset from Oak Ridge National Laboratory that enables differentiation of population by day and night (see Figures 1 and 2). This allows the assignment of population to non-residential areas. SVI data comes from the Center for Disease Control (CDC) and is used to identify populations that are especially at risk during public health emergencies because of factors like socioeconomic status, household composition, minority status, or housing type and transportation (see Figure 3).

The RFPGs are expected to confirm, update, and otherwise enhance the building data as appropriate to prepare the deliverables required for their flood risk analyses tasks.

Figure 1: Snapshot of TWDB Buildings (by Use) and LandScan USA night population

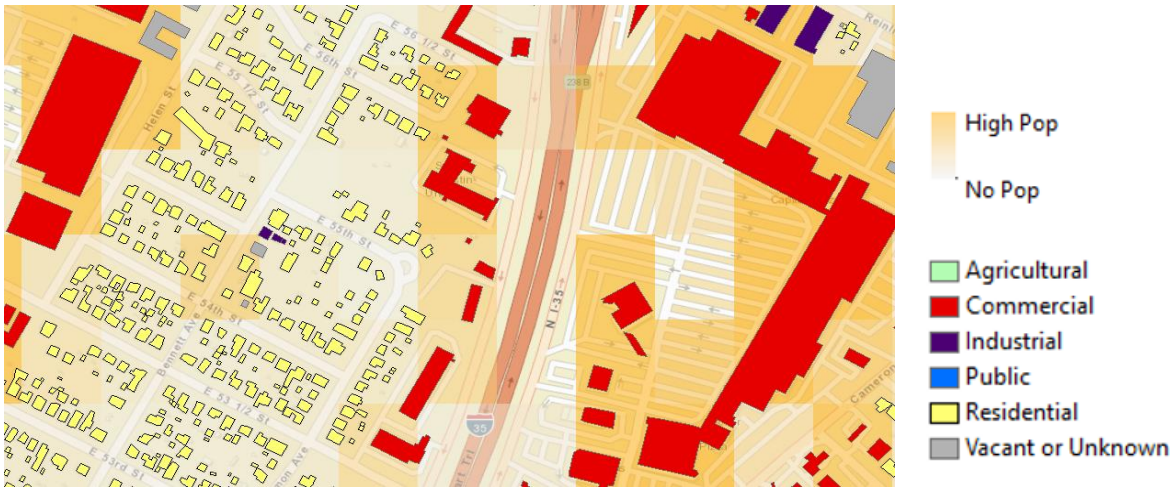

Figure 2: Snapshot of TWDB Buildings (by Use) and LandScan USA day population

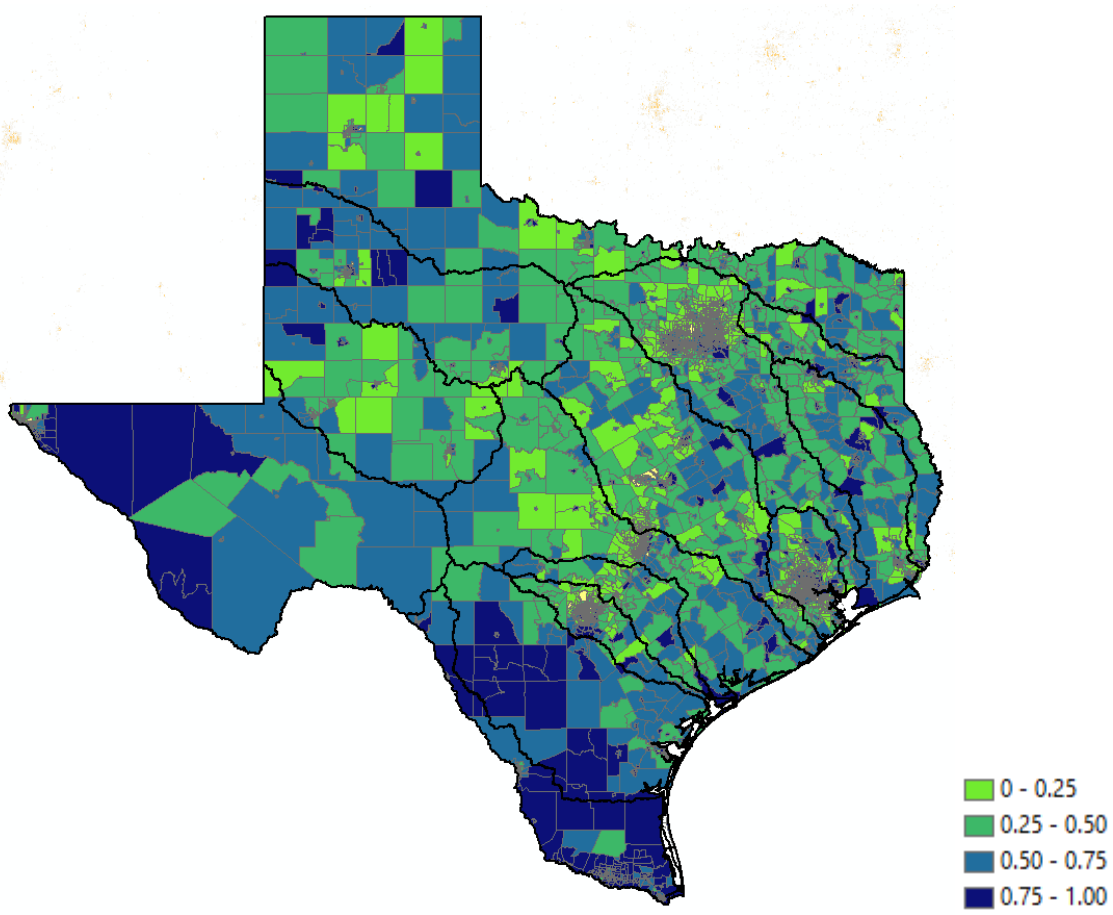

Figure 3: CDC Social Vulnerability Index (SVI) by US Census Tract

What's Included?

This building data includes data integrated from various sources, including county appraisal districts (CADs), the LandScan USA population database, and SVI from the CDC. See below for a complete field list.

Field Name (Alias)	Field Type	Source	Notes
BuildingID	Long	TWDB	Unique ID for each footprint
RFPG_NUM	String	TWDB	
RFPG_NAME	String	TWDB	
MERGE_SRC (Source)	String	TWDB	Identifies source data for each footprint
TAX_YEAR (Tax Year)	Long	TWDB / CADs	
IMP_VALUE (Improvement Value)	Long	TWDB / CADs	
Simp_Type (Simplified Type)	String	TWDB / CADs	
SVI	Double	CDC	A unit-less value between 0 and 1 where higher numbers are more vulnerable.
Shape_Length	Double	GIS	Perimeter length in meters
Shape_Area	Double	GIS	Footprint area in square meters
NightPop	Double	LandScan USA	
DayPop	Double	LandScan USA	

Methods for Data Production

Construction of the footprint feature class and the integration of data from various sources required processing by TWDB. This section details the methods and assumptions made by TWDB.

Building Footprints

Building footprints come from one of two sources, as identified in the MERGE_SRC field:

- Open Street Map (OSM) buildings (user-generated): https://osmbuildings.org/data/
- Microsoft Artificial Intelligence buildings (photogrammetry): https://usbuildingdata.blob.core.windows.net/usbuildings-v2/Texas.geojson.zip

Neither dataset is 100% accurate, so both were combined to produce a single, more comprehensive layer. The OSM buildings dataset has greater coverage in urban areas, while suburban and rural areas tended to have better coverage from the Microsoft data. This is because there are fewer users in less developed areas, so photogrammetry produces better results where population density is lower. However, the Microsoft data still has gaps with no buildings, and sometimes crops or agricultural fields are falsely denoted as buildings. TWDB conducted quality control of the data, attempting to remove such buildings. Despite this, users must understand that the footprints are only a best estimate, a snapshot in time. There are current efforts to clean this dataset by using a measurement of the shape, the Fractal Dimension, to flag and remove false positives. Additionally, this dataset will be gradually upgraded with more accurate LiDAR-derived building footprints.

County Appraisal Data

Data on improvement value and land use come from the StratMap 2019 Land Parcel data managed by TNRIS.

"Simplified_Type" is based on the building type. Different CAD systems use different categories for the land parcel use and type, some of them can be very diverse and cover tens of classes of residential for example. The data was assigned to simplified groupings that align with Building Category Definitions (as per Texas Property Classification Guide, https://comptroller.texas.gov/taxes/property-tax/docs/96-313.pdf). See below for a summary:

- Residential property includes townhouses, condominiums, row-houses, duplexes, mobile homes, multi-family, and single-family residential improvements.
- Commercial improvements are associated with businesses that sell goods or services to the general public. Some examples of commercial businesses are wholesale and retail stores, shopping centers, office buildings, restaurants, hotels and motels, gas stations, parking garages and lots, auto dealers, repair shops, finance companies, insurance companies, savings and loan associations, banks, credit unions, clinics, nursing homes, hospitals, marinas, bowling alleys, golf courses, and mobile home parks.
- Industrial improvements of businesses that add value to a product through development, manufacturing, fabrication, or processing of that product. Some examples of industrial businesses are cotton gins, processing plants, paper mills, steel mills, refineries, warehouse storing for a manufacturing facility, cement plants, chemical plants, canning companies, and clothing manufacturers.

- Agricultural improvements include all barns, sheds, silos, garages, and other buildings associated with farming or ranching other than residences.
- Public improvements owned by the government, private schools, school districts, religious organizations, charitable organizations, and other property not reported elsewhere.
- Vacant or Unknown properties are improvements not occupied. Because property use determines classification, there is no minimum or maximum size requirement for this category.

The IMP_VALUE comes from the improvement value of the parcel dataset. This data is incomplete with many null values. Some of these are public or religious buildings for which CADs do not assess value. For the buildings that have an Assessed Improvement value, a spatial join was made with the Land Parcels dataset to join the two. In cases such as apartment complexes and parcels with guest houses, the value of the improvement was apportioned to each building using the footprint area. For example in a parcel with an improvement value of $600,000, if the main house has a square footage of 2,000 while the guest house has only 1,000 sq. ft. then 2/3 of the value was applied to the main building and the remainder to the guest house.

For the buildings that do not have an assessed value, average assessed values were given by Improvement type and by County, since different counties have different prices per sq. ft. for each building type. This table has about 4,000 different combinations of Improvement_type-County and each has a mean Improvement Value, a mean number of Square Feet, and a Mean_Price_SQFT. Those were used to calculate an estimation of the value of the building lacking a CAD Assessment Value.

Combining with LandScan

LandScan USA is produced by the Oak Ridge National Laboratory and is intended for a variety of uses. It uses dasymetric analysis to spatially allocate the US Census population into a ~90-meter raster grid. The TWDB building data leverages this spatially explicit population dataset to assign both a night and daytime population to the building footprints.

Population assignment was achieved by first converting LandScan USA raster data into polygons (using a fishnet). Next, the building footprints were intersected with the LandScan derived polygons, yielding building-cell intersections that form the base unit-of-analysis for assigned population numbers. After filtering out buildings under 500 square feet (assumed to be primarily sheds, etc.; calculated using NAD_1983_2011_Texas_Centric_Mapping_System_Lambert projection), the following procedure is used to allocate populations:

1. Calculate the total area of all building-cell intersections within a given LandScan cell
2. Divide the area of each building-cell intersection by #1 to provide an allocation percentage
3. Multiply the allocation percentage against the LandScan cell value for day and then for night population
4. Sum all building-cell intersection population values for each building

Population Quality Assessment

The aim of the combination of LandScan USA and TWDB Building data was to locate the population as accurately as possible for use in flood risk analysis. Toward this end, a quality assessment was conducted to assess the completeness of building-based populations, see below for a summary:

			LandScan USA Population		Assigned to Buildings			Delta (total)			Delta (%)		
FPR	Total Footprints	Under500	Night	Day	Bldg#	Night	Day	Bldg#	Night	Day	%	%	%
1	336,983	16,147	604,878	604,963	320,836	585,704	566,448	(16,147)	(19,174)	(38,515)	95%	97%	94%
2	292,110	11,976	513,222	494,072	280,134	489,094	476,370	(11,976)	(24,128)	(17,702)	96%	95%	96%
3	2,306,324	40,159	7,864,784	7,931,060	2,266,165	7,514,859	7,679,916	(40,159)	(349,925)	(251,144)	98%	96%	97%
4	309,707	13,284	587,287	545,809	296,423	544,858	521,567	(13,284)	(42,429)	(24,242)	96%	93%	96%
5	443,874	16,163	916,397	929,563	427,711	875,514	896,343	(16,163)	(40,883)	(33,220)	96%	96%	96%
6	1,661,412	49,902	6,307,928	6,361,911	1,611,510	3,692,032	3,530,645	(49,902)	(2,615,896)	(2,831,266)	97%	59%	55%
7	304,905	12,738	633,698	639,738	292,167	585,413	615,486	(12,738)	(48,285)	(24,252)	96%	92%	96%
8	1,085,151	50,322	2,721,116	2,497,365	1,034,829	2,514,813	2,346,139	(50,322)	(206,303)	(151,226)	95%	92%	94%
9	286,360	11,672	630,172	637,615	274,688	600,859	607,807	(11,672)	(29,313)	(29,808)	96%	95%	95%
10	834,062	105,313	1,873,270	1,851,724	728,749	1,753,405	1,771,661	(105,313)	(119,865)	(80,063)	87%	94%	96%
11	260,891	14,654	616,717	572,380	246,237	573,115	541,294	(14,654)	(43,602)	(31,086)	94%	93%	95%
12	663,207	14,604	2,231,721	2,219,261	648,603	2,047,776	2,116,634	(14,604)	(183,945)	(102,627)	98%	92%	95%
13	359,748	13,658	776,038	770,739	346,090	726,839	710,732	(13,658)	(49,199)	(60,007)	96%	94%	92%
14	286,591	7,062	937,133	979,007	279,529	789,005	879,112	(7,062)	(148,128)	(99,895)	98%	84%	90%
15	513,945	13,009	1,775,808	1,745,068	500,936	1,649,275	1,587,678	(13,009)	(126,533)	(157,390)	97%	93%	91%
	9,945,270	390,663	28,990,169	28,780,275	9,554,607	24,942,562	24,847,833	(390,663)	(4,047,607)	(3,932,442)	96%	86%	86%

Table 1: Validation Summary by Flood Planning Region (FPR)

As demonstrated by Table 1, no region's population can be entirely allocated to a building footprint. There are several reasons for this, to be discussed below. For most regions, this is because of a temporal mismatch (#1 below). The missing population closely aligns with the expected population growth in the region that occurred between the production dates for the building footprints (variable) and LandScan (2019). However, region #6 has a significant portion of the region without footprint data.

Four types of known errors:

1. Temporal mismatch – LandScan USA represents the 2019 population, but some of the building data is older (2016). Thus, newer growth on the urban periphery shows up in LandScan but not in the building data.
2. Low density rural – LandScan USA and Buildings do not always match within 90m in low-density areas.
3. Missing Buildings – the two sources used for this iteration cover the entire state, but there are a few locations (particularly within the Houston, TX, metro region) where data is still missing.
4. County Edge Overstatement – the method of geoprocessing used to allocate population to building footprints required statewide data to be clipped and processed as smaller units (county and Flood Planning Region intersections). In doing so, it made it possible for buildings on the edge of a county to receive overstated population counts because not all buildings in a LandScan cell were within the processing unit.

Figure 4: Images of TWDB building footprints (green) against LandScan (left) and current base map (right). The base map shows an under-construction subdivision with partial building coverage.

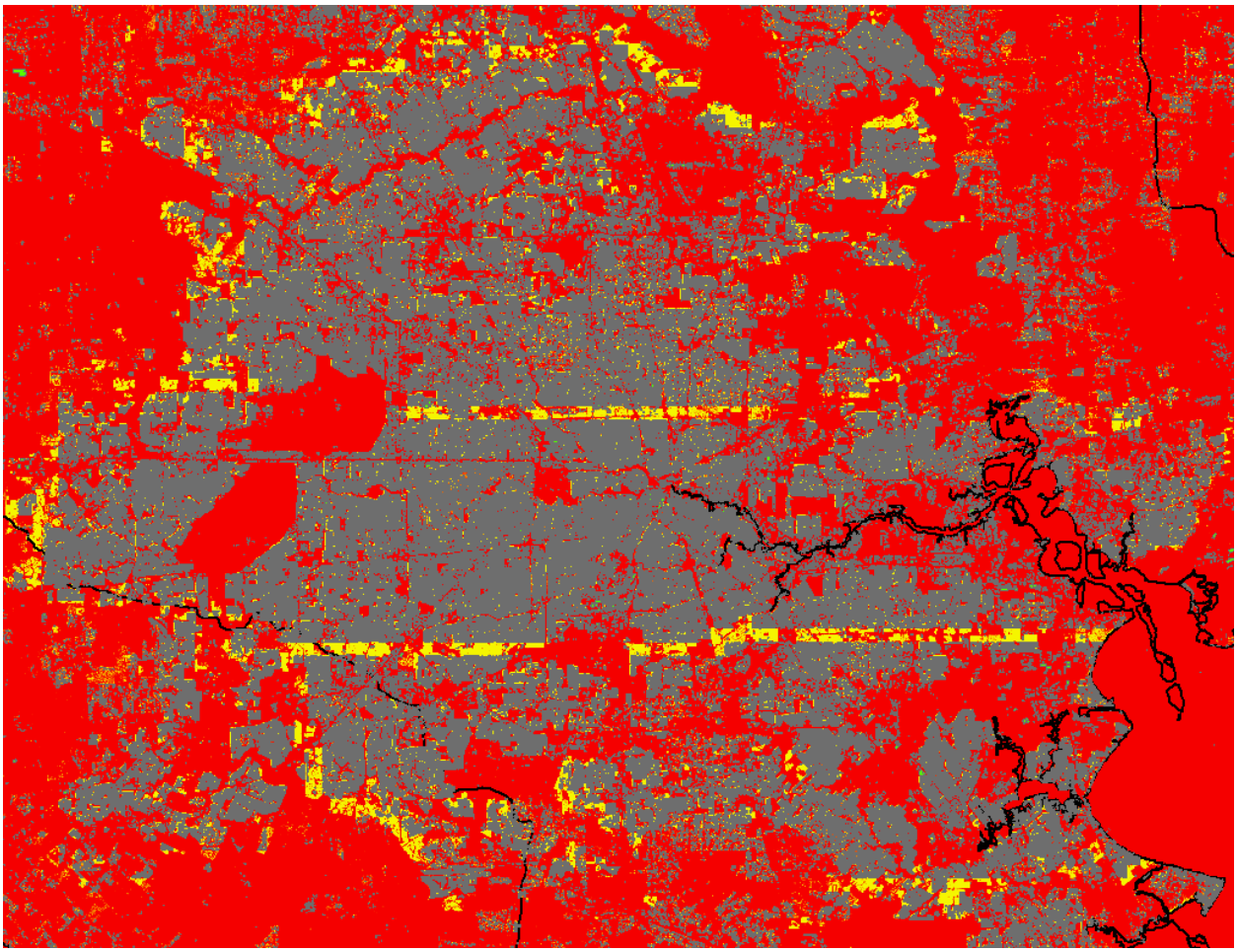

Figure 5: Houston, TX LandScan USA (yellow) overlaid by TWDB building footprints. Where yellow is visible identifies population areas without building coverage.