UCDAVIS
Graduate Program of
Environmental Policy and Management

Practicum Project Report

Incorporating Racial and Ethnic Considerations in the Prioritization of California's Maximum Contaminant Levels (MCLs) for Drinking Water

Federico Pacheco Oreamuno
MS Graduate Student
Environmental Policy and Management

Practicum project conducted at the Regulatory Development Unit,
Division of Drinking Water, State Water Resources Control Board,
Sacramento, California

September 2024

Table of Contents

1. Introduction

1.1 Environmental Policy and Management (EPM) Practicum

The practicum is an integral part of the Graduate Program in Environmental Policy and Management (EPM) at UC Davis and an essential component of the program's Comprehensive Exam. It provides students with an opportunity to gain practical experience in a professional setting by working at a partnering organization off-campus on a policy-related project or analysis for a minimum of 180 hours while applying relevant knowledge and skills acquired throughout the EPM program.

1.2 Regulatory Development Unit of the Division of Drinking Water

The EPM practicum was carried out at the Regulatory Development Unit (RDU) of the Division of Drinking Water (DDW) at the State Water Resources Control Board (SWRCB) in downtown Sacramento. Among many other tasks, the RDU is responsible for leading the rulemaking process to establish California's maximum contaminant levels (MCLs) for drinking water. MCLs are enforceable primary standards[1] that water systems supplying domestic water to the public must meet to protect human health.

1.3 Project Background and Objectives

In 2023, the SWRCB published the 2023–2025 Racial Equity Action Plan (REAP), which included strategic goals to address racial inequities and integrate racial equity considerations into the agency's decision-making processes (SWRCB, 2023).

One action under Goal 1b directed the DDW and the RDU to "incorporate racial equity analysis when developing maximum contaminant levels using available data and as data and methods allow" (SWRCB, 2023).

Currently, there are no agency guidelines or established procedures for conducting racial equity analyses. Additionally, datasets within the RDU or the DDW do not support spatial analysis of demographic data in relation to the locations of public water systems (PWS) and the occurrence of regulated contaminants. Furthermore, California statute, outlined in Health and Safety Code (HSC) sections 116365(a) and 116365(b), specify the criteria for MCL development but do not mandate racial equity analysis.

According to HSC section 116365(a) and 116365(b), MCLs must be set as close to the corresponding public health goals (PHGs) as is technologically and economically feasible. A PHG

[1] Primary drinking water standards are enforceable standards designed to protect human health and include all MCLs but also other contaminants that are regulated through treatment techniques. Secondary drinking water standards are established to improve drinking water aesthetics (color, taste, odor, etc.) for consumer acceptance.

is a non-enforceable, health-based standard established by the Office of Environmental Health Hazard Assessment (OEHHA). It indicates the concentration of a given contaminant at which no significant health risk or adverse health effects are expected if people consume water containing the contaminant over a lifetime of 70 years.

Instead of integrating racial equity analysis into the regulated MCL development process, it would be more practical to incorporate racial equity considerations into the prioritization process for selecting chemicals. The development or revision of MCLs often requires years and extensive staff resources due to the significant variability in contaminants' health impacts, prevalence in California water sources, and treatment and control costs. As a result, only a small fraction of regulated contaminants or new contaminant candidates can be addressed at any given time.

Although no formal prioritization schemes are in place, the RDU team has evaluated various criteria to prioritize contaminants, such as improved detection limits, expected health benefits, and contaminant-specific occurrences. The RDU also continues to explore additional criteria to enhance drinking water safety when selecting contaminants for MCL revision.

With the goal of integrating racial equity considerations into the prioritization of MCLs, three objectives were proposed for this practicum project:

1. Conduct a literature review to identify how racial equity analysis is applied in the field of drinking water.

2. Develop a spatial method to create a dataset that links racial, ethnic, and age demographic information to California's drinking water systems. This dataset will provide additional insights for analyses such as contaminant occurrence assessments.

3. Use the dataset to recommend potential approaches for incorporating racial and ethnic data into MCL prioritization.

2. Literature Review

2.1 Peer-reviewed Literature

A recent study by Rosenblum et al. (2024) proposed a population-based prioritization approach for evaluating and ranking both regulated and unregulated chemicals in drinking water, aimed at informing the development and revision of MCLs at the U.S. national level. This approach involved calculating a "health-reference normalized" hazard index (HI) for each chemical by dividing measured environmental concentrations[2] (MEC) detected at PWS by the most conservative health reference level (HRL) available for each chemical (Rosenblum et al., 2024). Most HRL values were selected from the Regional Screening Level Tables for Resident Tap Water[3], which are non-regulatory screening values developed by the U.S. Environmental Protection Agency (U.S. EPA) from toxicity data.

Mean HI values were then calculated for every chemical reported at each PWS from the MEC data. Concentrations below the chemical's detection limit (DL) were set at zero or at one half of the DL depending on the specific analysis. The authors focused primarily on HI values greater than one, which indicate that the concentrations measured at the PWS exceeded the corresponding HRLs. Using the population data associated with each PWS, a population-based ranking of chemicals was developed by calculating the percentage of people in the U.S. exposed to HIs greater than one. Figure 1 shows the ranked results for 50 chemicals. The top 10 constituents included three inorganic chemicals (hexavalent chromium, uranium, and arsenic) and seven disinfection byproducts.

[2] Chemical concentrations were obtained from values reported by public water systems across the U.S. serving at least 500 people from 2010 to 2020.

[3] Screening Levels are risk-based concentrations derived from exposure equations and toxicity data by U.S. EPA's Superfund program under the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) of 1980. These concentrations are considered to be universally protective over a human lifetime. https://www.epa.gov/risk/regional-screening-levels-rsls

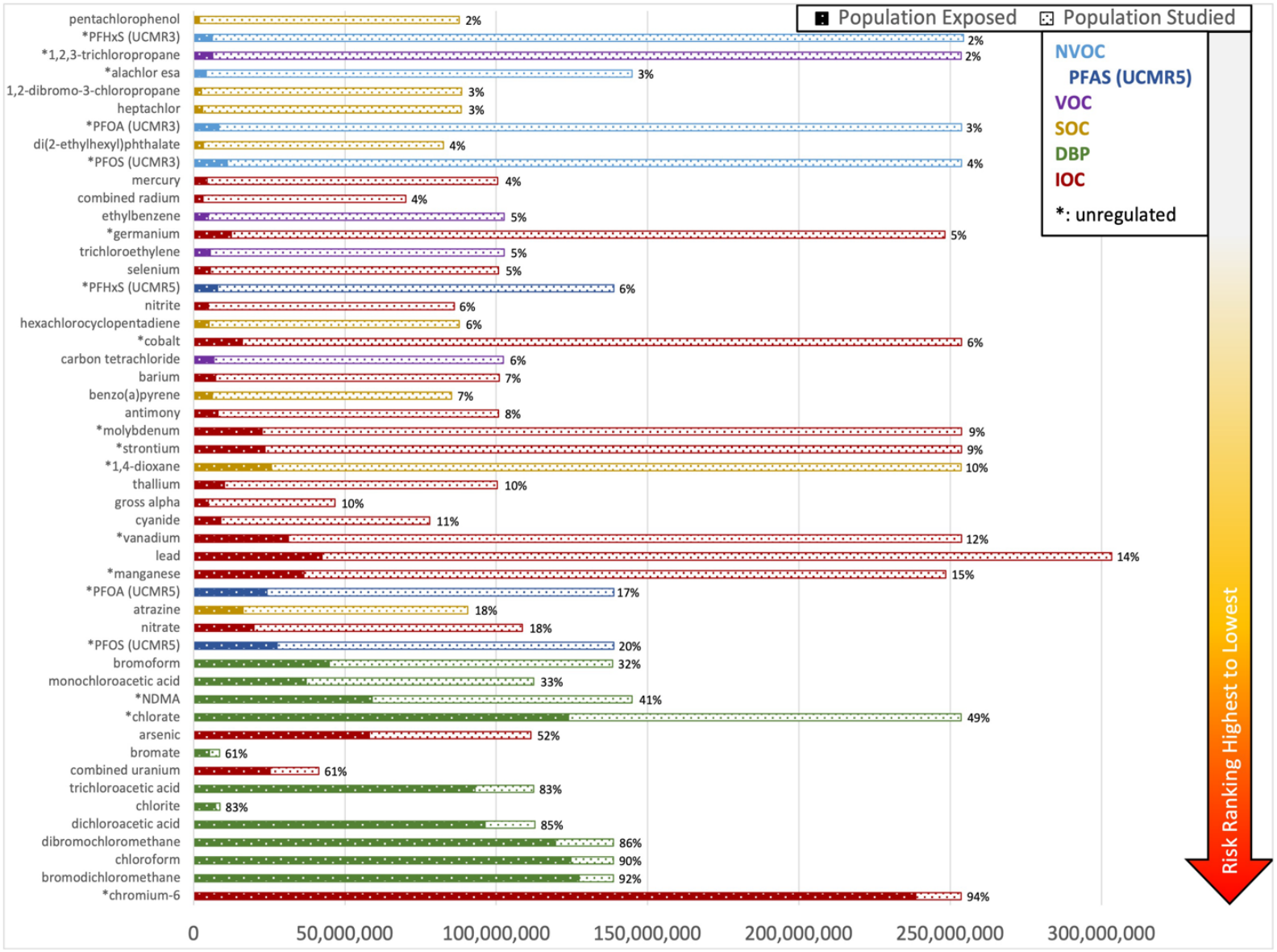

Figure 1. Prioritization and ranking of chemicals in drinking water based on a population risk approach. For every chemical, occurrence concentrations at PWS were normalized by a health reference level (HRL) to calculate individual hazard indices (HI). Population served by PWS with HI greater than one were tallied to determine exposure estimates (Figure 2 in Rosenblum et al., 2024).

The approach proposed by Rosenblum et al. is essentially a variation of similar screening methods already in use by U.S. EPA. Generally, they involve calculating the ratio of a health-based reference concentration to actual measured concentrations to assess the relative hazard of a chemical. For example, during the development of Contaminant Candidate List[4], U.S. EPA evaluates and compares the relative risk of each chemical by calculating the ratio of a non-regulatory HRL to the corresponding environmental concentration (Rosenblum et al., 2024).

While the authors only considered overall population percentages as the criterion for ranking contaminants, this analysis could have been expanded to include demographic factors such as race and ethnicity, provided that the relevant data is assigned to each PWS. No examples were found of any studies that included racial considerations in the prioritization of drinking water contaminants.

The topic of racial and ethnic disparities related to drinking water quality has been explored in the academic literature (Balazs & Ray, 2014; Switzer & Teodoro, 2017; McDonald & Jones, 2018; McDonald et al., 2022; Scanlon et al., 2023; Pace et al., 2022; Mueller et al., 2024; Uche et al., 2021; Allaire & Acquah, 2022; Acquah & Allaire, 2023). These studies primarily investigated the relationships between the occurrence of regulated contaminants, as indicated by health-based water quality violations in (PWS), and various demographic and socioeconomic factors of the communities served by those PWS (McDonald et al., 2022). Consequently, the racial and ethnic disparities reported in these studies were attributed to differences in water quality and PWS infrastructure, such as the size of the water system.

Generally, these reports identified disproportionate exposure to drinking water contaminants among low-income groups and communities of color, highlighting environmental justice concerns (Acquah & Allaire, 2023). Because these assessments required linking demographic information geographically to PWS locations, several studies also emphasized the crucial need for precise service area boundaries for all PWS. Such detailed boundaries would enable high-resolution and more reliable geographical analyses at the water system boundary level (McDonald et al., 2022; Mueller et al., 2024; Uche et al., 2021).

None of the reviewed articles considered the potential implications of setting new maximum contaminant levels (MCLs) or modifying existing ones on racial and ethnic disparities related to drinking water. However, Allaire and Acquah (2022) demonstrated that changes to MCLs were often followed by increases in water quality violations as PWS adapted to new regulations. MCLs are regulatory benchmarks designed to protect human health to the highest degree that is economically and technologically feasible. Consequently, MCLs are not viewed as tools for addressing socioeconomic inequities. Instead, policy and financial interventions are directed towards improving or modernizing PWS to ensure compliance with MCLs (McDonald & Jones, 2018).

[4] The Contaminant Candidate List (CCL) consists of a list of unregulated drinking water contaminants that, based on their current or anticipated occurrence in PWS, may require future regulation. The Safe Drinking Water Act mandates U.S. EPA to issue a new CCL every five years.

Since no research papers were found addressing regulatory prioritization of chemicals in drinking water with respect to racial and ethnic considerations, it was necessary to examine studies that assessed water quality violations to find examples of racial equity analyses.

Using a regression model, Switzer and Teodoro (2017) investigated the effects of class, race, and ethnicity on drinking water health-based violations represented by non-compliance with MCLs or treatment techniques. The dataset included 12,972 local government-owned utilities serving populations of 10,000 people or more across the United States between 2010 and 2013 (Switzer & Teodoro, 2017). Demographic data was obtained from the U.S. Census Bureau at the city level. As racial and ethnic variables, the model included percent Black population and percent Hispanic population, as well as a series of other variables describing income and education levels of the community.

The main finding was that racial and ethnic factors (percent Black or percent Hispanic) were statistically significant only in relatively poor communities. This indicated that, on average, poor Hispanic or Black communities were more likely to receive water from utilities prone to violations compared to similarly poor communities that were predominantly White (Switzer & Teodoro, 2017). In contrast, for relatively well-off communities, race and ethnicity did not correlate with the number of water quality violations. The authors explained that poverty amplified the effects of race and ethnicity when predicting health-based violations (Switzer & Teodoro, 2017).

In a more recent study, Allaire and Acquah (2022) analyzed data from approximately 1,693 community water systems[5] (CWS) in California between 2000 and 2018 to investigate how socioeconomic factors such as race, ethnicity, and income relate to compliance with drinking water regulations. Their hypothesis was that CWS serving communities of color and low-income groups were more likely to experience health-based violations due to capacity limitations, thereby leading to disparities (Allaire & Acquah, 2022; 2023).

The authors also noted that few studies had previously addressed health-based drinking water inequities at the water system level, primarily due to the lack of geographic data about the water systems' service areas. As a result, most of these earlier investigations used demographic data aggregated over larger geographic units beyond water system service areas such as counties or zip codes, which tend to mask differences in water quality (Allaire & Acquah, 2022).

The service areas of the selected CWS were obtained from datasets provided by the Tracking California's Water Boundary Tool[6] (WBT) and tribal governments. Demographic information, including total population, racial and ethnic composition, median household income, and housing density, was sourced from the U.S. Census Bureau at the zip code level. These data were interpolated to the CWS boundaries using an areal weighting method. The authors acknowledged the limitations of areal interpolation, noting that population errors might occur for water systems with very small service areas. Consequently, the final dataset of 1,693 CWS excluded hundreds of systems with fewer than 200 connections (Allaire & Acquah, 2022).

[5] U.S. EPA defines a community water system (CWS) as a system that delivers water to at least 25 people year-round. CWS comprise a subset of public water systems (PWS), distinguished by the requirement to provide water year-round compared to a minimum of 60 days per year for PWS.

[6] The Water Boundary Tool is the predecessor dataset to the current California Drinking Water System Boundaries Layer (SABL) dataset maintained by the DDW of the SWRCB, which was used for this project.

The same authors followed this study with a second publication (Acquah and Allaire, 2023), which expanded the demographic factors used in the regression models (e.g., age categories) and employed a population-based interpolation approach instead of areal weighting. Most findings aligned closely with those of the first study. The inclusion of age information provided additional nuanced insights, revealing that, for example, young children were proportionally more affected than adults by the occurrence of CWS arsenic violations (Acquah and Allaire, 2023).

Uche and colleagues found that, on average, cumulative cancer risk from drinking water supplied by CWS in California and Texas was higher for water systems serving areas with predominantly Hispanic and African American populations (Uche et al., 2021). For their analysis, they interpolated demographic information at the U.S. Census Bureau tract level to the service area boundaries of CWS using areal weighting. The authors noted significant limitations when using tract areas for small CWS serving fewer than 500 people and recommended using block-level information to better match the interpolated population numbers with the values reported by each CWS (Uche et al., 2021). For small CWS, they found low correlation between interpolated population numbers and the population values reported by the water systems.

Perhaps the most comprehensive study to date is the one by Pace et al. (2022), which used linear regression models to examine sociodemographic disparities related to health-based drinking water violations in California from 2011 to 2019. In addition to CWS service areas, this study also assessed domestic well areas to include the estimated 1.3 million Californians who rely on domestic wells for their drinking water (Pace et al., 2022). The authors created an unprecedently high-resolution dataset of demographic information at the subblock level for both CWS and domestic well service areas using various population-based weighting approaches (Pace et al., 2022). For drinking water contaminant occurrence—focusing on arsenic, nitrate, and hexavalent chromium—they calculated nine-year concentration averages for each CWS and used estimates derived from the SWRCB's Groundwater Ambient Monitoring and Assessment (GAMA) program for domestic well areas. While their findings are too extensive to summarize here, the overall conclusion, consistent with other studies, was that communities of color are disproportionately affected by water quality issues in areas served both by CWS and domestic wells.

The study by Mueller et al. (2024) offers another example of high-resolution geographical analysis, examining disparities in New Jersey at the census block group level in relation to the occurrence of polyfluoroalkyl substances (PFAS) from 2019 to 2021. CWS were characterized based on the number and type of overburdened communities[7] located within their boundaries. They reported that, statistically, overburdened communities were more likely to be exposed to PFAS compounds than other communities.

The studies included in this review demonstrated that linking racial and ethnic information to PWS service area boundaries is essential for conducting racial equity analyses. This approach can reveal water quality disparities by examining the racial profiles associated with water systems violating MCLs. However, none of these studies explored how MCL prioritization and development could be informed by racial or ethnic factors or whether MCLs could be used to mitigate drinking water

[7] The New Jersey Department of Environmental Protection defines overburdened communities as any census block group where at least 35% of the households qualify as low income, or at least 40% of the residents identify as people of color, or at least 40% of the households have limited English proficiency. A household is classified as low-income if it is at or below twice the poverty threshold determined annually by the U.S. Census Bureau.

inequities. While identifying the locations and severity of water quality disparities is essential for addressing them, the development of MCLs already takes into account technological and economic feasibility and should not be influenced by the potential for future water quality violations.

2.2 Examples of Racial Equity Analysis in State Agency Reports

2.2.1 CalEnviroScreen

The Office of Environmental Health Hazard Assessment (OEHHA) has conducted stand-alone analyses of race and ethnicity to supplement versions 3.0 (OEHHA, 2018) and 4.0 (OEHHA, 2021a) of CalEnviroScreen, California's geospatial environmental justice tool. CalEnviroScreen documents and evaluates cumulative pollution burdens from various sources, calculated from 19 indicators, and assesses the resulting vulnerabilities in communities across the state (OEHHA, 2021b; 2021c). The tool is also employed to identify disadvantaged communities[8] (SB 535, 2012; OEHHA, 2022). Race and ethnicity are not included as factors in the calculation of the pollution burden scores (OEHHA, 2018).

Consistent with the results reported in the peer-reviewed literature, OEHHA's analyses found that communities of color, particularly Latino and Black residents, are disproportionately impacted by pollution burden and vulnerability (OEHHA, 2021a).

CalEnviroScreen uses U.S. Census Bureau tracts as the geographical unit of analysis, calculating scores for each tract. The advantage of this approach is that it allows for seamlessly linking the scores with demographic and socioeconomic information for further analysis. However, census tracts are less suitable for linking racial demographic information to PWS, as PWS service areas are generally much smaller. Consequently, community differences may be missed when using relatively large administrative units.

Figure 2 shows pollution burden disparities with respect to race and ethnicity (OEHHA, 2021a). census tracts were divided evenly into ten categories (deciles) based on increasing CalEnviroScreen 4.0 Scores, and the racial composition was determined for each decile. If pollution burden were distributed equally the profiles would reflect the overall racial makeup for the state (bottom bar). However, the figure clearly shows that Latino and Black groups represent a higher proportion of residents in the most impacted deciles (OEHHA, 2021a). Figure 3 from CalEnviroScreen 3.0 (OEHHA, 2018) provides an alternative way to visualize racial disparities by plotting the fraction of each race or ethnicity residing in each decile.

[8] For the purposes of SB 535, the California Environmental Protection Agency (Cal/EPA) designated disadvantaged communities in 2022 as those census tracts: a) with the highest 25% CalEnviroScreen 4.0 Scores (i.e., highest pollution burdens); b) with the highest 25% CalEnviroScreen 3.0 Scores as of 2017; c) with highest 5% pollution burden indicator scores; or d) federally recognized tribal lands (OEHHA, 2022).

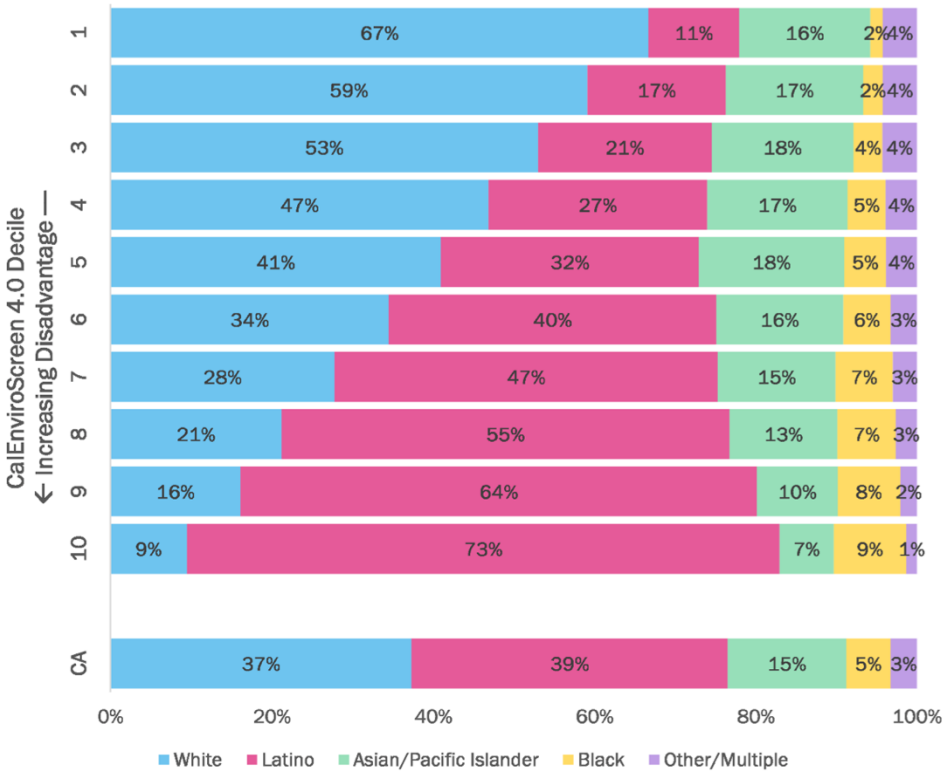

Figure 2. Racial composition of census tracts arranged by decile of increasing pollution burden (i.e., increasing CalEnviroScreen 4.0 scores). The bottom bar represents the overall racial composition for California (Figure 3 in OEHHA, 2021a).

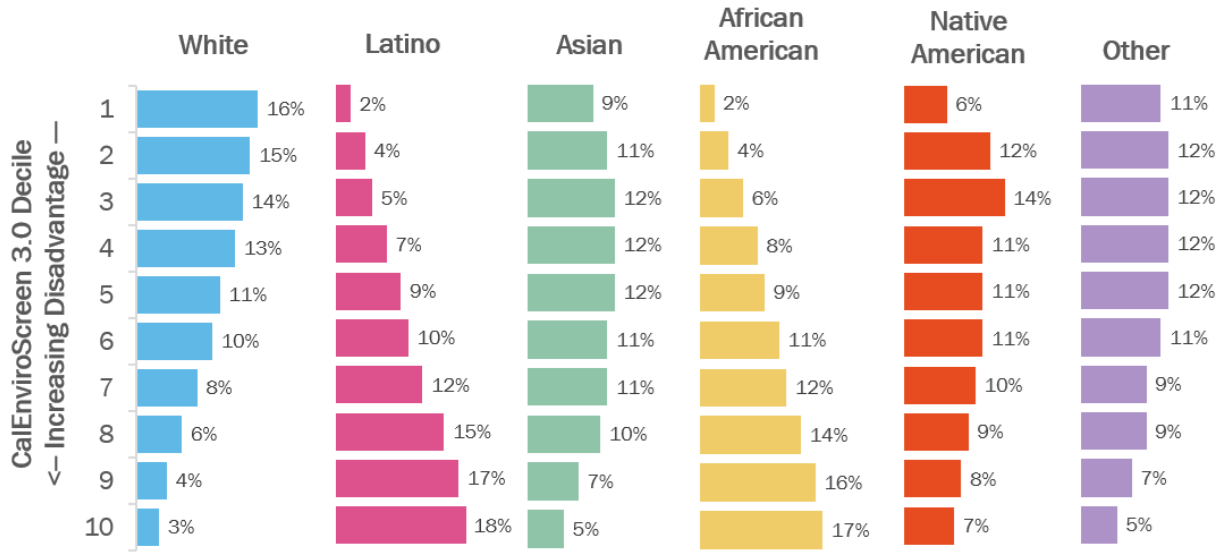

Figure 3. Race and ethnicity fractions for each census tract decile arranged by increasing pollution burden as determined from CalEnviroScreen 3.0 Scores. (Figure 5 in OEHHA, 2018).

2.2.2 Human Right to Water Framework

As part of the state's efforts to meet the human right to water[9] goals, OEHHA developed the Human Right to Water Framework and Data Tool (OEHHA, 2021d). Similar to CalEnviroScreen, this tool uses various indicators to comprehensively evaluate the quality of drinking water in California.

While this framework did not explicitly address racial equity, it examined equity considerations related to drinking water quality and financially disadvantaged communities[10] (OEHHA, 2021d). Similar to the violation studies in the peer-reviewed literature, this analysis required linking demographic data (e.g., median household income) to the service areas of CWS (OEHHA, 2021d). The results revealed that, on average, CWS in financially disadvantaged communities faced more water quality challenges (OEHHA, 2021d). Although a racial analysis was not conducted in this report, it could have been performed by linking the relevant demographic data to CWS service area boundaries.

[9] Assembly Bill 685 was passed in 2012 recognizing that in the State of California "every human being has the right to safe, clean, affordable, and accessible water adequate for human consumption, cooking, and sanitary purposes."
[10] For purposes of the Human Right to Water Framework, disadvantaged community status is based solely on California's statewide median household income (MHI). A disadvantaged community has an MHI at or below 80% of the statewide MHI. Similarly, a severely disadvantaged community has a MHI at or below 60% of the statewide MHI.

3. Methodology

The comprehensive study by McDonald et al. (2022), which reviewed geospatial data on CWS in the United States as of 2020, highlighted the crucial need for accurate service area boundaries for every water system to better understand socioeconomic disparities related to drinking water quality. Without CWS service boundaries, researchers frequently aggregate data to larger administrative units, such as counties or zip codes, which can lead to erroneous assumptions and reduce the ability to conduct higher-resolution analysis. According to the authors, 96% of all U.S. counties have more than one CWS (McDonald et al., 2022), making counties a less ideal administrative unit for this analysis. Furthermore, the absence of service area information complicates the assessment of how CWS health violations, such as MCL exceedances, impact specific communities, as racial and ethnic profiles can vary significantly within a given county.

When McDonald et al. conducted their research in 2020, California was one of the 24 states with available water system service area information (McDonald et al., 2022). California's database at the time, known as the Water Boundary Tool, was initially created in 2016 by the California Environmental Health Tracking Program of the Public Health Institute who maintained it until July 2020, when the Division of Drinking Water of the SWRCB integrated it into their newly developed System Area Boundary Layer (SABL) dataset. As of July 2024, the SABL dataset includes the service areas of 4,803 PWS serving over 95% of the state's population (SWRCBb, 2024).

In a 2018 report, the National Environmental Justice Advisory Council recommended that the U.S. EPA implement a mapping initiative of CWS infrastructure to generate more granular information, which could support project funding in environmental justice (EJ) communities (National Environmental Justice Advisory Council, 2018). Having precise service area boundaries would permit more accurate assignment of demographic variables for conducting socioeconomic analyses, and it could eventually inform where to conduct water testing within a specific community (McDonald et al., 2022).

One of the objectives of this project was to generate a new dataset that assigns racial and age demographic data to PWS in California, particularly those with service areas already included in the SABL dataset. The dataset was built in R. The complete R code is included as Appendix A.

3.1 Data Sources

3.1.1 Census Data

Demographic data from the most recent 5-year American Community Survey (ACS) (2018–2022) was downloaded from the U.S. Census Bureau directly into R using functions available in the R package *tidycensus* (Walker, 2024).

Data was retrieved at the block group (BG) level, which is a subdivision of census tracts typically containing between 600 and 3,000 people and is the smallest geographic area used in the ACS. According to the 2020 U.S. Census, the state of California is divided into 25,607 BGs.

Racial and ethnic data was obtained from ACS Table B03002: 'Hispanic or Latino Origin by Race.' Data elements from Table B03002 were consolidated into seven categories, as summarized in Table 1.

Table 1. Racial and ethnic categories built from data available in ACS Table B03002: 'Hispanic or Latino Origin by Race.'

Category	Table Elements
Total BG population	B03002.001
White population (non-Hispanic)	B03002.003
Black population (non-Hispanic)	B03002.004
Asian, Hawaiian, and Pacific Islander	B03002.006, B03002.007
Native American and Alaskan Native	B03002.005
Latino (all Hispanic)	B03002.012
Multiple races, and other	B03002.008, B03002.010, B03002.012

Age data was obtained from ACS Table B01001: 'Sex by Age.' Data elements from Table B01001 were consolidated into six categories, as summarized in Table 2.

Table 2. Age categories built from data available in ACS Table B01001: 'Sex by Age.'

Category	Table Elements
Total BG population	B01001.001
Total male population	B01001.002
Total female population	B01001.026
Total population under 10 years	B01001.003, B01001.004, B01001.027, B01001.028
Total population 10 years to 64 years of age	B01001.005–B01001.019; B01001.029–B01001.043
Total population 65 years of age and older	B01001.020–B01001.025; B01001.044–B01001.049

Similar information from Tables B03002 and B01001 was also retrieved for other administrative units such as counties, as well as at the state level. Shapefiles of blocks, BGs, and other geographic units were also downloaded using *tidycensus*. Demographic data from these tables was joined to the corresponding shapefiles using R functions.

Population data for California per block unit (i.e., BG subdivision) was downloaded from the 2020 U.S. Census to be used with population-weighted interpolation.

3.1.2. Drinking Water Systems

PWS information was collected from two publicly available datasets created by SWRCB.

3.1.2.1. System Area Boundary Layer (SABL) Dataset

The SABL dataset is maintained by the Division of Drinking Water, and it contains the geographic service areas and additional characteristics of numerous PWS serving most of the state's population. As of July 2024, the SABL dataset included 4,803 water systems.

The SABL shapefile was downloaded from the California State Geoportal repository at: https://gis.data.ca.gov/datasets/fbba842bf134497c9d611ad506ec48cc_0/explore

3.1.2.2. California Drinking Water System Locations (DWSL) Dataset

The DWSL dataset provides the best-known geographical locations for 8,425 water systems (as of July 2024) including PWS and "state small" drinking water systems. According to HSC section 116275(n) "state small water systems" have between 5 and 14 connections and do not serve more than an average of 25 people daily for more than 60 days.

The DWSL shapefile was downloaded from the California State Geoportal repository at: https://gis.data.ca.gov/datasets/346d649d1e654737ac5b6855466e89b2_0/explore?location=36.792025%2C-120.674407%2C9.94

3.2 GIS Spatial Analysis

3.2.1 Data Preparation

BGs with a total population of 13 or fewer individuals were excluded from the analysis. This threshold value corresponds to the margin of error for a BG with a population of zero. In total, 98 BGs were excluded, 91 of which already had a population of zero as they cover areas above the ocean and other water bodies.

The demographic information extracted from the ACS tables was joined to the corresponding BG shapefile in R. For every BG, percentages of every racial and age category were calculated using the total population of the BG.

Thirty five PWS from the SABL dataset were excluded from the analysis. Of these, 29 had jurisdictional system boundaries[11] listed in the SABL dataset, which, in most cases, represented duplicate entries of PWS with active service areas. Additionally, 6 PWS were located within BGs that had a population of zero, and therefore no demographic information could be assigned. In total, 4,768 PWS were selected from the SABL dataset.

From the DWSL, 7,969 water systems were selected, with 456 excluded as duplicate entries or because they did not contain data on the serving population. From the selected group, 4,725 were already included in the SABL dataset and were processed separately. However, the DWSL dataset contained additional variables not included in the SABL dataset, so these variables were incorporated into the final dataset.

A boundary layer for the service areas of the remaining 3,244 water systems in the DWSL dataset is currently unavailable. However, PWS coordinates were geocoded as small circular polygons (i.e., buffers) allowing them to be intersected with the BG shapefile.

In total, 8,012 water systems were included in the final dataset. Of these, 7,027 comprise PWS and 985 are classified as "state small water systems".

3.2.2 Spatial Operations

The SABL and DWSL datasets were processed separately.

3.2.2.1. System Area Boundary Layer (SABL) Dataset

PWS from the SABL dataset were intersected with the BG shapefile containing demographic data from the ACS. Of the 4,768 PWS included in the analysis, 3,143 intersected with only 1 BG, while 1,625 intersected with 2 or more BGs. The largest PWS[12] intersected with 2,968 BGs. However, approximately 90% of PWS in the dataset intersected with 10 or fewer BGs.

For PWS intersecting with only 1 BG, demographic percentages for each category were directly assigned from the corresponding BG in which the PWS is located. These BG percentages were previously calculated from each category value and the total population of the BG.

Using the assigned percentages and the service population value of each PWS in the SABL dataset, new values were subsequently calculated for each demographic category. Therefore, these values correspond to the total number of people in each category based on the populations served by each PWS.

[11] Jurisdictional boundaries include zones that are assigned to the PWS, but that are currently not served. These zones may include areas with issued "will serve" letters, or zones that the PWS is legally obligated to serve but it is not doing so.

[12] Los Angeles Department of Water and Power with 708,607 connections and serving 3.8 million people is the largest PWS in the dataset.

For PWS intersecting with multiple BGs, interpolation methods were used to assign population values to each PWS based on the intersected BGs. Both areal-weighting and population-weighting interpolations were done for each PWS in this subset, and the result closest to the PWS population value in the SABL dataset was selected. Block-level population data from the 2020 U.S. Census was used to calculate the population weights.

Demographic percentages were calculated from the interpolated data, and later combined with the PWS population value in the SABL dataset to arrive at the number of people in each demographic category for every PWS. From this subset of 1,625 PWS, 622 were interpolated using areal-weighting, while 1,003 where interpolated using the population-weighting approach.

3.2.2.2 California Drinking Water System Locations (DWSL) Dataset

The remaining 3,244 water systems did not have real service areas and were only represented as small, circular polygons in the shapefile. Therefore, it was expected that they would intersect with only 1 BG. While this was true for most of them, 779 PWS were near BG edges, and therefore intersected with up to 4 different BGs.

For systems intersecting with 1 BG, demographic percentages were assigned directly from the corresponding BG. For the rest of the PWS, percentages were assigned only from the BG contributing the largest area. Interpolation methods were not practical for this subset as their service areas are arbitrary representations.

3.3. Data Compilation and Output Files

The SABL and DWSL datasets had various overlapping variables, and others that were unique to each dataset. An effort was done to retain as much information from both datasets, while consolidating the rest of the information.

As mentioned above, 4,725 PWS were found in both the SABL and DWSL datasets. However, PWS information of each system was not always consistent between the two datasets. For example, there were discrepancies in the population and service connection values for 560 PWS and 708 PWS, respectively. Other differences were found in variables such as system name or system classification type. For PWS found in both datasets, the SABL information took precedence as it is updated on a regular basis.

The final dataset was saved as a shapefile for GIS analysis, as well as in a CSV file version (without the geographic information).

4. Results

4.1 Dataset and Interpolation Results

Two versions of the shapefile were created. The full version contains 61 fields (columns), which, in addition to the demographic categories (population values and percentages), retains most of the fields from the original SABL and DWSL datasets. The second version has 36 fields and includes only the essential information about the water systems, along with the newly added demographic category fields.

The shapefiles cover information on 8,012 water systems. Table 3 summarizes the state's classification of these water systems based on the original information from the SABL and DWSL datasets. The population column lists the total population served by each system type.

Table 3. California water systems included in the new dataset organized by state's classification.

Classification	System Count	Served Population
Community	2,846	39,658,483
Transient, non-community	2,783	857,941
Non-transient, non-community	1,398	909,986
"State small" water systems [a]	972	18,038
Non-public	3	93
Recycled water	10	2
Total	**8,012**	**41,444,543**

[a] The following classifications used in the original SABL and DWSL were consolidated: "state small water system", "state small", and "local state small system".

The population values correspond to estimates reported for each water system in the original SABL and DWSL datasets. Note that the total population exceeds California's official estimate of 39,356,104 from the 2020 U.S. Census. One reason for this discrepancy is that various systems in the SABL and DWSL datasets have overlapping service areas, including systems with identical service areas but unique system IDs[13] (PWSID field). Consequently, double counting likely led to overestimating the total population. Furthermore, wholesaler water systems were not excluded from the final dataset, as the primary objective was to assign demographic information to water system boundaries. However, these systems likely contributed to the overcount. Correcting the original datasets and identifying the most recent and accurate entries among possible duplicates was beyond the scope of this project. Depending on the specific objectives, future analyses using this dataset will require additional data filtering and cleanup.

[13] Every water system in California is assigned a unique PWSID number.

Table 4 categorizes the water systems according to size, measured by the number of service connections. While most systems are considered small systems (~95%), large systems serve the majority of the population.

Table 4. California water systems based on size as determined by the number of service connections.

Size	System Count [a]	Served Population
Small (≤ 3,300 connections)	7,589	4,868,060
Medium (3,300 – 9,999 connections)	181	4,083,046
Large (≥ 10,000 connections)	228	32,493,317
Total	**7,998**	**41,444,423**

[a] Excluded: "non-public", "recycled water", and systems with missing service connection information.

One challenge in assigning demographic data to water systems is the wide variability in service area sizes. As shown in Table 4, most water systems are classified as small, with service areas often significantly smaller than the BGs. In contrast, some large water systems cover hundreds of BGs, requiring interpolation methods to accurately weigh their contributions.

Figure 1 shows the PWS service areas in a section of Yolo County, including the cities of Davis, West Sacramento, and Woodland. The county is divided into BGs as indicated by gray lines. Blue areas represent the service areas of PWS that intersect with a single BG, while the tan areas depict the service areas of larger PWS, which intersect multiple BGs. The black dots mark the centroids of census blocks, illustrating the non-uniform distribution of population density across the county.

To assign demographic data (i.e., racial and age percentages) to the PWS that intersect multiple BGs, both areal-weighted and population-weighted interpolation methods were used. The approach providing the closest estimate to the PWS population value was selected. Table 5 summarizes the results of the interpolation selection for the 10 PWS in Yolo County that intersect multiple BGs.

The results in Table 5 indicate that population-weighted interpolation is more accurate, particularly for small PWS serving relatively small populations, where areal interpolation significantly underestimated the populations. For example, for PWS CA5700571 with a reported serving population of 876, the areal method estimate was only 4 individuals, essentially rendering the result useless for further population analysis. This is why studies in the literature relying on areal interpolation methods often excluded most smaller water systems (Allaire & Acquah, 2022). For relatively larger PWS serving larger populations, the differences between the methods were less pronounced, yet population-weighted interpolation provided closer results in all but one case.

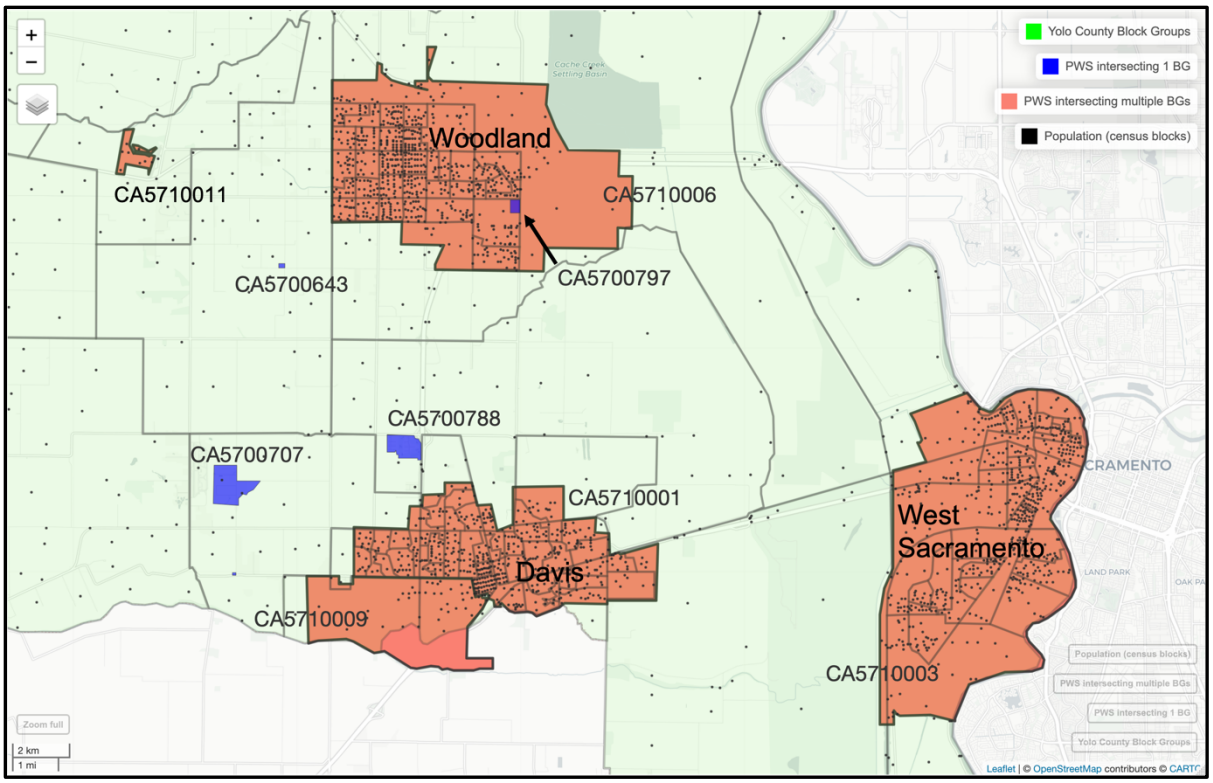

Figure 1. A section of Yolo County, California showing the service areas of PWS that intersect with a single BG (blue) and multiple BGs (tan). Black dots represent the centroids of census blocks (block group subdivisions), illustrating the population distribution within each BG. The map also shows the PWSIDs of individual PWS.

Table 5. Results of the interpolation selection for the 10 PWS in Yolo County that intersect multiple BGs. Some of these systems are shown in Figure 1. The selected method is underlined.

PWSID	System Population[a]	Areal-weighted Estimate	Difference	Population-weighted Estimate	Difference
CA5700571	876	4	872	<u>523</u>	353
CA5700712	602	300	302	<u>576</u>	26
CA5710001	67,217	61,982	5,235	<u>68,886</u>	-1,669
CA5710003	53,355	<u>53,090</u>	265	54,646	-1,291
CA5710004	944	5	939	<u>1,082</u>	-138
CA5710005	7,115	4,979	2,136	<u>6,755</u>	360
CA5710006	61,462	60,891	571	<u>61,325</u>	137
CA5710007	3,574	1,376	2,198	<u>2,589</u>	985
CA5710009	49,616	8,190	41,425	<u>8,528</u>	41,088
CA5710011	1,115	52	1,063	<u>634</u>	481

[a] As reported by the each PWS and captured in the original SABL and DWSL datasets.

For PWS CA5710009, which serves southwest Davis (see Figure 1), both methods significantly underestimated the population reported by the PWS by approximately 41,000 people. Given that Davis has a population of almost 66,000 people[14] and the City of Davis's PWS (CA5710001) serves this entire population, it is likely that the reported population for PWS CA5710009 was inaccurate, resulting in an overcount of approximately 50,000 people for Davis. A detailed analysis of Davis would require additional investigation into these PWS to resolve the discrepancies. However, since the demographic data was assigned as percentages, the relative fractions are likely representative of the city's population.

4.2 Demographic Comparisons

Having demographic data associated with California's PWS allows for straightforward mapping of these systems according to any of the categories. Figure 2 displays PWS in the Los Angeles area based on the percentage of Latino (Hispanic) population in each PWS service area. Similarly, Figure 3 illustrates the percentage of senior citizens in the same region.

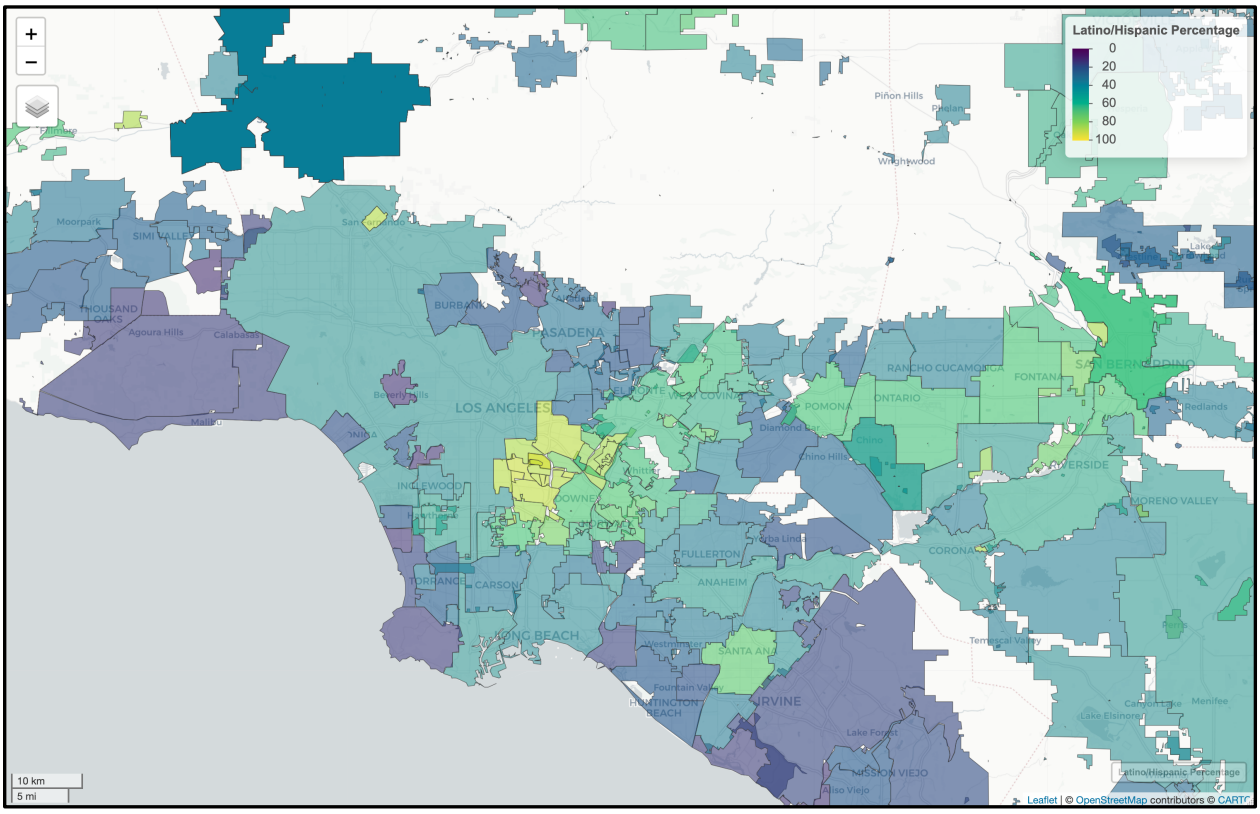

Figure 2. Percentage Latino/Hispanic population based on the service areas of different PWS in the Los Angeles area.

[14] U.S. Census Bureau: https://www.census.gov/quickfacts/table/PST045216/0618100,00

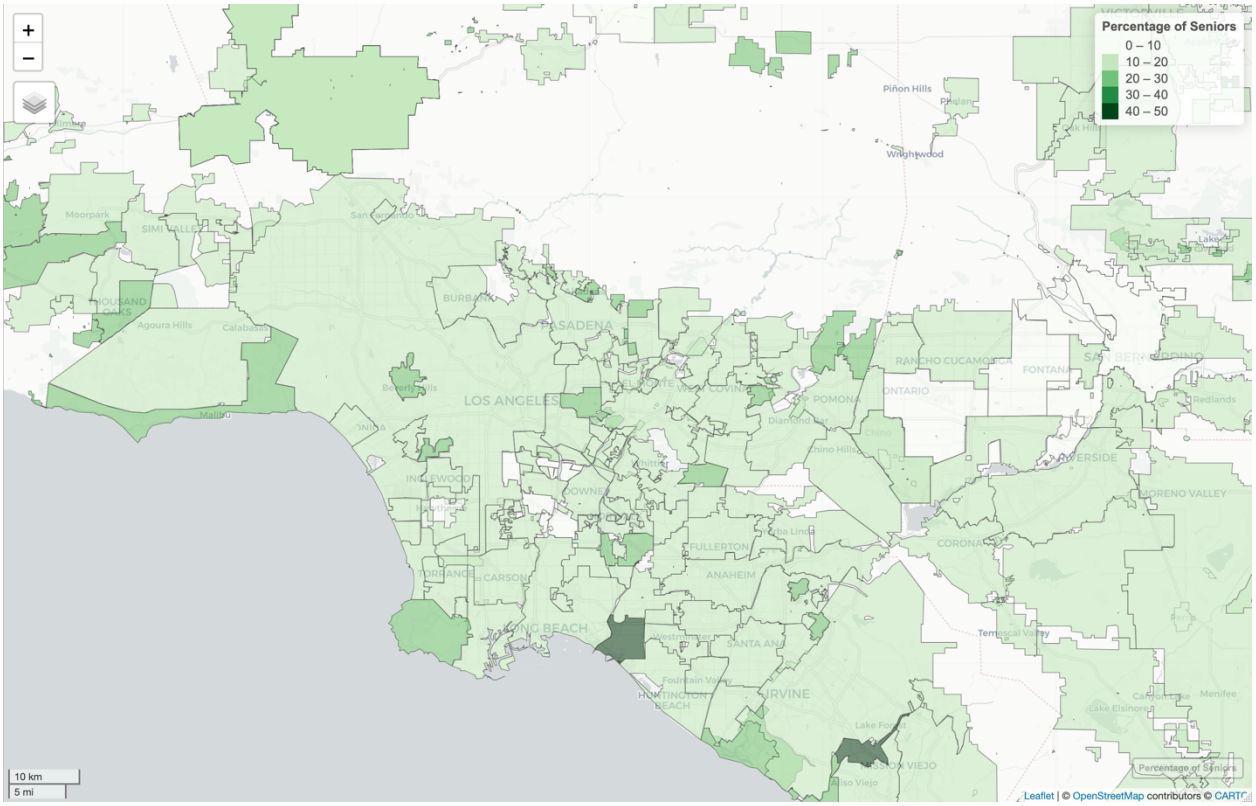

Figure 3. Percentage of senior citizens based on the service areas of different PWS in the Los Angeles area.

Furthermore, racial profiles can be constructed based on specific characteristics of the PWS. For example, Figure 4 summarizes the racial profiles of the populations served by PWS of varying sizes, as defined by their number of connections (see Table 4).

While maps like those in Figures 2 and 3 and the profiles in Figure 4 provide interesting insights into the racial distribution and composition of populations related to PWS, the reviewed literature suggests that establishing a baseline is required for conducting racial equity analyses to identify potential disparities. Typically, the overall demographic profiles for the state serve as an appropriate baseline for these comparisons. Figure 5 displays the racial and age group proportions for California based on data from the 2018–2022 ACS.

Comparing the profiles in Figure 4 with California's overall racial composition in Figure 5 indicates that the population served by the large PWS closely aligns with the state's racial composition. Even if the population values are overestimated, Table 4 shows that the majority of the state's population is served by the largest PWS, which consequently approximates the state's racial demographics.

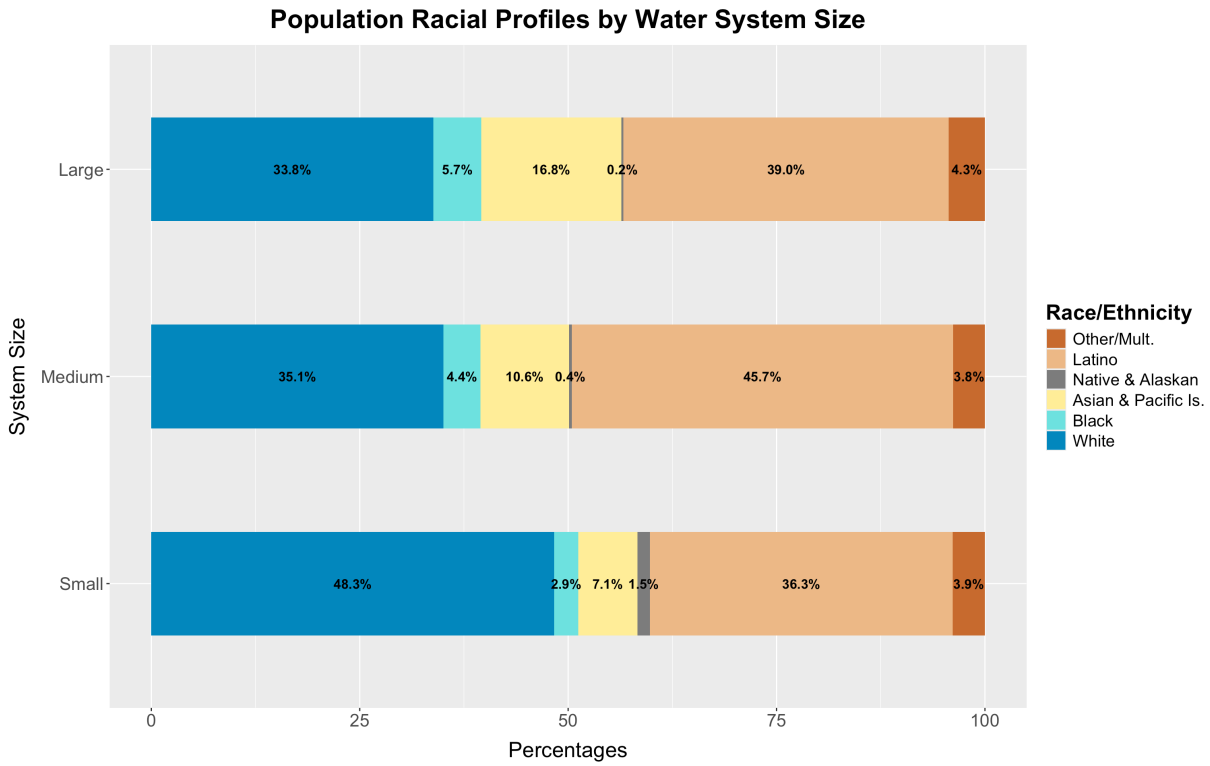

Figure 4. Racial profiles of populations served by PWS of varying sizes. Small PWS are water systems with up to 3,300 connections; medium systems have between 3,300 and 10,000 connections, and large systems have more than 10,000 connections.

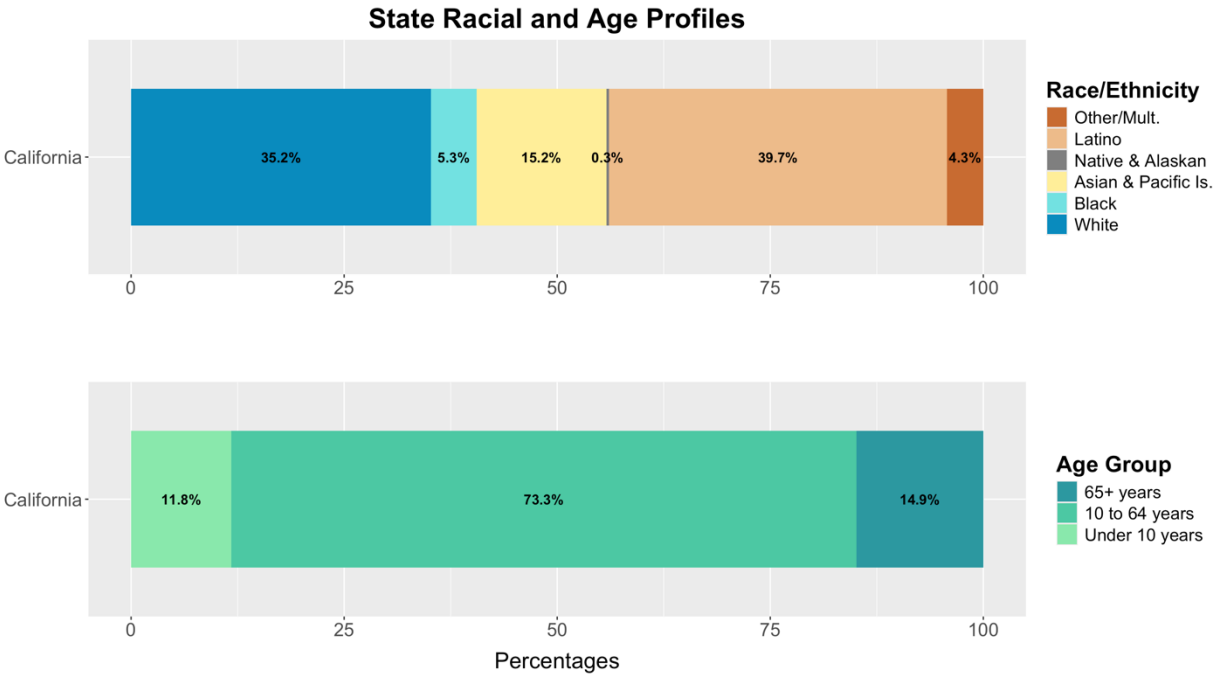

Figure 5. Overall racial and age group profiles for the state of California.

4.3 Contaminant Occurrence Assessment

The MCL development process typically involves assessing the occurrence of the corresponding contaminant based on historical monitoring reports from PWS across the state. This analysis provides information about the sources, distribution, and concentration levels of the contaminant, as well as the population that may potentially be exposed to it in their drinking water. Having demographic data associated with each water system allows for an expanded analysis, which can provide additional insights.

Occurrence data for arsenic from 2010 to 2020, compiled by the RDU, was used as an illustrative example. The PHG and MCL for arsenic are 0.004 μg/L and 10 μg/L, respectively, while the detection limit for purposes of reporting (DLR) is 2 μg/L. From this occurrence data, the maximum value detected during the entire period was recorded for each PWS. In total, 3,061 PWS reported at least one measurement at or above the DLR of 2 μg/L during this period. Figure 6 displays a region of the state, including the San Francisco Bay Area and sections of the Central Valley. PWS with maximum concentrations detected between the DLR and the MCL are shown in green, while those with maximum concentrations above the MCL appear in red.

It is important to emphasize that these values do not necessarily indicate non-compliance or that the served populations were exposed to these levels of arsenic. They simply reflect the maximum concentration detected within the PWS, such as in a source well. However, this source may not have been used, or the water may have undergone treatment or been blended with other source waters before being delivered to customers. Therefore, the data represents a hypothetical worst-case-scenario assessment. Generally, an occurrence analysis involves using other concentration benchmarks, such as running annual averages, as well as more detailed evaluations of individual treatment flows in PWS that detected particularly elevated concentrations.

Figure 7 presents the racial profiles of PWS that detected arsenic at least once within their system boundaries at concentrations equal to or above the DLR of 2 μg/L. For systems with concentrations below the MCL (top), the racial profile is very similar to the state's profile (see Figure 5). However, for the 579 systems that detected concentrations above the MCL (bottom), the relative proportions of Latino and Black populations increase, while the White and Asian populations decrease.

These profiles indicate that PWS that measured arsenic concentrations above the MCL serve a relatively larger fraction of Latino and Black populations compared to those with concentrations below the MCL. Qualitatively, this suggests that these racial and ethnic groups might be disproportionately exposed to higher levels of arsenic. However, this preliminary analysis does not imply that these populations received drinking water that was out of compliance. It only provides a rough assessment of risk based on contaminant measurements from the PWS.

Figure 8 summarizes similar information related to age groups. Both profiles closely align with the overall state distribution (see Figure 5). There is only a slight decrease in the proportion of the senior population among the PWS that detected arsenic concentrations above the MCL, with the difference being distributed among the other age groups.

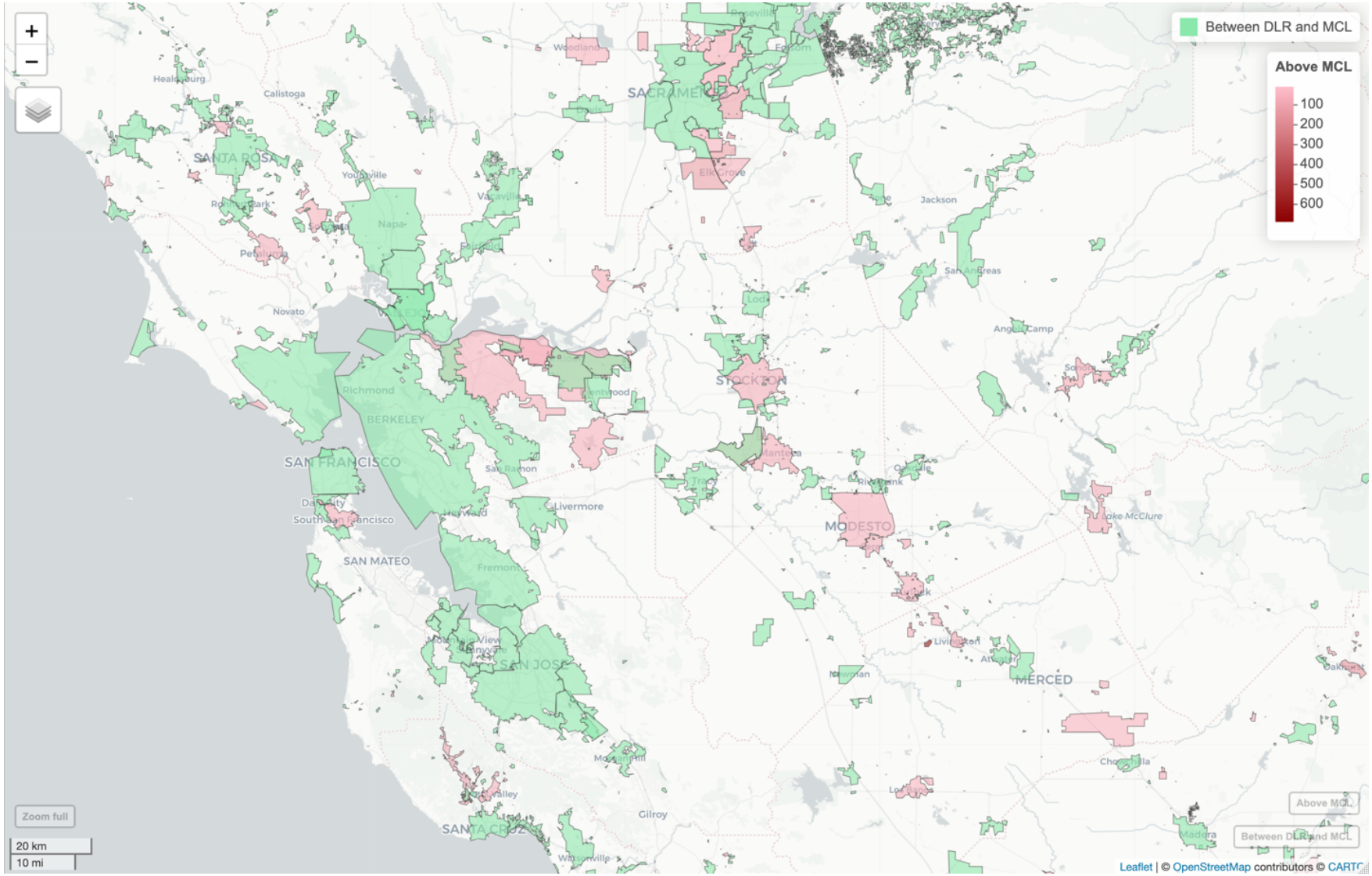

Figure 6. PWS that measured arsenic at least once during the period 2010–2020 at concentrations between the DLR (2 µg/L) and the MCL (10 µg/L) (green), and above the MCL (red).

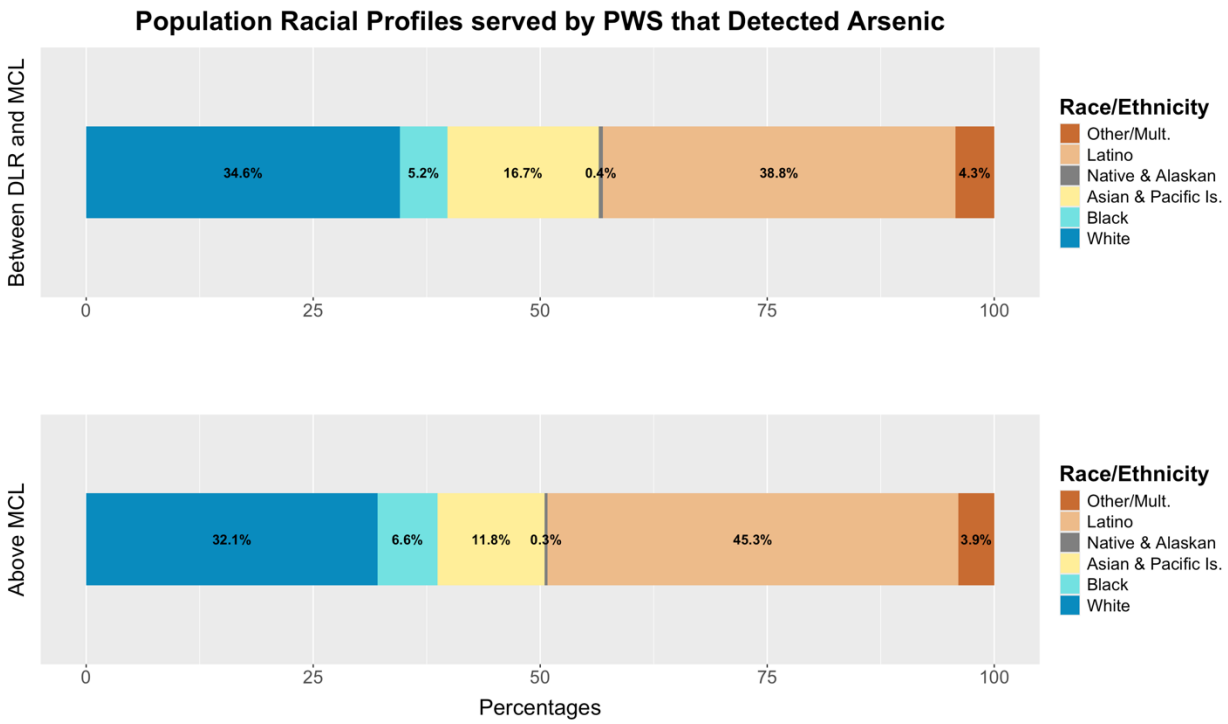

Figure 7. Racial profiles of all PWS that detected arsenic during 2010–2020 at maximum concentrations between the DLR (2 µg/L) and the MCL (10 µg/L) (top), and above the MCL (bottom). Based on the PWS population data, the cumulative populations are 23,169,522 people for the top group (2,481 systems) and 12,701,725 for the bottom group (579 systems). The population numbers are likely overestimated due to PWS overlap and double counting.

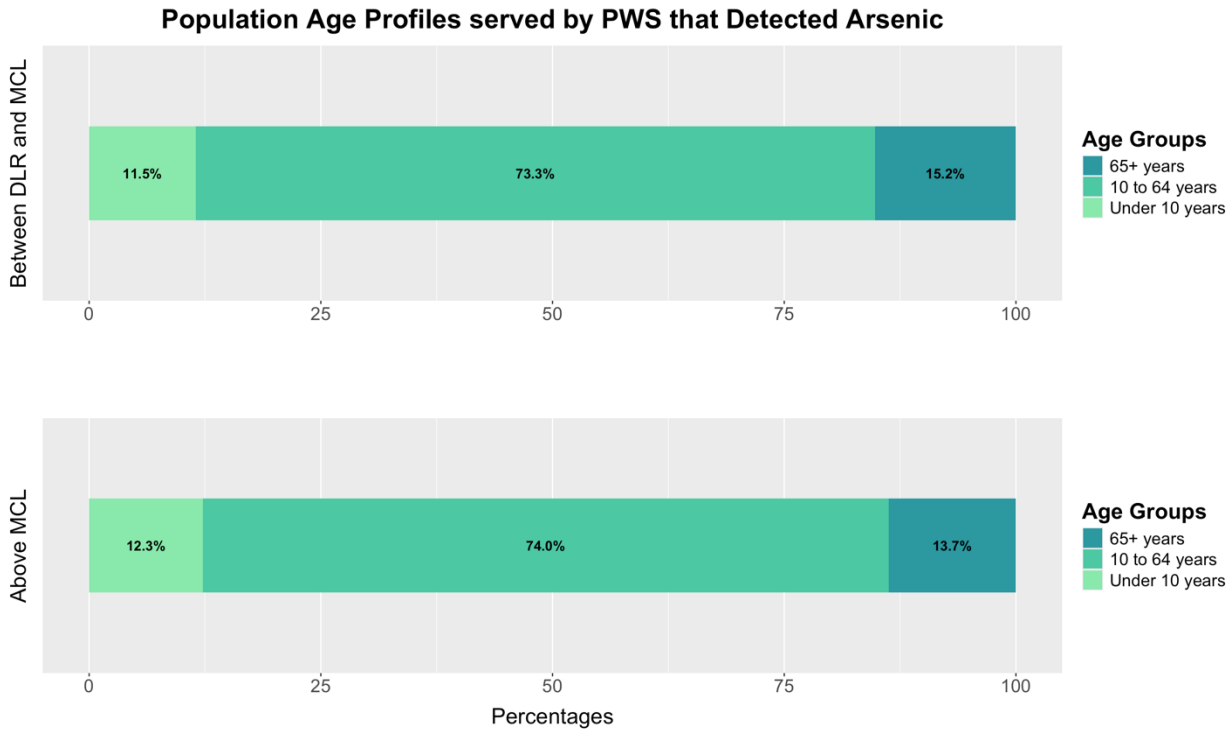

Figure 8. Population age groups served by PWS that detected arsenic during 2010–2020 at maximum concentrations between the DLR (2 µg/L) and the MCL (10 µg/L) (top), and above the MCL (bottom).

5. Implications for Prioritization based on Race and Ethnicity

The dataset developed in this project provides a tool for further informing contaminant occurrence analysis during the MCL process and may help identify potentially elevated exposure risks among certain demographic groups. However, MCLs are primarily established to universally protect human health, and the RDU already evaluates potential impacts on overall population numbers when prioritizing contaminants for the MCL program. Therefore, additional considerations of race, ethnicity, age, or other demographic characteristics that could further guide this decision-making process should be supported by health-based data relating these factors to contaminant exposure through drinking water. Unfortunately, conclusive public health data linking sensitivities related to race and ethnicity to regulated contaminants in drinking water is largely unavailable.

Arsenic is a widespread contaminant in California. Hypothetically, if the MCL were lowered from 10 μg/L to the current DLR of 2 μg/L (assuming technological and economic feasibility), it could provide additional health protection to a significant portion of the state's population by reducing the theoretical cancer risk by 80% and bringing it closer to its PHG. In a previous internal prioritization analysis exercise, the RDU ranked arsenic as number one for potential cancer risk reductions and eighth based on the state's population that would benefit from lowering the MCL. However, without further specific knowledge on differentiated health effects among racial or ethnic groups, currently available data can only support the general conclusion that health benefits would be equitable for the entire population should the MCL be lowered.

The deviation from the state's average racial profile identified for those PWS that detected arsenic above the MCL (see Figure 7) neither stems from the MCL providing differentiated health protection among racial and ethnic groups nor can it be resolved by changing the MCL. Instead, this disparity likely indicates potential water quality and compliance issues. If these elevated concentrations were found only in raw water measurements but each PWS was able to treat the water to meet the MCL, then no racial or ethnic group would be at a disadvantage, and everyone's health would be protected equally. However, if these concentrations reflect actual water quality violations, a true racial disparity exists that must be addressed by correcting the underlying causes of the violation.

California HSC section 116365(c) mandates that OEHHA conduct a risk assessment based solely on public health and scientific considerations for every contaminant considered for regulation. This assessment determines the PHG, which is the concentration of the contaminant in drinking water that would not pose a significant health risk if consumed daily over a human lifetime. Additionally, HSC section 116365(c)(1)(C)(ii) requires OEHHA to account for health effects that adversely impact sensitive subgroups of the general population, such as children, pregnant women, or the elderly. Although this section does not explicitly list racial and ethnic populations as examples of sensitive subgroups, if scientific information were to indicate that a specific racial group's health is disproportionately affected by a given contaminant, this group must be included in the risk assessment like any other sensitive subgroup.

To determine a PHG, OEHHA uses the most sensitive health endpoints and adverse effects (i.e., those occurring at the lowest dose), including considerations for potentially sensitive subgroups. By focusing on worst-case scenarios, this approach maximizes public health protection.

OEHHA's risk assessments and PHGs form the public health foundation for setting MCLs, as they involve a rigorous review of the best available information. For example, in the recent risk assessment of perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS), OEHHA concluded that "[c]urrently it is unknown whether PFOA or PFOS cancer risks might vary by race, ethnicity, or any other related factor" (OEHHA, 2024, p. 226). Therefore, if these risk assessments do not identify specific racial or ethnic subgroups as sensitive populations that could suffer disproportionate harm from a given contaminant, it would be impractical for the RDU to seek additional racial and ethnic criteria to prioritize MCLs.

Based on these findings, it is recommended that the RDU continue to prioritize contaminants whose MCLs can be lowered closer to their PHGs. This should involve using metrics like those tested in the internal prioritization exercise, such as evaluating cancer risk reductions and other health improvements, as well as using occurrence analyses to estimate the overall population numbers that would benefit from these changes. These approaches are also similar to the prioritization scheme proposed by Rosenblum et al. (2024) (see Figure 1).

While racial and ethnic considerations should be supported by health-based data relating drinking water contaminants to these factors, the dataset created in this project can be used to obtain further insights into the relative demographic characteristics and geographical distribution of the populations that may benefit from a new or revised MCL. Additionally, if future research reveals that certain adverse health effects disproportionately impact racial or ethnic groups, this dataset could inform the development of new prioritization criteria and help identify areas with higher proportions of these sensitive subgroups.

5.1 Other Data Limitations

5.1.1. Contaminant Monitoring

Another significant data limitation is related to insufficient monitoring of contaminants across the various stages of PWS, from raw water sources, through treatment plants, and especially along distribution systems to individual households. This data gap severely limits the ability to conduct comprehensive public health analyses to accurately assess the true exposure that different populations may be experiencing to specific contaminants. This limitation is not unique to the objectives of the RDU; other programs within DDW and the SWRCB would benefit from having such detailed information. However, addressing this issue is complex and would require significant modifications to current monitoring practices and requirements, including establishing new measurement locations and managing the associated costs and logistics, which could be prohibitive for some PWS.

When assessing contaminant occurrence, the RDU generally focuses on water source measurements, as this is where most contaminants enter PWS. Source water measurements often correspond to higher concentrations, providing a more conservative estimate of potential exposure. However, without comprehensive monitoring, it is difficult to determine whether issues arise at other stages of the PWS, such as in the distribution system, that could affect water quality and alter the conclusions of occurrence analyses for these contaminants.

There are other contaminants, such as lead, copper, pathogens, and disinfectant byproducts, that are specifically monitored in the distribution system or in the tap water. However, additional monitoring locations may also be necessary for these contaminants if the objective is to identify potential water quality differences within the PWS.

To conduct more accurate racial equity analyses and identify potential disparities in water quality or public health among communities and individuals, more extensive contaminant monitoring is essential at various locations within the PWS. The lack of sufficient monitoring results in data limitations that restrict occurrence analyses to averages of population numbers and contaminant concentrations across the entire PWS.

5.1.2. Population Estimates

The serving population estimates reported by each PWS through SDWIS[15] are crucial for any demographic analysis related to PWS. The method developed in this project uses data from the ACS to assign demographic percentages to each PWS. However, errors, such as population overestimations, may be introduced and propagated in any subsequent calculations if the values are inaccurate. In the future, it would be useful to have PWS verify their serving populations regularly, similar to how service boundaries are also checked and verified when updating the SABL dataset.

It is unlikely that PWS would be willing or have the capacity to start collecting and reporting demographic information about their customers. Therefore, census-based methods, such as the one used in this project, are the best available options for gathering the necessary data for analyses related to race and ethnicity. If population estimates are unreliable or not provided by PWS, they would need to be determined using interpolation methods, which can introduce additional uncertainty, especially if areal interpolation is used.

5.1.3. Demographic Differences based on PWS Size

The method developed in this project assigns average demographic percentages to each PWS based on information obtained at the census block group level. While this approach is likely accurate for most small and medium-sized PWS with relatively small service areas, it would not capture local community and demographic variations within the boundaries of large PWS that supply the majority of California's population. Consequently, the assigned demographic percentages are assumed to be uniformly distributed throughout the PWS boundary, which represents a rough approximation.

This averaging of demographic characteristics is less critical in cases where all customers within the PWS receive identical water quality. However, some large PWS may use different local water sources or operate multiple systems, which can lead to variations among the served population.

[15] Safe Drinking Water Information System (SDWIS): collects information on every PWS, including basic characteristics about the system, as well as violation and enforcement information.

Therefore, a more accurate assessment of these systems would require knowledge of both water quality differences (through more granular monitoring) and demographic information at a finer PWS subdivision level.

While these data limitations are not easily resolved, future pilot projects could be established in partnership with one or more large PWS to test more comprehensive monitoring schemes. Additionally, such projects could help develop more accurate racial equity studies to better inform the MCL development process and address potential water quality disparities.

6. References

Acquah, S., M. Allaire. 2023. “Disparities in Drinking Water Quality: Evidence from California.” *Water Policy* 25 (2): 69–86. https://doi.org/10.2166/wp.2023.068.

Allaire, M., S. Acquah. 2022. “Disparities in Drinking Water Compliance: Implications for Incorporating Equity into Regulatory Practices.” *AWWA Water Science* 4 (2): e1274. https://doi.org/10.1002/aws2.1274.

Balazs, C., J. J. Goddard, C. Chang, L. Zeise, J. Faust. 2021. “Monitoring the Human Right to Water in California: Development and Implementation of a Framework and Data Tool.” *Water Policy* 23 (5): 1189–1210. https://doi.org/10.2166/wp.2021.069.

Balazs, C. L., I. Ray. 2014. “The Drinking Water Disparities Framework: On the Origins and Persistence of Inequities in Exposure.” *American Journal of Public Health* 104 (4): 603–11. https://doi.org/10.2105/AJPH.2013.301664.

McDonald, Y. J., N. E. Jones. 2018. “Drinking Water Violations and Environmental Justice in the United States, 2011–2015.” *American Journal of Public Health* 108 (10): 1401–7. https://doi.org/10.2105/AJPH.2018.304621.

McDonald, Y. J., K. M. Anderson, M. D. Caballero, K. J. Ding, D. H. Fisher, C. P. Morkel, E. L. Hill. 2022. “A Systematic Review of Geospatial Representation of United States Community Water Systems.” *AWWA Water Science* 4 (1): e1266. https://doi.org/10.1002/aws2.1266.

Mueller, R., D. Salvatore, P. Brown, A. Cordner. 2024. “Quantifying Disparities in Per- and Polyfluoroalkyl Substances (PFAS) Levels in Drinking Water from Overburdened Communities in New Jersey, 2019–2021.” *Environmental Health Perspectives* 132 (4): 047011. https://doi.org/10.1289/EHP12787.

Office of Environmental Health Hazard Assessment (OEHHA). 2018. "Analysis of Race/Ethnicity, Age and CalEnviroScreen 3.0 Scores." https://oehha.ca.gov/media/downloads/calenviroscreen/document-calenviroscreen/raceageces3analysis.pdf

Office of Environmental Health Hazard Assessment (OEHHA). 2021a. "Analysis of Race/Ethnicity and CalEnviroScreen 4.0 Scores." https://oehha.ca.gov/media/downloads/calenviroscreen/document/calenviroscreen40raceanalysisf2021.pdf

Office of Environmental Health Hazard Assessment (OEHHA). 2021b. "CalEnvironScreen 4.0." https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40

Office of Environmental Health Hazard Assessment (OEHHA). 2021c. "CalEnvironScreen 4.0 Report". https://oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf

Office of Environmental Health Hazard Assessment (OEHHA). 2021d. "Achieving the Human Right to Water in California. An Assessment of the State's Community Water Systems."https://oehha.ca.gov/media/downloads/water/report/hrtwachievinghrtw2021f.pdf

Office of Environmental Health Hazard Assessment (OEHHA). 2022. "Final Designation of Disadvantaged Communities Pursuant to Senate Bill 535." https://calepa.ca.gov/wp-content/uploads/sites/6/2022/05/Updated-Disadvantaged-Communities-Designation-DAC-May-2022-Eng.a.hp_-1.pdf

Office of Environmental Health Hazard Assessment (OEHHA). 2024. "Public Health Goals. Perfluorooctanoic Acid and Perfluorooctane Sulfonic Acid in Drinking Water." https://oehha.ca.gov/media/downloads/water/chemicals/phg/pfoapfosphgfinaldraft040524.pdf

Pace, C., C. Balazs, K. Bangia, N. Depsky, A. Renteria, R. Morello-Frosch, L. J. Cushing. 2022. "Inequities in Drinking Water Quality Among Domestic Well Communities and Community Water Systems, California, 2011–2019." *American Journal of Public Health* 112 (1): 88–97. https://doi.org/10.2105/AJPH.2021.306561.

Rosenblum, J. S., A. Liethen, L. Miller-Robbie. 2024. "Prioritization and Risk Ranking of Regulated and Unregulated Chemicals in US Drinking Water". *Environmental Science & Technology* 58 (16): 6878–6889. https://doi.org/10.1021/acs.est.3c08745.

Scanlon, B. R., R. C. Reedy, S. Fakhreddine, Q. Yang, G. Pierce. 2023. "Drinking Water Quality and Soci0al Vulnerability Linkages at the System Level in the United States." *Environmental Research Letters* 18 (9): 094039. https://doi.org/10.1088/1748-9326/ace2d9.

Senate Bill 535. 2011-2012. "California Global Warming Solutions Act of 2006: Greenhouse Gas Reduction Fund." Chapter 830 (California Statutes of 2012). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201120120SB535

State Water Resources Control Board (SWRCB). 2023. "2023–2025 California State Water Resources Control Board Racial Equity Action Plan." https://www.waterboards.ca.gov/racial_equity/docs/racial-equity-action-plan-final-en.pdf

State Water Resources Control Board (SWRCB). 2024. "System area layer look-up tool." https://gispublic.waterboards.ca.gov/portal/apps/webappviewer/index.html?id=272351aa7db14435989647a86e6d3ad8

Switzer, D., M. P. Teodoro. 2017. "The Color of Drinking Water: Class, Race, Ethnicity, and Safe Drinking Water Act Compliance." *Journal AWWA* 109 (9): 40–45. https://doi.org/10.5942/jawwa.2017.109.0128.

Uche, U. I., S. Evans, S. Rundquist, C. Campbell, O. V. Naidenko. 2021. “Community-Level Analysis of Drinking Water Data Highlights the Importance of Drinking Water Metrics for the State, Federal Environmental Health Justice Priorities in the United States.” *International Journal of Environmental Research and Public Health* 18 (19): 10401. https://doi.org/10.3390/ijerph181910401.

Walker K., H. M. 2024. “tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames.” R package version 1.6.5. https://walker-data.com/tidycensus/

Appendix A

R-based GIS Method for Assigning Racial and Age Information to California's Drinking Water System

APPENDIX A – Assigning Racial and Age Information to California's Drinking Water Systems

Federico Pacheco

Last updated: August 2024

This document describes a GIS method based on R to assign racial and age demographic information from the American Community Survey (ACS) to individual public water systems (PWS) in California based on their geographic location. The code generates new shapefiles and accompanying CSV files containing the demographic information for each PWS. In total, the resulting dataset includes 8,012 water systems.

Census Information

Demographic information was obtained from the most recent 5-year ACS database (2018-2022) at the census block group (BG) level. According to the 2020 U.S. Census, California is divided into 25,607 BGs. BGs with a total population of 13 or fewer individuals were excluded from this analysis. This threshold was chosen because 13 corresponds to the margin of error for a BG with a population of zero. In total, 98 BGs were removed, 91 of which had already a population of zero.

Water Systems Datasets

PWS information was compiled from two publicly available datasets created by the California State Water Resources Control Board:

1. System Area Boundary Layer (SABL) dataset:

The SABL dataset is maintained by the Division of Drinking Water and contains the geographic service areas and characteristic details of numerous PWS. As of July 2024, the SABL dataset included 4,803 water systems.

The shapefile of this dataset was downloaded from the public ArcGIS Hub at:
https://hub.arcgis.com/datasets/fbba842bf134497c9d611ad506ec48cc_0/about

Alternatively, it is also available from the California State Geoportal repository at:
https://gis.data.ca.gov/datasets/fbba842bf134497c9d611ad506ec48cc_0/explore

35 PWS from the SABL dataset were excluded from the analysis. Of these, 29 corresponded to PWS with jurisdictional system boundaries, and the remaining 6 PWS were located in areas where the census BG had a population of zero. Jurisdictional boundaries include areas that the PWS does not currently serve, such as areas where "will serve" letters have been issued or areas where the PWS is legally obligated to serve but it is not doing so. Most of the PWS listing jurisdictional areas are supplemental entries in the SABL dataset for PWS with existing service areas.

The remaining 4,768 PWS were divided into two groups:

(1) 3,143 PWS whose service areas intersected with only 1 BG (with the majority entirely located within the BG), and

(2) 1,625 PWS whose service areas intersected with two or more BGs (up to 2,968 BGs for the PWS with the largest service area).

For the PWS in the first group, racial and age group values were assigned directly from the corresponding BG where the PWS is located.

For the remaining 1,625 PWS spanning more than 1 BG, racial and age group values were calculated using either area-weighted or population-weighted interpolation methods to determine the relative contribution of each BG to the water system's service area. For population-weighted interpolation, block-level population data from the 2020 U.S. Census was used to determine the weights.

For each PWS, the interpolation method selected was the one that provided the closest result to the population value already included in the SABL dataset. In general, population-weighted interpolation appeared more accurate.

2. California Drinking Water System Locations (DWSL) dataset:

As of July 2024, the DWSL dataset provided the best-known geographical locations for 8,425 water systems including PWS and state small drinking water systems, which serve fewer than 25 people and are not regulated by the State.

The shapefile of the DWSL dataset was downloaded from the California State Geoportal repository at: https://gis.data.ca.gov/datasets/346d649d1e654737ac5b6855466e89b2_0/explore?location=36.792025%2C-120.674407%2C9.94

The dataset was reduced to 7,975 water systems after removing duplicate entries and systems without population information.

Of the remaining subset, 4,725 PWS were already included in the cleaned SABL dataset. The SABL dataset contained 43 water systems that were unique to that set. Assignment of demographic information for these PWS was done directly when processing the SABL dataset.

Virtually all remaining 3,250 water systems do not have documented service areas and are currently not in the SABL dataset. However, these systems are geocoded as small circular polygons (i.e., buffers) based on their geographic coordinates.

A spatial join was used to assign demographic information from the ACS by intersecting the location of the water system with the census BG where each system is located. 6 systems did not intersect with any BG, as they were located in areas where the BG had a population of zero. These were the same 6 PWS that were excluded from the SABL dataset.

Since systems in these subset are mapped as small nominal areas, the majority (2,465 systems) intersected with only one BG. The racial and age composition of the corresponding BG was then assigned to each system.

779 systems intersected with 2 to 4 BGs. Because the polygons of these systems are relatively small, arbitrary, and not true representations of their service areas, applying a weighting approach during the interpolation of demographic information was not practical. Instead, the racial profile of the BG with the largest area contribution was directly assigned to the water system.

R Libraries:

```
# Load potentially necessary libraries:
library(tidyverse)
library(stringr)
library(dplyr)
library(tidycensus)
library(tigris)
library(sf)
library(tmap)
library(RColorBrewer)
library(maptiles)
library(patchwork)
library(mapview)
library(crsuggest)
library(areal)

# To retrieve U.S. Census and ACS data using tidycensus commands, an API key
# is necessary. It only needs to be installed/run once using the code below.
# The current API key installed on this computer is below. It was obtained
# using the UC Davis email address by requesting it from this site:
# https://api.census.gov/data/key_signup.html

# Remove the "#" to activate the code. To run the code in a different computer
# would require a new API key using the above link.

# census_api_key("1a450b064c7b4572f5534b02e0137542f856e4bd", install = TRUE)

options(tigris_use_cache = TRUE)
```

ACS 2018-2022 Datasets

Racial information for California was obtained from the 5-year ACS 2018-2022 at the block group level.

Census data was retrieved using the R-based *tidycensus* package.

Table B03002: "Hispanic or Latino Origin by Race" was used as the data source.

The convenient geometry argument available through the *get_acs()* function in *tidycensus* was not used because the assigned BG geometry (i.e., the geographical information of the polygon) was not identical to that of the actual BG shapefile retrieved separately. The reason for this discrepancy was unknown.

Instead, ACS data was joined to the BG shapefile manually (see below).

```
# Get Table B03002 "Hispanic or Latino Origin by Race" from ACS 2018-2022
# for California at the Block Group level.

acs.race.bg <- get_acs(geography = "block group",
                       table = "B03002",
                       state = "CA",
                       year = 2022,
                       output = "wide",
                       geometry = FALSE,
                       cache = TRUE)

# Using geometry = FALSE returns a tibble of the data instead of a sf object (shapefile).
# The argument survey = "acs5" may be needed when retrieving
# larger census geographies in order to differentiate it from the 1-year ACS dataset.
```

The racial data of each BG was consolidated into the following 7 columns:

- Total population: *total_pop*.
- White (non-Hispanic): *white_*.
- African American (Black): *black_*.
- Asian, Native Hawaiian, and Pacific Islanders: *asian_pac*.
- Native American (American Indian), which includes Alaska Natives: *nat_am*.
- Latino or Hispanic of any race: *latino*.
- Other races, including multiple races: *other_mult*.

```
# Prepare the racial dataset.
# Create new columns with the desired racial categories and remove all other info
# (except the GEOID column):
```

```
race.bg.data <- acs.race.bg %>%
  mutate(total_pop = B03002_001E, # Total population per block group
         white_ = B03002_003E, # White, non-Hispanic/Latino
         black_ = B03002_004E, # Black, non-Hispanic/Latino
         asian_pac = B03002_006E + B03002_007E, # Asian + Hawaiian + Pacific Islander
         nat_am = B03002_005E, # Native American (American Indian + Alaska Native)
         latino = B03002_012E, # Hispanic/Latino
         other_mult = B03002_008E + B03002_010E + B03002_011E) %>% # Others + Multiple races
  select(-starts_with("B03002"), -NAME) # drops all original ACS columns
```

Similarly, aggregated racial information for California from the 5-year ACS 2018-2022 was retrieved at the state level. Additional columns were added to contain the percentages of each racial category.

```
# Get Table B03002 "Hispanic or Latino Origin by Race" from ACS 2018-2022
# for California at the state level.

acs.race.state <- get_acs(geography = "state",
                          table = "B03002",
                          state = "CA",
                          year = 2022,
                          output = "wide",
                          survey = "acs5",
                          geometry = FALSE,
                          cache = TRUE)

# Using geometry = FALSE returns a tibble of the data.
# The argument survey = "acs5" may be needed when retrieving
# larger census geographies in order to differentiate it from the 1-year ACS dataset.
```

```
# Prepare the racial dataset for CA at the state level.
# Create new columns with the desired racial categories and percentages:

race.state.data <- acs.race.state %>%
  mutate(total_pop = B03002_001E, # Total population for CA
         white_ = B03002_003E,
         whi_pct = 100 * B03002_003E/B03002_001E,
         black_ = B03002_004E,
         bla_pct = 100 * B03002_004E/B03002_001E,
         asian_pac = B03002_006E + B03002_007E,
         as.pa_pct = 100 * (B03002_006E + B03002_007E)/B03002_001E,
         nat_am = B03002_005E,
         nat_pct = 100 * B03002_005E/B03002_001E,
         latino = B03002_012E,
         lat_pct = 100 * B03002_012E/B03002_001E,
         other_mult = B03002_008E + B03002_010E + B03002_011E,
         mult_pct = 100 * (B03002_008E + B03002_010E + B03002_011E)/B03002_001E) %>%
  select(-starts_with("B03002")) # drops all original ACS columns
```

Age information for California was obtained from the 5-year ACS 2018-2022 at the block group level. The data was obtained from **Table B01001: "Sex by Age"**.

```
# Get Table B01001 "Sex by Age" from ACS 2018-2022
# for California at the Block Group level.

acs.age.bg <- get_acs(geography = "block group",
                      table = "B01001",
                      state = "CA",
                      year = 2022,
                      output = "wide",
                      geometry = FALSE,
                      cache = TRUE)

# Using geometry = FALSE returns a tibble of the data.
# The argument survey = "acs5" may be needed when retrieving
# larger census geographies in order to differentiate it from the 1-year ACS dataset.
```

The age data of each BG was consolidated into the following 6 columns:

- Total population: *total_pop*.
- Total male population: *total_male*.
- Total female population: *total_female*.
- Total population under 10 years of age (children): *under_10*.
- Total general population between 10 and 64 years of age: *gen_pop*.
- Total population of senior citizens of age 65 and above: *seniors*.

```
# Modify age dataset.
# Create new columns with the desired data:

age.bg.data <- acs.age.bg %>%
  mutate(total_pop = B01001_001E, # Total population per block group
         total_male = B01001_002E, # Total male population per block group
         total_female = B01001_026E, # Total female population per block group
         # Total under 10 years age group
         under_10 = B01001_003E + B01001_004E + B01001_027E + B01001_028E,
         # Total 10-64 age group
         gen_pop = B01001_005E + B01001_006E + B01001_007E + B01001_008E + B01001_009E +
           B01001_010E + B01001_011E + B01001_012E + B01001_013E + B01001_014E +
           B01001_015E + B01001_016E + B01001_017E + B01001_018E + B01001_019E +
           B01001_029E + B01001_030E + B01001_031E + B01001_032E + B01001_033E +
           B01001_034E + B01001_035E + B01001_036E + B01001_037E + B01001_038E +
           B01001_039E + B01001_040E + B01001_041E + B01001_042E + B01001_043E,
         seniors = B01001_020E + B01001_021E + B01001_022E + B01001_023E + # Total 65+ age group
```

```
          B01001_024E + B01001_025E + B01001_044E + B01001_045E + B01001_046E +
          B01001_047E + B01001_048E + B01001_049E) %>%
  select(-starts_with("B01001"), -NAME) # drops all original ACS columns
```

Age information for California from ACS 2018-2022 at the state level:

```
# Get Table B01001 "Sex by Age" from ACS 2018-2022
# for California at the State level.

acs.age.state <- get_acs(geography = "state",
                         table = "B01001",
                         state = "CA",
                         year = 2022,
                         output = "wide",
                         survey = "acs5",
                         geometry = FALSE,
                         cache = TRUE)

# Using geometry = FALSE returns a tibble of the data.
# The argument survey = "acs5" may be needed when retrieving
# larger geographies in order to differentiate it from the 1-year ACS dataset.
```

```
# Modify age dataset.
# Create new columns with the desired data and percentages:

age.state.data <- acs.age.state %>%
  mutate(total_pop = B01001_001E,
         total_male = B01001_002E,
         male_pct = 100 * B01001_002E/B01001_001E,
         total_female = B01001_026E,
         fem_pct = 100 * B01001_026E/B01001_001E,
         under_10 = B01001_003E + B01001_004E + B01001_027E + B01001_028E,
         und_10_pct = 100 * (B01001_003E + B01001_004E + B01001_027E + B01001_028E) / B01001_001E,
         gen_pop = B01001_005E + B01001_006E + B01001_007E + B01001_008E + B01001_009E +
           B01001_010E + B01001_011E + B01001_012E + B01001_013E + B01001_014E +
           B01001_015E + B01001_016E + B01001_017E + B01001_018E + B01001_019E +
           B01001_029E + B01001_030E + B01001_031E + B01001_032E + B01001_033E +
           B01001_034E + B01001_035E + B01001_036E + B01001_037E + B01001_038E +
           B01001_039E + B01001_040E + B01001_041E + B01001_042E + B01001_043E,
         ge_pop_pct = 100 * (B01001_005E + B01001_006E + B01001_007E + B01001_008E +
                             B01001_009E + B01001_010E + B01001_011E + B01001_012E +
                             B01001_013E + B01001_014E + B01001_015E + B01001_016E +
                             B01001_017E + B01001_018E + B01001_019E + B01001_029E +
                             B01001_030E + B01001_031E + B01001_032E + B01001_033E +
                             B01001_034E + B01001_035E + B01001_036E + B01001_037E +
                             B01001_038E + B01001_039E + B01001_040E + B01001_041E +
                             B01001_042E + B01001_043E) / B01001_001E,
         seniors = B01001_020E + B01001_021E + B01001_022E + B01001_023E +
           B01001_024E + B01001_025E + B01001_044E + B01001_045E + B01001_046E +
           B01001_047E + B01001_048E + B01001_049E,
         sen_65_pct = 100 * (B01001_020E + B01001_021E + B01001_022E + B01001_023E +
           B01001_024E + B01001_025E + B01001_044E + B01001_045E + B01001_046E +
```

```
          B01001_047E + B01001_048E + B01001_049E) / B01001_001E) %>%
  select(-starts_with("B01001"), -NAME) # drops all original ACS columns
```

Compiled Age and Race Profile for the State of California

```
age.data <- age.state.data %>% select(-c(GEOID, total_pop))

state.race.age <- cbind(race.state.data, age.data)

rm(age.data) # Remove intermediate variable.
```

California Census Geometry Shapefiles

All files were projected to the following coordinate system:

EPSG: 6414 [NAD83(2011) / California Albers (equal area)]

Block groups (BGs):

```
# Retrieve CA Census Block Groups.
# cb = TRUE downloads the Cartographic Boundary shapefile, which is clipped to
# the shorelines. This version is better for mapping.

CA.bg <- block_groups("CA") %>% st_transform(CA.bg, crs = 6414)
CA.bg.cb <- block_groups("CA", cb = TRUE) %>% st_transform(CA.bg.cb, crs = 6414)
```

Blocks:

```
# Retrieve CA blocks that will be used for the population-weighted interpolation:

CA.blocks <- tigris::blocks(state = "CA", year = 2022) %>% st_transform(CA.blocks, crs = 6414)
```

Census tracts and counties:

```
# Additional census geographies in case they are needed.
# cb = TRUE downloads the Cartographic Boundary shapefile, which is clipped to
# the shorelines. This version is better for mapping.

# Retrieve CA census tracts:
CA.tracts.cb <- tracts("CA", cb = TRUE) %>% st_transform(CA.tracts.cb, crs = 6414)
```

```r
# Retrieve CA counties:
CA.counties.cb <- counties(state = "CA", year = 2022, cb = TRUE) %>%
  st_transform(CA.counties.cb, crs = 6414)
```

SABL Dataset

Read in the SABL shapefile as downloaded from the SWRCB:

```
# Read in the SABL shapefile into R as a sf object:

sabl_raw <- st_read("Shapefiles/California_Drinking_Water_System_Area_Boundaries/California_Drinking_Wa
```

```
## Reading layer 'California_Drinking_Water_System_Area_Boundaries' from data source '/Users/pachefede/
##   using driver 'ESRI Shapefile'
## Simple feature collection with 4803 features and 34 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -124.3161 ymin: 32.53425 xmax: -114.1686 ymax: 41.99871
## Geodetic CRS:  WGS 84
```

```
# SABL file from July 5, 2024.
# Downloaded it from:
# https://hub.arcgis.com/datasets/fbba842bf134497c9d611ad506ec48cc_0/about
# The coordinate system is based on WGS 1984, EPGS: 4326. This is not a projection.
```

Exclude PWS with "Jurisdictional" boundaries or service areas.

Jurisdictional boundaries include areas that the PWS does not currently serve, such as areas where "will serve" letters have been issued or areas where the PWS is legally obligated to serve but it is not doing so. Most of these PWS with "Jurisdictional" areas are duplicates of PWS with regular service areas.

```
# Removing PWS with Jurisdictional areas:

sabl <- sabl_raw %>% filter(!BOUNDARY_T == "Jurisdictional")
```

29 PWS with Jurisdictional areas were excluded. This step eliminated duplicate entries with the same PWSID code.

Verify that there are no other duplicate PWS with identical PWSID codes:

```
# Check for PWS duplicates:

sabl.duplicates <- sabl %>%
  group_by(SABL_PWSID) %>%
  mutate(n = n()) %>%
  filter(n > 1) %>%
  arrange(SABL_PWSID)

# There are no duplicates.
```

Check also for PWS with equal areas but unique PWSID codes:

```r
# Check for area duplicates:

sabl.area.duplicates <- sabl %>%
  group_by(Shape__Are) %>%
  mutate(n = n()) %>%
  filter(n > 1) %>%
  arrange(Shape__Are)
```

16 PWS were identified:

- 5 PWS pairs, each with identical service areas.
- 6 PWS in Santa Clarita with identical service areas.

All these PWS were kept with their individual PWSID codes and system names. The code above saves this subset of PWS under the variable *sabl.area.duplicates*.

Projected Coordinate System

The SABL sf object (shapefile) was projected to the following coordinate system:

EPSG: 6414 [NAD83(2011) / California Albers (equal area)]

```r
sabl <- st_transform(sabl, crs = 6414)
```

Data Joins and Cleanup

After joining the demographic data with the BG geometry, BGs with 13 or fewer people in the total population were removed.

13 is the margin of error when the population value of the BG is 0.

98 BGs were excluded, 91 of which had a population of 0 (mostly BGs over the ocean).

Race only:

```
# Join CA.bg sf object (shapefile) and race.bg.data table.
# Remove BGs with population <= 13.

race.bg <- CA.bg %>%
  left_join(race.bg.data, by = "GEOID") %>%
  select(GEOID, total_pop:other_mult, ALAND:INTPTLON, geometry) %>%
  filter(!total_pop <= 13)
```

Age only:

```
# Join CA.bg sf object (shapefile) and age.bg.data table.
# Remove BGs with population <= 13.

age.bg <- CA.bg %>%
  left_join(age.bg.data, by = "GEOID") %>%
  select(GEOID, total_pop:seniors, ALAND:INTPTLON, geometry) %>%
  filter(!total_pop <= 13)
```

Race + Age:

```
# Join CA.bg sf object (shapefile) with both race and age datasets.

age.bg.temp <- age.bg %>% select(GEOID, total_male:seniors) %>% st_drop_geometry()

race_age.bg <- race.bg %>%
  left_join(age.bg.temp, by = "GEOID") %>%
  select(GEOID, total_pop:other_mult, total_male:seniors, ALAND:INTPTLON, geometry)
```

Spatial Join

All intersects between PWS in the SABL dataset and BGs:

```
# All PWS intersects

pws.all.intersects <- st_join(sabl,
                              race_age.bg,
                              join = st_intersects,
                              left = FALSE)

# The argument left = FALSE results in an inner spatial join returning only intersected PWS.
# Use join = within instead to find PWS completely within a single BG.
```

Intersected sections of each PWS were grouped by PWSID codes:

```
# This code summarizes the number of BGs intersected by each PWS:

pws.intersects_by_pwsid <- pws.all.intersects %>%
  st_drop_geometry() %>%
  group_by(SABL_PWSID) %>%
  summarize(BG = n()) %>%
  arrange(BG)
```

The following 6 PWS did not intersect with any BGs:

```
# These are the 6 PWS that did not intersect any BGs.
# Thus, they were not included in the analysis.

pws.non_intersected <- sabl %>% filter(!SABL_PWSID %in% pws.intersects_by_pwsid$SABL_PWSID)
```

Find the ID codes (i.e., PWSID values) of the PWS that intersected with only 1 BG, and those that intersected with 2 or more BGs:

```
# Filter for the IDs (SABL_PWSID column) of the PWS that intersect with only 1 BG (pws.ind.ids)
# and with multiple BGs (pws.mult.ids):

# 1 BG:
pws.ind.ids <- pws.intersects_by_pwsid %>% filter(BG == 1) %>% select(SABL_PWSID)

# Multiple BGs:
```

```r
pws.mult.ids <- pws.intersects_by_pwsid %>% filter(BG > 1) %>% select(SABL_PWSID, BG)

summary(pws.mult.ids$BG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.00    3.00   21.89   14.00 2968.00
```

- 3,143 PWS intersected with a single BG.
- 1,625 PWS intersected with more than 1 BG. The number of intersected BGs ranged from 2 to 2,968 with a median of 3 BGs and a mean of 21.9 BGs.

PWS intersecting 1 individual BG:

```r
# Use the IDs to filter for the PWS that only intersect with 1 BG:

pws.ind <- pws.all.intersects %>%
  filter(pws.all.intersects$SABL_PWSID %in% pws.ind.ids$SABL_PWSID)

# Create a file with less variables that should be equivalent to the variable
# interpolate.ind.aw obtained from the Spatial join (further below):

pws.ind.lim <- pws.ind %>%
  select(SABL_PWSID, POPULATION, SERVICE_CO, STATE_CLAS, total_pop:seniors)
```

PWS intersecting multiple BGs:

```r
# Similarly, subset the rest of PWS that intersected with multiple BGs:

pws.mult <- pws.all.intersects %>%
  filter(pws.all.intersects$SABL_PWSID %in% pws.mult.ids$SABL_PWSID)
```

For the 1,625 PWS intersecting multiple BGs, population numbers were interpolated using either area weights or population weights.

The sf objects (shapefiles) needed to be simplified as they could only have numerical variables (i.e., columns) for the interpolation functions to work.

Simplified SABL files:

```r
# Simplified SABL file to run the interpolation functions:

sabl.lim.all <- sabl %>%  select(SABL_PWSID, POPULATION, SERVICE_CO, STATE_CLAS, geometry) %>%
  filter(SABL_PWSID %in% pws.intersects_by_pwsid$SABL_PWSID)

sabl.lim.ind <- sabl.lim.all %>% filter(sabl.lim.all$SABL_PWSID %in% pws.ind.ids$SABL_PWSID)

sabl.lim.mult <- sabl.lim.all %>% filter(sabl.lim.all$SABL_PWSID %in% pws.mult.ids$SABL_PWSID)
```

```r
# race_age.bg:

# To use with the function st_interpolate_aw()
race_age.bg.aw <- race_age.bg %>%
  select(total_pop, white_, black_, asian_pac, nat_am, latino, other_mult,
         total_male, total_female, under_10, gen_pop, seniors, geometry)
```

Interpolation calculation for the PWS intersecting 1 BG. This should return the same information already contained in the sf object *pws.ind*.

```r
# Interpolate for the PWS within 1 BG:

interpolate.ind.aw <- st_interpolate_aw(race_age.bg.aw, sabl.lim.ind, extensive = FALSE) %>%
  mutate(SABL_PWSID = sabl.lim.ind$SABL_PWSID,
         POPULATION = sabl.lim.ind$POPULATION,
         SERVICE_CO = sabl.lim.ind$SERVICE_CO,
         STATE_CLAS = sabl.lim.ind$STATE_CLAS) %>%
  select(SABL_PWSID, POPULATION, SERVICE_CO, STATE_CLAS, total_pop:seniors)

# Choosing extensive = FALSE, interpolates the mean. However, since these only
# intersect with 1 BG it transfers the absolute demographic values for that BG.
```

Interpolation for the PWS intersecting multiple BGs. **Area-weighted** interpolation:

```r
# Interpolate for the PWS intersecting more than 1 BG:

interpolate.mult.aw <- st_interpolate_aw(race_age.bg.aw, sabl.lim.mult, extensive = TRUE) %>%
  mutate(SABL_PWSID = sabl.lim.mult$SABL_PWSID,
         POPULATION = sabl.lim.mult$POPULATION,
         SERVICE_CO = sabl.lim.mult$SERVICE_CO,
         STATE_CLAS = sabl.lim.mult$STATE_CLAS,
         total_pop_aw = total_pop) %>%
  select(SABL_PWSID, POPULATION, SERVICE_CO, STATE_CLAS, total_pop_aw, white_:seniors)
```

Population-weighted interpolation.

```
# Interpolate for the PWS intersecting more than 1 BG using population-based weights.

interpolate.mult.pw <- interpolate_pw(race_age.bg.aw,
                                      sabl.lim.mult,
                                      to_id = "SABL_PWSID",
                                      extensive = TRUE,
                                      weights = CA.blocks,
                                      weight_column = "POP20") %>%
  mutate(POPULATION = sabl.lim.mult$POPULATION,
         SERVICE_CO = sabl.lim.mult$SERVICE_CO,
         STATE_CLAS = sabl.lim.mult$STATE_CLAS,
         total_pop_pw = total_pop) %>%
  select(SABL_PWSID, POPULATION, SERVICE_CO, STATE_CLAS, total_pop_pw, white_:seniors)
```

Selecting between Area-weighted vs Population-weighted Interpolation

```
# Results comparison:

selection.table <- data.frame(PWSID = interpolate.mult.aw$SABL_PWSID,
                              POPULATION = interpolate.mult.aw$POPULATION,
                              area_weighted = interpolate.mult.aw$total_pop_aw,
                              pop_weighted = interpolate.mult.pw$total_pop_pw) %>%
  mutate(diff_aw = abs(area_weighted - POPULATION),
         diff_pw = abs(pop_weighted - POPULATION),
         choice = ifelse(is.na(diff_pw), "area",
                         ifelse(pop_weighted == 0, "area", # Needed this to catch one exception.
                                ifelse(diff_aw < diff_pw, "area", "pop"))),
         value = ifelse(diff_aw < diff_pw, area_weighted, pop_weighted)) %>%
  select(PWSID, POPULATION, area_weighted, diff_aw, pop_weighted, diff_pw, choice, value)
```

The interpolation method for PWS intersecting multiple BGs was chosen based on which method provided the closest population to the value already included in the SABL dataset.

This analysis relied on the key assumption that the population number associated with each PWS in the SABL dataset is accurate and representative of the community served by the PWS.

Using this approach, 1,003 of the 1,625 PWS in this subset were interpolated using population weights and the remaining 622 PWS were interpolated based on area weights.

Use the results included in the *selection.table* variable to select PWS in the subset:

```
# The 1,625 subset will be divided in two additional groups based on the interpolation method.
# First, select the corresponding PWSIDs.

selection.aw.ids <- selection.table %>% filter(choice == "area") %>% select(PWSID)
```

```r
selection.pw.ids <- selection.table %>% filter(choice == "pop") %>% select(PWSID)
```

```r
# PWS subgroup from area-weighted interpolation:

pws.mult.lim.aw <- interpolate.mult.aw %>%
  filter(interpolate.mult.aw$SABL_PWSID %in% selection.aw.ids$PWSID) %>%
  rename(total_pop = total_pop_aw)
```

```r
# PWS subgroup from population-weighted interpolation:

pws.mult.lim.pw <- interpolate.mult.pw %>%
  filter(interpolate.mult.pw$SABL_PWSID %in% selection.pw.ids$PWSID) %>%
  rename(total_pop = total_pop_pw)
```

Building a New SABL File with Demographic Information

These three R variables (files) contained the demographic information for the 4,768 PWS selected from the original SABL database:

- *pws.ind.lim*: 3,143 PWS that intersected with 1 block group.
- *pws.mult.lim.aw*: 622 PWS that intersected with multiple block groups (area-weighted interpolation).
- *pws.mult.lim.pw*: 1.003 PWS that intersected with multiple block groups (population-weighted interpolation).

These files were combined to create a new SABL file with demographic variables. However, these variables were initially expressed in absolute population numbers and needed to be converted to percentages. If absolute numbers are needed for other analyses, they should be calculated using these percentages and the *population* value of each PWS that is included in the SABL dataset.

```
# Combine the 3 files:

pws.compiled <- rbind(pws.ind.lim, pws.mult.lim.aw, pws.mult.lim.pw)
```

Calculate the demographic percentages:

```
pws.compiled <- pws.compiled %>%
  mutate(whi_pct = round(100 * (white_ / total_pop), 4),
         bla_pct = round(100 * (black_ / total_pop), 4),
         as.pa_pct = round(100 * (asian_pac / total_pop), 4),
         nat_pct = round(100 * (nat_am / total_pop), 4),
         lat_pct = round(100 * (latino / total_pop), 4),
         mult_pct = round(100 * (other_mult / total_pop), 4),
         und_10_pct = round(100 * (under_10 / total_pop), 4),
         ge_pop_pct = round(100 * (gen_pop / total_pop), 4),
         sen_65_pct = round(100 * (seniors / total_pop), 4),
         male_pct = round(100 * (total_male / total_pop), 4),
         fem_pct = round(100 * (total_female / total_pop), 4)) %>%
  select(SABL_PWSID, STATE_CLAS, SERVICE_CO, POPULATION, whi_pct:fem_pct,
         total_pop:seniors, geometry)
```

Drop the columns with the absolute numbers to clean it up:

```
sabl.pws.race <- pws.compiled %>% select(-(total_pop:seniors))
```

Build a sabl shapefile with all the original information in addition to the demographic percentages:

```
# Prepare the columns from the DSWL dataset that will be added to the SABL dataset.
sabl.temp <- sabl %>%
  st_drop_geometry() %>%
  filter(SABL_PWSID %in% pws.intersects_by_pwsid$SABL_PWSID) %>%
  select(SABL_PWSID, setdiff(colnames(sabl), colnames(sabl.pws.race)))
```

```
sabl.pws.race.all <- sabl.pws.race %>%
  left_join(sabl.temp, by = "SABL_PWSID") %>%
  select(OBJECTID_1, SABL_PWSID, BOUNDARY_T:WATER_SY_1, STATE_CLAS:fem_pct,
         CREATED_US:Shape__Len, geometry)
```

Drinking Water System Location (DWSL) Dataset

Read in the DWSL shapefile as downloaded from the SWRCB:

```
# Read in the SABL shapefile into R as a sf object:

dwsl_raw <-
  st_read("Shapefiles/California_Drinking_Water_System_Locations/California_Drinking_Water_System_Locati
          stringsAsFactors = FALSE)
```

```
## Reading layer 'California_Drinking_Water_System_Locations' from data source
##   '/Users/pachefede/Desktop/DOCUMENTOS/R-WORLD/Practicum/Shapefiles/California_Drinking_Water_System_
##   using driver 'ESRI Shapefile'
## Simple feature collection with 8425 features and 23 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -13838800 ymin: 3833645 xmax: -12705370 ymax: 5160786
## Projected CRS: WGS 84 / Pseudo-Mercator
```

```
# DWSL file from July 25, 2024.
# Downloaded it from the California State Geoportal at:
# https://gis.data.ca.gov/datasets/346d649d1e654737ac5b6855466e89b2_0/explore?location=36.792025%2C-120
# The coordinate system is based on WGS 1984, EPGS: 3857.
# This is a projection based on the WGS 1984 datum.
```

First, remove PWS that do not have population data (= NA's):

```
dwsl <- dwsl_raw %>% filter(!is.na(population))

# This step also filters out 29 systems that do not have PWSID values.
```

Check for duplicates, i.e., PWS with identical PWSID codes:

```
# Identify PWS duplicates:

dwsl.duplicates <- dwsl %>%
  group_by(pwsid) %>%
  mutate(n = n()) %>%
  filter(n > 1) %>%
  arrange(pwsid)

# There are several systems with multiple identical entries (up to 8).
```

Remove duplicate entries:

```
# Keep only unique entries of these duplicates.
# This function takes a couple of minutes to run.

dwsl <- dwsl %>% distinct(pwsid, .keep_all = TRUE)
```

This reduces the number of systems in the dataset from 8,425 systems to 7,975 systems.

Projected Coordinate System

The DWSL sf object was also projected to the following coordinate system:

EPSG: 6414 [NAD83(2011) / California Albers (equal area)]

```
dwsl <- st_transform(dwsl, crs = 6414)
```

Check how many systems are also in the SABL dataset:

```
# Also in the SABL dataset:
dwsl.in.sabl <- dwsl %>% filter(pwsid %in% sabl.pws.race$SABL_PWSID)

dwsl.not.in.sabl <- dwsl %>% filter(!pwsid %in% sabl.pws.race$SABL_PWSID)
```

4,725 PWS in the DWSL dataset are also in the SABL dataset. There are 43 systems in the SABL dataset that are unique (not found in the DWSL dataset).

3,250 PWS included in the DWSL dataset do not currently have documented service areas and are therefore not found in the SABL dataset. These systems are geocoded as circular polygons (buffers) around their geographic coordinates.

Spatial Join

Racial and age demographic information was assigned using a spatial join between the locations of the PWS and the BG. It was expected that most PWS would fall within a single BG because they did not have established service areas and the geocoded buffered polygons were quite small.

All intersects between PWS and BGs:

```
# All PWS intersects

dwsl.all.intersects <- st_join(dwsl.not.in.sabl,
                               race_age.bg,
                               join = st_intersects,
                               left = FALSE,
                               largest = TRUE)

# The argument left = FALSE results in an inner spatial join returning only intersected PWS.
```

```
# The argument largest = TRUE only assigns the data from the largest intersection
# (i.e., the BG that contributes the most area).
# This is for cases when the PWS intersects with 2 or more BGs.
```

Because the argument largest = TRUE was used in the spatial join, the effective result was that only the largest intersection was kept (i.e., only the BG contributing the most area to the intersection was considered).

Intersected sections of each PWS were grouped by PWSID codes:

```
# This code summarizes the number of BGs intersected by each PWS:

pws.intersects_by_dwsl <- dwsl.all.intersects %>%
  st_drop_geometry() %>%
  group_by(pwsid, shape_Area) %>%
  summarize(BG = n()) %>%
  arrange(BG)
```

In reality, there were 779 PWS that intersected with at least two and up to four BGs.

6 PWS did not intersect with any BG. These are the same 6 PWS that were excluded from the SABL dataset because they were located in areas where the corresponding BG had a population of zero. These 6 PWS were stored in the *dwsl.not.intersected* variable.

```
dwsl.not.intersected <- dwsl.not.in.sabl %>% filter(!pwsid %in% dwsl.all.intersects$pwsid)
```

In total, 3,244 PWS from the DWSL dataset were assigned racial and age demographic information.

Now calculate the demographic percentages:

```
dwsl.pws.race <- dwsl.all.intersects %>%
  mutate(whi_pct = round(100 * (white_ / total_pop), 4),
         bla_pct = round(100 * (black_ / total_pop), 4),
         as.pa_pct = round(100 * (asian_pac / total_pop), 4),
         nat_pct = round(100 * (nat_am / total_pop), 4),
         lat_pct = round(100 * (latino / total_pop), 4),
         mult_pct = round(100 * (other_mult / total_pop), 4),
         und_10_pct = round(100 * (under_10 / total_pop), 4),
         ge_pop_pct = round(100 * (gen_pop / total_pop), 4),
         sen_65_pct = round(100 * (seniors / total_pop), 4),
         male_pct = round(100 * (total_male / total_pop), 4),
         fem_pct = round(100 * (total_female / total_pop), 4)) %>%
  select(objectid:shape_Area, whi_pct:fem_pct, geometry)
```

Building the Complete Dataframe

As mentioned above, 4,725 PWS were found in both the SABL and DWSL datasets. However, PWS information of each system was not always consistent between the two datasets. For example, there were discrepancies in the population and service connection values for 560 PWS and 708 PWS, respectively. Other differences were found in variables such as system name or system classification type. For PWS found in both datasets, the SABL information was used as it appeared to be updated more recently.

The remaining PWS retained the information of their respective datasets. A field callled *dat_source* was introduced to specify the origin of the data for that particular system. The exception was the risk status and primary water source information, which was only found in the DWSL dataset. This additional information was added to the PWS in the SABL dataset.

First, use the SABL dataframe stored in the variable *sabl.pws.race.all* to start building the complete dataframe:

```
# Prepare the columns from the DSWL dataset that will be added to the SABL dataset.
dwsl.temp <- dwsl.in.sabl %>%
  st_drop_geometry() %>%
  rename(wat_source = primary_so,
         risk_stat = risk_statu) %>%
  select(pwsid, wat_source, risk_stat, service_ar, cpuc_regul, controllin)

# Add the columns from DWSL, rename columns, create additional columns and clean up.
pws.sabl <- sabl.pws.race.all %>%
  rename(objectid = OBJECTID_1,
         pwsid = SABL_PWSID,
         sys_name = WATER_SY_1,
         sta_class = STATE_CLAS,
         county = COUNTY,
         regulator = REGULATING,
         population = POPULATION,
         serv_conn = SERVICE_CO,
         admin_addr = ADDR_LINE_,
         admin_ad_1 = ADDR_LIN_1,
         admin_city = ADDRESS_CI,
         admin_stat = ADDRESS_ST,
         admin_zip = ADDRESS_ZI,
         phone = AC_PHONE_N,
         email = AC_EMAIL,
         owner_type = OWNER_TYPE,
         shape_Area = Shape__Are,
         shape_Leng = Shape__Len) %>%
  mutate(dat_source = "sabl") %>%
  left_join(dwsl.temp, by = "pwsid") %>%
  select(-WATER_SYST) %>%
  select(objectid, dat_source, pwsid, sys_name, sta_class, county, regulator, serv_conn,
         population, whi_pct:fem_pct, wat_source, risk_stat, service_ar,
         admin_addr:admin_zip, phone, email, owner_type, cpuc_regul, controllin,
         BOUNDARY_T, CREATED_US, CREATED_DA, LAST_EDITE, LAST_EDI_1, ACTIVITY_S,
         ACTIVITY_D, FEDERAL_CL, GLOBALID, BOUNDARY_F, DT_VERIFIE, VERIFIED_S,
         VERIFIED_N, VERIFIED_T, shape_Area, shape_Leng, geometry)
```

Second, rearrange the DWSL dataframe to look like the one above:

```
dwsl.pws.race.all <- dwsl.pws.race %>%
  rename(sys_name = system_nam,
         sta_class = system_typ,
         serv_conn = connection,
         wat_source = primary_so,
         risk_stat = risk_statu) %>%
  mutate(dat_source = "dwsl",
         BOUNDARY_T = NA,
         CREATED_US = NA,
         CREATED_DA = NA,
         LAST_EDITE = NA,
         LAST_EDI_1 = NA,
         ACTIVITY_S = NA,
         ACTIVITY_D = NA,
         FEDERAL_CL = NA,
         GLOBALID = NA,
         BOUNDARY_F = NA,
         DT_VERIFIE = NA,
         VERIFIED_S = NA,
         VERIFIED_N = NA,
         VERIFIED_T = NA) %>%
  select(objectid, dat_source, pwsid, sys_name, sta_class, county, regulator, serv_conn,
         population, whi_pct:fem_pct, wat_source, risk_stat, service_ar,
         admin_addr:admin_zip, phone, email, owner_type, cpuc_regul, controllin,
         BOUNDARY_T, CREATED_US, CREATED_DA, LAST_EDITE, LAST_EDI_1, ACTIVITY_S,
         ACTIVITY_D, FEDERAL_CL, GLOBALID, BOUNDARY_F, DT_VERIFIE, VERIFIED_S,
         VERIFIED_N, VERIFIED_T, shape_Area, shape_Leng, geometry)
```

Third, combine the two dataframes:

```
pws.race.age.all <- rbind(pws.sabl, dwsl.pws.race.all)
```

Fourth, calculate population values for each race and age group using the total **population values** from the SABL and DWSL datasets:

```
pws.race.age.all <- pws.race.age.all %>%
  mutate(white_ = round(population * (whi_pct/100), 0),
         black_ = round(population * (bla_pct/100), 0),
         asian_pac = round(population * (as.pa_pct/100), 0),
         nat_am = round(population * (nat_pct/100), 0),
         latino = round(population * (lat_pct/100), 0),
         mult_other = round(population * (mult_pct/100), 0),
         under_10 = round(population * (und_10_pct/100), 0),
         gen_pop = round(population * (ge_pop_pct/100), 0),
         seniors_65 = round(population * (sen_65_pct/100), 0),
         male = round(population * (male_pct/100), 0),
         female = round(population * (fem_pct/100), 0)) %>%
  select(objectid, dat_source, pwsid, sys_name, sta_class, county, regulator, serv_conn,
         population, white_, whi_pct, black_, bla_pct, asian_pac, as.pa_pct, nat_am,
```

```
          nat_pct, latino, lat_pct, mult_other, mult_pct, under_10, und_10_pct, gen_pop,
          ge_pop_pct, seniors_65, sen_65_pct, male, male_pct, female, fem_pct, wat_source,
          risk_stat, service_ar, admin_addr:admin_zip, phone, email, owner_type,
          cpuc_regul, controllin, BOUNDARY_T, CREATED_US, CREATED_DA, LAST_EDITE,
          LAST_EDI_1, ACTIVITY_S, ACTIVITY_D, FEDERAL_CL, GLOBALID, BOUNDARY_F, DT_VERIFIE,
          VERIFIED_S, VERIFIED_N, VERIFIED_T, shape_Area, shape_Leng, geometry)
```

The following is the same data with only the essential demographic information and fewer of the secondary columns (address, contact info, etc.):

```
pws.race.age <- pws.race.age.all %>%
  select(objectid, dat_source, pwsid, sys_name, sta_class, county, regulator,
         serv_conn, population, white_, whi_pct, black_, bla_pct, asian_pac, as.pa_pct,
         nat_am, nat_pct, latino, lat_pct, mult_other, mult_pct, under_10, und_10_pct,
         gen_pop, ge_pop_pct, seniors_65, sen_65_pct, male, male_pct, female, fem_pct,
         wat_source, risk_stat, shape_Area, shape_Leng, geometry)
```

Centroids

Convert all PWS polygons to points (centroids):

```
# Use st_centroid() function:

pws.centroids <- st_centroid(pws.race.age.all)

sabl.centroids <- st_centroid(sabl.pws.race.all)
```

These variables were created in case there was the need to plot the PWS represented by single points rather than using their service areas and/or buffers.

Output Files: Shapefiles and Accompanying Spreadsheets

PWS Dataset with Demographic Information

These files provide information on 8,012 PWS, including racial and age profile estimates of the populations served by each PWS.
PWS information was compiled from both the SABL and DWSL datasets.

1. File name: pws_race_age.shp Contains only PWS IDs, names, connections, population, risk status, water source, and demographic information. Coordinate system: WGS84 (EPSG: 4326) (not projected)
2. File name: pws_race_age_all.shp Contains all original field columns in the SABL and DWSL datasets, as well as the demographic information. Coordinate system: WGS84 (EPSG: 4326) (not projected)
3. File name: pws_race_age.csv Spreadsheet with the same columns as the pws_race_age shapefile (file #1 above). It does not include the geometry (geographical) information that converts into a shapefile.
4. File name: pws_race_age_all.csv Spreadsheet with all original field columns in the SABL and DWSL datasets (file #2 above), as well as the demographic information. It does not include the geometry (geographical) information that converts into a shapefile.

```
# 1.

# PWS shapefile with demographic (race and age) data.
# These files are not projected. The coordinate system is WGS84 (EPSG: 4326)

st_write(st_transform(pws.race.age, crs = 4326), "Output_Files/pws_race_age.shp",
         delete_layer = TRUE)
```

```
## Writing layer 'pws_race_age' to data source
##   'Output_Files/pws_race_age.shp' using driver 'ESRI Shapefile'
## Writing 8012 features with 35 fields and geometry type Multi Polygon.
```

```
# The argument "delete_layer = TRUE" appears to overwrite the data in any existing file.
# However, double check that the file contains the expected data. Otherwise delete the
# the old file in the folder before saving a new one.
```

```
# 2.

# PWS shapefile with all original field columns in the SABL and DWSL datasets
# in addition to the demographic data.
# These files are not projected. The coordinate system is WGS84 (EPSG: 4326)

st_write(st_transform(pws.race.age.all, crs = 4326), "Output_Files/pws_race_age_all.shp",
         delete_layer = TRUE)
```

```
## Writing layer 'pws_race_age_all' to data source
##   'Output_Files/pws_race_age_all.shp' using driver 'ESRI Shapefile'
## Writing 8012 features with 60 fields and geometry type Multi Polygon.
```

```
# The argument "delete_layer = TRUE" appears to overwrite the data in any existing file.
# However, double check that the file contains the expected data. Otherwise delete the
# the old file in the folder before saving a new one.
```

NOTE: When creating a new CSV file, it is **crucial** to delete any previous copies of the file in the Output folder. Otherwise the information of the new file will be appended to the old file instead of creating a brand new file.

```
# 3.
# Create accompanying csv file. It contains just essential information except for the
# geometry (geographical) info:

pws.race.age.df <- as.data.frame(st_drop_geometry(pws.race.age))

st_write(pws.race.age.df, "Output_Files/pws_race_age.csv", delete_layer = TRUE)
```

```
## Writing layer 'pws_race_age' to data source
##   'Output_Files/pws_race_age.csv' using driver 'CSV'
## Writing 8012 features with 35 fields without geometries.
```

```
# The argument "delete_layer = TRUE" IS NOT working for the csv file.
# Delete the old file in the folder before saving a new copy. Otherwise the new versions
# is appended to the old file.

rm(pws.race.age.df) # Remove intermediate variable
```

```
# 4.
# Create accompanying csv file. It contains all information column except for the
# geometry (geographical) info:

pws.race.age.all.df <- as.data.frame(st_drop_geometry(pws.race.age.all))

st_write(pws.race.age.all.df, "Output_Files/pws_race_age_all.csv", delete_layer = TRUE)
```

```
## Writing layer 'pws_race_age_all' to data source
##   'Output_Files/pws_race_age_all.csv' using driver 'CSV'
## Writing 8012 features with 60 fields without geometries.
```

```
# The argument "delete_layer = TRUE" IS NOT working for the csv file.
# Delete the old file in the folder before saving a new copy. Otherwise the new versions
# is appended to the old file.

rm(pws.race.age.all.df) # Remove intermediate variable
```

SABL Dataset + Demographic Information (percentages only)

These files include information on 4,768 PWS in the SABL dataset.

5. File name: sabl_race_age.shp Contains only PWS IDs, names, connections, population and demographic percentages. Coordinate system: WGS84 (EPSG: 4326) (not projected)

6. File name: sabl_race_age_all.shp Contains all original field columns in the SABL dataset and the additional demographic percentages. Coordinate system: WGS84 (EPSG: 4326) (not projected)

7. File name: sabl_race_age_all.csv Spreadsheet with all original field columns in the SABL dataset and the additional demographic percentages. It does not include the geometry (geographical) information that is included in the shapefile.

```
# 5.

# SABL shapefile with demographic data.
# These files are not projected. The coordinate system is WGS84 (EPSG: 4326)

st_write(st_transform(sabl.pws.race, crs = 4326), "Output_Files/sabl_race_age.shp",
         delete_layer = TRUE)
```

```
## Writing layer 'sabl_race_age' to data source
##   'Output_Files/sabl_race_age.shp' using driver 'ESRI Shapefile'
## Writing 4768 features with 15 fields and geometry type Multi Polygon.
```

```
# The argument "delete_layer = TRUE" appears to overwrite the data in any existing file.
# However, double check that the file contains the expected data. Otherwise delete the
# the old file in the folder before saving a new one.
```

```
# 6.

# SABL shapefiles with all original field columns in the SABL dataset and demographic data.
# These files are not projected. The coordinate system is WGS84 (EPSG: 4326)

st_write(st_transform(sabl.pws.race.all, crs = 4326), "Output_Files/sabl_race_age_all.shp",
         delete_layer = TRUE)
```

```
## Writing layer 'sabl_race_age_all' to data source
##   'Output_Files/sabl_race_age_all.shp' using driver 'ESRI Shapefile'
## Writing 4768 features with 45 fields and geometry type Multi Polygon.
```

```
# The argument "delete_layer = TRUE" appears to overwrite the data in any existing file.
# However, double check that the file contains the expected data. Otherwise delete the
# the old file in the folder before saving a new one.
```

NOTE: When creating a new CSV file, it is **crucial** to delete any previous copies of the file in the Output folder. Otherwise the information of the new file will be appended to the old file instead of creating a brand new file.

```
# 7.
# Create accompanying csv file. It contains all information column except for the
# geometry (geographical) info:

sabl.pws.race.all.df <- as.data.frame(st_drop_geometry(sabl.pws.race.all))

st_write(sabl.pws.race.all.df, "Output_Files/sabl_race_age_all.csv", delete_layer = TRUE)
```

```
## Writing layer 'sabl_race_age_all' to data source
##   'Output_Files/sabl_race_age_all.csv' using driver 'CSV'
## Writing 4768 features with 45 fields without geometries.
```

```
# The argument "delete_layer = TRUE" IS NOT working for the csv file.
# Delete the old file in the folder before saving a new copy. Otherwise the new versions
# is appended to the old file.

rm(sabl.pws.race.all.df) # Remove intermediate variable
```

The code below produces the same shapefiles as above, but the coordinate system is projected to the California Albers coordinate system (EPSG: 6414), which allows for area-based calculations.

8. File name: sabl_race_age_ESPG-6414.shp Contains only PWS IDs, names, connections, population and demographic percentages. Coordinate system: NAD83(2011) / California Albers (EPSG: 6414) (projected)
9. File name: sabl_race_all_age_ESPG-6414.shp Contains all original field columns in the SABL dataset and the additional demographic percentages. Coordinate system: NAD83(2011) / California Albers (EPSG: 6414) (projected)

```
# Create a new SABL shapefiles with demographic data.
# These files are projected to NAD83(2011) / California Albers (EPSG: 6414)

# 8.
st_write(sabl.pws.race, "Output_Files/sabl_race_age_ESPG-6414.shp", delete_layer = TRUE)
```

```
## Writing layer 'sabl_race_age_ESPG-6414' to data source
##   'Output_Files/sabl_race_age_ESPG-6414.shp' using driver 'ESRI Shapefile'
## Writing 4768 features with 15 fields and geometry type Multi Polygon.
```

```
# 9.
st_write(sabl.pws.race.all, "Output_Files/sabl_race_age_all_ESPG-6414.shp",
         delete_layer = TRUE)
```

```
## Writing layer 'sabl_race_age_all_ESPG-6414' to data source
##   'Output_Files/sabl_race_age_all_ESPG-6414.shp' using driver 'ESRI Shapefile'
## Writing 4768 features with 45 fields and geometry type Multi Polygon.
```

```
# The argument "delete_layer = TRUE" appears to overwrite the data in any existing file.
# However, double check that the file contains the expected data. Otherwise delete the
# the old file in the folder before saving a new one.
```

Removing Intermediate Variables from R Studio's Environment

If needed, the global environment in R Studio can be cleaned up by running the code in the chunk below to remove intermediate variables that are not useful as final outputs.

Warning: *Any changes to previous code lines, such as altering or deleting variables will result in the unintended removal of useful variables. This may also render the file creation functions below to fail.*

Additionally, executing this code will prevent individual code chunks that depend on these variables from running again. If updated files are needed after removing the intermediate variables, the entire code should be run in sequence from the beginning.

This code can be improved by compiling a growing list of intermediate variables as soon as they are generated. Currently, it depends on the order of variables in the environment, which can be obtained using the function ls().

```
# Delete intermediate variables to clean up the Global Environment
# To run the code, remove the "#" in the next line to activate the code line.

# rm(list = ls()[c(1:8, 15, 22:27, 29:38, 43:48, 51:54, 57:60)])
```

Other Interpolation Functions

Areal Package

This is another R package containing area-weighted interpolation functions. The results obtained with this package were not used in the analysis as they were inconsistent with the other methods.

For the 1,625 PWS intersecting with multiple BGs, the interpolation results obtained with the areal functions were much higher than those obtained with the equivalent area-weighted function *st_interpolate_aw* despite both methods using the the same approach.

For consistency, the areal results were not used. Interpolation results were selected from those obtained with either the area-weighted *st_interpolate_aw* or the population-weighted *interpolate_pw* functions as described previously.

These functions also took significantly longer times to execute.

```
# To use with aw_interpolate() from the areal package:

race_age.bg.areal <- race_age.bg %>%
  select(GEOID, total_pop, black_, white_, asian_pac, nat_am, latino, other_mult,
         total_male, total_female, under_10, gen_pop, seniors, geometry)
```

```
# Interpolate (area-weight) using the function from the areal package.
# For the PWS intersecting 1 BG use the intensive argument.

interpolate.ind.areal <-aw_interpolate(sabl.lim.ind,
                                       tid = "SABL_PWSID",
                                       source = race_age.bg.areal,
                                       sid = "GEOID",
                                       weight = "sum",
                                       output = "sf",
                                       intensive = c("total_pop", "white_", "black_",
                                                     "asian_pac", "nat_am", "latino",
                                                     "other_mult", "total_male",
                                                     "total_female", "under_10",
                                                     "gen_pop", "seniors"))
```

```
# Interpolate (area-weight) using the function from the areal package.
# For the PWS intersecting more than 1 BG use the extensive argument.
# (This function takes several minutes for this subset).

interpolate.mult.areal <- aw_interpolate(sabl.lim.mult,
                                         tid = "SABL_PWSID",
                                         source = race_age.bg.areal,
                                         sid = "GEOID",
                                         weight = "sum",
                                         output = "sf",
                                         extensive = c("total_pop", "white_", "black_",
                                                       "asian_pac", "nat_am", "latino",
                                                       "other_mult", "total_male",
                                                       "total_female", "under_10",
                                                       "gen_pop", "seniors"))
```

```
# The function ar_validate() below can be used to troubleshoot the aw_interpolate() function.

# ar_validate(race_age.bg.areal,sabl.lim.mult, c("total_pop", "white_", "black_",
# "asian_pac", "ind_alask", "latino", "other_mult", "total_male", "total_female",
# "under_10", "gen_pop", "seniors"), method = "aw", verbose = TRUE)
```