NSSAC Building Knowledge Base: Modeling and Implementation

Author(s):
Henning S. Mortveit
Dawen Xie
Madhav V. Marathe

NSSAC Technical Report:
No. 2021-016

Date of report:
February 13, 2023

Status:
Approved for NSSAC Internal Release

Contact:
Henning S. Mortveit
Henning.Mortveit@virginia.edu

Overleaf URL:
https://www.overleaf.com/project/6015ecbb984e794fbf702a20

Network Systems Science and Advanced Computing
Biocomplexity Institute and Initiative
University of Virginia

NSSAC Building Knowledge Base*

Henning S. Mortveit, Dawen Xie, Madhav V. Marathe

January 2021

Abstract

This work describes the NSSAC Building Knowledge Base (BKB) that is designed to capture all buildings located in the United States along with key information such as business and dwelling counts within each building, geo-spatial information, NAICS classifications (e.g., school, medical, government, retail), building characteristics, as well as a range of additional augmented data. By design, the knowledge base readily permits integration of additional data, thus serving as a coordinate system for building data. Here we describe (i) the design based on the notion of a core with components that augment the core, (ii) the current set of data sources that were used either directly in constructions or for model training, and (iii) technical validation and general data quality aspects.

We include an application of the knowledge base to the construction of synthetic populations for the United States. Using the BKB, we give an algorithm for generating precise data for residence locations and business locations, the places where the individuals of the synthetic population conduct their activities and where interactions happen. The resulting time-indexed map of people to locations provided by the resulting SP serves many purposes: it can support disaster- and resilience planning for evacuation, transportation planning, and infrastructure development, to mention some. At a very direct level, it can be used to construct heatmaps of population density throughout a day, thus providing a principled, bottoms-up alternative for construction of LandScan-type data.

1 Introduction

Capability. A broad range of scenarios involving for example smart and connected communities, evacuation, and infrastructure resilience closely rely on detailed information about where people are at what times during the day, as well as the properties of the locations that they visit. In this paper, we describe a novel approach for constructing a knowledge base of locations (NSSAC-BKB) covering almost all buildings of the United States. This can form the basis for *synthetic populations* where a mapping from the set of people of the U.S. to the set of buildings (see [2, 9]). The implementation of the NSSAC building knowledge base (NSSAC-BKB) is referred to as the *NSSAC Building Database* (NSSAC-BDB). Just like synthetic populations represent a coordinate system for attaching information to people and households, the NSSAC-BKB is a *coordinate system for buildings* allowing one to naturally fuse a broad range of data to buildings, including actual collected ("real") data as well as model-inferred data. Examples of such data include (a) solar value and climate zone [Swapna], (b) broadband and cellular coverage [data-set-ref], (c) flooding risk or risks to various natural disasters, (d) modeled energy consumption of households [reference to Swapna's work], and (e) connecting buildings to the electrical distribution networks, thus providing coverage/outage maps as well as forming a foundation for analyzing smart grid solutions with local generation [rounak-reference].

Novelty. To the best of our knowledge, no other similar, open resource exists with our flexible design for augmentation, thus making it a unique resource for studying a broad range of scenarios related to societal resilience at large. Parts of the data covered in this work are also captured in commercial- or confidential/privacy-regulated solutions limiting availability and use. Examples include county- and municipality building records and building-use records, business data such as Dun & Bradstreet [3], PRIZM

*Biocomplexity Institute and Initiative Technical Report no. 2021-016

Premier [14], SafeGraph [15] for points of interest (POIs), and HERE Premium Streets U.S. with POIs [7] (formerly NAVTEQ). There are valuable open solutions like OpenStreetMap that also have POIs [13], but sometimes the quality of such sources as well as coverage cannot compete with commercial solutions. On the other hand, many of the commercial solutions have a marketing segmentation/analysis focus in their data products and less of a focus on for example infrastructure at large, which can possibly cause biases in what is recorded if that is the intended application domain.

The NSSAC-BDB is an effort to fuse both open and commercial data with the limitations and restrictions imposed by the above. In this paper we detail our design which can flexibly integrate broad classes of data and where commercial data dependencies can be tracked in a transparent manner, thereby supporting data export and data sharing that cleanly respects license terms of commercial data by, e.g., avoiding this altogether in the export process and/or using open, but possibly less detailed, alternatives for such purposes.

Applications. Applications of the above models, in particular in conjunction with synthetic population data, are nearly limitless. For example, using the location assignment that maps each person of the synthetic population to a location for each of they daily activities, one can derive exposure maps of all types (e.g., exposure to pollution, hazards), as well as obtain highly accurate estimates of where people are throughout the day which would be helpful for evacuation planning for sudden natural and man-induced disasters.

Data sources. A broad range of data sources was used in this work. Some data sets were used directly, while otehr data sets (e.g., commercial data sets) were used in for example model training. For its modeling and data, the NSSAC-BKB currently uses sources such as the Microsoft Building Data [8], NCES [11], HERE [7], BuildingFootprint USA [1], U.S. Census data [18], and OpenStreetMap [13]. By design, NSSAC-BKB supports (a) full use of commercial data for licencees in accordance with contract terms, (b) use of commercial data in model training with subsequenct model-based classifications, and (c) and modeling and instantiation using only open data which may be shared under the same terms as the open data being used. The report includes technical validation covering (i) buildings, building types, and building counts, (ii) comparisons with data sources such as LandScan, GPW v.4, and WorldPop for population densities, and (iii) validation based on NSSAC's synthetic populations which contain a mapping of people to buildings for each point in time during a normal week and in turn permits construction of accurate heat maps of people densities at any time of day. The latter may even be of interest for e.g. LandScan's own calibration.

Features and highlights. Some of the features included in NSSAC-BKB are the following:

- Footprint geometry for essentially every building entity in the U.S. through an extension of the Microsoft Building Data [8];
- Area and centroid (latitude and longitude) of each building entity;
- An urban/rural classification of each building based on U.S. Census data [17];
- A classification of each building entity as residential, non-residential, or mixed;
- Modeled estimates for the number of apartments (dwellings) and businesses associated to each footprint
- An assignment of a public school to each building entity using data for each grade level based on data from NCES [11];
- An NAICS-derived classification of each business entity associated to a building entity (see [10]);
- A common index for all building entities across all states. For MS building data, uniqueness of IDs only hold within a state.

Paper organization. In Section 2 we describe the high-level design with the common base and modular fusion mechanisms for data augmentation. The section also includes an overview of the current open and commercial data sets currently used, and the export mechanisms ensuring that one can use both commercial data when licensed and open data for sharing. Technical validation follows in Section 4, while a description of standard data records and data fields are described in Section 5 for both open and restricted solutions. In

Section 6 we summarize current limitations and also list planned extensions and possibilities for future work. The appendix sections contain a detailed overview of the standard table structures that result by applying the modeling described in the paper.

2 Modeling and Design

This section details the design of the NSSAC-BKB which can be outline into the following steps, many of which are illustrated in Figure 1.

- Construction of the NSSAC-BKB `Core` (Section 2.1)
- First residential/non-residential modeling pass (Section 2.2)
- Construction of the mapping β: BFUSA $\longrightarrow$ `Core` (Section 2.3)
- Second residential/non-residential modeling factoring through β (Section 2.4)
- Construction of the mapping η: HERE $\longrightarrow$ `Core` (Section 2.5)
- Modeling and fusion of educational institutions (Section 2.6)
- General notes on modeling and fusion of other spatial data and entities (Section 2.7)

General Design for Data Fusion. The general mechanism for fusing spatial data with the NSSAC-BKB is through the geometry data present in the core typically using criteria involving point containment and/or intersections of geometries. Generally, these criteria may include modeling, possibly using machine learning and training as for commercial data with restrictions on use. To ensure a modular design, each new data set that is fused will generally not modify anything in the `Core`, but will augment and make reference to the `Core` using a separate namespace (e.g., a database table for the NSSAC-BDB) specific to the data being fused. Some data sources will be fused using a *chain of data sources*, e.g., $A \longrightarrow B \longrightarrow$ `Core`, through what effectively amounts to a join operation. In this example, there may be data common to A and B that allows one to fuse from A to the `Core` by exploiting the existing correspondence between B and the `Core`. Clearly, as we develop more and more correspondences, the NSSAC-BKB becomes increasingly more powerful. However, it does introduce dependencies, and one must also keep careful track of the collection of inferences and relations that are established through modeling.

Fusion Based on Data Sets that are Subject to License Restrictions. High quality data sets are often have terms restricting their use (e.g., they may be commercial or involve sensitive information). Using such data in data fusion, either directly or though modeling and training, required special care in the design of the NSSAC-BKB. We pay close attention to license restrictions of data sets and store augmentation results separately based on the input data. In this way, knowledge acquired through open data and commercial data are clearly distinguished. Whenever possible, we strive to establish open alternatives for any data set used with restricting licensing, thus keeping the option of still being able to fuse data, but possible at some cost for data quality.

2.1 The NSSAC-BKB `Core`

The notion of *core* is fundamental to NSSAC-BKB. The `Core` as it is currently constructed, is an extension of the Microsoft Building Database (v1.1) which was released in 2018 [8], containing more than 125 million building footprints for all 50 U.S. States and the District of Columbia.[1] Our extension is designed to be able to flexibly accommodate additional building entities[2] that are spatially embedded and that may (a) be constructed after 2018, and/or that (b) are not captured adequately or that are not represented in the

[1] DX: will update text related to MS data once we are ready to use their v2 release.
[2] We will use the terms footprint and building entity interchangably

Microsoft data. An important example of the latter category is the following: the U.S. Census reports that there are block groups containing households but for which the Microsoft data has no corresponding building footprints. Clearly, many scenarios and analyses need a remedy for this. Not using the Microsoft data directly as the `Core` allows us to flexibly accommodate application scenarios that may require addition of other footprints or entities specified by the project sponsor or derived through other means in support of projects.

The general modeling includes the following steps:

1. Map each Microsoft building entity m into the `Core` by assigning it a *footprint ID* or `fid` in addition to other data described below;
2. Missing residential footprints correction: for each block group b for which there are no footprints assigned after the completion of Step 1, but for which the U.S. Census reports that there are households, insert one additional point geometry placed randomly within the boundary of b to accommodate one or more dwelling spaces that can accommodate said households. The number of block groups for which this is the case using the 2017 U.S. Census/Microsoft 2018 data is 3,227.
3. Additional location classes (e.g., military installation, or general group quarters) that are clearly not captured by step 1 can be accommodated in a similar fashion, but see Sections 2.2 and 2.4 for the cases where the residential/non-residential modeling is adjusted.

As will become clear in the following sections, this `Core` is the central coordination point used to link all other data sources that are integrated. Each footprint is augmented with the following data – see Appendix Section B.1 for a complete description:

- `fid`: the index used with the `Core`;
- Geo-spatial coordinates of the `fid` geometry and its centroid (WGS-84);
- Footprint area (in square meter);
- Administrative codes based on 2017 blockgroup data from U.S. Census Bureau for spatial indexing;
- Urban/Rural classification based on 2010 census urban area data;
- Source information (e.g., MS table, or synthesized)

2.2 Initial residential/non-residential modeling

The residential/non-residential classification initially applied to the `Core` is done through an exclusion approach that uses two spatial filters $\mathcal{F}_C$ and $\mathcal{F}_P$ constructed and trained from, e.g., land-use polygons, and through a union of disks placed at selected points-of-interest (POI) categories. The filter $\mathcal{F}_C$ (categories) includes regions classified as for example shopping center, airport, cemetery, golf course, industrial park, rail-yard, seaport, hospital, and sports complex. The POI filter $\mathcal{F}_P$ (POIs) includes classifications such as auto service, parks and recreation, shopping, transportation hubs, and places of worship. It was trained using sources such as [16, 7, 1, 13].

In this initial modeling step, there are only two classes:

- `R`: residence
- `U`: unknown

A footprint (or building) is classified as non-residential if its centroid falls inside the geometry defined by the combined filter $\mathcal{F}_C \cup \mathcal{F}_P$, or if its footprint satisfies area $< m$ or area $> M$ for positive parameters m

and M. (In the current version, these values are $m = 20$ and $M = 5000$.) Moreover, a footprint is classified as non-residential if there are no roads classified as "residential roads", or the equivalent thereof, sufficiently close in any of the respective road network data used. A location or building for which none of the above holds is classified as residential. At the end of the intial classification, each footprint has classification `R` or `U`.

The classification method for deciding if a footprint is residential or non-residential currently handles a large diversity of cases when applied to any of the U.S. states. However, some areas such as large cities that have massive apartment complexes (some of which inadvertently are merged in [8]), are at risk of having locations incorrectly classified as non-residential since they have footprint area that exceed M. In Section 2.4 we describe the correction step that we apply to handle this and other cases. A list of data sources used and implementation details can be found in Appendix Section B.2.

2.3 The mapping β from BFUSA to the `Core`

One of the key data products used for constructing β is *BuildingFootprintUSA (BFUSA)*, a building footprint data product for the United States developed by LightBox [1]. Its key features for use with NSSAC-BKB can be described as follows. First, the BFUSA data contains four layers: (1) the *primary layer* is a polygon layer containing building footprints with known addresses; (2) the *secondary layer* is a point layer containing comprehensive address information including secondary unit addresses; (3) the *no address layer* is a polygon layer containing building footprints for which there is currently no address information; and (4) the *address points layer* which is a point layer containing any address point locations that are not be linked to any building footprint. For the NSSAC-BKB the primary and secondary layers of BFUSA data are used. These two layers are linked by a 12-character string ID. One record in the primary layer has at least one associated record in the secondary layer; in some cases there can be hundreds of associated records as in the case of a large apartment- or business complex.

The mapping β from BFUSA to the `Core` is constructed using the BFUSA primary layer and the centroid of associated geometry objects. The (partial) mapping is established using BFUSA centroid containment within `Core` geometry. For each mapping, we compute sums of apartments (bfusa_num_apt) and suites (bfusa_num_ste) for each FID based on BFUSA secondary layer. Of the 77,015,829 BFUSA primary objects 71,232,713 can be mapped to `Core` geometry using this approach. Initial experimentation using a geometry buffer of radius r associated to each `Core` geometry was abandoned based on the limited number of additional matches it produced as well as the one-to-many ambiguity that arose in this situation, even for very small values of r. We remark that the mapping β is not injective, a fact that partially attributed to slight discrepancies between underlying data dates and geometry objects in the Microsoft data and those of the primary layers of BFUSA, nor is it surjective, the latter being indicated by the subset B' of the `Core` on the left side in Figure 1. There are also cases where the centroid of a footprint falls outside its geometry, even when BFUSA and Microsoft foootprints are aligned, see Figure 2. Examination of a selection of regions indicates that this number is quite small. For each of the 71,232,713 footprints of the BFUSA primary layer that can be mapped to the `Core` we also compute the total number of apartments and total number of business suites based on the secondary unit addresses, see the Appendix Section B.3.

2.4 Adjusted modeling for residential/non-residential classification

With the approach for residential classification outlined in Section 2.2, about 112.4 million core entities are classified as residential. A limitation of this initial modeling arises when a core footprint contains a large number of dwellings as in the case of an apartment complex, or when multiple real footprints inadvertently get merged during processing in [8]. Moreover, the initial modeling does not provide information about non-residential building entities and what is located within those.

As described in Appendix Section B.2, each footprint in the image of the mapping β is augmented by the fields

`core_id`, `bfusa_id`, `initial_r_nr_class`, `bfusa_num_apt`, and `bfusa_num_ste` .

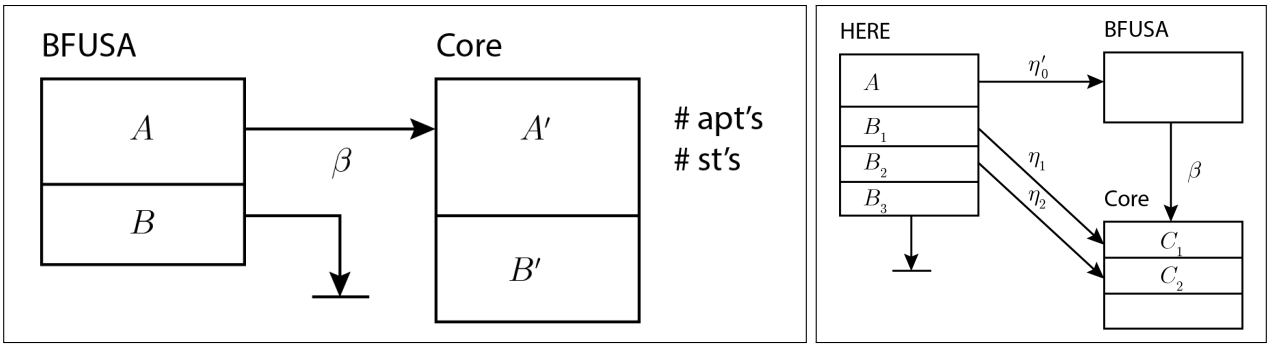

Figure 1: The mappings β and η providing the correspondence with the `Core`.

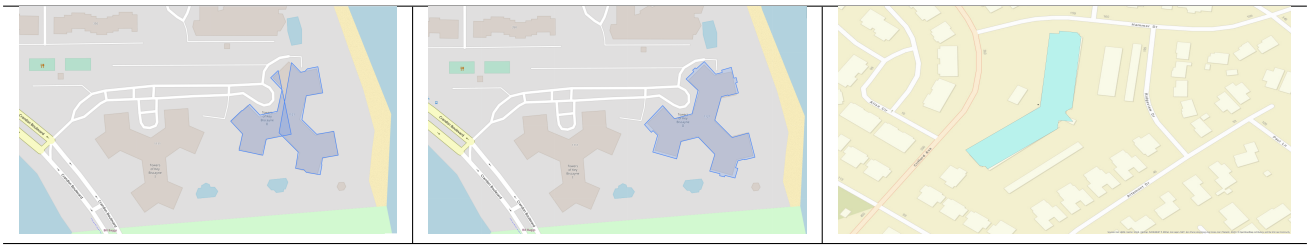

Figure 2: An illustration of some of the challenges faced when fusing data sources: (left) a building represented as two overlapping footprints in the MS data and (middle) the same building and its geometric representation in BFUSA; (right) a case where the MS footprint (blue) does not contain its own centroid (in black).

We expand the initial footprint classification from {`R`, `U`} to the following four classes

`R`: residence, `B`: business location, `RB`: mixed usage, and `U`: unknown .

More than one BFUSA footprint can be mapped to the same `Core` footprint, and we estimate the total count of apartments (apartment_counts), and total count of business suites (suite_counts) for each `Core` footprint by factoring through the mapping β, see Figure 1. The entities of the set `Core` $\setminus \beta$(BFUSA) retain their initial classification derived in Section 2.2. The updated modeling can be described as follows using an SQL like syntax:

```
R  if (initial_r_nr_class = R OR apartment_counts IS NOT NULL)
        AND (suite_counts IS NULL)
B  if (initial_r_nr_class = U AND apartment_counts IS NULL)
        AND (suite_counts IS NOT NULL)
RB if (initial_r_nr_class = R OR apartment_counts IS NOT NULL)
        AND (suite_counts IS NOT NULL)
U  if (initial_r_nr_class = U AND apartment_counts IS NULL)
        AND (suite_counts IS NULL)
```

2.5 The mapping η from HERE to the `Core` factoring through β

This part of the design integrates information about specific, geo-located building types, focusing on HERE [7], but could also include data such as OpenStreetMap [13], SLIPO [16], and D&B data [3].

For the 2021 Q1 release of HERE, there are around 19 million POIs. Each POI has data including facility type, address, and geometry, while data from the extended listings also include the NAICS type [10]. However, there is no footprint information for the POIs from HERE. In fact, a HERE POI is typically

represented as a point placed on the street (e.g., mail box location), unlike `Core` footprints that capture the actual geometry. Herein lies a big challenge: the misalignment in representation between `Core` and HERE makes it effectively impossible to spatially join data for HERE POIs and core footprints. Our multi-pass modeling approach reflects this, and can be detailed as follows, see Figure 1 and the diagram on the right.

Pass 1. Factoring through the mapping β. The first pass takes advantage of the fact that both HERE POIs and BFUSA data have addresses, and we establish a partial mapping $\eta_0' \colon \text{HERE} \longrightarrow \text{BFUSA}$ and then form the map $\eta_0 \colon \text{HERE} \longrightarrow \texttt{Core}$ as $\beta \circ \eta_0'$ thus using BFUSA data as a bridge, see Figure 1, right side. By using street name and street number for each address, stripping spaces between words and making all letters lower case, an address match is added when there's an exact match and both the HERE POI and the BFUSA footprint are in a same blockgroup. However, the addresses in both data sets are not standardized and we are only able to map 7.5 million HERE POIs to the `Core` using the following:

Remark: some HERE POIs are mapped to more than one core entity. This can be caused by inaccurate geo-locations for some POIs, and also situations like the overlapping footprint geometries as illustrated in Figure 2.

Pass 2. Matching of remaining HERE POIs within blockgroup to `fids` not classified as `'R'`. For the remaining HERE POIs not mapped in step 1 (about 11.5 millon) we identify each HERE POI's nearest `Core` footprint that is not modeled as `R` (residence location) and that is located in the same blockgroup as the POI; we also compute the distance `d` between the HERE POI and the `Core` footprint. The choice of cut-off value for `d` was based on the following measurements:

- For `d` = 200 meters, there are 2,690,737 remaining HERE POIs that cannot be mapped;
- For `d` = 300 meters, there are 1,937,474 remaining HERE POIs that cannot be mapped;
- For `d` = 1000 meters, there are only 380,738 remaining HERE POIs that cannot be mapped to a `Core` footprint.

Since the mapping of the second pass is restricted to footprints in the same blockgroup, and since we only use footprints that are classified as non-residential in both modeling passes, using `d` = 1000 meters as the criterion in the second pass was deemed reasonable, thus giving the mapping $\eta_1 \colon B_1 \longrightarrow C_1$ as shown on the right in Figure 1.

Pass 3. Matching of remaining POIs within blockgroup to `fids`. The $380,738$ HERE POIs that remain after the previous two passes using `d` = 1000 meters as threshold are mapped to the nearest core building within their blockgroup, establishing the mapping $\eta_2 \colon B_2 \longrightarrow C_2$ shown on the right in Figure 1.

We remark that the subset B_3 of the HERE data (see Figure 1) is not mapped. Moreover, the mapping η also records the number of the pass for which the correspondence was established. Details for fields, table names, and structure are given in Appendix Section B.4.

2.6 Educational institutions

Modeling of educational entities is done in great detail using data from the *National Center for Education Statistics* (NCES) [11]. The final results of the modeling described in this section are (i) a table that lists each entity contained in the NCES with relevant additional data described below, and (ii) a partial mapping from NCES entities to footprint IDs, and (iii) a mapping that assigns grade-level appropriate public school IDs to each `fid` of the `Core`. See Appendix Section B.5 for details and data structures.

The table in (i) captures the NCES school IDs (`ncessh`), teacher counts, students counts, and grade levels served. Precise description of data are given in Appendix Section B.5.

The partial mapping generated in (ii) relating NCES data and `Core` footprints uses the location data from NCES (latitude, longitude) and is constructed using a two-pass approach. In the first step, containment inside footprint geometry is used to establish the correspondence whenever this applies. In the second step,

the remaining NCES entities are assigned to their nearest `Core` footprint. We remark that a `Core fid` may contain dwelling units, business entities, and schools, something that is not uncommon in larger cities.

The final mapping in (iii) assigning NCES public school IDs[3] for each grade level for each footprint in the `Core`, uses the following approximation: for each block group and for each grade level, we determine which public school should serve that grade level based on the proximity of the schools' coordinates to the block group centroid. For this, we first attempt to find a grade level appropriate school within the county containing the block group. If there is such a school, the NCES school ID `ncessh` of the closest one is assigned to the footprint for that grade level. If no school is found within the county, the search is extended to the nearest school within the state containing the block group. This approach, which successfully assigns an NCES school ID to each (footprint, grade level) combination, was chosen in parts for computational efficiency and scalability, and parts because the school zone boundary data, albeit of high quality, contained too many missing pieces to allow for automated processing.

For reference, we remark that the for the data in [11], there are about 218,000 NCES entities, which should be compare to the 125 million building entities of the `Core`.

2.7 Modeling of other location entities and geo-spatial data

The approaches used to map school entities and POIs to `Core` footprints using a combination of geometry containment, nearest search with restrictions (e.g., be located within the same administrative region) can clearly be applied or extended to other data for which geometry is available.

Information such as broadband coverage[4], climate zones [12], solar values, and various hazard maps (e.g., hurricane, and flooding) can easily be fused with the `Core` to construct appropriate mappings stored as tables, see Appendix sections for examples.

HSM: should we have one example? Maybe simply point the existing urban/rural classification. Pretty much every simple augmentation will be of this kind.

3 Export of data from the NSSAC-BKB

The NSSAC-BKB is designed to support a broad class of applications. Here we describe how it is applied to construct synthetic populations (or digital twins) of the U.S. population, showcasing how it can be used to construct what is referred to as the *residence locations* and *activity locations* needed in that modeling. By the design of the `Core`, we can ensure that all the generated locations for the U.S. are given unique IDs. For reasons that will become clear below, activity locations are constructed before residence locations.

3.1 Activity locations:

The activity location data required for the U.S. detailed pipeline takes the following format:

```
alid,longitude,latitude,altitude,admin1,admin2,admin3,admin4,\
work,shopping,school,other,college,religion,designation
```

The first eight fields deal with book-keeping (ID, geo-coordinates, and administrative region specification), while the fields `work`, `shopping`, `other`,`school`,`college`, and `religion`, are attractor weights for the given location used. These attractor weights are used when mapping a person's activities to locations where the activity will take place. For schools, for example, the `school` attractor weight will be the number of students attending the given school, whereas for `shopping`, all locations supporting this activity are assigned the

[3] Private schools are currently not mapped.
[4] https://broadbandnow.com/national-broadband-map

uniform attractor weight of 1 in the current version. The final field, `designation`, is used in the location assignment of work activities: people have a matching designation field, and when assigning a person's work location, only locations matching their designation will be considered in the assignment.

The preparation of activity location data is organized in the following sequence of steps.

1. A mapping from the POI table to a new location table is constructed. This table, whose precise elements includes all the fields of activity location format (as well as others), is listed in Appendix Section C.
2. The `designation` field is updated next based on POI characteristics such as NAICS [10] codes, the source of the data in the POI table. The default value is `none`.
3. School and college attractor weights and designations.

Schools. In the preparation of school location data, we construct a mapping from school entities (indexed using NCES IDs (`ncessh`)) to footprint IDs (indexed using `fid`s). When generating activity locations for schools, the location ID `lid` constructed will also keep the field `ncessh` from which all required data can be gathered, including attractor weights for activities Work and School, and will also have access to the associated `Core` element through the mapping mentioned above. The assigned designation is `education`.

College. We use student enrollment information as a guide to generate activity locations for colleges. In particular, we create an activity location for every 1,000 students.

Religion. One activity location is constructed for every stored POI location for which the facility code matches that of a place of worship and we set attractor weight to 1 for religion only and designation `none:religion`.

3.2 Residence locations

The data fields needed to construct residence locations for use with digital twins for the U.S. population are the following:

```
rlid,longitude,latitude,altitude,admin1,admin2,admin3,admin4,\
area_sqm,associate_link_func_class,\
pub_pk,pub_kg,pub_01,pub_02,pub_03,pub_04,pub_05,\
pub_06,pub_07,pub_08,pub_09,pub_10,pub_11,pub_12
```

Those appearing on the first line are the same as those for activity locations, the ones on the second line can be ignored here, while those on the third and fourth lines are the school location IDs (`lid`s) as generated above.

Remark: for the U.S., the ID of residence locations are generated using an offset of 1,000,000,000. This ensures (i) that there is no overlap of IDs between activity locations for residence- and activity locations, and (ii) that if new activity locations need to be generated (e.g., as required by particular scenarios or analyses), this can be done without worrying about running into the residence location ID range.

Any element of the `Core` with classification `R` or `RB` from our modeling is a candidate for one or more residence location, the specific count can be modeled using the variable `bfusa_num_apt`. Each dwelling unit within an apartment complex is represented as a separate residence location. This is all right since the modeling within the computational pipeline for U.S. digital twins does not include any mixing between households residing within the same apartment complex. However, this can easily be updated when that modeling incorporates such features. Source table and source ID are kept to make sure we can link back to original data source.

4 Technical Validation

In the preparation of the NSSAC-BKB, extensive testing and calibration was done.

Location density plots. In Figure 3, a point plot of all locations for the state of Virginia are shown. Clearly, it captures major metropolitan areas correlates with road infrastructure. One would expect a strong correlation between this and the population counts in the gridded populations of data sources such as LandScan [5], WorldPop [21], and GPW [6]. However, mapping from buildings to counts of individuals are more accurately done through synthetic populations where a time-indexed map is constructed mapping people to locations. In fact, this mapping allows for a time-delineation vis-a-vis the before-mentioned data sources.

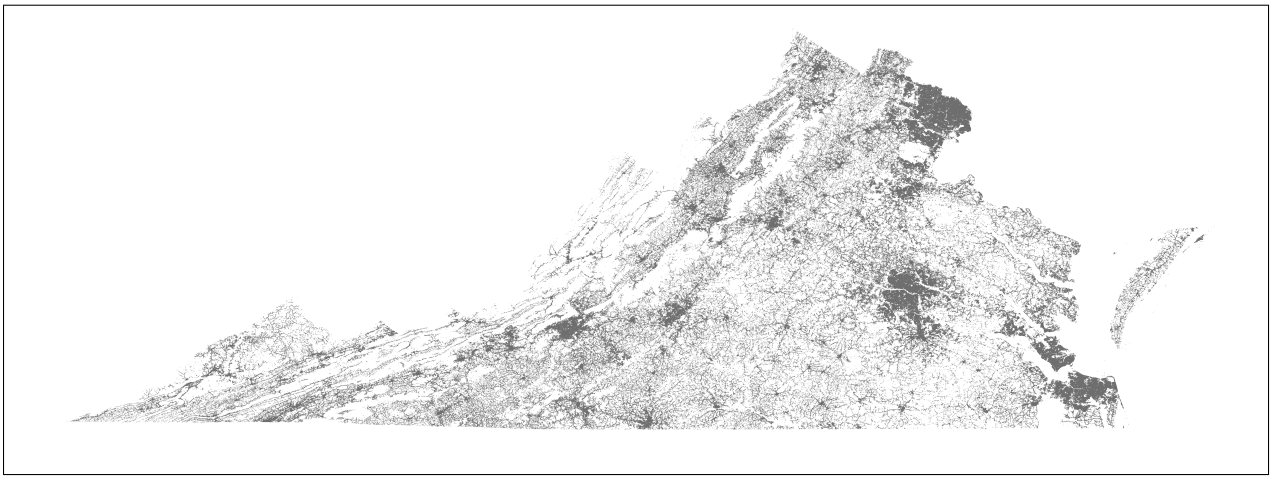

Figure 3: A point plot of the (longitude,latitude) of the centroid of each location in the NSSAC-BDB for Virginia.

Blockgroups without MS footprints. In Figure 4 we have shown the blockgroups of Virginia that do not have any building footprints as per [8], despite the Census reporting that there are several households registered at said blockgroups. Careful investigation of such cases revealed corner cases. As one example, we found a blockgroup with a marina where the single, main building structure was surrounded by many boats (ovals), leading to an incorrect classification, possibly as a harbor area, or maybe even a car cemetery, without any nearby building structure. Again, the number of such cases, and blockgroups, is very small indeed. In Table 1 we have given numbers for the of such blockgroup cases by state.

HSM: table content to be updated.

Residence location counts. This section displays counts of residence location by state and contrasts this with numbers reported by the Census for 2017 [19]. As can be seen in Table 2, the numbers agree well. Notable exceptions are Alaska and Hawaii. We assume that this is in large part determined by the source data (e.g., [8]), and assume that this will approve when updated version of such sources are made available.

Activity location counts by state.

HSM: table with counts of activity locations by state to be added; possibly broken down by designation type.

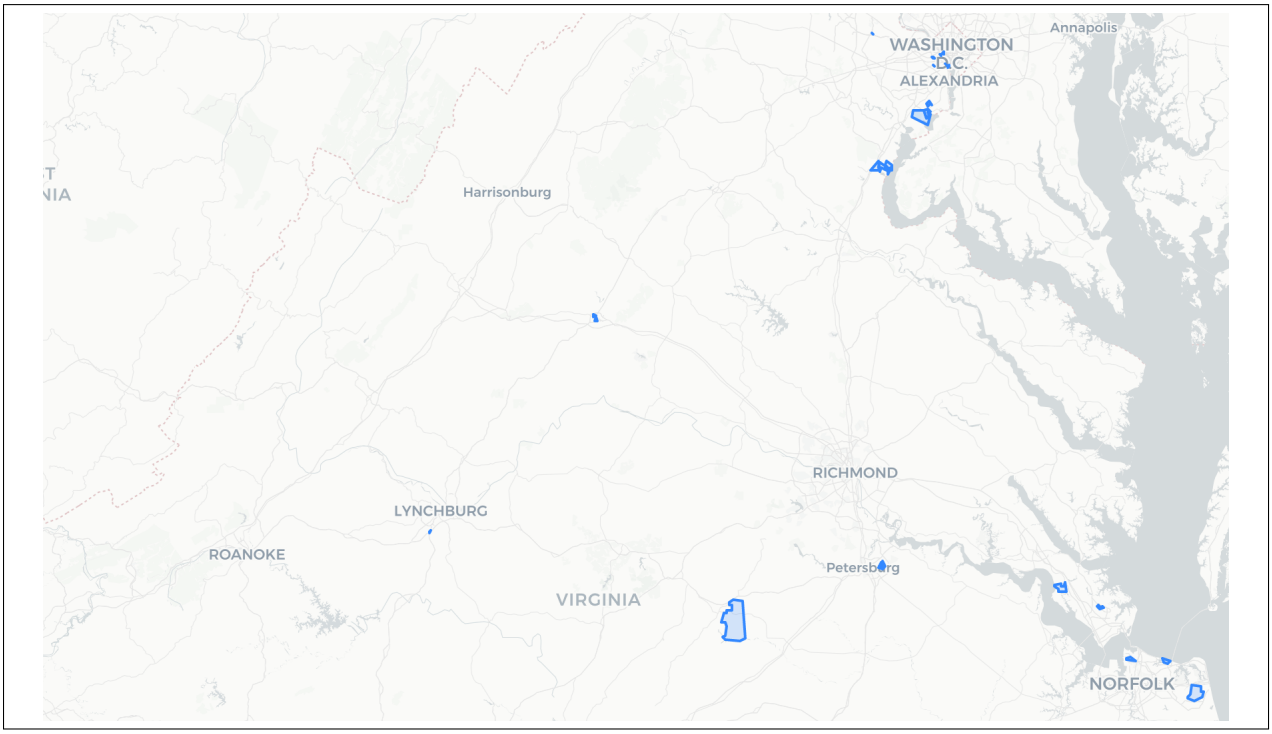

Figure 4: Illustration of blockgroups without footprints for VA.

id	fips	name	#(bgs)	#(no-res)	id	fips	name	#(bgs)	#(no-res)
1	01	Alabama	3438	4	27	30	Montana	842	1
2	02	Alaska	534	230	28	31	Nebraska	1633	6
3	04	Arizona	4178	7	29	32	Nevada	1836	5
4	05	Arkansas	2147	2	31	34	New Jersey	6320	40
5	06	California	23212	125	32	35	New Mexico	1449	4
6	08	Colorado	3532	13	33	36	New York	15463	1382
7	09	Connecticut	2585	4	34	37	North Carolina	6155	27
9	11	District of Columbia	450	16	35	38	North Dakota	572	5
10	12	Florida	11442	35	36	39	Ohio	9238	15
11	13	Georgia	5533	24	37	40	Oklahoma	2965	6
12	15	Hawaii	875	44	38	41	Oregon	2634	10
14	17	Illinois	9691	59	39	42	Pennsylvania	9740	50
15	18	Indiana	4814	8	40	44	Rhode Island	815	2
16	19	Iowa	2630	2	41	45	South Carolina	3059	11
17	20	Kansas	2351	8	42	46	South Dakota	654	1
18	21	Kentucky	3285	16	43	47	Tennessee	4125	5
19	22	Louisiana	3471	8	44	48	Texas	15811	31
20	23	Maine	1086	1	45	49	Utah	1690	5
21	24	Maryland	3926	14	46	50	Vermont	522	1
22	25	Massachusetts	4985	65	47	51	Virginia	5332	31
23	26	Michigan	8205	16	48	53	Washington	4783	31
24	27	Minnesota	4111	5	49	54	West Virginia	1592	1
25	28	Mississippi	2164	1	50	55	Wisconsin	4489	3
26	29	Missouri	4506	11	51	56	Wyoming	410	4

Table 1: The table shows the counts of block-groups by state and as well as the count of block-groups for which current modeling indicates that there are no residential buildings.

name	#(2017)	#(SI 2.0)	SI/2017	name	#(2017)	#(SI 2.0)	SI/2017
Alabama	2,254,615	2,321,508	103%	Montana	510,117	650,485	128%
Alaska	316,910	112,272	35%	Nebraska	837,246	1,107,654	132%
Arizona	3,000,102	2,545,305	85%	Nevada	1,249,298	997,731	80%
Arkansas	1,369,522	1,402,065	102%	New Hampshire	634,581	534,782	84%
California	14,172,348	12,722,095	90%	New Jersey	3,615,557	2,561,803	71%
Colorado	2,384,092	2,446,046	103%	New Mexico	937,601	968,076	103%
Connecticut	1,517,251	1,200,103	79%	New York	8,321,682	6,199,840	75%
Delaware	432,753	347,190	80%	North Carolina	4,621,314	4,382,041	95%
District of Columbia	314,702	527,026	167%	North Dakota	374,232	563,063	150%
Florida	9,438,687	9,076,003	96%	Ohio	5,201,015	5,395,308	104%
Georgia	4,280,497	3,692,972	86%	Oklahoma	1,732,215	1,936,186	112%
Hawaii	542,815	307,296	57%	Oregon	1,767,940	1,830,686	104%
Idaho	721,587	852,894	118%	Pennsylvania	5,693,469	4,804,867	84%
Illinois	5,358,908	5,530,319	103%	Rhode Island	468,235	432,545	92%
Indiana	2,884,772	3,111,133	108%	South Carolina	2,284,232	2,140,929	94%
Iowa	1,397,680	1,909,259	137%	South Dakota	392,519	582,807	148%
Kansas	1,273,536	1,515,892	119%	Tennessee	2,960,881	2,947,083	100%
Kentucky	1,983,422	2,286,791	115%	Texas	10,927,693	10,408,130	95%
Louisiana	2,060,665	1,960,525	95%	Utah	1,084,810	1,033,425	95%
Maine	742,490	722,415	97%	Vermont	335,183	341,247	102%
Maryland	2,448,604	2,675,135	109%	Virginia	3,512,201	3,789,345	108%
Massachusetts	2,896,402	2,668,923	92%	Washington	3,103,921	3,545,743	114%
Michigan	4,594,590	4,562,073	99%	West Virginia	892,048	974,270	109%
Minnesota	2,437,254	2,975,879	122%	Wisconsin	2,695,083	2,961,736	110%
Mississippi	1,322,503	1,366,950	103%	Wyoming	276,708	323,549	117%
Missouri	2,790,414	3,101,255	111%				

Table 2: The table shows estimates of housing units in 2017 [19] and residence locations exported from the NSSAC-BDB and their ratio.

5 Data Records, Export and Sharing

Data from the NSSAC-BKB/NSSAC-BDB is exported to match application and scenario needs. The example of export shown in Section 3 is just one example of this approach. The data records available as a starting point are described in Appendix Section A. We remark that some of the data in the NSSAC-BDB are derived from commercial or licence restricted data. We typically strive to have alternative data sources in such case, thereby permitting a data product to be export, albeit of somewhat lower data quality. Terms of use in our shared data generally fall under CC-BY-4.0.

6 Limitations, Future Work, and Extensions

While significant effort has gone into the modeling, development, validation and calibration, several features can benefit from improvement, either in modeling, or through availability additional types of data or data of higher resolution.

One of the key features that would be useful to include is height of the buildings represented by the footprints. There are some data products and tools at this time [Samarth reference] that allows one to model/assess building height. Having the building height, apartment count and business suite count would allow one to better assess the footprint area of each entity. It would also provide a better way to assess density of people in hallways, stairwells and elevators, aspects clearly relevant to spread of infectious diseases and the capacity for social distancing of a building complex.

For applications where only open versions of data exported from NSSAC-BKB can be used, the quality of the resulting data will often be reduced. Reasons include (i) limitations in alternative source data sets and (ii) limitations on what modeling such data sets would support when constructing for example the mappings β and η as described in Sections 2.3 and 2.5.

We have also chosen to not incorporate the Unique Building Identifier (UBID) [20] choosing instead to see how this will be accepted as a standard. We anticipate that it would be relatively straightforward to add this augmentation as BFUSA already has integrated this.

Particular extensions for augmented data include (i) addition of risk maps related to disasters involving flooding [**?**], and hurricanes [**?**], (ii) addition sun/solar ratings in support of studies involving PV cells on rooftops and renewable energy applications, (iii) addition of residential energy demand modeling (via synthetic populations linked to the building entities of NSSAC-BKB), (iv) broadband coverage in support of for example analyses of ability to work from home and/or social differences in availability, and (iv) housing costs [22].

Acknowledgements

To be added.

References

[1] BuildingFootprintUSA. BuildingFootprintUSA, 2021. Last accessed 15 September 2021.

[2] Jiangzhuo Chen, Stefan Hoops, Achla Marathe, Henning Mortveit, Bryan Lewis, Srinivasan Venkatramanan, Arash Haddadan, Parantapa Bhattacharya, Abhijin Adiga, Anil Vullikanti, Mandy Wilson, Gal Ehrlich, Maier Fenster, Stephen Eubank, Christopher Barrett, and Madhav Marathe. Prioritizing allocation of COVID-19 vaccines based on social contacts increases vaccination effectiveness, 2021. Preprint on medRxiv.

[3] Dun & Bradstreet data solutions. Last accessed: 7 Feb 2021.

[4] National Center for Education Statistics. School locations & geoassignments. Last accessed: 18 Jan 2022.

[5] Geographic Information Science & Technology, Oak Ridge National Laboratory. LandScan.

[6] Gridded Population of the World (GPW), v4.

[7] HERE Premium Streets Data set for the U.S., 2021.

[8] Microsoft. U.S. Building Footprints, 2018.

[9] Henning. S Mortveit, Abhijin Adiga, Chris L. Barrett, Jiangzhuo Chen, Youngyun Chungbaek, Stephen Eubank, Chris J. Kuhlman, Bryan Lewis, Samarth Swarup, Srinivasan Venkatramanan, Amanda Wilson, Dawen Xie, and Madhav V. Marathe. Synthetic populations and interaction networks for the U.S. Technical report, NSSAC Division of the Biocomplexity Institute and Initiative, University of Virginia, 2020. NSSAC Technical Report: #2019-025.

[10] NAICS Association. U.S. NAICS Manual.

[11] The National Center for Education Statistics (NCES). Last accessed: February 2020.

[12] Office of Energy Efficiency & Renewable Energy. Building America Best Practices Series, Volume 7.3: Guide to Determining Climate Regions by County. Accessed 18 January 2022.

[13] OpenStreetMap points of interests. Last accessed: 7 Feb 2021.

[14] PRIZM Premier. Last accessed: 7 Feb 2021.

[15] SafeGraph. Last accessed: 7 Feb 2021.

[16] SLIPO. World-scale OpenStreetMap POIs in CSV. Last accessed: June 2021.

[17] U.S. Census. 2010 Census Urban and Rural Classification. Last accessed: June 2021.

[18] U.S. Census. TIGER/Line Shapefiles. Last accessed: June 2021.

[19] U.S. Census. Vintage 2020 evaluation estimates. Last accessed: October 2021.

[20] Na Wang, Alex Vlachokostas, Mark Borkum, Harry Bergmann, and Sarah Zaleski. Unique building identifier: A natural key for building data matching and its energy applications. *Energy and Buildings*, 184:230–241, 2019.

[21] Last accessed: 17 February 2020.

[22] Zillow. Zillow's Transaction and Assessment Database (ZTRAX). Last accessed: 18 Jan 2022.

A Comments on data sources

A.1 Microsoft building footprints

In the Microsoft data, each U.S. state is represented using a GeoJSON format with two fields for each entity:[5] a polygon (geometry) and properties. There are no IDs for the entities.

B Detailed description of data of the NSSAC-BKB

This section contains precise descriptions and schema definitions for all the tables described earlier as well as for the mappings β and η.

B.1 The `Core`

In our design of the `Core`, we use a single table to store all the above data. In this core table, each entity has a `fid` that is unique for all footprints throughout the entire United States.

`Core` data fields:

Table: `nssac_buildings.core_ms_us_v1_1`:

```
id bigint,
the_geom geometry(Polygon,4326),
centroid geometry(Point,4326),
area_sqm integer,
blockgroup_id character varying(12),
urban_rural character varying(1),
initial_r_nr_class character varying(1),
source_table character varying(64),
source_id integer,
geometry_modeled boolean DEFAULT false
```

The `blockgroup_id` is the 12-digit FIPS code obtained through a geospatial join with U.S. Census polygons. Each `fid` is augmented with its footprint area (`area_sqm` square meters) and the coordinates of the footprint centroid (`centroid` WGS-84), as listed above.

Data sources used: HSM: fill in

B.2 Initial residential modeling

Refer to Section 2.2 for methods and modeling.

HSM: fill in where the mapping is stored.

Data sources used: this model used extensive data for both training, exploration, and production. In particular, this includes OpenStreetMap [13], SLIPO [16], U.S. Census geography data for administrative region boundaries 2017 [18], HERE (US/2021/Q1) [7], MS_US_BUILDINGS_V1_1 building footprints data [8].

[5] The raw data files were downloaded from [8] in September 2018 and stored in a PostgreSQL database with the PostGIS extension enabled

B.3 Data fields related to the mapping β

Table: nssac_buildings.core_bfusa_mapping_1_2

```
core_id bigint,
initial_r_nr_class character varying(1),
bfusa_id character varying(12),
bfusa_address character varying(64),
bfusa_resbus character varying(1),
bfusa_num_apt integer,
bfusa_num_ste integer,
blockgroup_id character varying(12)
```

B.4 Data fields related to the mapping η

This section provides details for tables constructed when establishing the mapping η in Section 2.5.

Table: nssac_buildings.core_here_mapping_1_2

```
id integer NOT NULL,
core_id bigint,
bfusa_id character varying(12),
here_poi_id bigint,
naics_type bigint,
blockgroup_id character varying(12),
mapping_method character varying(32),
distance numeric
```

B.5 NCES source data and curation

This section provides details on naming and table structure for the mapping established in Section 2.6. We remark the NCES is the primary federal entity that provides a wide range of data related to education within the U.S., covering all types and levels of schools in the education spectrum (e.g., public schools, private schools, and post secondary schools.) For each school it provides a wide range of information in different data products/surveys, in particular student and teacher counts. We collected geo-codes for public schools from Education Demographic and Geographic Estimates (EDGE)[6] for which source data is made available in the ESRI Shapefile format, well suited for processing and database loading. Data for student enrollment counts, teacher counts, and grade levels offered, were curated from three different surveys, (directory, staff and membership) from the Common Core of Data (CCD)[7] [4] NCES data on public schools with school ID (ncessch), geo-location (the_geom), and other fields are captured as follows.

Table: nces.edge_geocode_publicsch_1819

```
ncessch character varying(12),
leaid character varying(7),
name character varying(60),
opstfips character varying(2),
street character varying(60),
city character varying(30),
state character varying(2),
zip character varying(5),
```

[6] https://nces.ed.gov/programs/edge/Geographic/SchoolLocations
[7] https://nces.ed.gov/ccd/files.asp

```
stfip character varying(2),
cnty character varying(5),
nmcnty character varying(100),
locale character varying(2),
lat numeric,
lon numeric,
cbsa character varying(5),
nmcbsa character varying(100),
cbsatype character varying(1),
csa character varying(3),
nmcsa character varying(100),
necta character varying(5),
nmnecta character varying(100),
cd character varying(4),
sldl character varying(5),
sldu character varying(5),
schoolyear character varying(9),
the_geom geometry(Point,4326)
```

NCES data on public schools with school ID (ncessch) and number of teachers (teachers).

Table: `nces.ccd_sch_staff_1819`

```
school_year character varying(80),
fipst character varying(80),
statename character varying(80),
st character varying(80),
sch_name character varying(80),
state_agen character varying(80),
"union" character varying(80),
st_leaid character varying(80),
leaid character varying(80),
st_schid character varying(80),
ncessch character varying(80),
schid character varying(80),
teachers character varying(80),
total_indi character varying(80),
dms_flag character varying(80)
```

NCES data on public schools with school ID (ncessch) and number of students (student_count).

Table: `?????`

```
school_year character varying(80),
fipst character varying(80),
statename character varying(80),
st character varying(80),
sch_name character varying(80),
state_agen character varying(80),
"union" character varying(80),
st_leaid character varying(80),
leaid character varying(80),
st_schid character varying(80),
ncessch character varying(80),
schid character varying(80),
```

```
grade character varying(80),
race_ethni character varying(80),
sex character varying(80),
student_count character varying(80),
total_indi character varying(80),
dms_flag character varying(80)
```

Below is the curated data for all schools and partial mapping to the core.

Table: `us_locations.us_schools_1819`

```
source_id character varying(255),
source_info character varying(255),
name character varying(255),
state_fips character varying(2,
blockgroup_id character varying(12),
the_geom geometry(Point,4326),
designation character varying(255),
work integer DEFAULT 0,
school integer DEFAULT 0,
college integer DEFAULT 0,
core_id bigint,
distance numeric
```

Note: the designation field is set to "education". The partial mapping between NCES and the `Core` is stored in `core_id` and distance fields.

C Export for synthetic populations

This section provides the data structure details of the export example given in Section 3.

Schema for residential locations:

Table: `us_locations.us_res_locations_v1`

```
id bigint NOT NULL,
source_id character varying(255),
longitude numeric,
latitude numeric,
altitude integer DEFAULT '-1'::integer,
blockgroup_id character varying(12),
admin1 character varying(2),
admin2 character varying(3),
admin3 character varying(6),
admin4 character varying(1),
area_sqm integer NOT NULL,
pub_pk integer,
pub_kg integer,
pub_01 integer,
pub_02 integer,
pub_03 integer,
pub_04 integer,
pub_05 integer,
pub_06 integer,
```

```
pub_07 integer,
pub_08 integer,
pub_09 integer,
pub_10 integer,
pub_11 integer,
pub_12 integer,
the_geom geometry(Point,4326),
```

Schema for residential locations:

Table: us_locations.us_act_locations_v1

```
id bigint NOT NULL,
source_info character varying(255),
source_id character varying(255),
longitude numeric,
latitude numeric,
altitude integer DEFAULT '-1'::integer,
blockgroup_id character varying(12),
admin1 character varying(2),
admin2 character varying(3),
admin3 character varying(6),
admin4 character varying(1),
designation character varying(255),
naics_type bigint,
work integer DEFAULT 1,
shopping integer DEFAULT 0,
school integer DEFAULT 0,
other integer DEFAULT 1,
college integer DEFAULT 0,
religion integer DEFAULT 0,
the_geom geometry(Point,4326)
```