The purpose of this document is to provide technical background and details to complement the HDC visualization tools, and data downloads. We have made every attempt to provide this background in plain language for casual users while also providing sufficient detail for those who may want to evaluate or replicate. If you have questions that are not addressed here, please contact info@hawaiidata.org for more information.

Regions

Background

A primary challenge in linking multiple data sets to provide a comprehensive overview of well being in Hawaii is that various sources provide data using different region schemes (the details of the region schemes used to report data in the state are provided below). As such, for some data sources it is necessary to convert data from one region type to another in order to be able to compare data points across well being domains.

Additionally, we understand that users may have different needs with regards to the region scheme used to display the data. As such, it is also critical to be able to switch between region schemes based on use case.

In order to accommodate multiple data sets and be able to provide for more flexibility and utility across a broad range of stakeholders, HDC has employed multiple approaches to estimate indicator values across different region schemes. While this serves to minimize barriers to comparing various data points, there are limitations to converting data from one region to another. These limitations are also provided below.

Region Types

In this section, we describe each of the region types used to create the HDC dataset. Some are regions that are reported in the various data tools, while others are source regions that are used in the region conversion process (i.e. aggregation and/or crosswalking - described below).

Census Tract - Boundaries developed by the U.S. Census Bureau in cooperation with local government. Tracts are generally between 1,200 an 8,000 in population (optimally 4,000), and can be split as the population within the tract grows.[1]

Zip Code and ZCTA - Zip codes are maintained by the US Postal Service (USPS) to facilitate the delivery of letters and parcels to specific addresses. While the USPS does not define zip codes based on a regional area, geographic boundaries can be created by drawing a perimeter line around all addresses with the same zip code. In response to this limitation, the U.S. Census Bureau has created Zip Code Tabulation Areas (ZCTA) which serve as more consistent geographic boundaries that (for the most part) correspond to the postal addresses that carry that zip code.[2] Not all postal zip codes refer to a specific region - for example post office (P.O.)

[1] More information can be found at the U.S. Census Bureau site here.
[2] Refer to this document for a decent overview of the differences between zip code areas and ZCTAs.

boxes. Additionally, zip codes may be exclusive to a specific entity, such as a business, education institution or government agency. Therefore only a subset of all zip codes (with corresponding ZCTAs) in an area can be used to describe geographic regions.

For Hawaii, in addition to PO box zip codes, those specific to Federal and U.S. Military agencies are excluded:

- 96850 - Prince Kuhio Federal Building
- 96853 - Joint Base Pearl Harbor-Hickam
- 96857 - Schofield Barracks
- 96859 - Tripler Army Medical Center
- 96860 - Joint Base Pearl Harbor-Hickam
- 96863 - Marine Corps Base Hawaii

Lastly, while zip codes and ZCTAs can be convenient reference geographies because they are generally well known, they are not ideal for population research. One specific reason for this is that population sizes can vary dramatically, which is problematic for region to region comparisons. For example, in Hawaii in 2016 census ZCTA population estimates ranged from over 76,000 (96797 - Waipahu, CCH) to 70 (96751 - Kealia, Kauai).[3]

State of Hawaii Department of Health (DOH) Communities - In order to provide more granular estimates of health behaviors and outcomes from the annual Behavioral Risk Factor Surveillance System (BRFSS) survey, the DOH has created 24 communities based on combinations of respondent reported **zip codes**. Details of this regional scheme are detailed in the reports provided here.

State Legislative Districts - The State of Hawaii is organized into 25 upper chamber (senate) and 51 lower chamber (house) districts. These districts can be further subdivided into **Election Precincts**. Details of these geographies can be found here.

State of Hawaii Department of Education (DOE) School Complex Areas and School Catchment Areas - The State DOE is organized around 15 Complex Areas, which are further divided into 40 Complexes which include one high school, and the intermediate and elementary schools attached to it. In turn, each school has a geographic Area that it serves - for high schools this is the same as the Complex, while for intermediate and elementary schools this may represent a sub area of the complex. This preview document provides maps of each type of DOE geography.[4]

State Department of Business, Economic Development, and Tourism (DEBDT), Office of Planning Census Districts - Previously, the State DEBDT created “Districts” based on 2000 U.S. Census tracts. It is unclear the extent to which these maps are currently in use, with the

[3] Interested readers are referred to this article for more details regarding the limitations of zip codes for epidemiological research.
[4] Additionally, detailed GIS files can be found here, under ‘Administrative & Political Boundaries.’

exception of State Department of Human Services (DHS), which uses this region scheme to organize child abuse data (see below). Detailed maps can be found here, and a file (internally generated) providing correspondence between 2000 Census tract and District can be downloaded here.

Aggregation

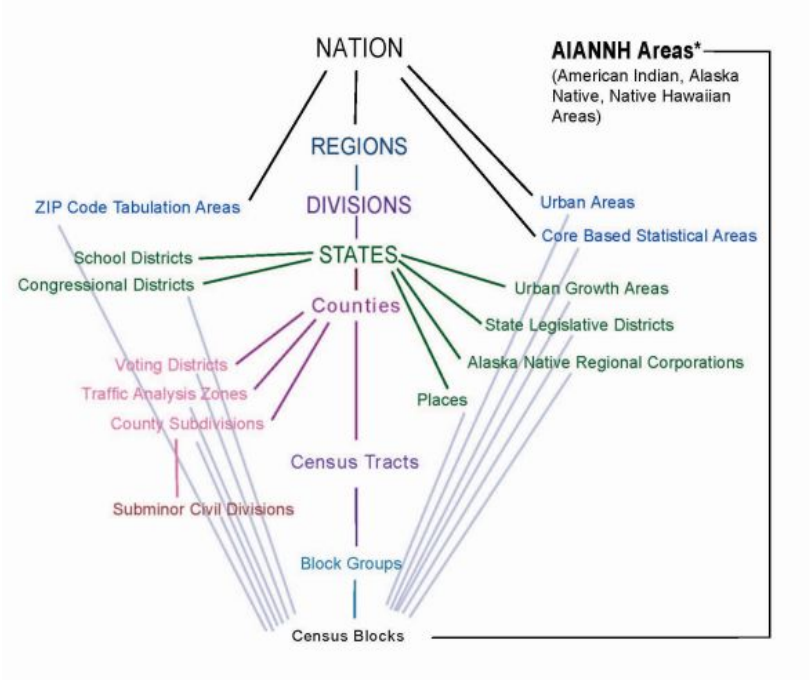
In some cases, when there is a hierarchical relationship between region types, converting smaller region types to larger region types can be achieved by simple aggregation - for counts, this involves summing the values of multiple small areas into one larger area. When combing proportional variables such as percentages, incident rates or ratios, population weighted averages must be calculated. This involves multiplying each sub-region value by the relevant population count for that region, summing the resulting values of all of the sub regions, and then dividing that number by the sum total population of the sub regions. This approach is only appropriate when the subregions correspond evenly to the higher level regions. For example:

- Census Tract → County
- Zip Code or ZCTA → DOH Community
- Election Precinct → State Legislative District
- DOE School Complex → DOE School Complex Area

Additionally, for the purposes of crosswalking (see below) it is possible to *disaggregate* higher level *proportional* (i.e. non-count) values across subordinate regions. Essentially, this is achieved by applying the proportional value of the larger area to all of the sub-regions within that area. The sub-region values can then be used to crosswalk to a different region scheme. However, caution is warranted with this kind of disaggregation, as this assumes homogeneity across the larger region. In other words, it is not appropriate to draw conclusions regarding the sub region values and caution should be used in interpreting the crosswalked values generated from disaggregation (see below).

Crosswalking

Unfortunately, as the illustration below illustrates, there are many region types that are not hierarchically related meaning that aggregation or disaggregation is not possible:



Source: TIGER/LineR Shapefiles - Technical Documentation 2012, p 3-12[5]

The illustration above describes all of the region types reported by the census. The middle vertical bar represents a hierarchy of region types that have perfect correspondence: The Nation is composed of States, which can be further broken down into counties. Further, Census tracts are drawn to fit within counties, and Census block groups and blocks fit within tracts.

However, there are other important region schemes reported by the Census, such as state legislative districts, school complexes, voting (election) districts, and ZCTAs that do not fit neatly into this scheme. There are some exceptions, such as election districts falling within counties, but otherwise this presents a challenge for pulling together two or more datasets with incompatible region schemes.

In order to address these challenges, we employed crosswalking methodology as one tool for converting data types into common region schemes. The Missouri Census Data Center has developed the MABLE/Geocorr14: Geographic Correspondence Engine, which provides correspondence or crosswalk tables between any two Census region types. Essentially, the crosswalk tables indicate the percent overlap of the population between two region types. For example, a portion of the table for crosswalking Census tract values to house districts is provided below:

Census Tract	House District	County	Population	Allocation Factor

[5] https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2012/TGRSHP2012_TechDoc.pdf

15001020100	1	Hawaii HI	5,213	100.0%
15001020202	1	Hawaii HI	2,070	80.6%
15001020202	3	Hawaii HI	498	19.4%
15001020300	1	Hawaii HI	2,485	63.2%
15001020300	2	Hawaii HI	1,449	36.8%
15001020400	2	Hawaii HI	3,294	100.0%

Looking at rows two and three we see tract 15001020202 covers house districts 1 and 3. The 'Population' column assigns 2,070 individuals from that tract to district 1 and 498 individuals to district 3, representing 80.6% and 19.4% of that tract respectively (Allocation Factor). Converting non-count variables from tract to house district is accomplished by multiplying the value of the tract by the population for each row, summing all of the products for each house district, and then dividing by the total population for that district. For count variables, the value is multiplied by the allocation factor, and then summed for each district.

Much like the limitations to disaggregation described above, a primary limitation of the crosswalking approach is it assumes the value assigned to the source region (e.g. tract 15001020202) is homogenous across the entire tract (i.e. the approach assumes the population of the tract residing within district 1 is the same as the population residing in district 3). For this reason, care should be taken when interpreting values generated from crosswalk tables. As a general rule, crosswalking is more reliable for larger population areas.

Data Sources

Center for Neighborhood Technology (CNT) - Housing and Transportation Affordability Index

Domains:

- Transportation

Refer to our Data Dictionary for the data points that rely on this source.

- The Center for Neighborhood Technology (CNT) has created the Housing and Transportation Affordability Index, which provides multiple relevant community-level transportation metrics. The Index relies on ACS, U.S. Census Longitudinal Employer-Household Dynamics and CNT AllTransitTM data, and current transportation

system and transportation behavior models. Currently, only 2015 estimates are available.

Data for Hawaii were downloaded at the tract level, and crosswalked using the methods described above. More details related to these indicators can be found in CNT's methodology report.

Location Inc. (property and violent crime data)

Domains:

- Safety & Security

Refer to our Data Dictionary for the indicators that rely on this source.

We use data from Location Inc. to provide estimates of property and violent crime. Location Inc. uses a proprietary algorithm to integrate crime data from multiple sources (primarily county and municipal police records, and the FBI Uniform Crime Reporting (UCR) progarm) to estimate crime rates at the Census tract level. Their estimation approach is also useful for small areas as it is not as vulnerable to anomalies due to small populations. Details of Location Inc's methodology can be found here.

(Unfortunately, due to the nature of our license data agreement, we are not able to share this data set beyond what is available in our visualization tools. Those interested in more information are encouraged to contact Location Inc.)

State of Hawaii Department of Education

Domains:

- Education

Refer to our Data Dictionary for the indicators that rely on this source.

DOE makes public school data available in two ways. The online Strategic Plan Dynamic Report, which provides school outcome measures by DOE complex area, and the School Status & Improvement Reports, which provides outcomes by individual school. Census Block population estimates were converted to School Complexes using the US EPA's Dasymetric

Toolbox for ArcMap. Dasymetric mapping techniques were utilized with land cover and land use from the NOAA Office for Coastal Management's C-CAP High-Resolution Land Cover and Change dataset as ancillary sources. When a Census Tract is split between different School Complex Areas, information about land cover and land use was used to proportionally allocate Census Block estimates to each School Complex Area. School-level metrics were also converted to Census tracts. For each tract, this was accomplished by calculating the simple point average of the school catchment area values that overlap the tract.

State of Hawaii Department of Health

Domains:

- Health

Refer to our Data Dictionary for the indicators that rely on this source.

Currently, the Department of Health publishes the results of the yearly Behavioral Risk Factor Surveillance System survey at the State, County, and Community level in the Hawaii's Indicator Based Information System (IBIS). The DOH describes the process of deriving the sub-county communities:

The four counties of the State of Hawaii are subdivided into communities or sub-areas by grouping zip codes such that each group has at least one school complex. These zip codes were given names as shown in the Community column in the following community maps by county.

A detailed breakdown of the DOH communities, and the zip codes that compose them, can be found in the yearly reports posted here. Additionally, we have created a shapefile for DOH communities that can be downloaded here.

BRFSS suppresses data for communities where the total number of responses to the question is less than 50 or the relative standard error is greater than 0.30.

Zip code-level data was then used to calculate DOH Community values for non-BRFSS indicators, using American Community Survey zip code adult population estimates to derive weighted averages for each community.

State of Hawaii Department of Human Services

Domains:

- Safety & Security

Refer to our Data Dictionary for the indicators that rely on this source.

At this time, we include one data point from DHS - Confirmed Cases of Child Abuse (unique count), which was extracted by year and community from the Child Abuse and Neglect Reports. DHS provides regional counts of child abuse cases based on the 2000 Census District Maps created by the Department of Business, Economic Development and Tourism (DEBDT), and based on clusters of Census Tracts. Some child abuse cases are not able to be assigned to a Census District and are classified as "unspecified". These unspecified cases are accounted for at the State and County level, but not at the sub-county level.

The Child Abuse and Neglect Reports provide counts of unique cases, not rates of abuse. To calculate abuse rates, it was necessary first to aggregate tract level population estimates for individuals 17 years of age and under to the DEBDT Census District scheme. Then the abuse rate by DEBDT Census District and year was calculated based on the following formula:

*(# of cases (unique count) / Census District under 17 years old population) * 1,000*

From there, estimates were derived by applying the district rates to corresponding tracts, and then crosswalking to the regions provided in the visualization tools.

State of Hawaii Office of Election Management

Domains:

- Civic Engagement

Refer to our Data Dictionary for the indicators that rely on this source.

The Hawaii Office of Elections provides general election data by precinct by year. In addition to voting outcomes, this data includes two important numbers: the number of registered voters in a precinct, and the number of ballots cast. Dividing the number of ballots cast by the number of registered voters provides an indicator of registered voter turnout (i.e. the percent of registered voters in the precinct who voted).

Voting age population (VAP) turnout was calculated by dividing the number of ballots cast by the VAP. VAP refers to the population 18 years and older, and was sourced from the American Community Survey (ACS).

Further, voting precincts serve as sub regions within State legislative districts (both upper and lower), which makes it possible to calculate voter turnout by legislative district. Correspondence tables can be found here.

U.S. Census Bureau - American Community Survey (ACS)

Domains:

- < Demographics >
- Economic Opportunity
- Education
- Housing
- Health
- Transportation

Refer to our Data Dictionary for the full list of indicators that rely on this source.

The United States Census Bureau collects population data for the entirety of the United States every ten years, with the most recent census conducted in 2010. Participation in the census is mandatory.

The Census Bureau also employs surveys to collect data on an ongoing basis through the American Community Survey (ACS), American Housing Survey (AHS), Current Population Survey (CPS), Annual Business Survey (ABS) and others. Currently, data for the HDC is sourced entirely from the ACS and is retrieved using the Census API and the American Factfinder tool.

Sampling and Estimation. Through the ACS the Census conducts approximately 10,000 interviews in Hawaii (housing units and group quarters) each year, with updated data being released the following year: 1 year estimates in September and 5 year estimates in December. The 1 year estimates provide state and county level values[6], and are based on data collected during the stated calendar year (i.e. 2017 1-year estimates are based on data collected in 2017). The 5 year estimates are provided for sub-county regions (e.g. block, tract, ZCTA, state legislative districts, etc.), and are based on data collected during the stated year and the four years preceding (i.e. 2017 5-year estimates are based on data collected from 2013 through 2017). This is required in order to achieve sufficient sample sizes to estimate values for small areas.

[6] Generally, the cut-off for 1-year estimates is regions with populations of 65,000 or higher.

So this presents an important caveat regarding the Census data we report: Although data points are offered yearly, it is not appropriate to interpret changes year to year, because of the overlaps in the data. For example, 2017 estimates are based on data collected from 2013-2017, while 2016 estimates are based on data collected from 2012-2016. So, both estimates share the same data from four years (2013-2016). In order to appropriately examine trends using ACS data, it is necessary to use non-overlapping years. So for 2017 (2013-2017), the most recent previous data point appropriate for comparison would be 2012 (2008-2012).

Internally Generated Indicators

We are currently working on multiple indicators that are not currently available from other sources. As we derive these indicators, the methodology we employ will be provided here.

Error Estimates

All measurement involves some level of error. Measurement of people is especially prone to error for multiple reasons. As such, indicator values should be considered estimates, and when available, margins of error should be consulted to understand the likelihood that the indicator value provided is representative of the phenomena of interest.

Essentially, the *margin of error* (MOE) represents the maximum difference between the reported value, and the range of potential values that would occur if the same measurement were conducted multiple times. MOE is usually reported as (+/-): For example 60% +/-5%, means that the range of expected values is between 55% and 65% (when values are expressed as a range, this is referred to as the *confidence interval*).

Currently, the HDC is only reporting MOE's directly reported by the data source. However, since we have employed various forms of crosswalking and aggregation in order to align indicators along specific region schemes, many of the values contained in our dataset do not have source MOE's. While there are methods for aggregating MOEs (for example, from the Census), we believe these weight too heavily the number of input regions used in the aggregation, and represent less the population size of the regions. So, for example, a region calculated from four source regions will have a higher MOE than one composed of three regions simply because of the number of input regions, even if the average MOE across the input regions is the same.

We are currently working to develop a solution to this challenge. In the meantime, we encourage those interested to consult the source region margins of error when assessing a region crosswalk or aggregation. Additionally, please contact us at info@hawaiidata.org, if you have further questions.

Validation of Estimates

All data points provided in our visualization tools are estimates of regional values. For those regional estimates that are the result of transformations from the original source, we have worked to validate and error check those values as best as possible. For validation, we conducted population-weighted aggregations to derive county and state-level estimates. Those estimates were then compared against official State and county (if reported) values for consistency. Further, for each indicator and region type we conducted analyses to identify outliers in the indicator set. Any identified outliers where are compared against relevant source values to determine if there was an error in the calculation.