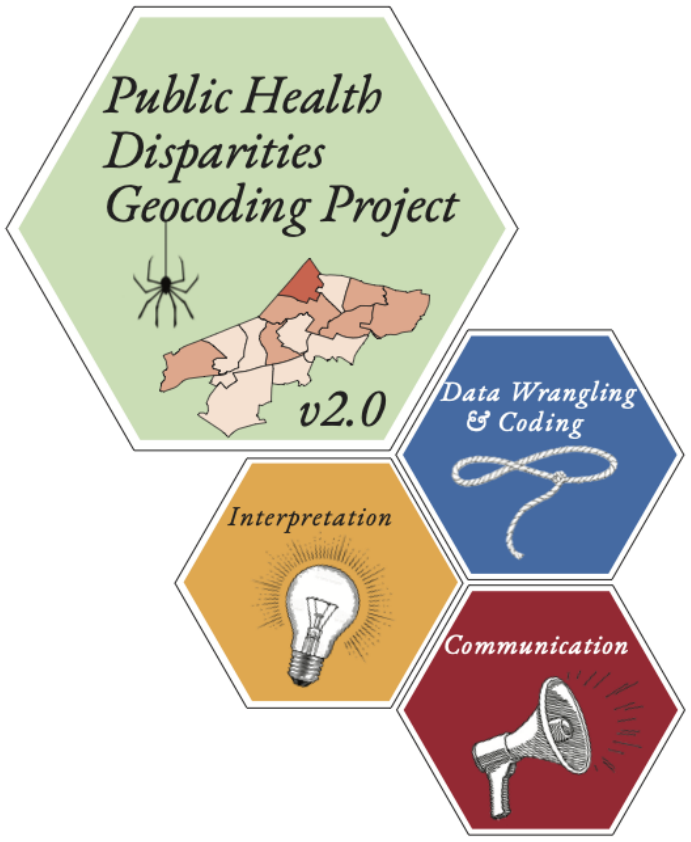
Public Health Disparities Geocoding Project 2.0 Training Manual

From the Harvard T.H. Chan School of Public Health, Boston MA

Authors: Christian Testa, Jarvis T. Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D. Waterman, Nancy Krieger

November 2022



HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

Visit the Public Health Disparities Geocoding Project website!

Source of funding: American Cancer Society Clinical Research Professor Award to N. Krieger

Please cite this work as: Testa C, Chen JT, Hall E, Javadi D, Morgan J, Rushovich T, Saha S, Waterman PD, Krieger N. The Public Health Disparities Geocoding Project 2.0. Training Manual. Available as of October 30, 2022

Preface »

Preface

By: Nancy Krieger PhD, Dena Javadi

In June 2004, a team based at the Harvard School of Public Health published a monograph (Krieger et al, 2004) based on the research and trainings we created for our *Public Health Disparities Geocoding Project* (Krieger et al, 2004). This project built on and systematized approaches the team members had been developing, since the early 1990s, for using geocoding and area-based socioeconomic measures to overcome the absence of socioeconomic data in most US health records (Krieger, 1992; Krieger, 1998, Krieger et al, 2004).

The objective was to boost efforts to document and inform efforts to address socioeconomic inequities in health, overall and in relation to US racialized health inequities.

A note on capitalization convention used for racial groups: "Mindful and respectful of different views among anti-racist scholars and activists regarding conventions for designating US racialized groups,[1-3] and specifically over whether to use a "w" or "W" for the group "white" / "White" (in contrast to no disagreement about capitalizing the "b" for "Black"), in this manual we have opted to employ the terminology and conventions of the primary source of the data for our case studies: the US Census Bureau, which capitalizes the names of each group. One set of arguments in favor of an upper-case "w" for "White"[1, 2] are that "[t]o not name 'White' as a race is, in fact, an anti-Black act which frames Whiteness as both neutral and the standard.[...] and removes accountability from White people's and White institutions' involvement in racism."[2] Conversely, arguments in favor of the lower-case "w" for "white" and against white supremacy note that: (1) White-supremacists capitalize the "w" in "White,"[1] and (2) "leaving white in lowercase represents a righting of a long-standing wrong and a demand for dignity and racial equity."[1,3]

[1] Appiah KA. The Case for Capitalizing the B in Black. The Atlantic. June 18, 2020. Accessed on 10/25/2022 at https://www.theatlantic.com/ideas/archive/2020/06/time-to-capitalize-blackand-white/613159/

[2] Thúy Nguyễn A. and Pendleton M. Recognizing Race in Language: Why We Capitalize "Black" and "White". Center for the Study of Social Policy. March 23, 2020. Accessed on 10/25/2022 at https://cssp.org/2020/03/recognizing-race-in-language-why-we-capitalize-black-and-white/

[3] Price A. Spell It with a Capital "B". Insight Center for Community Economic Development. Oct 1, 2019. Accessed on 10/25/2022 at https://insightcced.medium.com/spell-it-with-a-capital-b-9eab112d759a.

Oriented to academic researchers, health department staff, cancer registries, and public health students, this training offered a solution to the problem of lack of socioeconomic data in US public health surveillance systems: the geocoding of public health surveillance data and using census-derived area-based socioeconomic measures (ABSMs) to characterize both the cases and populations in the catchment area, thereby enabling computation of rates stratified by the area-based measure of socioeconomic position. In addition to informing the analyses of numerous scientific investigations, the work of the Public Health Disparities Geocoding Project was adopted by numerous US state and local health departments and cancer registries, and also informed the decisions of the US National Cancer Institute's cancer registry system to geocode their data to the census tract level and enable researchers to access not only county-level ABSMs (public access data) but also census tract level ABSMs (restricted data).

Since 2004, there has been a huge increase in conceptual and methodological work regarding use of ABSMs to document and analyze health inequities, and the computing tools and statistical methods to measure them have evolved. In particular, the availability of software that facilitates data access, mapping, visualization, and fitting of multilevel and spatial models has greatly enhanced the accessibility of these analytic methods for public health scientists, advocates, activists, and policy makers interested in advancing health equity. The updated *Public Health Disparities Geocoding Project 2.0* training (PHDGP2.0, 2022), held in June and July 2022 and now available through this manual, builds on our prior 2004 training, and expands on why & how to analyze population health and health inequities in relation to census tract, county, and other georeferenced societal and environmental data.

This online manual serves as a guide to help users explore the following topics:

- the history and context of, and rationale for, conducting this type of work
- getting data from the Census and other sources
- visualizing geocoded data and social metrics

- conducting analyses with data aggregated to a specified level of geography vs. multi-level analyses with 2 or more levels of geography
- interpreting data through a health equity lens

The PHDGP 2.0 manual also offers three case examples that go through analysis of mortality data (including aggregation and spatial analysis), and two case examples to illustrate the use of aggregated data from the American Community Survey (ACS) and the CDC's PLACES dataset. To access the data for these case studies, please visit the PHDGP 2.0 website (https://www.hsph.harvard.edu/thegeocodingproject/save-the-date-the-public-health-disparities-geocoding-project-2-0/).

The theory informing the work of this project is the ecosocial theory of disease distribution, first proposed in 1994 and elaborated since (Krieger, 1994; Krieger 2011, Krieger 2021). This theory not only provides conceptual tools for analyzing multilevel spatiotemporal processes of embodying (in)justice, but also calls attention to who and what is responsible for health inequities, and who holds or blocks agency and accountability to achieve health justice (Krieger, 1994; Krieger 2011, Krieger 2021). These course materials are written with as much emphasis on monitoring health for accountability and action as they are for etiological studies.

REFERENCES

Krieger N. Ecosocial Theory, Embodied Truths, and The People's Health. New York: Oxford University Press, 2021.

Krieger N. Epidemiology and The People's Health: Theory and Context. New York: Oxford University Press, 2011.

Krieger N, Waterman PD, Chen JT, Rehkopf DH, Subramanian SV. The Public Health Disparities Geocoding Project Monograph. Available as of June 30, 2004 at: http://www.hsph.harvard.edu/thegeocodingproject

Krieger N, Waterman PD, Chen JT, Rehkopf DH, Subramanian SV. The Public Health Disparities Geocoding Project Monograph: Publications. Available as of June 30, 2004 at: https://www.hsph.harvard.edu/thegeocodingproject/publications/

Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med. 1994 Oct;39(7):887-903. doi: 10.1016/0277-9536(94)90202-x.

Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. Am J Public Health. 1992 May;82(5):703-10. doi: 10.2105/ajph.82.5.703.

Krieger N (PI). Area-based socioeconomic measures for health data. NIH (NICHD) R01 HD3865-01 (1998-2003).

PHDGP.The Public Health Disparities Geocoding Project 2.0. Available as of March 15, 2022 at: https://www.hsph.harvard.edu/thegeocodingproject/save-the-date-the-public-health-disparities-geocoding-project-2-0/

« Public Health Disparities Geocoding Project 2.0 Training Manual

1 Background and History of Analytic Methods »

1 Background and History of Analytic Methods

By: Nancy Krieger PhD, Dena Javadi

In 2004, a team based at the Harvard T.H Chan School of Public Health's Department of Social and Behavioral Sciences published a Project Monograph describing the motivation behind the *Public Health Disparities Geocoding Project* and its analytic approaches, both conceptually and methodologically. The monograph dove into methods of geocoding, generating area-based social metrics (ABSMs), Multi-level Modeling, data visualization, and basic epidemiologic methods for generating descriptive statistics, with the intent of providing population health scientists, health department staff, cancer registries, policy makers, and advocates and activists for health justice with tools to put on the map – literally – rampant but underreported socioeconomic inequities in health and their links to racialized and spatial health inequities (Krieger et al, 2005; Krieger, 2009).

Key publications on these topics can also be found on the project's publications page.

The *Public Health Disparities Geocoding Project* is informed by the ecosocial theory of disease distribution, developed by Dr. Nancy Krieger in 1994, specifically "its focus on how people literally biologically embody their societal and ecological context at multiple levels, across the life course and historical generations" (Krieger, 2012).

Administrative health data, disease surveillance systems, and routine health surveys are important tools in understanding disease distribution and informing public health programming, advocacy, and policy development. However, the social patterning of disease distribution is often obscured by a lack of robust social metrics, including socioeconomic data, pertinent to understanding health inequities - defined as unfair, avoidable, and preventable health differences across social groups (Krieger, 2011). These social groups, co-defined by social relationships involving power, are among the many groups which together comprise the "populations" that embody health, experience health injustice (or health justice), and are the focus of public health monitoring, research, and action. As defined by Krieger (2012), "populations are dynamic beings constituted by intrinsic relationships both among their members and with the other populations that together produce their existence." Addressing health inequities across social groups and within populations accordingly requires data on the population-defined and defining relationships and characteristics that create and are created by structures and systems. Further, theories of disease distribution and the underlying agendas, ideologies, and motivations contributing to their implicit or explicit use in turn shape what data get analyzed, how the analysis is interpreted, and what visualizations are used to disseminate findings (Krieger, 2011). Misuse or poor use of data analysis and visualization tools can contribute to obscuring health inequities, leaving out certain subpopulations or misrepresenting trends in disease distribution, resulting in poor policy decisions and inadequate or misleading data to inform community and advocacy organizing for health justice. Therefore, not only is the availability of robust health information systems important, so too is the use of appropriate methods and a health equity lens in their analyses.

Using routine information systems to inform disease prevention is not a 21st or even 20th century concept. In 1829, William Farr, a "Compiler of Abstracts" at the General Register included a letter to the Register's first report which stated that

> *"Diseases are more easily prevented than cured, and the first step to their prevention is the discovery of their exciting causes. The Registry will show the agency of these causes by numerical facts and measure the intensity of their influence and will collect information on the laws of vitality with the variation in these laws in the two sexes at different ages and the influence of civilisation, occupation, locality, seasons and other physical agencies whether in generating diseases and inducing death or in improving the public health" (Whitehead, 2000).*

In the US, linking of public health data to US census-based socioeconomic data was carried out by the National Tuberculosis Association in the 1920s and 1930s (Green, 1932; Nathan, 1932). Similarly, cancer epidemiologists have used geocoded data to generate and stratify cancer incidence, categorizing social groups using variables defined in relation to "race/ethnicity," sex, and socioeconomic position, for many decades (Krieger, 2001). Of note, US health data have long been reported by US government agencies (federal, state, and local) stratified by "race" and "sex," informed by a long history of biological essentialism that treats these variables as a matter of innate biology, with no attention to inequitable racial, gender, or class relations (Krieger, 2021; Hunter et al, 2005). Adding socioeconomic data to the mix can aid with understanding the contribution of socioeconomic inequities to racialized and gender health inequities, but with the caveat that the 20th century CE framework of eugenics (whose shadow continues to be cast well into the 21st c CE) also has held that people's socioeconomic position reflects their genetic inheritance (Krieger, 2018; Levine, 2017).

Despite these early emphases on social metrics as critical in understanding inequitable and differential population rates and distributions of morbidity and mortality, the integration of socioeconomic data in national surveillance systems has been slow. A recent OECD report on national monitoring systems for health inequalities by socioeconomic status found that only seven of the 26 high-income countries included in the study had national routine monitoring systems with regular reports on socioeconomic inequalities in health over time (Frank and Matsunaga, 2020).

When the first *Public Health Disparities Geocoding Project* was launched, it presented a solution in the form of area-based socioeconomic measures (ABSMs) where multilevel approaches to understanding area-based measures, classified by socioeconomic characteristics, could be used to calculate stratified rates and render the invisible, visible. The project articulated the lack of a standardized approach in the choice of geographic levels and types of ABSMs used for monitoring disease distribution, making comparison across heterogeneous methods difficult.

The project took on the task of identifying which ABSMs would be most apt for monitoring socioeconomic inequalities in health and at what geographic level. Findings suggested that census tract poverty level - defined as "percent of persons below poverty" - was most apt (Krieger et al, 2003).

Since then, there has been significant development in both the conceptualization of geocoded health disparities, types of ABSMs, and the technologies available to capture, analyze, and visualize them.

Many ABSMs have been developed around the world. Globally, the Gini coefficient is one of the standards for measuring income inequality (with caveats around its use beyond larger aggregations and issues around spatial social polarization) (Shaw et al, 2007; Krieger et al, 2016). In Canada, examples include the Socioeconomic Factor Index (SEFI), the General Deprivation Index (GDI), and the Deprivation Index for Health and Welfare Planning for Quebec (DIHWPQ) (Schuurman et al, 2007). Starting in the early 20th c CE, the UK began using the Registrar General's social class classification systems (an ad hoc approach based on skill-level demarcations in occupational class), which in was replaced in 2000 by the theoretically-grounded National Statistics Socio-economic Classification (NS-SEC), which emphasizes employment relations and the conditions of occupations (UK Office of National Statistics, 2022). Also commonly used is the English Index of Multiple Deprivation (McLennan et al, 2019). In the US, studies have generated or used composite indices of deprivation or social vulnerability based on selected census variables (O'Campo et al, 2008; Messer et al, 2006; Hu et al, 2021; ATSDR 2022).

However, one problem common to many indices, and also single-variable measures (such as percent below poverty), is that they do not provide insight into the power relations and spatial social polarization driving health inequities (Krieger et al, 2016; Krieger et al, 2017). For example, the metric "percent below poverty," while useful for describing socioeconomic gradients in health, notably provides no information on the income distribution of those "above poverty," which can range from barely above poverty to extremely affluent. Similarly, a commonly used variable in the US, such as "percent of population classified as being Black Americans" says nothing about the distribution of other racialized groups and the social relationships that are core to racialized economic segregation. An additional problem is that diverse metrics intended to measure inequality across the full population distribution, such as the Gini coefficient or the dissimilarity index (which measures the proportion of a population that would need to move within a geographic area to achieve evenness of distribution), is that they are only meaningful at higher geographic levels (Krieger et al, 2016; Krieger et al, 2017). At issue is how, within the US, policies and practices, past and present, to generate and enforce racialized economic segregation have worked to buttress neighborhood boundaries, especially to keep some areas White and affluent and relegate lower-

income populations, disproportionately Black, Latinx, Indigenous and immigrant in the US to underdeveloped neighborhoods lacking resources for people to thrive (Krieger et al, 2016; Krieger et al, 2017; Rothstein, 2017; Bailey et al, 2017).

A new approach to capturing the extreme range of concentrations of economic deprivation and privilege, termed the "Index of Concentration at the Extremes" (ICE), was developed in 2001 by Douglass Massey, a leading US scholar on residential segregation (Massey, 2001; Massey, 1996; Massey, 2012). This measure, which ranges from -1 to 1 and captures the extent to which an area's population is concentrated into one end of the other of extremes of privilege and deprivation, notably can be used meaningfully at multiple levels of geography, from census block on up to counties and higher. In recent years, members of our *Public Health Disparities Geocoding Project* team have produced work promoting use of the ICE in public health research, and also extending Massey's original work to develop ICE measures that quantify not only racialized residential segregation but also racialized economic segregation, with the latter comprising the first metric of its kind (Krieger et al, 2016; Krieger et al, 2016B; Scally et al, 2018; Krieger et al, 2018; Chen & Krieger, 2021; Krieger et al, 2015). The intent is to provide insight into who and what drives health inequities, not just focus solely on those harmed (Krieger et al, 2010; Beckfield, 2018; Bambra et al, 2021; Bailey et al, 2017).

INDEX OF CONCENTRATION AT THE EXTREMES (ICE) (Massey, 2001)

Formula:

$$ICE_i = (A_i - P_i)/T_i$$

where, say,

A_i = N of high income persons in neighborhood
P_i = N of low income persons in neighborhood
T_i = total N with known income in neighborhood

range: -1 (total deprivation) to 1 (total privilege)

-- typically computed for income, also education
-- simultaneously provides information on the groups at both extremes (not just one or the other)

ICE and monitoring & analyzing health inequities: novel extension to ICE for racialized segregation and ICE for racialized economic segregation

- Krieger N, Waterman PD, Spasojevic J, Li W, Maduro G, Van Wye G. Public health monitoring of privilege and deprivation using the Index of Concentration at the Extremes (ICE). *Am J Public Health* 2016; 106: 256-253
- Krieger N, Waterman PD, Gryparis A, Coull BA. Black carbon exposure, socioeconomic and racial/ethnic spatial polarization, and the Index of Concentration at the Extremes (ICE). *Health & Place* 2015; 34:215-228.
- Krieger N, Feldman JM, Waterman PD, Chen JT, Coull BA, Hemenway D. Local residential segregation matters: stronger association of census tract compared to conventional city-level measures with fatal and non-fatal assaults (total and firearm related), using the Index of Concentration at the Extremes (ICE) for racial, economic, and racialized economic segregation, Massachusetts (US), 1995-2010. *J Urban Health* 2017; 94:244-258.
- Krieger N, Waterman PD, Batra N, Murphy JS, Dooley DP, Shah SN. Measures of local segregation for monitoring health inequities by local health departments. *Am J Public Health* 2017; 107:903-906.
- Krieger N, Kim R, Feldman J, Waterman PD. Using the Index of Concentration at the Extremes at multiple geographic levels to monitor health inequities in an era of growing spatial social polarization: Massachusetts, USA (2010-2014). *Int J Epidemiol* 2018; 47:788-819.
- Krieger N, Feldman JM, Kim R, Waterman PD. Cancer incidence and multilevel measures of residential economic and racial segregation for cancer registries. *JNCI Cancer Spectrum* 2018 Apr 25; 2(1):pky009.

The Covid-19 pandemic - due to the social and geographic patterning in the spread of the virus and associated hospitalizations and deaths - has highlighted the critical need for improved surveillance systems and systematic monitoring of health inequities (Presidential Task Force, 2021; Bambra, 2021). As a result, the *Public Health Disparities Geocoding Project* has compiled resources to support efforts in carrying out the analyses of health inequities in this context (Krieger, Chen, Waterman, 2020).

The 2020 update to the project, shared in May 2020 in the thick of the first months of the pandemic, provides:

- List of conceptual and empirical publications
- List of variables constructed using the US Census American Community Survey (ACS) data
- R code for extracting ABSMs from the ACS and replicating analyses in the published empirical papers

The *Public Health Disparities Geocoding Project 2.0* training (held online in June and July 2022) and now available through this manual builds on the work of our team throughout the pandemic and is offering an updated and revised training on why & how to analyze population health and health inequities in relation to census tract, county, and other georeferenced societal and environmental data. Because the area-based metrics we employ in this study include diverse social metrics, no longer restricted solely to economic measures, we employ the updated terminology of "area-based social metrics" – which we continue to abbreviate as "ABSMs."

This online manual will walk through each step of the training including:

- Chapter 2: Getting Set Up

- Chapter 3: Getting your Data
- Chapter 4: Visualizing your Data
- Chapter 5: Analyzing your data
- Chapter 6: Case Study - Premature Mortality
- Chapter 7: Case Study - Breast Cancer Mortality
- Chapter 8: Case Study - Cook County Covid-19
- Chapter 9: Case Study - Temporal Trends using American Community Survey (ACS) data (2012-2019)
- Chapter 10: Case Study - Comparing County Analyses of Inequities in Health Insurance using ACS vs. CDC PLACES data (2019) Survey
- Chapter 11: Conclusions

We hope that this resource is of use to you. If you have questions or comments, please reach out to: geoproj@hsph.harvard.edu

REFERENCES

Agency for Toxic Substances and Disease Registry (ATSDR). CDC/ATSDR Social Vulnerability Index. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html ; accessed June 14, 2022.

Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. Lancet. 2017 Apr 8;389(10077):1453-1463. doi: 10.1016/S0140-6736(17)30569-X.

Bambra C, Lynch J, Smith KE. The Unequal Pandemic: COVID-19 and Health Inequalities. Bristol, UK: Policy Press, University of Bristol, 2021.

Beckfield J. Political Sociology and The People's Health. New York: Oxford University Press, 2018.

Chen JT and Krieger N. Revealing the unequal burden of COVID-19 by income, race/ethnicity, and household crowding: US county versus zip code analyses. Journal of Public Health Management and Practice. 2021; 27(1), pp.S43-S56.

Frank JW and Matsunaga E. National monitoring systems for health inequalities by socioeconomic status–an OECD snapshot. Critical Public Health. 2020; pp.1-8. doi: 10.1080/09581596.2020.1862761

Green HW. Tuberculosis and economic strata, Cleveland's Five-City Area, 1928-1931. Cleveland, OH: Anti-Tuberculosis League, 1932.

Hu J, Bartels CM, Rovin RA, Lamb LE, Kind AJH, Nerenz DR. Race, Ethnicity, Neighborhood Characteristics, and In-Hospital Coronavirus Disease-2019 Mortality. Med Care. 2021 Oct 1;59(10):888-892. doi: 10.1097/MLR.0000000000001624. 1.

Hunter E, Friedman D, Parrish R (eds). Health statistics : Shaping policy and practice to improve the population's health. New York ; Oxford: Oxford University Press, 2005.

Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med. 1994 Oct;39(7):887-903. doi: 10.1016/0277-9536(94)90202-x.

Krieger, N. Socioeconomic data in cancer registries. Am J Public Health. 2001; 91(1), p.156.

Krieger, N. Chen, J.T., Waterman, P.D., Rehkopf, D.H. and Subramanian, S.V. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—the public health disparities geocoding project. Am J Public Health. 2003; 93(10), pp.1655-1671.

Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. Am J Public Health. 2005 Feb;95(2):312-23. doi: 10.2105/AJPH.2003.032482.

Krieger N. Putting health inequities on the map: social epidemiology meets medical/health geography—an ecosocial perspective. GeoJournal. 2009 Apr;74(2):87-97.

Krieger N, Alegría M, Almeida-Filho N et al. Who, and what, causes health inequities? Reflections on emerging debates from an exploratory Latin American/North American workshop. J Epidemiol Community Health. 2010;64(9):747–749.

Krieger N. Epidemiology and The People's Health: Theory and Context. New York: Oxford University Press, 2011.

Krieger N, 2012. Who and what is a "population"? Historical debates, current controversies, and implications for understanding "population health" and rectifying health inequities. The Milbank Quarterly, 90(4), pp.634-681.

Krieger N, Waterman PD, Gryparis A, Coull BA. Black carbon exposure, socioeconomic and racial/ethnic spatial polarization, and the Index of Concentration at the Extremes (ICE) Health Place. 2015;34:215–228.

Krieger N, Waterman PD, Spasojevic J, Li W, Maduro G. and Van Wye, G. Public health monitoring of privilege and deprivation with the index of concentration at the extremes. Am J Public Health. 2016; 106(2), pp.256-263.

Krieger N, Singh N, and Waterman PD. Metrics for monitoring cancer inequities: residential segregation, the Index of Concentration at the Extremes (ICE), and breast cancer estrogen receptor status (USA, 1992–2012). Cancer Causes & Control. 2016B; 27(9), pp.1139-1151.

Krieger N, Feldman JM, Waterman PD, Chen JT, Coull BA, Hemenway D. Local Residential Segregation Matters: Stronger Association of Census Tract Compared to Conventional City-Level Measures with Fatal and Non-Fatal Assaults (Total and Firearm Related), Using the Index of Concentration at the Extremes (ICE) for Racial, Economic, and Racialized Economic Segregation, Massachusetts (US), 1995-2010. J Urban Health. 2017 Apr;94(2):244-258. doi: 10.1007/s11524-016-0116-z.

Krieger N, Feldman JM, Kim R, and Waterman, PD. Cancer incidence and multilevel measures of residential economic and racial segregation for cancer registries. JNCI Cancer Spectrum. 2018; 2(1), p.pky009.

Krieger N. Inheritance and Health: What Really Matters? Am J Public Health. 2018 May;108(5):606-607. doi: 10.2105/AJPH.2018.304353.

Krieger N, Chen JT, Waterman PD. Using the methods of the Public Health Disparities Geocoding Project to monitor COVID-19 inequities and guide action for social justice. Available as of May 15, 2020 at: https://www.hsph.harvard.edu/thegeocodingproject/covid-19-resources/

Krieger N. Structural Racism, Health Inequities, and the Two-Edged Sword of Data: Structural Problems Require Structural Solutions. Front Public Health. 2021 Apr 15;9:655447. doi: 10.3389/fpubh.2021.655447.

Levine P. Eugenics: A Very Short Introduction. New York: Oxford University Press, 2017.

Massey DS. The age of extremes: concentrated affluence and poverty in the twenty-first century. Demography. 1996;33(4):395–412.

Massey DS. The prodigal paradigm returns: ecology comes back to sociology. In: Booth A, Crouter A, editors. Does It Take a Village? Community Effects on Children, Adolescents, and Families. Mahwah, NJ: Lawrence Erlbaum Associates; 2001. pp. 41–48.

Massey DS. Reflections on the dimensions of segregation. Soc Forces. 2012;91(1):39–43.

McLennan D, Noble S, Noble M, Plunkett E, Wright G, and Gutacker N. The English indices of deprivation 2019: technical report. 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019. Accessed June 5th, 2022.

Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, Elo I, Burke JG, and O'campo P. The development of a standardized neighborhood deprivation index. Journal of Urban Health. 2006; 83(6), pp.1041-1062.

Nathan WB. Health conditions in North Harlem 1923-1927. New York: National Tuberculosis Association, 1932.

O'Campo P, Burke JG, Culhane J, Elo IT, Eyster J, Holzman C, Messer LC, Kaufman JS, and Laraia BA. Neighborhood deprivation and preterm birth among non-Hispanic Black and White women in eight geographic areas in the United States. Am J Epidemiology. 2008; 167(2), pp.155-163.

Presidential COVID-19 Health Equity Task Force. Final reportand recommendations. HHS, Office of Minority Health. https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=2&lvlid=100. Updated November 10, 2021. Accessed June 14, 2022.

Rothstein R. The color of law : A forgotten history of how our government segregated America (First ed., Democracy and urban landscapes). New York ; London: Liveright Publishing Corporation, a division of W.W. Norton & Company, 2017.

Schuurman N, Bell N, Dunn JR, and Oliver L. Deprivation indices, population health and geography: an evaluation of the spatial effectiveness of indices at multiple scales. Journal of Urban Health. 2007; 84(4), pp.591-603.

Scally BJ, Krieger N. and Chen JT. Racialized economic segregation and stage at diagnosis of colorectal cancer in the United States. Cancer Causes & Control. 2018; 29(6), pp.527-537.

Shaw M, Galobardes B, Lawlor DA, Lynch J, Wheeler B, Davey Smith G. The Handbook of Inequality and Socioeconomic Position: Concepts and Measures. Bristol, UK: The Policy Press, 2007.

UK Office of National Statistics. The National Statistics Socio-economic classification (NS-SEC). https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010 ; accessed June 14, 2022.

Whitehead M. William Farr's legacy to the study of inequalities in health. Bulletin of the World Health Organization. 2000; 78(1), p.86.

« Preface

2 Author Bios »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

2 Author Bios

Nancy Krieger, PhD (she/her) (Principle Investigator for The Public Health Disparities Training Project 2.0) is Professor of Social Epidemiology and American Cancer Society Clinical Research Professor, in the Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health (HSPH), and also Director of the HSPH Interdisciplinary Concentration on Women, Gender, and Health. She is an internationally recognized social epidemiologist (PhD, Epidemiology, UC Berkeley, 1989), with a background in biochemistry, philosophy of science, and history of public health, plus 35+ years of activism involving social justice, science, and health. Dr. Krieger's work addresses: (1) conceptual frameworks to understand, analyze, and improve the people's health, including her ecosocial theory of disease distribution focused on embodiment and equity; (2) etiologic research on societal determinants of population health and health inequities, including structural racism and other types of adverse discrimination; and (3) methodologic research to improve monitoring of health inequities. She launched the initial Public Health Disparities Geocoding Project in the late 1990s to improve monitoring, analysis and action on the entangled impacts of social class and racism on population health and health inequities.

Jarvis Chen (he/him) is a social epidemiologist and Lecturer in Social and Behavioral Sciences at the Harvard T.H. Chan School of Public Health. His research focuses on methods for analyzing and understanding social inequities in health, particularly in relation to structural racism and socioeconomic deprivation. As a methodologist, Dr. Chen's interests include development of methods for geospatial and spatiotemporal analysis, disease mapping, and causal inference in social epidemiology. Dr. Chen is also Associate Director of the PhD in Population Health Sciences Program in Harvard University's Graduate School of Arts and Sciences and teaches several quantitative research methods courses at the school.

Enjoli Hall (she/her/hers) is a PhD student in the Department of Urban Studies and Planning at the Massachusetts Institute of Technology (MIT). Her work focuses on building infrastructures of collective care and action to understand and intervene in political and economic determinants of health.

Dena Javadi (she/her) is a PhD student in Population Health Sciences in the Department of Social and Behavioral Sciences at the Harvard T.H. Chan School of Public Health. Her prior work has been in Health Policy and Systems Research, with a focus on intersectoral action for health. Currently, her research explores the structural determinants of work-related health and wellbeing.

Justin Morgan (he/him) is currently pursuing a Ph.D. in Population Health Sciences in the Social and Behavioral Sciences department. His research interests center on power and the political determinants of health, with a focus on the practition of community engaged research to assess and address health equity.

Tamara Rushovich (she/her) is s a current PhD candidate in Population Health Sciences. Her research focuses on the ways that social factors and societal structures shape health. Prior to starting her PhD, Tamara worked in social services in Washington, DC and as an epidemiologist at the Chicago Department of Public Health. She has a BA in Sociology and an MPH in Epidemiology from the University of Michigan.

Sudipta Saha (he/him) is a Population Health Sciences PhD student in the Department of Social and Behavioral Sciences at Harvard University. His current research interests are at the intersection of social epidemiologic theories and infectious disease models. He is particularly interested in treating racial capitalism as a fundamental cause of health inequities to understand/illustrate how broader political-economic forces shape such inequities. He has a BSc in Microbiology from the University of Toronto, and a Master of Science in Global Health and Population at Harvard T.H. Chan School of Public Health.

Christian Testa (he/him) is a statistical analyst and programmer focused on modeling health outcomes and characterizing health inequities. His ongoing work is focused on COVID-19, epigenetic aging, survey data collection and analysis, discrimination, and area based social metrics. Christian is particularly interested in the application of flexible machine learning approaches in causal inference as well as the use of data visualization for the effective communication of scientific findings and their associated uncertainty.

« 1 Background and History of Analytic Methods

3 Getting Setup with R and RStudio »

3 Getting Setup with R and RStudio

By: Christian Testa

We recommend before moving on, readers should download and install R and RStudio and the recommended R packages and dependencies.

3.0.1 Principles for Reproducible Workflow and Programming

Two principles underlie why R and RStudio have been chosen to use throughout these course materials. First, as a matter of equity, the use of free (both free to use and freely licensed) software such as R and RStudio removes financial and administrative barriers to engaging with our work and facilitates the inclusion of people from more diverse backgrounds in science. Second, data analysis and science should be open and transparent wherever and whenever possible, so as to promote the external reproduction, verification, and validation of findings.

> The code examples in this book assume familiarity with the R programming language. You can still benefit from this book by reading the exposition without focusing on the code examples if you are not very familiar with R programming.
>
> If you want to become more familiar with R programming, you may want to start by familiarizing yourself with R. The *R for Data Science* book, available free and online here: https://r4ds.had.co.nz/, which provides a free, online, accessible introduction.

3.1 Downloading and Installing R and RStudio

In order to follow along with these resources, you will need to have R and RStudio installed and setup. You can download R from https://www.r-project.org/. You can download RStudio from https://www.rstudio.com/.

3.2 Basic Features of R

Throughout this text, the example code given will depend on various R packages which are all available free and open-source, easily installed in R with the `install.packages` function.

To work through the case examples here, you will need to install at least the following packages:

On this page

```r
# for data manipulation and visualization
install.packages("tidyverse")
install.packages("Hmisc") # mostly for the weighted quantile function
install.packages("fastDummies") # for creating "dummy"/ "indicator" variables


# for retrieving census data
install.packages("tidycensus") # note the special instructions below

# for mapping
install.packages("sf") # note the special instructions below
install.packages("mapview")
install.packages("tigris")
install.packages("leaflet")


# for visualization/color palettes
install.packages("RColorBrewer")
install.packages("viridis")
install.packages("cowplot")

# for multilevel modeling
```

3.2.1 Note about “compiling from source”

If, when installing these packages, R prompts you asking whether you would like to install these packages “from source,” you do not need to. Sometimes compiling from source can be more difficult than installing the pre-built packages from CRAN if there are compilation errors.

3.2.2 Tidycensus Special Instructions

You need to register for a Census API key as part of the setup procedures to use `tidycensus`.

See the instructions for installing on the tidycensus website.

Note that you only need to run the `census_api_key("YOUR API KEY GOES HERE")` command once.

3.2.3 sf Special Instructions

For the `sf` package, there are additional instructions here: https://r-spatial.github.io/sf/ which are OS (Mac, Windows, Linux) specific, which walk through getting setup with the `gdal` (or Geospatial Data Abstraction Library, a translator library for raster and vector geospatial data formats) which is a dependency of `sf`.

3.2.4 INLA Install Instructions

The code above installs the stable version of INLA. Should you ever need to upgrade your installation, instructions are online here: https://www.r-inla.org/download-install

3.3 References for Spatial Programming in R

As additional reference material to supplement the R programming code provided here, we would recommend:

- Geocomputation with R by Robin Lovelace, Jakub Nowosad, and Jannes Muenchow. https://geocompr.robinlovelace.net/.
- Spatial Data Science with Applications in R by Edzer Pebesma and Roger Bivand. https://r-spatial.org/book/.
- Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny by Paula Moraga. https://www.paulamoraga.com/book-geospatial/index.html.

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

4 Getting your data

By: Christian Testa, Jarvis Chen ScD, Enjoli Hall, Dena Javadi, Tamara Rushovich

4.1 High Level Overview

For the application of mapping area health outcome rates, georeferenced data are made up of three components: health outcome counts, population estimates, and the geographic boundaries of the areal units of interest. Each of these data may each come from different data sources, but they are linked together in practice by merging the data together by geographic area identifiers.

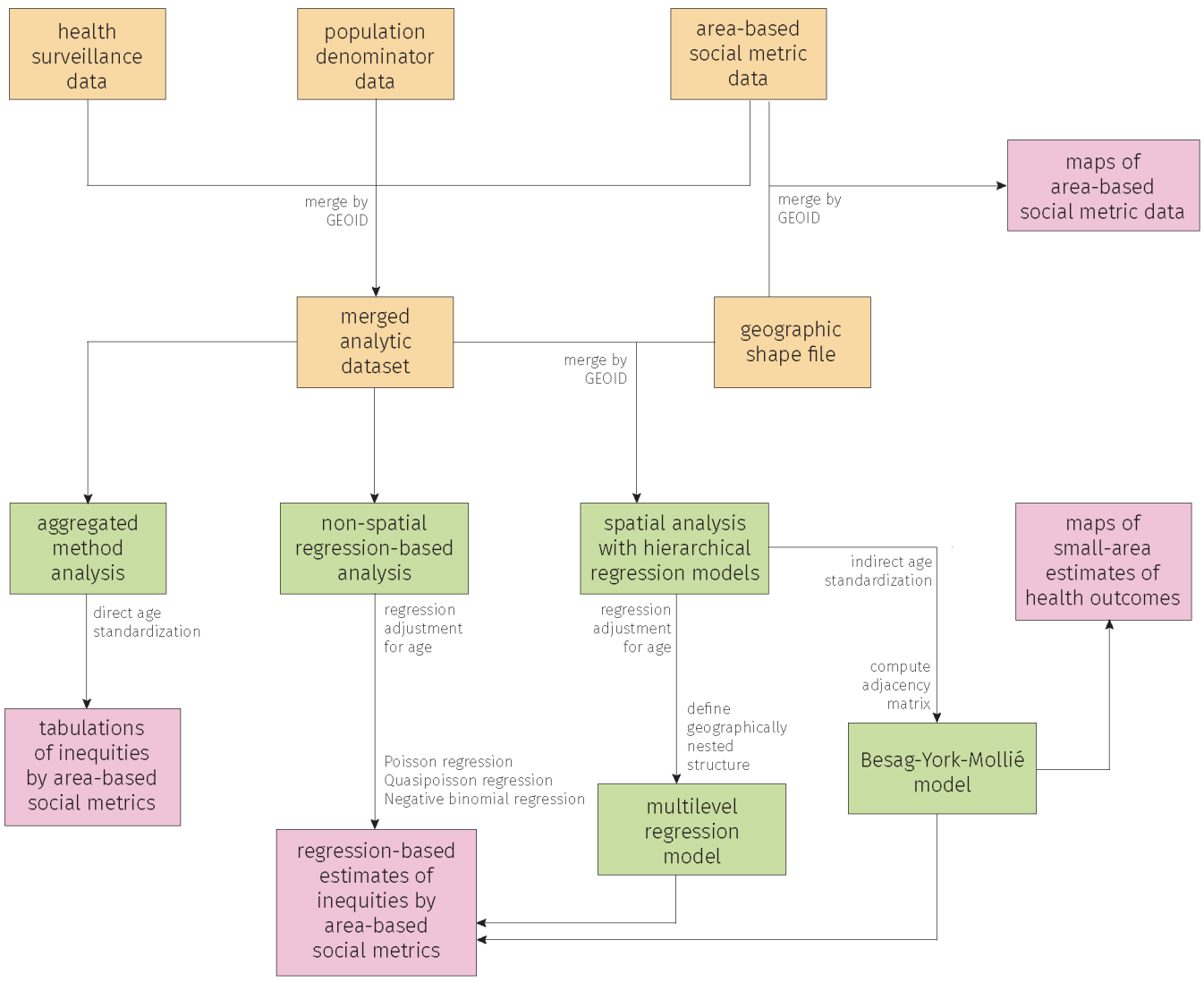


Figure 4.1: In this workflow schematic, the components for creating georeferenced data are shown in the first two rows.

Some examples of area identifiers common in the United States context include codes for ZIP codes, county FIPS codes, census block or census tract IDs, etc. As shown in Figure 4.2 (below), there are two major sets of geographies employed by the US Census Bureau. The first concerns fundamental administrative and political units (referred to as the “spine” of US Census geography) which are relevant to the primary rationale of the US census to allocate political representation (Krieger 2019); these areas include census block, block-groups, tracts, counties, and states, and corresponding units for American Indian, Alaska Native and Native Hawaiian areas, with the census block being the core area used to create voting districts (at the local, state, and federal levels). The second set includes an array of additional areas, which often cross the boundaries of these core Census geographic areas, such as ZIP Codes, school districts, traffic analysis zones, etc.

On this page

If you are studying area level health outcome rates, you may find that you must source these different necessary pieces from different places. Population estimates can often come from the Census or its related data products, health outcomes may come from healthcare providers, public health departments, electronic health records (EHR), or other sources, and geography shapefiles are similarly available from numerous sources including the Census.

4.2 Data Sources

Health data can be found in numerous formats, such as for download from a webpage as .csv or .xlsx files, accessible through querying an application programming interface (API), or available only through direct request to a health department or agency.

The sources for the data we've used in our case examples are as follows:

- US Census and American Community Survey data are retrieved through their API https://www.census.gov/data/developers/data-sets.html via the `tidycensus` package
- CDC Places data were downloaded directly as .csv files from https://www.cdc.gov/places/
- Cook County Medical Case Examiner Archive data on COVID-19 deaths were downloaded directly from https://datacatalog.cookcountyil.gov/Public-Safety/Medical-Examiner-Case-Archive-COVID-19-Related-Dea/3trz-enys
- Massachusetts mortality data were requested directly from the Massachusetts Department of Public Health
- The Social Vulnerability Index (SVI) is available online to download with csv and shapefiles available from https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

4.3 Loading Spreadsheet Data into R

If you have a .csv file that you have downloaded, you can use the `readr` package (which is part of the `tidyverse`) to load it into R.

```
library(readr)
example_df <- read_csv("filename.csv")

# learn more about the options read_csv has by running:
?read_csv
```

Health data often come in other kinds of delimited formats such as tab-delimited or fixed-width spaced, in which case you can use the `read_tsv` or `read_fwf` functions from the `readr` package similarly to how you would use `read_csv`. Learn more about `readr` here: https://readr.tidyverse.org/

To read Excel data, we recommend using the `readxl` package, also part of the `tidyverse`. Learn more here: https://readxl.tidyverse.org/

```
library(readxl)
example_df <- read_excel("filename.xlsx", sheet = 1)
# to learn more about the options in the read_excel function, run:
?read_excel

# one particularly helpful feature to know about is the range argument which
# allows the user to specify they want to read a dataframe from a specific
# range of cells using Excel-style range syntax:
example_df <- read_excel("filename.xlsx", sheet = 1, range = "A3:C17")
```

Even if your data are not a delimited text document or an Excel file, if they are in a common file format, it is still quite likely you can read your data into R using other packages. For example the `haven` package allows users to read SAS, SPSS, and Stata files. Read more here https://haven.tidyverse.org/

4.4 Connecting to Databases

Many online health datasets are accessible via query to a remote database server. References on how to interact with databases in R are available here: https://db.rstudio.com/, and the following reference shows how to interact with a remote database in the `tidyverse` style: https://dbplyr.tidyverse.org/

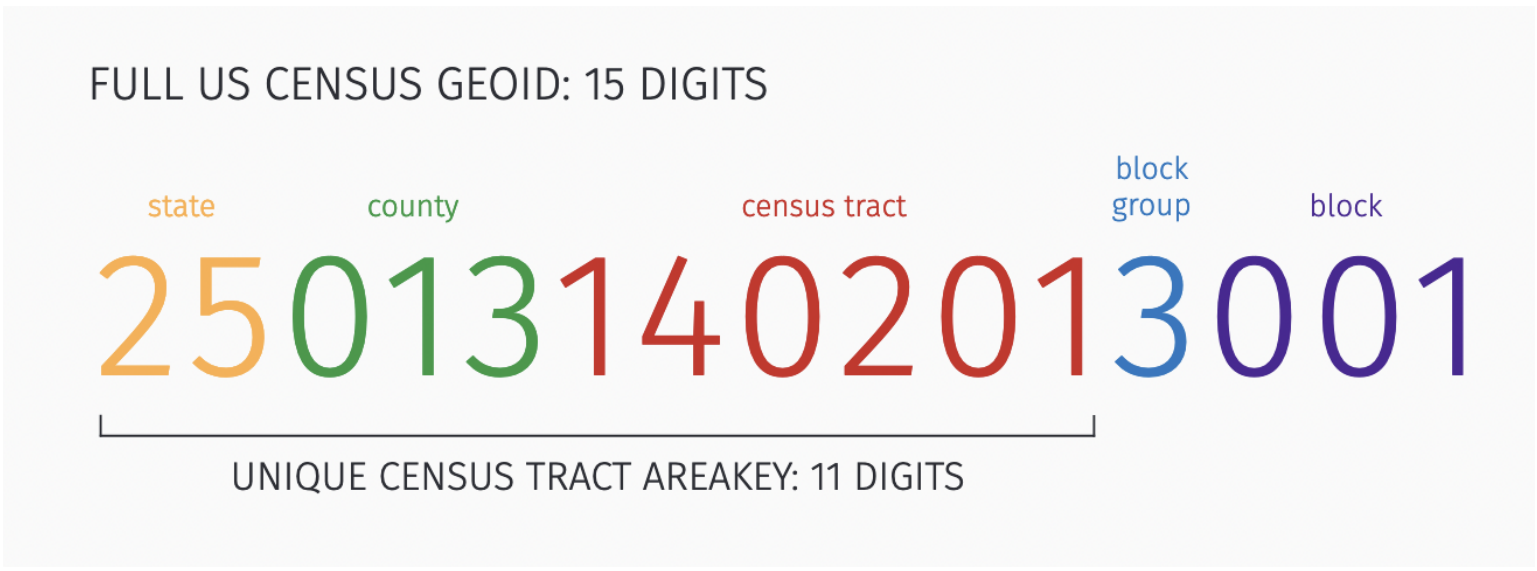
4.5 tidycensus

For some databases, R programmers have already written packages to help users submit their queries and get back their data of interest. One such example is the U.S. Census, for which the `tidycensus` package exists to automate fetching Census data in R.

The `tidycensus` package in R allows you to download data from the US Census Bureau products, including from the decennial Census and the 1-year, 3-year, and 5-year American Community Survey (ACS). Find detailed reference materials and an introduction to `tidycensus` here: https://walker-data.com/tidycensus/

In this section, we will walk you through example code that downloads the percent of residents under the poverty line and computing the Index of Concentration at the Extremes (ICE) for racialized economic segregation from the 2015-2019 ACS. As an example, we will demonstrate how to download these measures at the census tract level in Suffolk County, Massachusetts noting that this county includes the city of Boston, where the PHDGP 2.0 training team members are based!

You can find more of the variables available in the 5-year ACS at the following link, changing 2019 to your desired year starting with 2009 when the ACS began: https://api.census.gov/data/2019/acs/acs5/variables.html



(#fig:geoid_structure)The structure of Census tract GEOIDs

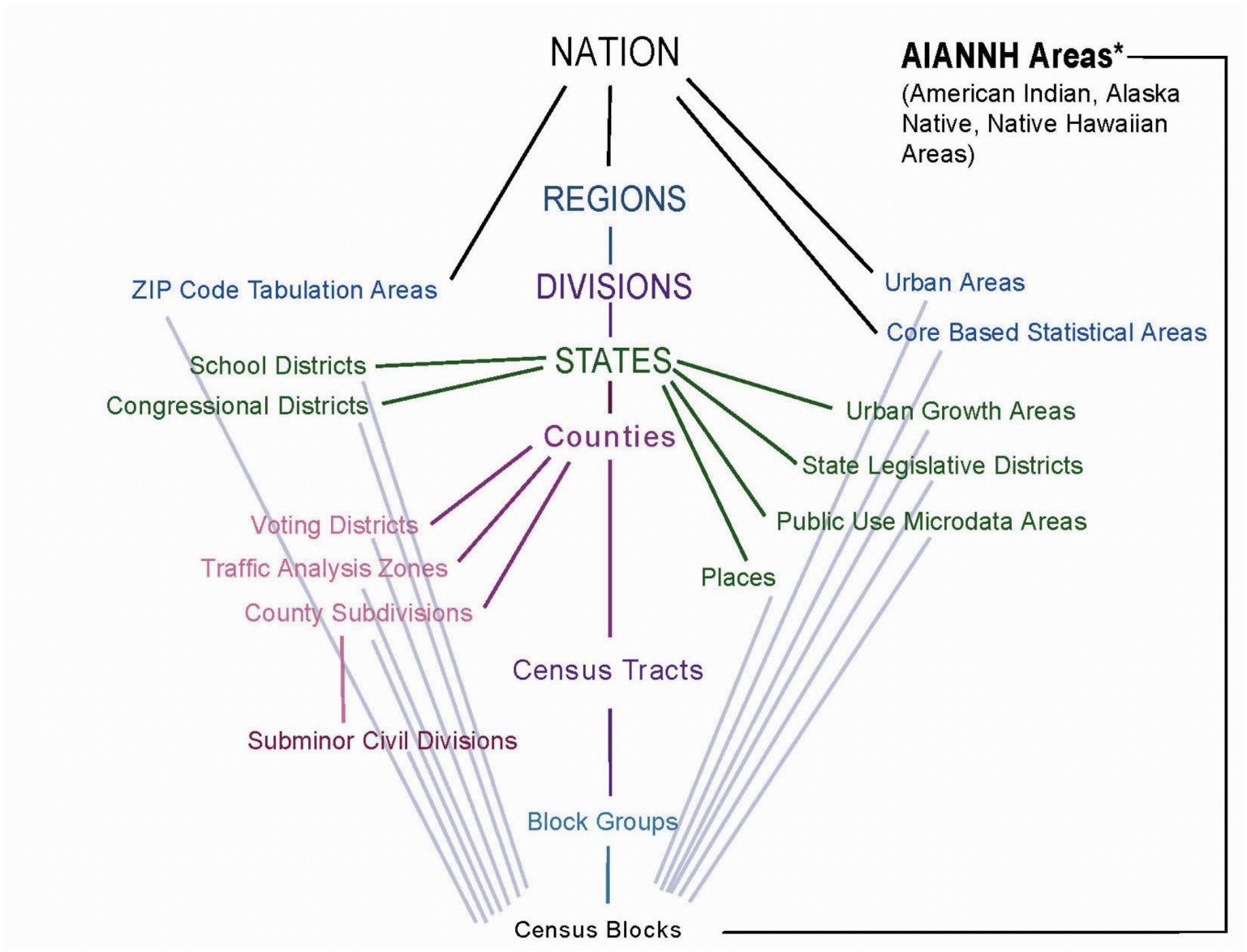
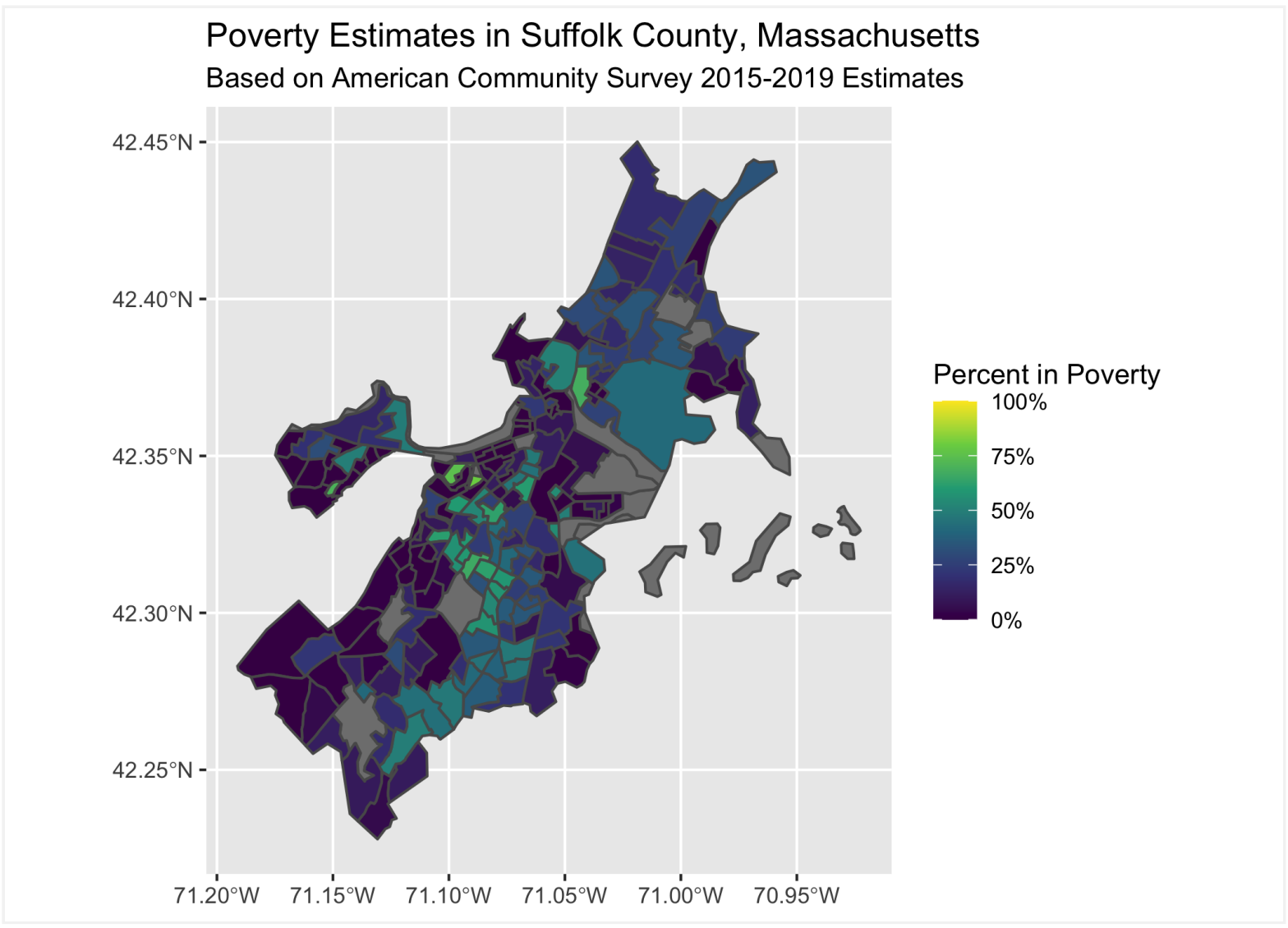
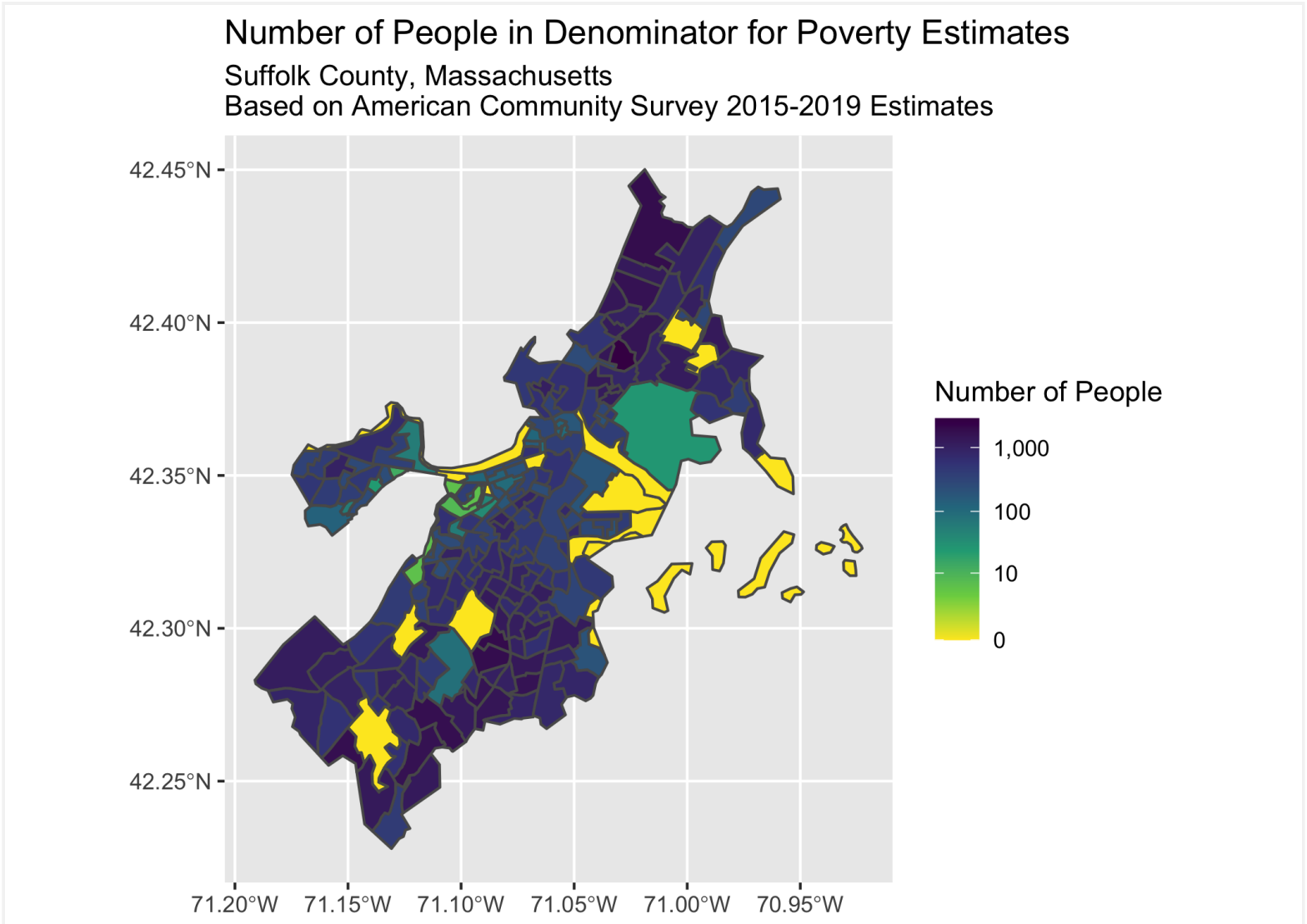


Figure 4.2: Hierarchy of geographic units as assignged by the US Census Bureau. Reproduced from https://www.census.gov/content/dam/Census/data/developers/geoareaconcepts.pdf

Here's example code for downloading data on the percent of people living in households with household income less than the poverty threshold. The poverty thresholds are defined by the US Census (see: https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html).

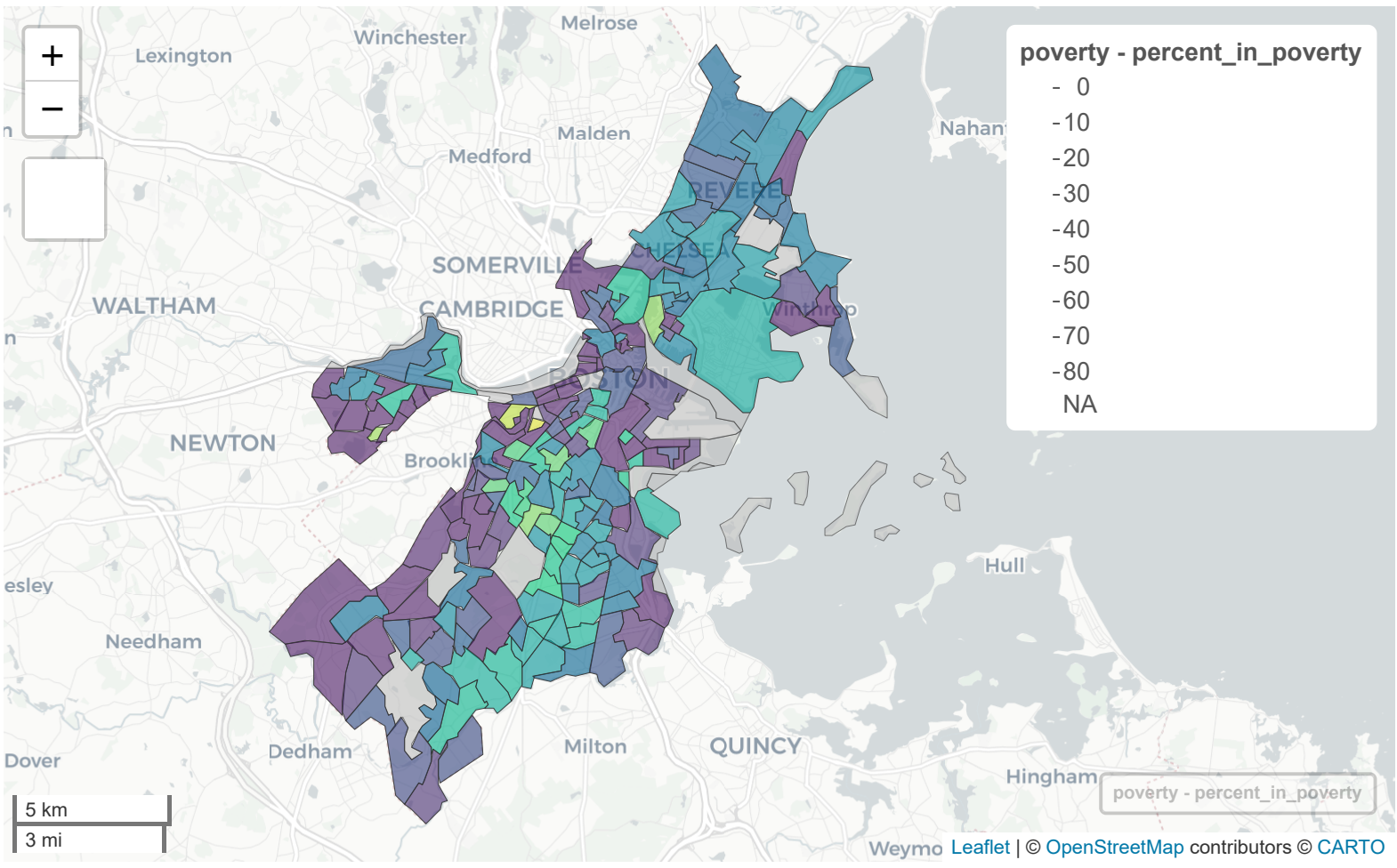
```r
# example code for downloading poverty measures from the American Community
# Survey through tidycensus and visualizing them through maps

# load the packages we'll use for this section
library(tidycensus)
library(tidyverse)
library(sf)
library(RColorBrewer)
library(mapview)

# download the data from the ACS using the get_acs method from tidycensus
#
# the B05010_002E variable refers to the count of residents who live in
# households with household income below the poverty line; the B05010_001E
# variable refers to the count of residents for whom household income was
# ascertained by the ACS, e.g. the relevant denominator.
#
poverty <- get_acs(
  state = 'MA',
county = '025', # this is the FIPS code for Suffolk County, MA
geography = 'tract',
year = 2019, # this indicates the 2015-2019 5-year acs
```





We can also use the `mapview` package to render interactive maps of the estimates and data we have computed/downloaded. This is particularly helpful in understanding how the geography relate to the estimates — especially to see where airports, green-space, schools, and other municipal zones may be located and which are not zoned to include residential units (hence the NaN or "not a number" estimates shown on the maps), but which may nevertheless be locations where unsheltered and other unhoused persons reside.

```
library(mapview)
mapview::mapview(poverty, zcol = 'percent_in_poverty')
```



Estimates of the percent of individuals in poverty based on the American Community Survey 2015-2019

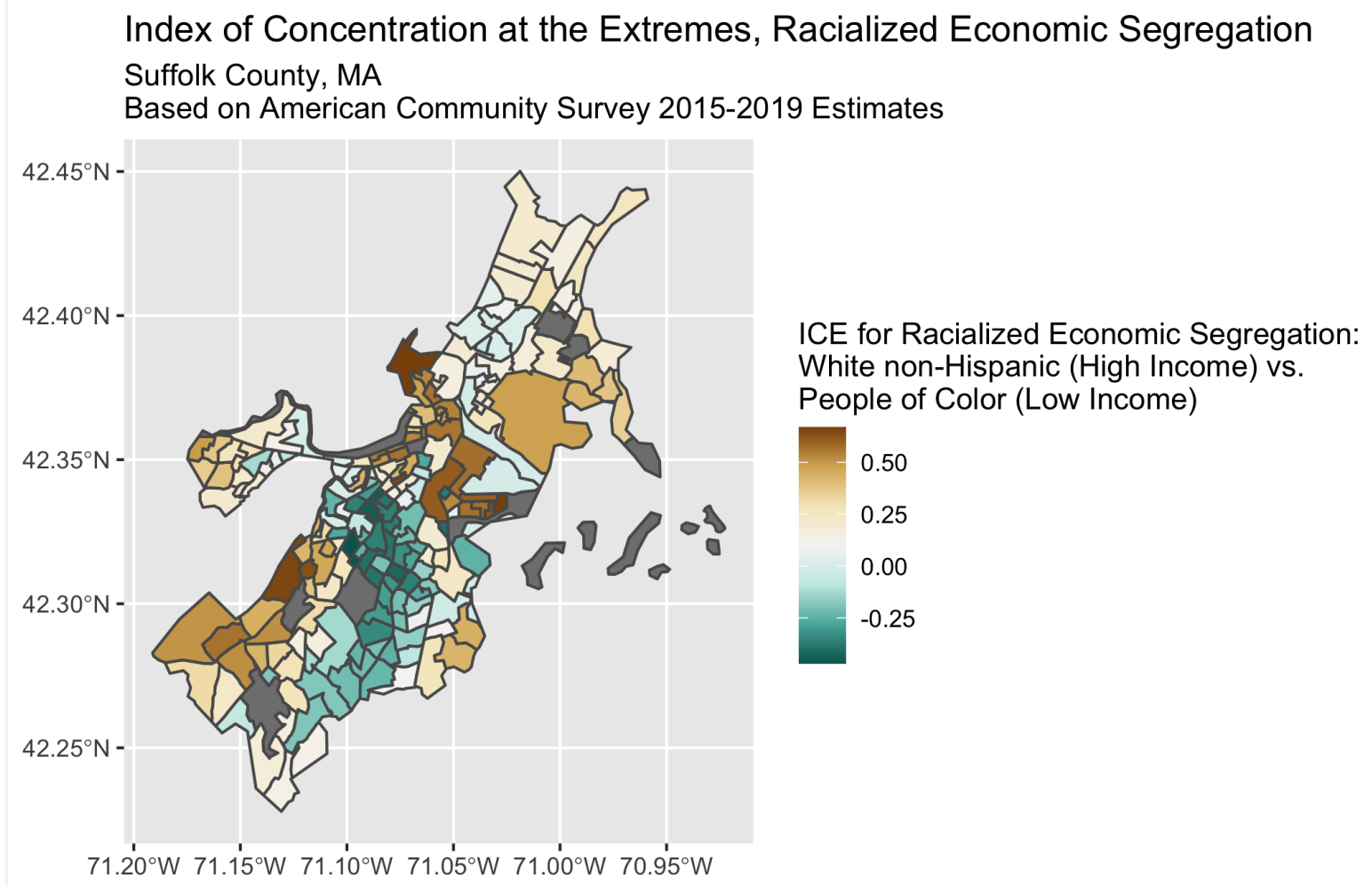
For reference, these are the poverty thresholds in US dollars for 2017 (i.e. the middle year of the 2015-2019 time-period) from the Census:

Size of family unit	Weighted average thresholds	Related children under 18 year: None	One	Two	Three	Four	Fiv
One person (unrelated individual):	12,488						
Under age 65.....	12,752	12,752					
Aged 65 and older.....	11,756	11,756					
Two people:	15,877						
Householder under age 65.....	16,493	16,414	16,895				
Householder aged 65 and older.....	14,828	14,816	16,831				
Three people.....	19,515	19,173	19,730	19,749			
Four people.....	25,094	25,283	25,696	24,858	24,944		
Five people.....	29,714	30,490	30,933	29,986	29,253	28,805	
Six people.....	33,618	35,069	35,208	34,482	33,787	32,753	32
Seven people.....	38,173	40,351	40,603	39,734	39,129	38,001	36
Eight people.....	42,684	45,129	45,528	44,708	43,990	42,971	41
Nine people or more.....	50,681	54,287	54,550	53,825	53,216	52,216	50

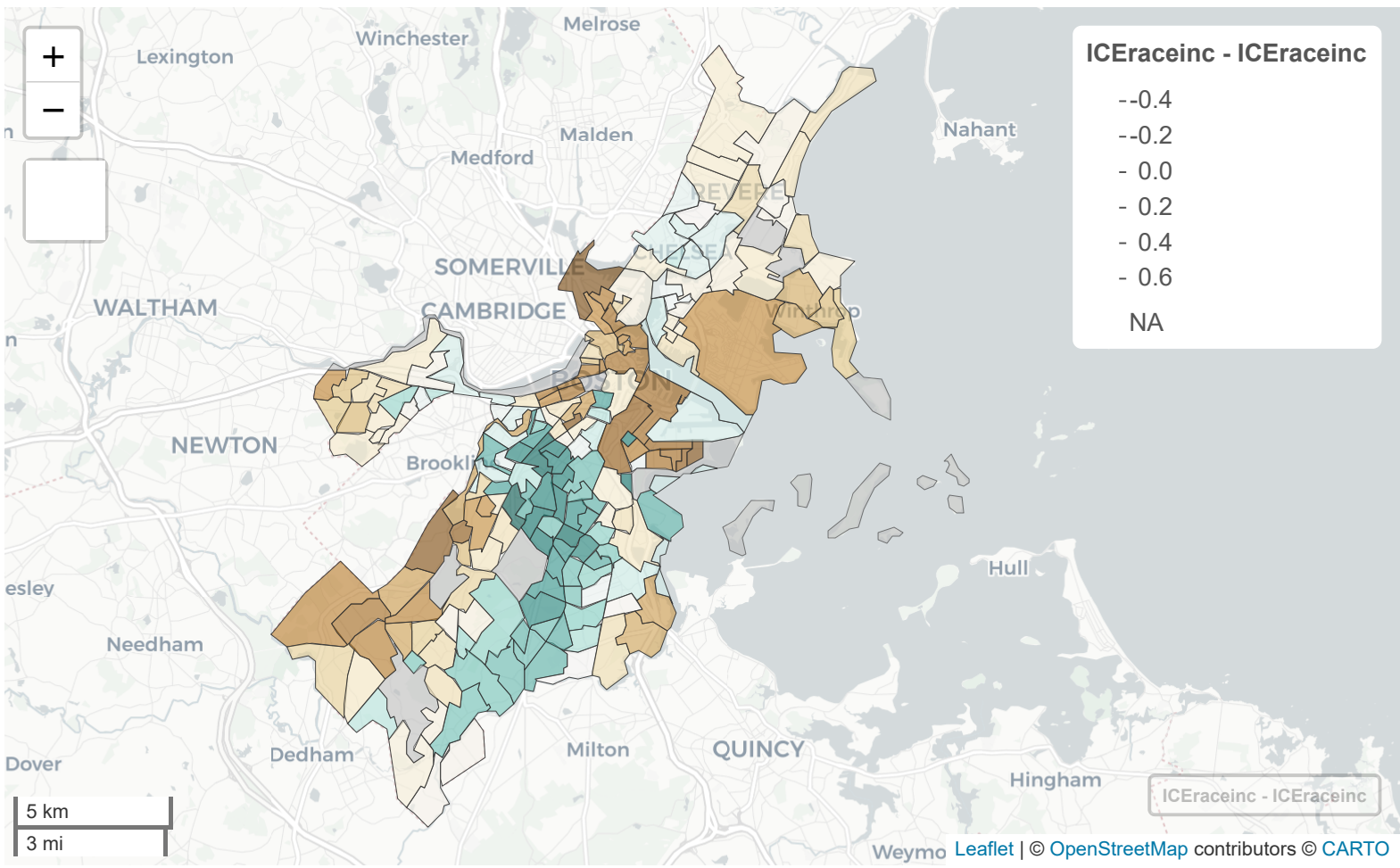
Source: U.S. Census Bureau.

Now that we've created static and interactive maps of the poverty estimates for Suffolk County, we can move on to downloading and computing the Index of Concentration at the Extremes for Racialized Economic Segregation (High Income White non-Hispanic High Income vs. Low Income People of Color).

```
# example code for creating the index of concentration at the extremes for the
# measure of racialized economic segregation (high income white non-hispanic
# vs. low income people of color) using tidycensus

# create a data dictionary detailing the variables we're going to use -
#
# associating each of the variables a more readable/friendly `shortname` and a
# description can help make the subsequent code more readable and thus easier
# to debug in case you run into any errors.
#
variables_dict <-
  tibble::tribble(
  ~var,          ~shortname,      ~desc,
  "B19001_001",  'hhinc_total',   "total population for household income estimates",
  "B19001A_002", 'hhinc_w_1',     "white n.h. pop with household income <$10k",
  "B19001A_003", 'hhinc_w_2',     "white n.h. pop with household income $10k-14 999k",
  "B19001A_004", 'hhinc_w_3',     "white n.h. pop with household income $15k-19 999k",
  "B19001A_005", 'hhinc_w_4',     "white n.h. pop with household income $20k-24 999k",
  "B19001A_014", 'hhinc_w_5',     "white n.h. pop with household income $100 000 to $124 999",
  "B19001A_015", 'hhinc_w_6',     "white n.h. pop with household income $125k-149 999k",
  "B19001A_016", 'hhinc_w_7',     "white n.h. pop with household income $150k-199 999k",
  "B19001A_017", 'hhinc_w_8',     "white n.h. pop with household income $196k+",
```




```
mapview(ICEraceinc, zcol = 'ICEraceinc',
        col.regions=rev(brewer.pal(11, "BrBG")))
```



Estimates of the Index of Concentration at the Extremes for Racialized Economic Segregation (white non-Hispanic with high income vs. People of Color with low income) based on the American Community Survey 2015-2019

4.6 **Cleaning Area Identifiers**

One of the most important aspects to pay attention to when cleaning georeferenced data is that area identifiers are stored as character or factor data types and not numeric types because if they are stored as numeric types leading 0s will be dropped and this may cause issues when merging multiple datasets together

if the area-keys are not coded in the same way (e.g. in one dataset area-keys might be coded numeric and in another dataset area-keys might be coded as a character or factor).

This can happen often with FIPS codes, like the 2-character state FIPS codes and 3-digit county FIPS code. For states like Alabama and Alaska with FIPS codes `01` and `02`, if these are mistakenly stored as numeric values, they will be truncated to as `1` and `2` which can introduce errors when trying to merge multiple datasets using FIPS codes.

> For county FIPS codes, the `tigris` package has a handy built-in reference.
>
> Once you've installed `tigris` (e.g. run `install.packages('tigris')`) and load the package (`library(tigris)`) you can access the built-in `fips_codes` data.frame.

state	state_code	state_name	county_code	county
AL	01	Alabama	001	Autauga County
AL	01	Alabama	003	Baldwin County
AL	01	Alabama	005	Barbour County
AL	01	Alabama	007	Bibb County
AL	01	Alabama	009	Blount County
AL	01	Alabama	011	Bullock County
AL	01	Alabama	013	Butler County
AL	01	Alabama	015	Calhoun County
AL	01	Alabama	017	Chambers County

Suppose, as can happen, that you download your data and find that due to a coding error the 5-digit combined state and county FIPS codes have been stored as numeric, causing leading zeroes to be truncated off. If you check the FIPS codes and are reasonably confident that the only error is that leading left-hand-side zeroes have been omitted, you could do the following to correct the mistake:

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(example_df)
```

```
## # A tibble: 6 × 3
##    FIPS Name    State
##   <int> <chr>   <chr>
## 1  1001 Autauga AL
## 2  1003 Baldwin AL
## 3  1005 Barbour AL
## 4  1007 Bibb    AL
## 5  1009 Blount  AL
## 6  1011 Bullock AL
```

```
library(stringr)
library(dplyr)
example_df <- example_df %>% mutate(FIPS = ifelse(
  nchar(FIPS) == 4,
  str_pad(
    FIPS,
    width = 5,
    pad = '0',
    side = 'left'
  ),
  FIPS
))

head(example_df)
```

```
## # A tibble: 6 × 3
##   FIPS  Name    State
##   <chr> <chr>   <chr>
## 1 01001 Autauga AL
## 2 01003 Baldwin AL
## 3 01005 Barbour AL
## 4 01007 Bibb    AL
## 5 01009 Blount  AL
## 6 01011 Bullock AL
```

The above code uses the `dplyr` and `stringr` packages, both part of the `tidyverse` and which have introductions here: https://dplyr.tidyverse.org/ and here: https://stringr.tidyverse.org/.

4.7 Numerators and Denominators

4.7.1 Epi Primer

When presenting data on disease distribution and measures of health disparities, the following types of measures are most often used:

- Prevalence
- Cumulative incidence
- Incidence rate

These measures serve to answer different types of research questions:

- Descriptive: describes the distribution of the outcome of interest
- Predictive: predicts who may experience the outcome of interest
- Causal: seeks to determine modifiable causes of the outcome of interest

Prevalence is defined as $Pr[Y = 1] = \frac{\#\text{ existing cases}}{\#\text{ individuals in study population at a point in time}}$. Prevalence is dimensionless, ranges from 0 to 1, and requires a time reference (i.e. prevalence at a certain year, age, etc.). Prevalence measures are useful for descriptive and not causal questions.

Cumulative incidence, also known as incidence proportion, risk, or attack rate (although not a rate), is defined as $Pr[Y = 1] = \frac{\#\text{ incident cases in a time period}}{\#\text{individuals at risk at a baseline}}$. Cumulative incidence is also dimensionless and ranges from 0 to 1, but it must state a time period and anyone included in the denominator must be eligible to move into the numerator by meeting the case definition. Cumulative incidence is limited by a lack of information on the exact timing of the outcome of interest, time-varying exposures, competing risks, and loss to follow-up.

Incidence rate, also known as incidence density or hazard rate, is defined as $\frac{\#\text{ incident cases during } t_0 \text{ to } t_1}{\sum \text{person-time at risk accumulated during } t_0 \text{ to } t_1}$. Person-time can be expressed in years, months, weeks, or days. Incidence rate is not a proportion and ranges from 0 to ∞.

Measures of disparity

- Ratio measures
- Difference measures
- Attributable fractions

Ratio measures include cumulative incidence ratios (CIR), incidence rate ratios (IRR), and odds ratios (OR). The range for these is 0 to ∞ and no association is indicated by a ratio of 1.

Odds ratios are defined by

$$\frac{(Pr[Y = 1|A = 1]/Pr[Y = 0|A = 1])}{(Pr[Y = 1|A = 0]/Pr[Y = 0|A = 0])}$$

Differences measures include cumulative incidence difference (CID), also known as attributable risk, and incidence rate difference (IRD), also known as attributable rate. The range for CID is -1 to 1 while the range for IRD is $-\infty$ to ∞. No association is indicated by a difference of 0.

Attributable Risk Percent (AR) or excess fraction is the proportion of disease burden among the exposed that is associated with the exposure and is defined by:

$$AR\% = \frac{CI_{Exposed} - CI_{unexposed}}{CI_{Exposed}} * 100$$

Population Attributable Risk Percent (PAR) is the proportion of disease burden among the total population that is associated with the exposure and is defined by:

$$AR\% * Pr[A = 1|Y = 1] = \frac{CI_{Total} - CI_{unexposed}}{CI_{Total}}$$

The analytic methods section of the *Public Health Disparities Geocoding Project* monograph demonstrates how the measures described above are identified and aggregated over areas and strata of area-based socioeconomic or other social metrics (ABSMs).

4.7.2 Numerator/Denominator Issues

When calculating population rates of health outcomes, it is important to consider and avoid bias resulting from a mismatch between the numerator and the denominator of the rate. Typically, this can occur when the data sources for the numerator and the denominator differ - e.g. to calculate the rate of deaths resulting from a drug overdose in a particular county, data for the numerator of the rate could come from death certificates and the data for the denominator from the national census. Using different data sources does not necessarily cause bias. It is when there is a mismatch between the two data sources that bias can result. For example, in the previous scenario, if the numerator data included all drug overdoses that occurred in a particular county, but the denominator only included residents of the county, bias could result as the numerator includes individuals that are not in the denominator.

For epidemiologic analyses of health disparities, data analysts generally rely on routinely collected data from health surveillance systems to define the numerators of rates. Such data are available on an ongoing basis with a lag of two or more years to allow for data compilation and cleaning. In contrast, data on population at risk

may be obtained from a variety of sources depending on the need for stratification by demographic variables, time frame, and the desired level of geography. For example, the US Decennial Census provides detailed population estimates by age, sex-gender, and selected racialized groups in decennial years. For intercensal years, the US Census uses demographic modeling that takes into consideration births, deaths, and migrations to provide estimates for these demographic groups at the national, state, and county levels via its Population Estimates Program (PEP) [US Census PEP]. For smaller levels of geography, the US Census Bureau's American Community Survey includes estimates of population for all census geographies larger than census blocks based on five year rolling averages. These are also typically made available after a two year delay.

Formally, the US Census Bureau recommends the use of PEP or decennial census counts as population size estimates in intercensal years, while recommending that ACS data be used for information about changing socioeconomic and demographic features. However, for estimation of small area disease rates or analyses of health disparities by ABSM in intercensal years, decennial counts are usually outdated and PEP estimates are not available at the desired geographic level. Thus, in practice, many epidemiologic studies will continue to rely on ACS small-area estimates for population denominators. While private companies are increasingly producing high-resolution gridded population estimates (typically for the total population, not stratified by sub-groups, whether by age or diverse social groups) that are based on machine-learning models combining census, remote sensing, land use, and other information, with the promise of providing population sizes at very small geographies in near real time, to date we have not found these data products to provide substantial improvement over using ACS small area estimates and in some cases they can induce bias (Nethery et al. 2021).

Because numerators (case data) and denominators (population estimates) come from different sources, and moreover, because intercensal population estimates in particular may be subject to uncertainty, it is possible to obtain sociodemographic strata in particular areas and years where the observed cases exceed the reported population at risk. This is particularly problematic when the number of cases is greater than zero but the reported population at risk is exactly zero. We term this phenomenon "numerator/denominator mismatch."

To ensure that there is no numerator/denominator mismatch and that all cases in the aggregated numerator are coming from the population pool in the aggregated denominator and that all those in the denominator have the potential to become cases, one must understand the sources of data being used. It is therefore critical to explore the data dictionaries and survey methodology for the various sources of data to better understand the eligibility criteria for who gets included in the denominator and who appears in the numerators to be used in aggregating data for analysis. This includes answering questions like:

- How are cases defined?
- Does the data source include just adults or all ages?
- Is the data source limited to non-institutionalized populations?
- Does the data source include solely residents of the jurisdiction?
- What happens to individuals who do not have an address of residence?
- What happens to individuals who are institutionalized, including incarcerated?

Additionally, one should exclude all records that are not geocoded, do not meet the case definition, or are missing data on important covariates (e.g. age, gender, race/ethnicity, etc. depending on the question being asked). However, any such exclusions need to be clearly documented, with careful thought given to how the resulting selection bias can affect interpretation of the results.

Numerator/denominator mismatch can also introduce bias when aggregating data over strata of ABSMs. Understanding the sources of bias arising from incompatibility and identifying population size estimates to limit bias is a critical part of designing the analytical approach. For example, temporally incompatible numerator and denominator data tends to result in greater bias for race-stratified models involving numerically smaller populations, with important implications for studying disparities (Nethery et al, 2021).

4.8 **Geocoding**

Address geocoding, or just geocoding, is the process of converting a location description (usually an address) into some form of geographic representation (usually latitude and longitude coordinates). A geocoding software or program will parse input data into standard, recognizable values, compare this information to an internal reference database of points, lines, or polygons, and then return the best match of values in the database to the input data. We use geocoding software regularly in everyday life – anytime you look up directions to a friends house, search for nearby restaurants for dinner, or hail a ride-sharing service, a geocoding tool is operating under the hood to provide the information you need.

When you need to geocode many addresses, which we often need to do when working with large datasets, this is called 'batch' geocoding. Many services offer batch geocoding, with varying degrees of speed, accuracy, and pricing. We recommend reviewing multiple services before deciding what is right for you and your team. Below we provide a table of some geocoding services you might be interested in.

There are a variety of geocoding services available on the internet ranging from free to very expensive. The table below contains a list of some of the available geocoding services.

There are several important questions to consider when choosing a geocoding services, including:

- Is it free? If not, what kind of subscription packages do they offer?
- Is caching or locally/permanently storing geocoding results permitted?
- Is using geocoding results with third-party basemaps permitted?
- Does service have the right to store your geocoding requests and results and transmit these data to third parties?
- What is their privacy and data protection policy?

Geocoder
ArcGIS World Geocoding Service
ArcGIS StreetMap Premium
Bing Maps
Census Geocoder
Geoapify
Geocode Earth
Geocodio
Google Maps
HERE
LocationIQ
Mapbox
MapQuest
OpenCage
TomTom

In order to access a geocoding service, you will have to interact with its application program interface (API). Some services, like ArcGIS, have interfaces that are internal to the software you might download for your computer. Others like Google, or Nominatum, require some programming to access. We utilize R packages to access APIs from within the R environment. We will walk you through this process using Google's Geocoding API as an example.

4.8.1 **Getting your Data in Shape**

In order for any batch geocoding process to run smoothly, we need to first get the data we want to be geocoded into the appropriate format to be processed. While each service may have different preferences for how the address data are formatted, they generally require all the address information combined into a single, string (text) variable. Some prefer commas separating different features of the string (Ellicott City, Maryland, 21042). If you are using R to clean your data, there are multiple functions in the `stringr` package that will prove useful.

When preparing your data, watch out for naming idiosyncrasies in your dataset that may confuse the geocoder and lead to mismatches or failed attempts. In the Massachusetts mortality data, for example, addresses often used shorthand for the street designation (road = RD, path = PA, CT = court). To a human reader, that is rarely a problem, but Google's Geocoding API often misidentified addresses ending in PA as being in Pennsylvania, and similarly identified CT as Connecticut. Several "drives" (DR) it struggled to identify it marked as doctor's offices (!). By reviewing and editing your data before running your geocoding program, you can avoid having to run it twice.

4.8.2 **Setting up the Service**

Depending on what geocoding service you use, you may have to take several steps gain access to the service. Google, not unlike other services, requires you to register for an API key through the Google Cloud Console. These API keys allow services to track who is using them (don't share yours!), and charge for services. Once you have the appropriate API key, you can save it to your R environment. This is recommended, as opposed to calling it in your code, as you may wish to share code without sharing your unique key.

4.8.3 **Geocoding**

Once your data are set up correctly, and you are credentialed to utilize the geocoding service, it's time to geocode your data. Here is an example of code from the package 'ggmap' which supports the Google geocoding API. Depending on the quality of your internet service, whether or not your service allows you to cache results, and how many items you have to geocode, you may want to split up the geocoding into smaller batches so that if it is interrupted and you have to start over, you aren't starting over from scratch.

Where possible, request as much output from your program as possible. In addition to latitude and longitude data, you can often receive information on the confidence of the program in the match, the precision of the match, additional geographic data, or other useful bits of information. And, importantly, when using services that cost money per request, be sure to save your results. If you find mistakes later, you can mark them, separate them from the main file, fix them, and rerun the geocoding service on the smaller sample.

4.8.4 **Checking your Results**

Once you've geocoded your results, it is important to check for accuracy. Some programs can tell you about the confidence of the match, and you can choose to review matches below a certain threshold to verify their accuracy. You can broadly verify the results by mapping the points to ensure things look appropriate. At a broad scale, like the premature mortality data, it may be hard to note small unexpected patterns in a sea of dots, but obvious issues (such as clusters of points being in a different state) can be identified using this method. You can also use some of the helpful output from the geocoding program. For example, Google offers a variable called "type" which tells you what kind of building the point represents. If you are mapping residences, it may be good to explore when some addresses come back as "electronics_store" or "dentist".

As you go through this process, be mindful of your source data. Addresses for our Massachusetts mortality data, for example, are recorded by human beings. If your geocoding service struggles to find a match, it may help to verify that the street name is spelled correctly or has the proper designation. One check we used, for example, was to verify that the address Google returned had the same zip code as the address submitted. Large clusters of data where the zip code was wrong may reveal a common misspelling.

The key to this process is to not nitpick every small mistake the geocoding service makes, but to identify broad issues that might seriously bias your results.

4.9 **Data Governance**

There is only so much data users can do to account for poor data or the absence of data. Therefore, improvements in the availability and quality of health data and, in particular, improvements to the use of social metrics in conceptualizing and analyzing this data, are necessary. **Data governance** is therefore a critical feature of addressing health disparities. Data governance is about "who has input into making the decisions about which data are required, informed by the tandem expertise of health equity researchers and other members of the communities whose data are at stake, affording the expertise of lived experience" (Krieger, 2021). See *Structural Racism, Health Inequities, and the Two-Edged Sword of Data: Structural Problems Require Structural Solutions* for a proposed "two-part institutional mandate regarding the reporting and analysis of publicly-funded work involving racialized groups and health data and documentation as to why the proposed mandates are feasible" (Krieger, 2021).

Part of demanding improved data governance is interrogating the sources of data currently available, including their sampling strategies, underlying theories that inform their design, and who is involved in their generation - in line with ecosocial theory's construct of accountability and agency. By exploring the origins of one's data (both the actual records and the history of the data systems and categories at issue) and the processes through which they have been generated, data analysts and proponents of health justice can better articulate the potential for bias in the data, identify means of improving data collection to mitigate this bias, and actively call for structural change in data collection and governance.

Another key component of equity-oriented data governance is to make data openly accessible, with relevant safeguards for data privacy and use, so as to enable communities to ask their own data-driven questions, and not just rely on what analyses get published by governmental agencies or academic researchers. This also requires more robust infrastructure for community science, transparent review of methods and findings, and strengthened capacity for community-oriented knowledge translation to enhance agency. Although it is beyond the scope of our technical PHDGP 2.0 training to provide guidance on the specifics of equity-oriented data governance for the particular projects in which diverse groups are engaged, issues of data governance, sovereignty, and privacy are increasingly the focus of critical analysis and practice (Else 2022, Carroll, Rodriguez-Lonebear and Martinez 2019, Committee on National Statistics 2019, US Census Bureau 2020, Krieger et al. 2020).

4.10 **References**

Krieger N. The US Census and the People's Health: Public Health Engagement From Enslavement and "Indians Not Taxed" to Census Tracts and Health Equity (1790-2018). Am J Public Health. 2019 Aug;109(8):1092-1100. doi: 10.2105/AJPH.2019.305017. Epub 2019 Jun 20.

Nethery, Rachel C., Tamara Rushovich, Emily Peterson, Jarvis T. Chen, Pamela D. Waterman, Nancy Krieger, Lance Waller, and Brent A. Coull. "Comparing denominator sources for real-time disease incidence modeling: American Community Survey and WorldPop." SSM-Population Health 14 (2021): 100786.

Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. Modern Epidemiology. Vol. 3. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.

Else H. African researchers lead campaign for equity in global collaborations. Nature. 2022 Jun 10. doi: 10.1038/d41586-022-01604-3. Epub ahead of print.

Carroll SR, Rodriguez-Lonebear D, Martinez A. Indigenous Data Governance: Strategies from United States Native Nations. Data Sci J. 2019;18:31. doi: 10.5334/dsj-2019-031.

Committee on National Statistics. The National Academies of Sciences, Engineering, and Medicine. Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations. Washington, DC: December 11–12, 2019. Available at: https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518. Accessed June 14, 2022.

US Census Bureau. Understanding differential privacy. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html ; accessed June 14, 2022.

Krieger N, Nethery RC, Chen JT, Waterman PD, Wright E, Rushovich T, Coull BA. Impact of Differential Privacy and Census Tract Data Source (Decennial Census Versus American Community Survey) for Monitoring Health Inequities. Am J Public Health. 2021 Feb;111(2):265-268. doi: 10.2105/AJPH.2020.305989. Epub 2020 Dec 22.

« 3 Getting Setup with R and RStudio

5 Visualizing your data »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

5 Visualizing your data

By: Christian Testa, Enjoli Hall

Data visualization is a critical component in communicating and advocating for health equity as it makes data accessible and transparent. However, it is not without its pitfalls, and in this chapter we will discuss some of the important points to consider when visualizing data for advancing health equity.

Firstly, as a matter of accessibility, we strive to use colorblind friendly color palettes when using color so that individuals with one of the different kinds of colorblindness can still interpret our visualizations. We encourage you to learn more about colorblindness and colorblind friendly palettes from a number of resources (Katsnelson, 2021) (Ou, 2021). The color vision deficiency simulator from the `colorblindr` package is especially helpful in testing if a visual you are creating is colorblind friendly.

> If you are looking for help learning how to create data visualizations in R, we recommend checking out the online, free book: ggplot2: Elegant Graphics for Data Analysis.
>
> If you are looking for help learning how to work with spatial data in R, we recommend the following free, online books:
>
> - Spatial Data Science with applications in R by Edzer Pebesma and Roger Bivand https://r-spatial.org/book/
> - Geocomputation with R by Robin Lovelace, Jakub Nowosad, and Jannes Muenchow https://geocompr.robinlovelace.net/
> - Analyzing US Census Data: Methods, Maps, and Models in R, by Kyle Walker https://walker-data.com/census-r/
> - Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny by Paula Moraga. https://www.paulamoraga.com/book-geospatial/index.html.

One of the main points we urge caution around with respect to the visualization and mapping of area based health outcome data are the presentation of rates which are unstable due to small population sizes. As a broader principle, we emphasize the need for careful choice of the area level at which results are presented. In part, this is because it is well established that changing the areal units into which data are aggregated and analyzed can change the relationships observed in the data. This is known as the Modifiable Areal Unit Problem, which can be read about in (Wong 2004) and (Buzzelli 2020). Of note, choice of area should not be arbitrary, but should be guided by a priori reasons, including which areas make the most sense to use to answer which questions. For example, analyses of population-wide health inequities within the entire US may wish to use census tract or county level data, whereas analyses more specifically focused on the political geography of health inequities may wish to use areas with political boundaries, e.g., state or congressional legislative districts (Keena et al, 2021; Krieger, 2019; Krieger et al, 2022).

For example at small area levels like the Census tract level, the underlying population may be so small that the health outcomes observed are quite noisy because a single additional case represents a large shift in the rate. In such a situation, it it could potentially be erroneous to infer that the Census tract with the highest observed rate has the greatest underlying risk because it could be an artifact of noise. This motivates the need for spatial smoothing, and we encourage the use of spatially smoothed estimates for choropleth maps showing health outcome rates to avoid such a potential pitfall. As an introduction to these pitfalls, we recommend the Pitfalls to avoid chapter from (Gimond 2022).

It is worthwhile to remark on how different areal units can either center or marginalize social groups based on the geographic boundaries employed. It is also crucial to be clear about which social groups are excluded or inaccurately represented in the area-based data, including but not limited to people who are

unhoused, incarcerated, otherwise institutionalized or experiencing other kinds of marginalization. Warranting scrutiny are the protocols employed by the data holders to assign addresses and/or georeferenced codes to the records of persons who are not living in non-institutional residences, as well as who is counted towards the population totals of the specified geographic areas.As an example, one lesson we have learned from analyzing Boston area mortality data is that it can be quite useful to know the locations of and areas containing homeless shelters, because we have found that individuals designated as experiencing homelessness at the time of their death have had their place of death in the residential field of the death certificate listed as the address of the homeless shelter. The net impact can be to inflate the mortality rate of the census tracts in which these shelters are located.

Oftentimes georeferenced and geocoded data will have anomalous or idiosyncratic features such as areas where the rate appears to be infinite because population estimates are zero for that area despite having observed health outcomes, especially if the data are for small areas or particular sub-populations. Although there is not a single solution to these problems, a key starting place is to examine the data critically to understand the socially-produced data protocols and social distribution of the populations at issue, so as to be on the lookout for these very real problems that arise from and manifest social spatial inequities. We accordingly encourage data analysts and visualizers to consider carefully the impacts the choice of areas, their boundaries, the processes by which the locations of people are assigned to these areas (whether as “numerators” or “denominators”), and the power relations affecting who is likely to be geocoded to these areas – and who is missed – whether as a “case” or member of the population from which the cases arise (i.e., “denominator”).

To start to develop a wider perspective on how data visualization and mapping can be used, we list below a range of recommended books and articles. Because data visualization and mapping are powerful tools to communicate (or miscommunicate) data, it is critical to be aware of how they can reflect bias and tell lies (Deluca and Nelson, 2017), (Fleckenstein 1991), (Monmonier 2021), as well as reveal powerful truths (Koch 2017). Tom Koch outlines the history of disease mapping in his book *Disease Maps, Epidemics on the Ground* (2011), and in his 2017 follow-up book, *Ethics in Everyday Places: Mapping Moral Stress, Distress, and Injury,* he takes a broader perspective.

5.1 **Health Equity**

To understand the distribution of health and disease in place, it is necessary to collect, analyze, and visualize health data and area-based social metrics. It is equally important, however, that when documenting and mapping health inequities, we contextualize such data with adequate analysis of social, political, and ecological context. Shared observations of disparities in health do not necessarily translate to common understandings of cause, especially when population patterns of disease and health mirror population distributions of deprivation and privilege. Mapping and visualizing the uneven social and spatial distribution of health and disease in areas – which can be a particularly effective way of communicating health inequities to decision makers – without offering some explanatory context of the allocation of resources and hazards in these areas can perpetuate harmful ideas and actions that actually undermine the goal of eliminating health inequities.

With the increasing availability of local health data such as the CDC PLACES program, as well as advancements in the availability and accessibility of geocoding services, there is ample opportunity to disaggregate health data, particularly to the neighborhood (census tract) level. Geographic disaggregation allows for more fine-grained analyses, including multilevel spatial modeling, which can inform more “targeted” interventions. But when presented by themselves with no explanatory context, such granular data can create or reinforce what sociologist Loïc Wacquant refers to as “territorial stigmatization,” whereby the characteristics or features of a place are associated with the moral character and behavior of its residents, or vice versa—especially for people and places who are already politically, economically, and socially marginalized and/or materially deprived of important resources (Chowkwanyun and Reed 2020). For example, if some places are found to have a high concentration of illness or disease, narratives and representations of those places as “diseased,” “contaminated,” could produce or reinforce existing stigma and lead to targeted interventions such as heightened policing and surveillance in an attempt to contain and control residents, reclamation or demolition of physical structures, and social neglect and abandonment (There are too many historical case studies of this, for these potential risks to be downplayed or ignored: Craddock 2004; Molina 2006; Roberts 2009; Lopez 2009; Krupar and Ehlers 2017).

There are various approaches to countering territorial stigmatization. Mapping place-based risks and resource deficits that might help explain the spatial distribution of disease, illness, and injury along racial and socioeconomic lines can focus public and policy attention on shifting the context for health rather than individual behaviors and attitudes. For example, in the case of Covid-19, this could look like mapping and visualizing the uneven geographic distribution of preventive health care facilities or the concentration of respiratory hazards in areas of racialized concentrated poverty. Furthermore, one could map historical and political variables such as historical redlining that can offer important insight into how and why the social and spatial patterning of life-enhancing and harmful resources exists in an area as a result of racist policies and practices (Rothstein 2017; Mapping Inequality 2022; Krieger et al 2020a, 2020b; Wright et al. 2022). Additionally, asset mapping can provide helpful information about the strengths and resources of a community to facilitate discussion and action around building on these assets to address community needs and improve community health.

Analyzing and visualizing patterns of White wealth and health is also important to understanding and addressing patterns of population health and health inequities. Mapping and visualizing "racially concentrated areas of affluence" can help move research and policy attention away from a predominant concern for racially concentrated areas of poverty and toward a more holistic consideration of the full range of health outcomes, resources, and hazards in an area (Goetz et al. 2019). A focus on racially concentrated areas of affluence underscores the reality that structural racism produces both racialized concentrations of poverty (and hazards) and racialized concentrations of wealth (and health-enabling resources). Other measures such as the Index of Concentration of the Extremes (ICE), which quantifies the distribution of persons at the extremes of relationships of privilege and deprivation, also bring the full population and power relations into view, and can be scaled for use at multiple levels of geography (e.g., census block, census block group, census tract, city/town, county, etc.) (Massey 2001; Krieger et al. 2015, 2016, 2017, 2018). Initially developed to measure spatial polarization in economic terms (i.e., economic residential segregation), in public health studies we have extended its use to include novel measures of racialized residential segregation and racialized economic segregation (Krieger et al. 2015, 2016, 2017, 2018).

In summary, addressing health inequities requires a relational understanding of how systems of power and resource allocation simultaneously produce poor health for some and good health for others. This approach may require analyzing and visualizing patterns in population health and health inequities at large geographic scales such as counties and regions, rather than at the city level for example, to capture a wider range of values for health outcome data and area-based social metrics.

REFERENCES

Buzzelli M. (2020) 'Modifiable Areal Unit Problem', International Encyclopedia of Human Geography, pp. 169–173. doi:10.1016/B978-0-08-102295-5.10406-8.

Chowkwanyun M and Reed Jr AL. (2020). Racial health disparities and Covid-19—Caution and context. New England Journal of Medicine, 383(3), 201-203. doi:10.1056/NEJMp2012910

Craddock, S. (2004). City of Plagues: Disease, Poverty, and Deviance in San Francisco. University of Minnesota Press.

Deluca E. and Nelson S. (2017) 'Lying With Maps'. Available at: https://open.lib.umn.edu/mapping/chapter/7-lying-with-maps/ (Accessed: 7 June 2022).

Dorling D, Fairbairn D (1997). Mapping: Ways of Representing the World. Old Tappen, UK: Routledge.

Fleckenstein L. (1991) 'How Maps Lie', Syracuse University Magazine, December. Available at: https://surface.syr.edu/cgi/viewcontent.cgi?article=1245&context=sumagazine.

Gimond M. (2022) Intro to GIS and Spatial Analysis. Available at: https://mgimond.github.io/Spatial/index.html (Accessed: 7 June 2022).

Goetz EG, Damiano A, and Williams RA. Racially concentrated areas of affluence: A preliminary investigation. Cityscape, 21(1), 99-123.

Katsnelson, A. (2021) 'Colour me better: fixing figures for colour blindness', Nature, 598(7879), pp. 224–225. doi:10.1038/d41586-021-02696-z.

Keena A, Latner M, McGann AJM, Smith CA. Gerrymandering the States: Partisanship, Race, and the Transformation of American Federalism. Cambridge, UK: Cambridge University Press, 2021.

Koch T. (2011) Disease Maps: Epidemics on the Ground. Chicago, IL: University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/D/bo8490164.html (Accessed: 7 June 2022).

Koch T. (2017) Ethics in Everyday Places: Mapping Moral Stress, Distress, and Injury | Esri Press (2017). Available at: https://www.esri.com/en-us/esri-press/browse/ethics-in-everyday-places-mapping-moral-stress-distress-and-injury (Accessed: 7 June 2022).

Krieger N, Van Wye G, Huynh M, Waterman PD, Maduro G, Li W, Gwynn C, Barbot O, Bassett MT. Historical redlining, structural racism, and preterm birth risk in New York City (2013-2017). Am J Public Health 2020; 110(7):1046-1053.

Krieger N, Wright E, Chen JT, Waterman PD, Huntley ER, Arcaya M. Cancer stage at diagnosis, historical redlining, and current neighborhood characteristics: breast, cervical, lung, and colorectal cancer, Massachusetts, 2001-2015. Am J Epidemiol 2020; 189(10):1065-1075.

Krieger N, Waterman PD, Spasojevic J, Li W, Maduro G, Van Wye G. Public health monitoring of privilege and deprivation using the Index of Concentration at the Extremes (ICE). Am J Public Health 2016; 106: 256-253

Krieger N, Waterman PD, Gryparis A, Coull BA. Black carbon exposure, socioeconomic and racial/ethnic spatial polarization, and the Index of Concentration at the Extremes (ICE). Health & Place 2015; 34:215-228.

Krieger N, Feldman JM, Waterman PD, Chen JT, Coull BA, Hemenway D. Local residential segregation matters: stronger association of census tract compared to conventional city-level measures with fatal and non-fatal assaults (total and firearm related), using the Index of Concentration at the Extremes (ICE) for racial, economic, and racialized economic segregation, Massachusetts (US), 1995-2010. J Urban Health 2017; 94:244-258.

Krieger N, Kim R, Feldman J, Waterman PD. Using the Index of Concentration at the Extremes at multiple geographic levels to monitor health inequities in an era of growing spatial social polarization: Massachusetts, USA (2010-2014). Int J Epidemiol 2018; 47:788-819.

Krieger N. The US Census and the people's health: Public health engagement from enslavement and "Indians Not Taxed" to census tracts and health equity (1790-2018). Am J Public Health 2019; 109(8):1092-1100.

Krieger N, Testa C, Chen JT, Hanage WP, McGregor AJ. Relationship of political ideology of US federal and state elected officials and key COVID pandemic outcomes during the vaccine era: April 2021-March 2022. Lancet – Regional Health Americas (in press).

Krupar S, & Ehlers N. (2017). Biofutures: Race and the governance of health. Environment and Planning D: Society and Space, 35(2), 222-240. doi:10.1177/0263775816654475

Lopez RP. (2009). Public health, the APHA, and urban renewal. American Journal of Public Health, 99(9), 1603-1611. doi:10.2105/AJPH.2008.150136

Lovelace R, Nowosad J, and Muenchow J. (2022) Geocomputation with R. Available at: https://geocompr.robinlovelace.net/ (Accessed: 7 June 2022).

Molina N. (2006). Fit to be Citizens? Public Health and Race in Los Angeles, 1879-1939. University of California Press.

Monmonier M. (2018) How to Lie with Maps, Third Edition. Chicago, IL: University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/H/bo27400568.html (Accessed: 7 June 2022).

Nelson R, Winling L, Connolly NDB, Madron J, Marciano, R. Mapping Inequality: Redlining in New Deal America. https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58&text=about ; accessed June 17, 2022.

Ou, J. (2021) Safe colorsets. Available at: https://cran.r-project.org/web/packages/colorBlindness/vignettes/colorBlindness.html (Accessed: 7 June 2022).

Roberts Jr SK. (2009). Infectious Fear: Politics, disease, and the Health Effects of Segregation. University of North Carolina Press.

Rothstein R. (2017). The Color of Law: A Forgotten History of How Our Government Segregated America. New York: Liveright Publishing Co., W.W. Norton & Co.

Tufte D. (2001). The Visual Display of Quantitative Information. 2nd ed. Cheshire, CT: Graphics Press.

Tufte ER. (2020). Seeing with Fresh Eyes: Meaning, Space, Data, Truth. Cheshire, CT: Graphics Press.

Wong D. (2004) 'The modifiable areal unit problem (MAUP)', in WorldMinds: Geographical Perspectives on 100 Problems. Springer Netherlands. Available at: https://link.springer.com/chapter/10.1007/978-1-4020-2352-1_93.

Wright E, Waterman PD, Testa C, Chen JT, Krieger N. Breast Cancer Incidence, Hormone Receptor Status, Historical Redlining, and Current Neighborhood Characteristics in Massachusetts, 2005-2015. JNCI Cancer Spectrum 2022; 6(2); https://doi.org/10.1093/jncics/pkac016; epub on-line: Feb 18, 2022.

« 4 Getting your data

6 Analyzing your data »

6 Analyzing your data

By: Jarvis Chen, ScD

6.1 Overview of Methods

In the years since we initially presented the **Public Health Disparities Geocoding Training** in 2004, there has been a huge increase in conceptual and methodological work regarding use of ABSMs to document and analyze health inequities, and the computing tools and statistical methods to conduct these analyses. In particular, the availability of software that facilitates data access, mapping, visualization, and fitting of multilevel and spatial models has greatly enhanced the accessibility of these analytic methods for public health scientists and advocates interested in advancing health equity. Our goal in presenting these updated methods is to illustrate their applicability to monitoring, reporting, and investigating health inequities. In doing so, we highlight many of the assumptions underlying the use of these methods and show how they can be applied to real-life data that arise in public health surveillance and health equity monitoring. We comment on some of the pitfalls that may arise when working with imperfect data in real time and discuss analytic and interpretive strategies keeping in mind the overarching goal of accurately documenting health inequities for accountability. Where appropriate, we also comment on ongoing methodologic work to improve analytic approaches.

6.2 Motivating Questions

As the choice of methods for any data analysis depends fundamentally on the goals of the analysis, we will focus our presentation on the following motivating questions:

1. "What is the **geographic distribution** of area-based social metrics across the study area?" This is of particular interest for exploratory analyses in order to understand who in what areas are most affected by social advantage and disadvantage as captured by area-based social metrics. Mapping and visualization tools can be used to communicate geographic patterns and to elicit local knowledge in order to spur hypothesis generation and prioritization of research questions.

2. "What are the **social inequities** in health associated with area-based social metrics across the study area?" We highlight the importance of **descriptive epidemiology** as a first step in understanding inequities in the health of populations living in areas characterized by area-based social metrics. That is, accurate description of where the burden of social inequities falls is a necessary pre-requisite to more causally-oriented investigations or evaluations of interventions. We discuss **aggregated analyses**, which have the advantages of being straightforward to implement and report using tabulation methods and avoiding the problems of small area estimation, as well as **non-spatial regression models**. We focus on health equity settings in which understanding the joint patterning of inequities by racialized group and area-based social metric are of interest, and comment on how analyses may need to consider heterogeneity across age strata.

3. "How can geographic variation in health outcomes be modeled in order to facilitate mapping of **small-area disease estimates** and **estimation of social inequities**?" We review concepts from the extensive statistical literature on small-area estimation and disease mapping with the goal of visualizing geographic patterns in health outcomes for studies of health equity. While these methods entail more complex modeling frameworks and computational details, we focus on their application to health equity and the interpretation and use of model outputs in the context of monitoring and reporting of health disparities. We consider settings in which **multilevel modeling** approaches and **spatial modeling** approaches may be preferred and discuss considerations contributing to the choice of modeling framework.

6.3 Choice of geographic level

- A pragmatic question and approach: at what level(s) are data available?

On this page

- For small area level analyses, do the number of events support small area-level analyses?
- census tract (CT) vs. city/town, e.g. breast cancer mortality analysis
- ZIP Codes and US Census defined ZIP Code Tabulation Areas (ZCTAs)
- Modifiable Areal Unit Problem[1]
- note that the smallest level of Census geography with intercensal population estimates is the county. The US Census Population Estimates Program (PEP) estimates rely on demographic modeling.
- CT level population estimates are in the US Census American Community Survey (ACS), but ACS specifically cautions against using these as population denominator estimates in small area estimation. However, we generally do not have any other sources of population denominators on which to rely. Five year average estimates may help, but there is still year to year variability. Margin of error estimates are available, but how to incorporate denominator uncertainty into analyses is still an area of active research.
- For a review of issues connected to numerator/ denominator mismatch, please visit Section 4.7 in the "Getting your Data" chapter.

6.3.1 **Example: Age-specific all-cause mortality by racialized group in Massachusetts, 2013-2017**

We compared the use of the aggregation method to a non-spatial Poisson model for analyzing deaths from all causes by age and racialized group. Using the aggregation method, deaths and population person-time at risk are aggregated into strata by age (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+) and racialized group (White non-Hispanic and Black), and mortality rates per 100,000 person-years are computed using the formulas listed in Section 6.5.1. For the non-spatial Poisson model, the inputs are age- and racialized-group specific deaths for each census tract in the study area, the log(population person-time at risk) as an offset, and a set of indicators for each stratum of age category by racialized group.

When the log(population person-time at risk) variable is zero, however, this presents a problem for the Poisson model-fitting algorithm as the rate is undefined. Census tracts where there are zero deaths and zero person-time at risk in a particular strata can be deleted from the dataset as these observations contribute no information. But situations where the count of deaths in a stratum is greater than zero but the population person-time at risk is zero continue to be an issue (numerator/denominator mismatch). It may be tempting to delete these observations from the dataset as well, reasoning that as these usually involve just one death occurring in a stratum where the population estimate is zero, deleting these observations will not unduly affect the analysis.

Unfortunately, it turns out that numerator/denominator mismatch occurs more frequently for Black observations in the dataset, due to patterns of racialized residential segregation and that fact that the Black population accounts for a much smaller proportion of the population compared with the White non-Hispanic population (10.1% vs. 72.9% in Massachusetts according to the ACS population estimates we used for our population denominators). That is, at the census tract level, there are more likely to be age strata where there are deaths reported by zero population person-time at risk for the Black population compared with the White non-Hispanic population. As a result, when we delete these observations, we end up deleting a larger proportion of the Black deaths across the whole state, which has an impact on our estimates of rates and rate ratios.

To see this, consider the following table where we show the number of deaths and population by age and racialized group aggregated across census tracts. In columns 2-8, we show the deaths, population, and estimated rates per 100,000 person-years for White Non-Hispanics and Blacks in the full dataset, while in columns 9-15 we show the deaths and population after deleting observations where the numerator is greater than zero but the denominator is zero. In columns 16-22 we show the percent bias by strata comparing the dataset with deleted observations to the full dataset. As is evident in column 19, the effect of deleting these observations is to reduce the Black death count in age strata (between 4-30% across strata) to a much greater degree than among the White non-Hispanic age strata (between 0-2%). As a result, the age-specific incidence rate ratios can be depressed by as much as 30% relative to the IRRs calculated using the full aggregated data.

	Full dataset							Ad		
Age	NHW deaths	NHW pop	NHW Rate	Black deaths	Black pop	Black Rate	IRR	NHW deaths	NHW pop	NHV Rat
0-4	769	1068730	5.0	316	177968	12.3	2.47	766	1068730	5.
5-9	90	1147677	0.6	14	172883	0.6	1.03	88	1147677	0.
10-14	114	1301178	0.6	29	176597	1.2	1.87	114	1301178	0.
15-19	371	1533896	1.7	100	198678	3.6	2.08	371	1533896	1.
20-24	1103	1611714	4.5	163	224987	4.8	1.06	1102	1611714	4.
25-29	1788	1615532	7.1	244	212885	7.4	1.04	1787	1615532	7.
30-34	2079	1518953	9.7	213	183138	8.3	0.85	2078	1518953	9.
35-44	4846	2873290	27.4	517	334861	25.1	0.92	4846	2873290	27.
45-54	12386	3756899	44.5	1165	325584	48.2	1.09	12386	3756899	44.
55-64	25334	3735786	59.2	1997	262045	66.5	1.12	25326	3735786	59.
65-74	38765	2527556	101.3	2282	139490	108.0	1.07	38752	2527556	101.
75-84	58096	1335819	195.0	2374	67943	156.7	0.80	58057	1335819	194.
85+	101687	711624	221.6	2571	23810	167.5	0.76	101585	711624	221.

If instead of deleting those strata where deaths>0 and denominator=0, we replace the denominator with the number of deaths, we end up with very slightly larger population counts overall, but the effect on the rates is to bring them more in line with the aggregated analysis. Here is the comparable table comparing the full aggregated dataset (columns 2-8) with a dataset in which the denominators are adjusted to increase the person time at risk to equal the number of deaths the affected strata.

	Full dataset							Ad		
Age	NHW deaths	NHW pop	NHW Rate	Black deaths	Black pop	Black Rate	IRR	NHW deaths	NHW pop	NHV Rat
0-4	769	1068730	5.0	316	177968	12.3	2.47	769	1068733	5.

	Full dataset							Ad		
Age	NHW deaths	NHW pop	NHW Rate	Black deaths	Black pop	Black Rate	IRR	NHW deaths	NHW pop	NHV Rat
5-9	90	1147677	0.6	14	172883	0.6	1.03	90	1147679	0.
10-14	114	1301178	0.6	29	176597	1.2	1.87	114	1301178	0.
15-19	371	1533896	1.7	100	198678	3.6	2.08	371	1533896	1.
20-24	1103	1611714	4.5	163	224987	4.8	1.06	1103	1611715	4.
25-29	1788	1615532	7.1	244	212885	7.4	1.04	1788	1615533	7.
30-34	2079	1518953	9.7	213	183138	8.3	0.85	2079	1518954	9.
35-44	4846	2873290	27.4	517	334861	25.1	0.92	4846	2873290	27.
45-54	12386	3756899	44.5	1165	325584	48.2	1.09	12386	3756899	44.
55-64	25334	3735786	59.2	1997	262045	66.5	1.12	25334	3735794	59.
65-74	38765	2527556	101.3	2282	139490	108.0	1.07	38765	2527569	101.
75-84	58096	1335819	195.0	2374	67943	156.7	0.80	58096	1335858	195.
85+	101687	711624	221.6	2571	23810	167.5	0.76	101687	711726	221.

An even better solution, and the one we ultimately recommend, is to add a small number (e.g. 0.001) to all of the population person-time denominators to allow the Poisson model fitting algorithm to incorporate these observations into the analysis. The resulting effect on the overall population denominators is negligible, but this allows for estimation of rates and rate ratios of interest. Note that this is only needed when fitting non-spatial regression models to the data. The aggregation method does not incur this problem as small discrepancies are averaged out over areas, and the spatial models we discuss in Section 6.6 address the problem of zero or infinite rates by smoothing.

To visualize the effect of these adjustments on estimates of the age-specific disparity by racialized group, we plot estimates from aggregated analyses and Poisson models with deletion and denominator adjustment in Figure 1.1. The plot confirms that either (a) adjusting denominators to match the number of observed deaths in problematic strata in census tracts or (b) adding a small number to all population person time estimates (preferred) yields comparable estimates to the aggregation method, whereas removing problematic observations results in bias.

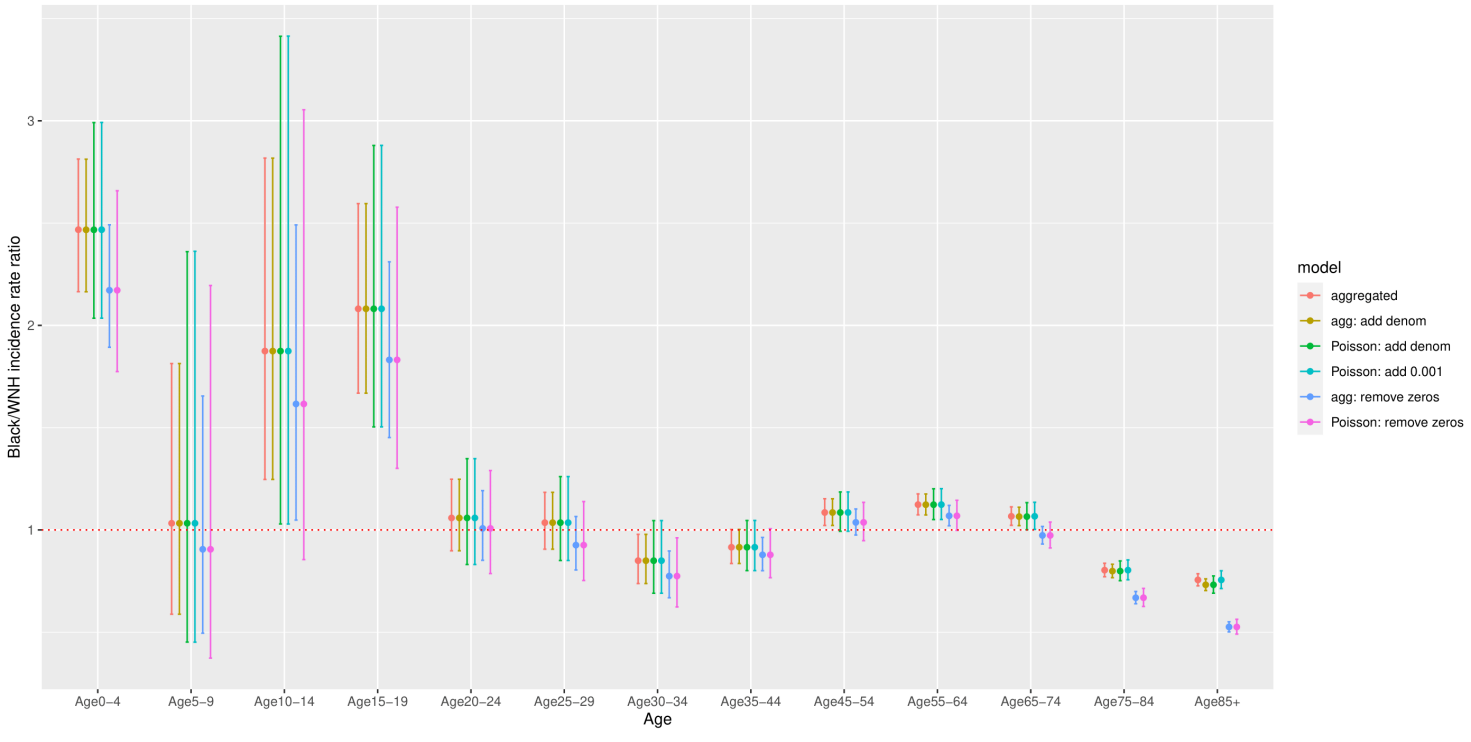


Figure 6.1: Comparison of estimated age-specific Black/White mortality rate ratios using different methods to address numerator/denominator mismatch

6.4 **Aggregation Method**

As we presented in the original *Public Health Disparities Geocoding Project Monograph* (Krieger et al, 2004), one of the most straightforward ways to incorporate area-based social metrics into analyses of health inequities is to use what we term the Aggregation Method. This method entails using geocodes to append area-based social metrics to health surveillance records, to stratify these records into discrete categories based on ABSM, and to aggregate numerators and denominators over areas, within levels defined by ABSM. Rates, including age-standardized rates, and measures of association (e.g. rate differences or rate ratios) can be easily computed using tabulation methods and formulae that are taught in all basic epidemiologic textbooks. The analyses are straightforward and do not require specialized software: they can even be done in Excel or other spreadsheet programs. One of the key advantages is that aggregation avoids problems with numerator/denominator mismatch since small discrepancies in case counts and population at risk tend to get averaged out in aggregating numerators and denominators over small areas within ABSM strata.

We can describe inequities by categories of ABSMs by aggregating deaths and population at risk from census tracts that share the same values of those ABSMs. This is analogous to what health departments typically do when reporting, for example, statewide cancer mortality rates by age or gender.

This method avoids the problem of unstable rates arising from small areas by assuming that cases and population denominators from areas with similar socioeconomic characteristics can be validly combined into the same strata.

The following steps are used to generate age-standardized disease rates stratified by area-based social metrics once the case data have been geocoded and appropriate ABSMs have been generated from census data.

1. Aggregate the case data into numerators (age cells within areas/geocodes).
2. Aggregate population denominator data into age cells within areas/geocodes.
3. Merge the numerators and denominators with ABSMs, by area/geocode.
4. Aggregate over areas into strata defined by categorical ABSM and age category.
5. Generate age-standardized rates and other summary measures.

Advantages:

- easy to implement using tabulation methods; can even be done in MS Excel
- formulae are taught in basic epidemiologic textbooks

- presentation and interpretation of social inequities is straightforward
- aggregation avoids problems with numerator-denominator mismatch
- age-standardization using the direct method has coherent interpretation even in the presence of effect heterogeneity by age
- Confidence intervals may be wider particularly when age strata with relatively less information get upweighted

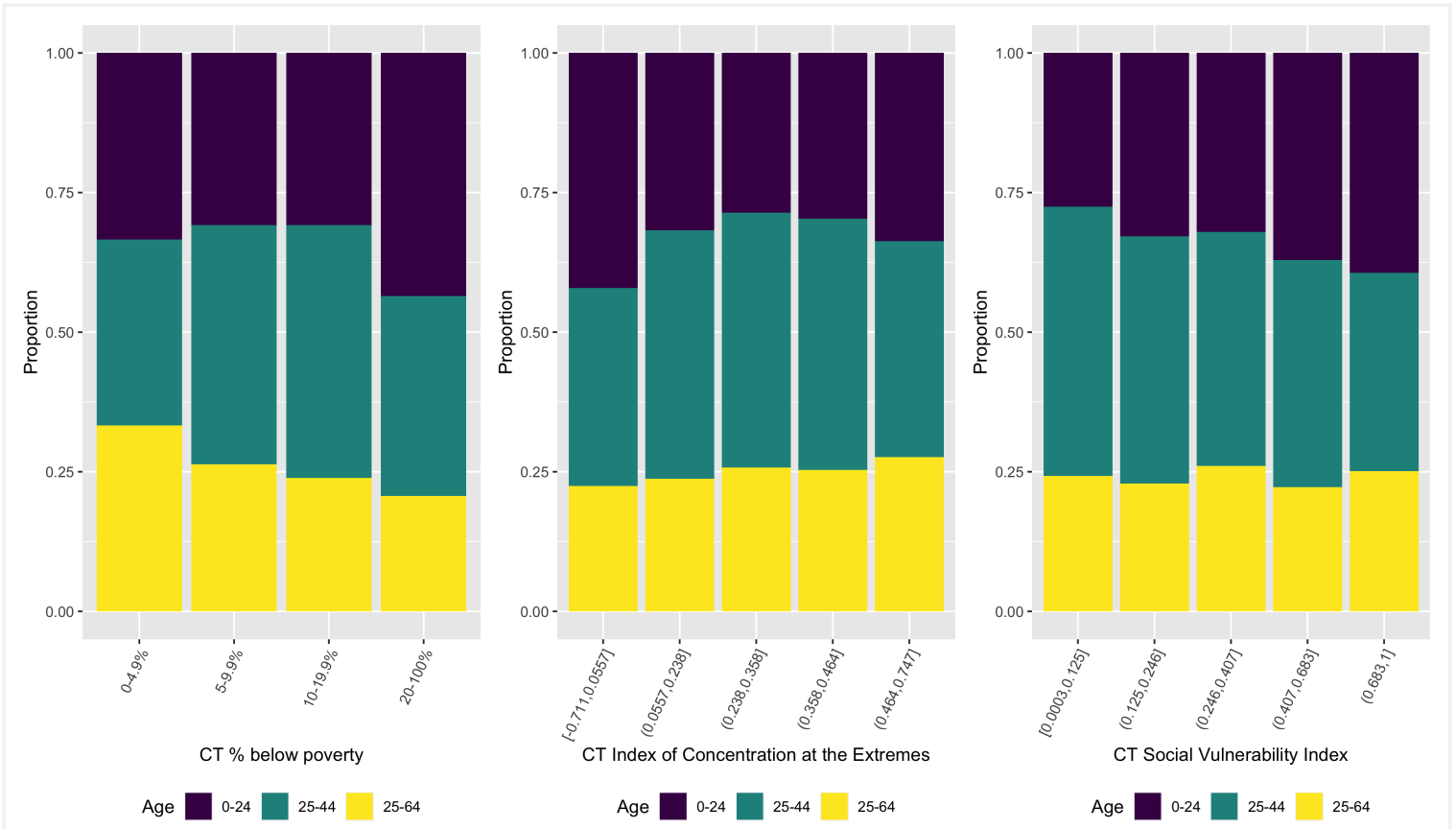
Disadvantages:

- precludes identification of small areas with unusually high or low disease rates
- may preclude further adjustment for compositional differences or area-level covariates that vary within ABSM strata

Throughout this section, we will be using data from the Greater Boston Area for the years 2013-2017, drawing on Massachusetts Mortality data and the American Community Survey (ACS) population data.

6.4.1 **Direct Age Standardization**

- rate estimation
- confidence intervals on the rates
- measures of disparity: incidence rate difference and incidence rate ratio
- population attributable fraction
- a comment on life expectancy



CT % below poverty	Age category	Deaths	Person-years	Weight	Rate per 100,000 p-y
0-4.9%	0-4	27	33162	0.069	5.63
0-4.9%	5-9	3	31593	0.073	0.69
0-4.9%	10-14	1	33800	0.073	0.22
0-4.9%	15-19	7	38060	0.072	1.33
0-4.9%	20-24	12	30408	0.066	2.62

CT % below poverty	Age category	Deaths	Person-years	Weight	Rate per 100,000 p-y
0-4.9%	25-29	20	45937	0.065	2.81
0-4.9%	30-34	31	43138	0.071	5.11
0-4.9%	35-44	51	77292	0.163	10.73
0-4.9%	45-54	145	85527	0.135	22.86
0-4.9%	55-64	330	81124	0.087	35.49
20-100%	0-4	156	107867	0.069	10.00
20-100%	5-9	15	94231	0.073	1.15
20-100%	10-14	11	94853	0.073	0.85
20-100%	15-19	64	209056	0.072	2.21
20-100%	20-24	115	308641	0.066	2.48
20-100%	25-29	168	254985	0.065	4.25
20-100%	30-34	176	182451	0.071	6.85
20-100%	35-44	422	232327	0.163	29.54
20-100%	45-54	867	208522	0.135	56.06
20-100%	55-64	1580	177590	0.087	77.62

CT % below poverty	Deaths	Person-years	Standardized Rate per 100,000 p-y	Var(Rate)	Sum of weights	Sum of weights^2	std_rate_lo9
0-4.9%	627	500041	100	0	0.873614	0.086	100.134
20-100%	3574	1870523	219	0	0.873614	0.086	218.646

6.5 Non-Spatial Regression Methods

6.5.1 Poisson regression

While the aggregation method is attractive for its conceptual simplicity and ease of implementation, some may prefer to take a regression approach to analysis of health inequities by ABSM. As we will describe below, this makes it possible to relax the strong assumption of a homogenous Poisson process within category of ABSM. We might therefore choose to model the data (deaths and population at risk by age stratum within census tracts) using a Poisson loglinear model – that is, a generalized linear model with a Poisson error distribution and a log link.

Let Y_{ij} be the count of deaths in age-stratum j in census tract i and n_{ij} be the corresponding person-time at risk. We fit a Poisson log linear model to the data and include dummy variables for ABSM categories, e.g. with x_1, x_2, x_3, and x_4 coded 1 if the census tract is in the corresponding category of CT % below

poverty and 0 otherwise. We model age categories as a set of dummy variables. We also include $\log(n_{ij})$ as an offset in the model for $\log(\mu_{ij})$.

$$
\begin{aligned}
Y_{ij} &\sim \text{Poisson}(\mu_i) \\
\log(\mu_{ij}) &= \beta_0 + \beta_1 I(pov = 5 - 9.9\%) + \beta_2 I(pov = 10 - 19.9\%) + \beta_3 I(pov = 20 - 100\%) \\
&\quad + \sum_{j=2}^{J} \alpha_j I(age = age_j) + \log(n_{ij})
\end{aligned}
$$

6.5.2 Quasi-poisson regression

A key assumption of the Poisson distribution is that the variance of the expected count equals the mean,

$$\text{Var}(Y_{ij}) = \mathbb{E}(Y_{ij}) = \mu_{ij}$$

In real life, however, we often find data where the empirical variance is larger than the mean. This is known as **overdispersion**.

To diagnose potential overdispersion, we can look at the standardized residuals, defined as

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$

If the Poisson model is true, then the $\{z_i\}$s should be approximately independent, each with mean 0 and standard deviation 1. If there is overdispersion, we would expect the $\{z_i\}$s to be larger in absolute value, reflecting the extra variation beyond what is predicted under the Poisson model.

We can test for overdispersion by computing the sum of squares of the standardized residuals and comparing this to the χ^2_{n-k} distribution where n is the sample size and k is the number of parameters in the Poisson log linear model.

```
> df_overdispersion <- df_forModel
> df_overdispersion$yhat <-  predict(poisson_apINDPOV,
+                                    newdata=df_forModel %>%
+                                    dplyr::select(denominator,
+                                                  apINDPOV_2,
+                                                  apINDPOV_3,
+                                                  apINDPOV_4,
+                                                  apINDPOV_NA,
+                                                  agecat), type="response")
>
> df_overdispersion$z <- (df_overdispersion$numerator -
df_overdispersion$yhat)/sqrt(df_overdispersion$yhat)
> n.obs <- length(df_overdispersion$numerator)
> k <- 4
> cat("overdispersion ratio is ",sum(df_overdispersion$z^2, na.rm=T)/(n.obs-k),"\n")
overdispersion ratio is  2.144269
> cat("p-value of overdispersion test is ", pchisq(sum(df_overdispersion$z^2, na.rm=T), n.obs-k,
lower.tail=FALSE), "\n")
p-value of overdispersion test is  0
```

We can also visualize the predicted counts vs. the standardized residuals.

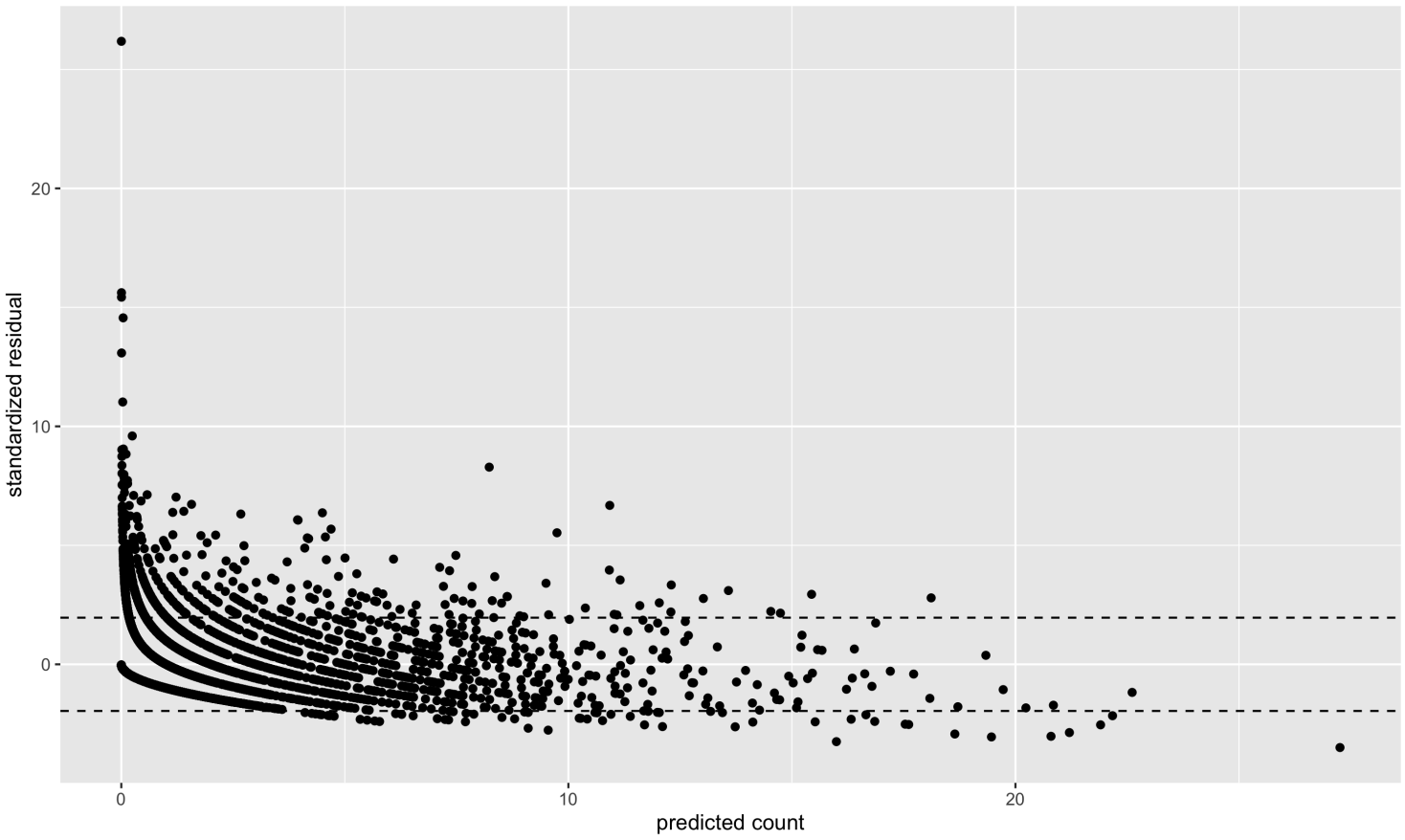


Figure 6.2: Predicted counts vs. standardized residuals. The standardized residuals should have mean 0 and standard deviation 1 (hence the dashed lines at ± 2 indicating approximate 95% error bounds). The variance of the standardized residuals is larger than 1, indicating in this case a moderate amount of overdispersion.

6.5.3 **Negative binomial regression**

- aggregation method and Poisson regression coincide in the analysis of crude rates
- Poisson regression assumes that the variance is equal to the mean, but real-life data often exhibit overdispersion (actually more complicated with age-specific data, where some age strata may be underdispersed and others may be overdispersed)
- Quasi-poisson regression estimates an extra scale parameter: results in wider confidence limits when there is overdispersion. ABSM estimates are identical to the Poisson model fit.
- Negative binomial model assumes a different distribution to allow for overdispersion. Can be conceived of as a mixture of Poisson distributions where the latent variable is gamma distributed
- Negative binomial model yields ABSM estimates that can vary from the Poisson and Quasi-Poisson fits

How do these models handle numerator/denominator mismatch? Zero numerator/zero denominator observations need to be deleted. Non-zero numerator/zero denominator areas present a problem. If we delete these areas, the undercount of cases can be differential by racialized group; a better approach is to add a small number (e.g., 0.1) (the effect on the total denominators is negligible, but this yields rates and ABSM estimates more similar to the aggregation method).

6.5.4 **Comparison of non-spatial regression estimates**

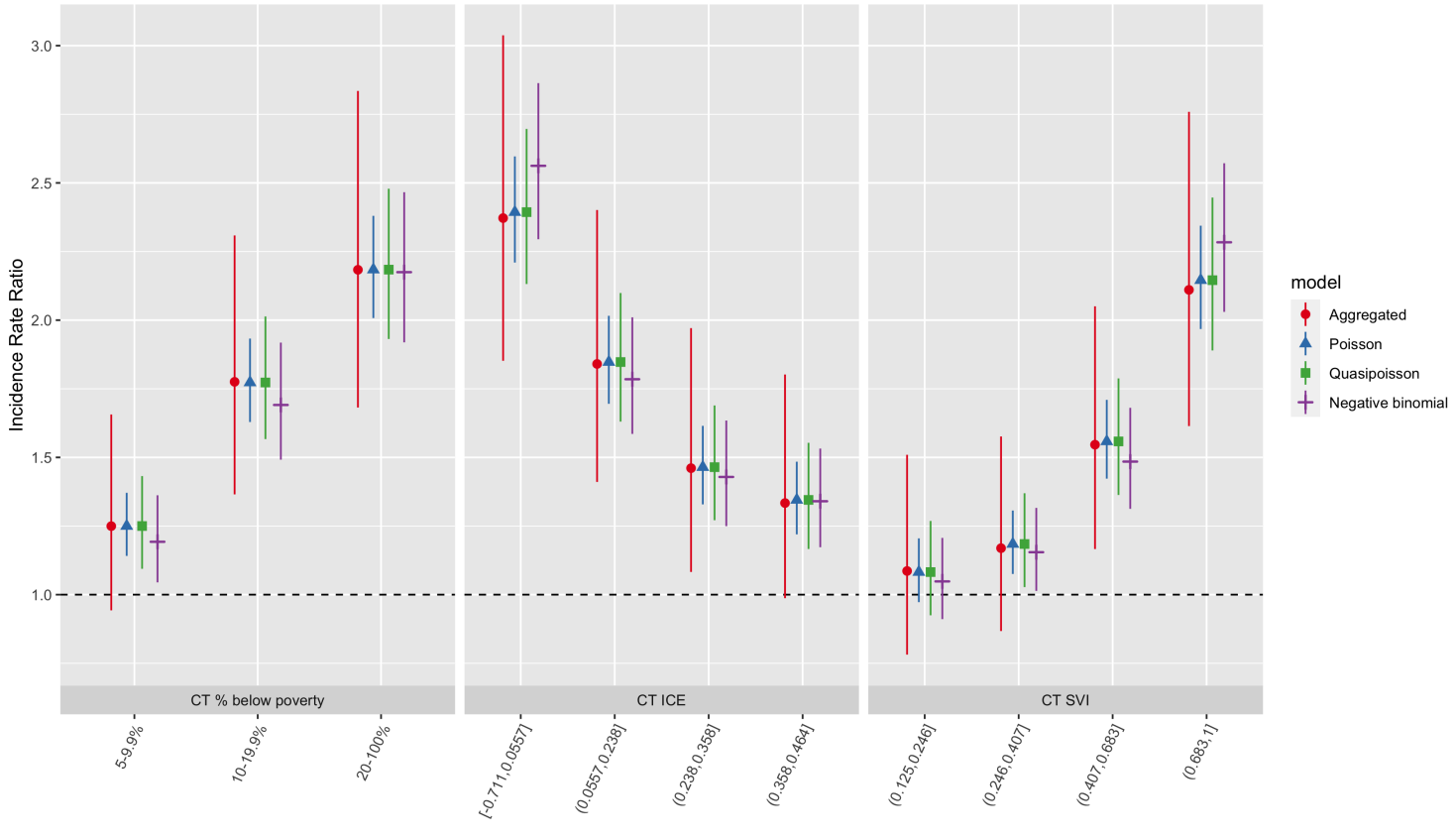


Figure 6.3: Comparison of estimates of social inequities by ABSM from the aggregated method and from non-spatial regression models

6.6 Small Area Estimation

6.6.1 Indirect Age Standardization

We observe y_{ij}, the count of deaths in area i in age stratum j. We define O_i as the observed number of deaths in area i, summed up over the age strata: $O_i = \sum_j y_{ij}$. The expected number of deaths, E_i, is computed by applying age-specific mortality rates from a reference population to the age-specific population counts in each area:

$$E_i = \sum_j N_{ij} \times R_j$$

where N_{ij} is the person-time at risk in age group j in area i (computed by taking the population in age-group j in area $i\times$ number of years) and R_j is the mortality rate age group j of a suitable reference population. The ratio O_i/E_i is known as the Standardized Incidence Ratio (SIR), or, in the case of mortality outcomes, the Standardized Mortality Ratio (SMR). This ratio is an estimate of the within each area, θ_i. For rare, non-communicable diseases, the standard statistical model for O_i is that of the Poisson distribution:

$$O_i \sim \text{Poisson}(\theta_i E_i).$$

The maximum likelihood estimator of θ_i is

$$\hat{\theta}_i = \text{SMR}_i = \frac{O_i}{E_i}$$

with $\text{Var}(\text{SMR}_i) = \theta_i/E_i$, estimated by O_i/E_i^2. Note that $\text{Var}(\text{SMR}_i)$ is inversely proportional to E_i. To translate the SMR into an age-standardized rate, we can multiply the SMR by the overall mortality rate from the reference population. In these examples, however, we will present the small area estimates on the SMR scale, since we are interested in the pattern of increased or decreased relative risk across small areas.

Indirect age standardization and SMRs are particularly useful when the age-specific counts of deaths are not available for each area, since all that is required are age-specific population estimates for each area, a set of age-specific reference rates, and the **total cases** in each area. Interestingly, this method also produces rate estimates with smaller asymptotic variance than the corresponding direct standardization method (Pickle and White, 1995).

As summary measures, SMRs also have certain drawbacks. They are based on ratio estimators, and are thus sensitive to small changes in E_i. In particular, when E_i is close to zero, the SMR will be very large for any positive count. As the estimate of $\mathrm{Var}(\mathrm{SMR}_i)$ is proportional to $1/E_i$, SMRs of zero do not distinguish variation in expected counts. Most importantly, interpretability and comparability of SMRs based on indirectly age standardized data across areas depends on the assumption of independent area and age effects with respect to the standard population. This is known as the proportionality assumption, and assumes that $r_{ij} = \theta_i \times \alpha_j$ where α_j is an age effect that does not vary over areas.

To determine whether an $SMR > 1$ is inconsistent with a chance occurrence, we can carry out a statistical hypothesis test:

$$\begin{aligned} \text{Null hypothesis} &: \theta_i = 1 \\ \text{Alternative hypothesis} &: \theta_i > 1 \end{aligned}$$

We test this hypothesis as follows: Under the null hypothesis, the mean of the Poisson distribution is E_i

$$O_i \sim \mathrm{Poisson}(E_i)$$

Calculate the probability of obtaining at least O_i cases by chance from a Poisson distribution with mean E_i :

$$\begin{aligned} \Pr(X \geq O_i | \mathrm{Exp} = E_i) &= 1 - \Pr(X < O_i | \mathrm{Exp} = E_i) \\ &= 1 - \sum_{X=0}^{O_i - 1} \frac{E_i^X e^{-E_i}}{X!} \\ &= p \end{aligned}$$

This probability p is the (1-sided) p-value. If $p \leq 0.05$, we usually reject the null hypothesis and accept the alternative. We say that the excess risk of disease in area i is **statistically significant** at the 5% level. Such a finding might warrant further epidemiological investigation.

Consider, however, that we have 306 census tracts in the Greater Boston mortality example. The statistical test described above was just for the comparison of one census tract's SMR relative to the null hypothesis. If we want to identify all of the census tracts in our study area with significantly elevated rates, we would have to repeat this tests many times, which would create a multiple hypothesis testing situation. Moreover, we could imagine that there would be dependence of tests for nearby areas if neighboring areas shared some risk factor which increased the SMR for these clusters of census tracts.

To overcome this variability, hierarchical models can be used to "smooth" the raw rates. When faced with the problem of making inferences on many parameters $\{\theta_i\} = \theta_1, \ldots, \theta_n$, measured on n areas, one can imagine two possible extreme assumptions:

- We could assume that all of the $\{\theta_i\}$ are **identical**, in which case all the data can be pooled, and the individual units ignored. This is what we typically do when presenting summary rates and rate ratios over the whole study area.
- At the other extreme, we could assume that all the $\{\theta_i\}$ are **independent** and entirely unrelated. In this case, the SMR from each area would have to be estimated independent of the data for other areas. As we have saw in the SMR example above, this leads to statistical instability if the numbers are small.

A third possible assumption lies somewhere between these two extremes. One could assume that the $\{\theta_i\}$ are "similar" in the sense that the area labels convey no additional information. This is known as **exchangeability**, and is equivalent to assuming that $\{\theta_i\}$ are drawn from a common prior distribution with unknown parameters.

6.6.2 Poisson gamma model

A classic example of this approach is presented by Clayton and Kaldor (1987), who developed a Bayesian analysis of a Poisson likelihood model. Their model is a useful introduction to the idea of hierarchical modelling of disease rates, as the second stage distribution for the area variability is analytically tractable and helps to build intuition about how smoothing works to stabilize the SMR estimates.

In the first stage of the hierarchy, we assume that the observed death counts for each area are Poisson distributed:

$$O_i \sim \mathrm{Poisson}(\theta_i E_i).$$

In the second stage, a hierarchical prior is placed on θ_i:

$$\theta_i \sim \text{Gamma}(\nu, \alpha).$$

Recalling that the gamma distribution with parameters ν and α has mean ν/α, this simply states that we expect the distribution of $\{\theta_i\}$ to follow a gamma distribution with mean ν/α and variance ν/α^2. Since the gamma distribution is the conjugate prior of the Poisson, the posterior distribution of $p(\theta_i|O_i, E_i)$ also follows a gamma distribution:

$$\text{Gamma}(\nu + O_i, \alpha + E_i)$$

with mean given by

$$\mathbb{E}(\theta_i|O_i, \nu, \alpha) = \frac{\nu + O_i}{\alpha + E_i} = w_i \text{SMR}_i + (1 - w_i)\frac{\nu}{\alpha}$$

where

$$w_i = \frac{E_i}{\alpha + E_i}.$$

The expression for $\mathbb{E}(\theta_i|O_i, \nu, \alpha)$ shows that the posterior mean of the relative risk for the ith area is a *weighted average* of the observed SMR for the ith area and the average relative risk (ν/α) over all areas. The weight is inversely proportional to the variance of the SMR. Accordingly, when E_i is small (for rare diseases or small population counts), the variance is large, so the weight w_i is small and the posterior mean is dominated by the prior mean, ν/α. In areas with abundant data, the posterior mean is close to the observed $SMR_i = O_i/E_i$. This feature, whereby the amount of smoothing is proportional to the amount of information available for a particular area, is known as **precision weighting**. It has an intuitive appeal in that, when one does not observe a lot of information about an area (because the sample size is small and the risk estimate is unstable), one's "best guess" concerning that area's mortality risk should be weighted towards what little is known from prior knowledge, i.e. that the risk is, on average, ν/α. In contrast, if one observes a lot of information for an area (e.g. because the sample size is large), one is more likely to believe what the data say about mortality risk in that particular area, and thus the "best guess" would reasonably be weighted towards the observed SMR for that specific area.

In the Empirical Bayes approach developed by Clayton and Kaldor (1987), ν and α are replaced by their estimates, $\hat{\nu}$ and $\hat{\alpha}$, which can be calculated by means of an iterative procedure using the following two equations:

$$\frac{\hat{\nu}}{\hat{\alpha}} = \frac{1}{n}\sum_i^n \frac{O_i + \hat{\nu}}{E_i + \hat{\alpha}} = \frac{1}{n}\sum_i \hat{\theta}_i$$

$$\frac{\hat{\nu}}{\hat{\alpha}^2} = \frac{1}{n-1}\sum_i^n \left(1 + \frac{\hat{\alpha}}{E_i}\right)\left(\hat{\theta}_i - \frac{\hat{\nu}}{\hat{\alpha}}\right)^2$$

where $\{\hat{\theta}_i\}$ are the empirical Bayes estimates. Together, these two equations can be used recursively to compute $\hat{\nu}$ and $\hat{\alpha}$. At each stage of the iteration, the $\{\hat{\theta}_i\}$ are calculated from the current estimates of ν and α, and then the right hand sides of the two equations are used to provide new estimates of ν and α (Clayton and Kaldor, 1987).

There is an interesting connection here to negative binomial regression in that the marginal posterior distribution of O_i (unconditional on θ_i) is negative binomial with size ν and probability $\alpha/(E_i + \alpha)$.

We can apply this algorithm to the observed and expected death counts in the census tracts of our study area to obtain empirical Bayes estimates of the census tract level SMRs, which we compare to the raw $SMR_i = O_i/E_i$:

```
# Use DCluster::empbaysmooth to fit the Poisson Gamma model as proposed by Clayton and Kaldor
(1987)
poisson_gamma <- DCluster::empbaysmooth(df_indirect_ordered$O, df_indirect_ordered$E)

# Append these estimates to the dataset and also calculate naive empirical Bayes credible
intervals
# using the gamma distribution.
df_eb <- df_indirect_ordered %>%
         mutate(nu1 = O + poisson_gamma$nu,
                alpha1 = E + poisson_gamma$alpha,
                empbayes_SMR = poisson_gamma$smthrr,
                empbayes_CI95low = qgamma(0.025, nu1, alpha1),
                empbayes_CI95up = qgamma(0.975, nu1, alpha1),
                eb_sig = factor(case_when(
                  empbayes_CI95low>1 & empbayes_CI95up>1 ~ 1,
                  empbayes_CI95low<1 & empbayes_CI95up<1 ~ -1,
                  TRUE ~ 0)),
                raw_SMR = O/E, # recenter rawSMRs by adding intercept from intercept only model
                raw_SMR_CI95low = pois.exact(x=O, pt=E, conf.level=0.95)[,4],
                raw_SMR_CI95up = pois.exact(x=O, pt=E, conf.level=0.94)[,5],
                raw_sig = factor(case_when(
```

From the caterpillar plots, we observe that the empirial Bayes estimates are substantially smoothed relative to the raw SMRs. The observations at the extremes of the distribution with huge confidence intervals are smoothed towards the mean, and the spread of smoothed SMRs is generally narrower than the raw SMRs.

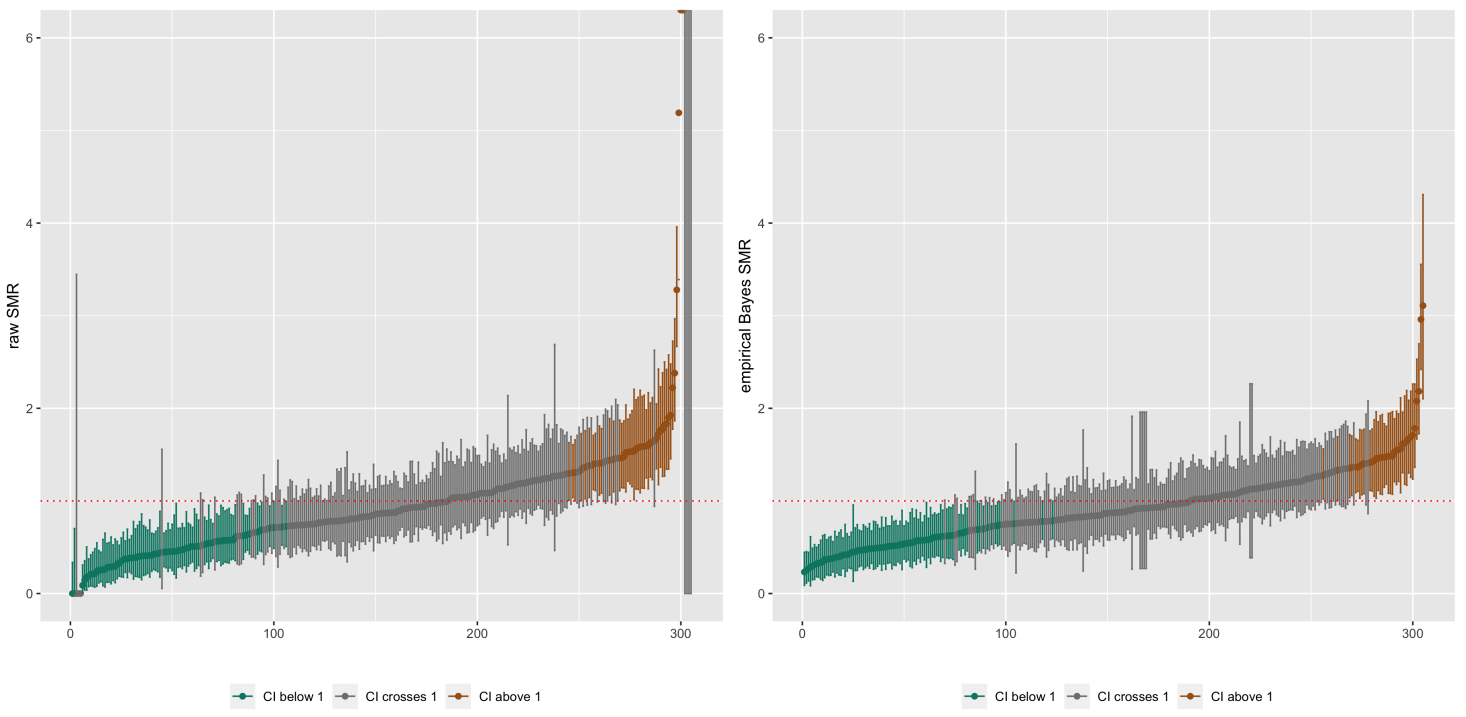


Figure 6.4: Caterpillar plots of raw SMRs with 95% CIs and smoothed SMRs from a Poisson gamma model with 95% credible intervals

This is also evident if we compare maps of the raw SMRs to the empirical Bayes SMRs, where the more extreme low or high SMRs are smoothed towards the middle values.

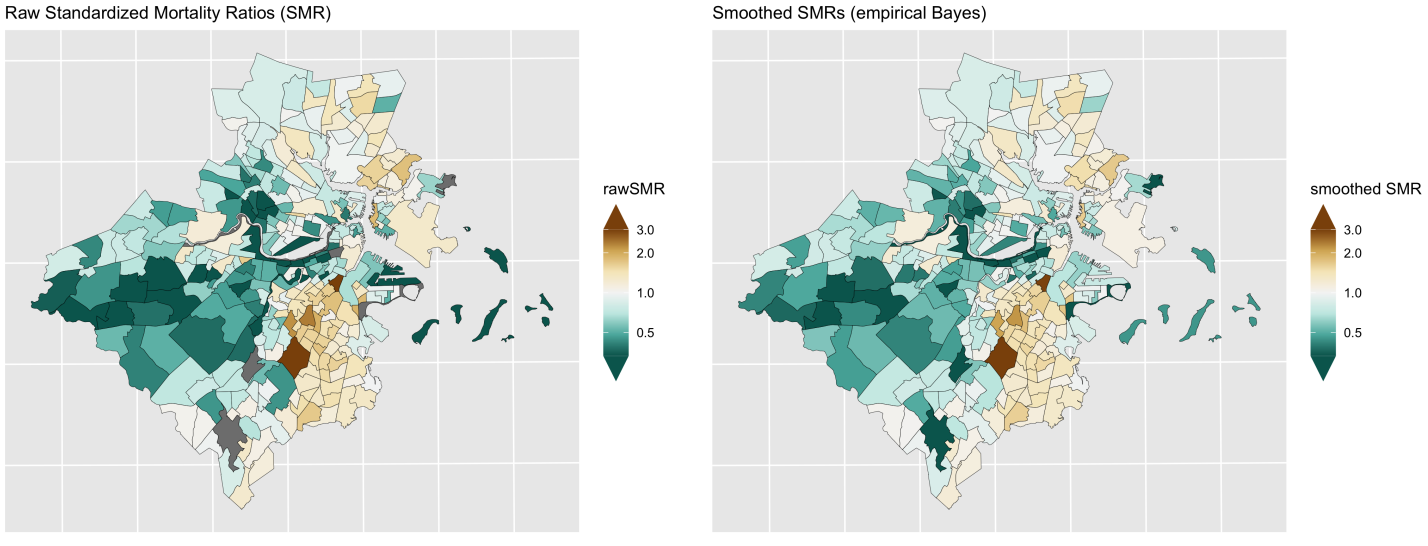


Figure 6.5: Maps of raw SMRs (left) and smoothed SMRs from a Poisson gamma model

6.6.3 **Poisson log normal model**

While a gamma prior for $\{\theta_i\}$ is mathematically convenient, it can be restrictive in that it is difficult to incorporate observed covariates into the model. A more flexible model makes use of a normal prior for $\log(\theta_i)$ (Wakefield et al., 2000; Lawson et al., 2003):

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \alpha + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

where α is an intercept term representing the overall log relative risk of disease in the whole study region compared to the reference rate and v_i is the residual log relative risk in area i compared with the average over the study region. Here, the $\{v_i\}$ are assumed to arise from a normal distribution with mean 0 and variance σ_v^2. You may recognize this model as a generalized linear mixed model, where "mixed" refers to the fact that the model contains both "random effects" (the $\{v_i\}$), as well as accommodating "fixed" covariate effects, e.g.

$$\log(\theta_i) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p + v_i.$$

In this case the $\{v_i\}$ are interpretable as *__residual__ area-specific effects conditional on the fixed covariates. Similarly, the βs are interpretable as covariate effects **conditional** on the area random effects.

Fixed vs Random Effects: in standard regression models, fixed and random effects refer to the type of statistical model being used. When using fixed effects analysis of variance (ANOVA), the assumptions being made are about the independent variable and the error distributions for the variable. Fixed effects are estimated with maximum likelihood (the traditional beta estimates we see in regression models). They are most appropriate when trying to generalize results to values for the fixed variables used in the study (fixed variables are "assumed to be measured without error … and assumed that the values of a fixed variable in one study are the same as values of the fixed variable in another study") (Newsom, 2019). If however, the researcher is seeking to make inferences beyond particular values of the independent variable, a random effects model is used. Random effects are estimated with shrinkage (partial pooling or linear unbiased predictions). Random effects models are accounting for "additional expected random variation on the independent variable" and instead of the value of the variable itself being of interest, the random variables "are assumed to be values that are drawn from a larger population of values and thus will represent them" (Newsom, 2019; Gelman, 2005)

It should be noted here that, as pointed out by Wolpert and Ickstadt (1998), the Poisson log normal model does not aggregate consistently. That is, if one specifies a log normal distribution for each of the relative risks and then combines two areas and specifies a log normal distribution for the relative risk of the combined area, then these distributions are inconsistent (because the sum of log normal distributions is not log normal). This can be understood as a form of aggregation bias whereby risk relationships do not remain constant across levels of aggregation (Wakefield et al., 2000). Nevertheless, a normal second-stage distribution has been observed empirically to provide a good model for log relative risks over a range of aggregations, and does present advantages with respect to model flexibility and ease of computation.

Unlike the Poisson gamma model, there is no analytically tractable closed-form solution for the posterior distributions. Instead, we will use INLA to fit this model.

```
# Use INLA to fit Poisson gamma model
model_form <- O ~ 1 + f(id_order, model="iid")
model_iid <- inla(model_form, family="poisson",
                  data=df_indirect_ordered, E=E, # E points to the expected count field
                  control.predictor=list(compute=TRUE), # computes transformed posterior marginals
                  control.compute=list(dic=TRUE, waic=TRUE)) # computes DIC for model fit
summary(model_iid)
```

We can see from the caterpillar plots of the raw SMRs vs. the smoothed SMRs from the Poisson lognormal model that, similar to the Poisson gamma model, the Poisson lognormal model smooths extreme SMRs towards the overall mean.

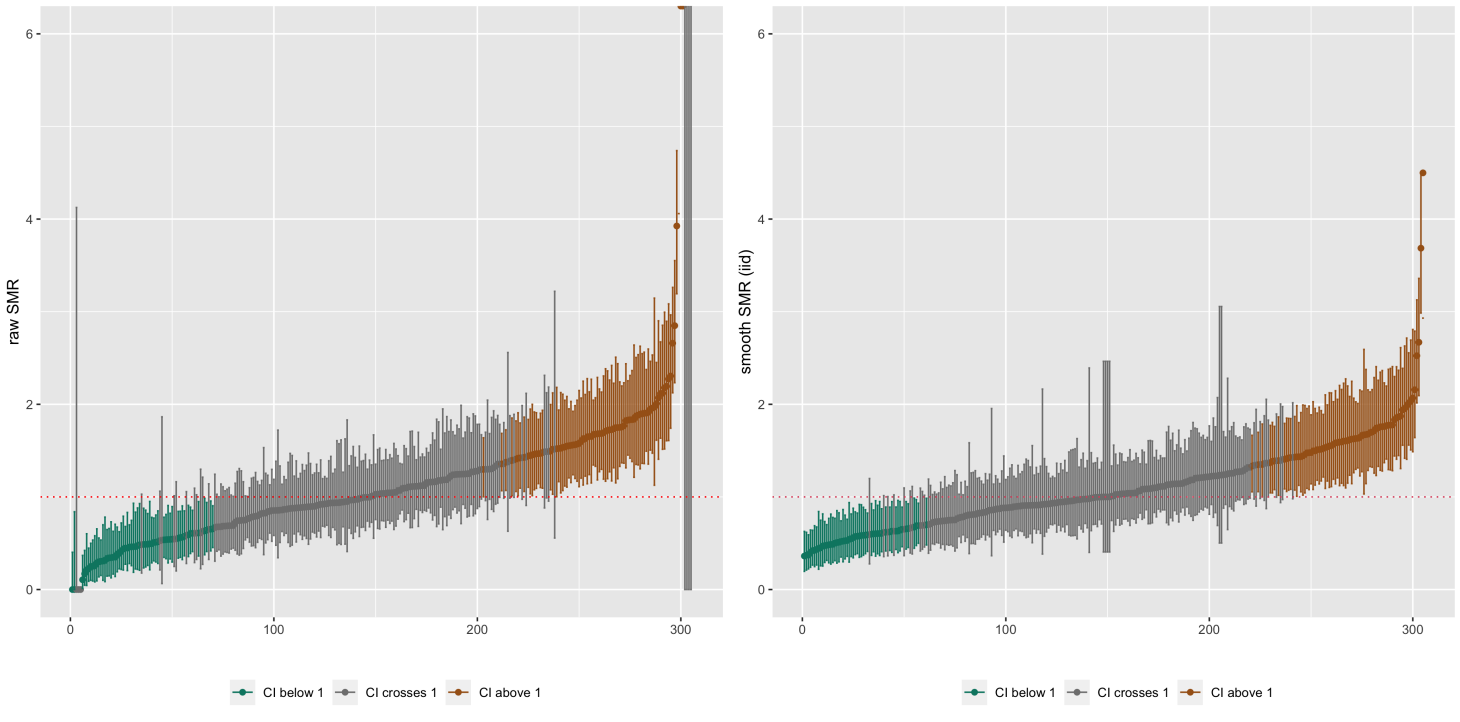


Figure 6.6: Caterpillar plots of raw SMRs with 95% CIs and smoothed SMRs from a Poisson lognormal model with 95% credible intervals

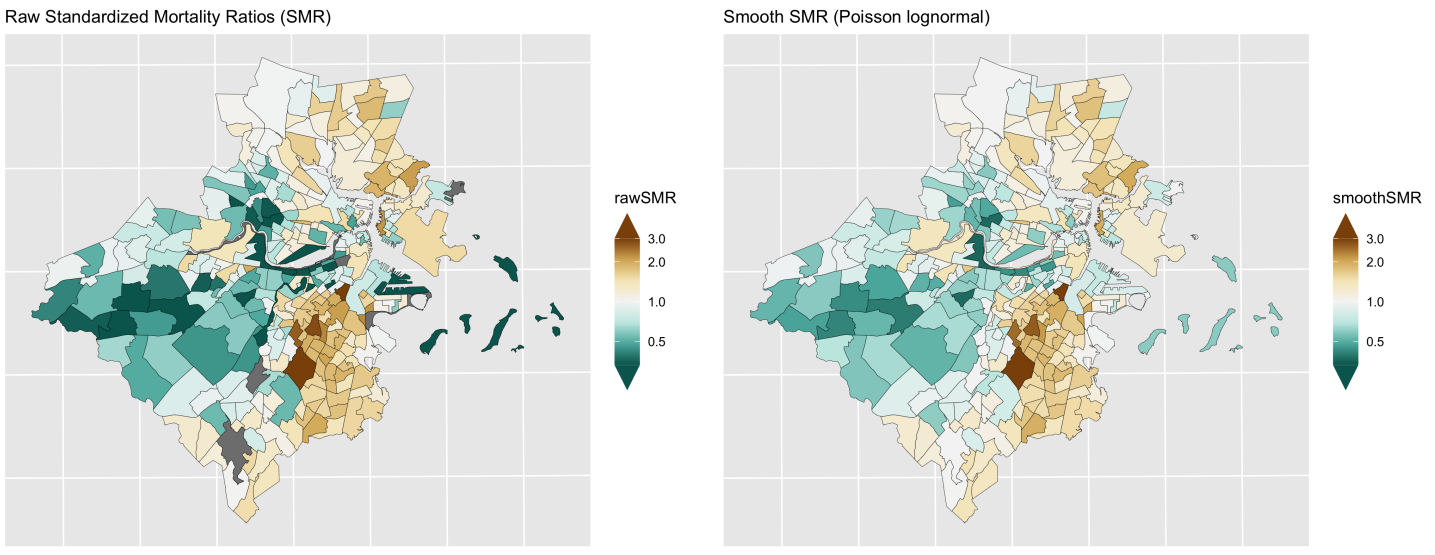


Figure 6.7: Maps of raw SMRs (left) and smoothed SMRs from a Poisson lognormal model

The estimate of the SMR for area i relative to the reference mortality rate (from the indirect age standardization) is $\exp(\alpha + v_i)$, while $\exp(v_i)$ is the residual relative risk for area i relative to the mean over the study area. Note that if an internal set of reference rates for indirect age standardization has been used, i.e. based on the observed rates over the whole study area itself, then α will generally be close to zero.

6.6.4 **Poisson multilevel model**

In both the Poisson gamma and Poisson log normal models, the smoothing is global: the random effects are treated as exchangeable and come from the same common distribution. As a result, all of the area specific relative risks are shrunk to the same overall mean. This kind of global smoothing does not allow for spatial correlation between risks in nearby areas, as might be expected if there is local clustering in the spatial pattern of risks. Local clustering of risks may be due to risk factors that are shared within and across census tracts: individuals who share these spatially varying risk factors would be expected to have similar outcomes. If such local clustering exists, this is a violation of the assumption of exchangeability. If these risk factors are known and observed, the most straightforward solution would be to include them as

covariates in the model. However, in most observational settings, one can rarely measure, or even know about, all the relevant risk factors. If one suspects that these risk factors vary by area, it will be necessary to consider ways to allow for local clustering in models for $\{\theta_i\}$.

In order to accommodate local smoothing, we can specify a Poisson multilevel model with random effects at the census tract and neighborhood levels. In the Greater Boston example, we will treat the neighborhoods of Boston as roughly equivalent as a level to the city/town designation in the communities outside of Boston that make up the greater Boston study area.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\theta_{ij} E_{ij}) \\
\log(\theta_{ij}) &= \alpha + v_{ij} + \eta_j \\
v_{ij} &\sim \text{Normal}(0, \sigma_v^2) \\
\eta_j &\sim \text{Normal}(0, \sigma_\eta^2)
\end{aligned}
$$

This can be in INLA using the following code:

```
model_form <- O ~ 1 + f(id_order, model="iid") + f(cityid, model="iid")
model_multi <- inla(model_form, family="poisson",
                data=df_indirect_ordered, E=E, # E points to the expected count field
                control.predictor=list(compute=TRUE), # computes transformed posterior marginals
                control.compute=list(dic=TRUE, waic=TRUE)) # computes DIC for model fit
summary(model_multi)
```

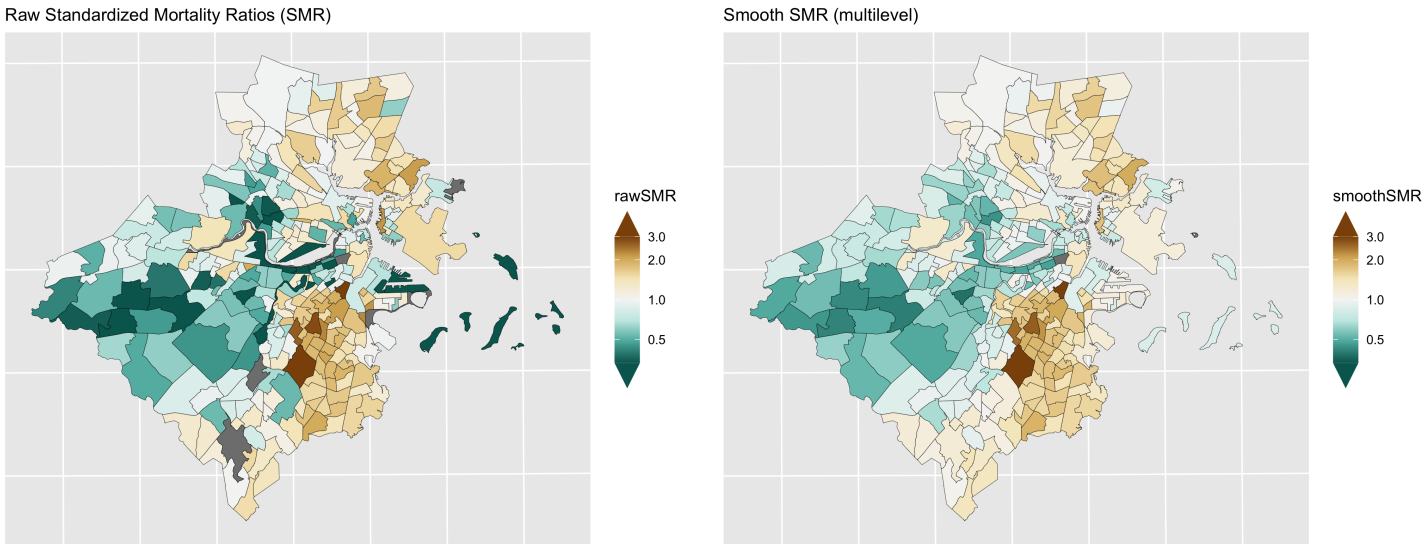


Figure 6.8: Maps of CT raw SMRs (left) and smoothed SMRs from a Poisson multilevel model

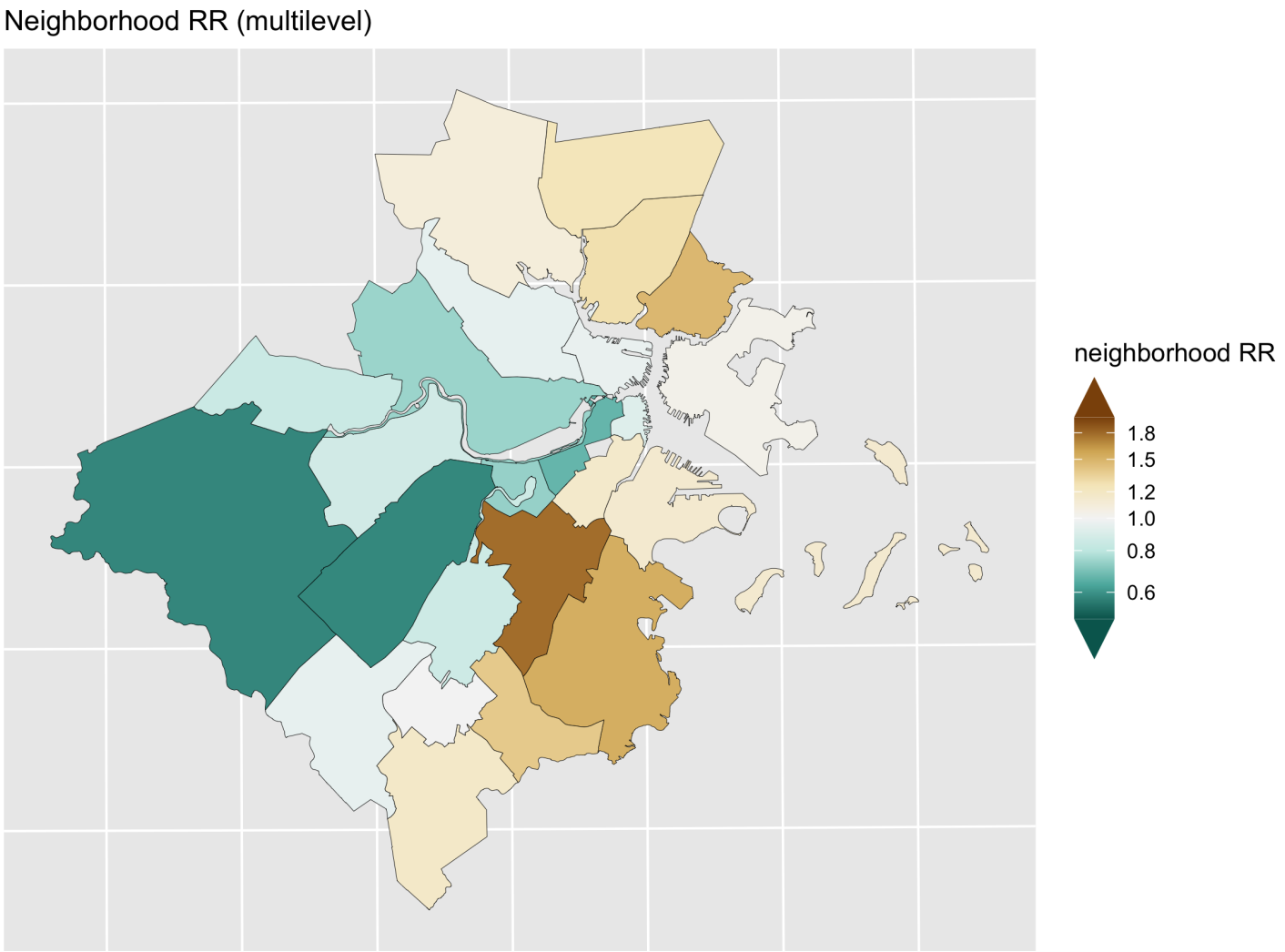


Figure 6.9: Map of the neighborhood level variation in relative risk (Level 2 of the Poisson multilevel model)

6.6.5 **Poisson BYM model**

In the multilevel Poisson log normal model with independent census tracts and neighborhoods, the neighborhood units were specified **a priori**. In a sense, the neighborhood boundaries are fixed, and crossing a boundary means that one's relative risk may jump by quite a lot. What if one wants to account

for spatial correlation in a **smoother** manner? One way to do this is to specify a multivariate normal prior distribution for all the area parameters with a spatially-structured covariance matrix. Many different ways have been proposed to specify spatially structured multivariate normal distributions for $\log(\theta_i)$ (Wakefield et al., 2000; Banerjee et al., 2004).

One particular form of the multivariate normal distribution that is commonly used is the **intrinsic Gaussian conditional autoregressive** (CAR) prior suggested by Clayton and Kaldor (1987) and developed by Besag et al. (1991). This is one of the most popular ways of dealing with spatial autocorrelation. The spatial structure is formulated through a set of conditional autoregressions, which uses the fact that if a vector of random variables has a multivariate normal distribution, then the distribution of each element of that vector conditional on all the other elements in the vector is also normal, with mean and variance that depend on the original multivariate mean and covariance matrix.

As with the other models, the first stage model assumes that the observed counts are Poisson distributed, and that an additive model for $\log(\theta_i)$ can be specified for accommodating covariate effects:

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \alpha + u_i \end{aligned}$$

Instead of the independent normal prior for the distribution of the census tract effects, one models

$$u_i | u_{j,j\neq i} \sim \text{Normal}(\mu_i, \tau_u^2/m_i)$$

where

$$\mu_i = \frac{\sum_j w_{ij} u_j}{\sum_j w_{ij}}, \qquad \sigma_i^2 = \frac{\tau_u^2}{\sum_j w_{ij}}$$

and

$$w_{ij} = \begin{cases} 1 & \text{if } i, j \text{ are adjacent,} \\ 0 & \text{if they are not} \end{cases}$$

and m_i is the number of adjacent areas. To understand this, consider the red highlighted census tract i in Figure 6.10. If ∂_i is the set of areas adjacent to i (shaded in yellow in the figure), and one sets w_{ij} to 1 for areas $j \in \partial_i$ and zero otherwise, then the prior distribution for u_i has conditional mean equal to the average of the neighbouring u_j's and variance inversely proportional to the number of adjacent neighbours. The effect is to smooth u_i toward the mean risk in the set of neighbouring areas. Note that τ_u^2 is the variance (scaled by $\sum_{j\in\partial_i} w_{ij} = m_i$, i.e. the number of neighbours); to emphasize that it is only interpretable conditionally, we have labeled it τ_u^2 instead of σ_u^2.

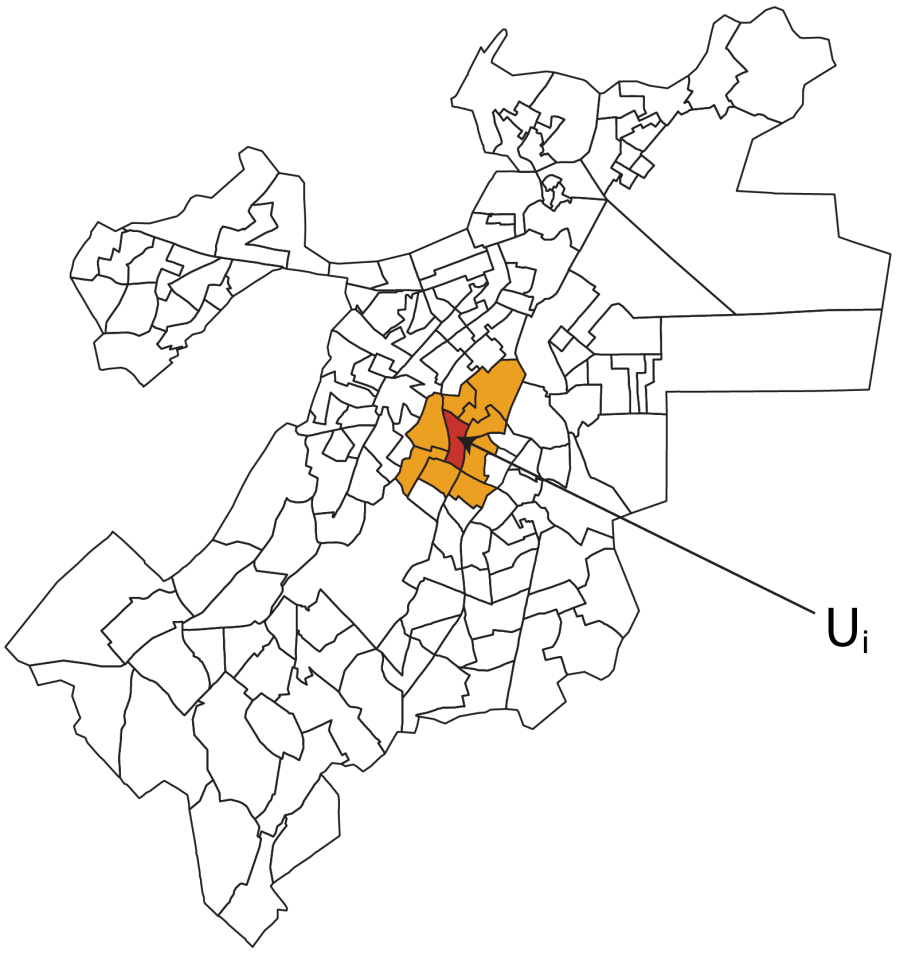


Figure 6.10: Illustration of the CAR idea

Besag, Yorke and Mollié (1991) recommend combining the CAR prior and the standard normal prior to allow for both **spatially unstructured** latent covariates **spatially correlated** latent covariates. Their model, called the BYM model or the convolution model in the literature, is thus

$$\begin{aligned} O_i &\sim \text{Poisson}(\theta_i E_i) \\ \log(\theta_i) &= \beta_0 + u_i + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \\ u_i | u_{j, j \neq i} &\sim \text{Normal}(\mu_i, \sigma_u^2 / m_i) \end{aligned}$$

Here, the unstructured v_i can be thought of as capturing correlation **within** areas and the spatially structured u_i capture spatial correlation **across** areas.

Note, however, that σ_v^2 (unstructured heterogeneity variance) and τ_u^2 (spatial variance) are not directly comparable: σ_v^2 reflects the variability of the unstructured random effects between areas, while τ_u^2 is the variance of the spatial effect in area i, conditional on values of neighboring spatial effects. No closed-form expression is available for the between-area variance of the spatial effects. However, in the Bayesian approach, the marginal spatial variance s_u^2 can be estimated empirically from the posterior samples of $\{u_i\}$.

$$s_u^2 = \sum_i (u_i - \bar{u})^2 / (n-1)$$

where $\bar{u}$ is the average of the $\{u_i\}$.

With an estimate of the marginal between-area variance, one can also characterize the relative contribution of spatial vs. unstructured heterogeneity:

$$\text{frac}_{\text{spatial}} = s_u^2 / (s_u^2 + \sigma_v^2)$$

When $\text{frac}_{\text{spatial}}$ is close to 1, spatial heterogeneity dominates. When $\text{frac}_{\text{spatial}}$ is close to 0, unstructured heterogeneity dominates (Best et al. 2005).

To aid with interpretation of the variability in the posterior estimates of $SMR_i = \exp(u_i + v_i)$, we define a quantity which we will call QR90 (Best et al. 2005)

$$QR_{90} = \exp(q_{95\%} - q_{5\%}) = \text{ relative risk between top and bottom } 5\backslash\% \text{ of areas}$$

where

$$\begin{aligned} q_{5\%} &= \text{ log relative risk for the area ranked at the 5th percentile} \\ q_{95\%} &= \text{ log relative risk for the area ranked at the 95th percentile} \end{aligned}$$

The advantage of QR90 is that it expresses the variability in the SMRs on the relative risk scale, which facilitates comparison of the disparities between areas to the estimated inequities (also on the relative risk scale) attributable to categories of ABSMs.

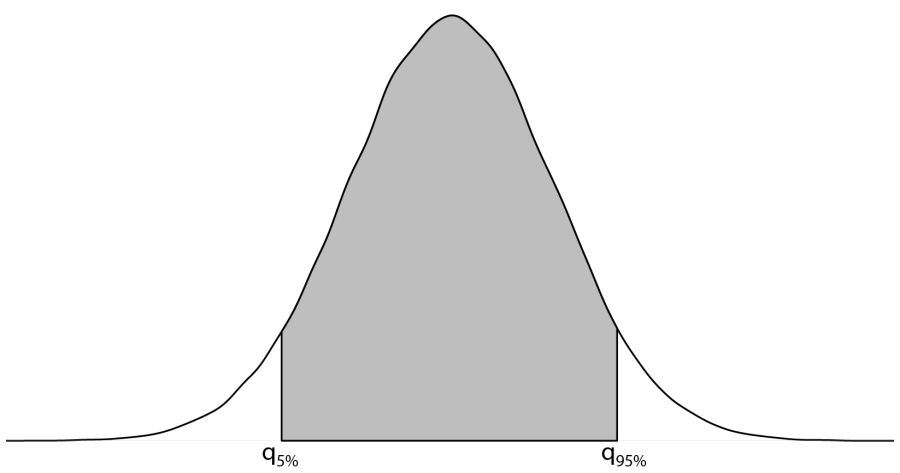


Figure 6.11: Depiction of QR90: the relative risk comparing the 95th and 5th quantiles of the random effects distribution

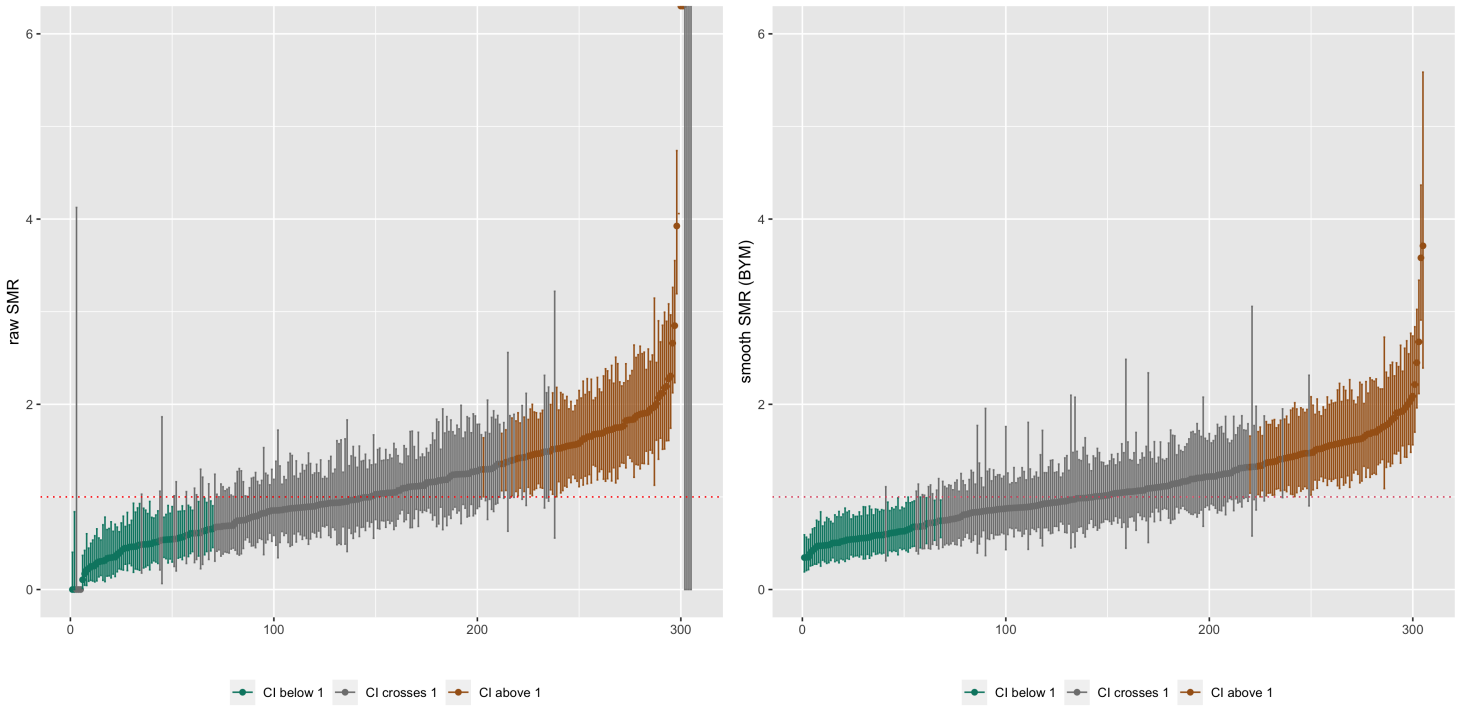


Figure 6.12: Caterpillar plots of raw SMRs with 95% CIs and smoothed SMRs from a Poisson BYM model with 95% credible intervals

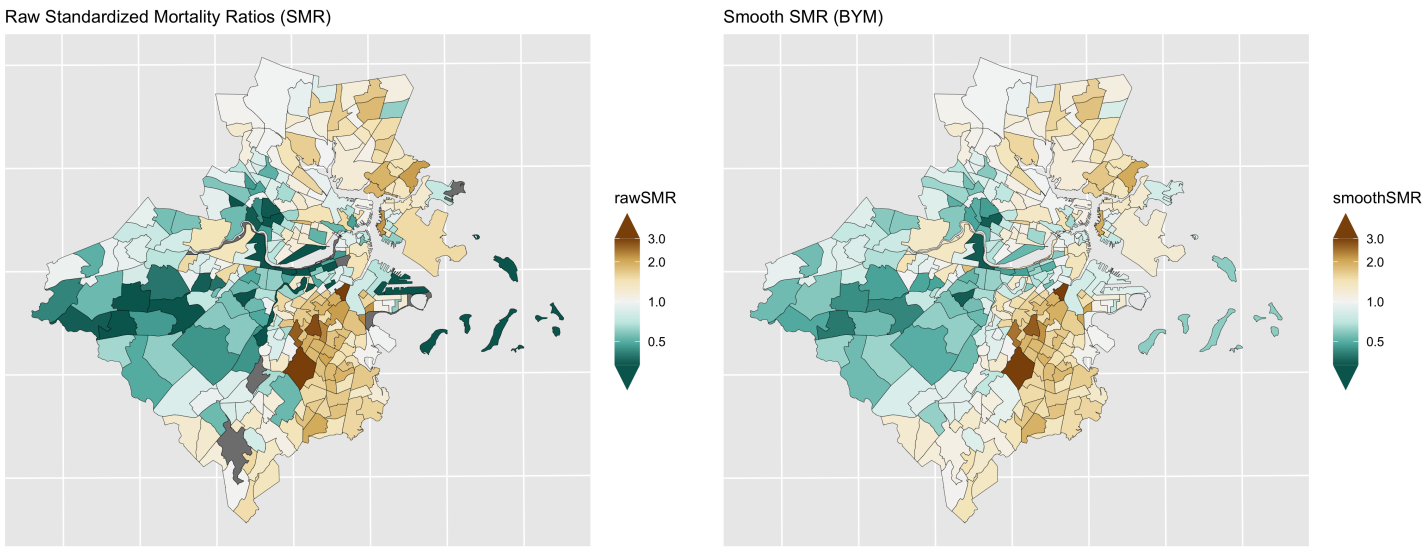


Figure 6.13: Comparison of premature mortality raw SMRs and smoothed (BYM) SMRs

6.7 **Estimating ABSM effects**

6.7.1 **Premature mortality**

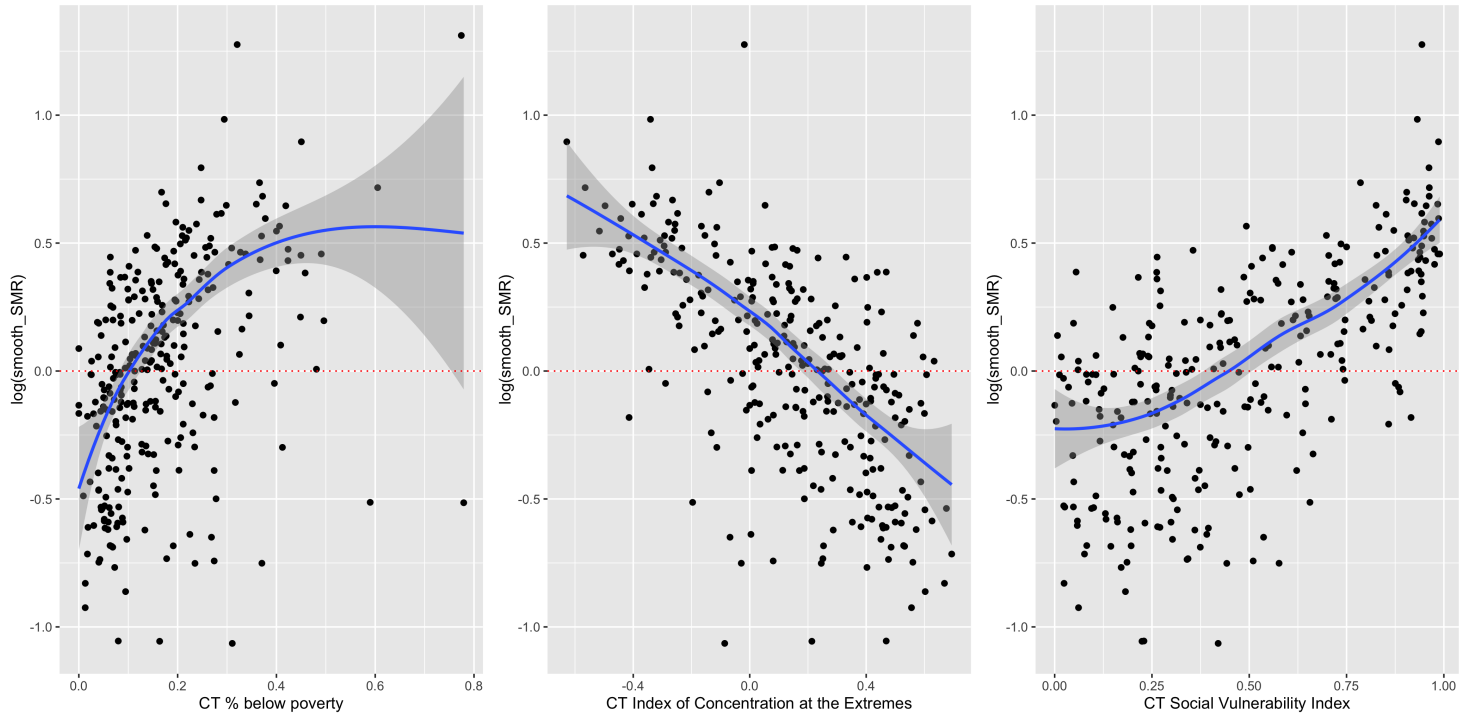


Figure 6.14: ABSMs vs. smoothed SMRs (BYM)

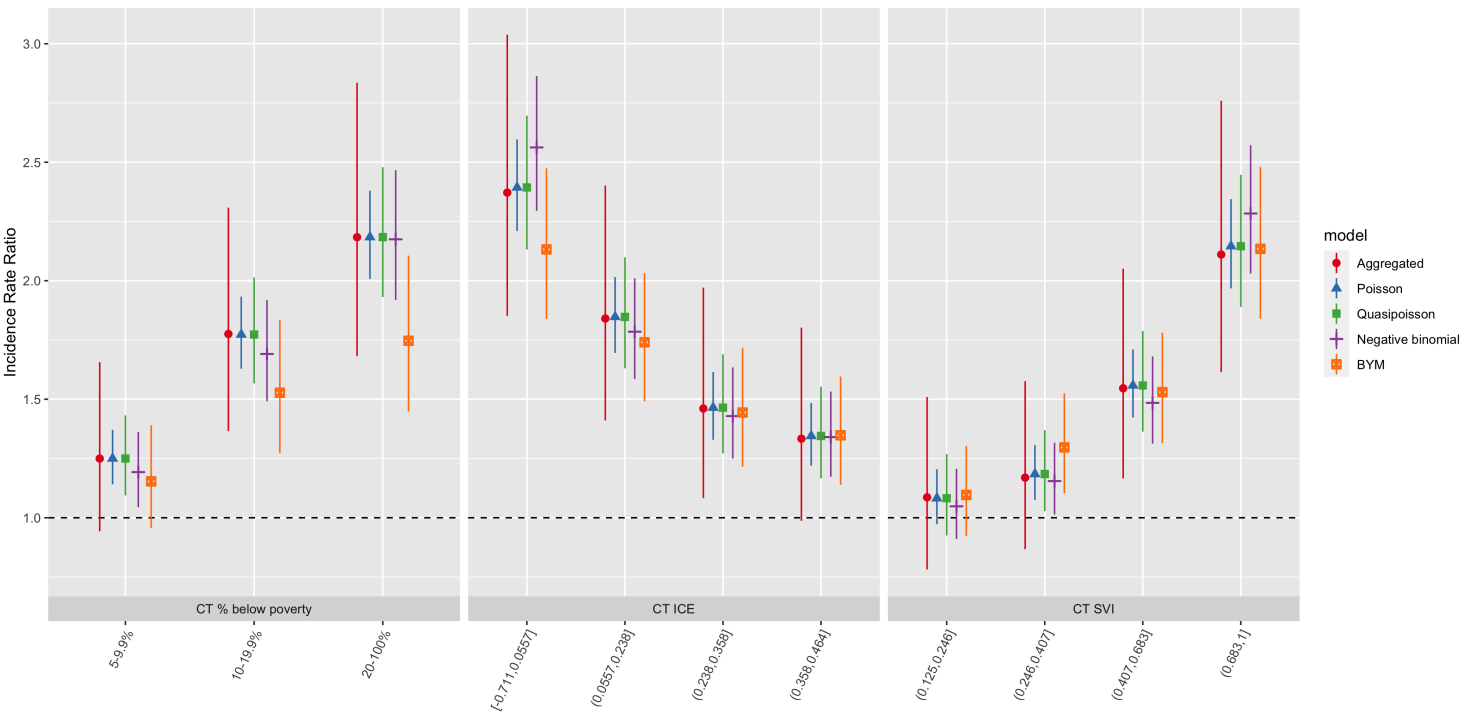


Figure 6.15: Comparison of premature mortality inequities by ABSM categories across all analytic methods and models

		Aggregated			Quasipoisson			
ABSM	**Category**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IF**
CT % below poverty								
apINDPOV	5-9.9%	1.25	0.94	1.66	1.25	1.09	1.43	1.
apINDPOV	10-19.9%	1.78	1.37	2.31	1.77	1.57	2.01	1.(
apINDPOV	20-100%	2.18	1.68	2.84	2.18	1.93	2.48	2.
CT Index of Concentration at the Extremes								
qICEwnhinc	[-0.711,0.0557]	2.37	1.85	3.04	2.39	2.13	2.70	2.
qICEwnhinc	(0.0557,0.238]	1.84	1.41	2.40	1.85	1.63	2.10	1.1

		Aggregated			Quasipoisson			
ABSM	**Category**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IR**
qICEwnhinc	(0.238,0.358]	1.46	1.08	1.97	1.46	1.27	1.69	1.4
qICEwnhinc	(0.358,0.464]	1.33	0.99	1.80	1.35	1.17	1.55	1.3
CT Social Vulnerability Index								
qsvi	(0.125,0.246]	1.09	0.78	1.51	1.08	0.92	1.27	1.0
qsvi	(0.246,0.407]	1.17	0.87	1.58	1.18	1.03	1.37	1.2
qsvi	(0.407,0.683]	1.55	1.17	2.05	1.56	1.36	1.79	1.4
qsvi	(0.683,1]	2.11	1.61	2.76	2.15	1.89	2.45	2.3

Table 6.1: Comparison of estimated inequities in premature mortality by category of CT ABSMs from aggregated analysi

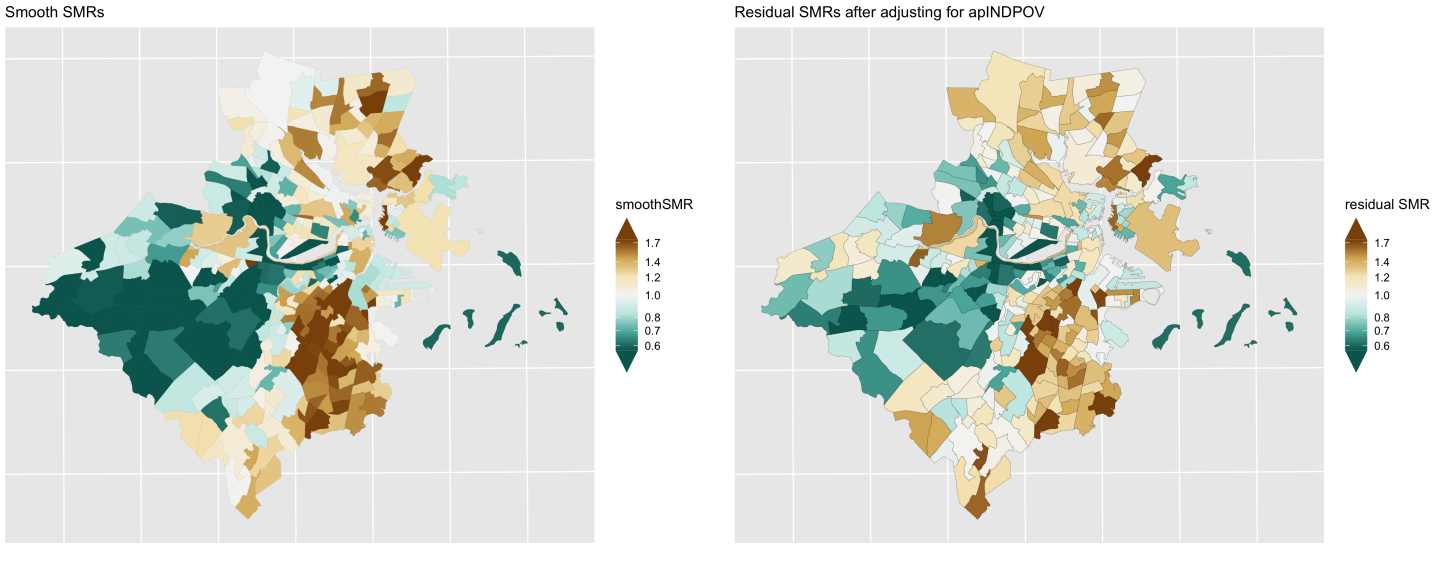


Figure 6.16: Premature mortality: comparison of smoothed SMRs (BYM) before and after adjustment for CT % below poverty

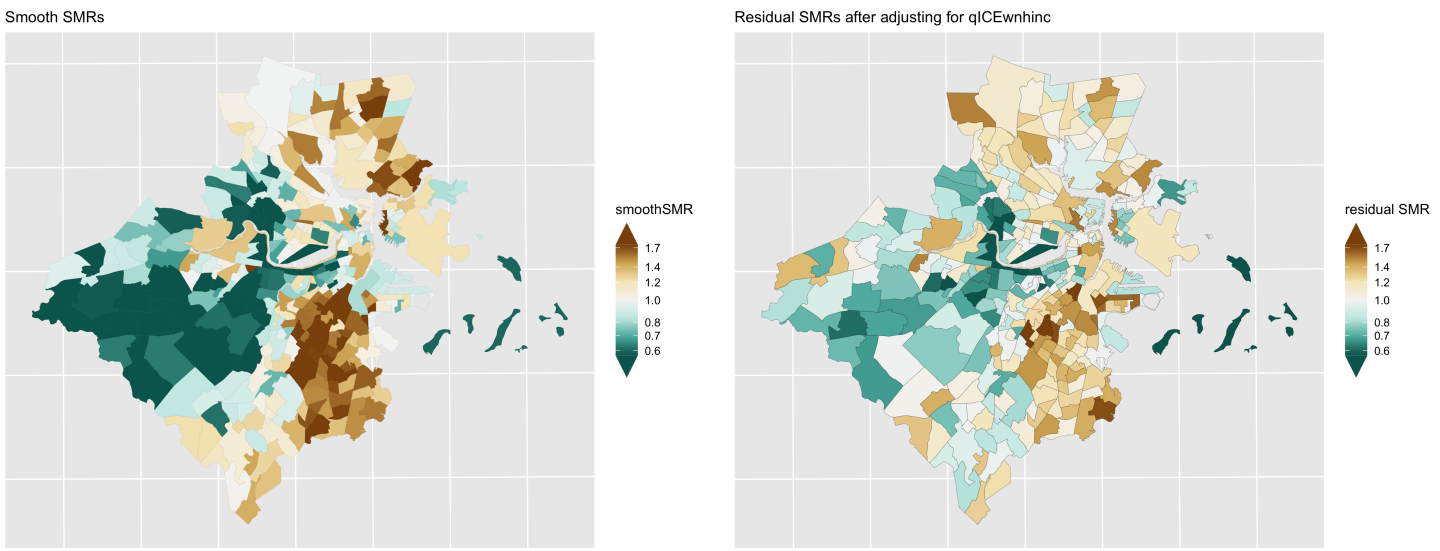


Figure 6.17: Premature mortality: comparison of smoothed SMRs (BYM) before and after adjustment for CT Index of Concentration at the Extremes (racialized economic segregation)

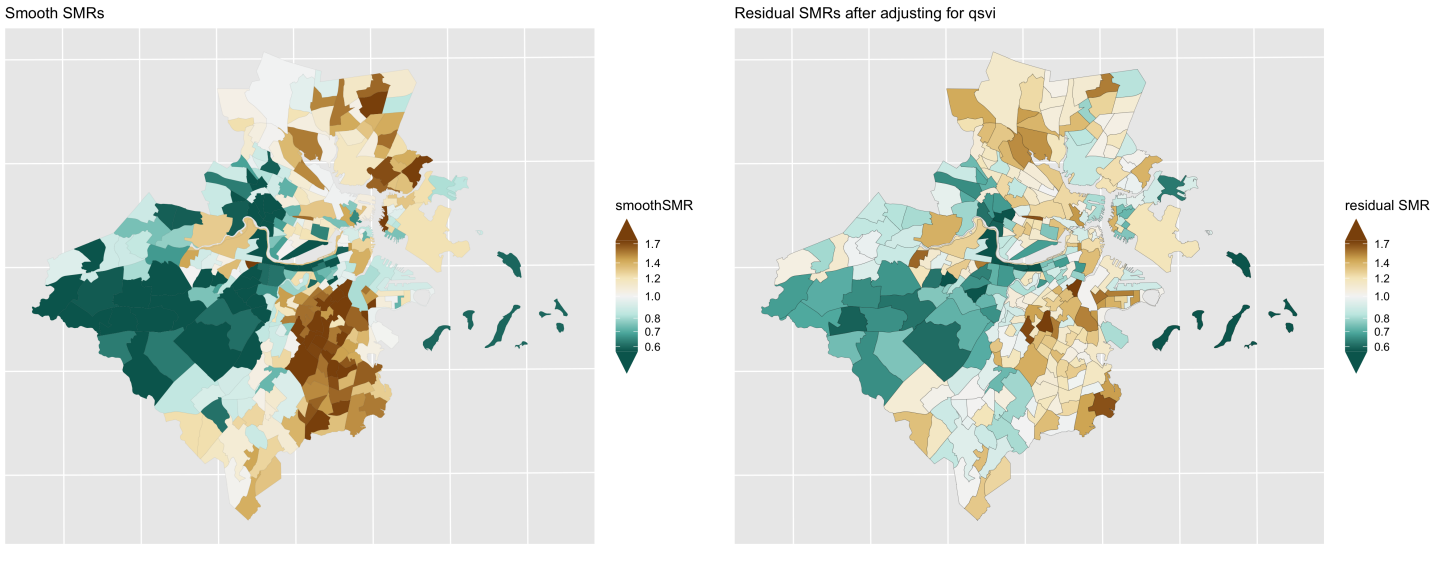


Figure 6.18: Premature mortality: comparison of smoothed SMRs (BYM) before and after adjustment for CT Social Vulnerability Index

6.7.2 **Lung cancer mortality**

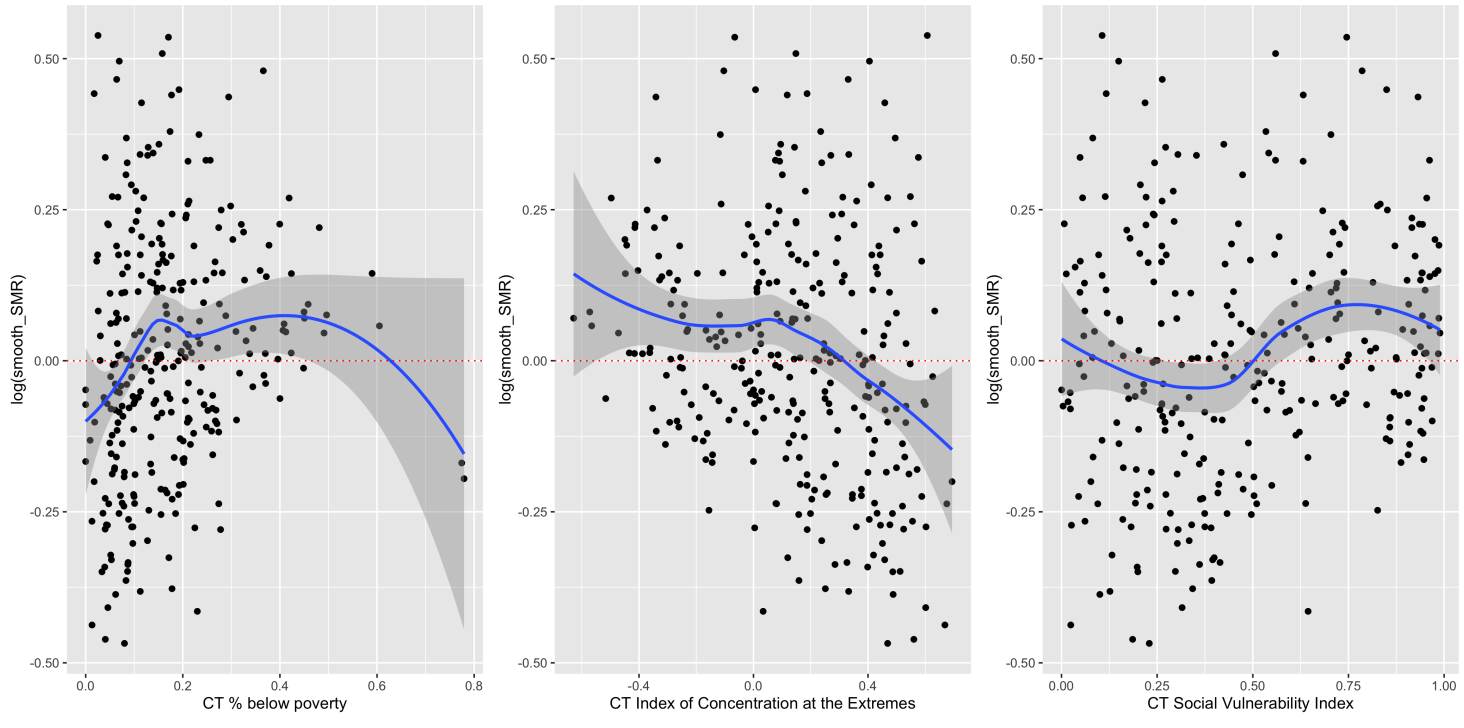


Figure 6.19: Lung cancer mortality: CT ABSMs vs. smoothed SMRs (BYM)

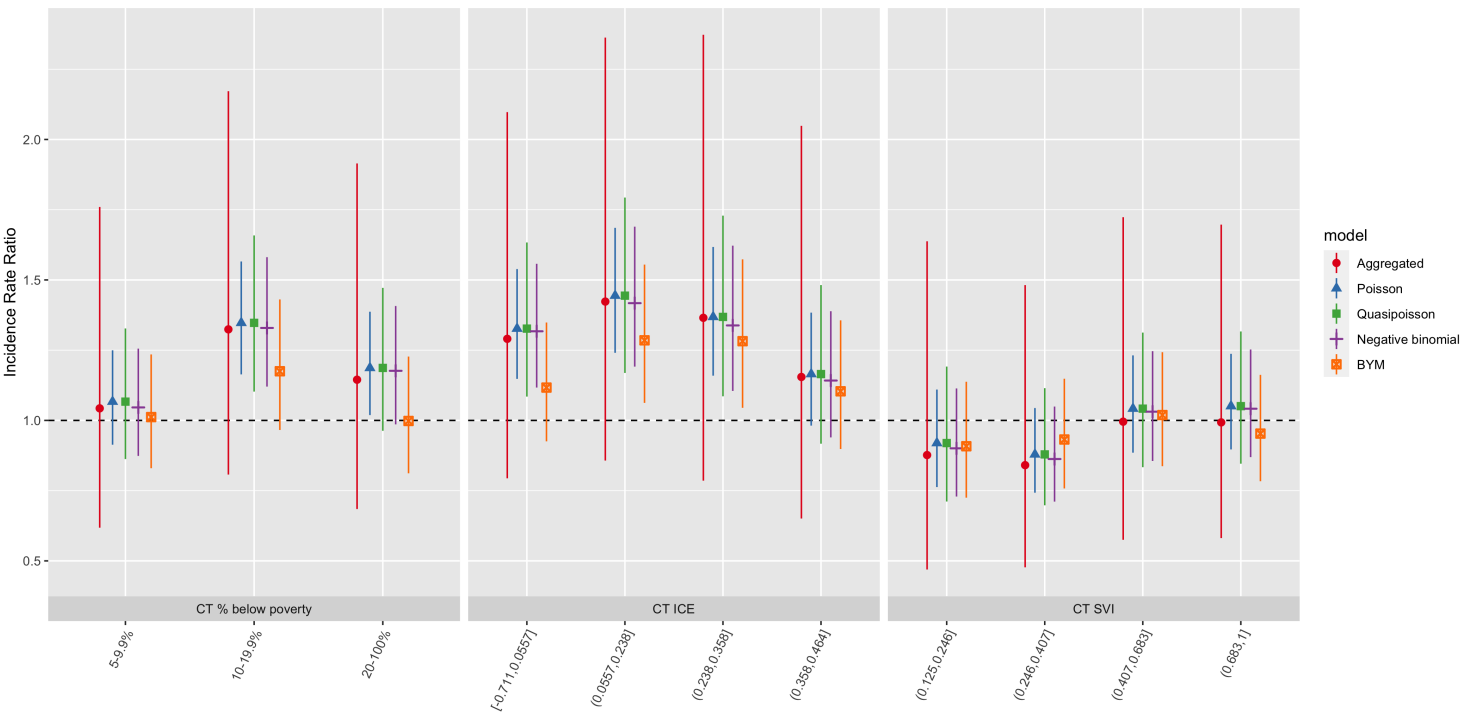


Figure 6.20: Comparison of lung cancer mortality inequities by CT ABSM categories across all analytic methods and models

		Aggregated			Quasipoisson			
ABSM	**Category**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IF**
CT % below poverty								
apINDPOV	5-9.9%	1.04	0.62	1.76	1.07	0.86	1.33	1.(
apINDPOV	10-19.9%	1.32	0.81	2.17	1.35	1.10	1.66	1.:
apINDPOV	20-100%	1.14	0.68	1.91	1.19	0.96	1.47	1.
CT Index of Concentration at the Extremes								
qICEwnhinc	[-0.711,0.0557]	1.29	0.79	2.10	1.33	1.08	1.63	1.:
qICEwnhinc	(0.0557,0.238]	1.42	0.86	2.36	1.44	1.17	1.79	1.4

		Aggregated			Quasipoisson			
ABSM	**Category**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IR**
qICEwnhinc	(0.238,0.358]	1.37	0.79	2.37	1.37	1.09	1.73	1.
qICEwnhinc	(0.358,0.464]	1.15	0.65	2.05	1.16	0.92	1.48	1.
CT Social Vulnerability Index								
qsvi	(0.125,0.246]	0.88	0.47	1.64	0.92	0.71	1.19	0.
qsvi	(0.246,0.407]	0.84	0.48	1.48	0.88	0.70	1.11	0.
qsvi	(0.407,0.683]	1.00	0.58	1.72	1.04	0.83	1.31	1.
qsvi	(0.683,1]	0.99	0.58	1.70	1.05	0.85	1.32	1.

Table 6.2: Comparison of estimated inequities in lung cancer mortality by category of CT ABSMs from aggregated analy

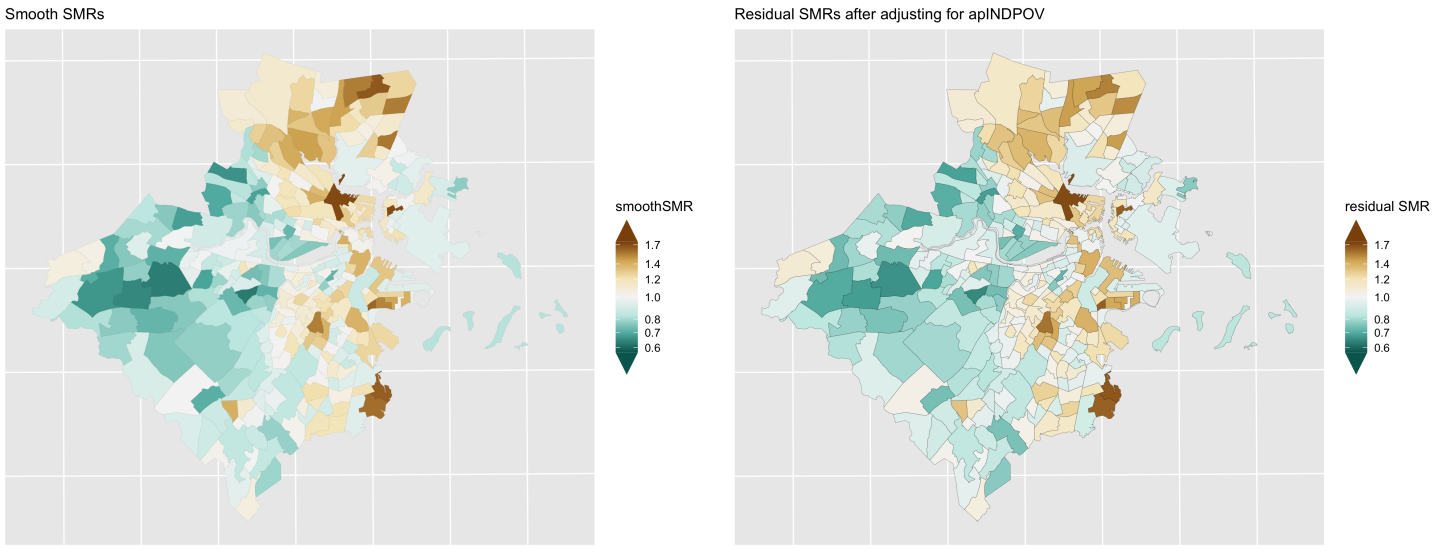


Figure 6.21: Lung cancer: comparison of smoothed SMRs (BYM) before and after adjustment for CT % below poverty

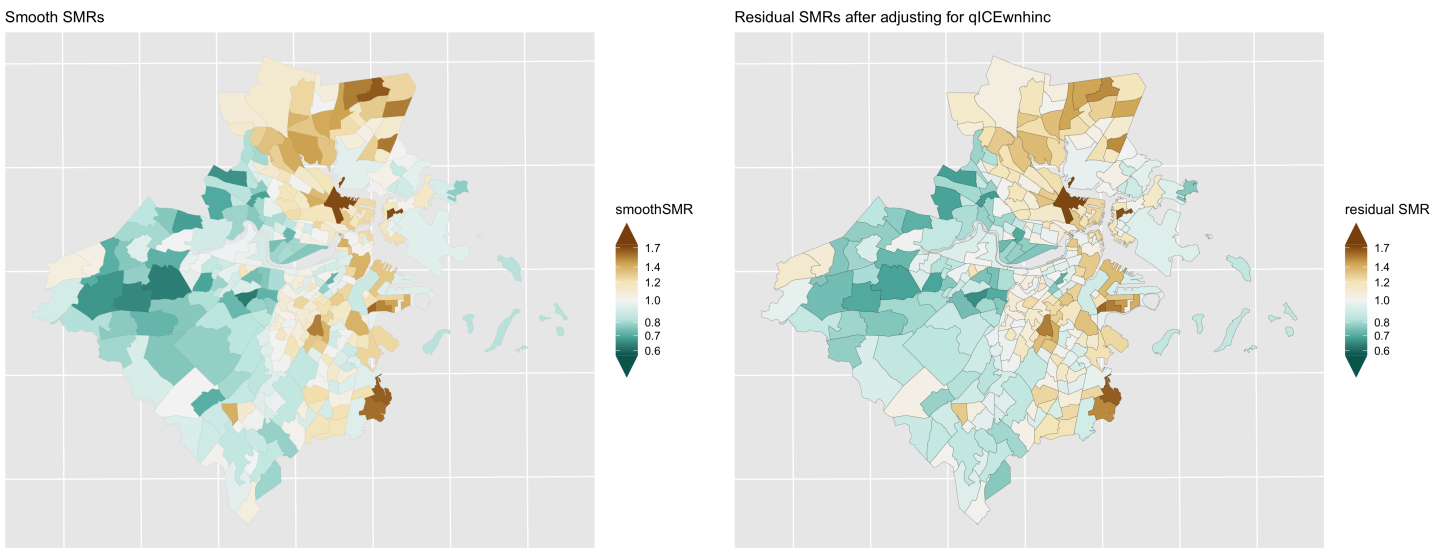


Figure 6.22: Lung cancer: comparison of smoothed SMRs (BYM) before and after adjustment for CT Index of Concentration at the Extremes (racialized economic segregation)

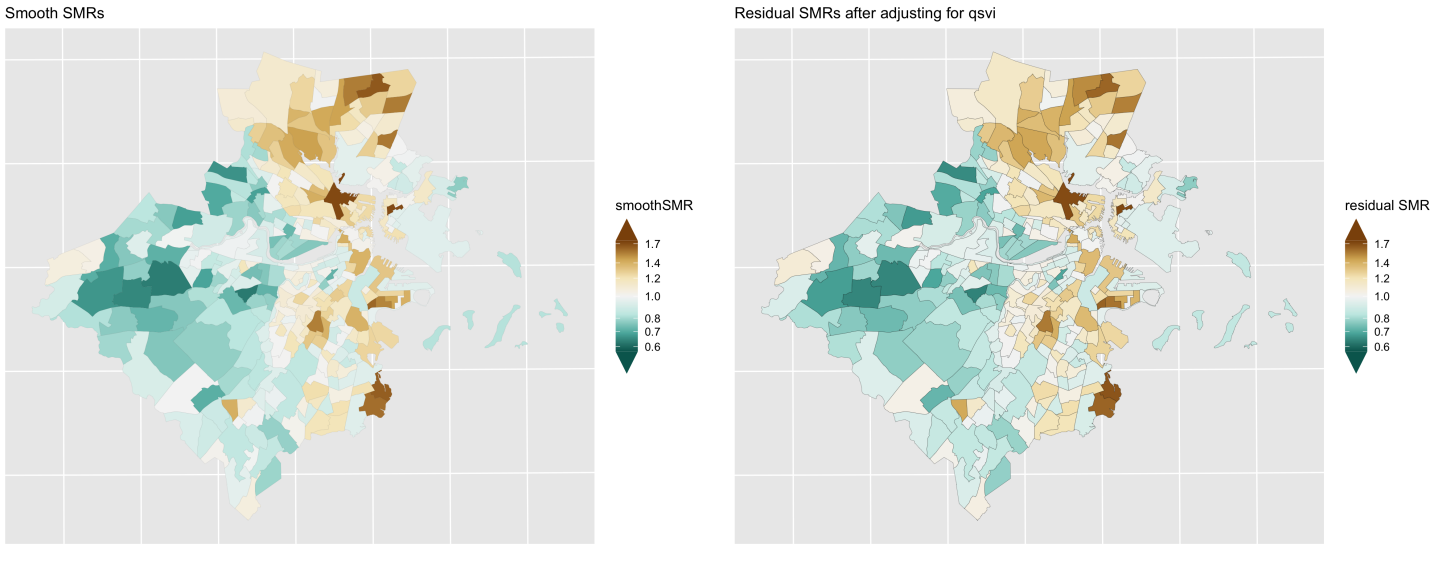


Figure 6.23: Lung cancer: comparison of smoothed SMRs (BYM) before and after adjustment for CT Social Vulnerability Index

6.8 Intersectional analysis of inequities by racialized group and CT ABSMs

In this section, we explore intersectional patterns of inequities in premature mortality by racialized group and CT ABSMs in the census tracts comprising the Greater Boston study area. That is, we focus on the joint patterning of inequities by membership in racialized groups (as captured by individual-level membership in racialized groups) and residence in census tracts characterized by area-based social metrics as indicative of **racialized societal inequities**. To accomplish this, we operationalize intersectional effects using methods for analyzing statistical interactions (Knol and VanderWeele 2012; Jackson et al. 2016; VanderWeele 2015). In contrast to analyses that focus on racial/ethnic disparities "controlling" for area-based social metrics or, conversely, disparities by ABSM "controlling" for racialized group membership, this framework emphasizes the importance of describing the joint patterning of racialized social inequities.

Drawing on recommendations on reporting interactions (Knol and VanderWeele 2012), we propose that health disparities researchers should be interested in reporting these interactions in three different ways:

- highlighting inequities by racialized group **within** categories of CT ABSMs
- highlighting inequities by CT ABSM **within** racialized groups
- highlighting the joint inequities by racialized group and CT ABSM relative to a *common reference group*

Moreover, a thorough reporting of interaction analyses should present measures of effect measure modification on both the additive and multiplicative scales.

6.8.1 Aggregated analysis

We begin by computing age-standardized premature mortality rates by racialized group and CT ABSMs. Similar to how we computed these age-standardized rates by CT ABSMs alone, we aggregate age-specific death counts and population person-time by racialized group across census tracts within categories of CT ABSM and apply direct age-standardization. We present theses estimated age-standardized premature mortality rates and 95% confidence intervals in Figure 6.24.

Several patterns are apparent from the visualization of the age-standardized rates. Firstly, gradients in inequities by CT ABSM are apparent for all three ABSMs, with premature mortality rates higher in more disadvantaged census tracts. This gradient is more pronounced among the non-Hispanic White population, with steeper gradients and tighter confidence limits. Among the Black population, inequities across CT ABSM categories are more shallow, with wider confidence limits in the more advantaged categories reflecting the smaller population sizes of Black individuals living in these census tracts.

Focusing on the Black/White inequities across CT ABSM categories, there is some suggestion of qualitative interaction. That is, Black/White inequities are largest in the most advantaged census tracts, but in more disadvantaged census tracts, non-Hispanic Whites have higher rates of premature mortality than Black populations living in the same areas, particular by quintile of CT ICE and SVI.

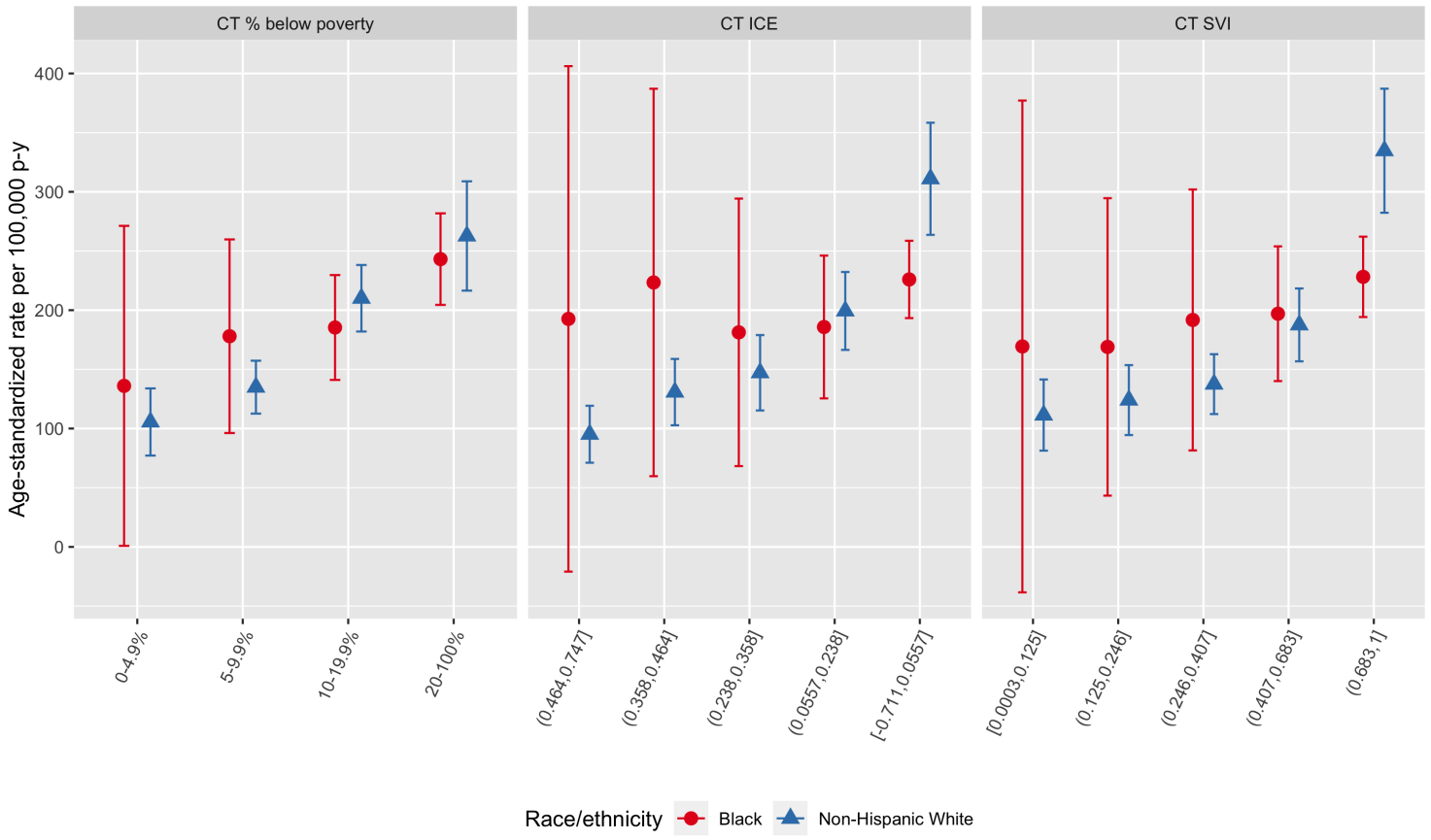


Figure 6.24: Age-standardized premature mortality rates by racialized group and CT ABSM, Greater Boston, 2013-2017, as computed by the aggregated method

To evaluate the magnitude of the CT ABSM inequities **within** racialized group, we present (a) age-standardized rate differences per 100,000 person-years and (b) age-standardized incidence rate ratios in Figure 6.25. On both the additive (incidence rate difference) and multiplicative scales (incidence rate ratio), ABSM gradients are much steeper among the non-Hispanic White population compared with the Black population. ABSM gradients among the Black population are attenuated towards the null and have much wider confidence limits.

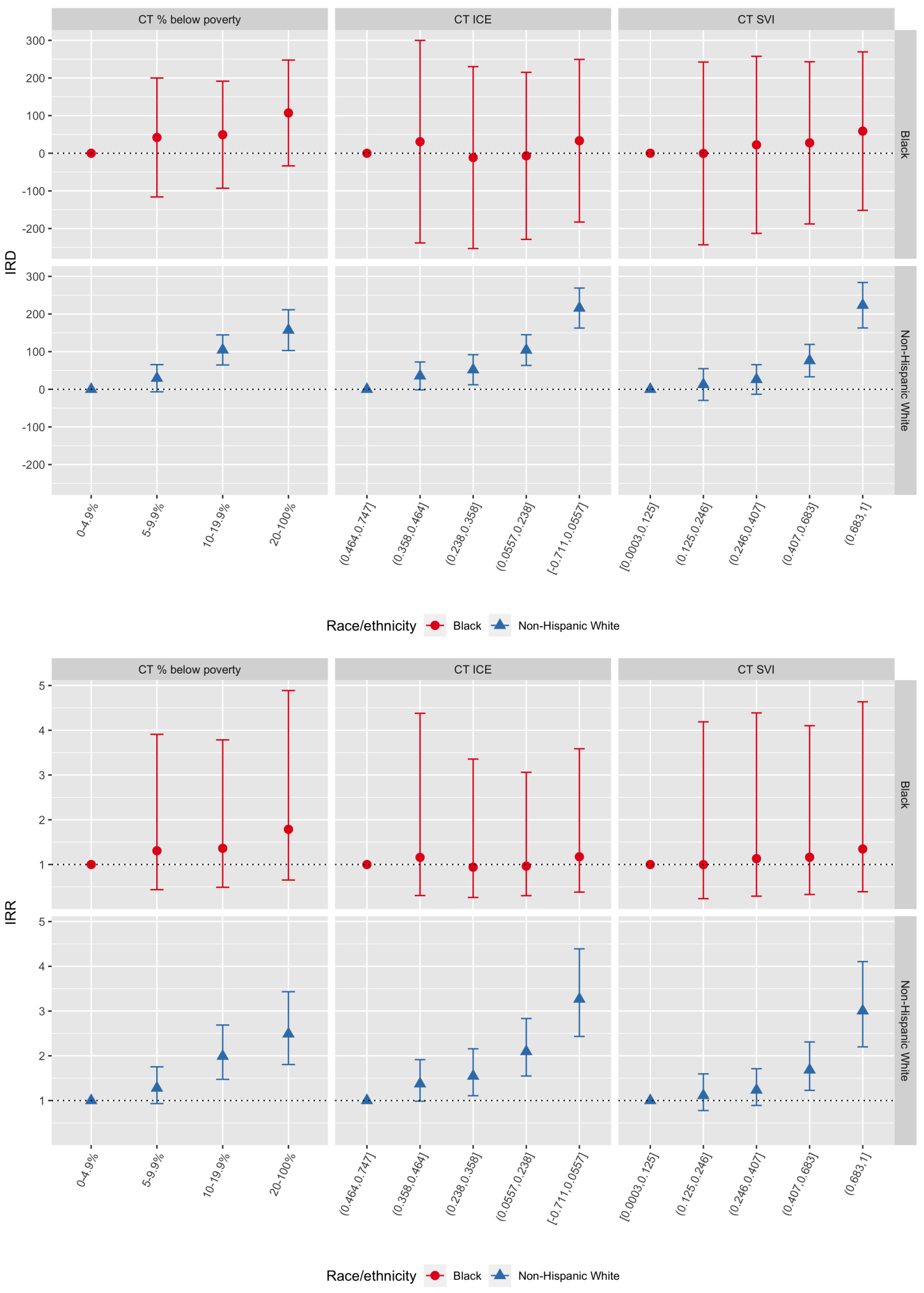


Figure 6.25: ABSM gradients in premature mortality by racialized group (Blacks vs. Non-Hispanic Whites): age-standardized rate differences per 100,000 person-years (top row) and rate ratios (bottom row) computed using the aggregation method, Greater Boston, 2013-2017. In each plot, the reference group is the most advantaged racialized group-specific ABSM category.

Meanwhile, Black/White inequities as depicted in Figure 6.26 confirm the pattern of Black/White crossover suggested in Figure 6.24, with non-Hispanic White premature mortality rates significantly elevated in the most disadvantaged census tracts by CT Index of Concentration at the Extremes and CT Social Vulnerability Index.

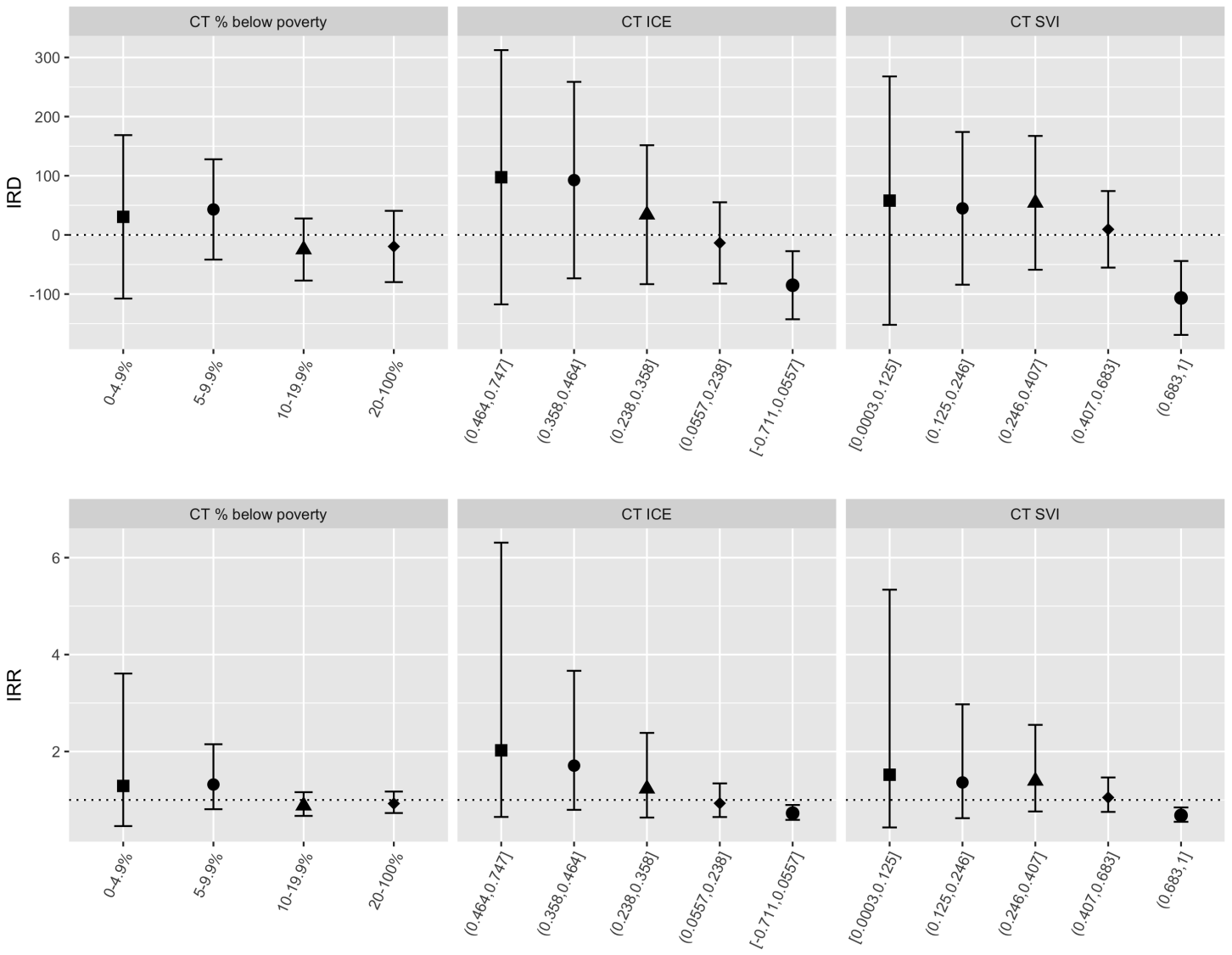


Figure 6.26: Black/White inequities in premature mortality by category of CT ABSM: age-standardized rate differences per 100,000 person-years (top row) and rate ratios (bottom row) computed using the aggregation method, Greater Boston, 2013-2017.

We visualize the joint pattern of inequities on the additive (incidence rate difference) and multiplicative (incidence rate ratio) scales relative to a common reference group in Figure 6.27. In general, it gives insights similar to visualization of age-standardized incidence rates by racialized group and CT ABSM in Figure 6.24 above.

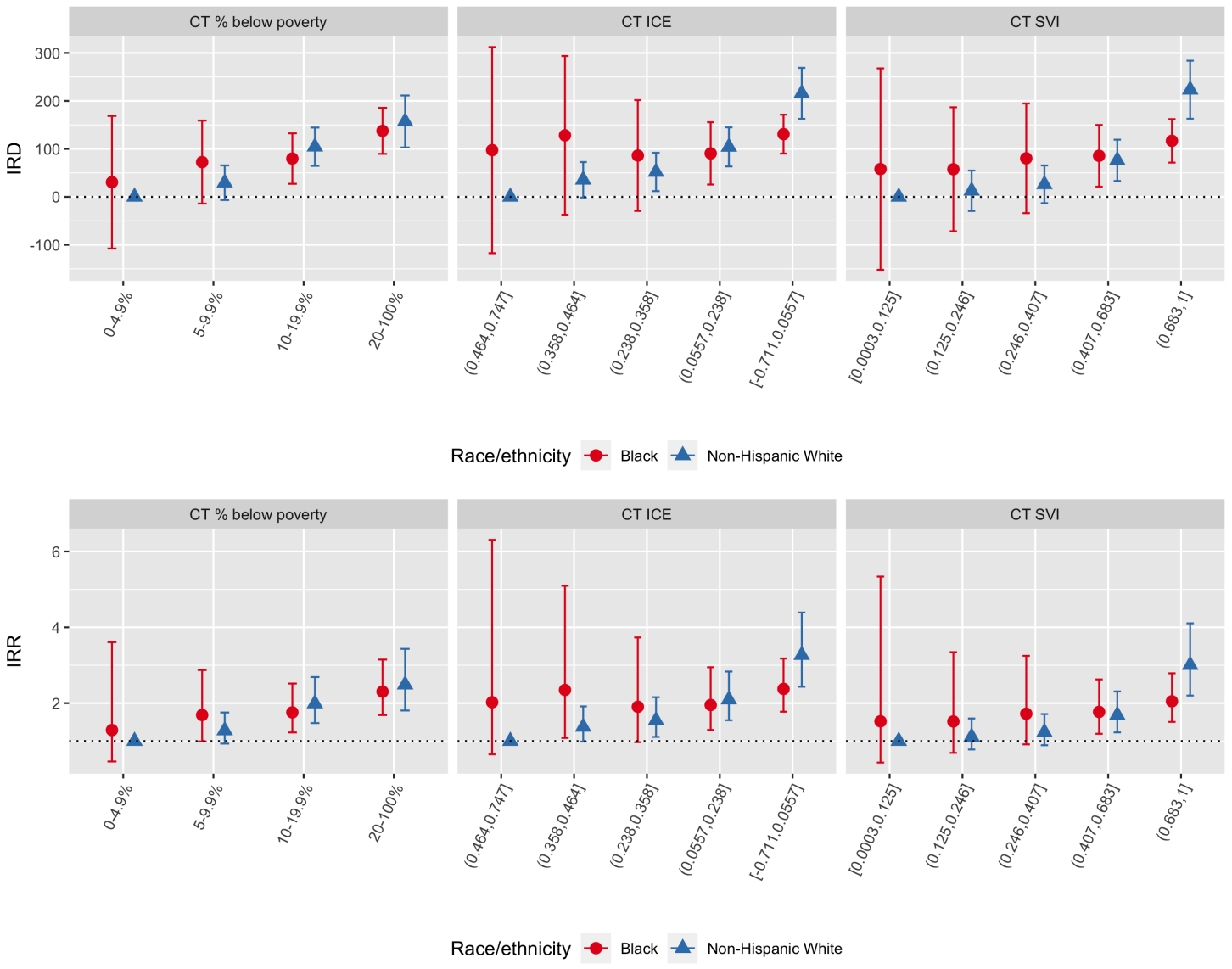


Figure 6.27: Intersectional age-standardized incidence rate differences per 100,000 person-years (top row) and age-standardized rate ratios (bottom row) by racialized group and category of CT ABSM computed using the aggregation method, Greater Boston, 2013-2017. Presented with a common reference group (non-Hispanic Whites in the most advantaged categories of CT ABSMs

ABSM gradients among the Black population are attenuated towards the null and have much wider confidence limits. Why do you think this might be the case?

6.8.2 **Intersectional inequities as estimated by non-spatial, multilevel, and spatial regression models**

Intersectional inequities by racialized group and CT ABSM can also be estimated using regression models.

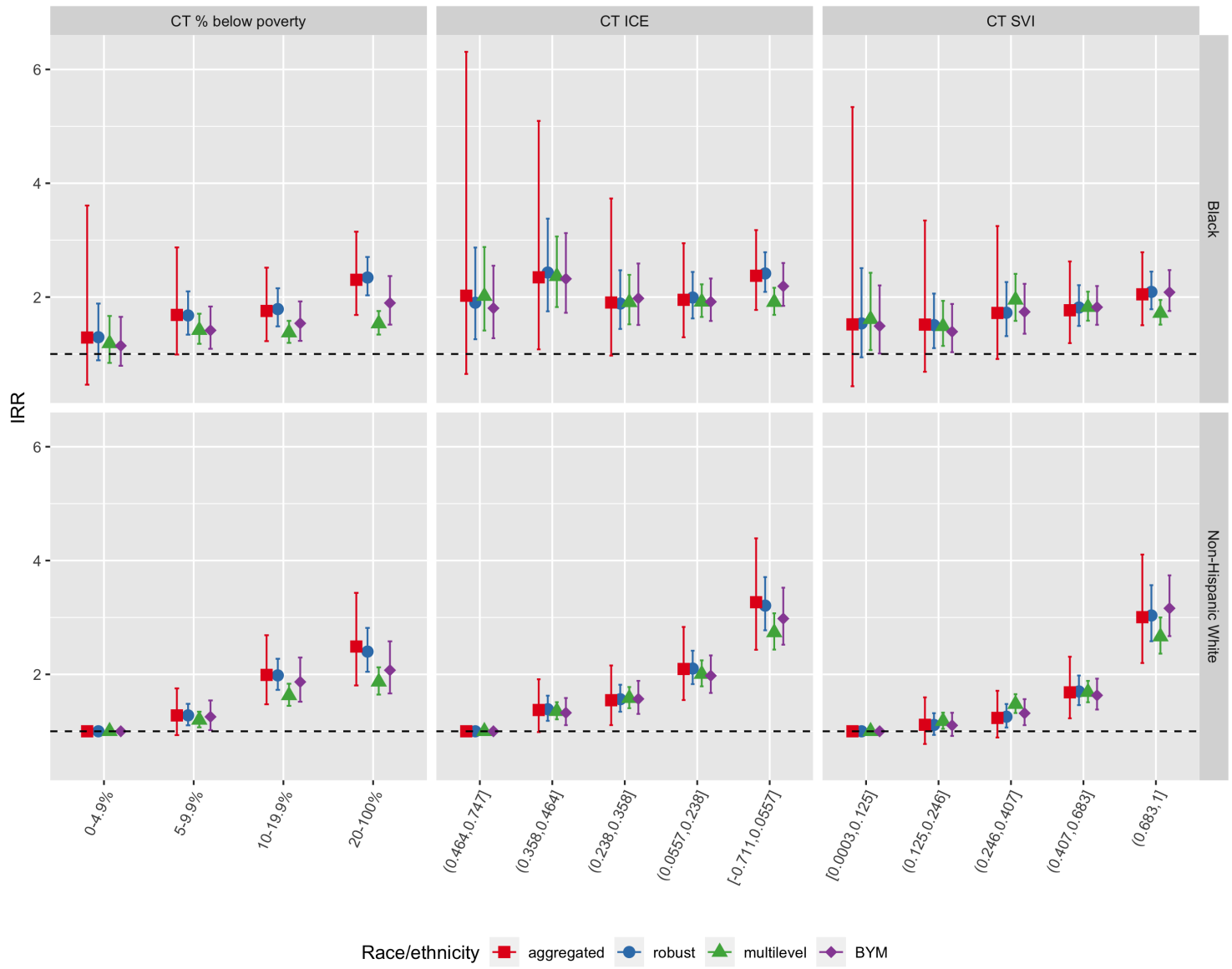


Figure 6.28: Comparison of inequities in premature mortality by racialized group and CT ABSM as estimated by the aggregated method, Poisson regression with robust variance estimator, multilevel Poisson regression (census tracts in city/town/neighborhoods), and spatial Poisson regression with BYM prior, Greater Boston, 2013-2017. Here, we focus on estimated incidence rate ratios with a common reference group (non-Hispanic Whites in the most advantaged census tracts)

		Aggregated			Poisson robust			
Racialized group	**ABSM**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IRR**
CT % below poverty								
NHW	0-4.9%	1.00			1.00			1.00
NHW	5-9.9%	1.28	0.93	1.75	1.28	1.10	1.48	1.20
NHW	10-19.9%	1.99	1.47	2.69	1.98	1.73	2.27	1.63
NHW	20-100%	2.49	1.81	3.43	2.40	2.05	2.82	1.87
Black	0-4.9%	1.29	0.46	3.61	1.29	0.89	1.89	1.19
Black	5-9.9%	1.69	0.99	2.87	1.68	1.34	2.10	1.42
Black	10-19.9%	1.76	1.23	2.52	1.79	1.49	2.15	1.38
Black	20-100%	2.30	1.69	3.15	2.34	2.03	2.70	1.53
CT Index of Concentration at the Extremes								
NHW	(0.464,0.747]	1.00			1.00			1.00

		Aggregated			Poisson robust			
Racialized group	**ABSM**	**IRR**	**conf.low**	**conf.high**	**IRR**	**conf.low**	**conf.high**	**IRR**
NHW	(0.358,0.464]	1.37	0.99	1.91	1.39	1.18	1.63	1.35
NHW	(0.238,0.358]	1.55	1.11	2.16	1.56	1.34	1.82	1.58
NHW	(0.0557,0.238]	2.09	1.55	2.83	2.10	1.83	2.42	2.01
NHW	[-0.711,0.0557]	3.27	2.43	4.39	3.21	2.77	3.71	2.74
Black	(0.464,0.747]	2.02	0.65	6.31	1.90	1.26	2.87	2.02
Black	(0.358,0.464]	2.35	1.08	5.10	2.43	1.75	3.38	2.36
Black	(0.238,0.358]	1.90	0.97	3.73	1.89	1.44	2.47	1.91
Black	(0.0557,0.238]	1.95	1.29	2.95	1.99	1.62	2.44	1.92
Black	[-0.711,0.0557]	2.37	1.77	3.18	2.41	2.09	2.79	1.91
CT Social Vulnerability Index								
NHW	[0.0003,0.125]	1.00			1.00			1.00
NHW	(0.125,0.246]	1.11	0.78	1.60	1.11	0.94	1.32	1.18
NHW	(0.246,0.407]	1.23	0.89	1.71	1.26	1.07	1.48	1.48
NHW	(0.407,0.683]	1.68	1.23	2.31	1.70	1.46	1.98	1.69
NHW	(0.683,1]	3.00	2.20	4.11	3.03	2.58	3.57	2.66
Black	[0.0003,0.125]	1.52	0.43	5.34	1.54	0.94	2.51	1.61
Black	(0.125,0.246]	1.52	0.69	3.35	1.51	1.10	2.06	1.49
Black	(0.246,0.407]	1.72	0.91	3.25	1.73	1.31	2.26	1.95
Black	(0.407,0.683]	1.77	1.19	2.63	1.82	1.49	2.21	1.82
Black	(0.683,1]	2.05	1.51	2.79	2.09	1.79	2.45	1.72

Table 6.3: Comparison of intersectional inequities in premature mortality by racialized group and category of CT ABSMs multilevel, and BYM models. Note the common reference group (non-Hispanic Whites in the most advantaged CTs)

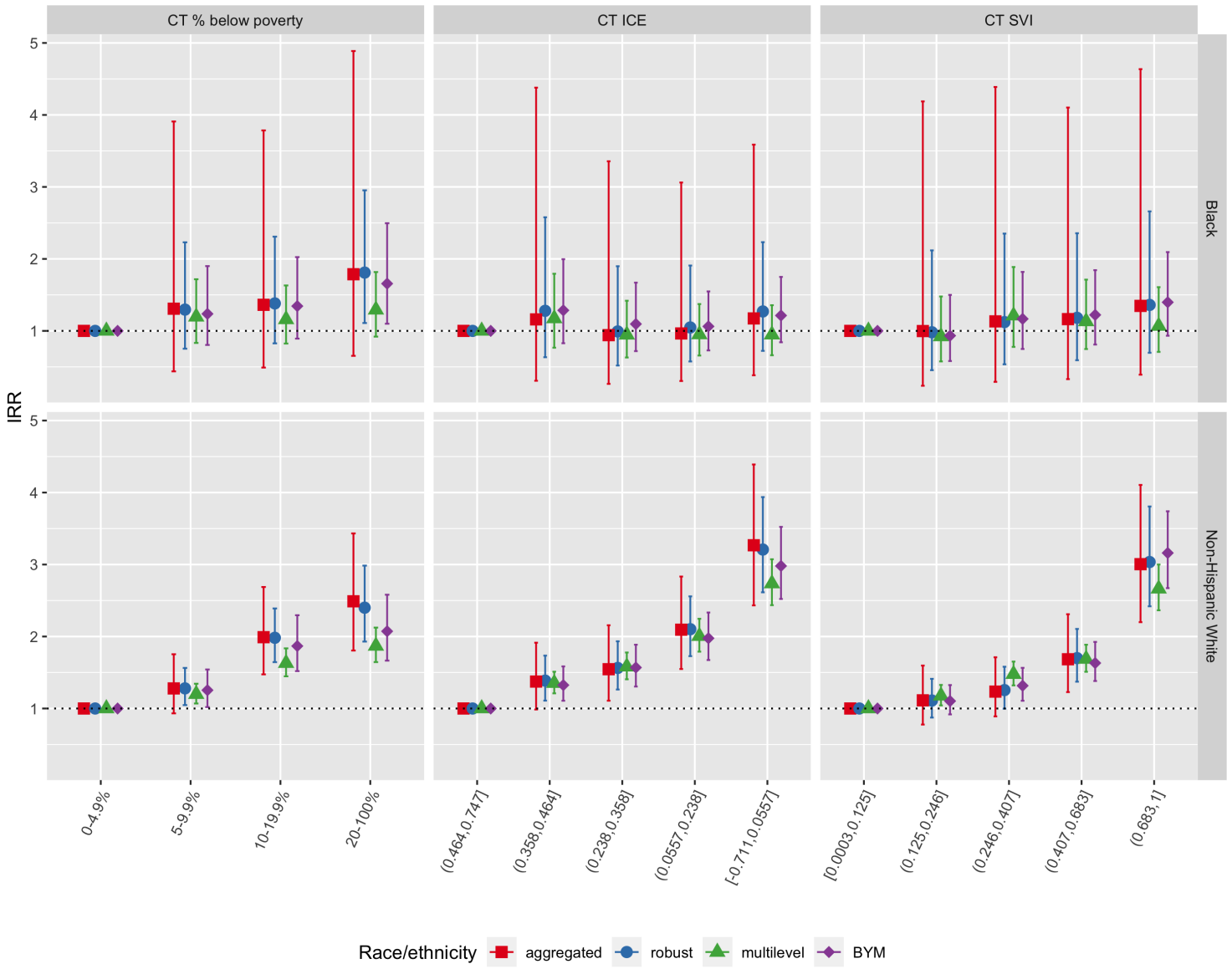


Figure 6.29: Comparison of CT ABSM inequities in premature mortality within racialized group (Black and Non-Hispanic White) as estimated by the aggregated method, Poisson regression with robust variance estimator, multilevel Poisson regression (census tracts in city/town/neighborhoods), and spatial Poisson regression with BYM prior, Greater Boston, 2013-2017. In each plot, the reference group is the most advantaged ABSM category **within** racialized group.

As seen in Figure 6.29, inequities in premature mortality across CT ABSM categories are much more pronounced for the non-Hispanic White population, whereas the gradient in age-adjusted incidence rate ratios for CT ABSM categories within the Black population is much shallower.

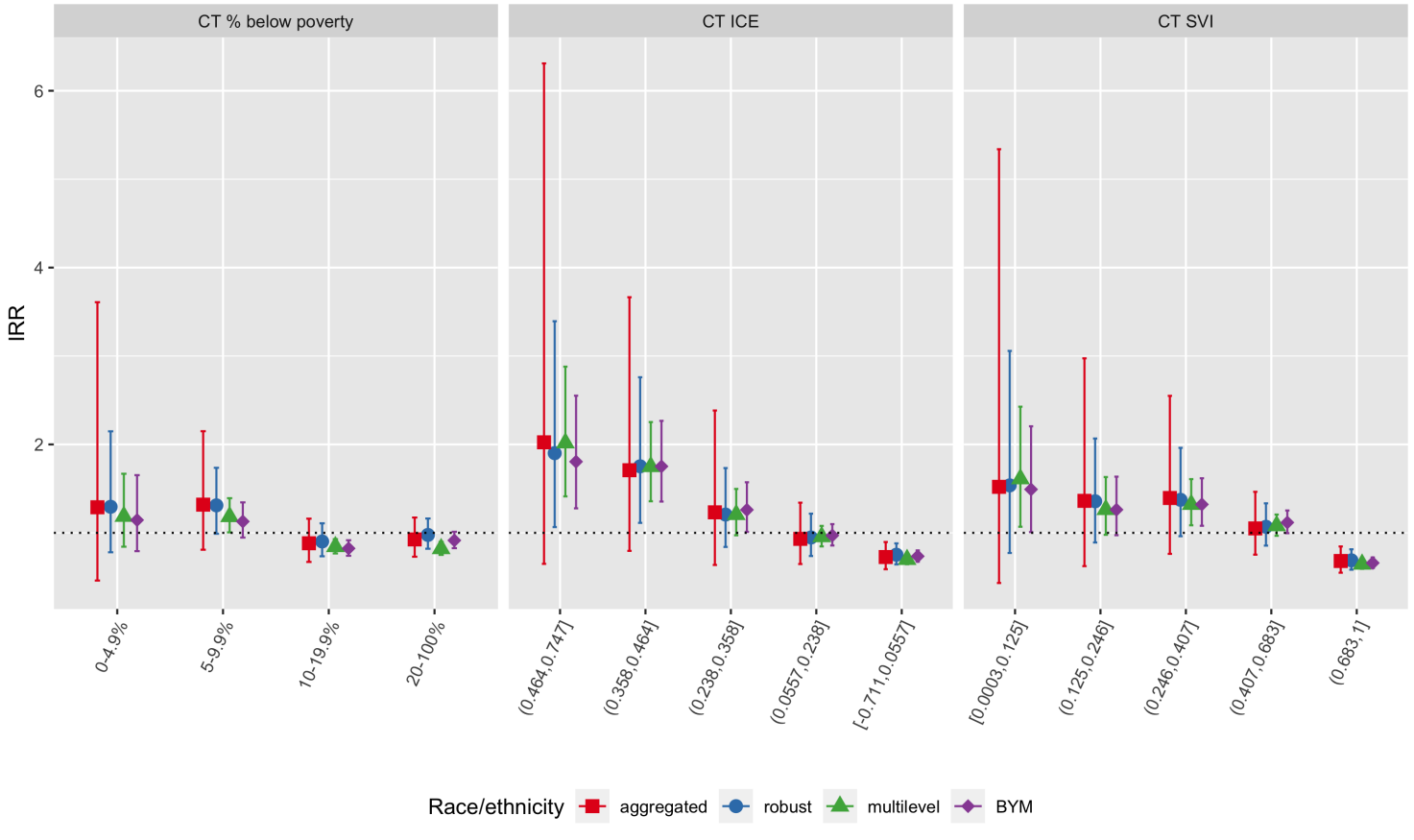


Figure 6.30: Comparison of Black vs. non-Hispanic White inequities in premature mortality within category of CT ABSM as estimated by the aggregated method, Poisson regression with robust variance estimator, multilevel Poisson regression (census tracts in city/town/neighborhoods), and spatial Poisson regression with BYM prior, Greater Boston, 2013-2017

As seen in Figure 6.30, Black-White inequities as measured by incidence rate ratios are largest in the most advantaged census tracts, whereas inequities by racialized group are smaller or even reversed in the most disadvantaged census tracts. That is, Non-Hispanic Whites living in the most disadvantaged census tracts have higher observed age-adjusted premature mortality rates than Black populations living in the same areas. Why do you think this might be the case?

Across methods, we note that the confidence limits on age-standardized rates, incidence rate differences, and incidence rate ratios are much wider for the aggregated method compared with model-based estimates. This reflects the impact of age-standardization, whereby age strata with relatively less information may nevertheless contribute strongly to the age-standardized estimates if those age strata are highly weighted in the standard population. The result is that standard errors and confidence limits from the aggregated method will usually be larger than their corresponding model-based estimates. Thus, the aggregated method, the most commonly used, has less power to detect statistically significant inequities.

6.9 REFERENCES

Banerjee A, Dhillon I, Ghosh J, Merugu S. An information theoretic analysis of maximum likelihood mixture estimation for exponential families. In Proceedings of the twenty-first international conference on Machine learning 2004 Jul 4 (p. 8). https://dl.acm.org/doi/10.1145/1015330.1015431 Accessed June 4, 2022

Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics. 1991 Mar;43(1):1-20.

Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Statistical methods in medical research. 2005 Feb;14(1):35-59.

Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 1987 Sep 1:671-81.

Gelman, A. Why I don't use the term "fixed and random effects." Post in Statistical Modeling, Causal Inference, and Social Science. 01,25,2005. Available at: https://statmodeling.stat.columbia.edu/2005/01/25/why_i_dont_use/

Jackson JW, Williams DR, VanderWeele TJ. Disparities at the intersection of marginalized groups. Soc Psychiatry Psychiatr Epidemiol. 2016 Oct;51(10):1349-1359. doi: 10.1007/s00127-016-1276-6. Epub 2016 Aug 16. PMID: 27531592; PMCID: PMC5350011.

Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012 Apr;41(2):514-20. doi: 10.1093/ije/dyr218. Epub 2012 Jan 9. PMID: 22253321; PMCID: PMC3324457.

Krieger N, Waterman PD, Chen JT, Rehkopf DH, Subramanian SV. The Public Health Disparities Geocoding Project Monograph. Available as of June 30, 2004 at: http://www.hsph.harvard.edu/thegeocodingproject

Lawson AB, Browne WJ, Rodeiro CL. Disease mapping with WinBUGS and MLwiN. John Wiley & Sons; 2003 Sep 12.

Nethery RC, Rushovich T, Peterson E, Chen JT, Waterman PD, Krieger N, Waller L, Coull BA. Comparing denominator sources for real-time disease incidence modeling: American Community Survey and WorldPop. SSM-Population Health. 2021 Jun 1;14:100786.

Newsom, J. Distinguishing Between Random and Fixed: Variables, Effects, and Coefficients. Psy 526 Lecture Material: Multilevel Regression. Spring 2019. Available at: http://www.web.pdx.edu/~newsomj/mlrclass/ho_randfixd.pdf

Pickle LW, White AA. Effects of the choice of age-adjustment method on maps of death rates. Statistics in medicine. 1995 Mar 15;14(5-7):615-27.

United States Census Bureau. Population and Housing Estimate Program (PEP). Updated 2022. Available at: https://www.census.gov/programs-surveys/popest.html

VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction. New York, NY: Oxford University Press, 2015.

Wakefield JC, Best NG, and Waller L. Bayesian approaches to diseasemapping, In: Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. SpatialEpidemiology: Methods and Applications, Oxford University Press, London, UK.pp 104–127, 2000.

Wolpert RL, Ickstadt K. Poisson/gamma random field models for spatial statistics. Biometrika. 1998 Jun 1;85(2):251-67.

« 5 Visualizing your data

7 Case Study 1: Premature Mortality in Massachusetts (2013 - 2017) »

7 Case Study 1: Premature Mortality in Massachusetts (2013 - 2017)

By: Justin W. Morgan

7.1 Introduction

In this case study, the outcome of interest is premature mortality (here defined as death occuring before age 65). This data, which was requested from the Massachusetts Registry of Vital Records and Statistics is for the years 2013-2017 and is merged with area-based social metrics (ABSMs) created from American Community Survey (ACS) 5-Year Estimates. For more information on how the ACA estimates these metrics from their survey, please visit their website. Vital Statistics registries require data on residential address at time of death, so they can be useful tools for monitoring health and health equity. These data were geocoded using the Google Maps API.

7.2 Motivation, Research Questions, and Learning Objectives

The goal of this case study is to develop familiarity with methods of exploring and visualizing racial disparities in health data. Our specific goals will be to:

- Download and merge health outcome and ABSM data
- Visualize and map estimates of ABSMs and premature mortality
- Identify the relationships between racialized group, ABSMs, and premature mortality
- Model the interaction effects between racialized group and ABSM.

The research questions we will seek to answer throughout this case study include:

1. What is the overall socioeconomic gradient in premature mortality?
2. What is the racialized disparity in premature mortality?
3. How does ABSM interact with individual level membership in racialized groups? (i.e., interactions between socioeconomic position and racialized groups, not just socioeconomic inequities within racialized groups)

7.3 Downloading and Wrangling Your Data

In this section, we will show you how to download ACS data by querying the census API and how to manipulate the data into the format we need for the rest of the analysis. This case study investigates a constructed variable: the Index of Concentration at the Extremes (for high-income Non-Hispanic White people vs low-income people of Color). See if you can replicate the analysis with a different ACS variable!

Note: Our outcome data cannot be shared due to privacy restrictions - for this case study, you will receive a pre-wrangled mortality dataset. Example code for how the wrangling could be done is included in this chapter's appendix.

7.3.1 Dependencies

You will need these packages throughout our case study:

```r
# "mission critical" packages
library(tidyverse)      # A collection of packages used for tidy data wrangling
library(readxl)         # A package for easy loading of Excel files
library(ggplot2)        # The most popular and flexible visualization package in R
library(tidycensus)     # A package to download data from the U.S. Census Bureau API
library(tigris)         # A package to download shapefiles from the U.S. Census Bureau
library(sf)             # A package with tools for simple (spatial) features
library(INLA)           # A package that allows for bayesian inference
inla.setOption(inla.mode = "experimental") # An option to use the experimental (beta version)
library(spdep)          # A package to allow for spatial weighting in analyses
# "nice-to-have" packages
library(cowplot)        # A ggplot2 add-on that will allow us to add sub plots
options(tigris_use_cache = TRUE) # An option within `tigris` to save what you download
library(viridis)        # A package including color-blind friendly color pallettes
library(Hmisc)          # A package containing useful functions for data analyses
library(fastDummies)    # A package that includes functions to create indicator variables
library(mapview)        # For interactive mapping, including topographical mapping
library(purrr)          # A programming toolkit for R
library(scales)         # An add-on package to help with scaling our maps appropriately
library(broom)          # A package that allows for extraction and wrangling of model output
# If this code does not run for you, you may need to run install.package("package_name")
```

7.3.2 **Your Health Outcome Data - Premature Mortality**

This data have been aggregated from individual observations into death counts by year, age, sex, racialized group, census tract, and town. When you receive unrestricted mortality files from government agencies for research, you will likely encounter files with one observation per death. After you have geocoded these individual-level observations, you will need to aggregate them up to the geographic level of interest for your analysis. The data we use is also aggregated into groups that correspond to common variables on the ACS, so that they can be aligned to population counts that will be used as denominators as rates are calculated. Aggregation to specific age groups and racial categorizations allows us to perform key aspects of our analysis, such as age-standardization of the data, and stratification of analysis by racialized group.

> NOTE: the variable for town refers to a constructed variable. Ideally when performing an analysis that might include two or more levels, the smaller level (here, census tracts) would be nested entirely within the larger level (towns). In Massachusetts though, there are several towns that have such small populations that they are smaller than census tracts. For these towns, we have created a crosswalk wherein neighboring small towns have been combined to create larger “super towns”. These super towns each make up one census tract, so that each super town in the analysis will now have at least one census tract “nested within it” in a multilevel analysis. If you are interested in attempting a multilevel analysis, this data may make good practice!

```r
ma_mort_ct <- readRDS("your_file_path/ma_mort_ct.RDS")
```

7.3.3 **Your Denominator Data and ABSM**

We download population data (to use as denominators to calculate mortality rates) and ABSMs from the U.S. Census Bureau API using the `tidycensus` package. This requires registering with the Bureau for an API key. The key is redacted here, but you can get your own from the Bureau. The ABSM we will use for this analysis is the Index of Concentration at the Extremes for Racialized Economic Segregation (ICE), which quantifies how persons in a specified area are concentrated into the top vs bottom of a specified societal distribution. The distribution we will construct is comparing high-income Non-Hispanic White people to low-income people of Color.

```
# Population Denominators
ma_denominator <- vector(mode = "list", length = 5)     # Create an empty list to store our
data in
names(ma_denominator) <- c(2013,2014,2015,2016,2017)    # Name the indices of the list for the
data years
# The following code creates a loop, and then runs the tidycensus package get_acs() function to
draw down population counts for a list of variables (using their Census Bureau designations).
You can see the full list of variables at
https://api.census.gov/data/2019/acs/acs5/variables.html.
for (nm in names(ma_denominator)) {
  ma_denominator[[nm]] <- get_acs(geography = "tract",
                            variables = c("B01001_003","B01001_004","B01001_005","B01001_006",
                                          "B01001_007","B01001_008","B01001_009","B01001_010",
                                          "B01001_011","B01001_012","B01001_013","B01001_014",
                                          "B01001_015","B01001_016","B01001_017","B01001_018",
                                          "B01001_019","B01001_027","B01001_028","B01001_029",
                                          "B01001_030","B01001_031","B01001_032","B01001_033",
                                          "B01001_034","B01001_035","B01001_036","B01001_037",
                                          "B01001_038","B01001_039","B01001_040","B01001_041",
                                          "B01001_042","B01001_043",
                                          "B01001B_003","B01001B_004","B01001B_005",
                                          "B01001B_006","B01001B_007","B01001B_008",
```

```
# Area Based Social Metric
ma_absm <- vector(mode = "list", length = 5)
names(ma_absm) <- c(2013,2014,2015,2016,2017)
for (nm in names(ma_absm)) {
  ma_absm[[nm]] <- get_acs(geography = "tract",
                           variables = c("B01003_001E",
                                         "B19001_001E",

"B19001_002E","B19001_003E","B19001_004E","B19001_005E",

"B19001_014E","B19001_015E","B19001_016E","B19001_017E",

"B19001H_002E","B19001H_003E","B19001H_004E","B19001H_005E",

"B19001H_014E","B19001H_015E","B19001H_016E","B19001H_017E"),
                           year = as.numeric(nm) + 2,
                           output = "wide",
                           state = "MA",
                           geometry = FALSE,
                           key = "4407a63721e192545e1e2a2fc7f6920477b10108",
                           moe_level = 95,
                           survey = "acs5",
```

7.4 **Approach**

Now that we have our data, let's revisit our questions of interest:

1. What is the overall socioeconomic gradient in premature mortality?
2. What is the racialized disparity in premature mortality?
3. How does ABSM interact with individual level membership in racialized groups? (i.e., interactions between socioeconomic position and racialized groups, not just socioeconomic inequities within racialized groups)

Let's first visualize the overall socioeconomic gradient in premature mortality.

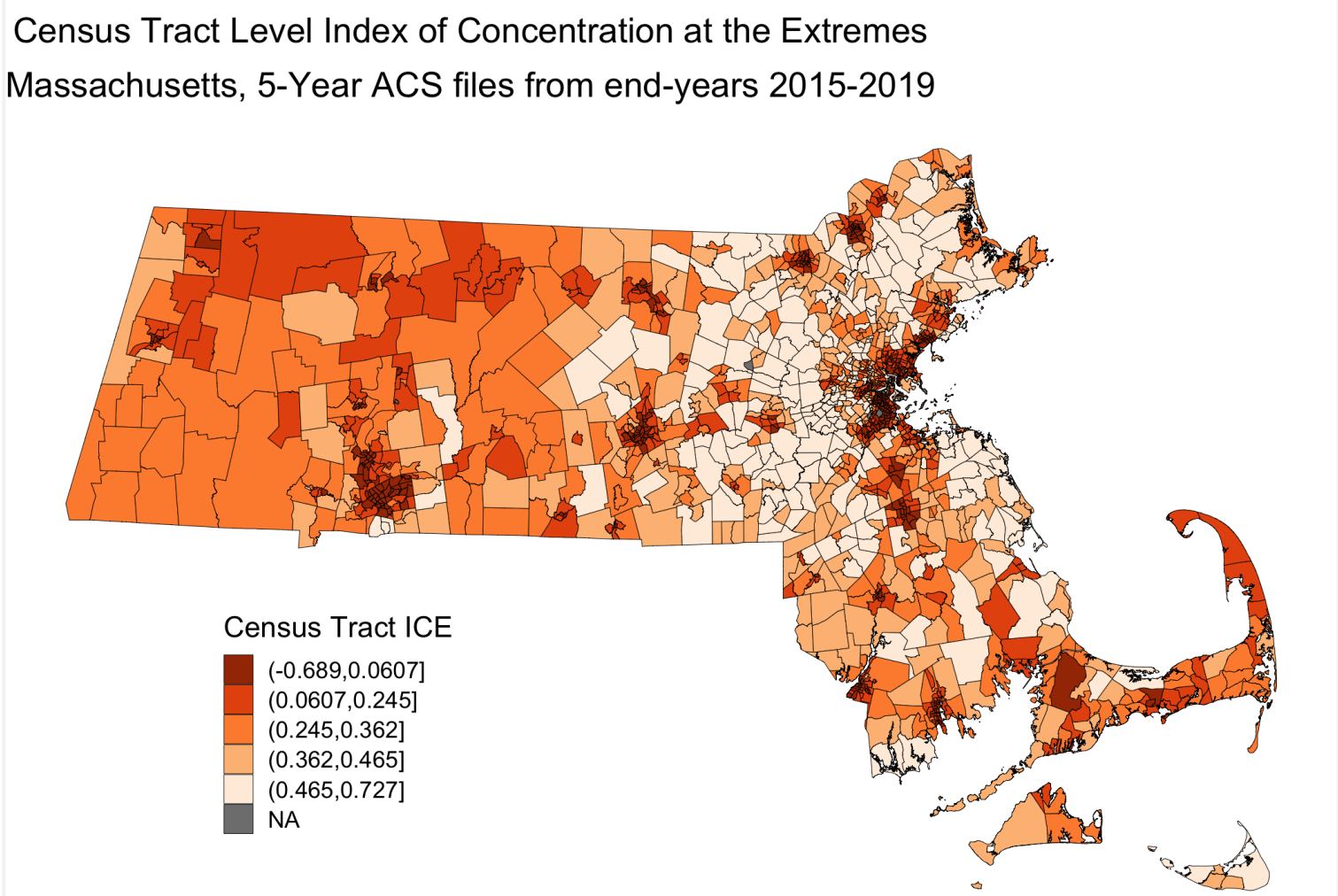
7.4.1 **What is the overall socioeconomic gradient in premature mortality?**

Before we try to see how our ABSM might impact premature mortality, let us first assess how the ICE measure is distributed across space in Massachusetts. We can visualize this using a map. In order to map our data, we will need to add geometries to our dataset. We will import a census tract shapefile produced by the government of Massachusetts.

```
tract_geometry <- st_read("CENSUS2010_BLK_BG_TRCT_SHP",
                          layer = "CENSUS2010TRACTS_POLY") %>%
  mutate(id_order = row_number()) %>%
  select(GEOID10, id_order)
```

This code adds the census tract geometries to our ICE data, and then creates a map. There are several lines, each referring to a different customization of the map. `ggplot2` has many such options, you can learn more about these here

```
map.ice <- ma_absm_sum %>%
  left_join(tract_geometry, by = "GEOID10") %>%
  ggplot(aes(geometry = geometry, fill=ICE_qt)) +
    geom_sf(col="black", size=0.1) +
    labs(fill="Census Tract ICE",
         x="", y="",
         title=expression(atop("Census Tract Level Index of Concentration at the Extremes",
                               "Massachusetts, 5-Year ACS files from end-years 2015-2019"))) +
    scale_fill_brewer(palette="Oranges", direction=-1, na.value="grey50") +
    theme_void() +
    theme(axis.text.x=element_blank(), #remove x axis labels
          axis.ticks.x=element_blank(), #remove x axis ticks
          axis.text.y=element_blank(),  #remove y axis labels
          axis.ticks.y=element_blank(),
          legend.position = c(0.25, 0.25),
          legend.key.size = unit(0.4, "cm"))
# ggsave("your_file_path/map.ice.png")
```



This map is showing us how the ICE measure is distributed across the state of Massachusetts. This particular iteration of ICE compares high-income Non-Hispanic White people with low-income people of Color. When ICE is low, those neighborhoods

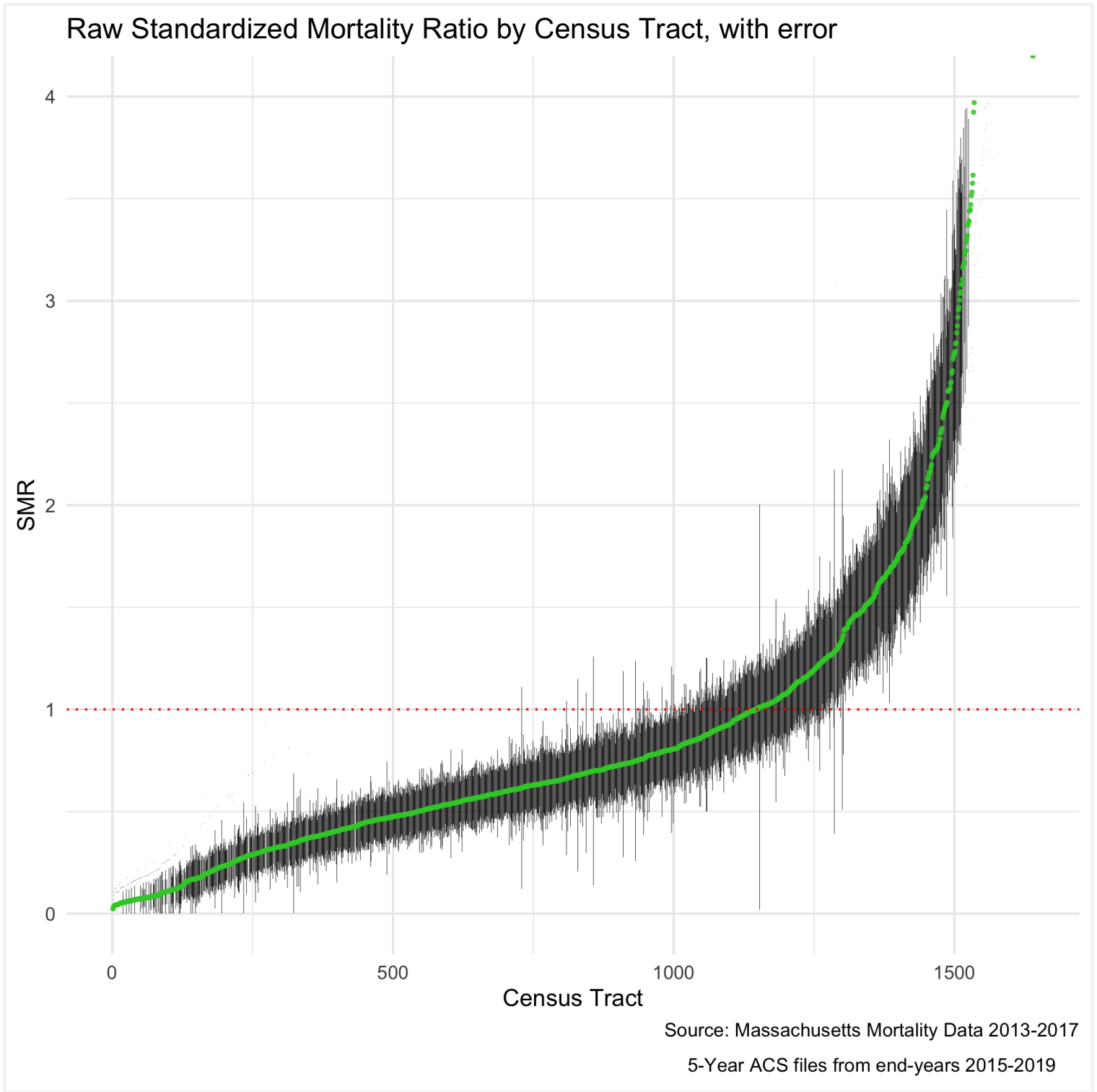
are concentrated with low-income people of Color. At 0, there is no extreme concentration (or, alternatively, the polarization is balanced). And when ICE is higher, there is a high concentration of high-income Non-Hispanic White people.

Next, we will visualize our premature mortality data, using spatial models. In order to do this, we will first utilize the indirect method to age-standardize the data. Once we have the data age-standardized we can calculate raw standardized mortality ratios (SMRs).

```
# Calculate reference rates by age for MA overall
# This will give us overall totals, and remove the race specific information we arent
interested in right now
ma_total_pop <- ma_denominator %>%
                filter(race_group == "Total")
# This will align the mortality counts with those denominators so we can create total
population expected death rates for each age group
reference_rates <- ma_mort_ct %>%
  filter(race_group == "Total",
         sex != "Total",
         age_cat != "Total") %>%
  group_by(GEOID10, year, sex, age_cat) %>%
  summarise(deaths=n()) %>%
  left_join(ma_total_pop,
            by=c("GEOID10", "year", "sex", "age_cat")) %>%
  mutate(age_cat = case_when(age_cat %in% c("35-39","40-44") ~ "35-44",
                             age_cat %in% c("45-49","50-54") ~ "45-54",
                             age_cat %in% c("55-59","60-64") ~ "55-64",
                             TRUE ~ age_cat),
         deaths = ifelse(is.na(deaths), 0, deaths)) %>%
  group_by(age_cat) %>%
  summarise(num = sum(deaths),
```

These ratios are susceptible to much of what is discussed in previous chapters with regards to infinity-magnitude rates, and smaller sample sizes inducing extreme results. But as we are going to eventually smooth this data with our model, we are not as concerned here about the state of the raw data. Here we can see our raw standardized mortality ratios extend from zero to infinity.

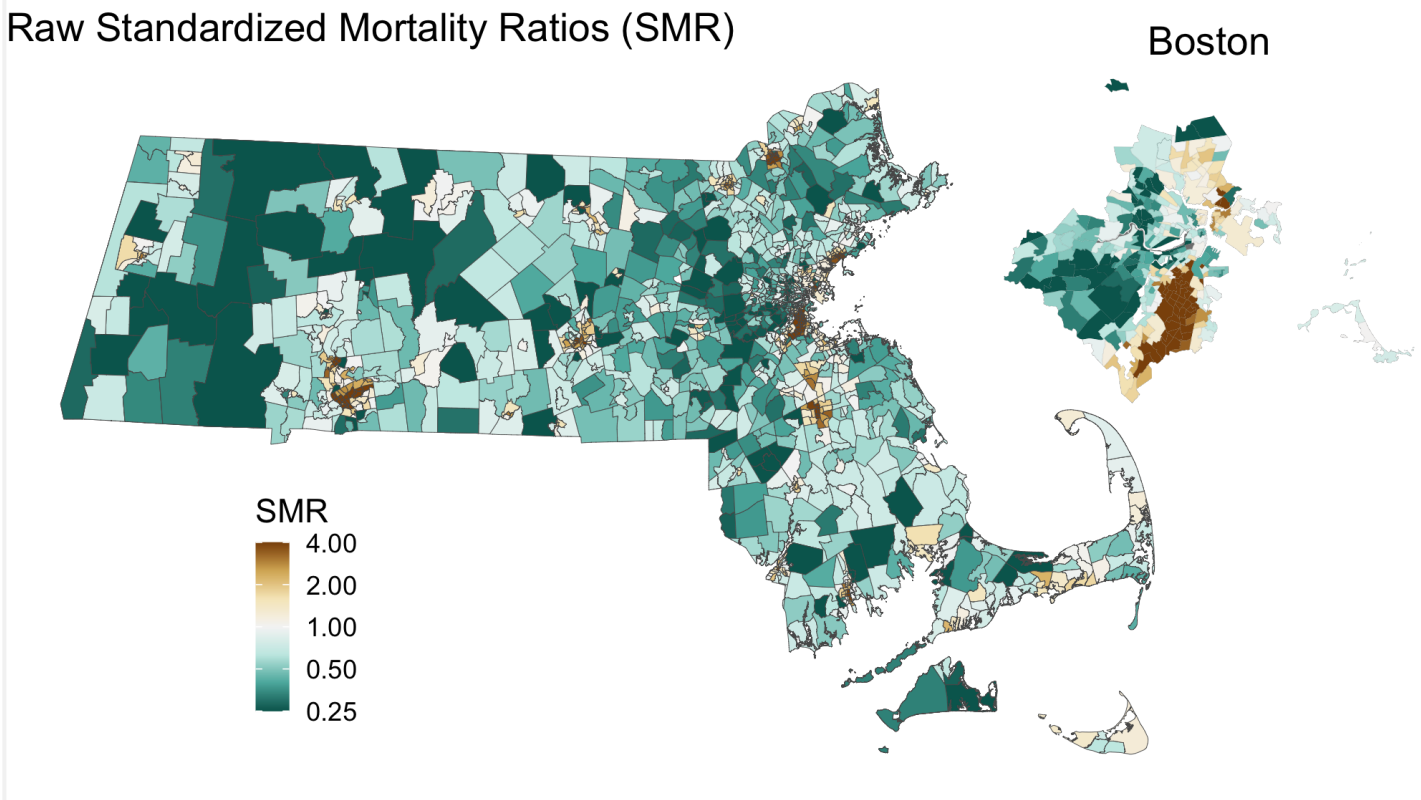
```
plot.raw_smr <-  ma_mort_istd %>%
  arrange(raw_smr) %>%
  mutate(orderID = row_number()) %>%
  ggplot(aes(x=orderID, y=raw_smr)) +
    geom_errorbar(aes(ymin = raw_smr_lo95, ymax=raw_smr_up95), size = 0.1) +
    geom_point(color = "limegreen", alpha = 0.8, size = 0.5) +
    geom_hline(yintercept = 1, col="red", linetype="dotted") +
    ylim(0,4) +
    labs(title = "Raw Standardized Mortality Ratio by Census Tract, with error",
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019"))) +
    xlab("Census Tract") +
    ylab("SMR") +
    theme_minimal()
# ggsave("your_file_path/plot.raw_smr.png")
```



The SMR shows us the ratio of mortality in a given census tract compared to what would be expected had that census tract had the same age_specific rates as the standard population. So an SMR greater than 1 means the area premature mortality is higher than expected. An SMR less than one means the opposite.

We can also map out this relationship:

```
# Here you will see code for two maps. This block of code creates one state level map, and then
a smaller, Boston area map to accompany it. This optional subsetting provides detail in an area
of interest. The cowplot package combines the maps to display both in one space
map.raw_smr_state <- ma_mort_istd %>%
  left_join(tract_geometry, by= "GEOID10") %>%
  ggplot() +
  geom_sf(mapping = aes(geometry=geometry,
                        fill=raw_smr),
          lwd = 0.1) +
    scale_fill_distiller(palette = "BrBG",
                         trans = scales::pseudo_log_trans(sigma=0.01),
                         limits = exp(c(-1,1)*log(4)),
                         breaks = c(0.25,0.5,1,2,4), oob=squish) +
    labs(title = "Raw Standardized Mortality Ratios (SMR)",
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019")),
         fill = "SMR", x="", y="") +
    theme_void() +
    theme(axis.text.x=element_blank(), #remove x axis labels
          axis.ticks.x=element_blank(), #remove x axis ticks
          axis.text.y=element_blank(),  #remove y axis labels
          axis.ticks.y=element_blank(),
```



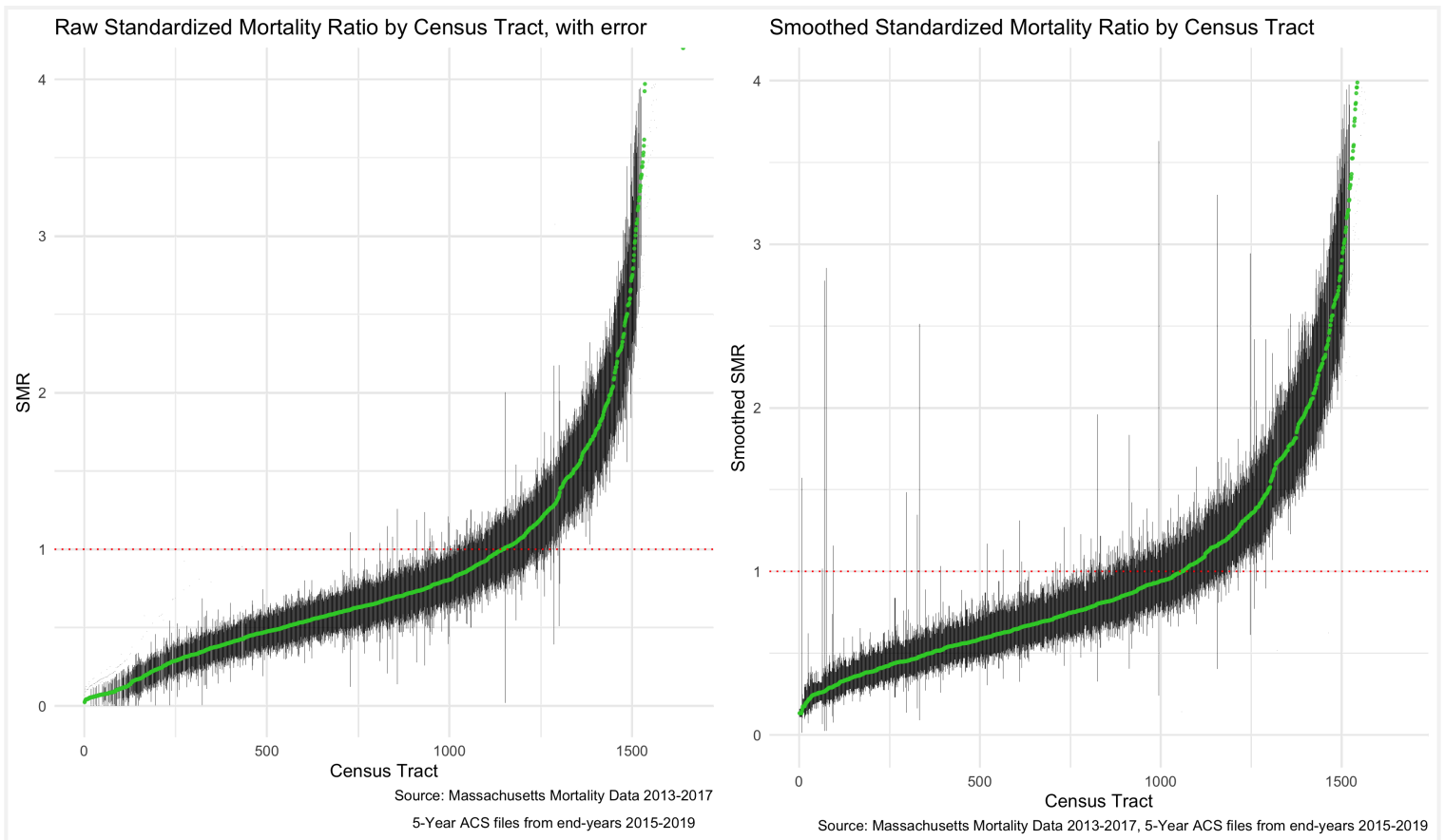
We can fit a spatial model using the `INLA` package which will "smooth" our results by fitting our data to a Besag York Mollié (BYM) model. As mentioned in previous chapters, a BYM model will adjust estimates based on the surrounding areas - instead of assuming neighboring geographies are independent, the model will assume they are similar. In order to implement this model in R, we first need to create an adjacency matrix, so that the package knows which census tracts are neighbors

```
# We first want to make sure that our data file has our geometries in the same order as our
shapefile - in other words, ensure that they are aligned.
ma_mort_istd_ordered <- ma_mort_istd %>%
  mutate(intercept = 1) %>%
  right_join(tract_geometry, by="GEOID10")%>%
  arrange(id_order) %>%
  select(-geometry)
n.tracts <- ma_mort_istd_ordered %>%
  select(GEOID10) %>%
  unique() %>%
  nrow()
# THis code will calculate the adjacency matrix
W.nb <- poly2nb(tract_geometry , snap=0.001)
W.list <- nb2listw(W.nb, style="B", zero.policy = TRUE)
# And this code will convert that matrix into a format INLA can understand
nb2INLA("INLA_adj_mat", W.nb) # this saves a file in the working directory
INLA_adj_mat <- "INLA_adj_mat"
# Intercept only ("null") BYM model
model_form_0 <- O ~ 1 + f(id_order, model="bym2", graph=INLA_adj_mat, scale.model=TRUE,
constr=TRUE)
model_0 <- inla(model_form_0, family="poisson",
                data=ma_mort_istd_ordered, E=E, # E points to the expected count field
```

Note the code above provides code to calculate the percentage of the variance in the data that is spatially correlated - in this case, about 45%.

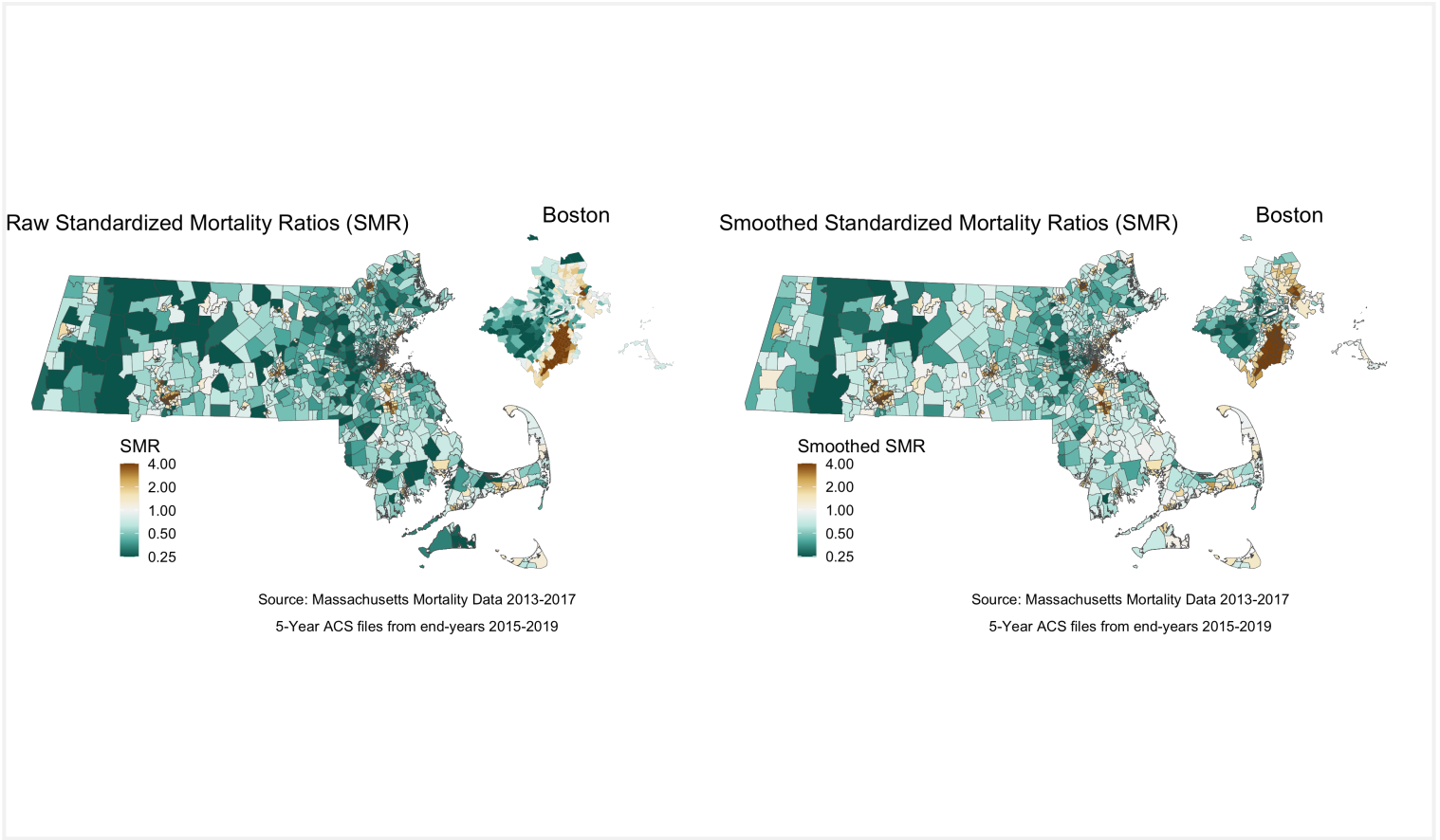
We can now visualize our smoothed SMR. First, let's look at a similar caterpillar plot to what we did with the raw SMR. The plot looks slightly flatter, and notably, the infinity and zero values have been pulled more towards the middle of the plot.

```
plot.smr_smooth <- ma_mort_istd_smoothed %>%
  ungroup() %>%
  arrange(smooth_smr) %>%
  mutate(orderID = row_number()) %>%
  ggplot(aes(x=orderID, y=exp(smooth_smr))) +
    geom_errorbar(aes(ymin = exp(smooth_smr_lo95), ymax=exp(smooth_smr_up95)), size = 0.1) +
    geom_point(color = "limegreen", alpha = 0.8, size = 0.5) +
    geom_hline(yintercept = 1, col="red", linetype="dotted") +
    ylim(0,4) +
    labs(title = "Smoothed Standardized Mortality Ratio by Census Tract",
         caption = "Source: Massachusetts Mortality Data 2013-2017, 5-Year ACS files from end-
years 2015-2019") +
    xlab("Census Tract") +
    ylab("Smoothed SMR") +
    theme_minimal()
plot.comp_smr <- plot_grid(plot.raw_smr, plot.smr_smooth,
                           ncol = 2)
# ggsave("your_file_path/plot.comp_smr.png"
```



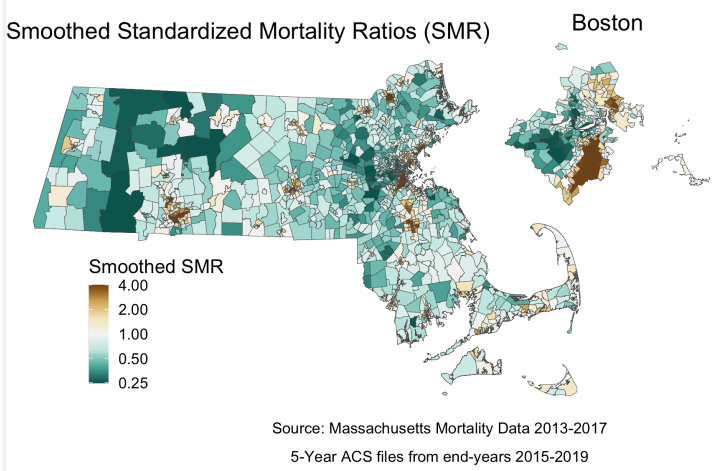
And in the map, we can see some of the color of extreme values lighten.

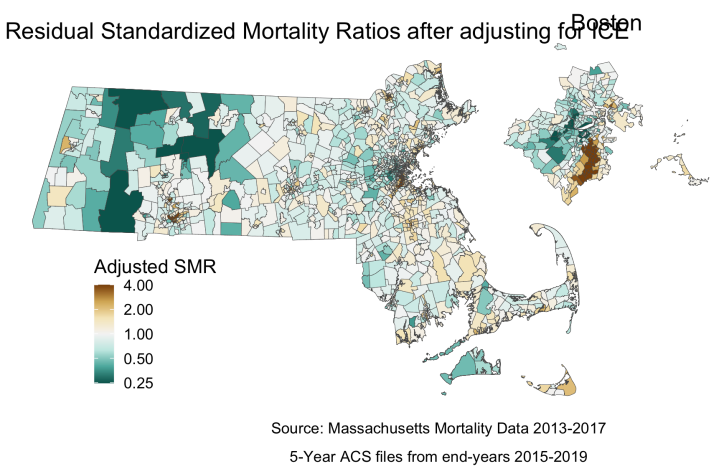
```
map.smooth_smr_state <- ma_mort_istd_smoothed %>%
  left_join(tract_geometry, by= "GEOID10") %>%
  ggplot() +
  geom_sf(mapping = aes(geometry=geometry,
                        fill= exp(smooth_smr)),
          lwd = 0.1) +
    scale_fill_distiller(palette = "BrBG",
                         trans = scales::pseudo_log_trans(sigma=0.01),
                         limits = exp(c(-1.01,1)*log(4)),
                         breaks=c(0.25,0.5,1,2,4), oob=squish) +
    labs(title = "Smoothed Standardized Mortality Ratios (SMR)",
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019")),
         fill = "Smoothed SMR", x="", y="") +
    theme_void() +
    theme(axis.text.x=element_blank(), #remove x axis labels
          axis.ticks.x=element_blank(), #remove x axis ticks
          axis.text.y=element_blank(),  #remove y axis labels
          axis.ticks.y=element_blank(),
          legend.position = c(0.25, 0.25),
          legend.key.size = unit(0.4, "cm"))
map.smooth_smr_boston <- ma_mort_istd_smoothed %>%
```



We can see from the plot and the map that while there is still variation in the spatially smoothed SMR - it is a bit less extreme in some areas than the raw SMR. To get a sense of how premature mortality varies with respect to ICE, we should look at a model that adjusts for our ICE data and see if there are any changes in variation.

```
#Code to include our ABSM
model_form_1 <- O ~ 1 + f(id_order, model='bym2', graph=INLA_adj_mat, scale.model=TRUE,
constr=TRUE) + ICE_qt_2 + ICE_qt_3 + ICE_qt_4 + ICE_qt_5 + ICE_qt_NA
model_1 <- inla(model_form_1, family="poisson",
                data = ma_mort_istd_ordered,
                E=E, # E points to the expected count field
                control.predictor = list(compute=TRUE), # computes transformed posterior
marginals
                control.compute   = list(dic=TRUE)) # computes DIC for model fit
# We can use this code to extract results of the model
fixed_results <- model_1$summary.fixed
dic_results <- summary(model_1$dic$dic)
# the first block of the summary.random output are the area effects (u + v)
random_results <- model_1$summary.random$id_order$mean[1:n.tracts]
# and this code pulls out the residual results after adjusting for ICE
risk_residuals <- data.frame(ice_residuals = random_results) %>%
  mutate(id_order = row_number())
# append to dataset of other model effects
ma_mort_istd_adj <- ma_mort_istd_smoothed %>%
  left_join(risk_residuals, by="id_order")
# Visualize map after adjusting for ICE
map.adj_smr_state <- ma_mort_istd_adj %>%
```





> Comparing the map adjusted for the ICE measure to the smoothed map reveals that concentration of polarized racial-economic populations did contribute to some of the impact. This is most clearly seen in the less pronounced SMRs (in both directions) after adjusting for ICE in areas of southern Boston.

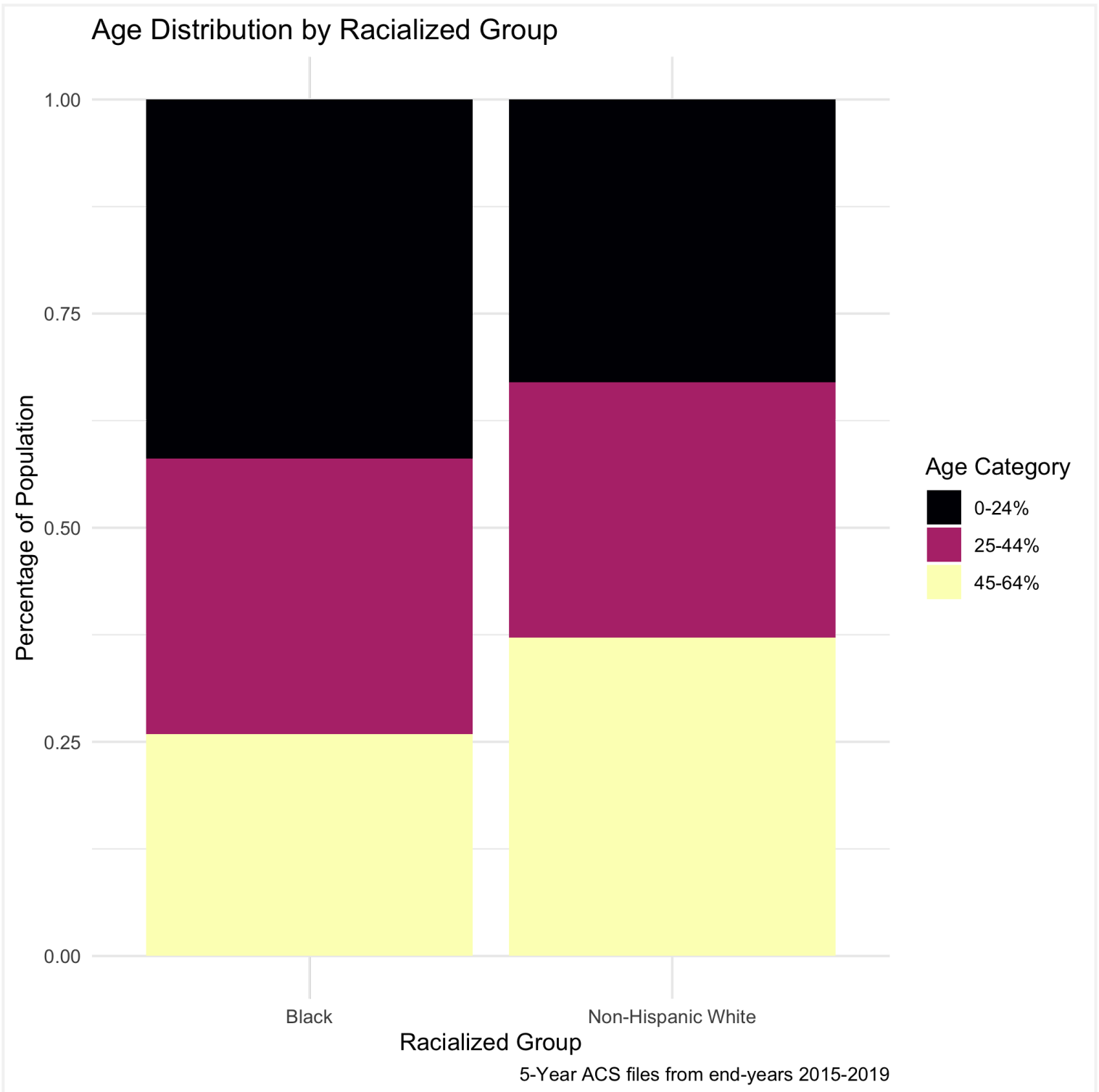
7.4.2 What is the racialized disparity in premature mortality overall?

We can describe inequities by racialized groups much the same way we describe inequities in ABSM - by aggregating death and population data. We already have these aggregated - now we can stratify our analysis to compare the White Non-Hispanic and Black populations. Throughout the example we may refer to these groups as White and Black as a shorthand.

> NOTE: You may have noticed that the description of the White population includes a “Non-Hispanic” designation, while the description of the Black population does not. This is a quirk of the ACS data - population counts are not readily available for the Black Non-Hispanic population. There is thus a slight mismatch in in the numerator from the mortality data (Non-Hispanic Black deaths) and denominator (Black population).

We will need to age-adjust our data here, as differences in premature mortality may be due to differential distribution of ages by racialized group seen here:

```
tab_age_race <- ma_denominator %>%
      filter(race_group %in% c("Non-Hispanic White","Black"),
         sex != "Total",
         age_cat != "Total") %>%
  group_by(GEOID10, race_group, sex, age_cat) %>%
  summarise(pop = sum(population, na.rm=T)) %>%
  inner_join(ma_absm_sum, by = c("GEOID10")) %>%
  mutate(age_cat_broad = case_when(age_cat %in% c("0-4","5-9","10-14","15-19","20-24") ~ "0-
24%",
                                   age_cat %in% c("25-29","30-34","35-39","35-44","40-44") ~
"25-44%",
                                   age_cat %in% c("45-49","45-54","50-54","55-59","55-64","60-
64") ~ "45-64%")) %>%
  group_by(race_group, age_cat_broad) %>%
  summarise(pop = sum(pop, na.rm=T)) %>%
  group_by(race_group) %>%
  mutate(percentage = pop/sum(pop)) %>%
  ggplot(aes(x=race_group, y=percentage, fill= age_cat_broad)) +
    geom_bar(position="stack", stat="identity") +
  scale_fill_viridis_d(option = "A") +
  labs(title ="Age Distribution by Racialized Group",
       fill = "Age Category",
```

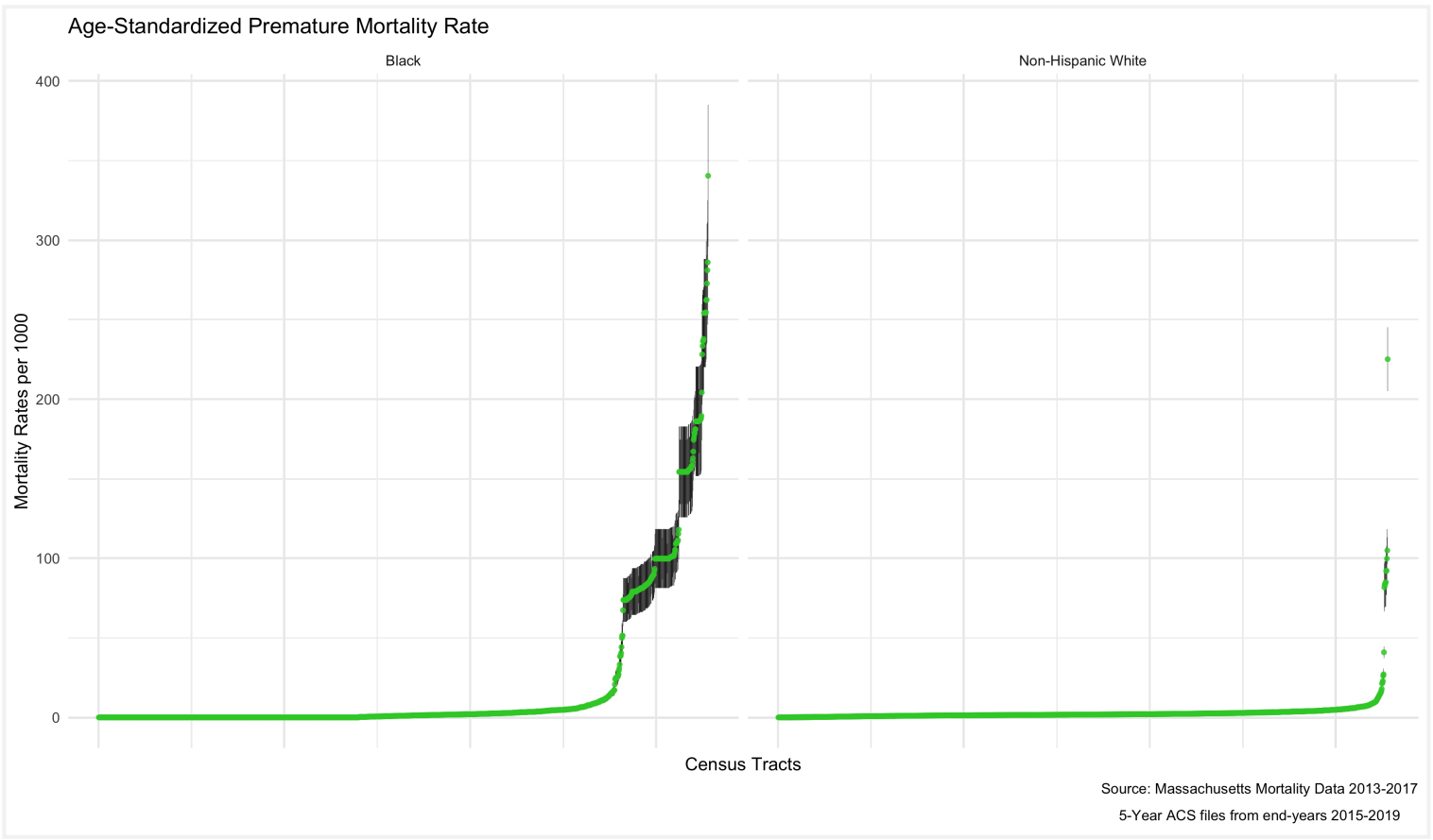


We will, in this case, do a direct age adjustment to get standardized mortality ratios. This will require a reference population, which we have downloaded (and wrangled) from the National Cancer Institute (https://seer.cancer.gov/stdpopulations/). This is preferential to indirect age standardization, as the Non-Hispanic White population would drive the expected rates using that method.

```r
seer_std <- readRDS("data/07-premature-mortality/seer_std.RDS")
```

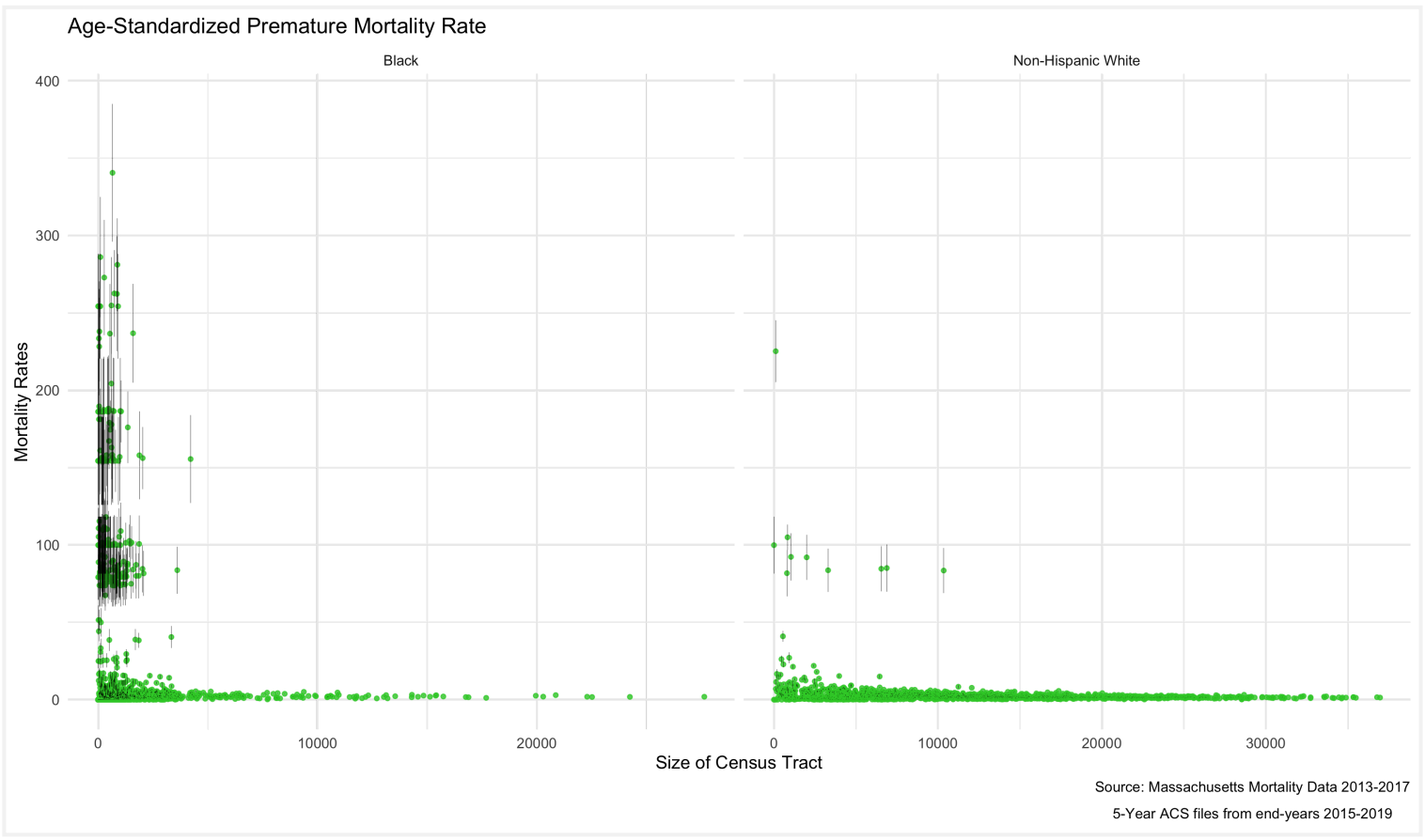
Let's now age adjust our aggregated data by racialized group so we can compare standardized mortality rates.

```r
ma_mort_dstd <- ma_mort_ct %>%
  mutate(race_group = case_when(race_group == "Non-Hispanic Black" ~ "Black",
                                TRUE ~ race_group)) %>%
  inner_join(ma_denominator, by = c("year","GEOID10","race_group","sex","age_cat")) %>%
  filter(race_group %in% c("Non-Hispanic White","Black"),
         age_cat != "Total") %>%
  group_by(GEOID10,super_town, race_group, age_cat) %>%
  summarise(num = sum(deaths, na.rm=T),
            den = sum(population, na.rm=T))%>%
  mutate(den = ifelse(den == 0, num, den)) %>% # this line takes areas with population counts as zero, and provides the count as the number of deaths. This is an approach to resolve when population data does not align with the outcome data
  left_join(seer_std, by="age_cat") %>%
  mutate(rate_i = wt*num/den,
         var_rate_i = (num*wt^2)/den^2) %>%
  group_by(GEOID10, super_town, race_group) %>%
  summarise(num = sum(num, na.rm=T),
            den = sum(den, na.rm=T),
            std_rate = sum(rate_i, na.rm=T),
            var_std_rate = sum(var_rate_i, na.rm=T),
            sumwt = sum(wt),
            sumwt2 = sum(wt^2)) %>%
```



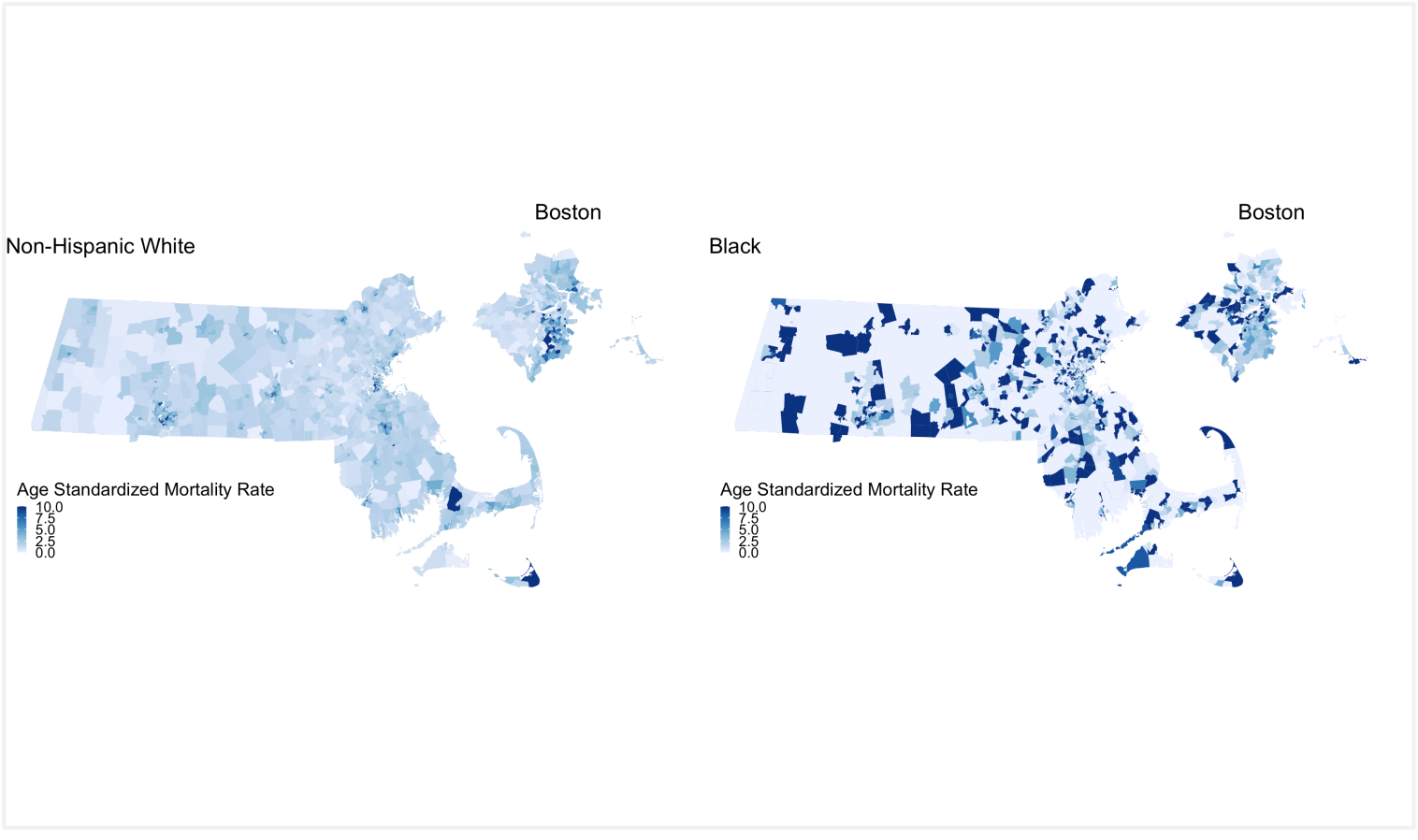
As mortality rates increase, there also seems to be an increase in the variance of the estimates. This is because the more extreme rates we are seeing are really a product of having very small sample sizes in some districts. We can see that demonstrated by ordering the results by the population size of the census tract.

```r
ordered.agg_rates <- ma_mort_dstd %>%
  filter(!is.na(std_rate)) %>%
  select(GEOID10, race_group, contains("std_rate"), den) %>%
  ggplot(aes(x=den, y=std_rate)) +
    geom_point(color = "limegreen", alpha = 0.8, size = 1) +
    geom_errorbar(aes(ymin = std_rate_lo95, ymax=std_rate_up95), size = 0.1) +
    labs(title = "Age-Standardized Premature Mortality Rate",
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019"))) +
         xlab("Size of Census Tract") +
         ylab("Mortality Rates") +
    facet_wrap(vars(race_group), ncol = 2, scales = "free_x") +
    theme_minimal()
# ggsave("your_file_path/ordered.agg_rates.png")
```



As we might expect, these extremely variable mortality rates are occurring in smaller populations. What do these maps look like?

```r
plotlist <- vector(mode = "list", length = 2)
names(plotlist) <- c("Non-Hispanic White","Black")
for (plt in names(plotlist)) { # This loop allows us to make the same map twice, one for Non-Hispanic White, and one for Black
  map.state <- ma_mort_dstd %>%
    filter(race_group == plt) %>%
    mutate(std_rate = ifelse(std_rate > 10, 10, std_rate)) %>%
    left_join(tract_geometry, by= "GEOID10") %>%
    ggplot() +
  geom_sf(mapping = aes(geometry = geometry,
                        fill = std_rate),
          lwd = 0) +
  scale_fill_distiller(palette = 'Blues', direction = 1, limits = c(0, 10)) +
  labs(title = plt,
       fill = "Age Standardized Mortality Rate", x="", y="") +
  theme_void() +
  theme(axis.text.x=element_blank(), #remove x axis labels
        axis.ticks.x=element_blank(), #remove x axis ticks
        axis.text.y=element_blank(),  #remove y axis labels
        axis.ticks.y=element_blank(),
        legend.position = c(0.25, 0.25),
        legend.key.size = unit(0.2, "cm"))
```



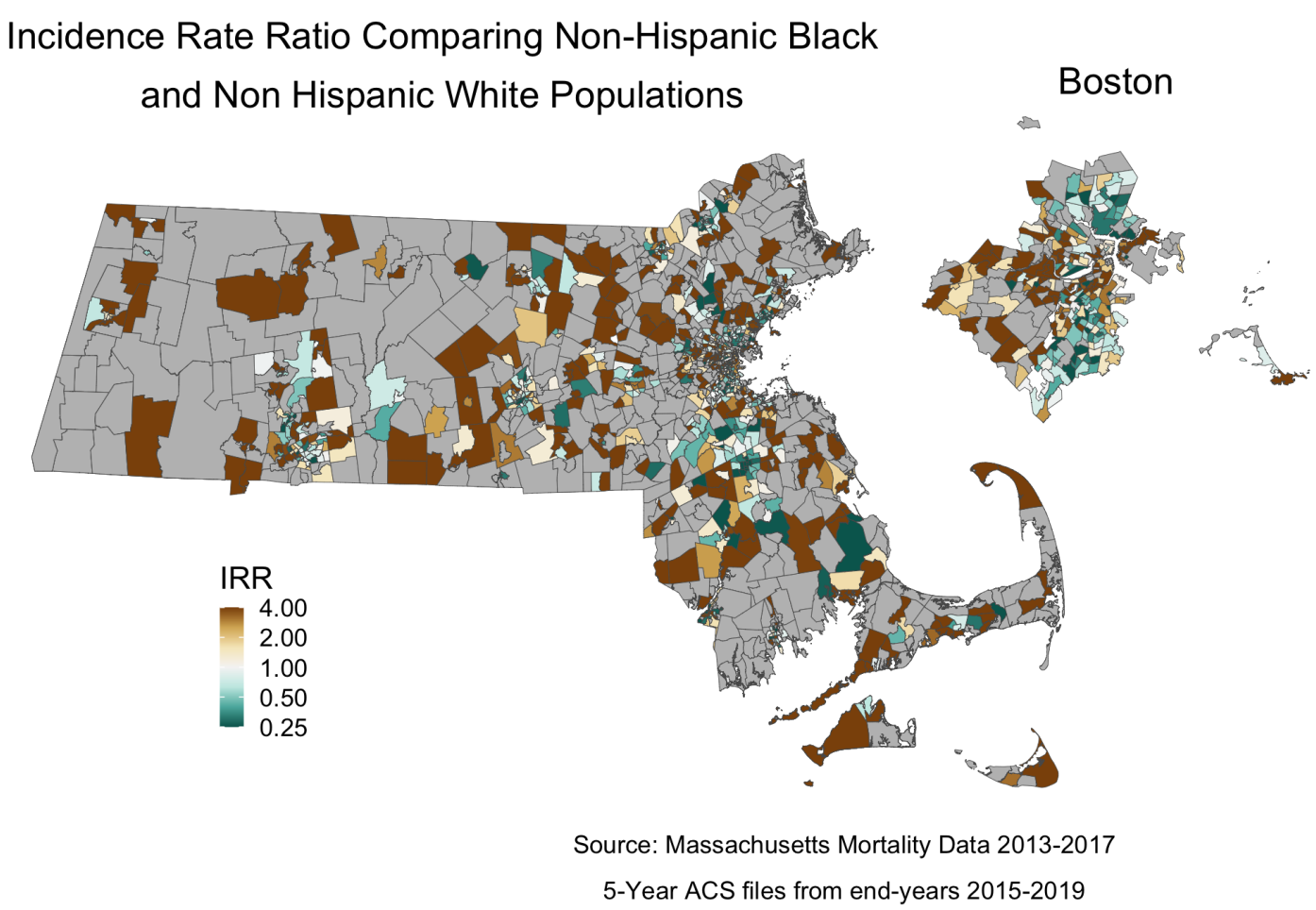
We are seeing some really high rates in more census tracts in the Black population, where we have smaller overall sample sizes. We are also seeing many census tracts with no black population. We can also display these two maps as one, by calculating the Rate Difference, or the Rate Ratio. Here we will show the crude Rate Ratio of the Age Standardized rates. Key to interpreting this map is understanding that these rates are influenced by the small sample sizes and large potential errors we have previously visualized. It's for researchers and communities to interpret how "real" the effects we see here are.

```
irr_data <- ma_mort_dstd %>%
  select(GEOID10,super_town,race_group, std_rate, var_std_rate) %>%
  pivot_wider(id_cols = c(GEOID10,super_town),
              names_from = race_group,
              values_from = c(std_rate, var_std_rate)) %>%
  mutate(irr = ifelse(`std_rate_Non-Hispanic White` == 0, NA_real_, `std_rate_Black` /
`std_rate_Non-Hispanic White`),
         irr_var = `var_std_rate_Black` + `var_std_rate_Non-Hispanic White`,
         irr_lo95 = irr - 1.96*sqrt(irr_var),
         irr_up95 = irr + 1.96*sqrt(irr_var))
irr_plot <- irr_data %>%
  arrange(irr) %>%
  mutate(orderID = row_number()) %>%
  ungroup() %>%
  ggplot(aes(x=orderID, y=irr)) +
    geom_point(color = "limegreen", alpha = 0.8, size = 1) +
    geom_errorbar(aes(ymin = irr_lo95, ymax=irr_up95), size = 0.1) +
    labs(title = expression(atop("Incidence Rate Ratio Comparing Black",
                                 "and Non Hispanic White Populations")),
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019"))) +
    xlab("Census Tracts") +
```

When we map the IRR, we can see the crude rate ratio between Black and Non-Hispanic White populations:

```
map.state.irr <- irr_data %>%
  mutate(irr = case_when(irr > 4  ~ 4,
                         std_rate_Black == 0 ~ NA_real_, # Setting as missing any areas with no
black deaths
                         TRUE     ~ irr)) %>%
  left_join(tract_geometry, by= "GEOID10") %>%
  ggplot() +
    geom_sf(mapping = aes(geometry = geometry,
                          fill = irr),
            lwd = 0.1) +
    scale_fill_distiller(palette = "BrBG",
                         trans = scales::pseudo_log_trans(sigma=0.01),
                         limits = exp(c(-1,1)*log(4)),
                         breaks = c(0.25,0.5,1,2,4), oob=squish,
                         na.value = "grey") +
    labs(title = expression(atop("Incidence Rate Ratio Comparing Non-Hispanic Black",
                                 "and Non Hispanic White Populations")),
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                   "5-Year ACS files from end-years 2015-2019")),
         fill = "IRR", x="", y="") +
    theme_void() +
    theme(legend.position = c(0.25, 0.25),
```

So we can see from this crude map that, in areas where there were Black deaths, there was some variation in the IRR, and some very high, concerning IRR values.



7.4.3 What are the associations with ABSM by racialized group?

In order to visualize differences in how our ABSM - ICE - interacts with racialized group, we can use a poisson model for the mortality rates, and include an interaction term between racialized group and poverty. We can plot this using `INLA' again or alternatively use a generalized linear model function. We

will have to recreate our indirect standardization - this time by race, in order to include both variables in the model.

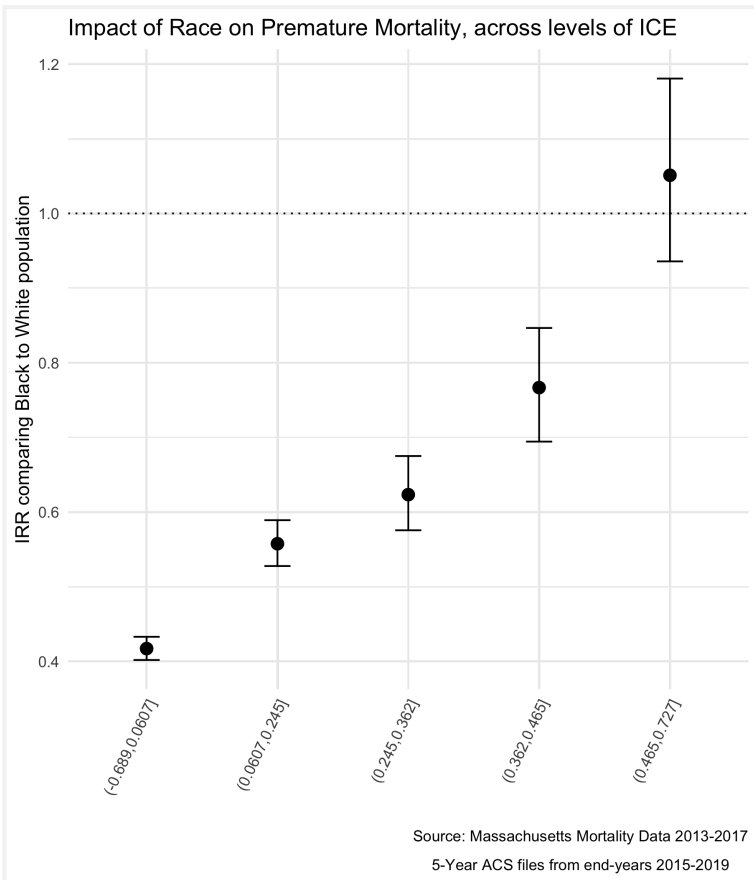
```
# Indirect standardization to prepare the data for spatial models
ma_mort_istd_byrace <- ma_mort_ct %>%
  filter(race_group != "Total",
         sex != "Total",
         age_cat != "Total") %>%
  mutate(race_group = ifelse(race_group == "Non-Hispanic Black", "Black", race_group)) %>%
  filter(race_group %in% c("Non-Hispanic White","Black")) %>%
  left_join(ma_denominator, by = c("GEOID10","year","age_cat","race_group","sex")) %>%
  mutate(age_cat = case_when(age_cat %in% c("35-39","40-44") ~ "35-44",
                             age_cat %in% c("45-49","50-54") ~ "45-54",
                             age_cat %in% c("55-59","60-64") ~ "55-64",
                             TRUE ~ age_cat)) %>%
  group_by(GEOID10, super_town, age_cat, race_group) %>%
  summarise(deaths = sum(deaths, na.rm = TRUE),
            population = sum(population, na.rm = TRUE))%>%
  right_join(reference_rates, by="age_cat") %>%
  mutate(expected = ref_rate*population) %>%
  group_by(GEOID10, super_town, race_group) %>%
  summarise(O = sum(deaths),
            E = sum(expected),
            raw_smr = O/E,
            var_raw_smr = O/E^2,
```

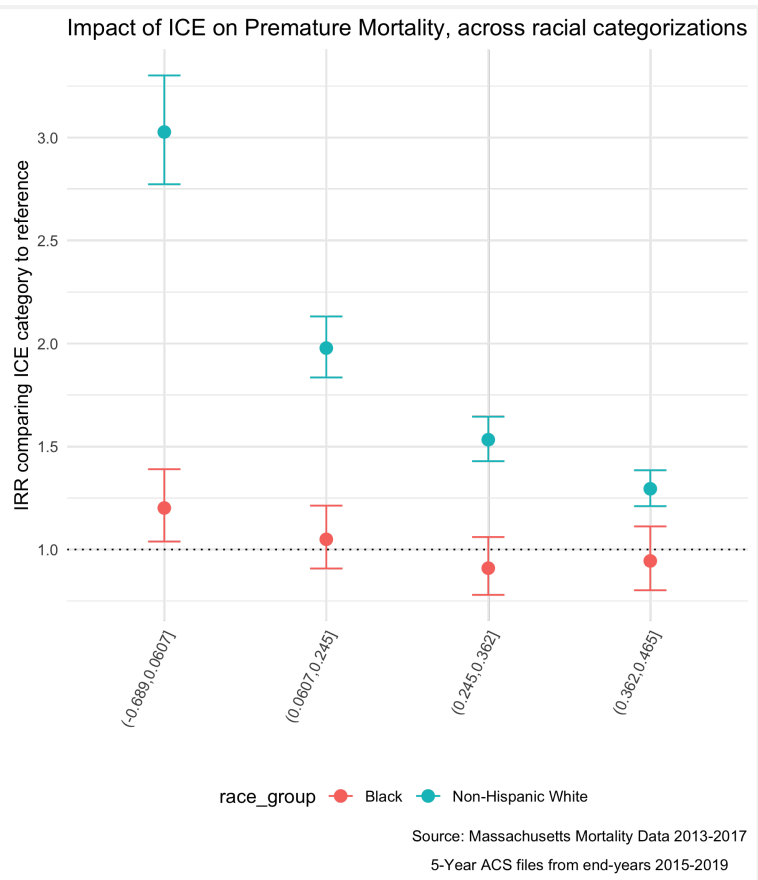
```
# Make sure that the data file has areas in the same order as the shape file
ma_mort_istd_byrace_ordered <- ma_mort_istd_byrace %>%
  filter(!(O == 0 & E == 0)) %>%
  mutate(intercept = 1) %>%
  left_join(tract_geometry, by="GEOID10") %>%
  select(-geometry) %>%
  arrange(id_order) %>%
  mutate(race_group = factor(race_group, levels = c("Non-Hispanic White", "Black")),
         ICE_qt = factor(ICE_qt, levels=c("(0.465,0.727]",
                                          "(0.362,0.465]",
                                          "(0.245,0.362]",
                                          "(0.0607,0.245]",
                                          "(-0.689,0.0607]")))
```

```
# BYM model with ICE-race interaction
model_form_2 <- O ~ 1 + f(id_order, model='bym2', graph=INLA_adj_mat, scale.model=TRUE,
constr=TRUE) + factor(race_group)*factor(ICE_qt)
#The following code outlines the linear combinations our model will make, so that we can
calculate intersectional inequities
linear_combinations <- inla.make.lincombs(
"factor(race_group)Black"                          = c(1,1,1,1,0,0,0,0,1,1,1,1),
"factor(ICE_qt)(0.362,0.465]"                      = c(0,0,0,0,1,0,0,0,1,0,0,0),
"factor(ICE_qt)(0.245,0.362]"                      = c(0,0,0,0,0,1,0,0,0,1,0,0),
"factor(ICE_qt)(0.0607,0.245]"                     = c(0,0,0,0,0,0,1,0,0,0,1,0),
"factor(ICE_qt)(-0.689,0.0607]"                    = c(0,0,0,0,0,0,0,1,0,0,0,1),
"factor(race_group)Black:factor(ICE_qt)(0.362,0.465]"   = c(1,0,0,0,1,0,0,0,1,0,0,0),
"factor(race_group)Black:factor(ICE_qt)(0.245,0.362]"   = c(0,1,0,0,0,1,0,0,0,1,0,0),
"factor(race_group)Black:factor(ICE_qt)(0.0607,0.245]"  = c(0,0,1,0,0,0,1,0,0,0,1,0),
"factor(race_group)Black:factor(ICE_qt)(-0.689,0.0607]" = c(0,0,0,1,0,0,0,1,0,0,0,1))
model_results_lc <- inla(model_form_2,
                         family="poisson",
                         data=ma_mort_istd_byrace_ordered, E=E,
                         control.predictor=list(compute=TRUE),
                         control.compute=list(dic=TRUE, waic=TRUE),
                         lincomb = linear_combinations)
# Function to extract the fixed results
```

We can look at some of these results in plots, to get a sense of the way that ICE is interacting with race to impact mortality in Massachusetts.

```r
# Plot black/white disparities within ICE categories
race_effect_within_ICEqt <- interaction_effects_bym %>%
  filter(which == "byICE_qt") %>%
  ggplot(aes(x=qval,
             y=exp(estimate),
             ymin=exp(conf.low),
             ymax=exp(conf.high))) +
  geom_point(position=position_dodge(width=0.5), size=3) +
  geom_errorbar(width=0.2, position = position_dodge(width=0.5)) +
  geom_hline(yintercept=1, linetype="dotted") +
  scale_color_brewer(palette="Set1") +
  labs(title = "Impact of Race on Premature Mortality, across levels of ICE",
       caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                 "5-Year ACS files from end-years 2015-2019")),
       x = "", y = "IRR comparing Black to White population") +
  theme_minimal() +
  theme(legend.position="bottom",
        axis.text.x=element_text(angle=65, hjust=1, vjust=1))
# Plot ABSM disparities within racial/ethnic categories, for aggregated method and Poisson
robust
ICEqt_effect_within_race <- interaction_effects_bym %>%
  filter(which=="byRace") %>%
```





We can see from these plots that as areas become more heavily concentrated with Non-Hispanic White, high-income individuals, the impact of racialized group on Premature Mortality increases. In populations that are heavily concentrated with low-income People of Color (POC), the mortality rates for Non-Hispanic White individuals is higher than that of Black individuals. At the other end of the spectrum the results are insignificant (possibly because Massachusetts has less census tracts in that category) but trend the other direction.

When we look at the impact of ICE within racial categories we actually see very little impact for the Black population, with the impact of ICE only significantly different in the most concentrated low-income POC category. But for the Non-Hispanic White population we see larger differences from the reference, and a clear pattern where in Non-Hispanic White individuals fare much worse in neighborhoods with more low-income POC.

Why would this be? What have we learned about the differential impact of neighborhood context by individual racialized group membership?

7.5 Appendix: Wrangling Your Mortality Data

Below is some code (for example) to turn a geocoded mortality file into what has been provided to you, by merging census tracts and towns to the data, and then aggregating into groups that align with the age and racialized groups that match your denominator data.

```
# Importing the raw mortality data, with the geocoded coordinates, converted into spatial
feature (sf) data
ma_mort <- read_rds("ma_mort_geo.RDS") %>%
  st_as_sf(coords=c("lon","lat"), crs = 4269)
# Importing a MA tract-level shapefile using tigris package
tracts_sf <- tracts(state = "25", year = "2010")
# Merging the Census Tracts to our crosswalk of towns so we know which ones need to be combined
supertowns <- read_excel("ma ct to supertinytowns.xlsx", col_names = FALSE, col_types = "text")
%>%
  transmute(GEOID10 = ...1,
            super_town = str_to_lower(...2))
tracts_sf <- left_join(tracts_sf,supertowns, by = "GEOID10")
# Merging to geocoded mortality and dropping geometries
# The geometries (actual stored location information) of a shapefile can be very unwieldy and
slow data wrangling commands. Since we only need them to visualize our maps, we will remove
them for now
ma_mort_agg <- st_join(tracts_sf, ma_mort, left = FALSE) %>%
  st_drop_geometry() %>%
  arrange(id) %>%
  mutate(super_town = case_when(super_town == "stt1"   ~ "Super Town 1",
                                super_town == "stt2"   ~ "Super Town 2",
                                super_town == "stt3"   ~ "Super Town 3",
```

« 6 Analyzing your data

8 Case Study 2: Breast Cancer Mortality in Massachusetts »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

8 Case Study 2: Breast Cancer Mortality in Massachusetts

By: Enjoli Hall

8.1 Overview

Our outcome of interest is breast cancer mortality. We investigate disparities in breast cancer mortality by racialized/ethnic group and socioeconomic position. We primarily focus on comparisons of risk in Black non-Hispanic and White non-Hispanic populations, motivated by longstanding health inequities between these groups and sufficiently large population sizes in Massachusetts to support the analyses. We will be working with mortality data from the Massachusetts Registry of Vital Records and Statistics [https://www.mass.gov/lists/death-data] for the years 2013-2017. Vital statistics and disease registries are how states (which submit their data to the federal government) keep an official enumeration of deaths (Friedman et al., 2005; Hetzel, 1997; Krieger, 2019); because the death certificates require data on residential address at time of death, they can be useful tools for monitoring trends in health and health equity. Having data on the residential address of each individual allows us to geocode each observation to the physical and social environment in which the person lived. Death records also typically include demographic information about the deceased, including categories for racialized groups that conform to the 1997 US Office of Management and Budget (OMB) standards for the classification of federal data on "race and ethnicity" (OMB, 1997). We will explore how social membership in racialized groups and area-based social metrics (ABSMs) might be associated with breast cancer mortality across Massachusetts. To do this, we will pair our mortality data with demographic and socioeconomic data from the U.S. Census Bureau's 5-year American Community Survey (ACS) files.

8.2 Background and Significance

Breast cancer is the most commonly diagnosed cancer worldwide and is one of the leading causes of cancer death. In the United States, breast cancer incidence and mortality rates vary widely across geographic regions and racialized groups. For example, while breast cancer deaths in the United States have declined overall by 42 percent over the last 30 years, there is a persistent mortality gap between Black women and White women. Breast cancer incidence rates among Black and White women are close, but mortality rates are markedly different, with Black women having a 41 percent higher death rate from breast cancer compared to White women in the United States (Breast Cancer Research Foundation, 2022).

Nevertheless, the routine stratification and presentation of cancer data by "race" in the absence of socioeconomic data such as occupation, educational level, or income, perpetuates the view that "race"—wrongly construed as a biological variable—explains racialized inequities in breast cancer mortality and other cancer outcomes. Hidden from view are ways that economic forms of discrimination and inequality might drive "racial"/ethnic inequities in breast cancer mortality. Rather than taking an either/or approach to analyzing and interpreting cancer data by "race/ethnicity" and socioeconomic position, it is important for public health researchers to stratify cancer data by both "race/ethnicity" and socioeconomic position as neither by itself is sufficient to capture how membership in racialized groups and class relations, separately and together, affect the health of populations. However, U.S. public health surveillance systems typically do not include data relating health status to socioeconomic position for individual records. One possible solution to these gaps is to combine data from health surveillance systems with socioeconomic data derived from the U.S. Census to analyze breast cancer mortality in relation to area-based socioeconomic measures for domains of socioeconomic position such as income and poverty, thereby permitting calculation of population-based breast cancer mortality rates stratified by area-based socioeconomic position and making visible socioeconomic gradients in breast cancer mortality.

We can also monitor racialized and socioeconomic cancer and other health inequities using not only conventional individual- and area-level socioeconomic measures but also measures of racialized and economic segregation and polarization at the neighborhood, city or town, and regional levels; these latter

On this page

measures bring into focus the full range of concentrations of privilege and deprivation in an area. Over 20 years ago, Williams and Collins (2001) explained how racial residential segregation acts as a fundamental cause of racial disparities in health by exposing Black communities to less healthy neighborhood and housing conditions, fewer economic and educational opportunities, and lower quality health care resources compared to White communities. Since that time, hundreds of empirical studies have examined segregation as a key driver of patterns of population health and health inequities, but relatively few studies have examined cancer outcomes. Together, these studies suggest that residential segregation can generate racialized economic inequities across the cancer continuum, in part through producing differential access to medical care and unequal exposures to social and environmental cancer risks.

8.3 Motivation, Research Questions, and Learning Objectives

The goal of this case study is to develop familiarity with methods of exploring and visualizing racialized and socioeconomic inequities in health. Our specific goals will be to:

- Download and merge different datasets for our analysis
- Visualize and map estimates of area-based social metrics and breast cancer mortality
- Identify relationships between racialized group, area-based social metrics, and breast cancer mortality
- Model how space may impact inequities in breast cancer mortality

The research questions we will seek to answer throughout this case study include:

1. What is the overall socioeconomic gradient in breast cancer mortality? (hint: we can visualize this with a spatial model)
2. What is the racialized disparity in breast cancer mortality overall? (hint: we can visualize this with stratified aggregate analyses)
3. How do area-based socioeconomic measures interact with individual-level membership in racialized groups to affect patterns of breast cancer mortality (i.e., interactions between socioeconomic position and racialized groups, not just socioeconomic inequities within racialized groups)? (hint: we can explore this with a Poisson model using an interaction term)

8.4 Getting and Wrangling Your Data

We are providing datasets for you to use throughout these case studies that have been wrangled and reshaped for your use, and we also provide code to show how you could go through that process on your own. You can look at whole datasets in RStudio using the `view()` command, or look at summaries of the datasets by simply typing the dataset name into the console window. You can skip ahead to the "Approach" section and follow along with the case study without issue.

8.4.1 Running RStudio and Setting Up Your Working Directory

Download the Breast Cancer Mortality project folder [https://hu-my.sharepoint.com/:f:/g/personal/denajavadi_g_harvard_edu/EpyUmRZB-hBGvdygLgRLTmMByYr_lmlP6pqY09f_bz8QBg?e=tDxIIg] that contains all of the data files and geographic shapefiles as well as maps and figures that we will use for this case study. Save the folder to your Desktop or another easily accessible location on your computer. Note: Do not edit or change any of the file names as our code corresponds to the file names in the folder.

Next, you will need to open RStudio and create a new R Project file to work on the case study. Create a new R Project file by selecting `File > New Project…` from the menu bar. Select `New Directory` from the popup window. Next, select `New Project` . Pick a meaningful name for your project folder, i.e. the `Directory Name` . Ensure this project folder is created in the right place. You can change the `subdirectory` by clicking on `Browse…` . The subdirectory should be the place where the Breast Cancer Mortality folder and files that you just downloaded are saved on your computer. Lastly, tick `Open in new session` . This will open your R Project in a new RStudio window. Once you are happy with your choices, you can click `Create Project` . This will open a new R Session, and you can start working on the case study. To make sure all of your project files for the case study are properly loaded and to avoid potential errors when running the

code, navigate to your project folder and files in the `Files/Plots/Packages/Help` window in the bottom-right corner of your RStudio window and go through and double-click each of the file names to open and run all of the files for this project.

8.4.2 Dependencies

Run the lines of code below to load the R packages that you will need throughout the case study. If this code does not run for you, you may need to run `install.packages("package_name")` .

```
# Libraries - if this code does not run for you, you may need to run
install.package("package_name")
library(knitr)
library(tidyverse)
library(readxl)
library(ggplot2)
library(cowplot)
library(tidycensus)
library(tigris)
options(tigris_use_cache = TRUE)
library(sf)
library(spdep)
library(viridis)
library(Hmisc)
library(fastDummies)
library(lme4)
library(INLA)
library(broom)
```

8.4.3 Health Surveillance Data

Data on breast cancer deaths were obtained from the Massachusetts Department of Public Health. Each mortality record included data on the decedent's age, gender, racialized group (using U.S. census categories), residential address at the time of death and coded cause of death following the International Classification of Diseases 10th Revision (ICD-10). We employed R software and Google Maps API to geocode the residential address of each case to its latitude and longitude, which were used to assign census tract and city/town geocodes.

The mortality data has been aggregated from individual observations into death counts by age, racialized group, census tract, and city/town. When you get unrestricted mortality files from government sources for research, you will likely receive files with one observation per death. After you have geocoded these observations, you will need to aggregate them up to the level of interest for your analysis. This requires aggregation not only up to the census tract level, but also age groups so we can do an appropriate age standardization, sex groups and racialized groups so we can stratify our analyses by these groups, and towns so we can explore a second areal level of analysis.

Below is some code (for example) to turn a raw mortality file into what has been provided to you, by merging census tracts and towns to the data, and then aggregating into groups. The vast majority of our area-based social metrics come from the U.S. Census, so the most detailed level for analysis available will be the census tract (however for this case study, because we are analyzing a health outcome with relatively small numbers of cases, we need to perform our analysis at a larger geography such as city/town instead of census tract to ensure that there are a sufficient number of cases, especially inclusive of different racialized groups, in our units of analysis). We can use the `tigris` package to download census tract geometries and the sf package to link our geocoded observations to the appropriate census tracts.

Ideally when performing an analysis that might include two or more levels the smaller level (e.g., census tracts) would be nested entirely within the larger level (e.g., towns). In Massachusetts though, there are many towns that have such small populations that they are smaller than census tracts. For these towns, we have created a crosswalk wherein small towns have been combined to create larger super towns. These super towns each make up one census tract, so that each super town in the analysis will now have at least one census tract nested within it. Please note that for this analysis we will be focusing on breast cancer mortality in women, designated as such in the MA cancer registry records.

```r
# the raw mortality data, with the geocoded coordinates converted into spatial feature (sf)
data
ma_mort <- read_rds("data/08-breast-cancer-mortality/ma_mort_geo.RDS") %>%
  st_as_sf(coords=c("lon","lat"), crs = 4269)

# Importing tract shapefile using tigris package
tracts_sf <- tracts(state = "25", year = "2010")

# Linking Census Tracts and Super Towns File
supertowns <- read_excel("data/08-breast-cancer-mortality/ma ct to supertinytowns.xlsx",
col_names = FALSE, col_types = "text") %>%
  transmute(GEOID10 = ...1,
            super_town = str_to_lower(...2)) %>%
  mutate(super_town = case_when(GEOID10   == "25023500101" ~ "hull",
                                super_town == "stt1"       ~ "Super Town 1",
                                super_town == "stt2"       ~ "Super Town 2",
                                super_town == "stt3"       ~ "Super Town 3",
                                super_town == "stt4"       ~ "Super Town 4",
                                super_town == "stt5"       ~ "Super Town 5",
                                super_town == "stt6"       ~ "Super Town 6",
                                super_town == "stt7"       ~ "Super Town 7",
                                super_town == "stt8"       ~ "Super Town 8",
```

```
## # A tibble: 6 × 6
##   super_town  year age_cat sex    race_group         deaths
##   <chr>      <dbl> <fct>   <fct>  <fct>               <int>
## 1 abington    2013 70-74   Female Non-Hispanic White      1
## 2 abington    2014 75-79   Female Non-Hispanic White      1
## 3 abington    2016 40-44   Female Non-Hispanic White      2
## 4 abington    2016 50-54   Female Non-Hispanic White      1
## 5 abington    2016 85+     Female Non-Hispanic White      1
## 6 abington    2017 70-74   Female Non-Hispanic White      1
```

8.4.4 Population Denominator Data and Area-Based Social Metric Data

We download population data to use as denominators in our rates and area-based social metrics (ABSMs) from the U.S. Census Bureau using the `tidycensus` package. This requires registering with the U.S. Census for an API key. The key is redacted here, but you can get your own from https://api.census.gov/data/key_signup.html. For this case study, we are using the 2015-2019 American Community Survey (ACS) 5-year estimates counts for our population estimates and for constructing our ABSMs.

The ABSMs we will use for this analysis are the percentage of the population in poverty (for more information on how the Census Bureau measures poverty, see https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html) and the Index of Concentration at the Extremes (ICE) for racialized economic segregation (i.e., race/ethnicity + income), which measures the extent to which an area's population is concentrated into extremes of deprivation and privilege. The ICE measure for racialized economic segregation is scaled from -1 to 1: a value of -1 means that 100% of the population is concentrated in the most deprived group (in this analysis, conceptualized as the Black non-Hispanic population in low-income households), and a value of 1 means that 100% of the population is concentrated into the most privileged group (in this analysis, conceptualized as the White non-Hispanic population in high-income households). For more information on the formula for the ICE measure and the construction of the specific ICE variables, see Krieger et al., 2016. Both of our area-based social metrics (poverty and racialized economic segregation) are calculated at the city/town level.

```r
# Population Denominators
ma_demo_acs_bc <- vector(mode = "list", length = 5)
names(ma_demo_acs_bc) <- c(2013,2014,2015,2016,2017)
for (nm in names(ma_demo_acs_bc)) {
  ma_demo_acs_bc[[nm]] <- get_acs(geography = "tract",
                                  # These are the myriad variables that make up the variables we
intend to use
                                  variables = c("B01001_027","B01001_028","B01001_029",

"B01001_030","B01001_031","B01001_032","B01001_033",

"B01001_034","B01001_035","B01001_036","B01001_037",

"B01001_038","B01001_039","B01001_040","B01001_041",

"B01001_042","B01001_043","B01001_044","B01001_045",
                                                  "B01001_046","B01001_047",
"B01001_048","B01001_049",
                                                  "B01001B_017","B01001B_018","B01001B_019",

"B01001B_020","B01001B_021","B01001B_022","B01001B_023",
```

```
## # A tibble: 6 × 6
##   super_town  year age_cat sex    race_group
population
##   <chr>      <dbl> <chr>   <chr>  <chr>
<dbl>
## 1 abington    2013 0-4     Female Hispanic
45
## 2 abington    2013 0-4     Female Non-Hispanic Asian or Pacific Islander
0
## 3 abington    2013 0-4     Female Non-Hispanic Black
45
## 4 abington    2013 0-4     Female Non-Hispanic Native American, Alaskan Native, Other
0
## 5 abington    2013 0-4     Female Non-Hispanic White
456
## 6 abington    2013 0-4     Female Total
529
```

```r
# Area-Based Social Metrics
ma_absm_acs <- vector(mode = "list", length = 5)
names(ma_absm_acs) <- c(2013,2014,2015,2016,2017)
for (nm in names(ma_absm_acs)) {
  ma_absm_acs[[nm]] <- get_acs(geography = "tract",
                               variables =
c("B01003_001E","B02001_001E","B02001_002E","B02001_003E",

"B02001_004E","B02001_005E","B02001_006E","B02001_007E",

"B02001_008E","B02001_009E","B02001_010E","B03001_001E",
                                                  "B03001_003","B01001H_001E"),
                               year = as.numeric(nm) + 2,
                               output = "wide",
                               state = "MA",
                               geometry = FALSE,
                               moe_level = 95,
                               survey = "acs5",
                               cache_table = TRUE) %>%
    # Transforming ACS variables into the ABSMs we want to use for our dataset
    mutate(GEOID10 = GEOID,
           percBlack = B02001_003E/B02001_001E,
```

```
## # A tibble: 6 × 19
##   super_town percBlack percHisp percColor ICEwbinc ICEwnhinc tractPov tractCrowd
tractSevereCrowd
##   <chr>          <dbl>    <dbl>     <dbl>    <dbl>     <dbl>    <dbl>      <dbl>
<dbl>
## 1 abington    0.0243     0.0232    0.101     0.378     0.373   0.0384    0.0120
0.00224
## 2 acton       0.0155     0.0236    0.295     0.465     0.448   0.0307    0.00933
0.00130
## 3 acushnet    0.000574   0.0331    0.0926    0.309     0.303   0.0546    0.0380
0.00675
## 4 adams       0.00715    0.0121    0.0379    0.183     0.179   0.104     0.0103
0.00124
## 5 agawam      0.0155     0.0609    0.117     0.300     0.271   0.0822    0.0126
0.00606
## 6 amesbury    0.00675    0.0239    0.0668    0.368     0.356   0.0552    0.0111
0.000323
## # … with 10 more variables: ice_wnh_highinc <dbl>, ice_wnh_lowinc <dbl>, ice_poc_lowinc <dbl>,
## #   ice_poc_highinc <dbl>, pop_total <dbl>, pov_cat <fct>, pov_qt <fct>, ICE_qt <fct>,
perc_Color_qt <fct>,
## #   perc_BLACK_qt <fct>
```

8.5 Approach

Now that we have our data, let's revisit our questions of interest: 1. What is the overall socioeconomic gradient in breast cancer mortality?

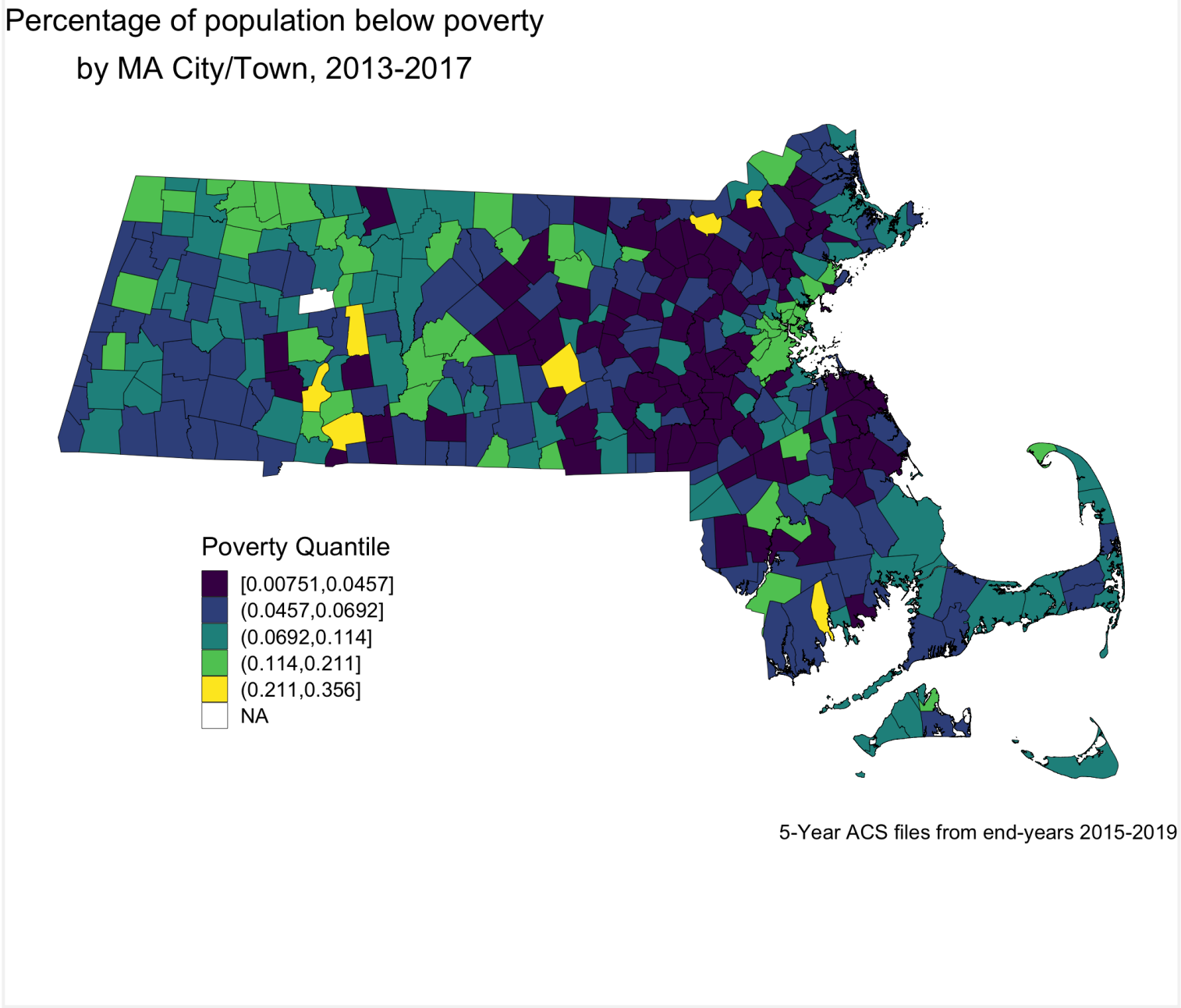
2. What is the racialized disparity in breast cancer mortality overall?

3. How do area-based socioeconomic measures interact with individual-level membership in racialized groups to affect patterns of breast cancer mortality (i.e., interactions between socioeconomic position and racialized groups, not just socioeconomic inequities within racialized groups)?

8.5.1 Question 1: What is the overall socioeconomic gradient in breast cancer mortality?

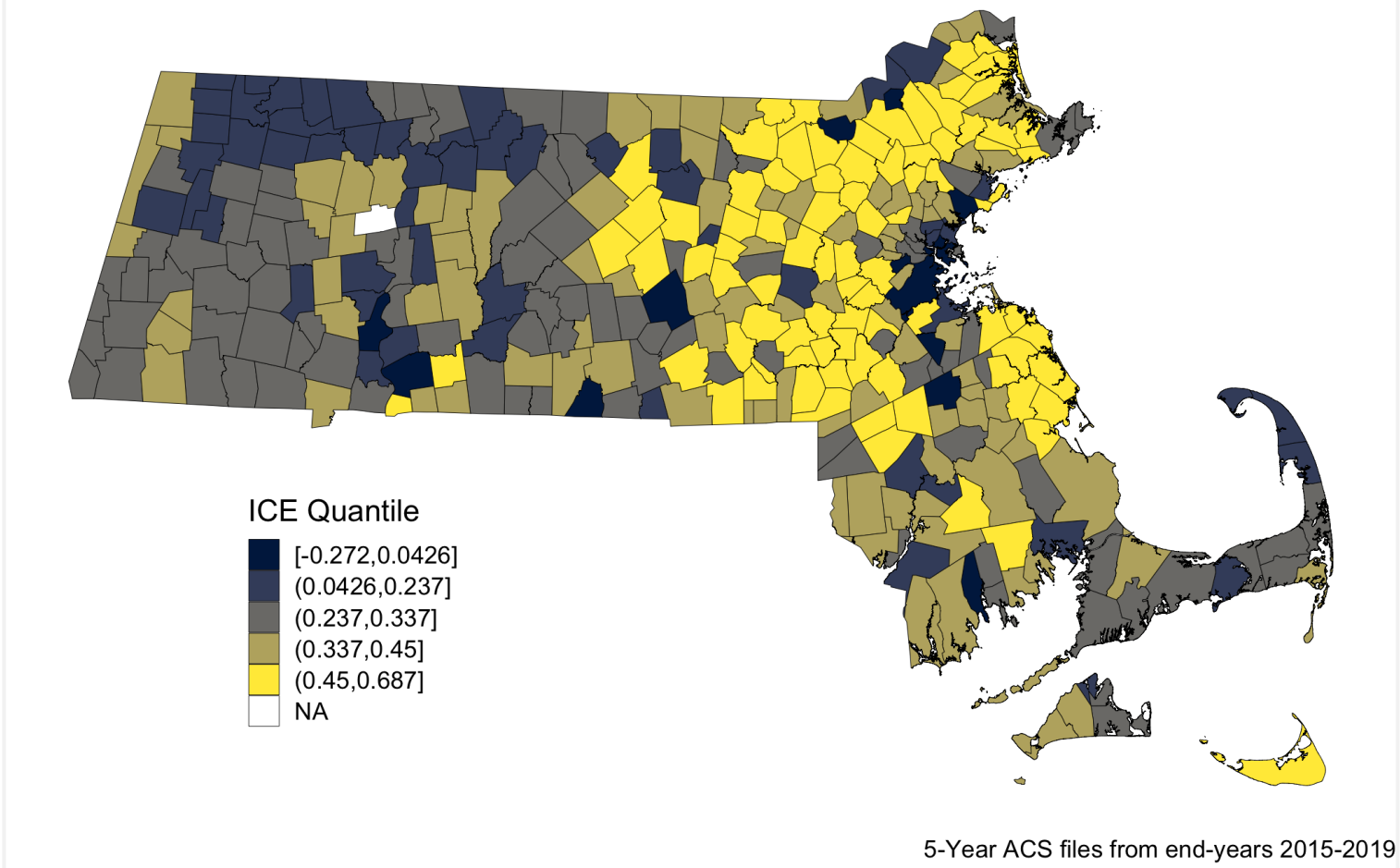
First, let's just look at the spatial distribution of the area-based socioeconomic measure of poverty that we constructed: the percentage of the city/town population below poverty. Here is a map of the percentage of population below poverty by Massachusetts city/town poverty quantile for years 2013-2017.

```
ma_poverty_map <- town_ma_absm_sum %>%
  left_join(town_geometry, by= "super_town") %>%
  ggplot() +
  geom_sf(mapping = aes(geometry=geometry,
                        fill=pov_qt),
          lwd = 0.1,
          color = "black") +
  scale_fill_viridis_d() +
  labs(title = expression(atop("Percentage of population below poverty", "by MA city/town, 2013-
2017")),
       caption = "5-year ACS files from end-years 2015-2019",
       fill = "Poverty quantile", x="", y="") +
  theme_void() +
  theme(axis.text.x=element_blank(), #remove x axis labels
        axis.ticks.x=element_blank(), #remove x axis ticks
        axis.text.y=element_blank(),  #remove y axis labels
        axis.ticks.y=element_blank(),
        legend.position = c(0.25, 0.25),
        legend.key.size = unit(0.4, "cm"))

ggsave("ma_poverty_map.png")
```



Now, let's look at the spatial distribution of the other area-based socioeconomic measure that we constructed: the ICE measure for racialized economic segregation. Compare this map of city/town ICE (racialized group + income) quantile to the first map of city/town poverty quantile. Do you notice any differences? Do you find one of these area-based socioeconomic measures to be more visually clear and compelling than the other, and why or why not?

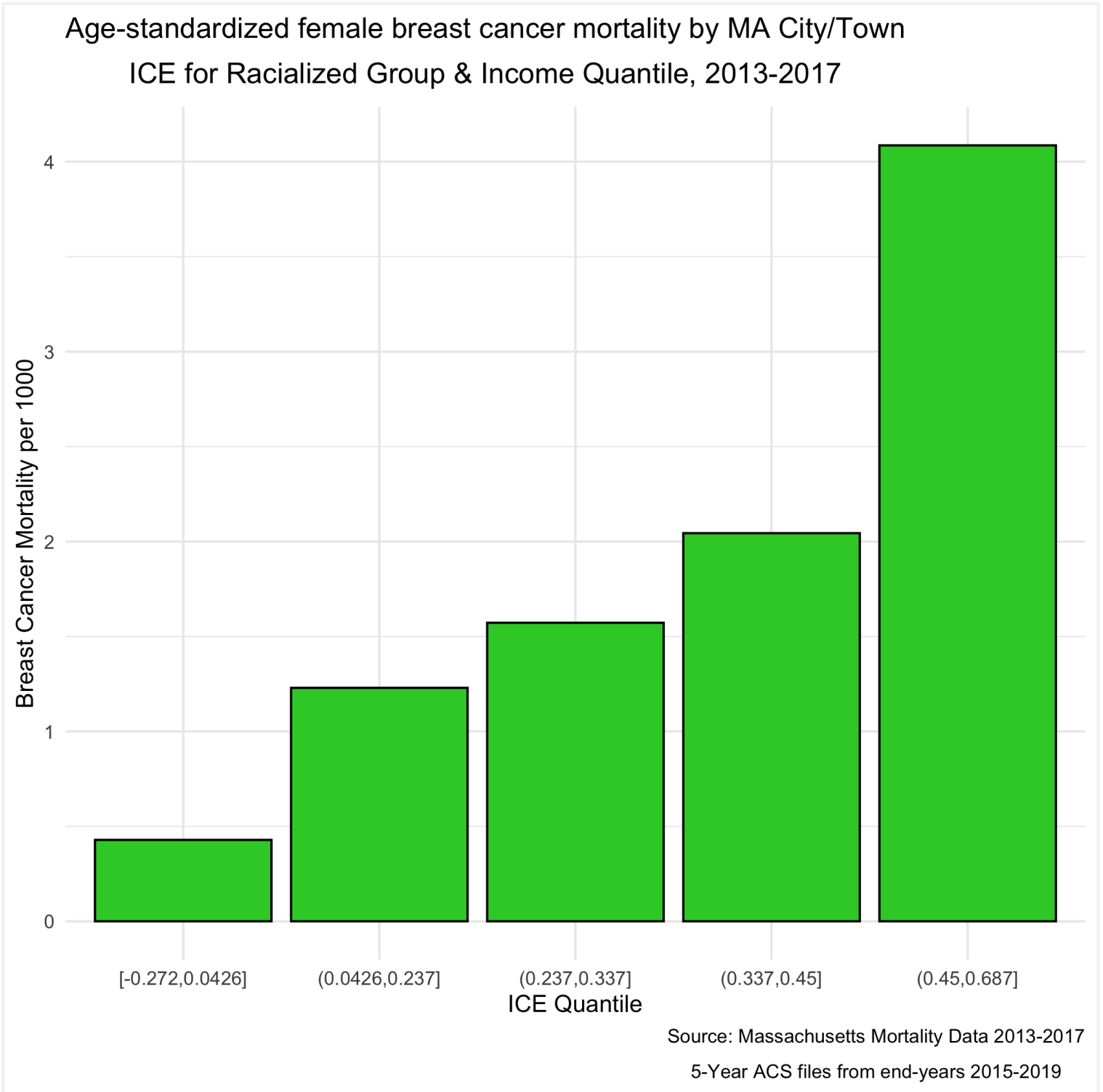
```
ma_ice_map <- town_ma_absm_sum %>%
  left_join(town_geometry, by= "super_town") %>%
  ggplot() +
  geom_sf(mapping = aes(geometry=geometry,
                        fill=ICE_qt),
          lwd = 0.1,
          color = "black") +
  scale_fill_viridis_d(option = "E") +
  labs(title = expression(atop("Index of Concentration at the Extremes for Racialized Economic
Segregation", "by MA city/town, 2013-2017")),
       caption = "5-year ACS files from end-years 2015-2019",
       fill = "ICE quantile", x="", y="") +
  theme_void() +
  theme(axis.text.x=element_blank(), #remove x axis labels
        axis.ticks.x=element_blank(), #remove x axis ticks
        axis.text.y=element_blank(),  #remove y axis labels
        axis.ticks.y=element_blank(),
        legend.position = c(0.25, 0.25),
        legend.key.size = unit(0.4, "cm"))

ggsave("ma_ice_map.png")
```



Next, we will look at how breast cancer mortality varies across the ICE measure for racialized economic segregation. Using the 2015-2019 American Community Survey (ACS) 5-year estimates counts, we compute age- and sex-standardized 2013-2017 breast cancer mortality rates (per 1000) for each MA city/town ICE quantile using the direct method. This method will require a reference population, which we have downloaded (and wrangled) from the National Cancer Institute. (https://seer.cancer.gov/stdpopulations/).

```
## # A tibble: 13 × 3
##    age_cat    pop     wt
##    <chr>    <dbl>  <dbl>
##  1 0-4      69135 0.0691
##  2 10-14    73032 0.0730
##  3 15-19    72169 0.0722
##  4 20-24    66478 0.0665
##  5 25-29    64529 0.0645
##  6 30-34    71044 0.0710
##  7 35-44   162613 0.163
##  8 45-54   134834 0.135
##  9 5-9      72533 0.0725
## 10 55-64    87247 0.0872
## 11 65-74    66037 0.0660
## 12 75-84    44841 0.0448
## 13 85+      15508 0.0155
```

```
ma_mort_ice_qt <- town_ma_mort_bc %>%
  inner_join(town_ma_demo_acs_bc, by = c("year","super_town","race_group","sex","age_cat")) %>%
  left_join(town_ma_absm_sum, by= "super_town") %>%
  filter(age_cat != "Total") %>%
  group_by(ICE_qt, age_cat) %>%
  summarise(num = sum(deaths, na.rm=T),
            den = sum(population, na.rm=T))%>%
  mutate(den = den + 0.001) %>%
  left_join(seer_std, by="age_cat") %>%
  mutate(rate_i = wt*num/den,
         var_rate_i = (num*wt^2)/den^2) %>%
  group_by(ICE_qt) %>%
  summarise(num = sum(num, na.rm=T),
            den = sum(den, na.rm=T),
            std_rate = sum(rate_i, na.rm=T),
            var_std_rate = sum(var_rate_i, na.rm=T),
            sumwt = sum(wt),
            sumwt2 = sum(wt^2)) %>%
  mutate(std_rate = std_rate / sumwt *1000,
         var_std_rate = var_std_rate / sumwt2 *1000,
         std_rate_lo95 = std_rate - 1.96*sqrt(var_std_rate),
         std_rate_up95 = std_rate + 1.96*sqrt(var_std_rate)) %>%
```



We can see a clear socioeconomic gradient in breast cancer mortality rates at the city/town level in Massachusetts for years 2013-2017. How would you interpret this socioeconomic gradient? Is it what you expected, why or why not?

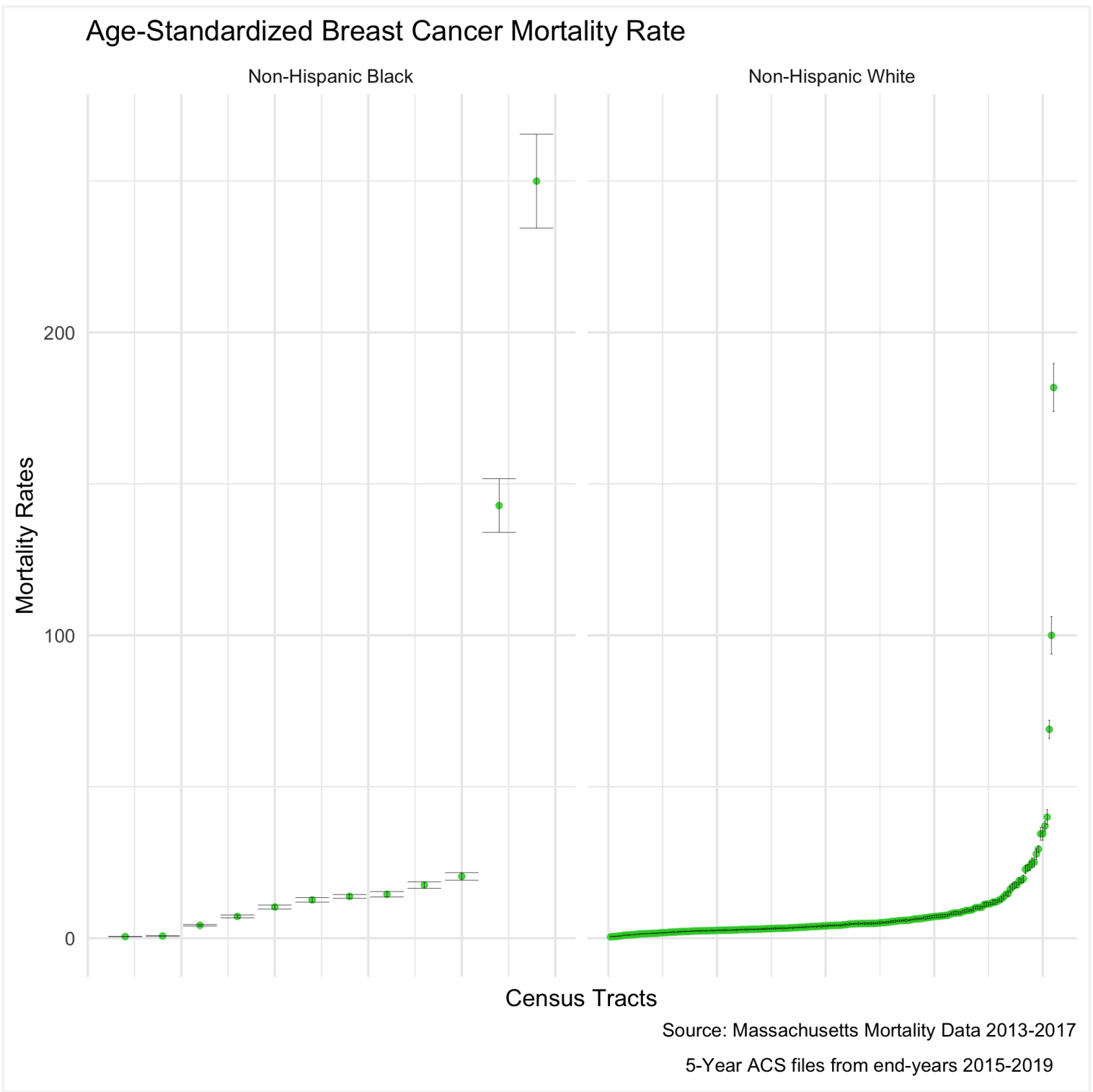
8.5.2 **Question 2: What is the racialized disparity in breast cancer mortality overall?**

We can describe inequities by racialized groups much the same way we describe inequities by ABSMs: by aggregating death and population data. We've already got this data aggregated, so we can stratify our analysis to compare Black non-Hispanic and White non-Hispanic populations. Note: Throughout the case

study we may refer to these groups as Black and White as a shorthand (see footnote in preface regarding capitalization conventions used throughout the manual).

As we did for the first question, we age-adjust our aggregated data by racialized group to compare standardized mortality rates for Black and White populations. We can see lots of variation here as the mortality rates increase - what is causing that?

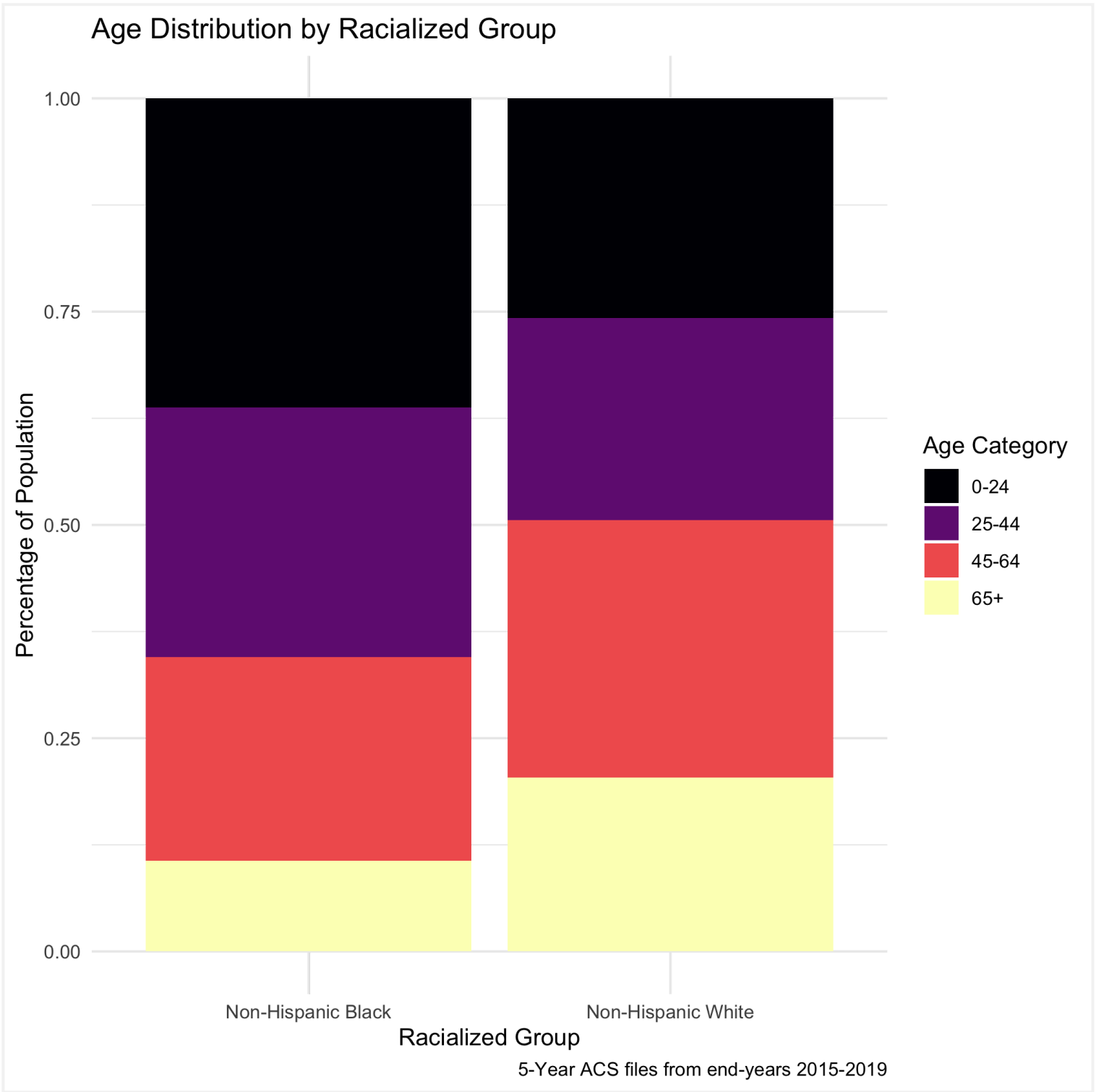
```
ma_mort_stratified <- town_ma_mort_bc %>%
  inner_join(town_ma_demo_acs_bc, by = c("year","super_town","race_group","sex","age_cat"))%>%
  filter(race_group %in% c("Non-Hispanic White","Non-Hispanic Black"),
         age_cat != "Total") %>%
  group_by(super_town, race_group, age_cat) %>%
  summarise(num = sum(deaths, na.rm=T),
            den = sum(population, na.rm=T))%>%
  mutate(den = den + 0.001) %>%
  left_join(seer_std, by="age_cat") %>%
  mutate(rate_i = wt*num/den,
         var_rate_i = (num*wt^2)/den^2) %>%
  group_by(super_town, race_group) %>%
  summarise(num = sum(num, na.rm=T),
            den = sum(den, na.rm=T),
            std_rate = sum(rate_i, na.rm=T),
            var_std_rate = sum(var_rate_i, na.rm=T),
            sumwt = sum(wt),
            sumwt2 = sum(wt^2)) %>%
mutate(std_rate = std_rate / sumwt *1000,
       var_std_rate = var_std_rate / sumwt2 *1000,
       std_rate_lo95 = std_rate - 1.96*sqrt(var_std_rate),
       std_rate_up95 = std_rate + 1.96*sqrt(var_std_rate)) %>%
```



Before we explore that variation, as a reminder, we age-adjust our aggregated data by racialized group so we can compare standardized mortality rates, as differences in breast cancer mortality between the two groups may be due to differential distribution of ages by racialized group. We can see that the age distributions for each racialized group are quite different. How would you describe the differences?

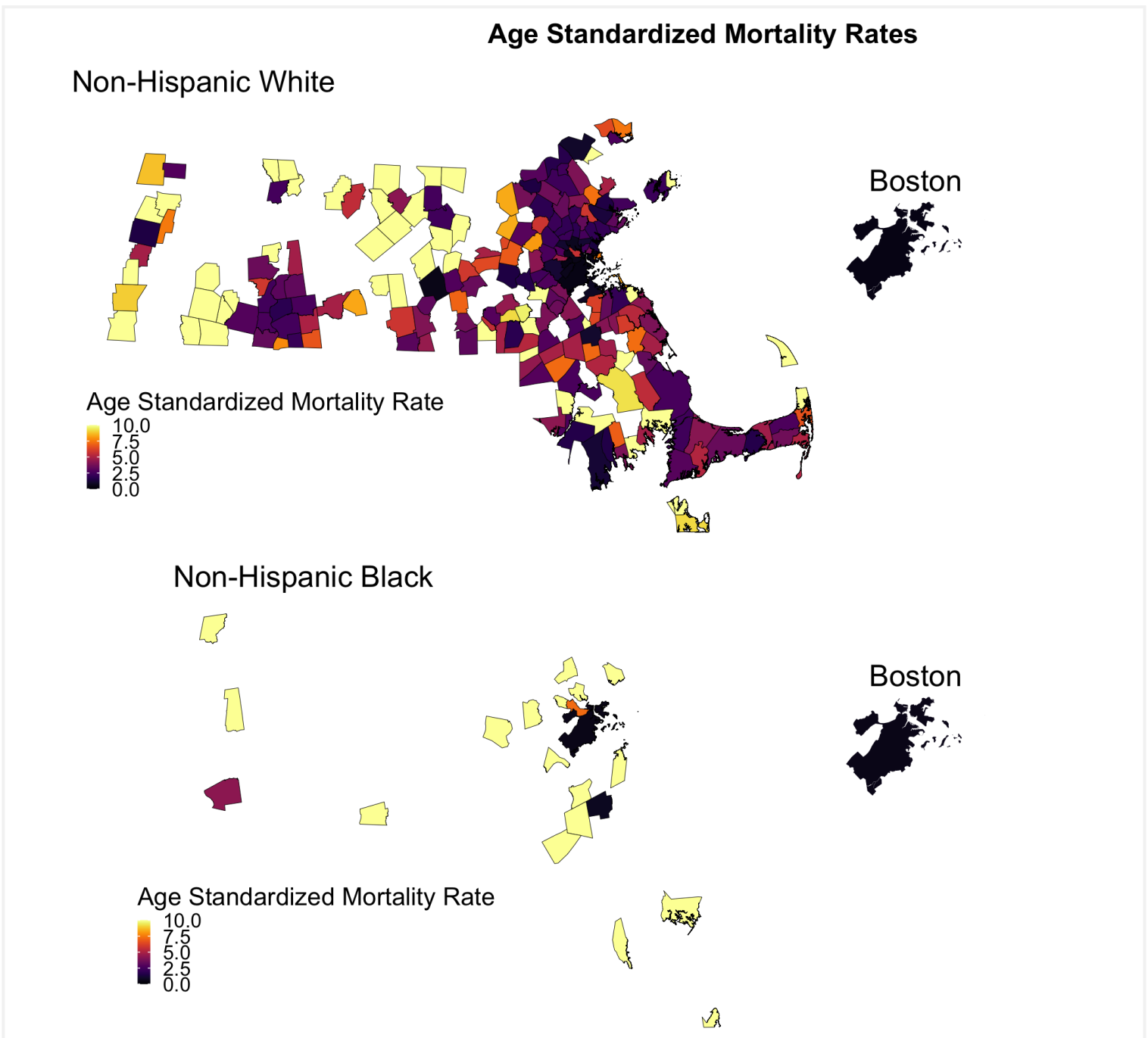
```
bc_tab_age_race <- town_ma_demo_acs_bc %>%
      filter(race_group %in% c("Non-Hispanic White","Non-Hispanic Black"),
         sex == "Female",
         age_cat != "Total") %>%
  group_by(super_town, race_group, sex, age_cat) %>%
  summarise(pop = sum(population, na.rm=T)) %>%
  inner_join(town_ma_absm_sum, by = c("super_town")) %>%
  mutate(age_cat_broad = case_when(age_cat %in% c("0-4","5-9","10-14","15-19","20-24") ~ "0-
24",
                                   age_cat %in% c("25-29","30-34","35-39","35-44","40-44") ~
"25-44",
                                   age_cat %in% c("45-49","45-54","50-54","55-59","55-64","60-
64") ~ "45-64",
                                   age_cat %in% c("65-69","65-74","70-74","75-59","75-84","80-
84","85+") ~ "65+")) %>%
  group_by(race_group, age_cat_broad) %>%
  summarise(pop = sum(pop, na.rm=T)) %>%
  group_by(race_group) %>%
  mutate(percentage = pop/sum(pop)) %>%
  ggplot(aes(x=race_group, y=percentage, fill= age_cat_broad)) +
    geom_bar(position="stack", stat="identity") +
    scale_fill_viridis_d(option = "A") +
```

```
include_graphics("images/08-breast-cancer-mortality/bc_tab_age_race.png")
```



Now, to return to this question of why we see lots of variation in mortality rates. As we might expect, these extremely variable mortality rates are occurring in small populations. What do these maps look like?

```
plotlist <- vector(mode = "list", length = 2)
names(plotlist) <- c("Non-Hispanic White","Non-Hispanic Black")
for (plt in names(plotlist)) {

  map.state <- ma_mort_stratified %>%
    filter(race_group == plt) %>%
    mutate(std_rate = ifelse(std_rate > 10, 10, std_rate)) %>%
    left_join(town_geometry, by= "super_town") %>%
    ggplot() +
      geom_sf(mapping = aes(geometry = geometry,
                            fill = std_rate),
              lwd = 0.1,
              color = "black") +
      scale_fill_viridis(option = "B", limits = c (0, 10)) +
      labs(title = plt,
           fill = "Age Standardized Mortality Rate", x="", y="") +
      theme_void() +
      theme(axis.text.x=element_blank(), #remove x axis labels
            axis.ticks.x=element_blank(), #remove x axis ticks
            axis.text.y=element_blank(),  #remove y axis labels
            axis.ticks.y=element_blank(),
            legend.position = c(0.25, 0.25),
```



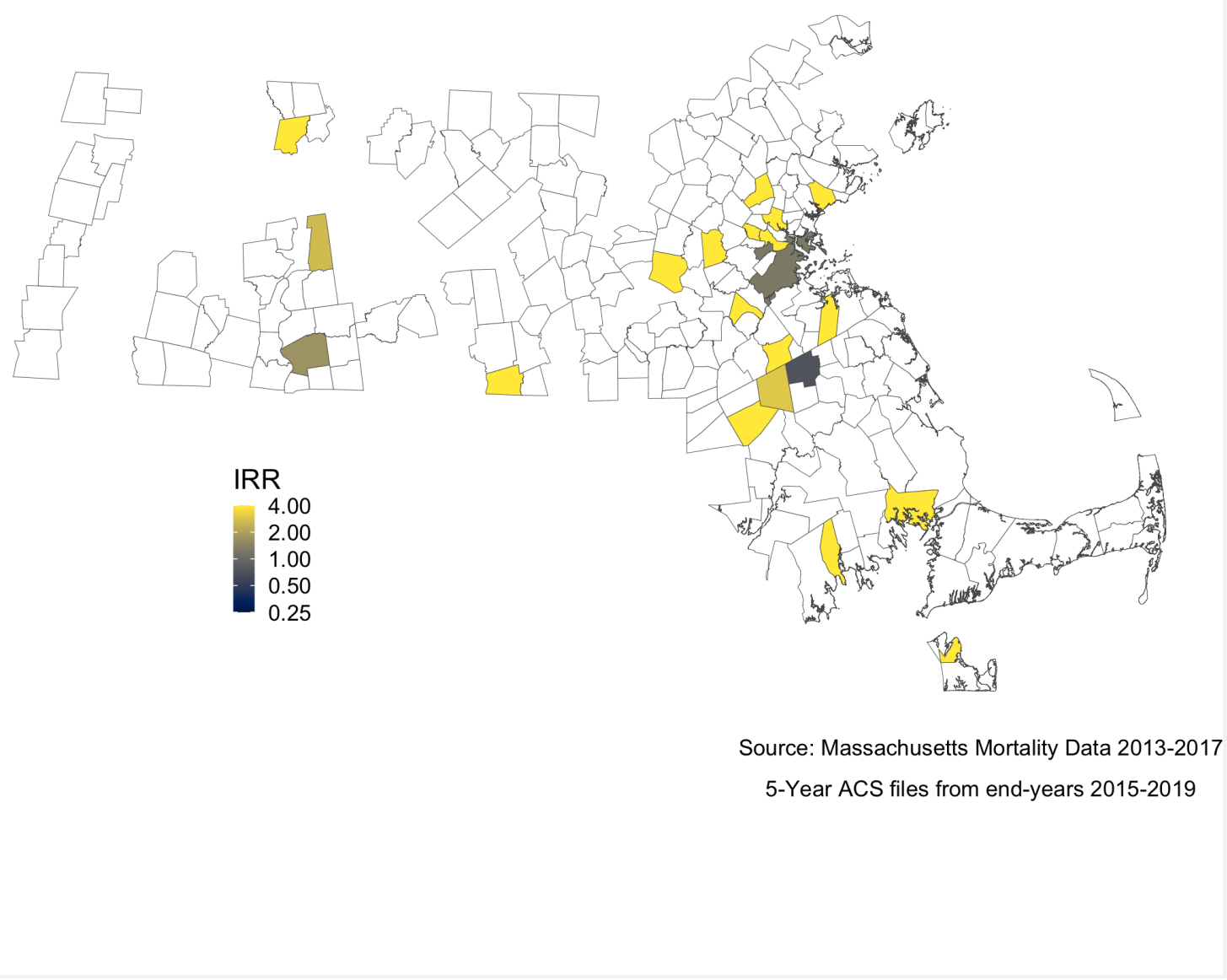
We are seeing some really extreme rates in the Black population, where we have smaller overall sample sizes. We are also seeing many cities and towns with no recorded breast cancer deaths for both Black and White populations, but this is particularly the case for the Black population (it is likely that these are cities and towns with very small or no resident Black populations due to racialized economic segregation). What else do you observe?

We can also display these two maps as one, by calculating the rate difference, or the rate ratio. We will calculate and display the incidence rate ratio comparing the age-standardized mortality rates for the Black non-Hispanic and White non-Hispanic populations, as we visualized at the start of this section.

```
irr_data <- ma_mort_stratified %>%
  select(super_town,race_group, std_rate, var_std_rate) %>%
  pivot_wider(id_cols = super_town,
              names_from = race_group,
              values_from = c(std_rate, var_std_rate)) %>%
  mutate(irr = ifelse(`std_rate_Non-Hispanic White` == 0, NA_real_, `std_rate_Non-Hispanic
Black` / `std_rate_Non-Hispanic White`),
         irr_var = `var_std_rate_Non-Hispanic Black` + `var_std_rate_Non-Hispanic White`,
         irr_lo95 = irr - 1.96*sqrt(irr_var),
         irr_up95 = irr + 1.96*sqrt(irr_var))

irr_plot <- irr_data %>%
  arrange(irr) %>%
  filter(!is.na(irr)) %>%
  mutate(orderID = row_number()) %>%
  ungroup() %>%
  ggplot(aes(x=orderID, y=irr)) +
    geom_point(color = "limegreen", alpha = 0.8, size = 1) +
    geom_errorbar(aes(ymin = irr_lo95, ymax=irr_up95), size = 0.1) +
    labs(title = expression(atop("Incidence Rate Ratio Comparing Non-Hispanic Black",
                                 "and Non Hispanic White Populations")),
         caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
```

We can also map this like the other maps.

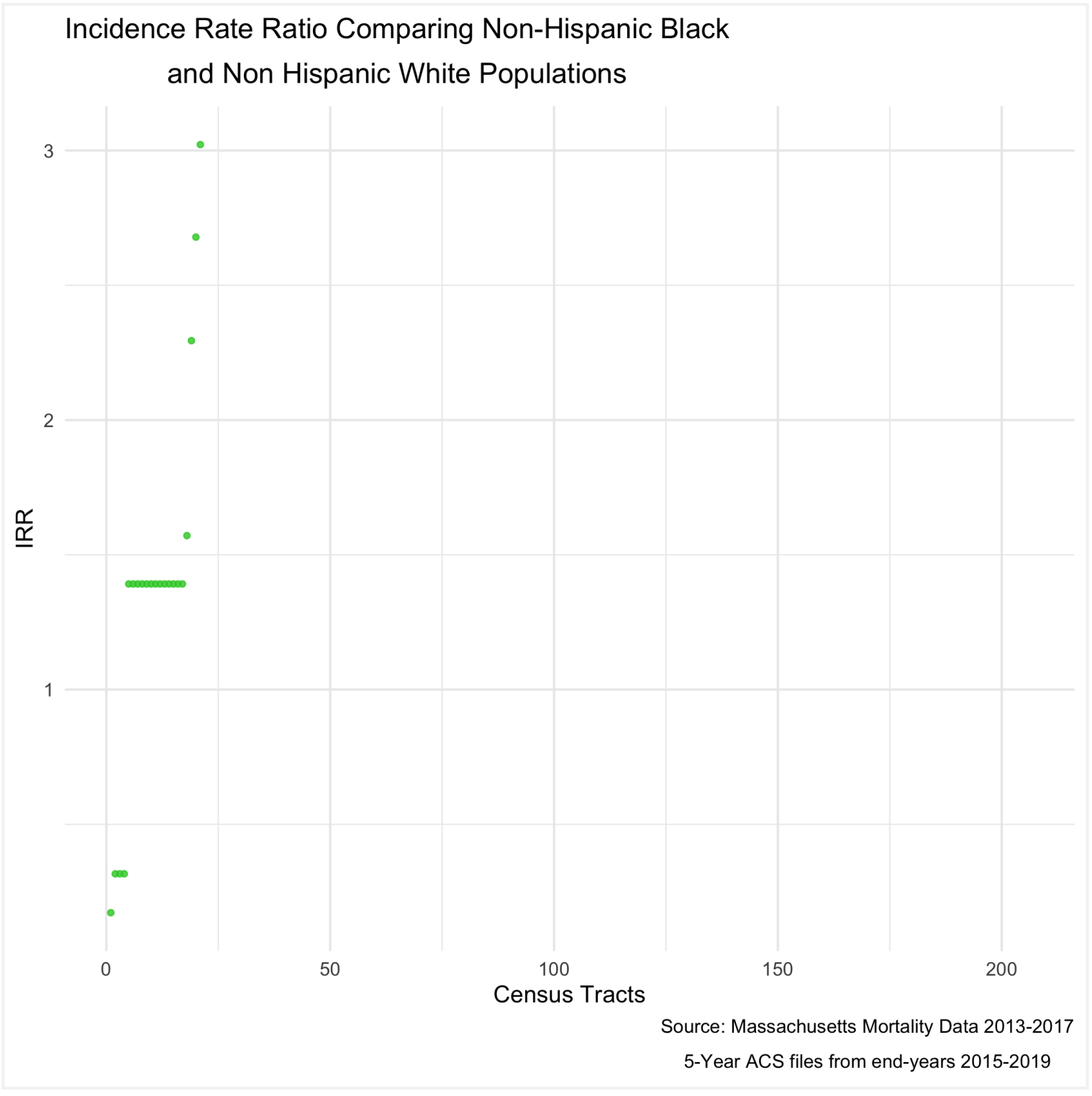
```
map.irr.bc <- irr_data %>%
  mutate(irr = case_when(irr > 4  ~ 4,
                         TRUE     ~ irr)) %>%
  left_join(town_geometry, by= "super_town") %>%
  ggplot() +
    geom_sf(mapping = aes(geometry = geometry,
                          fill = irr),
            lwd = 0.1) +
  scale_fill_viridis(option = "E",
                     trans = scales::pseudo_log_trans(sigma=0.01),
                     limits = exp(c(-1,1)*log(4)),
                     breaks=c(0.25, 0.5,1,2,4),
                     na.value = "white") +
  labs(title = expression(atop("Incidence Rate Ratio Comparing Non-Hispanic Black",
                               "and Non Hispanic White Populations")),
       caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                                 "5-Year ACS files from end-years 2015-2019")),
       fill = "IRR", x="", y="") +
theme_void() +
  theme(legend.position = c(0.25, 0.25),
        legend.key.size = unit(0.3, "cm"),
        plot.title      = element_text(hjust = 0.1),
```



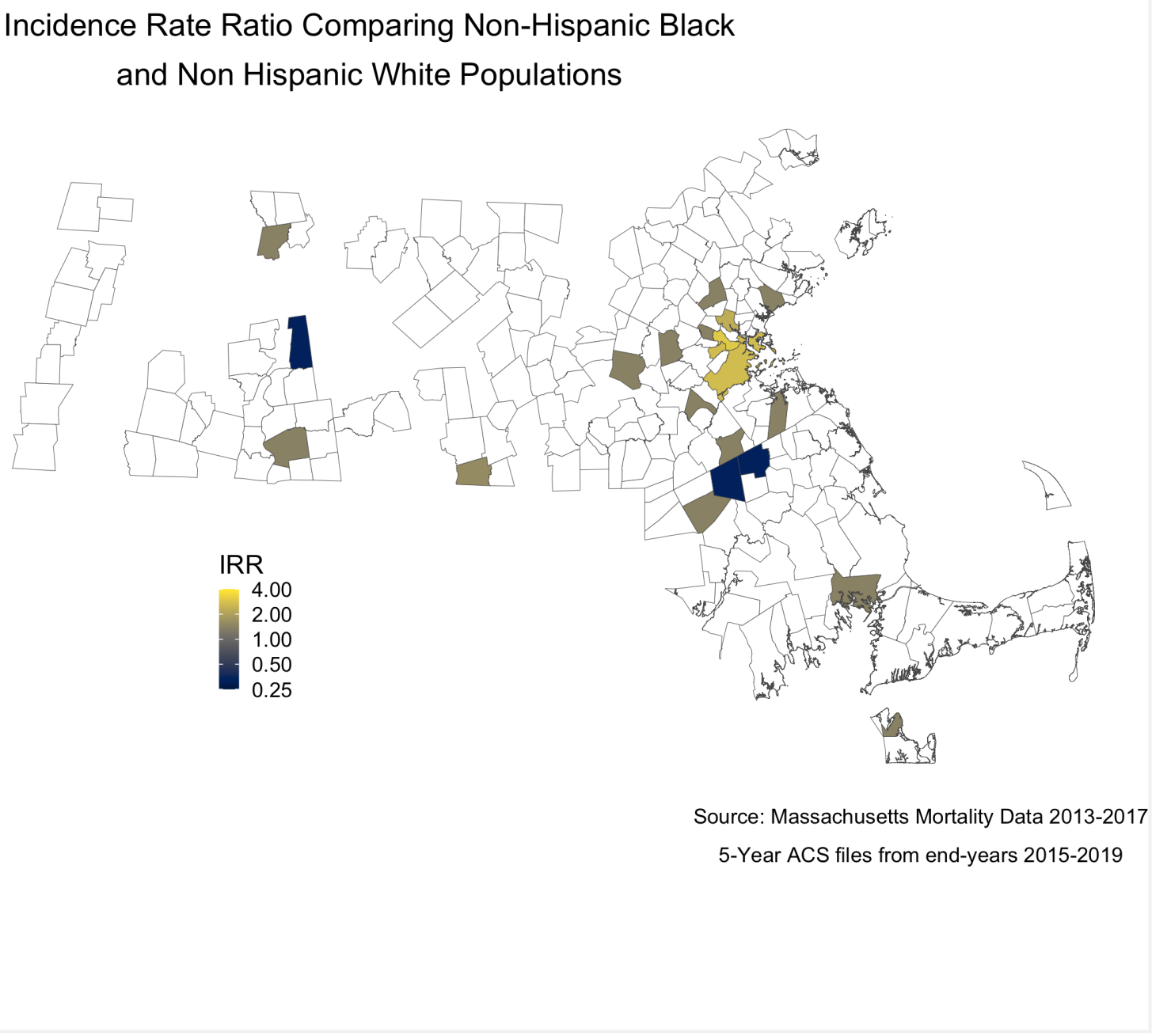
Note that the boundaries for some cities and towns are completely blank. These areas did not have any recorded breast cancer deaths for Black or White women so they don't get visualized because there is literally no data. Cities and towns that are visualized, but are white or "empty," only have data for White women. Again, we see that only a handful of cities and towns recorded breast cancer deaths for Black women, so these cities and towns get visualized as white or "empty" because an IRR can't be calculated. Key is understanding that these rates are influenced by the small sample sizes and large potential errors we have previously visualized. It's for researchers and communities to interpret how "real" the effects we see are.

8.5.3 **Question 3: How do area-based socioeconomic measures interact with individual-level membership in racialized groups to affect patterns of breast cancer mortality (i.e., interactions between socioeconomic position and racialized groups, not just socioeconomic inequities within racialized groups)?**

```r
poisson_data <- town_ma_mort_bc %>%
  inner_join(town_ma_demo_acs_bc, by = c("year","super_town","race_group","sex","age_cat")) %>%
  filter(race_group %in% c("Non-Hispanic White","Non-Hispanic Black"),
         age_cat != "Total") %>%
  group_by(super_town, race_group, age_cat) %>%
  summarise(num = sum(deaths, na.rm=T),
            den = sum(population, na.rm=T))%>%
  mutate(den = den + 0.001,
         race_group = factor(race_group)) %>%
  inner_join(town_ma_absm_sum, by = c("super_town")) %>%
  fastDummies::dummy_cols(select_columns=c("pov_cat", "ICE_qt")) %>%
  rename(pov_cat_1 = "pov_cat_0-4.9%",
         pov_cat_2 = "pov_cat_5-9.9%",
         pov_cat_3 = "pov_cat_10-19.9%",
         pov_cat_4 = "pov_cat_20-100%",
         ICE_qt_1 = "ICE_qt_[-0.272,0.0426]",
         ICE_qt_2 = "ICE_qt_(0.0426,0.237]" ,
         ICE_qt_3 = "ICE_qt_(0.237,0.337]" ,
         ICE_qt_4 = "ICE_qt_(0.337,0.45]" ,
         ICE_qt_5 = "ICE_qt_(0.45,0.687]")


# Null poisson model
```



And of course, now we can map that value as well.

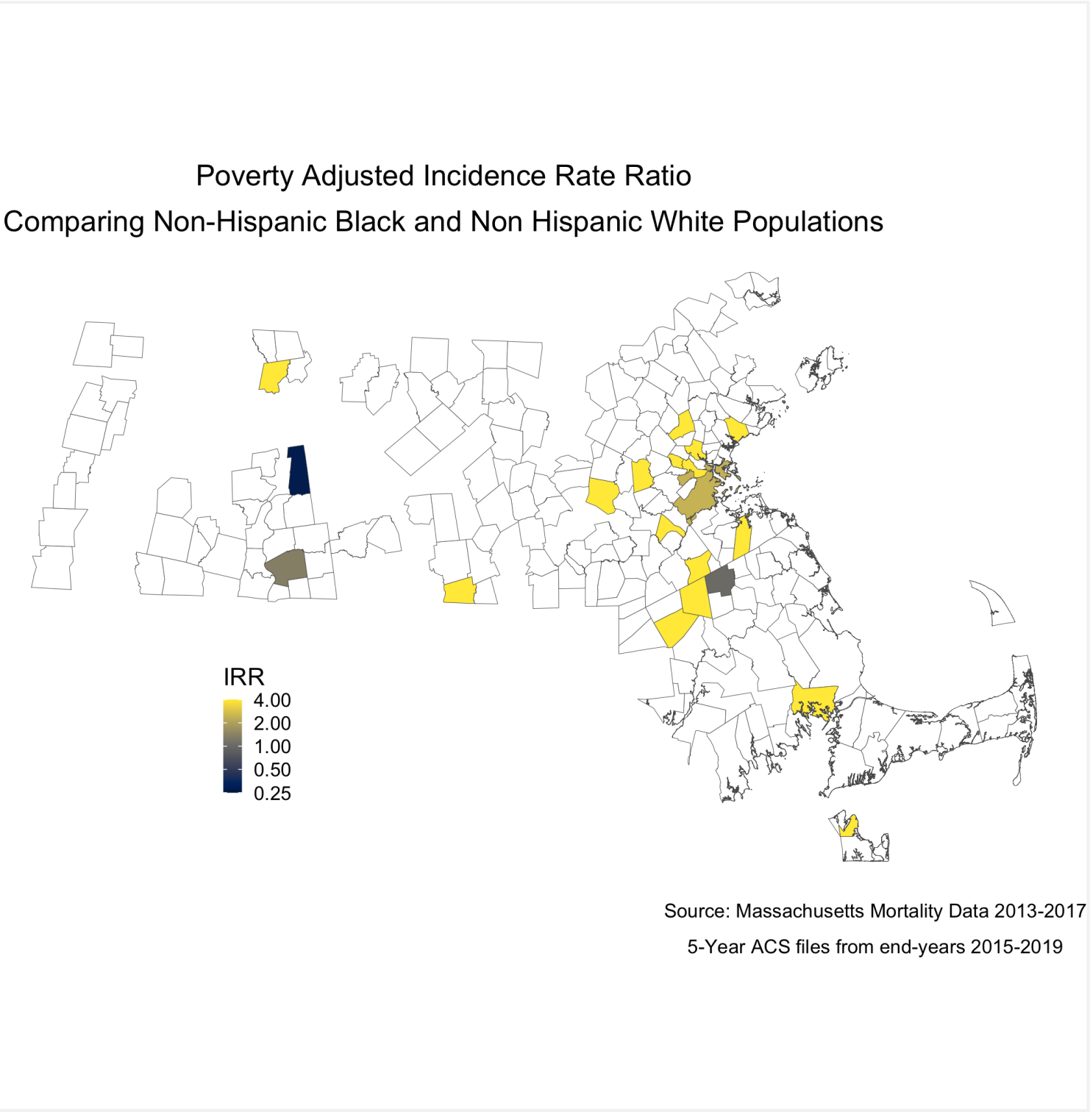


An adjusted model would allow for us to see how poverty rates are impacting the incidence rate ratio.

```
# Adjusted Poisson Model
model1 <- glm(num ~ race_group + factor(age_cat) + (race_group * factor(pov_cat, exclude =
NULL)) + offset(log(den)),
                      family=poisson(link=log),
                      data=poisson_data)

summary.model1 <- summary(model1)
saveRDS(summary.model1, file = "poisson1_bc.rds")

poisson_data$adj_fit <- as_tibble(predict(model1)) %>%
  transmute(adj_fit = exp(value))

adj_poisson_irr <- poisson_data %>%
  group_by(super_town, race_group) %>%
  summarise(num = sum(adj_fit, na.rm=T),
            den = sum(den, na.rm=T)) %>%
  mutate(adj_fit_rate = num/den * 1000) %>%
  pivot_wider(id_cols = super_town,
              names_from = race_group,
              values_from = adj_fit_rate) %>%
  mutate(adj_fit_irr = ifelse(`Non-Hispanic White` == 0, NA_real_, `Non-Hispanic Black` / `Non-
Hispanic White`))
```

And now the map.

```
adj_map.state.p.irr <- adj_poisson_irr %>%
  mutate(adj_fit_irr = case_when(adj_fit_irr > 4  ~ 4,
                         TRUE     ~ adj_fit_irr)) %>%
  left_join(town_geometry, by= "super_town") %>%
  ggplot() +
    geom_sf(mapping = aes(geometry = geometry,
                          fill = adj_fit_irr),
            lwd = 0.1) +
  scale_fill_viridis(option = "E",
                     trans = scales::pseudo_log_trans(sigma=0.01),
                     limits = exp(c(-1,1)*log(4)),
                     breaks=c(0.25, 0.5,1,2,4),
                     na.value = "white") +
  labs(title = expression(atop("Poverty Adjusted Incidence Rate Ratio",
                               "Comparing Non-Hispanic Black and Non Hispanic White
Populations")),
       caption = expression(atop("Source: Massachusetts Mortality Data 2013-2017",
                               "5-Year ACS files from end-years 2015-2019")),
       fill = "IRR", x="", y="") +
  theme_void() +
    theme(legend.position = c(0.25, 0.25),
          legend.key.size = unit(0.3, "cm"),
```



To understand statistically how city/town poverty level is impacting the relationship between racialized groups and breast cancer mortality, we can compare the models with and without poverty. We can see the impact of social membership in racialized group is reduced with the presence of the city/town level poverty.

```
##
## Call:
## glm(formula = num ~ race_group + factor(age_cat) + offset(log(den)),
##     family = poisson(link = log), data = poisson_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4188  -0.0709   0.7356   1.3605   7.0549
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -6.1048     1.0000  -6.105 1.03e-09 ***
## race_groupNon-Hispanic Black   0.3307     0.1640   2.016   0.0438 *
## factor(age_cat)25-29          -2.7375     1.0541  -2.597   0.0094 **
## factor(age_cat)30-34          -1.2184     1.0240  -1.190   0.2341
## factor(age_cat)85+             0.2629     1.0006   0.263   0.7928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 962.55  on 250  degrees of freedom
```

```
##
## Call:
## glm(formula = num ~ race_group + factor(age_cat) + (race_group *
##     factor(pov_cat, exclude = NULL)) + offset(log(den)), family = poisson(link = log),
##     data = poisson_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5916  -0.2185   0.4568   1.2308   7.0101
##
## Coefficients:
##                                                                 Estimate Std. Error z
value Pr(>|z|)
## (Intercept)                                                    -6.00791    1.00397
-5.984 2.17e-09 ***
## race_groupNon-Hispanic Black                                    3.60000    0.74960
4.803 1.57e-06 ***
## factor(age_cat)25-29                                           -2.21154    1.05829
-2.090 0.036641 *
## factor(age_cat)30-34                                           -1.13313    1.02978
-1.100 0.271174
## factor(age_cat)85+                                              0.42021    1.00158
```

8.5.4 References

Breast Cancer Research Foundation. (2022). Black women and breast cancer: Why disparities persist and how to end them. https://www.bcrf.org/blog/black-women-and-breast-cancer-why-disparities-persist-and-how-end-them/; accessed June 14, 2022.

Friedman D, Hunter E, Parrish R (eds). (2005). Health statistics. New York: Oxford University Press.

Hetzel AM. (1997). History and organization of vital statistics systems. Bethesda, MA: National Center for Health Statistics. https://www.cdc.gov/nchs/data/misc/usvss.pdf; accessed June 14, 2022.

Krieger N. (2019). The US Census and the People's Health: Public Health Engagement From Enslavement and "Indians Not Taxed" to Census Tracts and Health Equity (1790-2018). Am J Public Health. 2019 Aug;109(8):1092-1100. doi: 10.2105/AJPH.2019.305017. Epub 2019 Jun 20.

Krieger N, Singh, N, Waterman, PD. (2016). Metrics for monitoring cancer inequities: residential segregation, the Index of Concentration at the Extremes (ICE), and breast cancer estrogen receptor status (USA, 1992-2012). Cancer Causes Control. 2016 Sep;27(9):1139-51. doi: 10.1007/s10552-016-0793-7. Epub 2016 Aug 8.

US Office of Management and Budget. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Federal Register 1997; 62(210):58782-58790. https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf; accessed June 14, 2022.

Williams DR, Collins C. (2001). Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Rep. Sep-Oct 2001;116(5):404-16. doi: 10.1093/phr/116.5.404.

« 7 Case Study 1: Premature Mortality in Massachusetts (2013 - 2017)

9 Case Study 3: COVID-19 Mortality in Cook County (March 2020 - March 2022) »

9 Case Study 3: COVID-19 Mortality in Cook County (March 2020 - March 2022)

By: Sudipta Saha, Christian Testa

9.1 Introduction

In this case study, the outcome of interest is COVID-19 mortality in Cook County. The impact of COVID-19 has been marked by inequities by racialized/ethnic group and socioeconomic position. Here we investigate disparities in COVID-19 mortality, with a focus on comparing risks between Black Hispanic and non-Hispanic, White non-Hispanic, and Hispanic populations.

The Cook County Medical Examiner has made a dataset with COVID-19 related deaths publicly available with the following intent:

> Cook County Government has created the Medical Examiner COVID-19 Dashboard to provide direct, transparent access to critical information about COVID-19 deaths in the County for public health agencies, medical professionals, first responders, journalists, policymakers and residents. The Medical Examiner's Office encourages visitors to use this data to explore trends, identify areas of concern and take appropriate action. The data can be utilized to identify communities that are most severely impacted by the virus and can inform proactive public policy.

Read more: https://datacatalog.cookcountyil.gov/stories/s/ttk4-trbu

Data Source: https://datacatalog.cookcountyil.gov/Public-Safety/Medical-Examiner-Case-Archive-COVID-19-Related-Dea/3trz-enys

The death records in this dataset includes individual-level fields for "Race" and "Latino", from which we obtain social membership in racialized/ethnic groups. The data set also includes residential ZIP Codes, which can be linked to US Census ZIP Code Tabulation Areas (ZCTAs), which we can use to link to area-based social measures (ABSMs) from the American Community Survey (ACS) 2015-19. We will explore how membership in racialized/ethnic groups and ABSMs is associated with COVID-19 mortality.

9.2 Motivation, Research Questions, and Learning Objectives

This case study focuses on the following research questions:

- What are the differences in COVID-19 mortality rates by racialized group accounting for age?
- What are the gradients in overall COVID-19 mortality in relation to ABSMs measuring: racialized group composition, the ICE for racialized economic segregation, percent of population living below the poverty line, percent of people living in crowded housing, and median income?
- What are the gradients in COVID-19 mortality by racialized group in relation to the ABSMs?
- What is the spatial variation in COVID-19 mortality by racialized group?

In today's case example using these data, we'll show you how you can use the `tidycensus` package to download relevant area based population estimates and sociodemographic measures from the ACS.

> We'll show you how we cleaned the data and merged in data from the ACS in the **Cleaning the Data** section. However, we're also providing you a cleaned dataset that should allow you to pick up and follow along from the **Visualizing Your Data** section onwards.

9.3 Approach

- To look at overall differences in COVID-19 mortality by racialized group, we will model aggregate mortality rates across Cook County by racialized group
- To look at gradients in relation to ABSMs, we will model overall ZCTA-level mortality rates for each ABSM of interest.
- To look at gradients in relation to ABSMs by racialized group, we will model ZCTA-level mortality separately for each racialized group and each ABSM.
- To explore spatial variation we will use spatial models.

9.4 Dependencies

These are the packages that you will need to run the code in this case example. Once you copy and run the code below to load the dependencies, you can jump ahead to the **Visualizing Your Data** Section if you want.

```
# mission critical packages
library(tidycensus)
library(tidyverse)
library(sf)
library(tigris)
library(mapview)
library(INLA)
library(spdep)

# nice to have packages
library(magrittr) # for the %<>% and %$% pipe
library(janitor)
library(purrr)
library(Hmisc)
library(epitools)
library(leaflet)
library(scales)
```

9.5 Cleaning the Data

9.5.1 Denominator Data

Here we will use `tidycensus` to download relevant variables from the 2015-19 ACS dataset. The complete list of variables can be viewed online here: https://api.census.gov/data/2019/acs/acs5/variables.html. You will need a U.S. Census API key to download the data, which you can obtain from here: https://api.census.gov/data/key_signup.html. The key is for personal use only, so it has been redacted here.

We will be looking at a range of different ABSMs, and we also require age-stratified populations by racialized group. You can browse the variables to identify which tables contain the variables you need.

> It can be helpful to identify patterns in the variable names to efficiently query the Census API. In the code below you will see an approach that does not require us to type out all the variable names. Can you think of more efficient ways to query the API?

```r
# census_api_key("Your API KEY goes here")

# get population sizes from ACS ----------------------------------------

#Get a list of the variable names
acs_vars <- tidycensus::load_variables(2019, dataset = 'acs5')



# the racialized group and age-group stratified population estimates from ACS are
# in the B01001 table.
#
# B01001_001 through _049 are the sex/gender overall population size estimates,
# and B01001A through B01001I are the "race/ethnicity" specific tables.  For the
# "race/ethnicity" specific tables the age-groups suffixes range from _001 to _031
#
# in the following three steps, we programmatically construct the population
# size variables that we want to retrieve from the ACS since otherwise there are
# a lot of them to type out.
race_chars <-
  c(
    white = 'H', # In the 2015-19 ACS data, a suffix of H indicates non-Hispanic White, but be
careful because these suffixes may change from year to year!
```

Since the Cook County Medical Examiner Case Archive data includes records on deaths where the decedents' residential ZCTAs are inside as well as outside Cook County, Illinois, we want to restrict our dataset to deaths where their county of residence was in Cook County. An issue is that ZCTAs are not neatly nested within the county borders - often part of a ZCTA can lie outside the county.

> How would you work with this issue of ZCTAs crossing county borders? Here we calculated the percentage of an area in a ZCTA that also falls within the Cook County borders, and retained those with 90% overlap (an arbitrary threshold). Would you use a different threshold? We also had to deal with issues of parts of Lake Michigan (bordering Cook County's east coast) being included in some shapefiles but not others.

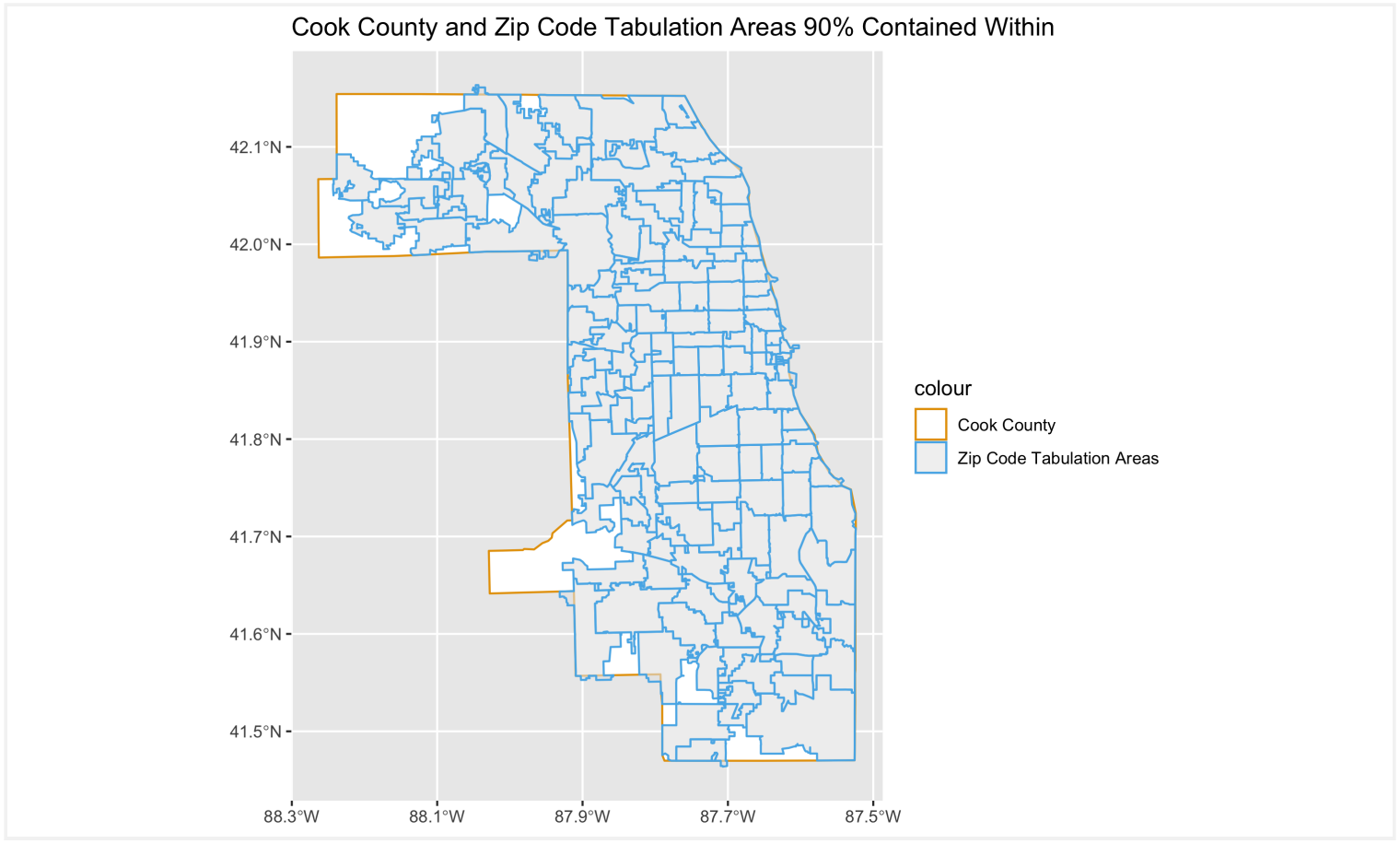
> Note that the unit of geography for analysis is the US census-defined Zip Code Tabulation Area (ZCTA), which is related to but not identical with the US Postal Service ZIP Code (which is a unit of mail delivery, reflecting postal carrier routes, such that a given spatial area can encompass several ZIP Codes). To create ZCTAs, the US Census assigns each census block the most frequently occurring ZIP Code within the block. For technical information about ZCTAs, see: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html. This means that ZCTAs and zip codes may not cover the same area exactly. A discussion of pitfalls to keep in mind when using ZCTAs linked to residential Zip Code can be found in this paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447194/ .

```r
# identify relevant zip codes ----------------------------------------

# we originally downloaded the county shapefiles using tigris, but
# we found that those contained water area we wanted to remove.
#
# we tried using the tigris::erase_water function but it took a very
# long time to run for us, so we found that the alternative was easier
# to implement:  downloading a county shapefile directly from the census
# and using that one which came already with the water areas removed.
#
# We got our county shapefile for Cook county in what follows from here:
# https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

########## Change to file path where the shape file is downloaded ############
counties <- read_sf('./data/09-cook-county-
covid/cb_2018_us_county_5m/cb_2018_us_county_5m.shp')

# get the map for cook county
cook_county <- counties %>% filter(
  # get the 5-digit FIPS or GEOID for Cook County programmatically by filtering
  # the tigris::fips_codes dataframe, or if you happen to know it's 17031
  # you could code it explicitly instead
```

Now that we have obtained the ZCTAs that we will keep in the study, we can now focus on these to prepare demographic and denominator data.

```r
# group population sizes together by age groups and racialized group and zip code
zips_with_population_estimates_by_race_ethnicity_and_age <-
  popsizes %>%
  filter(! is.na(age) & concept != 'SEX BY AGE') %>% # filter for overall population
  filter(geoid %in% zips_with_over_90pct_overlap$geoid) %$% #Filter to keep only the zips
selected above
  left_join(zips_with_over_90pct_overlap, .) #join with the zip dataframe to add the geometry
column

# Change names of the Census racialized groups
zips_with_population_estimates_by_race_ethnicity_and_age %<>% mutate(
  race_ethnicity = case_when(
    concept ==  "SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE)" ~ "Black, Hispanic or non-
Hispanic",
    concept ==  "SEX BY AGE (HISPANIC OR LATINO)" ~ "Hispanic",
    concept ==  "SEX BY AGE (WHITE ALONE, NOT HISPANIC OR LATINO)" ~ "White, non-Hispanic"
  ))
```

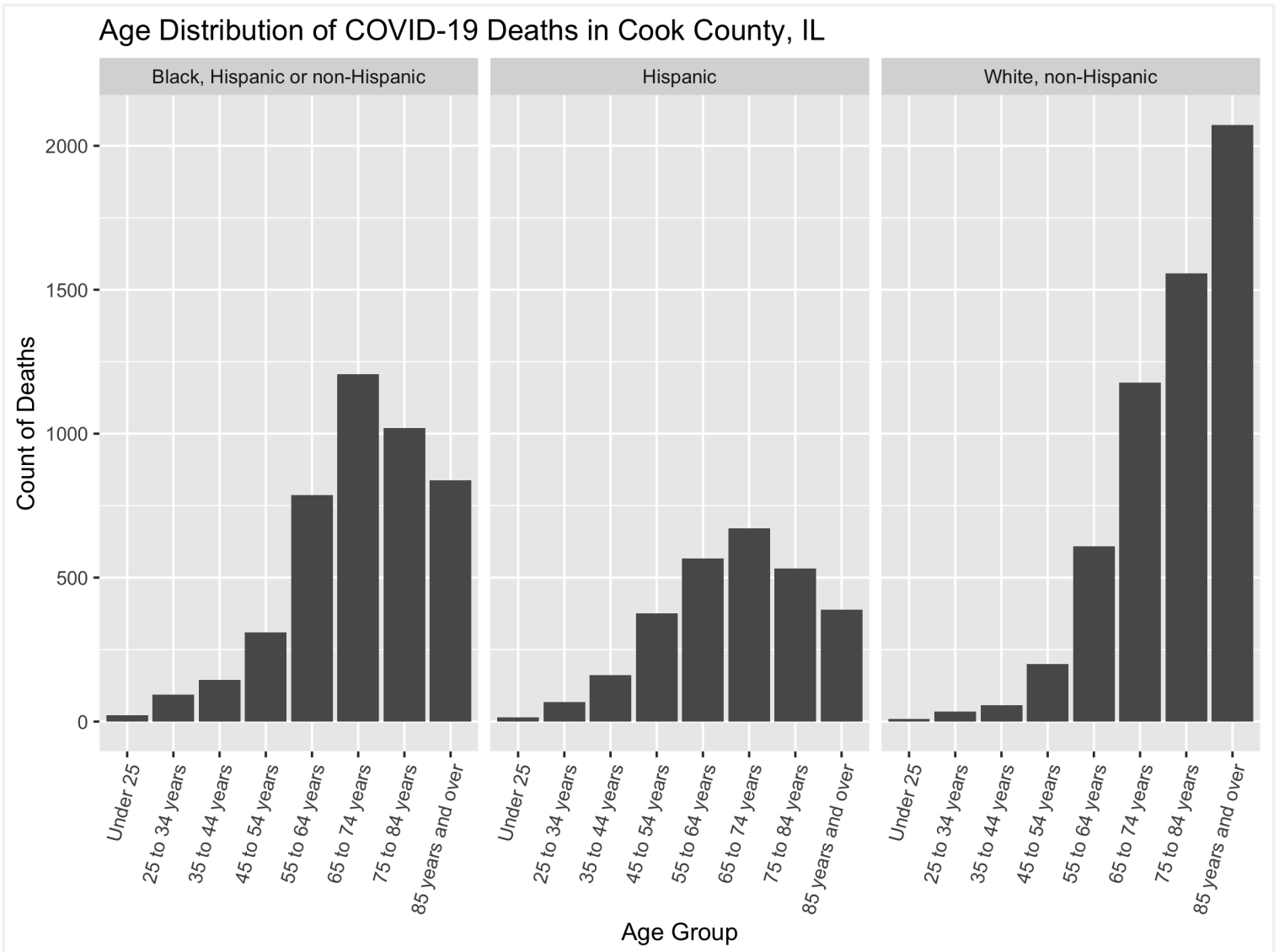
```r
# remove totals (across age groups) observations
zips_with_population_estimates_by_race_ethnicity_and_age %<>% filter(! is.na(age))
zips_with_population_estimates_by_race_ethnicity_and_age %<>% filter(! is.na(race_ethnicity))
```



9.5.2 Case Data

Now we're ready to load the case data, and add the population size estimates for each racialized group (White non-Hispanic, Black non-Hispanic and Hispanic, and Hispanic) and age-group by ZIP code.

```
# read in your data. This has been donwloaded from the Cook County Medical Examiner's office website, as mentioned above. https://datacatalog.cookcountyil.gov/Public-Safety/Medical-Examiner-Case-Archive-COVID-19-Related-Dea/3trz-enys

cook_county_deaths <-
  readr::read_csv("./data/09-cook-county-covid/Medical_Examiner_Case_Archive_-_COVID-19_Related_Deaths.csv")

# merge racialized group and age-group specific denominators ----------------

# use janitor::clean_names to standardize column name syntax into snake_case
cook_county_deaths %<>% janitor::clean_names()

# check distribution of deaths if needed
#ggplot(cook_county_deaths, aes(x = age)) + geom_bar()
#ggplot(cook_county_deaths, aes(x = race)) + geom_bar()
#ggplot(cook_county_deaths, aes(x = latino)) + geom_bar() + facet_wrap(~race)
#ggplot(cook_county_deaths, aes(x = latino)) + geom_bar()

# code racialized groups into the following:
#  Black (Hispanic or non-Hispanic)
#  Hispanic and non-Hispanic
```



We can now add ABSMs that we are interested in to the data. We'll add the Index of Concentration at the Extremes for Racialized Economic Segregation, the proportion of the population under the poverty line, and the median income in each ZCTA.

```
# add area based socioeconomic measures  ----------------------------------

# get zip code rates for
#   - poverty
#   - ICEraceinc
#   - median income

# We create a data dictionary for ABSMS. The first column indicates the total variable code,
the second the variable name, and the third the description.
absms_dictionary <- tibble::tribble(
  ~var, ~varname, ~description,
  # total population
  "B01001_001",  "total_popsize", "total population estimate",

  # racial composition
  'B01003_001',  "race_ethnicity_total", "race_ethnicity_total",

  # ICEraceinc
  "B19001_001",  'hhinc_total',   "total population for household income estimates",
  "B19001A_002", 'hhinc_w_1',     "white n.h. pop with household income <$10k",
  "B19001A_003", 'hhinc_w_2',     "white n.h. pop with household income $10k-14 999k",
  "B19001A_004", 'hhinc_w_3',     "white n.h. pop with household income $15k-19 999k",
```

We have added discretized (cut, i.e. categorical) versions of the different ABSMs. This allows the model to fit nonlinear responses with increasing levels in these covariates. Alternative approaches could involve fitting the models with smoothing splines on these variables instead, but we aren't including these approaches here. We used Illinois-wide distribution of ABSMs at the ZCTA level to create quantiles for all variables except poverty. For poverty, we use pre-specified cutpoints.

> Here we set the cutpoints for the continuous ABSMs based on the state-wide distribution of ZCTAs. How might this affect our interpretation? How might your research question and context affect how you decide what the ABSM cutpoints are?

> How might the selection of cutpoints affect your visualizations? How might your cutpoints change depending on what you want to communicate through the map?

We have a few remaining necessary cleaning steps to perform before we're ready to model these data.

We have to make sure our factor variables have appropriate reference levels set. Typically we set the reference level as the most privileged group so that we can frame the results as "the _____ group has X times the mortality rate of the reference group."

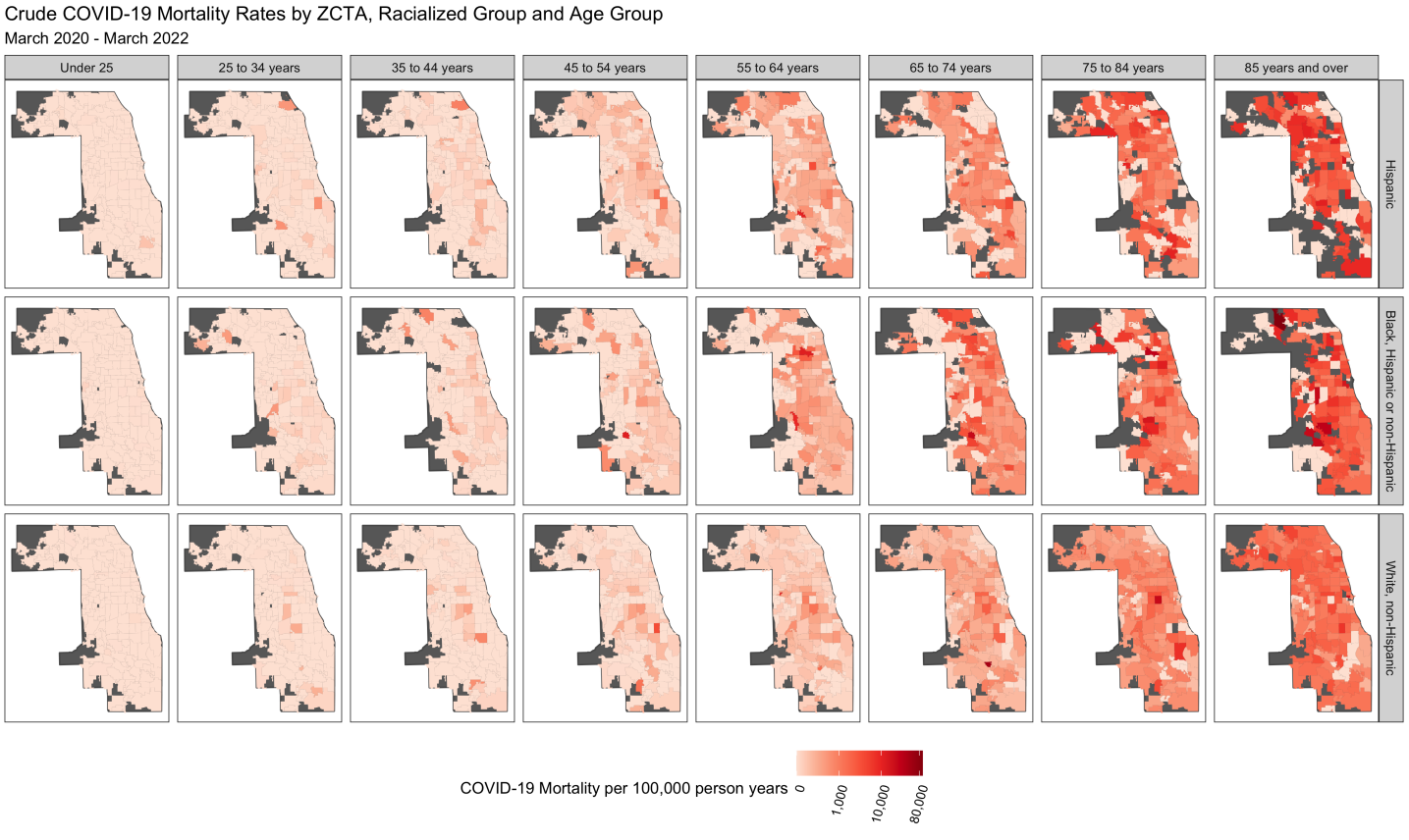
```
# prepare data for modeling  ----------------------------------------------

# remove infinite or NA mortality rates since they will cause errors in trying
# to fit the models
df_prepped <- df %>% filter(
  is.finite(mortality_per100k_py) &
    ! is.na(mortality_per100k_py))

# ungroup
df_prepped %<>% ungroup()

# make the most privileged the reference category
df_prepped %<>% mutate(
  race_ethnicity = forcats::fct_relevel(factor(race_ethnicity), "White, non-Hispanic"),
  age_group = age_group,
  ICEraceinc_cut = forcats::fct_rev(ICEraceinc_cut),
  median_income_cut = forcats::fct_rev(median_income_cut)
)

# rename the estimate variable to 'population_estimate' to be more clear
df_prepped %<>% rename(population_estimate = estimate)
```

9.6 Visualizing Your Data

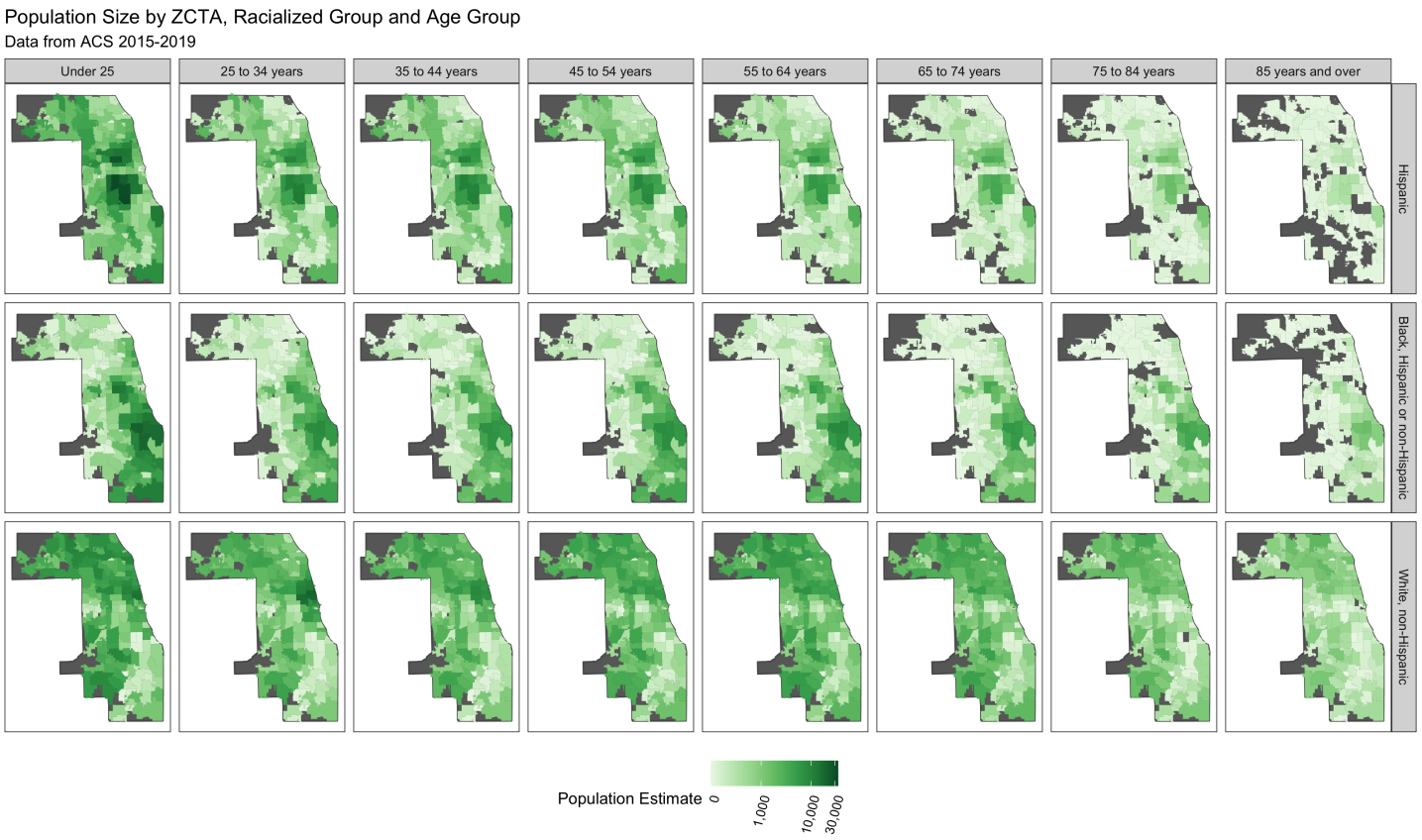
If you want to start by using the clean dataset you can start from this point on. Now that we have a clean dataset, we can create some exploratory visualizations. In the code chunk below we do some exploratory visualizations of the data. The first series of maps show the crude mortality rates by ZCTA, racialized group and age group. **To see a larger version of this or any other image right click and select open image in new tab.**

```
#change this to point to where your downloaded file is
df <- readRDS("./data/09-cook-county-covid/cook_county_mortality_cleaned.rds")

#Read in the county boundary shapefile to plot maps
counties <- read_sf('./data/09-cook-county-
covid/cb_2018_us_county_5m/cb_2018_us_county_5m.shp')

# get the map for cook county
cook_county <- counties %>% filter(
  # get the 5-digit FIPS or GEOID for Cook County programmatically by filtering
  # the tigris::fips_codes dataframe, or if you happen to know it's 17031
  # you could code it explicitly instead
  GEOID == tigris::fips_codes %>%
    filter(county == 'Cook County', state == 'IL') %$%
    paste0(state_code, county_code))

# let's start by mapping the crude mortality rates in each racialized group and
# age strata
df %<>% mutate(deaths = ifelse(is.na(deaths), 0, deaths)) # Convert missing cells to 0
df %<>% mutate(mortality_per100k_py = deaths / person_time * 1e5) # Calculate mortality rate
per 100,000 person years
```

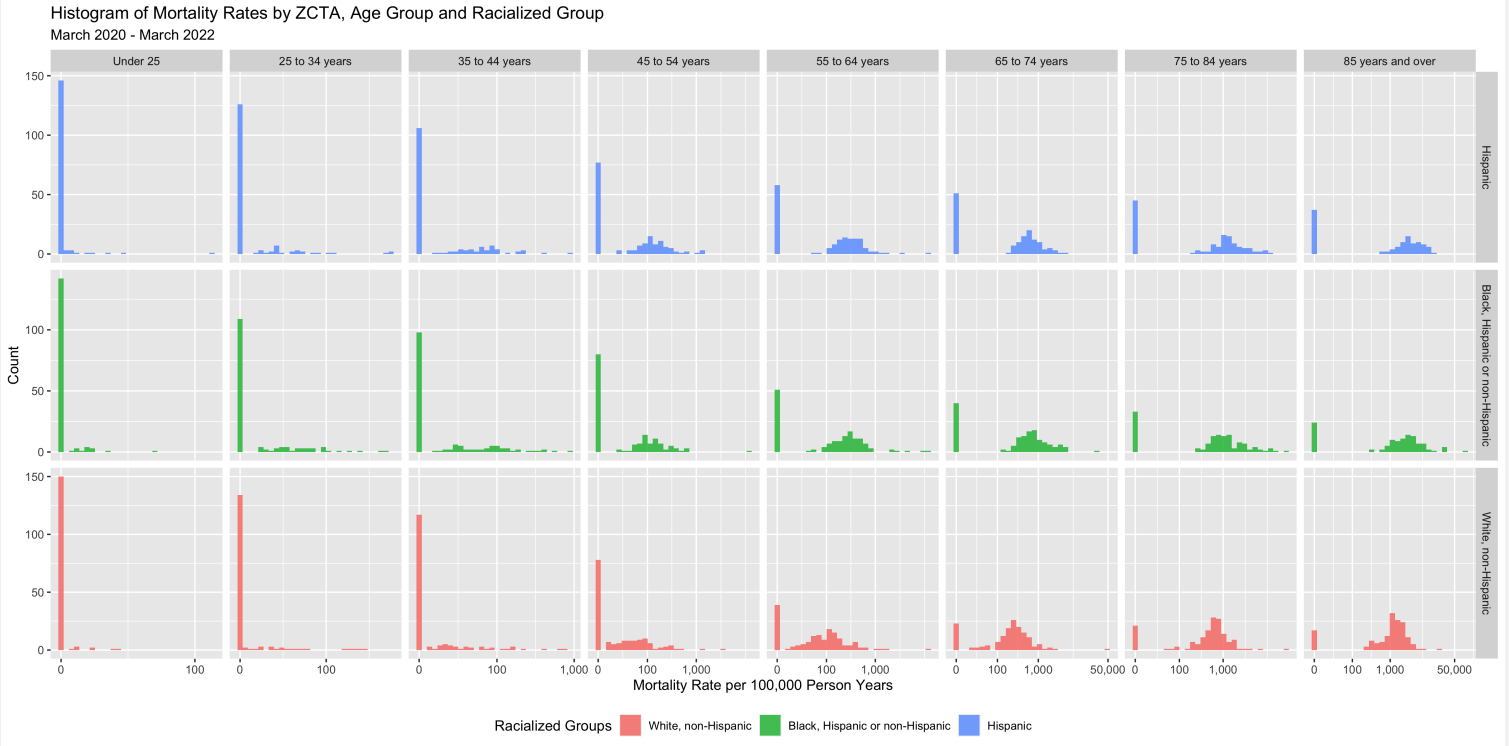
This figure shows a series of maps of the COVID-19 mortality rate per 100,000 person years between 2020-2022, by age and racialized groups.



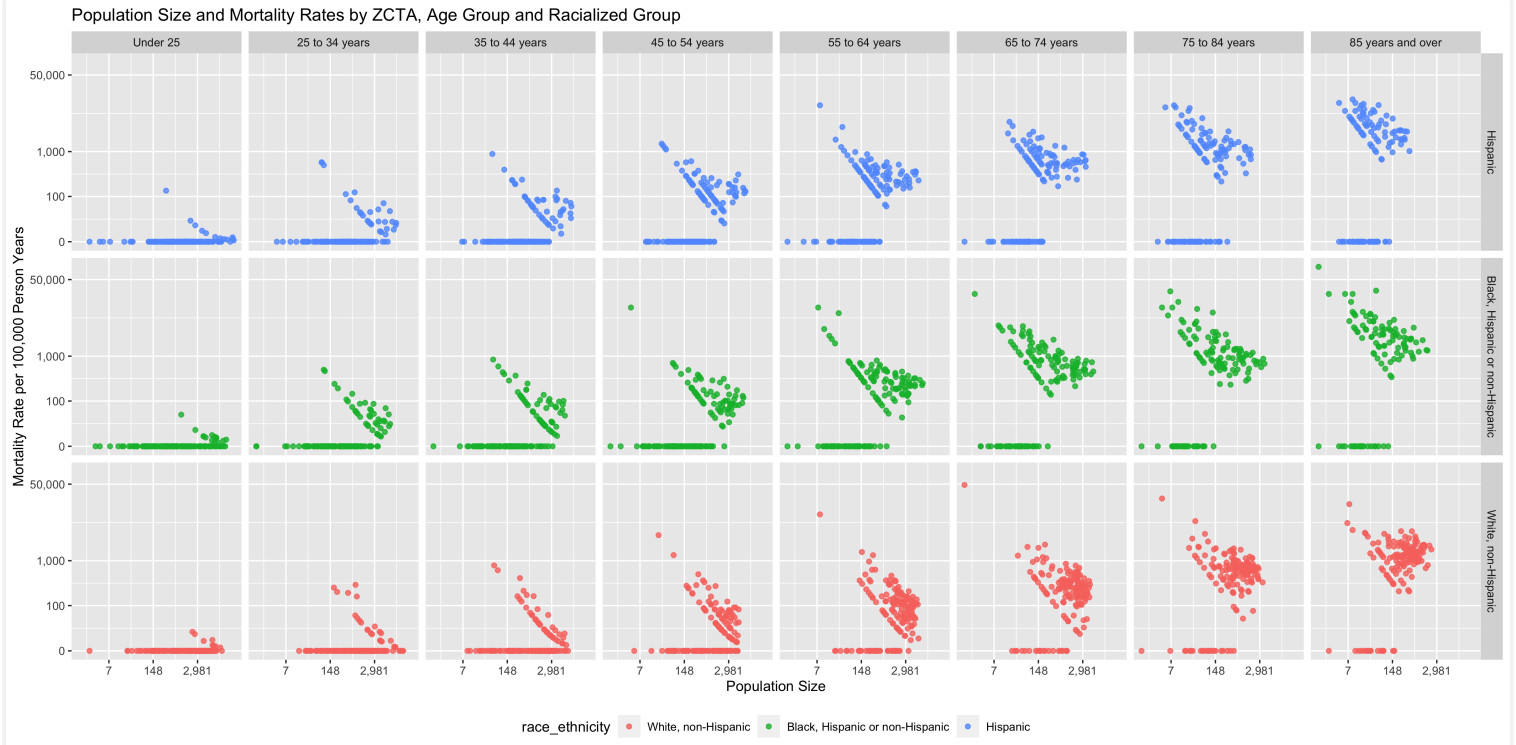
This figure shows a series of maps highlighting the underlying population at risk in each strata.



This figure shows the histogram of these rates to visualizate the distribution of the data.



This figure shows the relationship between mortality rates and population size in each strata.



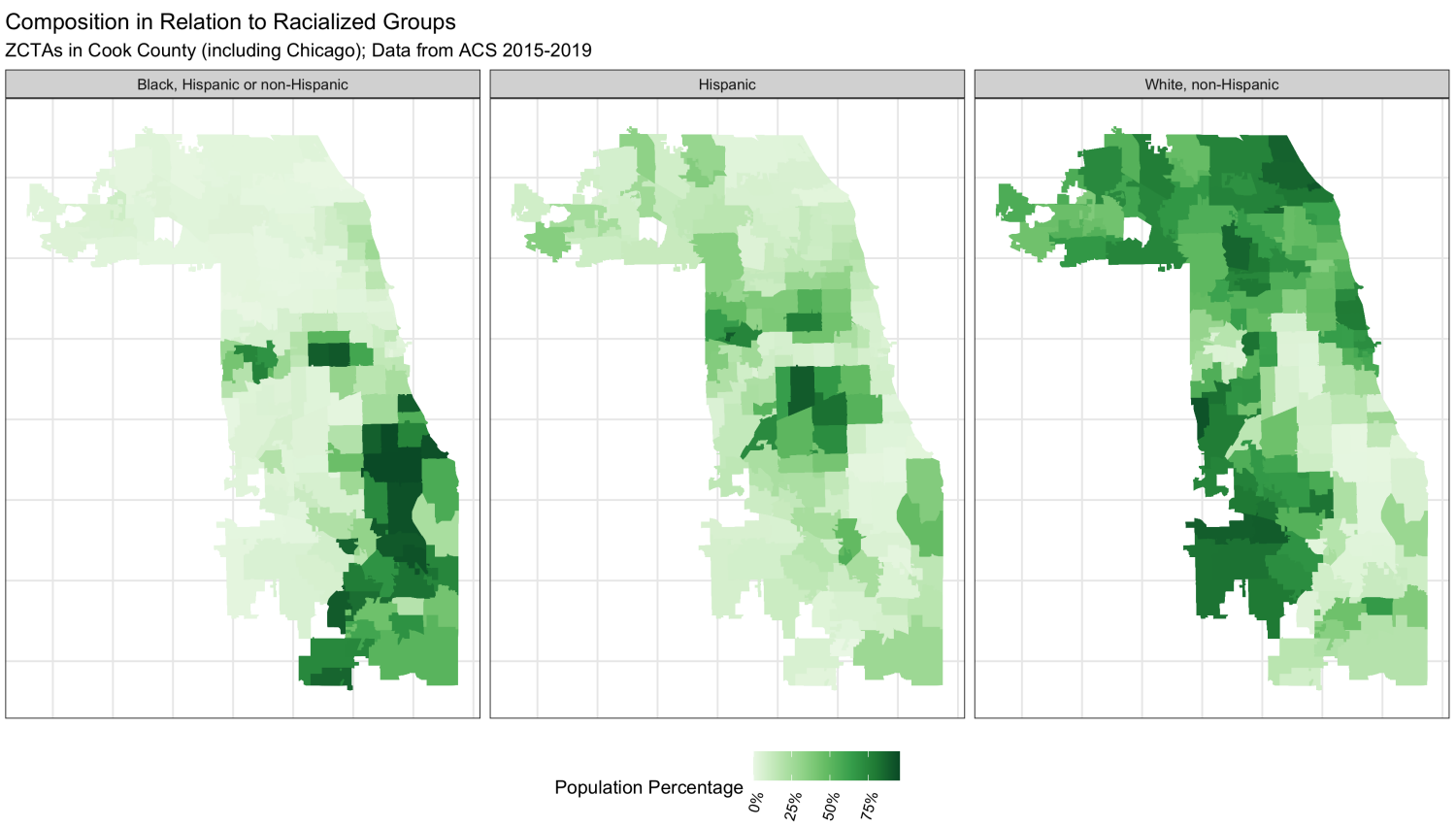
You can see that there are a number of zero-counts, and that rates are noisier when population sizes are smaller.

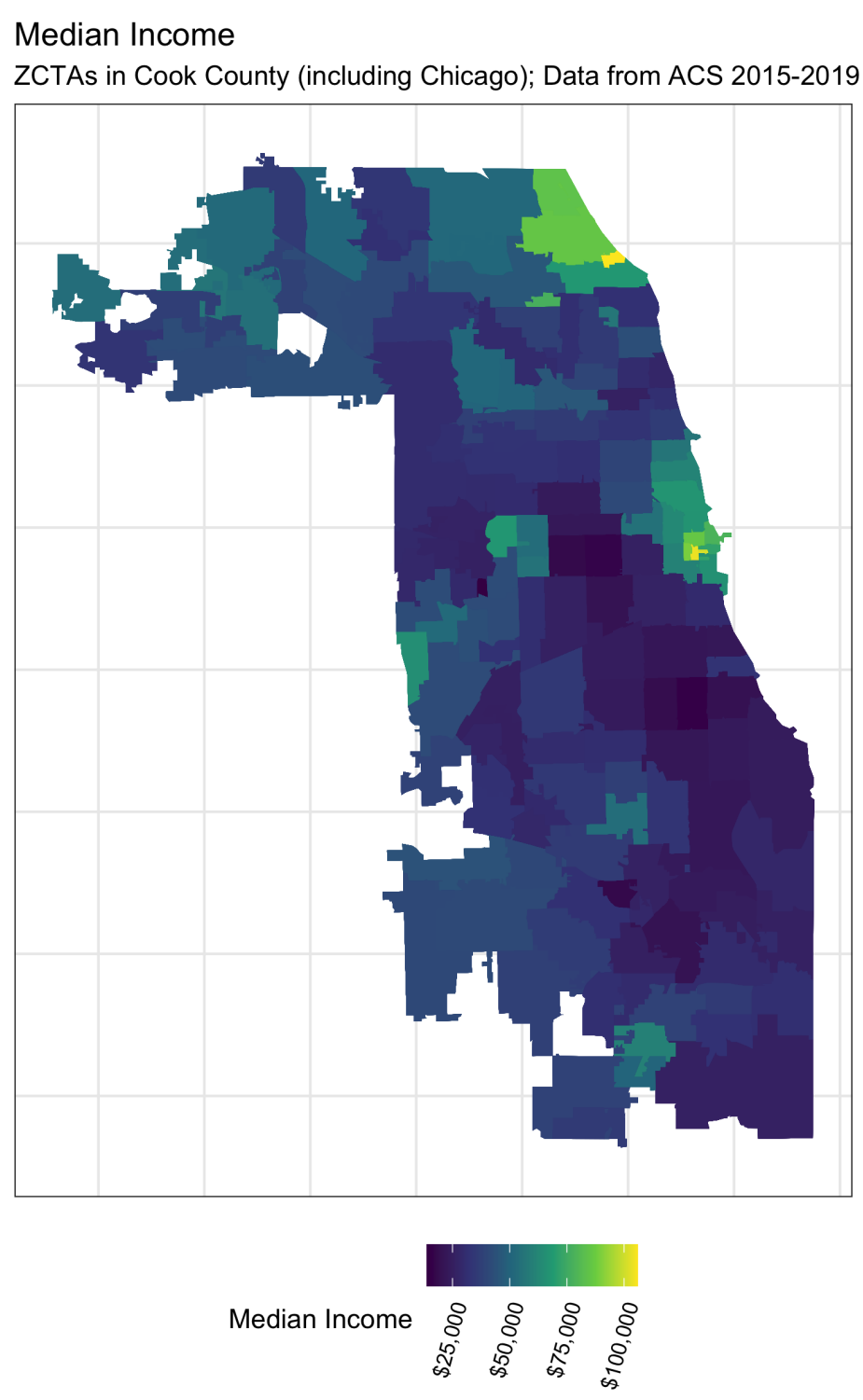
> The figure above illustrates that there is a lot more variability of mortality rates at the ZCTA-level when the population count is small. Thus we need to be careful when interpreting extreme values obtained in small areas.

> A map of crude rates of small areas with low person-time at risk may be misleading. It may identify some areas as having extreme rates, but this may be just due to chance and the small size of the population at risk.

```
# in the following visualizations the motivating principle behind the direction of the
# color palette is to show (for sequential palettes) higher density with darker colors;
# for the ICEraceinc variable a divergent color palette is used to draw attention to the
# extreme ends of the scale

# visualize the proportional racial/ethnic breakdown
df %>% dplyr::select(geoid, prop_black, prop_white_nh, prop_hispanic) %>%
  tidyr::pivot_longer(
    cols = c(prop_black, prop_white_nh, prop_hispanic),
    names_to = "race_ethnicity",
    values_to = 'proportion'
) %>%
mutate(
  race_ethnicity = recode(race_ethnicity,
                          prop_white_nh = 'White, non-Hispanic',
                          prop_black = 'Black, Hispanic or non-Hispanic',
                          prop_hispanic = 'Hispanic'
                          )
) %>%
ggplot(aes(fill = proportion)) +
geom_sf(size = 0) +
facet_grid(~race_ethnicity) +
```

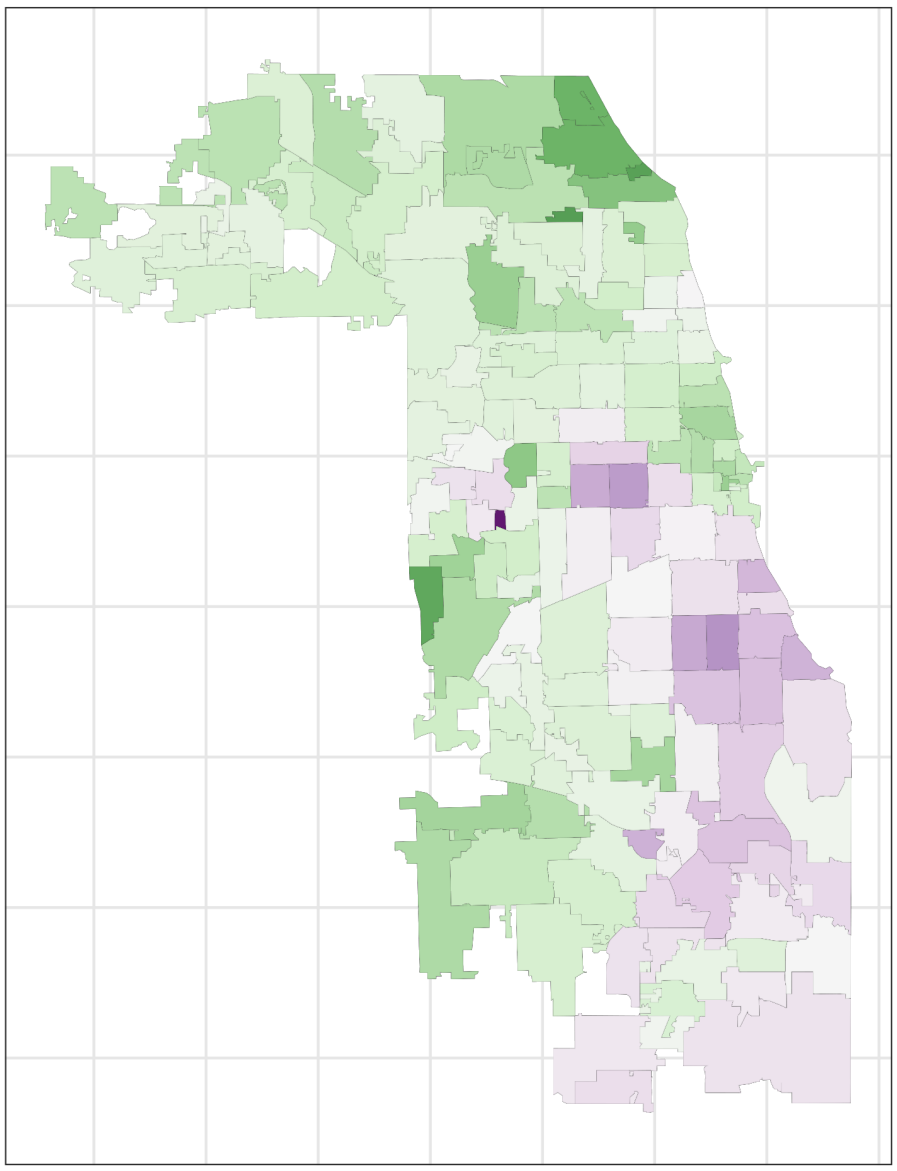
The code above produces a series of maps that visualize the different ABSMs we will use in this analysis. Note that Chicago is situated in Cook County.





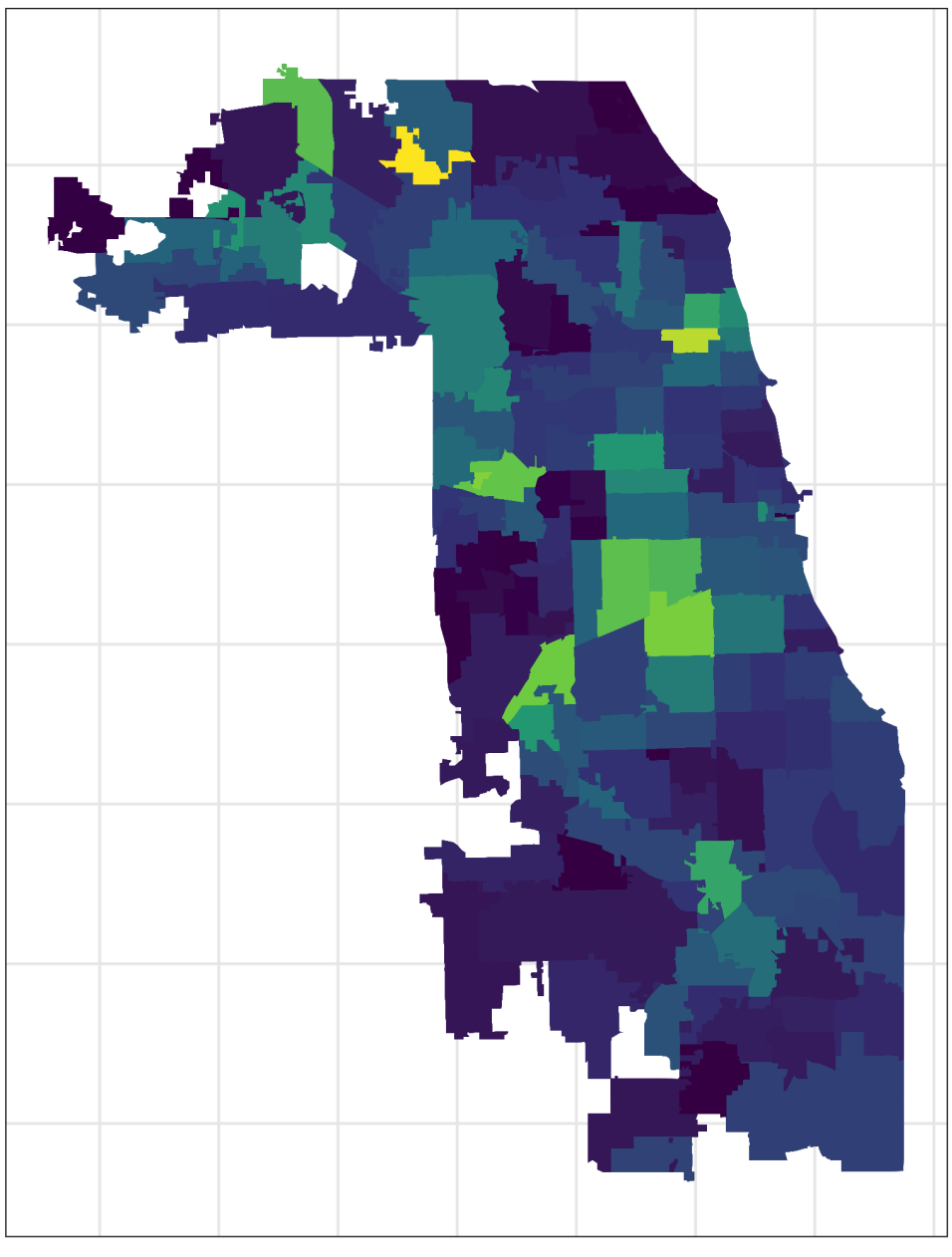
Index of Concentration at the Extremes for Racialized Economic Segregation

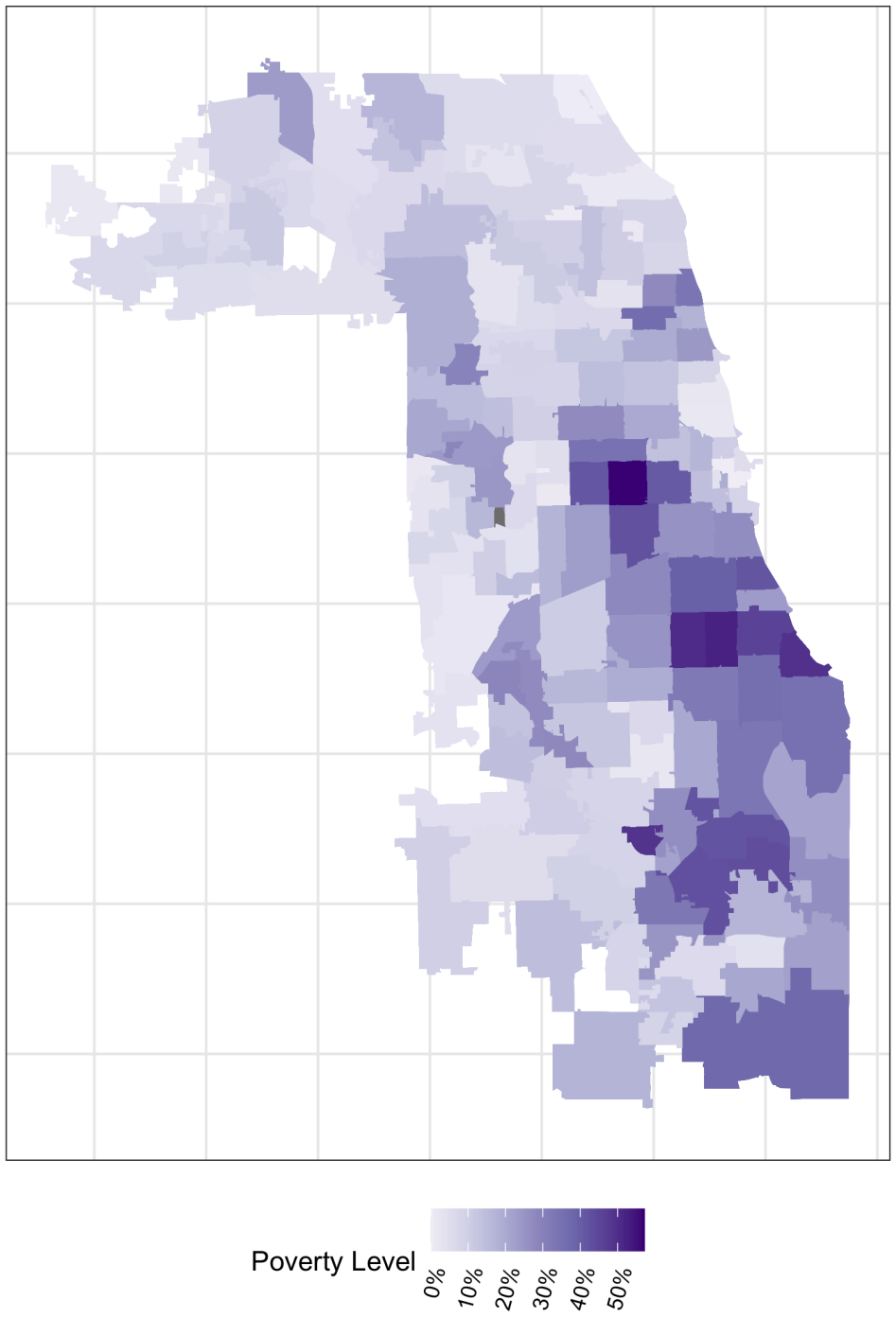
ZCTAs in Cook County (including Chicago); Data from ACS 2015-2019



Household Crowding

ZCTAs in Cook County (including Chicago); Data from ACS 2015-2019



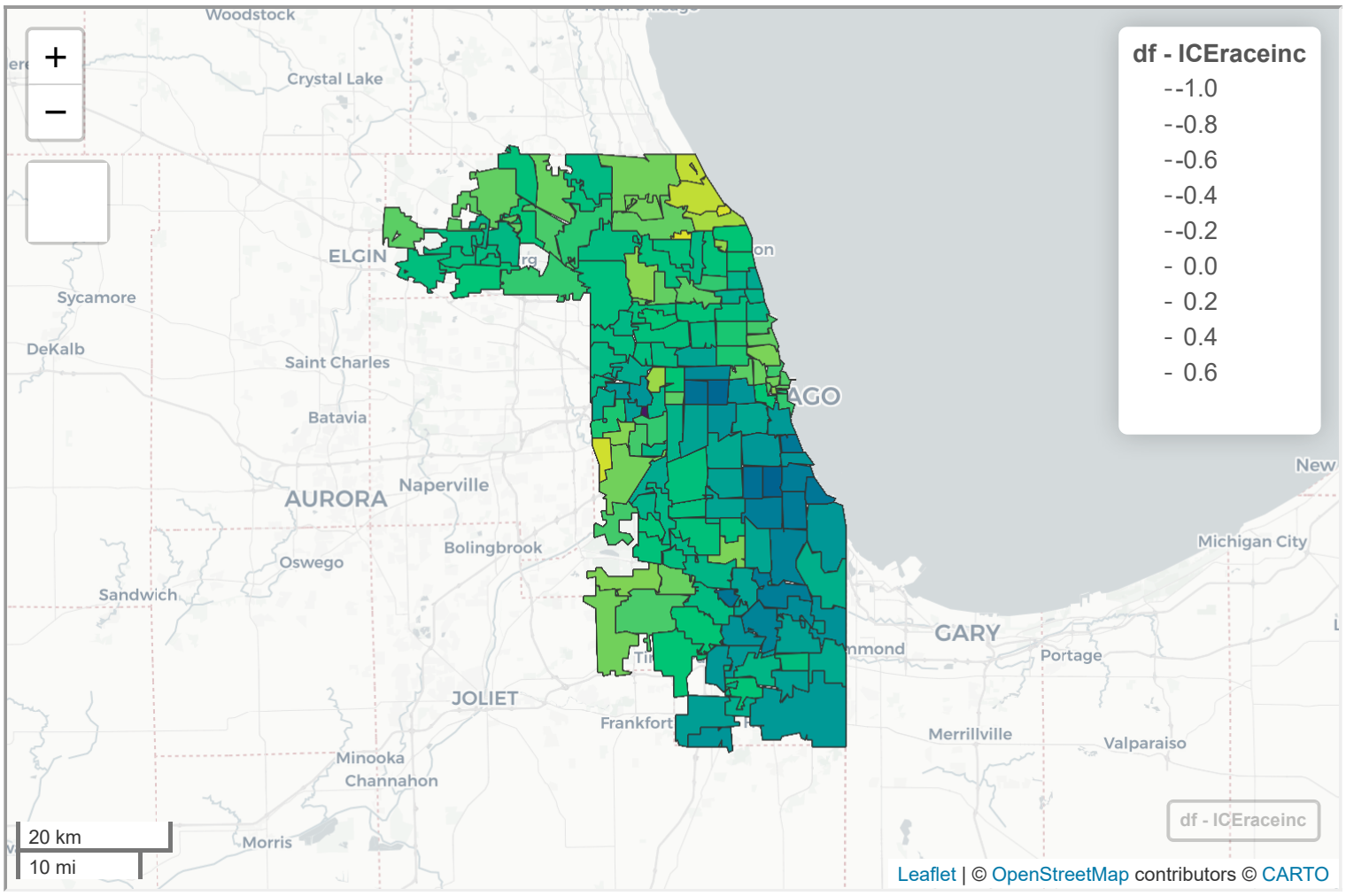


In contrast to the other ABSMs, the Index of Concentration at the Extremes (ICE) highlights racialized economic segregation (i.e., race/ethnicity + income). It measures the extent to which an area's population is concentrated into extremes of deprivation and privilege. It is scaled from -1 to 1: a value of -1 means that 100% of the population is concentrated in the most deprived group (in this analysis, conceptualized as the population of color in low-income households), and a value of 1 means that 100% of the population is concentrated into the most privileged group (in this analysis, conceptualized as the White non-Hispanic population in high-income households).

The maps above utilize various color scales. How do you think this can affect communication through maps? Do you have a preference? When might you use a diverging versus continuous color scale? Sometimes it can also be useful to plot an interactive map, shown below.

We can utilize the mapview package to plot interactive graphs.

```
# launch an interactive map to view Cook County ICEraceinc by zip code
# mapview::mapview(df, zcol = "ICEraceinc")
```



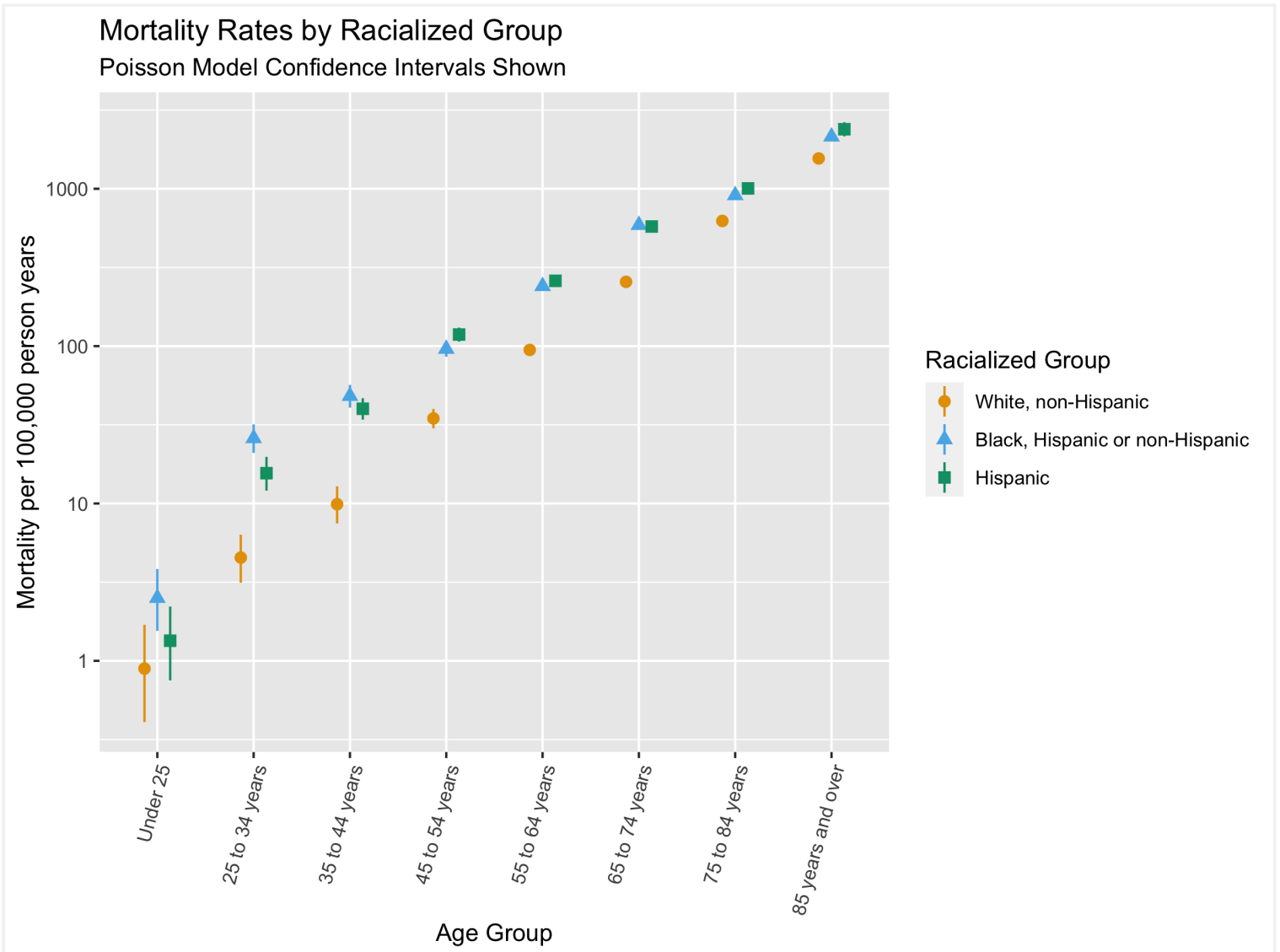
9.7 Analyzing the data

9.7.1 Differences in COVID-19 mortality rates by racialized group

First, we can take a look at the COVID-19 mortality rates by racialized group in aggregate. Because we are aggregating up to the County level, we are not looking at gradients by ABSMs.

```
#Since we are not taking into account ABSMs in this first analysis, aggregate the data and
visualize
crude_mortality <- df %>% st_drop_geometry() %>% #dplyr works faster if we drop the geometry
column.
  group_by(age_group, race_ethnicity) %>%
  dplyr::summarize(population_estimate = sum(population_estimate),
            deaths = sum(deaths),
            person_time = sum(person_time)) %>%
  mutate(mortality_per100k_py = (deaths / person_time)*100000)

#plot clude mortality and confidence intervals
ggplot() +
  geom_pointrange(
    data = crude_mortality,
    aes(
      x = age_group,
      color = race_ethnicity,
      y = mortality_per100k_py,
      ymin = epitools::pois.exact(x=deaths, pt=person_time, conf.level=0.95)[,4]*1e5,
      ymax = epitools::pois.exact(x=deaths, pt=person_time, conf.level=0.95)[,5]*1e5, #CIs
calculated using pos.exact from epitools package
      shape = race_ethnicity
```

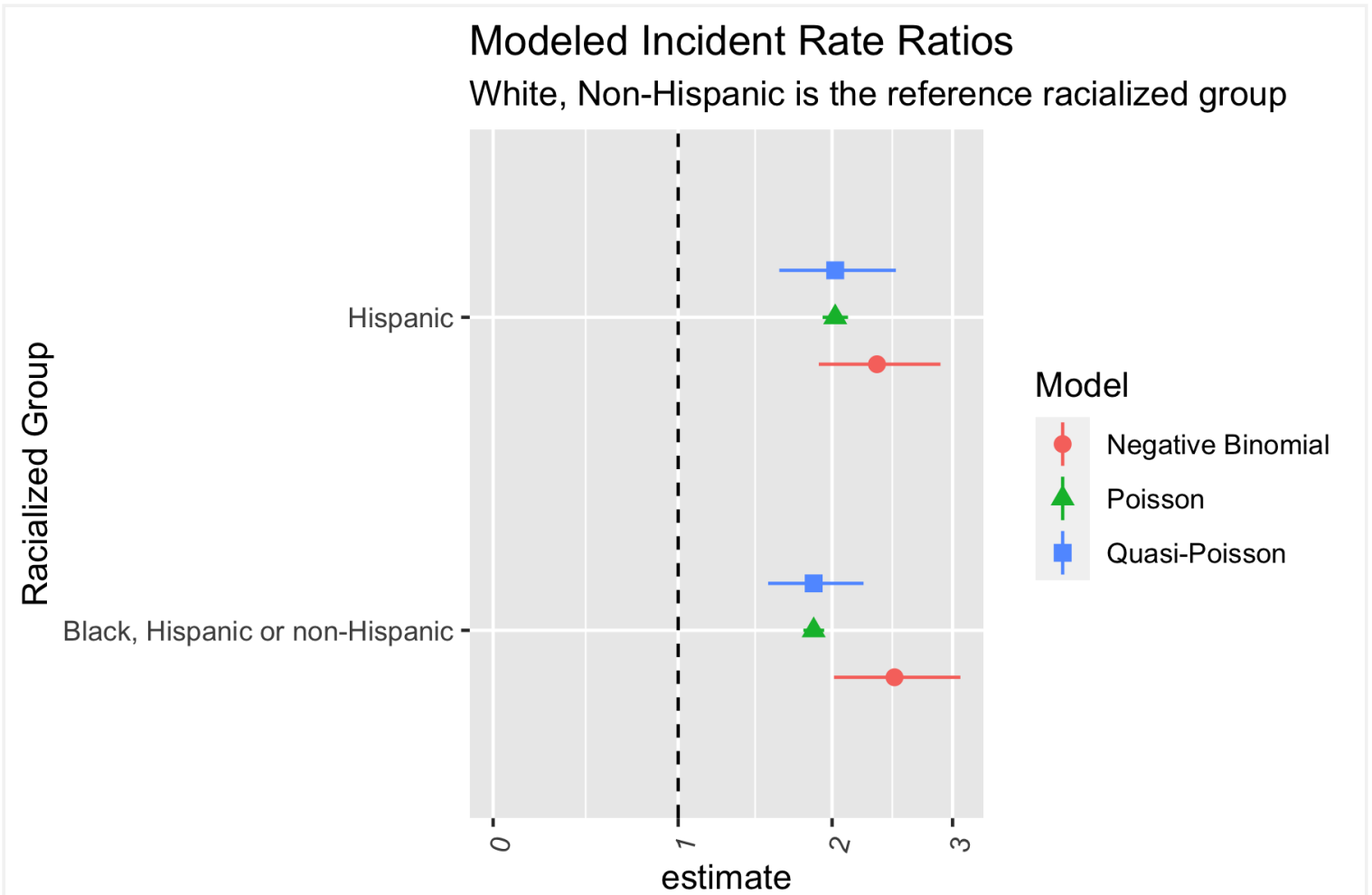
There are racialized disparities in mortality across all age groups. How would you interpret this, and is this what you expected to see? Why or why not?



We can fit models to the aggregate data to adjust for age and obtain relative risk estimates.

> We can fit Poisson, Quasi-Poisson, and Negative Binomial models here, depending on our assumptions about overdispersion. The code below fits the different models and plots them together to demonstrate how modeling assumptions may affect the results. Doing this repeatedly for various analysis can be tedious. One idea could be to write a function that fits all three models given a dataset and model specifications.

```
#Fit a Poisson model with offset for person time to get rate
crude_poisson_model <- glm(deaths ~ race_ethnicity + age_group + offset(log(person_time /
1e5)),
                           data = crude_mortality, family = poisson(link = "log"))

#Quasipoisson model
crude_quasipoisson_model <- glm(deaths ~ race_ethnicity + age_group + offset(log(person_time /
1e5)),
                           data = crude_mortality, family = quasipoisson(link = 'log'))

#Negative binomial model from MASS package
crude_negbin_model <- MASS::glm.nb(deaths ~ race_ethnicity + age_group + offset(log(person_time
/ 1e5)),
                           data = crude_mortality) ##The MASS package is used to fit the
negative binomial model

# extract the model coefficients
crude_results <- bind_rows(
  #The tidy function from broom makes it easy to extract useful model outputs
  broom::tidy(crude_poisson_model, exponentiate = TRUE, conf.int = TRUE) %>% mutate(Model =
"Poisson"),
  broom::tidy(crude_quasipoisson_model, exponentiate = TRUE, conf.int = TRUE) %>% mutate(Model
```



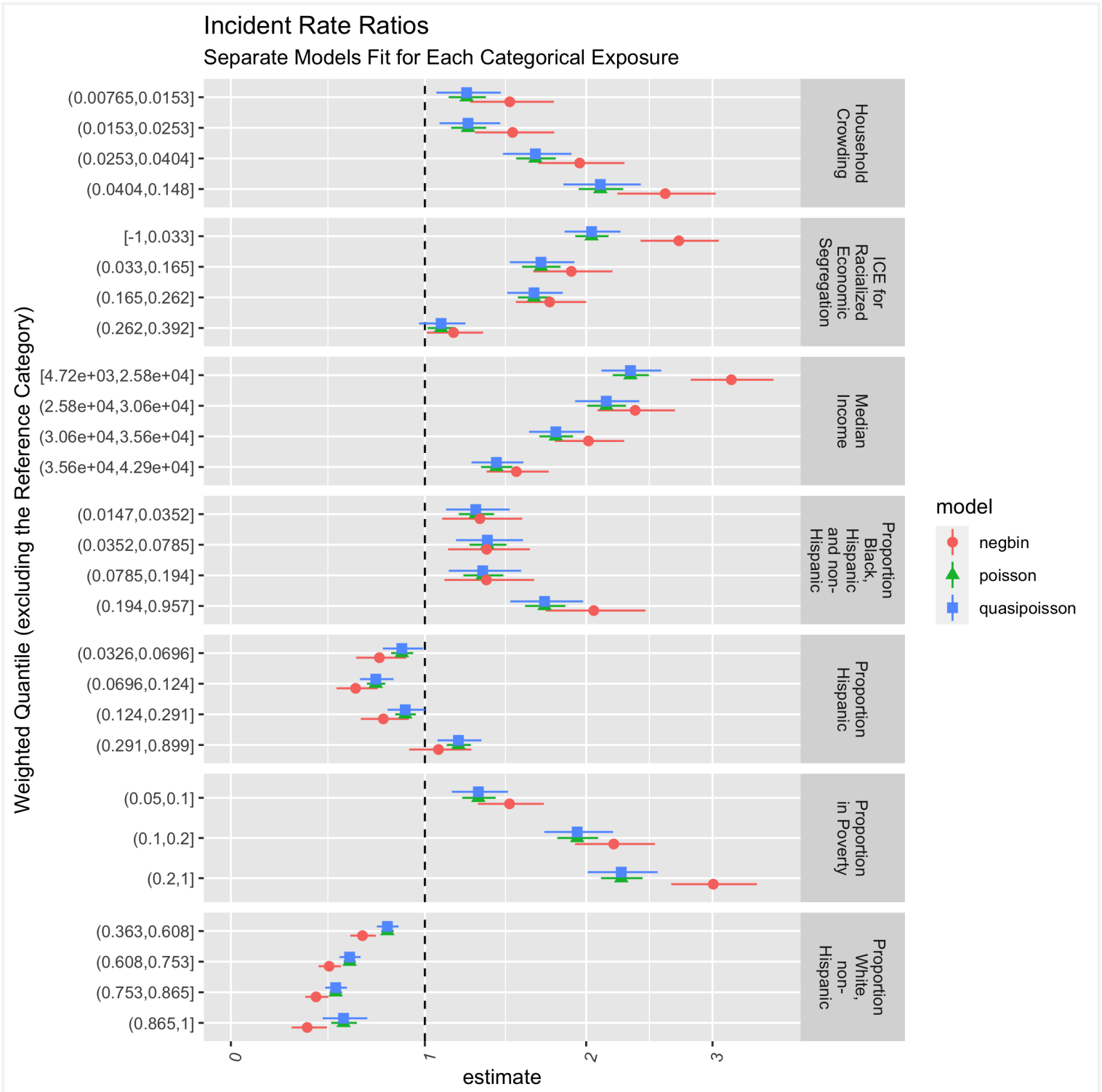
The Poisson and Quasi-Poisson models have the same point estimate but the confidence interval from the Quasi-Poisson model is much greater. The Negative Binomial models have different point estimates and confidence intervals. Is this what you expected? Why or why not?

When presenting, it would likely be easier to communicate these results by showing results from one type of model.

9.7.2 **Gradients in COVID-19 mortality by in relation to ABSMs**

Having looked at overall racialized disparities in COVID-19 mortality, let us now take a look at how mortality varies by ABSMs. Like in the section above, we take age into account by adjusting for age in the model. We fit separate models for each ABSM we are considering.

```
## We will aggregate data across racialized groups since we are looking at overall
relationships
age_ZCTA_mortality <- df %>% st_drop_geometry() %>%
  group_by(geoid, age_group) %>%
  dplyr::summarize(population_estimate = sum(population_estimate),
            deaths = sum(deaths),
            person_time = sum(person_time),
            ICEraceinc_cut = first(ICEraceinc_cut), #The ABSMs are the same for every strata in
each ZCTA, so we can just take the first one.
            prop_in_poverty_cut = first(prop_in_poverty_cut),
            median_income_cut = first(median_income_cut),
            crowding_cut = first(crowding_cut),
            prop_black_cut = first(prop_black_cut),
            prop_hispanic_cut = first(prop_hispanic_cut),
            prop_white_nh_cut = first(prop_white_nh_cut)) %>%
  mutate(mortality_per100k_py = (deaths / person_time)*100000) %>%
  mutate(race_ethnicity="All")

# specify our variables of interest
variables_of_interest <- c("ICEraceinc_cut",
      "prop_in_poverty_cut",
      "median_income_cut",
```

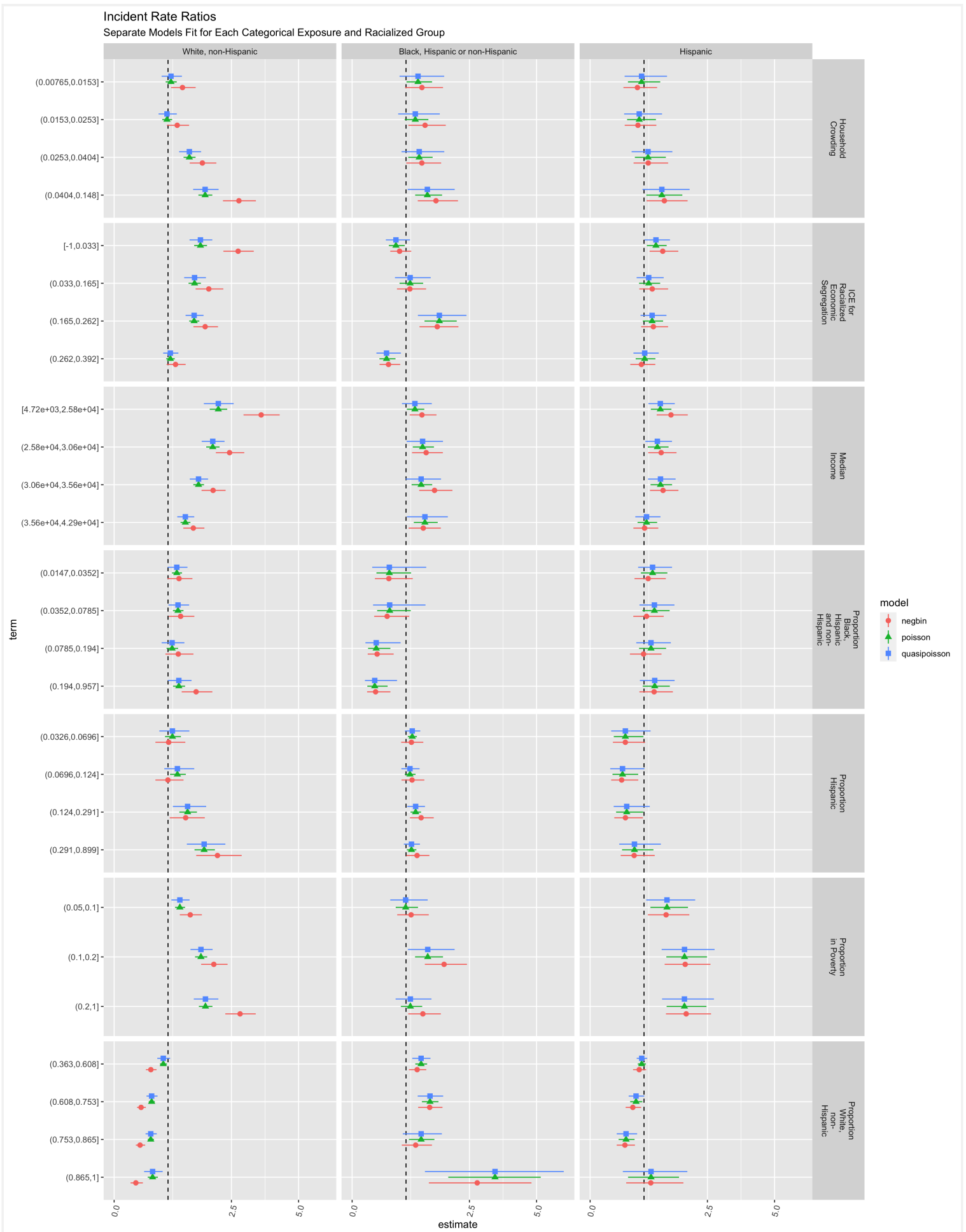


9.7.3 **Gradients in COVID-19 mortality by Racialized Group in relation to ABSMs**

Now we can take a look at how mortality varies in relation to ABSMs, by racialized group. We build on the code in the section above, but now have models for each racialized group and each ABSM.

> Here we fit the three types of models, for three racialized groups, and for seven ABSMs. Writing out 3x3x7 models would take up space and time, and be prone to mistakes. The code below uses the `map` function from the `purrr` package to do this more efficiently. Do you agree with this approach? What are some other ways you would approach this?
>
> You can change to code to focus in on fewer ABSMs or investigate other ABSMs of interest for the presentation.

```r
# specify our variables of interest
variables_of_interest <- c("ICEraceinc_cut",
      "prop_in_poverty_cut",
      "median_income_cut",
      "crowding_cut",
      "prop_black_cut",
      "prop_hispanic_cut",
      "prop_white_nh_cut")

# specify our model formulae with each variable of interest
model_formula <-
  purrr::map(variables_of_interest,
    ~ paste0("deaths ~ ",
             ., " + age_group + offset(log(person_time / 1e5))")
  )

# set the names of the model formulae accordingly
names(model_formula) <- variables_of_interest

# estimate race/ethnicity stratified models
race_ethnicity_stratified_models <-
  df %>%
```



The various models and ABSMs are shown for demonstration. However, it may be better to focus on one particular aspect (e.g. a particular ABSM and model), driven by a specific question and understanding of the context. What would you focus on and why?

9.7.4 Hierarchical and Spatial Models

9.7.4.1 Background

In the examples above, we modeled the associations between individual- and area-based measures with the outcome in aggregate across the county. However, we may also be interested in modeling the risk of mortality in each ZCTA. This can help us better understand the spatial distribution of risk.

We may also want to take into account the fact that these ZCTAs are likely not independent of each other, but rather related to each other in a spatially structured fashion. That is, ZCTAs that are next to each other are more likely to have similar rates than ZCTAs that are farther apart.

One of the *statistical* challenges of estimating the risk in each ZCTA (or any other small areal unit), is that, due to small underlying populations and sparse events interspersed in time, small-area estimation (SAE) of risk and incidence rates can be quite unstable. This has led to the use of various strategies to smooth estimates, by borrowing information from nearby spatial units. Thus, by accounting for the fact that ZCTAs that are closer together are more likely to be similar, we can obtain spatially smoothed risk estimates.

Put another way: let us say that a particular part of the county has high mortality rates. If we look at observations over a limited period of time, in each specific ZCTA separately, we may miss this just by chance due to the fact that the populations in individual ZCTAs are small, and thus fewer events are observed. However, if we pool information from neighboring ZCTAs, we would be able to obtain a more stable estimate of the risk.

Hierarchical Bayesian models are a popular set of tools to conduct such estimations. Often, in such approaches, we will model the Standardized Mortality Ratio (SMR), standardized by age. The relative risk or SMR in area i, (θ_i) is obtained by dividing the expected count (E_i) by the observed count O_i). How we calculate the "expected count" depends on the objectives of our study and analysis. For example, if we know that there are strong age effects, and want to estimate excess risk after accounting for age-composition differences between ZCTAs, we can take the age-strata-specific mortality rates for all of Cook County, and apply it to the age-distribution of each ZCTA. This tells us what the "expected" mortality count would be if each ZCTA experienced Cook County's average age-specific mortality rates. One could also standardize by composition in relation to racialized groups, but then we would not be able to estimate how membership in racialized groups was associated with mortality. Similarly, age-standardization precludes the ability to look at the interaction between age and membership in racialized groups, for example.

9.7.4.2 SMR Calculation

In this case, we will estimate the SMR using age-standardization. For each racialized group in each ZCTA, we obtain the age standardized SMR.

```r
## We will need to have a ZCTA specific ID number that starts from 1 eventually, so let's first make that
df <- df %>% group_by(geoid) %>%
  mutate(id = cur_group_id()) %>% ungroup()


## Calculate overall age-specific mortality
overall_mortality <- df %>% st_drop_geometry %>%
  group_by(age_group) %>%
  summarise(deaths = sum(deaths),
            person_time = sum(person_time)) %>%
  mutate(mortality_per_py = (deaths/person_time))


## Apply this to each ZCTA
smr_df <- df %>% st_drop_geometry %>%
  left_join(overall_mortality[, c("age_group", "mortality_per_py")], by="age_group") %>% #Join the overall age-specific mortality rates
  mutate(Exp = mortality_per_py * person_time, #Multiply the overall age-specific mortality to the person-time in each ZCTA-race-age strata %>%
         Obs = deaths) %>%
  group_by(geoid, id, race_ethnicity) %>%
```

So, in our dataset we can calculate the Observed count O_i, the Expected count E_i (based on age standardization), and the SMR θ_i. The observed count approximately follows a Poisson distribution, such that:

$$O_i \sim Poisson(\theta_i * E_i)$$

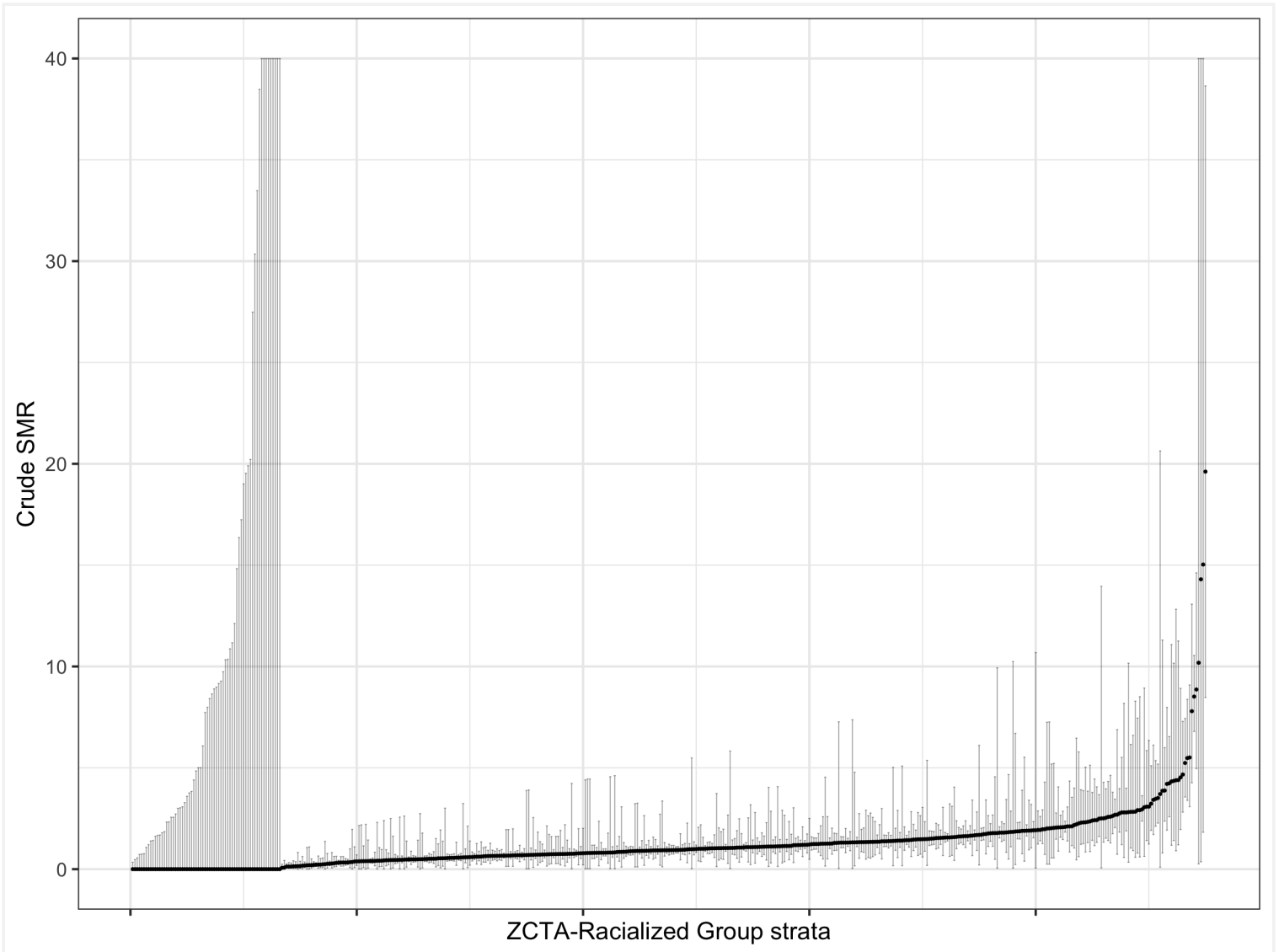
That is, the observed county in any one ZCTA follows a Poisson distribution whose mean is the expected county times a relative risk / SMR. We are interested in this last component since it quantifies the "excess" or "reduced" risk in each area beyond what is "expected" by an age-composition. In our particular case, since we actually have strata of racialized groups within each ZCTA, we can create a hierarchical structure such that:

$$O_{ij} \sim Poisson(\theta_{ij} * E_{ij})$$

Now, we are looking at observed, expected counts and SMR within each racialized group strata j, within each ZCTA i.

Let us visualize these crude SMRs for each ZCTA-racialized group strata, along with their confidence intervals

```r
ggplot(data = smr_df %>% arrange(raw_SMR, desc(Exp)) %>% ungroup() %>%
         mutate(order_id = row_number(),
                raw_CI95up = ifelse(raw_CI95up > 40, 40, raw_CI95up))) + # Since there are large confidence intervals, we are going to cut them off at 40 for easier visualization.
  geom_point(aes(x = order_id, y = raw_SMR), size = 0.2) +
  geom_errorbar(aes(x = order_id, ymin = raw_CI95low, ymax = raw_CI95up),
                size=0.2, alpha=0.4) +
  ylim(0,40) +
  ylab("Crude SMR") +
  xlab("ZCTA-Racialized Group strata") +
  theme_bw() +
  theme(axis.text.x = element_blank())

# ggsave(here("images/09-cook-county-covid/raw_SMR_caterpillar.png"), height = 6, width = 8)
```



Here we can see that there is large variability in these strata-specific SMRs, as well as large confidence intervals for many strata — especially when the predicted rate is zero.

9.7.4.3 **Fitting Models**

Instead of just calculating the crude SMRs, we can model them using a Poisson log-normal model such that:

$$log(\theta_{ij}) = \beta_0 + v_i$$
$$where\ v_i \sim Normal(0, \sigma_v^2)$$

Here there is a county-wide intercept β_0, and ZCTA random effects v_i that quantify the excess/reduced risk in each area from the county-wide average. Note that this is not a spatial model and we are still assuming that the ZCTAs are independent of each other.

However, this model does not have any fixed effects for individual-level variables, so we are estimating the same SMR for each racialized group strata within each ZCTA. We may do just this if we did not have any individual-level data. However, in our case, we are interested in estimating the effect of membership in racialized groups, and we can include that in the fixed part of the model such that:

$$log(\theta_{ij}) = \beta_0 + \beta(race\_ethnicity_j) + v_i$$
$$where\ v_i \sim Normal(0, \sigma_v^2)$$

Here, there is a county-wide intercept for a reference racialized group β_0, βs for each other racialized group. Note that here we are modeling a constant effect for each racialized group across all ZCTAs.

Let us fit this model

```r
formula_poisson_mod1 <- Obs ~ 1 + race_ethnicity + f(geoid, model="iid")

model_poisson_mod1 <- inla(formula_poisson_mod1, family="poisson", data=smr_df, E=Exp,
control.predictor=list(compute=TRUE), control.compute = list(dic = TRUE), verbose = F)
#The control.compute option here calculates the Deviance Information Criteria which we could use
to assess model fit, especially when comparing different formulations
#The control.predictor option here computes the predicted values

# Save the fitted SMRs and their confidence intervals for each ZCTA into the dataframe we have.
smr_df <- smr_df %>% ungroup() %>%
  mutate(poissonmod1_SMR = model_poisson_mod1$summary.fitted.values$mean,
         poissonmod1_CI95low = model_poisson_mod1$summary.fitted.values$`0.025quant`,
         poissonmod1_CI95up = model_poisson_mod1$summary.fitted.values$`0.975quant`) %>%
  left_join(model_poisson_mod1$summary.random$geoid[, c("ID", "mean")], by=c("geoid"="ID")) %>%
  mutate(mean = exp(mean)) %>%
  rename(poissonmod1_RE = mean)
```

Next, as mentioned before, we may want to take into account the fact that neighboring ZCTAs are more closely related. One way to do this is to fit a "Besag-York-Mollie" or BYM model. This builds on the previous model, and partitions the random effects into two components - a spatially structured one, and a spatially unstructured one. A BYM model would be:

$$log(\theta_{ij}) = \beta_0 + \beta(race/ethnicity) + u_i + v_i$$
$$where\ v_i \sim Normal(0, \sigma_v^2)\ and$$
$$u_i \sim Conditional\ Autoregressive(W, \sigma_u^2)$$

What this specifies is that part of the residual variation across the ZCTAs is spatially structured such that neighboring ZCTAs have similar excess risks. This part is indicated by the new u_i introduced above. After accounting for the spatial structure, further residual variation is indicated by v_i. After fitting such models, we can also calculate the proportion of the residual variance that is spatially structured (that is, u_i), and this is known as the spatial fraction.

Let us fit this model:

```r
#Calculate Neighbours matrix
W.nb <- poly2nb(unique(df %>% dplyr::select(id)), snap=0.001)
#W.list <- nb2listw(W.nb, style="B", zero.policy = TRUE)

# Visualize the neighbourhood matrix
coords <- st_coordinates(st_centroid(st_geometry(unique(df %>% dplyr::select(id)))))
plot(st_geometry(unique(df %>% dplyr::select(id))), border="grey")
plot(W.nb, coords, add=TRUE)

#Make adjacency matrix in format INLA can understand
nb2INLA("INLA_adj_mat", W.nb) #this saves a file in the working directory
INLA_adj_mat <- "INLA_adj_mat"

formula_bym_mod1 <- Obs ~ 1 + race_ethnicity + f(id, model="bym2", graph=INLA_adj_mat,
scale.model=TRUE, constr=TRUE)

model_bym_mod1 <- inla(formula_bym_mod1, family="poisson", data=smr_df, E=Exp,
                       control.predictor=list(compute=TRUE),
                       control.compute = list(dic = TRUE), verbose = F)

#Save the fitted model estimates and Confidence intervals into the dataframe
smr_df <- smr_df %>% ungroup() %>%
```

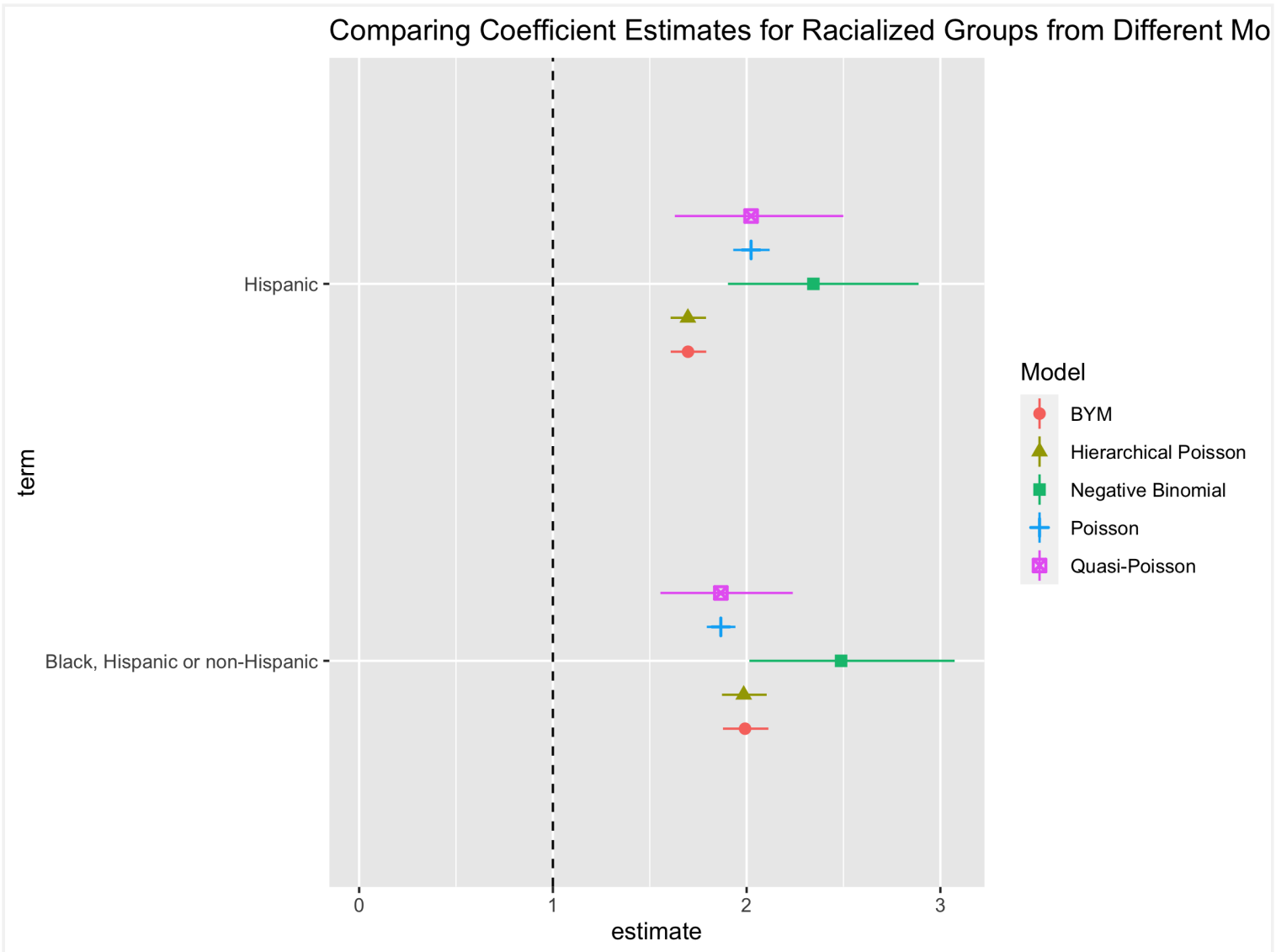
9.7.4.4 Comparing Fitted SMRs

Having fit all of these models, let us visualize all of their components and see what we gain from using these different types of models, and how they differ.

First, we can look at the the estimates of the fixed effects. Since our non-hierarchical models also estimated fixed effects, we can include them in this comparison, with the caveat that we are modeling SMRs in the hierarchical models.

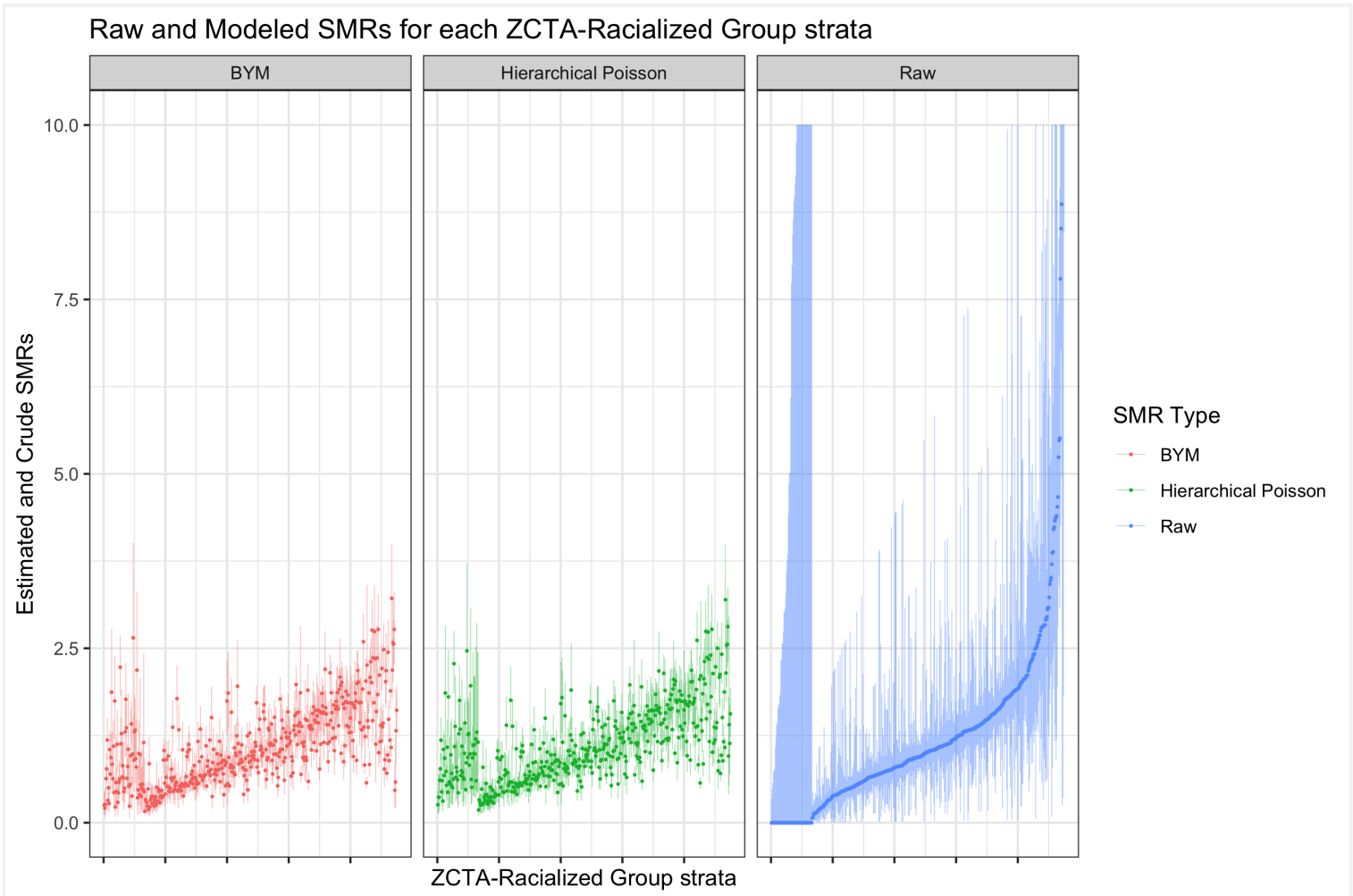
```r
#The tidy functions from broom work on most commonly used models in R, but INLA is not one of them.
#This is a helper function to extract coefficients and credible/"confidence" intervals.
tidy.inla <- function(x){

  # x = model_inla
  term_names <- rownames(x$summary.fixed)

  tibble::as_tibble(x$summary.fixed) %>%
    dplyr::mutate(terms = term_names) %>%
    dplyr::rename(term = terms,
                  estimate = mean,
                  std.error = sd,
                  conf.low = `0.025quant`,
                  conf.high = `0.975quant`) %>%
    dplyr::select(term, estimate, std.error,
                  conf.low, conf.high)
}

coefficients_together <-
  bind_rows(crude_results,
            tidy.inla(model_poisson_mod1) %>%
```

```r
# visualize the effects in each model
coefficients_together %>%
  filter(term != '(Intercept)') %>%
  filter(! stringr::str_detect(term, "age_group")) %>%
  ggplot(aes(x = estimate, xmax = conf.high, xmin = conf.low, y = term, color = Model, shape = Model)) +
  geom_vline(xintercept = 1, linetype = 'dashed') +
  geom_pointrange(position = position_dodge(width=.45)) +
  scale_x_continuous(limits = c(0, NA), n.breaks = 4) +
  ggtitle("Comparing Coefficient Estimates for Racialized Groups from Different Models")

# ggsave(here("images/09-cook-county-covid/coefficient_comparison.png"), height = 6, width = 8)
```



One of the motivations for fitting the hierarchical models is to get stable and smoothed estimates of each ZCTA. Let us visualize the modeled SMRs for each ZCTA-racialized group strata, along with their confidence intervals, and compare them to the crude SMRs we calculated in the beginning.

```r
# Creating a caterpillar plot to visualize the unsmoothed and smoothed SMRs.
caterpillar_plot <- smr_df %>% arrange(raw_SMR, desc(Exp)) %>% ungroup() %>%
  mutate(order_id = row_number(),
         raw_CI95up = ifelse(raw_CI95up > 10, 10, raw_CI95up)) %>%
  dplyr::select(!ends_with("_RE")) %>%
  #Pivot longer to help with ggplot
  pivot_longer(cols = raw_SMR:bymmod1_CI95up,
               names_to = c("SMR_type", ".value"),
               names_sep = "_") %>%
  mutate(SMR_type = ifelse(SMR_type == "bymmod1", "BYM",
                           ifelse(SMR_type == "poissonmod1", "Hierarchical Poisson",
                                  "Raw")))

ggplot(data = caterpillar_plot) +
  geom_point(aes(x = order_id, y = SMR, color = SMR_type,
                 group = SMR_type), size = 0.2) +
  geom_errorbar(aes(x = order_id, ymin = CI95low, ymax = CI95up, color = SMR_type,
                 group = SMR_type),
                size=0.2, alpha=0.4) +
  facet_wrap(~SMR_type) +
  ylim(0,10) +
  ylab("Estimated and Crude SMRs") +
```

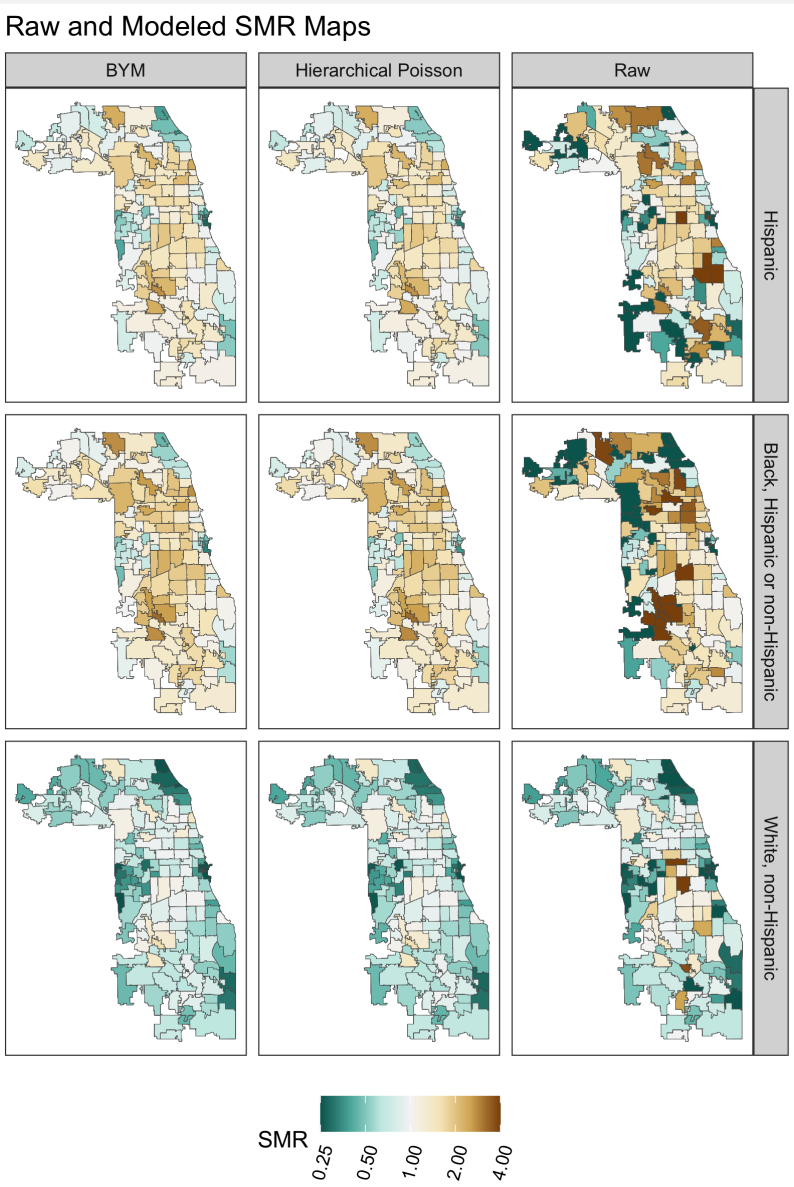


These caterpillar plots demonstrate the noisiness of the raw SMRs. Note that we cut off the y-axis range here to allow comparison within a limited range. Both models smooth the SMRs.

These plots also demonstrates the pitfalls of identifying small areas with extreme rates based on crude SMRs only. They may have a small underlying population, and have high confidence intervals.

The hierarchical models can be most useful when we visualize the estimated ZCTA-specific effects. Since we have racialized group strata within ZCTAs, we can plot maps for each group.

```
# Similar code as for the caterpillar plot above, but we are going to plot just the point
estimates on a map, and not the CIs
smr_df %>%
  arrange(raw_SMR, desc(Exp)) %>% ungroup() %>%
  mutate(order_id = row_number(),
         raw_CI95up = ifelse(raw_CI95up > 10, 10, raw_CI95up)) %>%
  dplyr::select(!ends_with("_RE")) %>%
  pivot_longer(cols = raw_SMR:bymmod1_CI95up,
               names_to = c("SMR_type", ".value"),
               names_sep = "_") %>%
  left_join(df %>% dplyr::select(geoid, geometry), by="geoid") %>% st_as_sf(sf_column_name =
"geometry") %>%
  mutate(SMR_type = ifelse(SMR_type == "bymmod1", "BYM",
                           ifelse(SMR_type == "poissonmod1", "Hierarchical Poisson",
                                  "Raw"))) %>%
  ggplot(aes(fill = SMR)) +
  geom_sf(size = 0.1) +
  facet_grid(forcats::fct_rev(race_ethnicity)~SMR_type) +
  scale_fill_distiller(palette = "BrBG",
                       trans = scales::pseudo_log_trans(sigma=0.01),
                       limits = exp(c(-1,1)*log(4)),
                       breaks = c(0.25,0.5,1,2,4),
```



The spatial fraction is 72%.

> What we see here is that the crude calculated SMRs are very noisy. Using the Poisson hierarchical and BYM models has clear benefits in smoothing the rates. However, there does not seem to be that much difference, at a high level, between the Poisson hierarchical and BYM model.

** IMPORTANT NOTE **

> When we look at these smoothed maps, we can see that the risk is lowest for the White non-Hispanic group. However, the spatial patterning of risk is similar for each racialized group in the maps. We should be cautious about concluding from this that the spatial

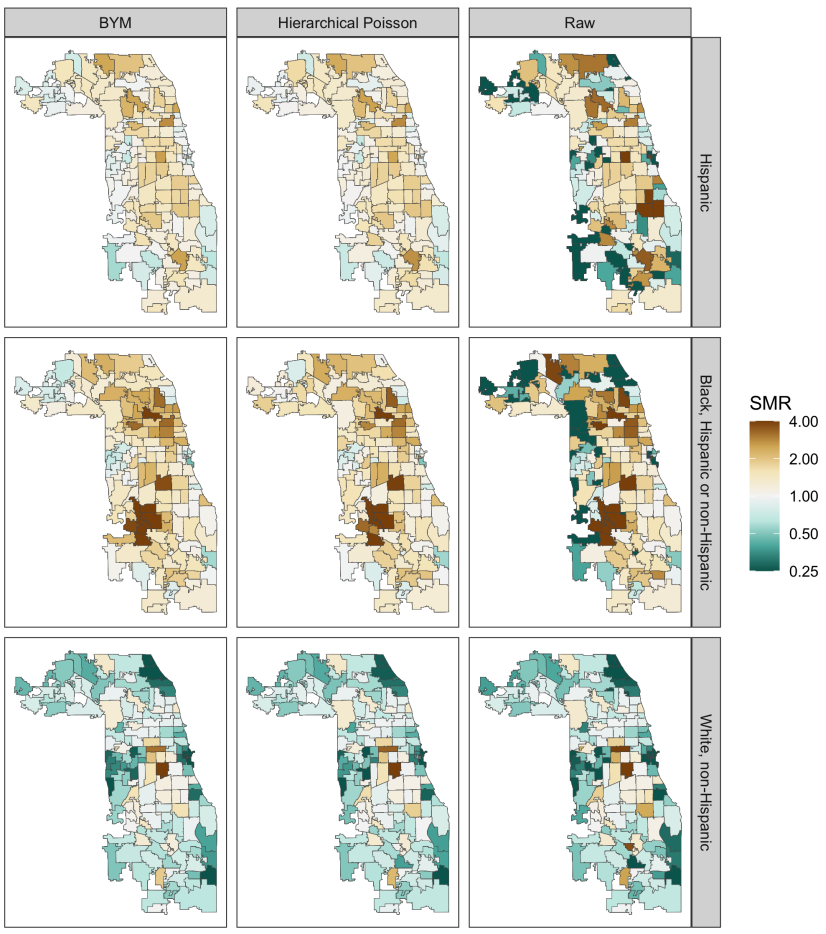
patterning is the same. It is the result of how we constrained the model, as we did not model the spatial component separately for each group. We are going to do this next by fitting a separate models for each racialized group.

```
## Fitting the models for the White Non-Hispanic group
smr_df_wnh <- smr_df %>%
  filter(race_ethnicity == "White, non-Hispanic")
#Poisson
formula_poisson_wnh <- Obs ~ 1 + f(geoid, model="iid")

model_poisson_wnh <- inla(formula_poisson_wnh, family="poisson",
                          data = smr_df_wnh, E=Exp,
                          control.predictor=list(compute=TRUE),
                          control.compute = list(dic = TRUE), verbose = F)

# As above for the overall models, save the fitted values and CIs
smr_df_wnh <- smr_df_wnh %>% ungroup() %>%
  mutate(poissonmodspecific_SMR = model_poisson_wnh$summary.fitted.values$mean,
         poissonmodspecific_CI95low = model_poisson_wnh$summary.fitted.values$`0.025quant`,
         poissonmodspecific_CI95up = model_poisson_wnh$summary.fitted.values$`0.975quant`) %>%
  left_join(model_poisson_wnh$summary.random$geoid[, c("ID", "mean")], by=c("geoid"="ID")) %>%
  mutate(mean = exp(mean)) %>%
  rename(poissonmodspecific_RE = mean)


#BYM
```

The spatial fraction of the residual variance is 61% for the White, non-Hispanic model; 67.8% for the Black; Hispanic and non-Hispanic model; and 77% for the Hispanic model.

We can visualize the smoothed SMRs from these separate models.

```
#Combine all of the racialized group specific results into one
smr_df <- smr_df %>%
  left_join(
    bind_rows(smr_df_wnh, smr_df_bhnh, smr_df_h) %>%
      dplyr::select(geoid, race_ethnicity, poissonmodspecific_SMR:bymmodspecific_RE),
    by = c("geoid", "race_ethnicity"))

# Similar to how we plotted the maps for overall fitted SMRs in each ZCTA, we do the same by
racialized group
smr_df %>%
  arrange(raw_SMR, desc(Exp)) %>% ungroup() %>%
  mutate(order_id = row_number(),
         raw_CI95up = ifelse(raw_CI95up > 10, 10, raw_CI95up)) %>%
  dplyr::select(!ends_with("_RE")) %>%
  pivot_longer(cols = c(raw_SMR:raw_CI95up, poissonmodspecific_SMR:bymmodspecific_CI95up),
               names_to = c("SMR_type", ".value"),
               names_sep = "_") %>%
  left_join(df %>% dplyr::select(geoid, geometry), by="geoid") %>% st_as_sf(sf_column_name =
"geometry") %>%
  mutate(SMR_type = recode(SMR_type,
                           bymmodspecific = "BYM",
                           poissonmodspecific = "Hierarchical Poisson",
```

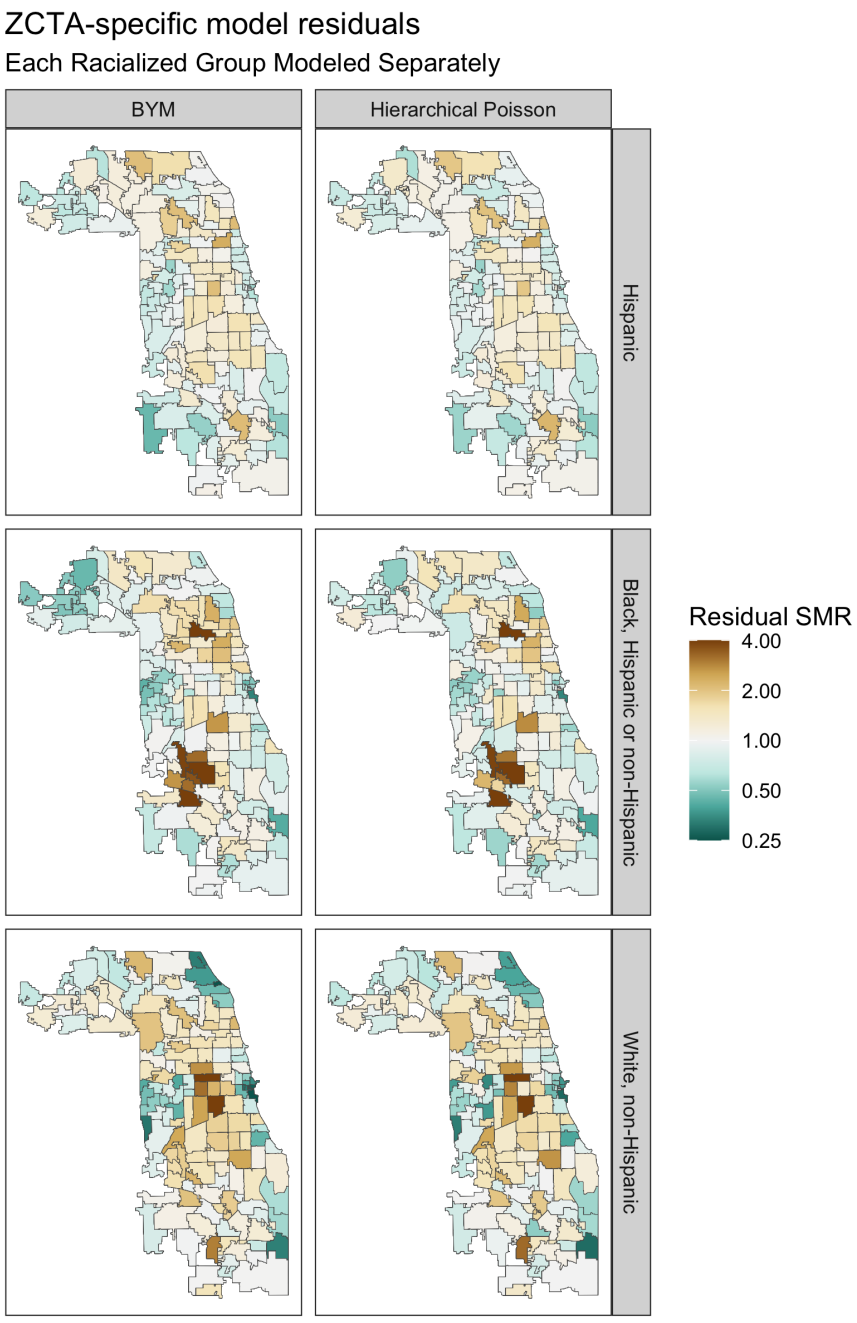


This approach reflects the different spatial patterning of risks for each racialized group, while also smoothing rates.

9.7.4.5 **Additional Analysis**

We may also be interested in looking at the area-specific **residuals**. These give us a sense of how much each area's risk deviates from the county-wide mean, conditional on any covariates we include in the model. We can visualize the residuals of the models above (where no covariates were included).

```
smr_df %>%
  arrange(raw_SMR, desc(Exp)) %>% ungroup() %>%
  mutate(order_id = row_number(),
         raw_CI95up = ifelse(raw_CI95up > 10, 10, raw_CI95up)) %>%
  dplyr::select(!ends_with(c("_SMR", "_CI95up", "_CI95low"))) %>%
  pivot_longer(cols = c(poissonmodspecific_RE:bymmodspecific_RE),
               names_to = c("RE_type", ".value"),
               names_sep = "_") %>%
  left_join(df %>% dplyr::select(geoid, geometry), by="geoid") %>% st_as_sf(sf_column_name =
"geometry") %>%
  mutate(SMR_type = recode(RE_type,
                           bymmodspecific = "BYM",
                           poissonmodspecific = "Hierarchical Poisson")) %>%
  ggplot(aes(fill = RE)) +
  geom_sf(size = 0.1) +
  facet_grid(forcats::fct_rev(race_ethnicity)~SMR_type) +
  scale_fill_distiller(palette = "BrBG",
                       trans = scales::pseudo_log_trans(sigma=0.01),
                       limits = exp(c(-1,1)*log(4)),
                       breaks = c(0.25,0.5,1,2,4),
                       oob = scales::squish) +
  theme_bw() +
```



Finally, as in the non-hierarchical models above, we can also add ABSMs to these hierarchical models. The model would be as follows:

$$log(\theta_{ij}) = \beta_0 + \beta(ABSM_i) + u_i + v_i$$
$$where\ v_i \sim Normal(0, \sigma_v^2)\ ,\ and$$
$$u_i \sim Conditional\ Autoregressive(W, \sigma_u^2)$$

We won't explore these models here, but we could simply add the covariates to the model equations we fit in the code above.

« 8 Case Study 2: Breast Cancer Mortality in Massachusetts

10 Case Study 4: Case Study on Temporal Trends using American Community Survey (ACS) data (2012-2019) »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

10 Case Study 4: Case Study on Temporal Trends using American Community Survey (ACS) data (2012-2019)

By: Dena Javadi, Tamara Rushovich, Christian Testa

10.1 Introduction

The American Community Survey (ACS), conducted by the U.S. census bureau is a yearly survey (monthly samples providing annual updates) that provides vital information on educational attainment, jobs, occupations, veterans, house ownership, disability, poverty status, demographics, and more.

To learn more about the survey methodology used, please visit the ACS website.

10.2 Motivation, Research Questions, and Learning Objectives

The goal of this case study is to gain familiarity with the ACS, visualize trends over time, model the effect of different Area-Based Social Metrics (ABSMs) on trends in a particular outcome.

The outcome of interest in this case study is health insurance. The specific learning objectives are to:

- Download health outcome and ABSM data from the ACS
- Visualize (plot) and map data
- Characterize trends in outcome of interest by ABSMs
- Work with estimate uncertainty (margin of error) data to:
    - Aggregate it up from stratified estimates (like aggregating private and public insurance rates together into overall insurance rates)
    - Visualize uncertainty for individual areal units over time
    - Model trends over time

Although this case study will follow insurance rates over time from specified counties, the skills learned should be generalizable to pulling other ACS variables at other geographic levels.

In this case study, we are particularly interested in observed differences in health insurance coverage across four US states (Massachusetts, California, Texas, and Florida) from 2012-2019. Our research questions are:

1. What are the trends in health insurance coverage from 2012-2019 in each state and what is the level of certainty in these trends?
2. How is county-level poverty associated with health insurance coverage from 2012-2019 in each state?
3. How is county-level racialized economic segregation as measured by the Index of Concentration at the Extremes (ICE) associated with health insurance coverage from 2012-2019 in each state?

> In this next section, **Downloading your Data**, we will show you how to download ACS data by querying the census API and how to manipulate the data into the format we need for the rest of the analysis. This case study investigates the ACS variable of health insurance in four states. See if you can replicate the analysis with a different ACS variable or in different states!

10.3 **Downloading your data**

Explore the ACS data dictionary to identify the variables of interest and to create your own data dictionary. Here is an example of the 2016 dictionary.

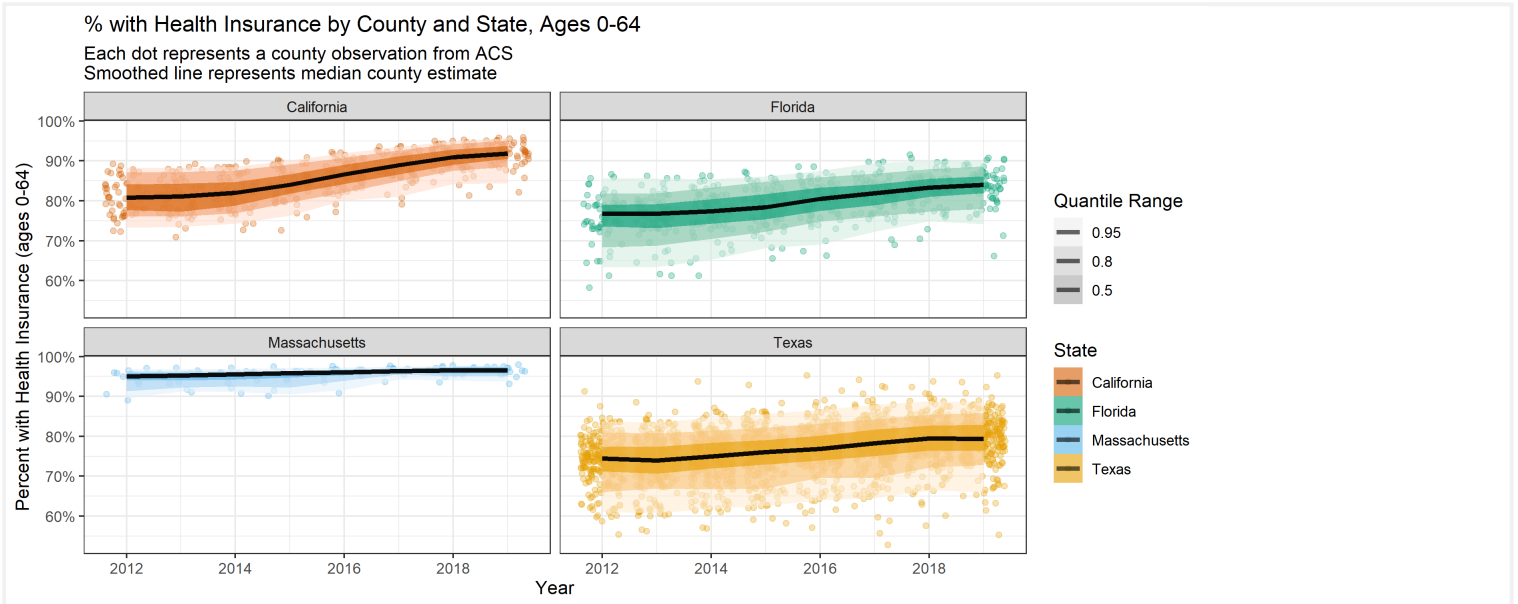
```
# We need to load the following packages:
library(tidycensus)
library(tidyverse)
library(tidyr)
library(magrittr) # magrittr defines the %<>% operator, read more here:
https://magrittr.tidyverse.org/
library(patchwork)

# build a data dictionary - based on ACS 2013
data_dictionary <- tibble::tribble(
  ~variable,     ~shortname,                       ~description,
  "B18135_002",  'under18_denom',                   "Estimate!!Total!!Under 18 years",
  "B18135_004",  'under18_insured_with_disability', "Estimate!!Total!!Under 18 years!!With a
disability!!With health insurance coverage",
  "B18135_009",  'under18_insured_no_disability',   "Estimate!!Total!!Under 18 years!!No
disability!!With health insurance coverage",
  "B18135_013",  'adult_denom',                     "Estimate!!Total!!18 to 64 years",
  "B18135_015",  'adult_insured_with_disability',   "Estimate!!Total!!18 to 64 years!!With a
disability!!With health insurance coverage",
  "B18135_020",  'adult_insured_no_disability',     "Estimate!!Total!!18 to 64 years!!No
disability!!With health insurance coverage",
  # ICEraceinc variables
```

Congrats! You've downloaded your dataset!

10.4 **Visualizing your Data**

Now on to visualizing estimates:

```
library(colorblindr)
# create plot of county level % with health insurance by year and smooth with geom_smooth
method = 'loess' by year
ggplot(health_insurance, aes(x = year, y = estimate_zero_to_64_insurance_prop, color = state))
+
  geom_jitter(alpha = 0.6, height = 0) +
  geom_smooth(method = 'loess', color = 'dimgrey') +
  facet_wrap(~state) +
  scale_fill_manual(
    values = setNames(palette_OkabeIto[c(1,2,3,6)],
                      c('Texas', 'Massachusetts', 'Florida', 'California'))
                      ) +
  scale_color_manual(
    values = setNames(palette_OkabeIto[c(1,2,3,6)],
                      c('Texas', 'Massachusetts', 'Florida', 'California'))
                      ) +
  xlab("Year") +
  ylab("Percent with Health Insurance (age 0-64)") +
  scale_y_continuous(labels = scales::percent_format()) +
  ggtitle("% with Health Insurance by County and State",
          "Each dot represents a county observation from ACS\nSmooth fit shows average county
observation")
```



Now, we might want to map our quantities of interest!

Since we didn't download the geometry with our queries to the Census API through `tidycensus`, we can now download our county geometry from `tigris` and merge it in.

We didn't want to download the geometry data in our calls to `tidycensus` for two reasons:

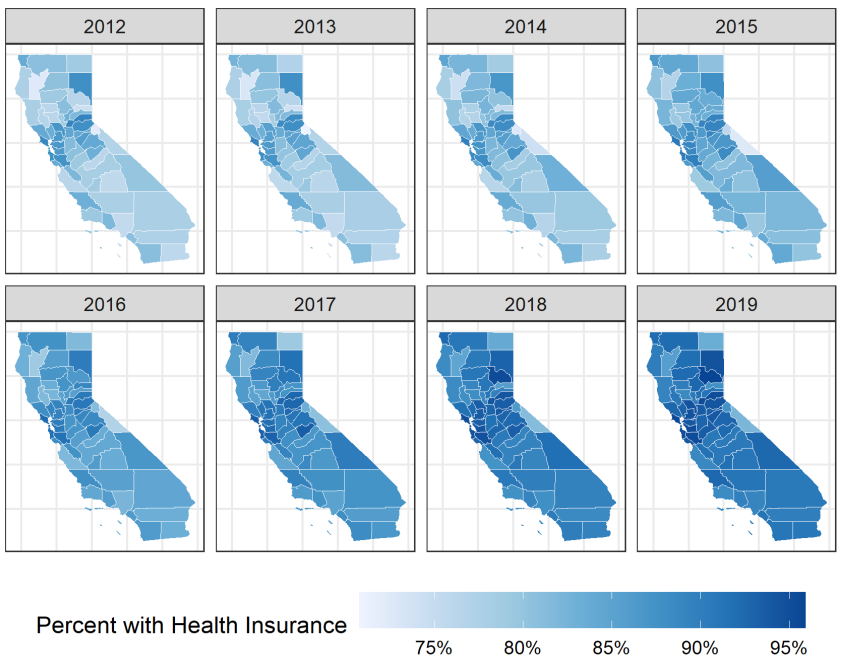
1. Because the `tidycensus` package returns tidy formatted data (one variable and observation per row), our geometry data would be repeated many times taking up more space than necessary; and
2. Because it's easier to do our aggregation and data manipulation without the spatial/geometry data and to add them in afterwards as some functions from `dplyr` or the `tidyverse` may not work well with spatial data frames.

When you look at the maps, make sure to note the scales presented in the legend because sometimes they differ (i.e. pay attention to what the range in the legend represents)!

```
counties_sf <-
  tigris::counties(
    state = c('CA', 'FL', 'MA', 'TX'),
    cb = TRUE,
    resolution = '20m') # here we're downloading 1:20m resolution data

# make sure to do an inner join since we have more than 1 row per county
# for more information on different types of joins: https://statisticsglobe.com/r-dplyr-join-inner-left-right-full-semi-anti
health_insurance_sf <- inner_join(
  counties_sf %>% select(GEOID, STATE_NAME), # when joining spatial data in, put the spatial data argument first so
  # the output inherits the spatial data class
  health_insurance,
  by = c('GEOID' = 'GEOID'))

# Function to make maps for each of California, Florida, Massachusetts, Texas
# counties over time
map_health_insurance_over_time <- function(state_name, outcome_var,
                                           fill_label = 'Percent with Health Insurance') {
  health_insurance_sf %>%
    filter(STATE_NAME == state_name) %>%
```

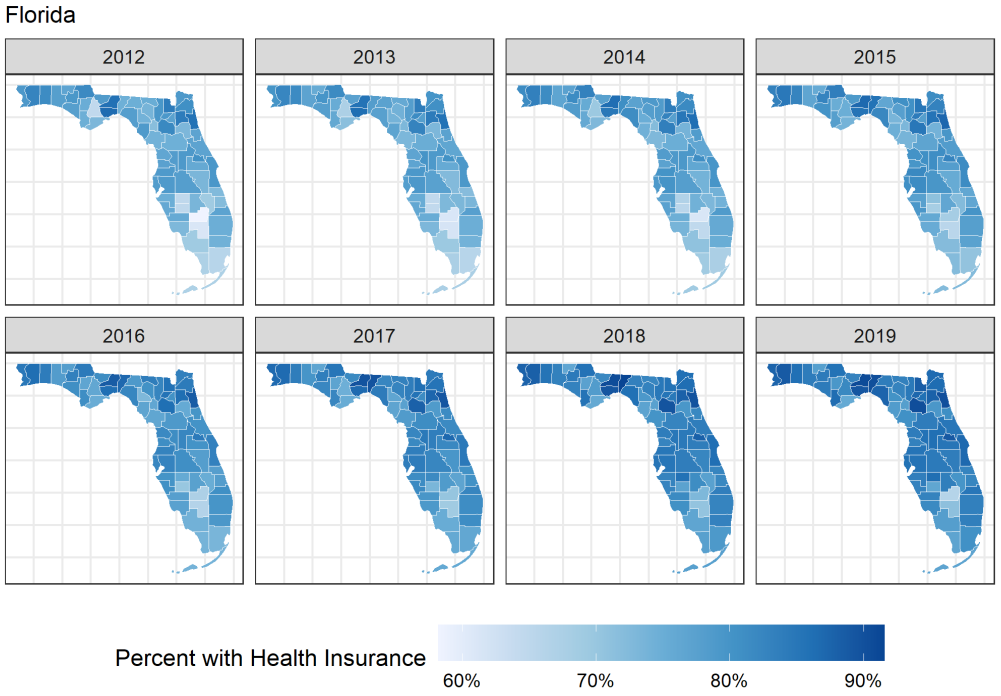
Percent with Health Insurance by County over Time

California



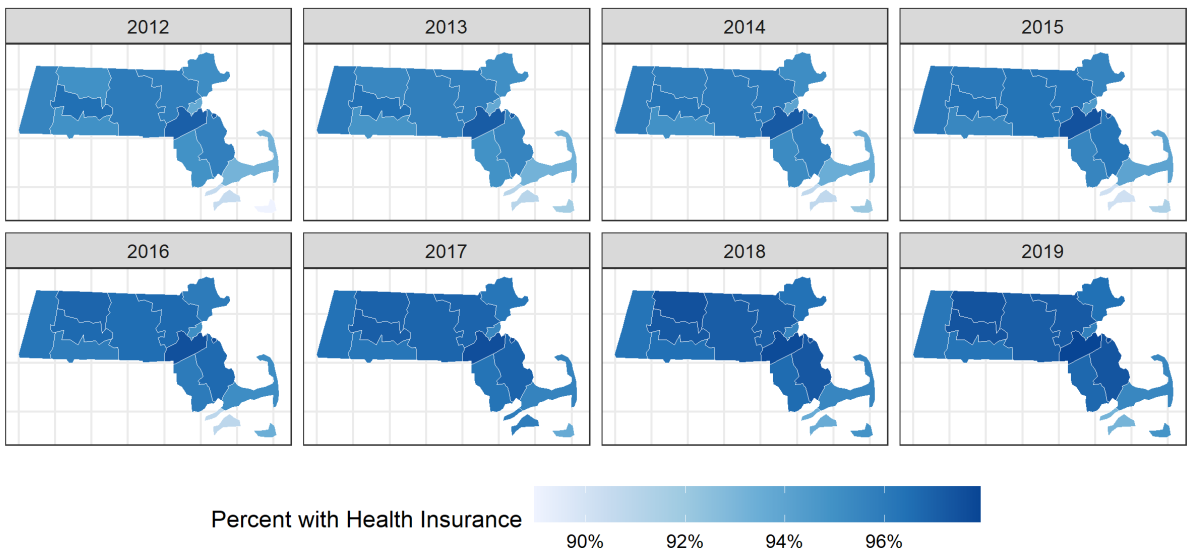
Percent with Health Insurance by County over Time

Florida



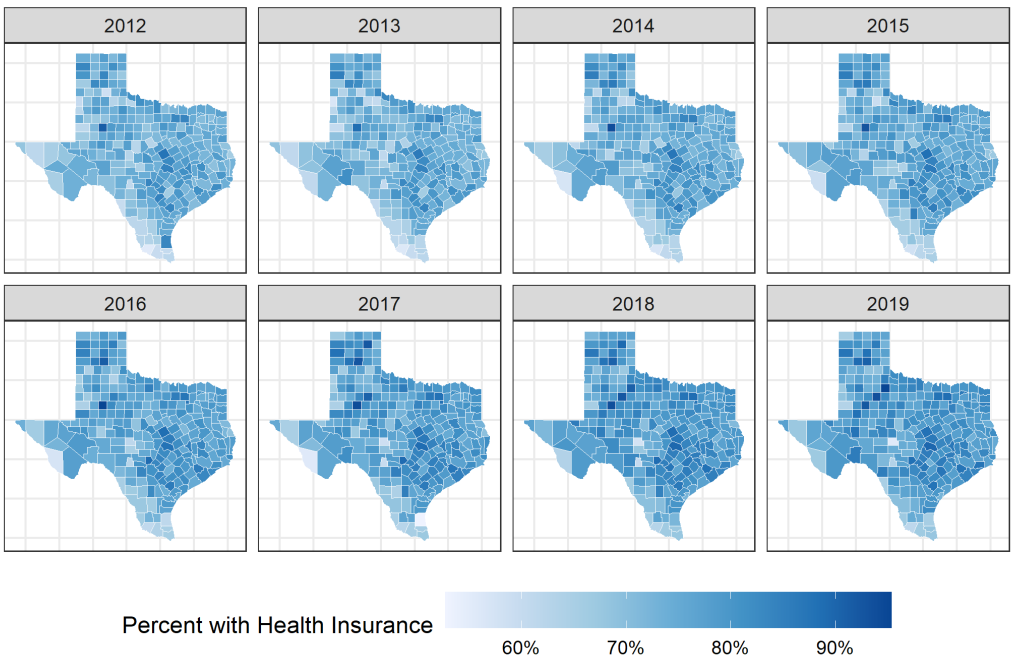
Percent with Health Insurance by County over Time

Massachusetts



Percent with Health Insurance by County over Time

Texas



Notice, in Texas there is one county that seems to experience a decrease in insurance coverage over time. This seems strange! What do you think is the reason for this? Let's investigate further.

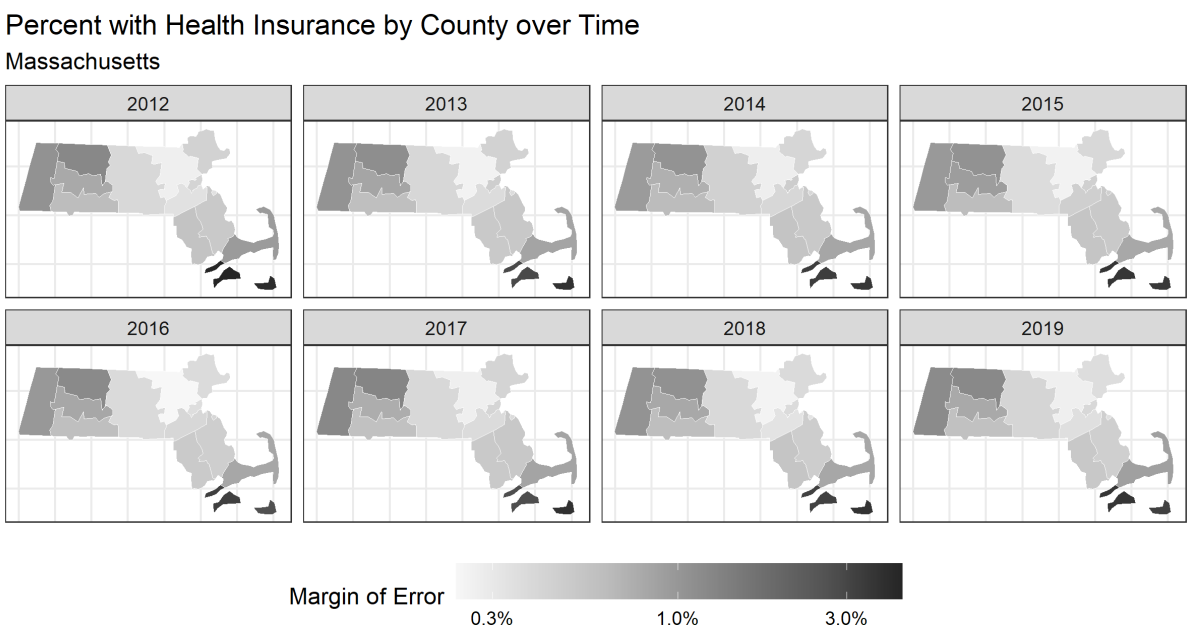
One way to identify which county in Texas appears to be decreasing is by using the R package plotly. The function ggplotly, produces a map that will allow you to hover your cursor over a county to see the name of the county and the percent with health insurance. This is how we can identify that the percent with health insurance in McCulloch County appears to be decreasing over time. Try it out below!

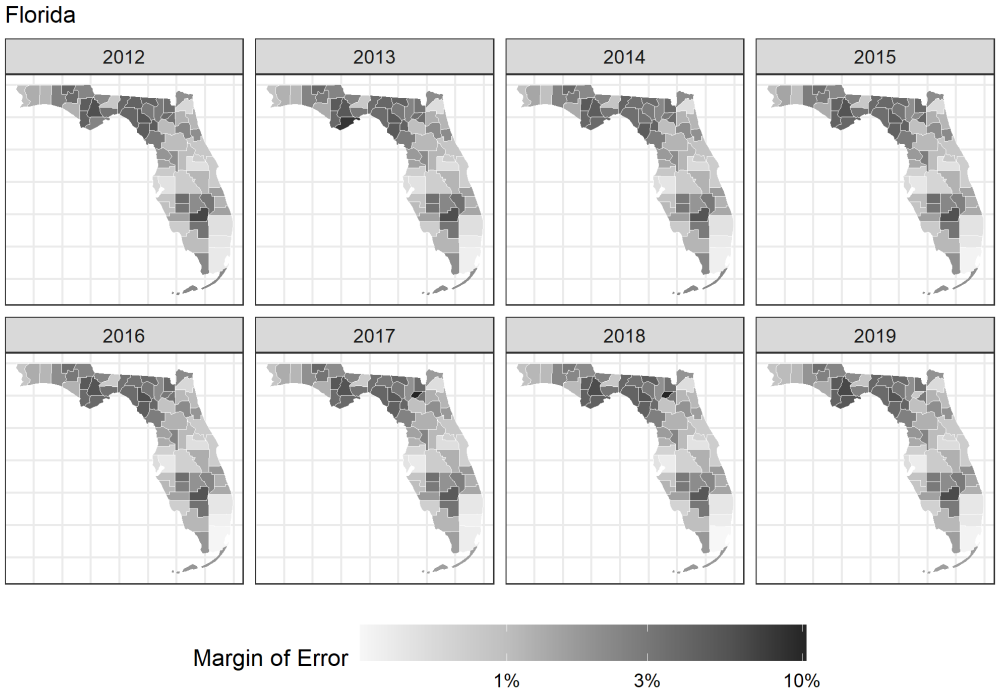
```
library(plotly)
ggplotly(map_health_insurance_over_time('Texas', estimate_zero_to_64_insurance_prop))

#Now we can take a closer look at McCulloch County, TX and see that it has a large margin of error!
health_insurance %>%
  filter(NAME == 'McCulloch County, Texas') %>%
  ggplot(aes(x = year, y = estimate_zero_to_64_insurance_prop,
             ymax = pmin(moe_zero_to_64_insurance_prop/2 + estimate_zero_to_64_insurance_prop, 1),
             ymin = pmax(estimate_zero_to_64_insurance_prop - moe_zero_to_64_insurance_prop/2), 0)) +
  geom_ribbon(fill = '#3182BD', alpha = .8, size = 0) +
  geom_line() +
  geom_point() +
  scale_y_continuous(labels = scales::percent_format()) +
  ylab("Percent with Health Insurance") +
  ggtitle("% with Health Insurance, McCulloch County, Texas",
          "90% Margin of Error and Estimates Shown")

# One reason that a county would have a large margin of error is if it has a small population size. Let's see what the population size is of McCulloch County, TX.
```

For more information on how the ACS calculates margin of error, see https://www.census.gov/data/academy/webinars/2020/calculating-margins-of-error-acs.html

The margin of error is a measure of uncertainty of the estimate. Often the width of the margin of error has to do with the sample size, whereby very small sample sizes yield estimates with large margins of error. Why do you think it is important to communicate uncertainty in the data to the public?

```
# visualize the margin of error data.
map_health_insurance_over_time('Texas', outcome_var = moe_zero_to_64_insurance_prop,
                               fill_label = 'Margin of Error') +
  scale_fill_distiller(
    palette = 'Greys',
    direction = 1,
    trans = 'log10',
    labels = scales::percent_format()
  )

ggsave("images/10-temporal-health-insurance/texasmoe.png", width = 8, height = 5)

map_health_insurance_over_time('Massachusetts',
                               outcome_var = moe_zero_to_64_insurance_prop,
                               fill_label = 'Margin of Error') +
  scale_fill_distiller(
    palette = 'Greys',
    direction = 1,
    trans = 'log10',
    labels = scales::percent_format()
  )
```





Congrats! You can now visualize your estimates and uncertainty in plots and maps.

10.5 **Adding Area Based Social Metrics**

Now let's look at our ABSMs. We are going to assess the association between poverty and health insurance and between racialized economic segregation (as measured by the ICE) and health insurance.

First, we will assess the association between poverty and health insurance by plotting scatter plots of the county percent poverty and the county percent with health insurance. We will add a smoothed line to the scatter using linear regression. In plot 1, we will use unweighted data, and in plot 2, will use the inverse of the margin of error to weight the data. Can you see a difference when the inverse margin of error weights are included?

Next, we will plot the trend in heath insurance status by quintiles of poverty and ICE using loess smoothing.

```
# look at relationship between poverty and insurance proportion
plt1 <-
health_insurance %>%
  filter(state == 'Texas') %>%
  ggplot(aes(x = poverty_prop, y = estimate_zero_to_64_insurance_prop)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  facet_wrap(~year, nrow = 2) +
  xlab("Percent in Poverty") +
  ylab("Percent with Health Insurance") +
  scale_x_continuous(labels = scales::percent_format()) +
  scale_y_continuous(labels = scales::percent_format()) +
  ggtitle("Relationship Between Percent in Poverty and Percent with Health Insurance",
          "County Observations, Texas -- Unweighted")

# the same as above, but with inverse margin of error weighting
plt2 <-
health_insurance %>%
  filter(state == 'Texas') %>%
  ggplot(aes(x = poverty_prop, y = estimate_zero_to_64_insurance_prop, weight =
1/moe_zero_to_64_insurance_prop)) +
  geom_point() +
```

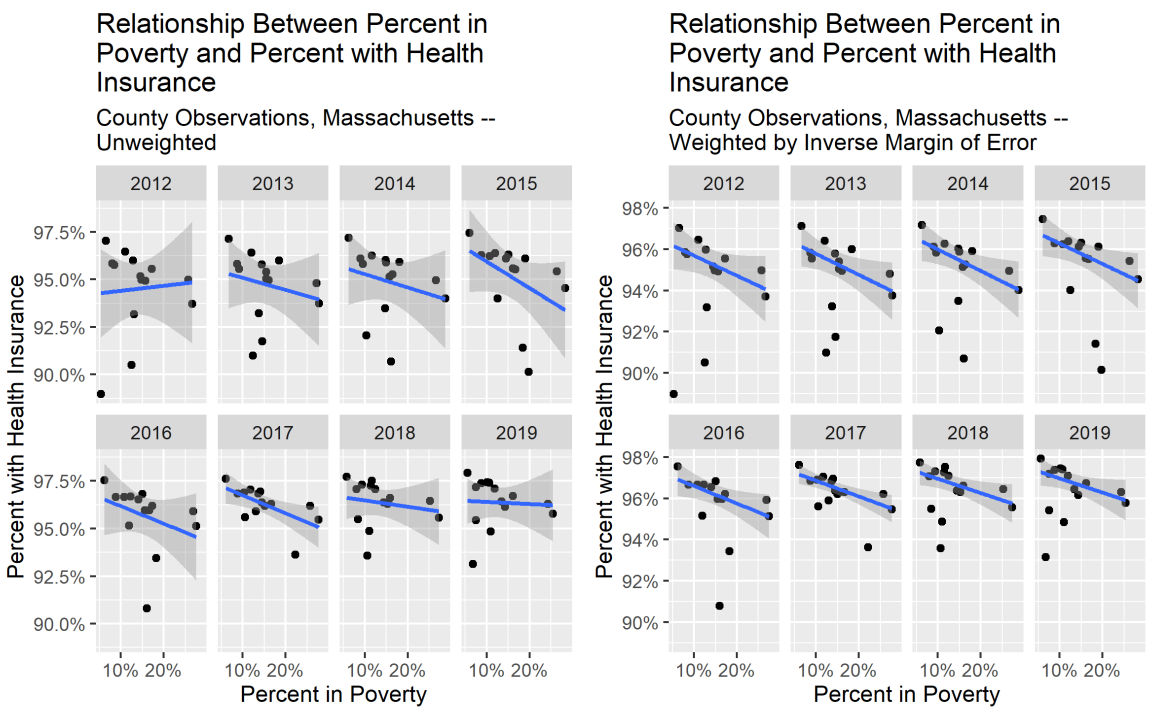


Figure 10.1: County observations of adult (age 19-64) health insurance rates in California

Notice the difference in weighted vs unweighted estimates.

Now we will explore the relationships between ICE and health insurance, weighted by the inverse margin of error as well as poverty and health insurance, weighted by the inverse margin of error.

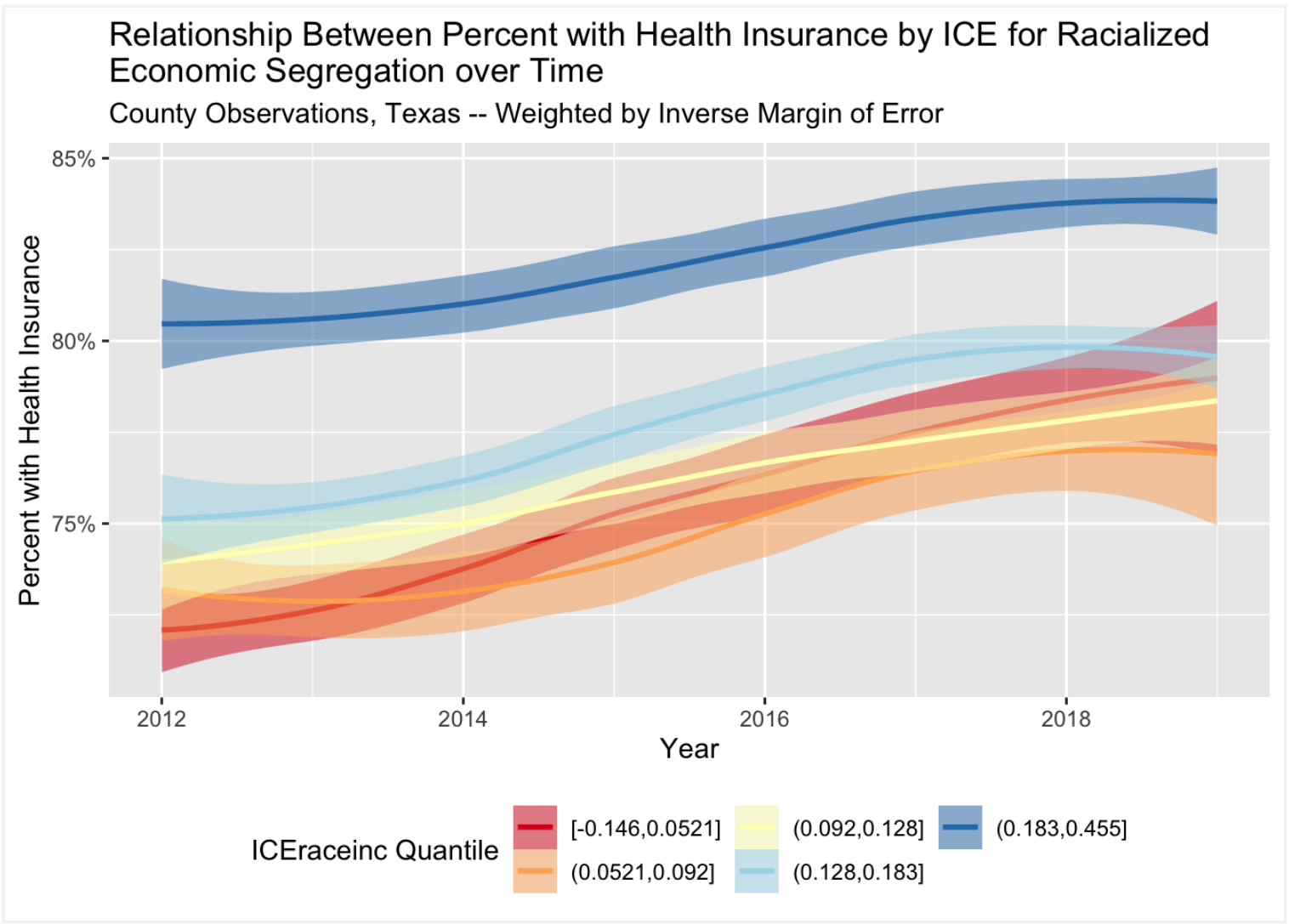
```
# look at the relationship with ICEraceinc and insurance, weighted by inverse margin of error
health_insurance %>%
  filter(state == 'Texas') %>%
  ggplot(aes(x = ICEraceinc, y = estimate_zero_to_64_insurance_prop, weight =
1/moe_zero_to_64_insurance_prop)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  facet_wrap(~year, nrow = 2) +
  xlab("ICEraceinc") +
  ylab("Percent with Health Insurance") +
  scale_y_continuous(labels = scales::percent_format()) +
  ggtitle("Relationship Between ICE for Racialized Economic Segregation and \nPercent with Health
Insurance",
          "County Observations, Texas -- Weighted by Inverse Margin of Error")
```

```
## `geom_smooth()` using formula 'y ~ x'
```


```
# smoothed insurance relationship with ICEraceinc quantiles over time
# using loess smoothing
health_insurance %>%
  filter(state == 'Texas') %>%
  ggplot(aes(x = year, y = estimate_zero_to_64_insurance_prop,
             color = cut(ICEraceinc, quantile(ICEraceinc, seq(0, 1, .2)), include.lowest = T),
             fill = cut(ICEraceinc, quantile(ICEraceinc, seq(0, 1, .2)), include.lowest = T),
             weight = 1/moe_zero_to_64_insurance_prop)) +
  # geom_jitter(height = 0) +
  geom_smooth(alpha = 0.5, alpha = .5) +
  scale_color_brewer(palette = 'RdYlBu') +
  scale_fill_brewer(palette = 'RdYlBu') +
  xlab("Year") +
  ylab("Percent with Health Insurance") +
  guides(fill = guide_legend(nrow = 2), color = guide_legend(nrow = 2)) +
  labs(fill = "ICEraceinc Quantile", color = "ICEraceinc Quantile") +
  theme(legend.position = 'bottom') +
  scale_y_continuous(labels = scales::percent_format()) +
  ggtitle("Relationship Between Percent with Health Insurance by ICE for Racialized \nEconomic
Segregation over Time",
          "County Observations, Texas -- Weighted by Inverse Margin of Error")
```

```
## Warning: Duplicated aesthetics after name standardisation: alpha
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

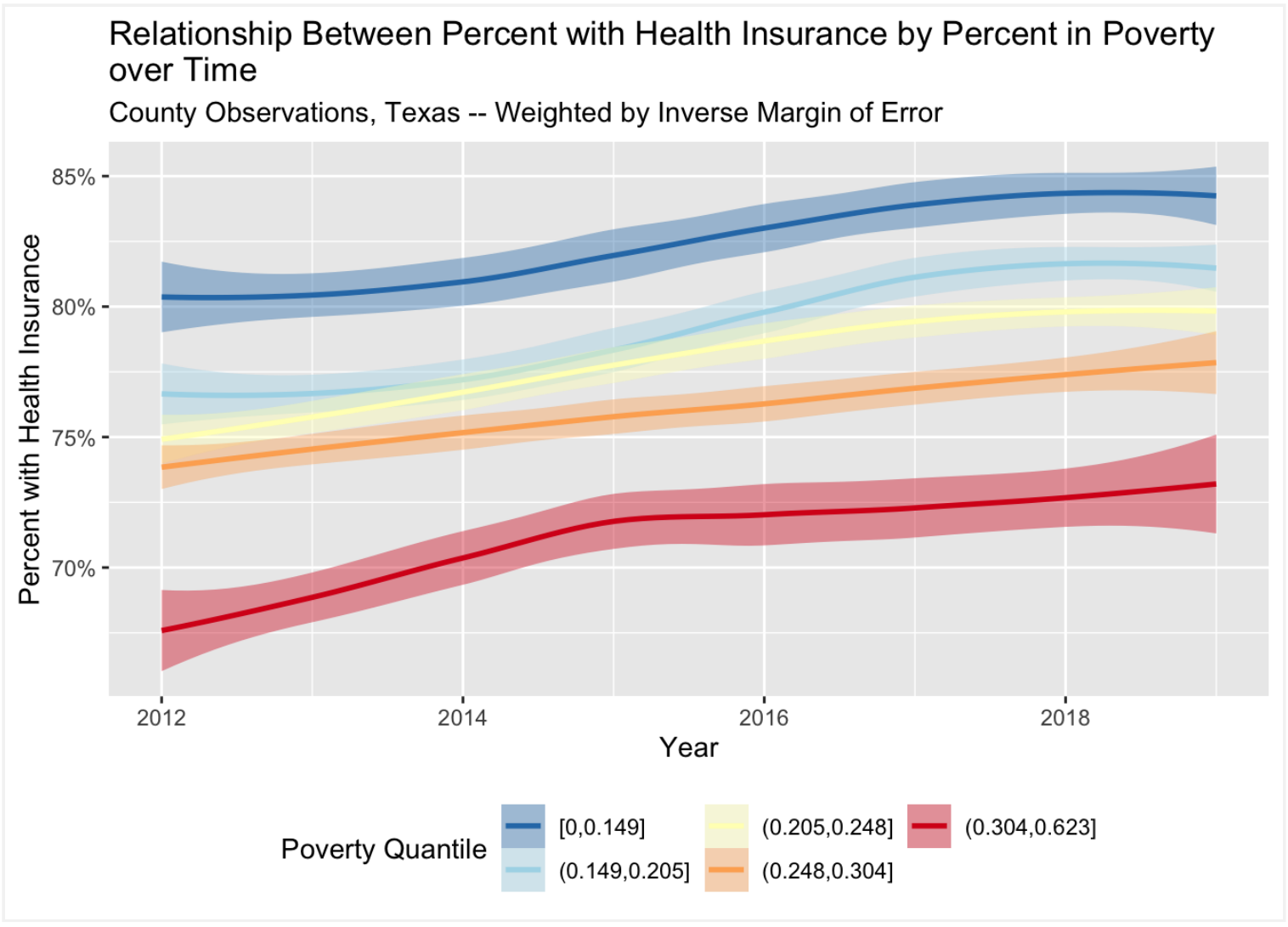



```
ggsave("images/10-temporal-health-insurance/healthinsurance_ICE_time.png", width = 8, height = 5)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
# smoothed relationship between poverty (cut in quantiles) and insurance rates
# using loess smoothing
health_insurance %>%
  filter(state == 'Texas', ! is.na(poverty_prop)) %>%
  ggplot(aes(x = year, y = estimate_zero_to_64_insurance_prop,
             color = cut(poverty_prop, quantile(poverty_prop, seq(0, 1, .2),na.rm=T),
                         include.lowest = TRUE),
             fill = cut(poverty_prop, quantile(poverty_prop, seq(0, 1, .2),na.rm=T),
                         include.lowest = TRUE),
             weight = 1/moe_zero_to_64_insurance_prop
             )) +
  geom_smooth() +
  scale_color_brewer(palette = 'RdYlBu', direction = -1) +
  scale_fill_brewer(palette = 'RdYlBu', direction = -1) +
  xlab("Year") +
  ylab("Percent with Health Insurance") +
  scale_y_continuous(labels = scales::percent_format()) +
  guides(fill = guide_legend(nrow = 2), color = guide_legend(nrow = 2)) +
  labs(fill = "Poverty Quantile", color = "Poverty Quantile") +
  theme(legend.position = 'bottom') +
  ggtitle("Relationship Between Percent with Health Insurance by Percent in Poverty \nover
Time",
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
ggsave("images/10-temporal-health-insurance/healthinsurance_poverty_time.png", width = 8, height = 5)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Now, we are going to use linear regression with inverse margin of error weighting to assess the trends in insurance coverage by county. We will categorize each county as increase, stable/uncertain, or decreasing. We will visualize these results with both maps and histograms.

```
# cleaning before modeling
health_insurance %<>% rename(county = NAME)
health_insurance %<>% mutate(year_centered = year - mean(c(2012, 2019)))

# create linear models with inverse margin of error weighting for each county
# regressing insurance rates on year
model_df <-
  health_insurance %>%
  nest_by(GEOID, county, state) %>%
  mutate(
  # for each county, fit a linear model and store it in a list column
  model = list(
    lm(
      estimate_zero_to_64_insurance_prop ~ year_centered,
      data = data,
      weights = 1 / data$moe_zero_to_64_insurance_prop
    )
  ),
  # use broom::tidy to extract coefficients
  coefs = list(broom::tidy(model, conf.int=T)),
) %>%
rowwise() %>%
```

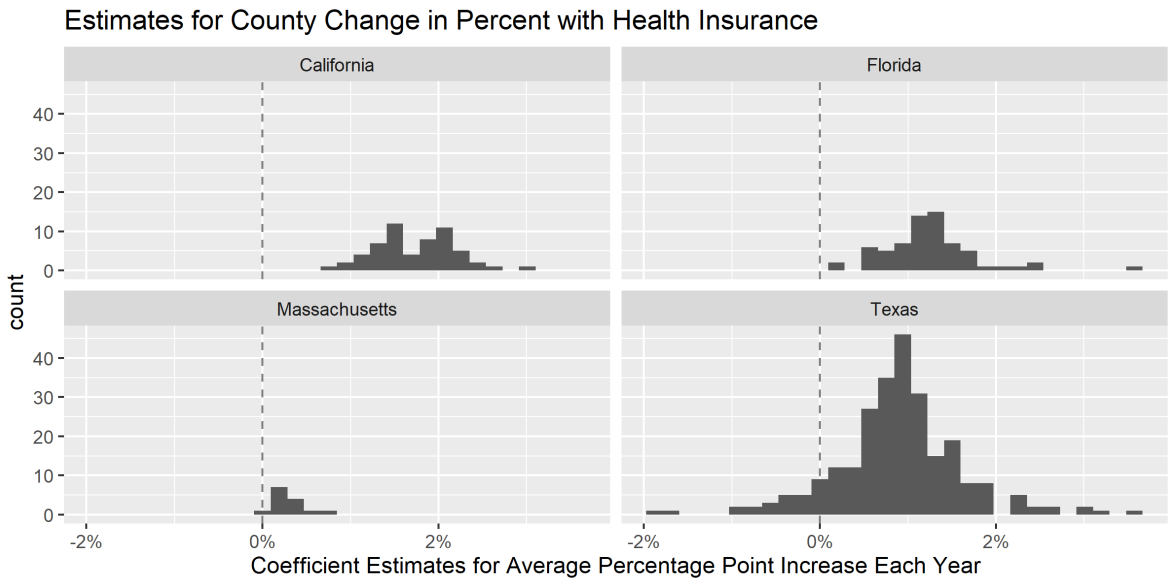


Figure 10.2: County observations of adult (age 19-64) health insurance rates in California

What additional analyses might you be interested in?

« 9 Case Study 3: COVID-19 Mortality in Cook County (March 2020 - March 2022)

11 Case Study 5: Case Study Comparing County Analyses of Inequities in Health Insurance using ACS vs. CDC PLACES data (2019) »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

11 Case Study 5: Case Study Comparing County Analyses of Inequities in Health Insurance using ACS vs. CDC PLACES data (2019)

By: Christian Testa, Dena Javadi

11.1 Introduction

The CDC PLACES dataset and American Community Survey (ACS) represent two distinct approaches to collecting and reporting on small area level data: the CDC PLACES uses a method called small area estimation (SAE) to produce estimates at the Census tract level, while the ACS provides survey based estimates directly at the Census tract level.

The methodology page from CDC PLACES provides further explanation as to exactly how they implemented SAE.

In their article *Multilevel Regression and Postratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System*, Zhang et al. describe how SAE was performed to estimate chronic obstructive pulmonary disease (COPD) in CDC PLACES.

11.2 Motivation, Research Questions, and Learning Objectives

The goal of this case study is to gain familiarity with the ACS and CDC's PLACES, understand how SAEs are created, and visualize comparisons between SAEs. The outcome of interest in this case study is health insurance.

The specific learning objectives are to:

- Download specific ACS and CDC's PLACES data
- Visualize (plot) and map estimates Characterize trends in the estimates by area based social measures (ABSMs)
- Understand limitations of SAEs Explore validity of model-based estimates
- Understand the effect of smoothing estimates in very small areas

Given that both CDC PLACES and the ACS provide estimates of the Census tract prevalence of health insurance, we are interested in assessing what difference there is between the two datasets and how that difference may vary according to ABSM.

11.3 Understanding Small Area Estimates

What is Small Area Estimation?

Goal: To provide finer geographic detail of published statistics for various subpopulations.

Problem: Surveys generally do not provide large enough samples for reliable direct estimates for small areas (counties, and especially census tracts).

Challenges: model misspecification and failure to account for informative sampling.

How valid and reliable are the PLACES county level estimates?

- Depends on quality of underlying data

- Depends on assumptions and appropriateness of statistical models used
- smaller uncertainty ranges for model-based estimates as compared to direct estimates
- smoothing out of local geographic variation which results in potentially underestimating small areas with high prevalence estimates and overestimating small areas with lower prevalence estimates
- higher discrepancies for behavior indicators than for diagnosed chronic diseases potentially due to different biases such as self-report or recall bias

Model based estimates are also not sensitive to local area interventions or shocks, meaning that prevalence may be over or underestimated based on prior trends without accounting for what local surveys may capture - such as a targeted local intervention or a disaster event. An example of this can be seen in Zhang et al's study where model-based estimates of current smoking prevalence for two adjacent counties in Missouri were 24.2% and 25.0% while direct survey estimates from a local survey were 25.3% and 13.5% due to the latter county's local tobacco control initiative.

What is Multilevel Regression and Post-stratification (MRP)?

Multilevel statistical modeling framework linking geocoded health surveys and high spatial resolution population demographic and socioeconomic data

To conduct MRP, we need:

- Surveys with outcome of interest, demographic data, and geographic indicator
- The geographic and demographic data will be used as predictors in the first stage model to define geographic and demographic levels for the second stage
- Continuous variables will need to be split into intervals to create levels
- Factors and levels must match those in the poststratification table (so individual level demographic and geographic variables must match population level variables)

What is a postratification table?

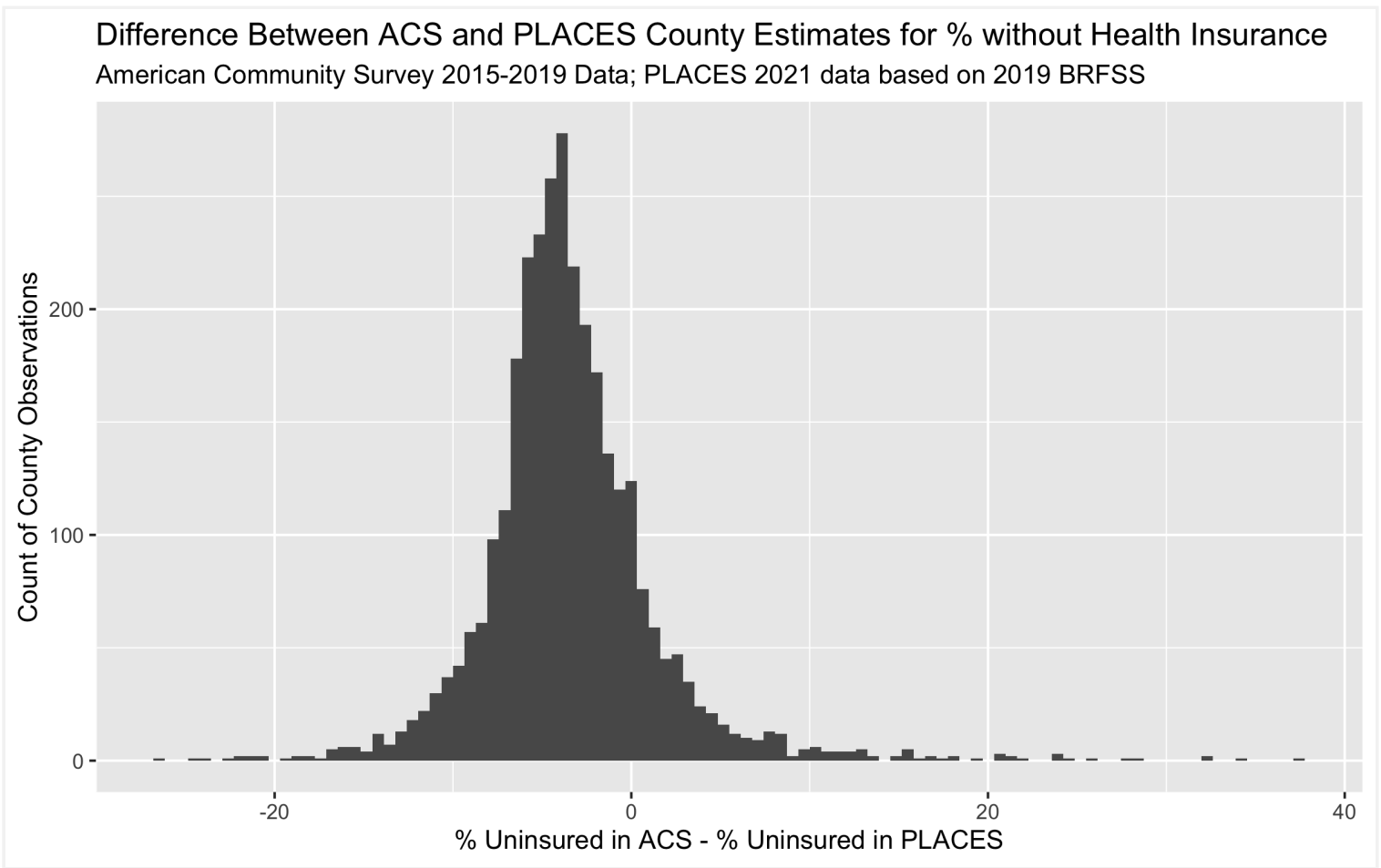
- Representation of the combination of demographic and geographic factors that correspond to the number of individuals in a population of interest
- The number of rows in a poststratification table would correspond to the product of the number of options for each of the variables included
  - I.e. 6 age groups * 3 gender categories * 8 racialized categories * 50 states = 7200 rows
- First stage
  - Multilevel logistic regression model
- Second stage
  - Weight model predictions for each subgroup by population estimated frequency of these subgroups
  - Considerations: missing variables for post-stratification table, nonresponse, missing data, inclusion of interaction terms (useful for studying demographic subgroups within geographic levels) and use of structured priors to share information across levels of a factor, violation of positivity assumption

11.4 Create our Cleaned Dataset

```r
# dependencies
library(tidyverse)
library(magrittr)
library(tidycensus)
library(patchwork)

# get our data from PLACES
# download the county level data from https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb
# rename it to "places_2021_release_county_level.csv"
places <- readr::read_csv("data/11-health-insurance-comparison/places_2021_release_county_level.csv")


# filter for our measure of interest
places %<>% filter(Measure == "Current lack of health insurance among adults aged 18-64 years")

# download ACS data
# remember you can check the variable table here:
# https://api.census.gov/data/2019/acs/acs5/variables.html
acs <- tidycensus::get_acs(
  geography = 'county',
```

11.5 Visualizing the Data

```r
# load our cleaned dataset
uninsurance <- readRDS(here("data/11-health-insurance-comparison/county_uninsurance.rds"))

# calculate the difference in the
uninsurance %<>% mutate(
  difference_in_acs_minus_places = pct_uninsured_19_64_acs - pct_uninsured_18_64_places
)

# plot the distribution of difference in ACS vs. PLACES
ggplot(uninsurance, aes(x = difference_in_acs_minus_places)) +
  geom_histogram(bins = 100) +
  ylab("Count of County Observations") +
  xlab("% Uninsured in ACS - % Uninsured in PLACES") +
  ggtitle("Difference Between ACS and PLACES County Estimates for % without Health Insurance",
      subtitle = "American Community Survey 2015-2019 Data; PLACES 2021 data based on 2019 BRFSS")
```



```r
ggplot(uninsurance,
       aes(x = pct_uninsured_19_64_acs, y = pct_uninsured_18_64_places))+
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0) +
  geom_smooth(method = 'lm') +
  scale_x_continuous(breaks = seq(0, 60, 10), limits = c(0, NA)) +
  scale_y_continuous(breaks = seq(0, 60, 10), limits = c(0, NA)) +
  ylab("PLACES Estimate of % without Health Insurance") +
  xlab("ACS Estimate of % without Health Insurance") +
  ggtitle("Comparing ACS vs. PLACES Estimate of % without Health Insurance by County",
    subtitle = "American Community Survey 2015-2019 Data; PLACES 2021 data based on 2019 BRFSS")
```

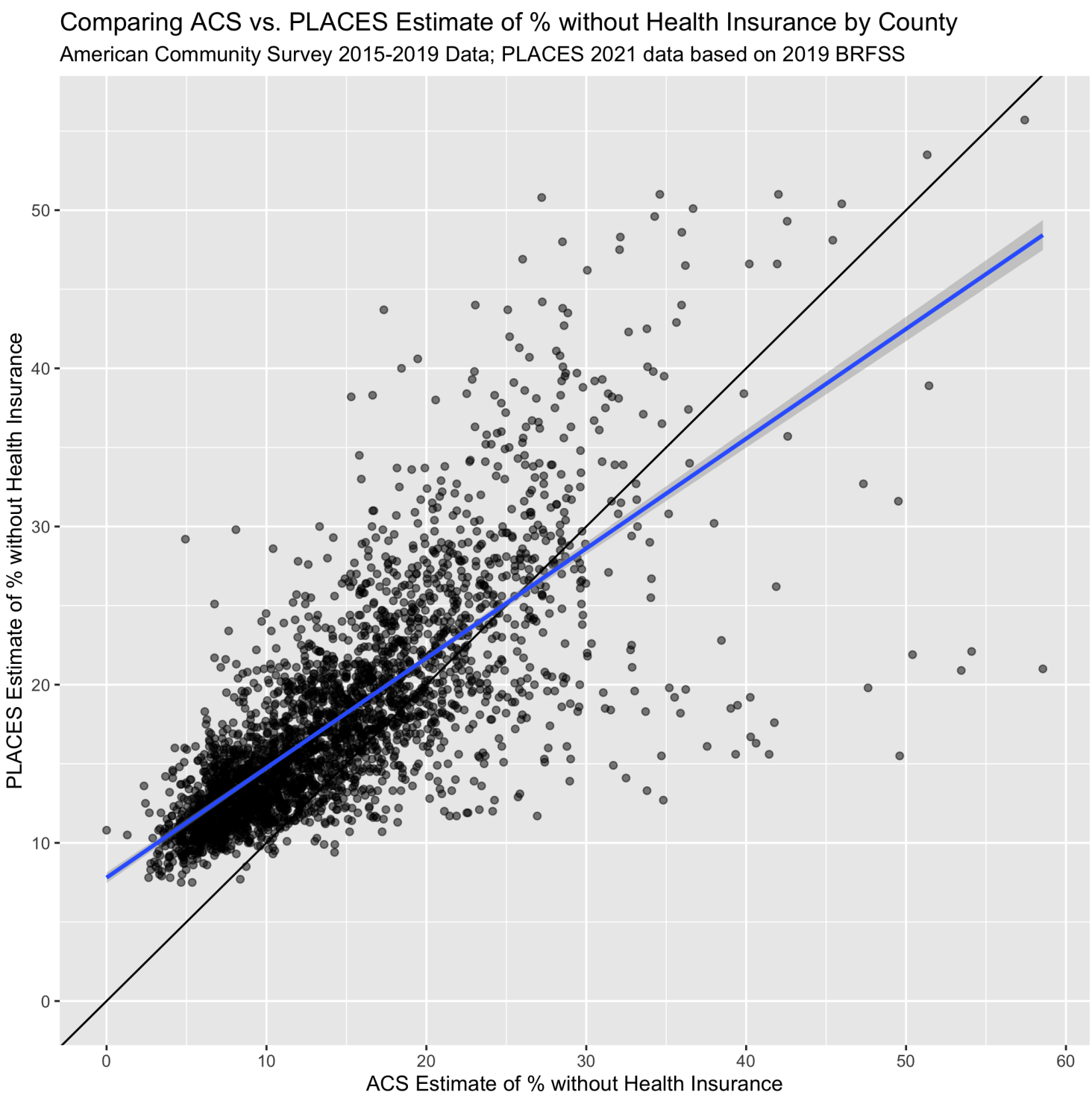
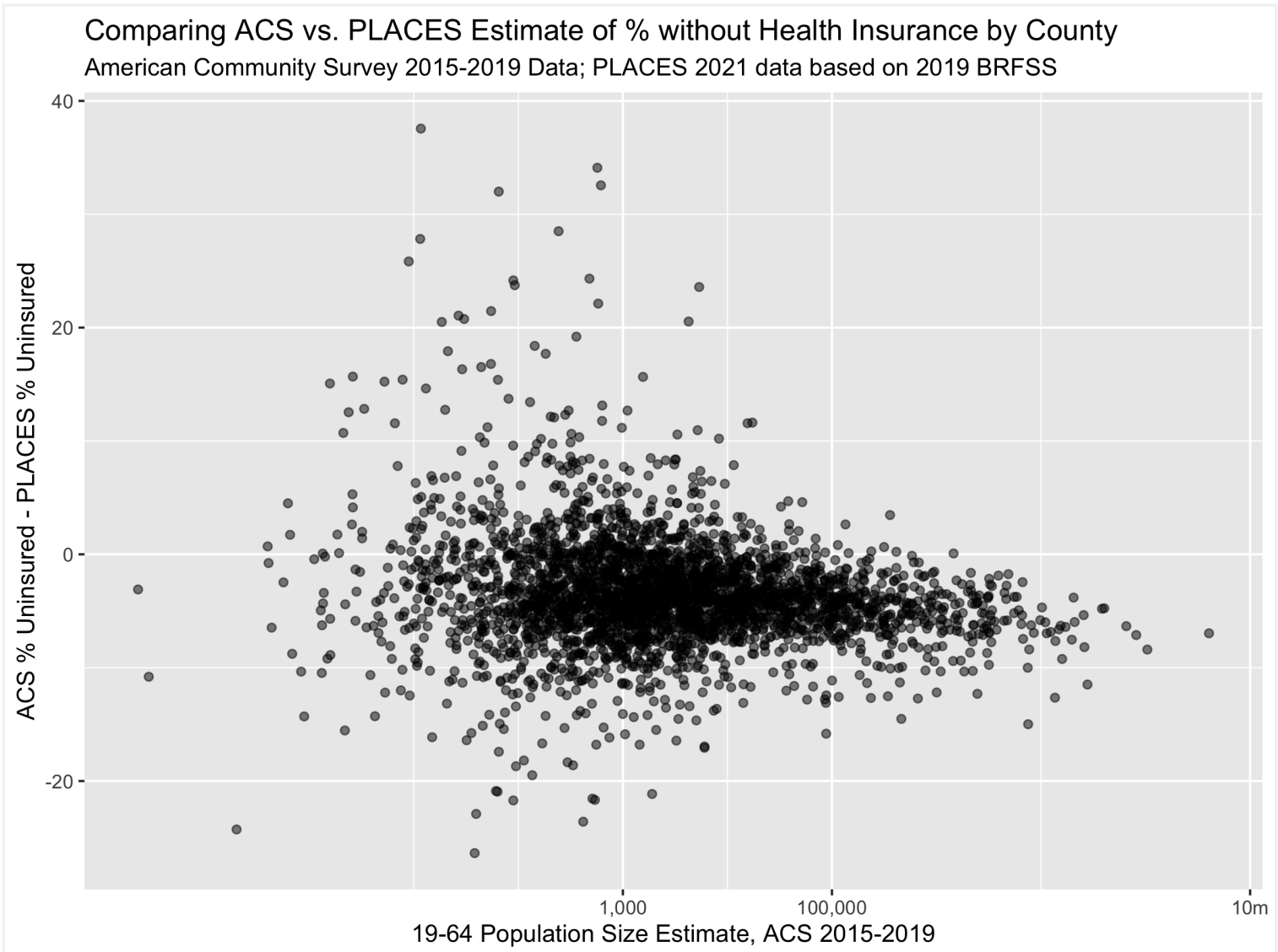


Figure 11.1: A scatterplot of ACS vs. PLACES estimates for the % without health insurance. The solid line shoes the line with slope = 1 and intercept = 0, while the line in blue shows a linear model line of best fit.

```r
ggplot(uninsurance,
       aes(x = estimate_total_19_64_population, y = difference_in_acs_minus_places ))+
  geom_point(alpha = 0.5) +
  scale_x_log10(
    breaks = c(1e4,1e5,1e7),
    labels = c("1,000", "100,000", "10m")
    ) +
  labs(x = "19-64 Population Size Estimate, ACS 2015-2019",
       y = "ACS % Uninsured - PLACES % Uninsured") +
  ggtitle("Comparing ACS vs. PLACES Estimate of % without Health Insurance by County",
    subtitle = "American Community Survey 2015-2019 Data; PLACES 2021 data based on 2019 BRFSS")
```



The last three figures help us to see the on-average differences between ACS and PLACES estimates of health uninsurance rates. In the histogram figure, we can see that the PLACES estimates of uninsurance are often higher than those in ACS, while the scatter-plot figures show us how this relationship changes with population size.

11.6 Adding ABSMs

```
variables_dict <-
  tibble::tribble(
  # ICEraceinc
  ~var, ~varname, ~description,
  # total population
  "B01001_001",  "total_popsize", "total population estimate",

  # racial composition
  'B01003_001',  "race_ethnicity_total", "race_ethnicity_total",

  # ICEraceinc
  "B19001_001",  'hhinc_total',   "total population for household income estimates",
  "B19001A_002", 'hhinc_w_1',     "white n.h. pop with household income <$10k",
  "B19001A_003", 'hhinc_w_2',     "white n.h. pop with household income $10k-14 999k",
  "B19001A_004", 'hhinc_w_3',     "white n.h. pop with household income $15k-19 999k",
  "B19001A_005", 'hhinc_w_4',     "white n.h. pop with household income $20k-24 999k",
  "B19001A_014", 'hhinc_w_5',     "white n.h. pop with household income $100 000 to $124 999",
  "B19001A_015", 'hhinc_w_6',     "white n.h. pop with household income $125k-149 999k",
  "B19001A_016", 'hhinc_w_7',     "white n.h. pop with household income $150k-199 999k",
  "B19001A_017", 'hhinc_w_8',     "white n.h. pop with household income $196k+",
  "B19001_002",  'hhinc_total_1', "total pop with household income <$10k",
  "B19001_003",  'hhinc_total_2', "total pop with household income $10k-14 999k",
```

```r
# relationship between ICE and percent without insurance in PLACES
model <- lm(
  formula = pct_uninsured_18_64_places ~ ICEraceinc_cut,
  data = uninsurance
)

broom::tidy(model)
```

```
## # A tibble: 5 × 5
##   term                          estimate std.error statistic   p.value
##   <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                      13.3      0.371     35.9  4.82e-236
## 2 ICEraceinc_cut(0.207,0.276]      1.38     0.463      2.98 2.88e-  3
## 3 ICEraceinc_cut(0.144,0.207]      2.80     0.427      6.55 6.54e- 11
## 4 ICEraceinc_cut(0.106,0.144]      5.22     0.462     11.3  5.01e- 29
## 5 ICEraceinc_cut[-0.476,0.106]     8.07     0.422     19.1  3.79e- 77
```

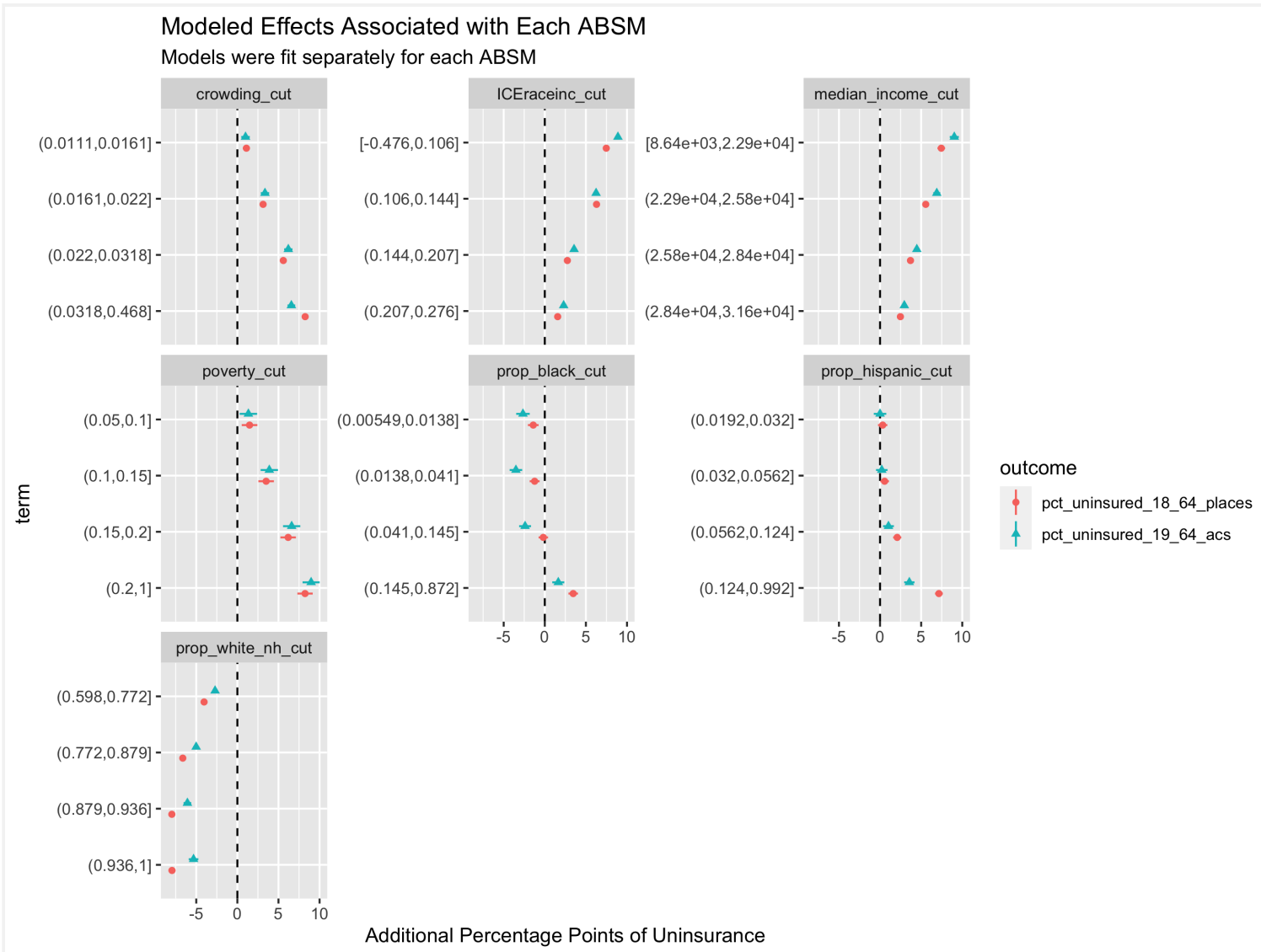
```r
# relationship between ICE and percent without insurance in ACS
model <- lm(
  formula = pct_uninsured_19_64_acs ~ ICEraceinc_cut,
  data = uninsurance
)

broom::tidy(model)
```

```
## # A tibble: 5 × 5
##   term                          estimate std.error statistic   p.value
##   <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                      9.11      0.386     23.6  1.17e-113
## 2 ICEraceinc_cut(0.207,0.276]      1.65      0.484      3.41 6.50e-  4
## 3 ICEraceinc_cut(0.144,0.207]      3.09      0.445      6.94 4.76e- 12
## 4 ICEraceinc_cut(0.106,0.144]      5.98      0.484     12.4  2.47e- 34
## 5 ICEraceinc_cut[-0.476,0.106]     9.60      0.441     21.8  3.60e- 98
```

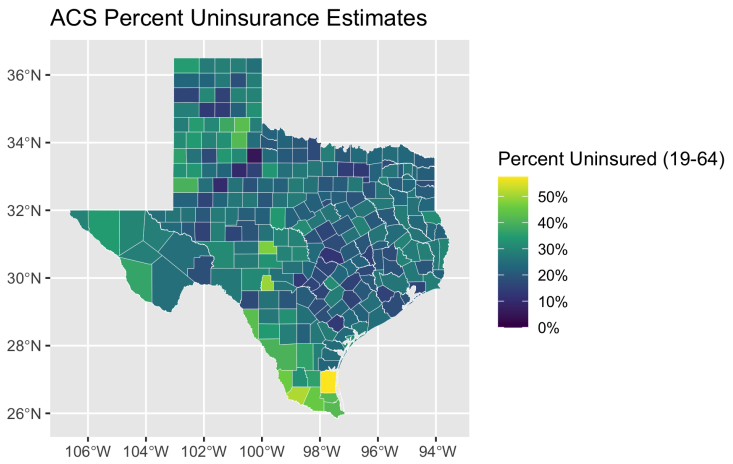
```r
# relationship between poverty and percent without insurance in ACS
model <- lm(
  formula = pct_uninsured_19_64_acs ~ poverty_cut,
  data = uninsurance
)

broom::tidy(model)
```

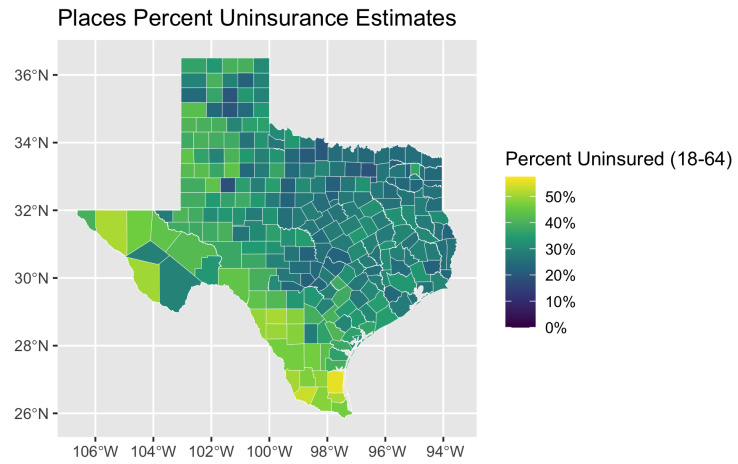
```
## # A tibble: 5 × 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)              10.5       0.692    15.2   3.91e-50
## 2 poverty_cut(0.05,0.1]   -0.325     0.776    -0.419 6.75e- 1
## 3 poverty_cut(0.1,0.15]    0.731     0.747     0.978 3.28e- 1
## 4 poverty_cut(0.15,0.2]    2.90      0.740     3.91  9.28e- 5
## 5 poverty_cut(0.2,1]       6.67      0.716     9.32  2.22e-20
```

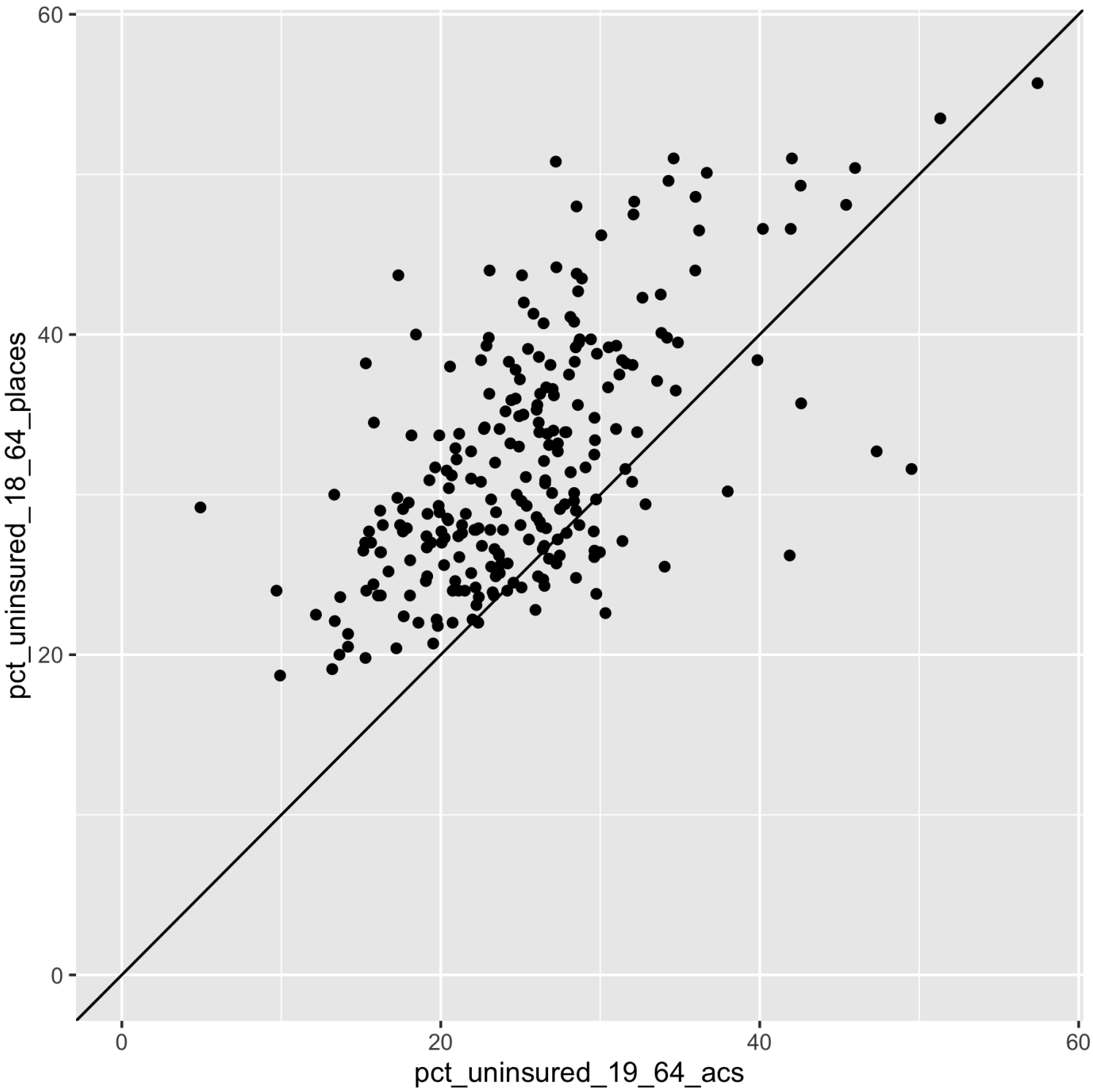
```r
# create a data frame with the names for each combination of the percent
# uninsured in PLACES or ACS variables in one column and each of the categorical
# exposure variables in another column
model_formulae <-
  tidyr::expand_grid(
    exposure = c('ICEraceinc_cut',
      'poverty_cut',
      'median_income_cut',
      'crowding_cut',
      'prop_black_cut',
      'prop_hispanic_cut',
      'prop_white_nh_cut'),
    outcome = c('pct_uninsured_19_64_acs', 'pct_uninsured_18_64_places'))

# create a column that puts the uninsurance variables on the left-hand-side
# and the exposure variables on the right hand side
model_formulae %<>% mutate(formula = paste0(outcome, " ~ ", exposure))

# fit regression models using the robust linear model rlm method from
# the MASS package
model_formulae %<>% rowwise() %>% mutate(
  model = list(
```

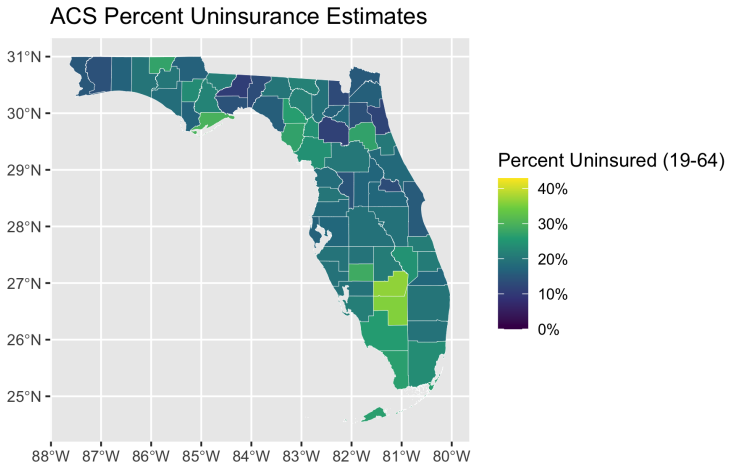


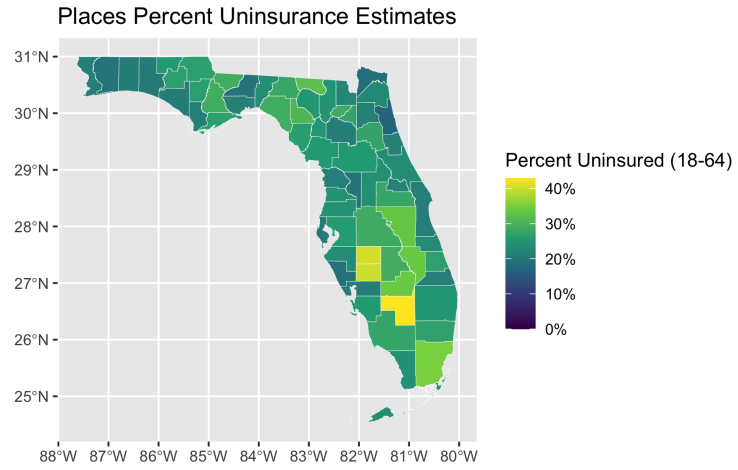
```r
# let's take a look at some specific states

# extract the maximum uninsurance rate across both ACS and PLACES so that we can
# use it in making the color palettes consistent between the two maps for ACS
# and PLACES
max_uninsurance_rate <-
  pmax(max(uninsurance %>%
    filter(state_abbr == 'TX') %>% pull(pct_uninsured_19_64_acs)),
  max(uninsurance %>%
    filter(state_abbr == 'TX') %>% pull(pct_uninsured_18_64_places)))

# create a map showing the ACS direct estimates of uninsurance rates
acs_figure <- uninsurance %>%
  filter(state_abbr == 'TX') %>%
  ggplot(aes(fill = pct_uninsured_19_64_acs)) +
  geom_sf(size = .1, color = 'white') +
  scale_fill_viridis_c(limits = c(0, max_uninsurance_rate),
                       labels = scales::percent_format(scale = 1)) +
  labs(fill = "Percent Uninsured (19-64)") +
  ggtitle("ACS Percent Uninsurance Estimates")

# create a map showing the PLACES estimates of uninsurance rates
```

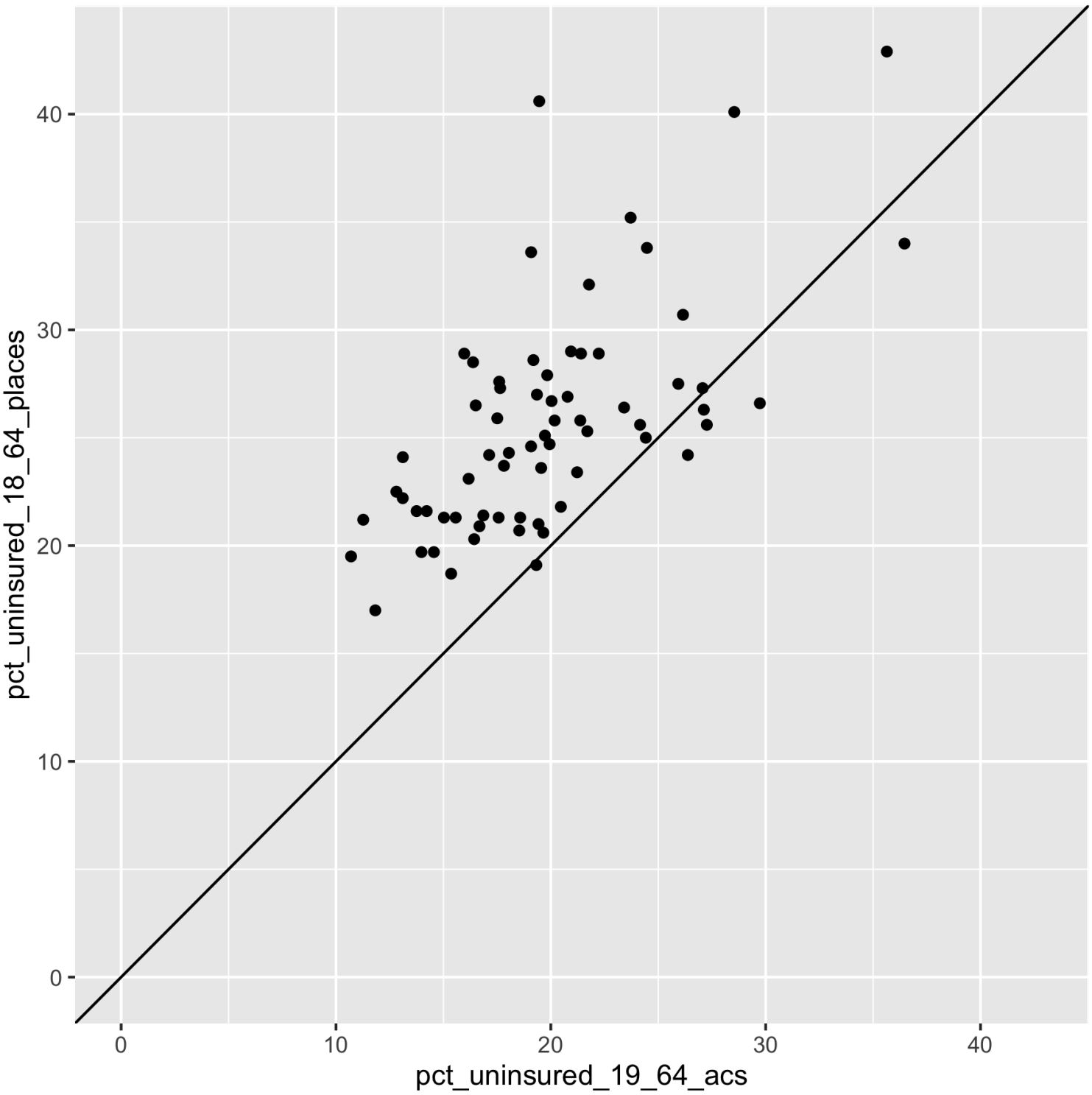

ACS Percent Uninsurance Estimates
36°N
34°N
32°N
30°N
28°N
26°N
106°W
104°W
102°W
100°W
98°W
96°W
94°W
Percent Uninsured (19-64)
50%
40%
30%
20%
10%
0%

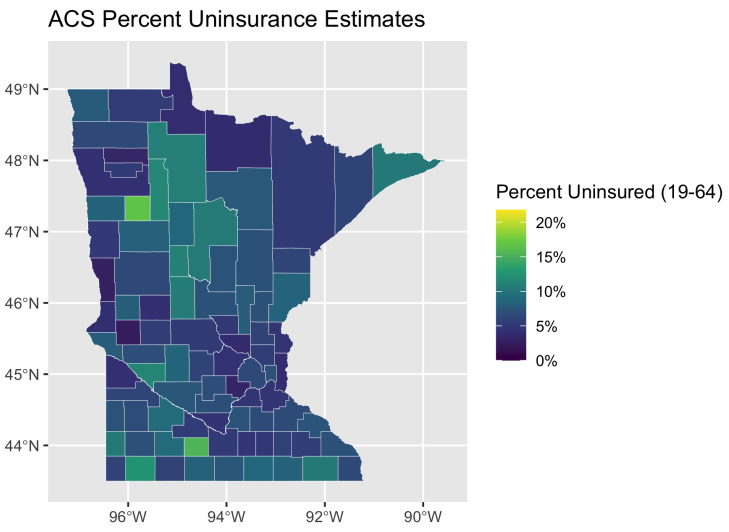

Places Percent Uninsurance Estimates
36°N
34°N
32°N
30°N
28°N
26°N
106°W
104°W
102°W
100°W
98°W
96°W
94°W
Percent Uninsured (18-64)
50%
40%
30%
20%
10%
0%

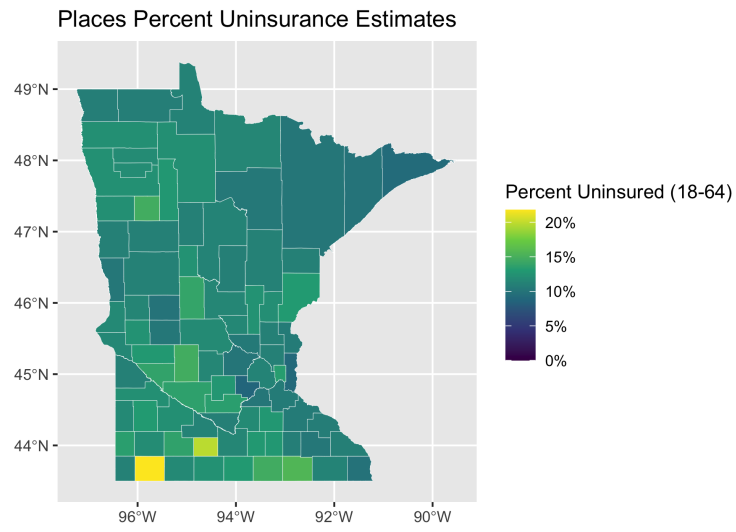

60
40
20
0
pct_uninsured_18_64_places
0
20
40
60
pct_uninsured_19_64_acs

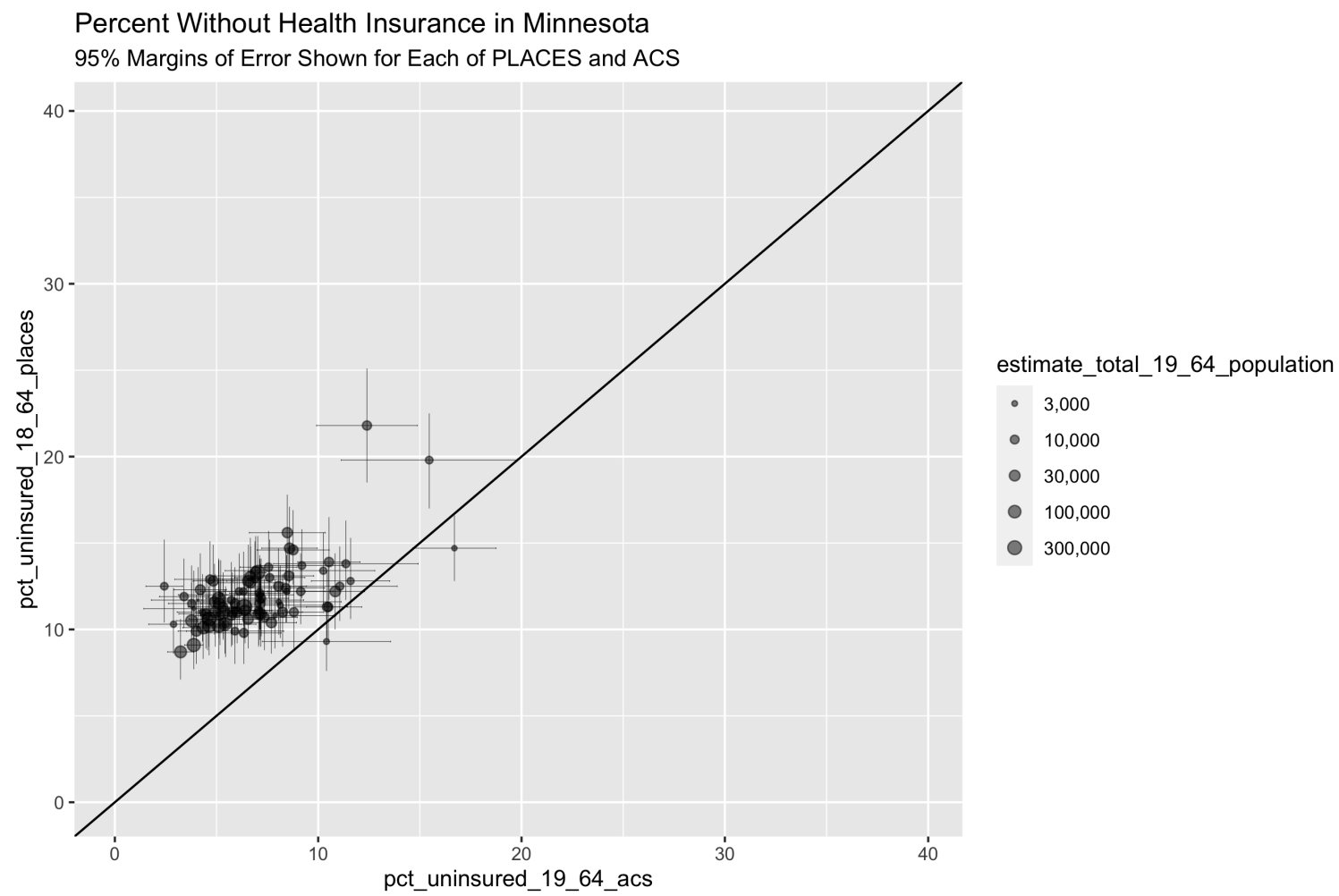

ACS Percent Uninsurance Estimates
31°N
30°N
29°N
28°N
27°N
26°N
25°N
88°W
87°W
86°W
85°W
84°W
83°W
82°W
81°W
80°W
Percent Uninsured (19-64)
40%
30%
20%
10%
0%

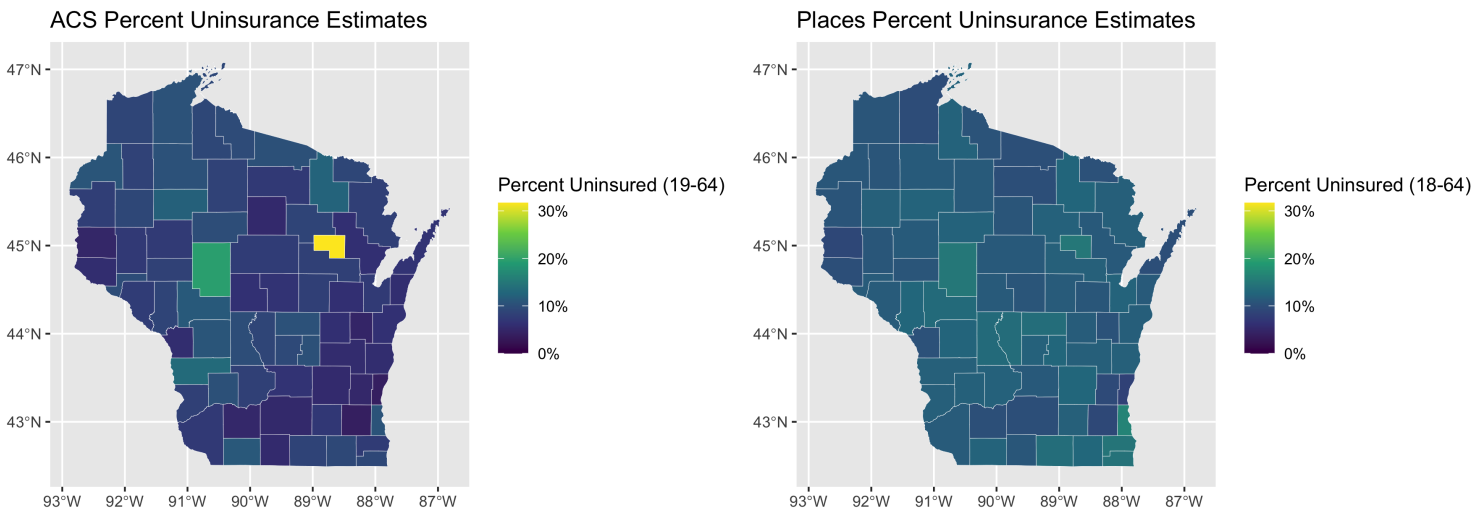

Places Percent Uninsurance Estimates
31°N
30°N
29°N
28°N
27°N
26°N
25°N
88°W
87°W
86°W
85°W
84°W
83°W
82°W
81°W
80°W
Percent Uninsured (18-64)
40%
30%
20%
10%
0%

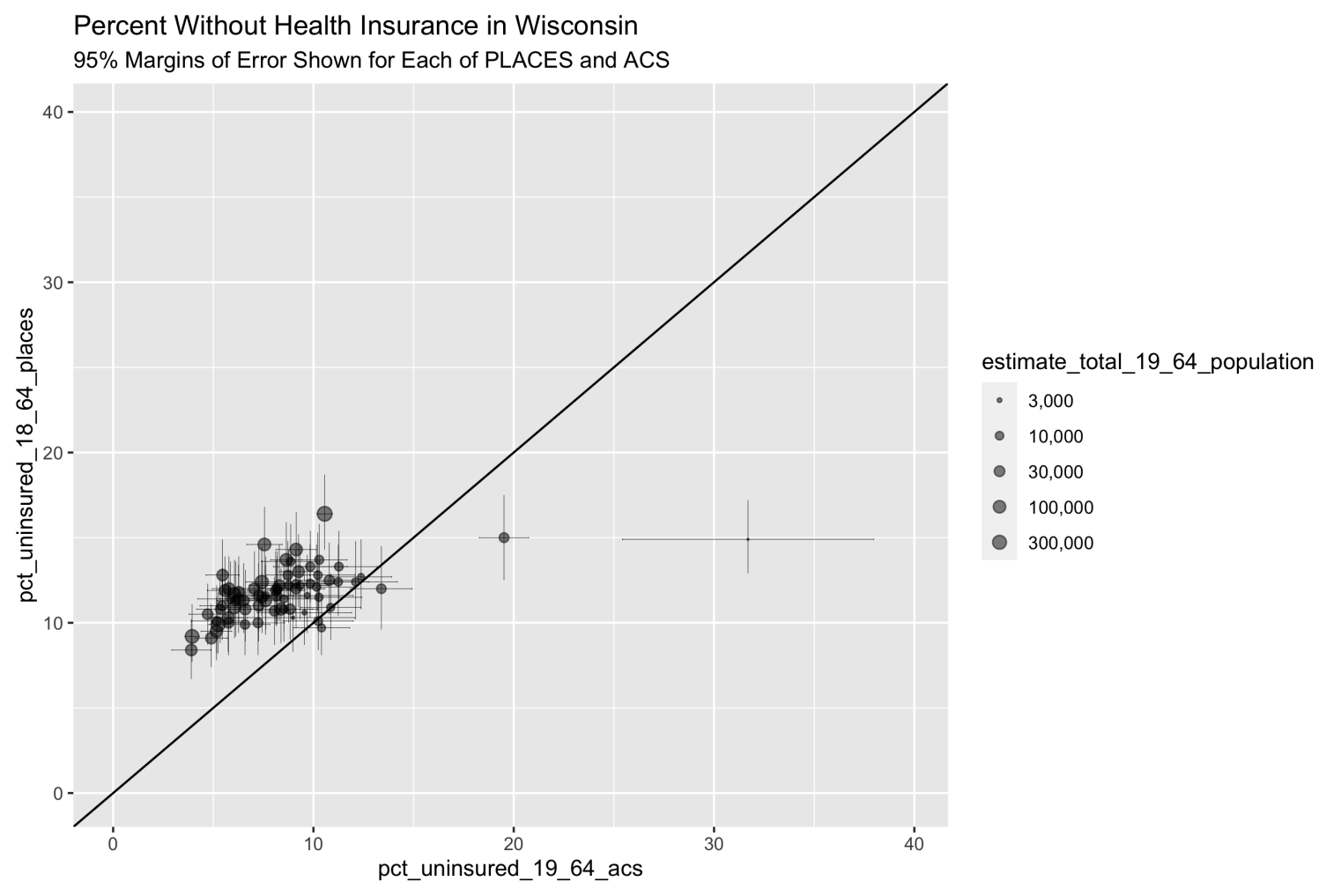

pct_uninsured_18_64_places
40
30
20
10
0
0
10
20
30
40
pct_uninsured_19_64_acs


ACS Percent Uninsurance Estimates
49°N
48°N
47°N
46°N
45°N
44°N
96°W
94°W
92°W
90°W
Percent Uninsured (19-64)
20%
15%
10%
5%
0%


Places Percent Uninsurance Estimates
49°N
48°N
47°N
46°N
45°N
44°N
96°W
94°W
92°W
90°W
Percent Uninsured (18-64)
20%
15%
10%
5%
0%







What do these graphs and analyses reveal about the value of looking at the relationships between the PLACES and ACS health insurance data in different states?

What do they show about how this relationship differs by the level of state's overall value for health insurance coverage as well as the distribution of uninsurance?

What does this make you ask about why and how place matters?

« 10 Case Study 4: Case Study on Temporal Trends using American Community Survey (ACS) data (2012-2019)

12 Conclusion »

"**Public Health Disparities Geocoding Project 2.0 Training Manual**" was written by Christian Testa, Jarvis T Chen, Enjoli Hall, Dena Javadi, Justin Morgan, Tamara Rushovich, Sudipta Saha, Pamela D Waterman, Nancy Krieger.

This book was built by the bookdown R package.

12 Conclusion

This manual has highlighted the history, context, rationale, and approaches for using spatial analysis tools and area based social metrics (ABSMs) to describe, analyze, visualize and communicate health inequities across geographic levels.

The case studies serve as opportunities to apply the topics highlighted in the manual to specific datasets and research questions. During the workshops held in the Summer of 2022, participants worked through the case studies using the following presentation template.

This template offers a systematic way to work through key questions in carrying out spatial analyses using ABSMs while applying a health equity lens. We highlight the key questions and elements of this template below to serve as a guide when developing research projects and studies of this nature. We hope that this may be of use in your current and future work!

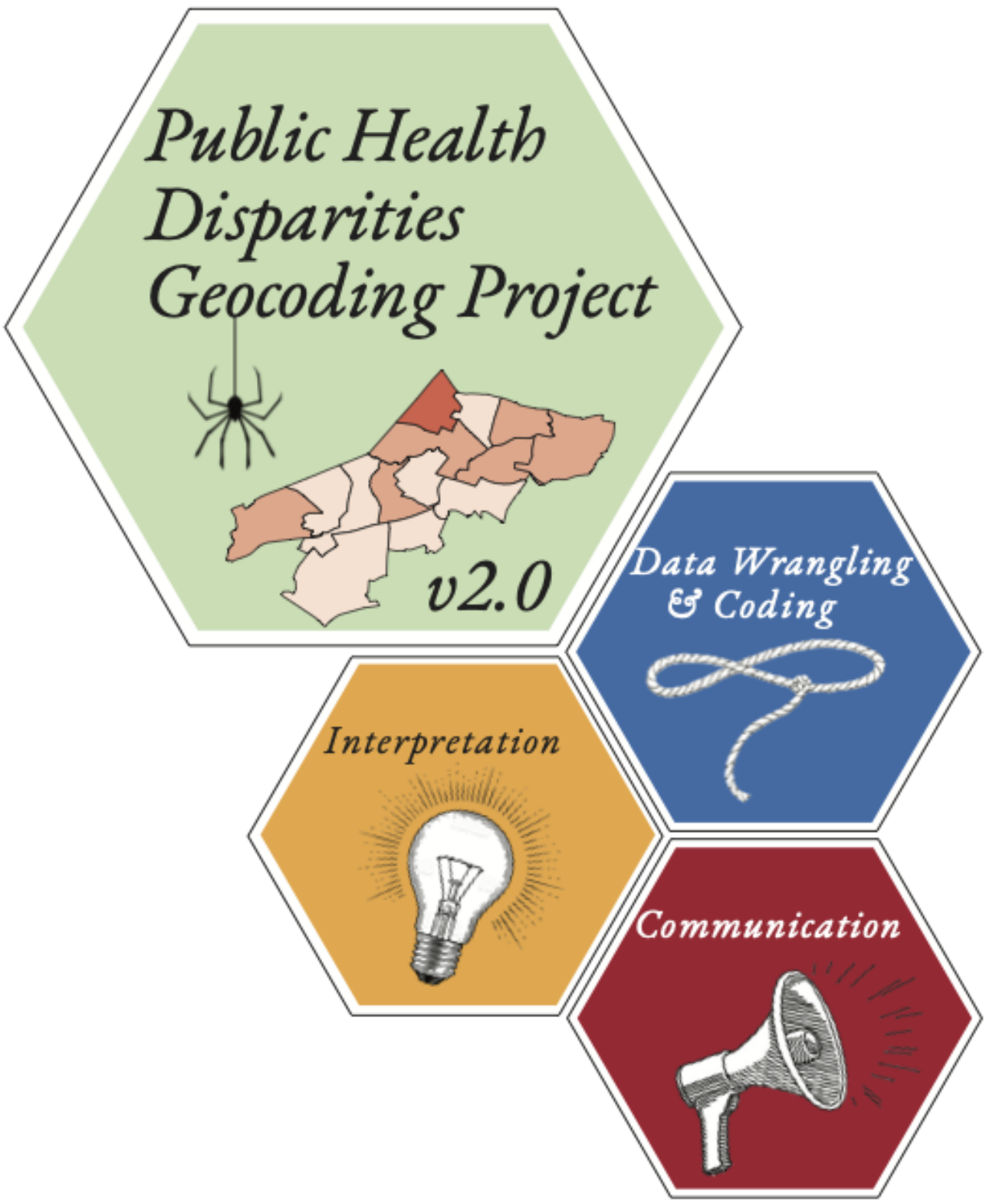
12.1 Key questions to ask

- State the study objective include:
    a. *Population*
    b. *Timeframe*
    c. *Why this is of interest?*
    d. *Whom do you want to do what, with whom, to use the knowledge generated by your study to advance health justice?*
- What are the key health equity concerns?
- What ABSMs are most relevant to your objective?
    a. *What is the ABSM intended to measure?*
        b. *Who is included in this measurement (numerator/denominator)?*
    b. *What is the data source?*
    c. *What are key concerns?*
- What geographic level is used? Why?
- What is the analytic approach?
    a. *What is your outcome of interest*
    b. *How are you characterizing the relationships between your health outcome of interest and the ABSM of interest? (e.g. Risk ratio? Risk difference? Rate difference? Rate ratio?)*
    c. *Are you doing age adjustment and if so how?*
    d. *Are you using a regression model? If so, what type?*
    e. *Are you going to model spatial effects? (e.g. Multilevel or spatial modeling approach)*
- What are key findings?
    a. *Present exploratory (univariate) tables, figures and/or maps that help contextualize your results*
    b. *Present regression models if relevant*
    c. *Summarize association between ABSMs and the health outcome*
    d. *Comment on the effect of age adjustment*
    e. *Present maps and visualizations*

Finally, if you have found the ideas and examples in this training manual to be of use, we encourage you to share widely with your colleagues and networks the links to this manual and our Public Health Disparities Geocoding Project. The more researchers, health agencies, and community-based organizations rigorously

On this page

use these methods, informed by the concepts woven throughout our manual, the richer and better the evidence base to inform analysis and action to rectify health inequities. Let us each do our part!



« 11 Case Study 5: Case Study Comparing County Analyses of Inequities in Health Insurance using ACS vs. CDC PLACES data (2019)

13 Glossary »

13 Glossary

ABSM see “Area-based socioeconomic measure”/”Area-based social metric”

address cleaning The process of taking an original address and retaining only key elements of that address (building number, street and street type), as well as correcting spelling errors and standardizing abbreviations.

age stratum One category of age in a series of age categories.

American Community Survey A new national survey administered by the US Census Bureau that provides yearly data on states and counties between the decennial censuses and which, by 2008, should provide these data for census tracts as well. For more information see https://www.census.gov/programs-surveys/acs/ .

area A geographic region whose boundaries may be defined socially, topographically, or ecologically (singly or in combination).

area-based measure see “area-based socioeconomic measure”/”area-based social metric”

area-based socioeconomic measure/area-based social metric A specifically defined measure that is used to characterize the social and contextual conditions of an area (as opposed to the social or economic characteristics of individuals). An “area-based socioeconomic measures,” for example, might pertain to the “percent of persons living below poverty”; an “area-based social metric” is a broader construct that can include but not be limited to economic data, e.g., a metric to measure racialized segregation or racialized economic segregation.

block group “A subdivision of a census tract, generally containing between 600 and 3,000 people, with an optimum size of 1,500 people. Most block groups were delineated by local participants as part of the U.S. Census Bureau’s Participant Statistical Areas Program. It is the lowest level of the geographic hierarchy for which the U.S. Census Bureau tabulates and presents sample data. (from Appendix A. Census 2000 Geographic Terms and Concepts. https://www2.census.gov/geo/pdfs/reference/glossry2.pdf)

case record see case report

case report Data on an individual that indicates the incidence or prevalence of a morbidity or mortality outcome.

cdf see cumulative distribution function

cell A basic unit of aggregation based on the cross-classification of a number of categorical variables. For example, all cases occurring among women ages 40-44 in a given census tract are aggregated into a single cell defined by gender, age, and area.

census geography A scheme of classification of areas used by the U.S. census. For example, census tract and block group are both types of areas by which data are classified in U.S. census data.

census tract “A small relatively permanent statistical subdivision delineated by local participants as part of the U.S. Census Bureau’s Participant Statistical Areas program. When first delineated they are designed to be relatively homogenous with respect to population characteristics, economic status and living conditions. They average in size between 1,500 and 8,000 people, with an optimum size of 4,000 people. The geographic size varies considerably depending on population density. (from Appendix A. Census 2000 Geographic Terms and Concepts. http://www.census.gov/geo/www/tiger/ glossry2.pdf)

census variable Items of data organized by the U.S. Census bureau. Data for these variables is structured in the form of census tables, that may include one or more census variables.

class see social class

comma-delimited file A text file format where data fields are separated by commas. The Microsoft Excel file extension for this type of data is .csv .

composite index see composite measure

composite measure A measure that combines information on more than one component variable. For example, the Townsend index consists of percent unemployment, percent renters, percent not owning a car, and percent crowding.

compositional factors Attributes of areas that derive from the characteristics of individuals.

construct A theoretical concept or idea.

contextual factors Attributes of areas that derive from structural or social characteristics of the area.

CT see census tract

cumulative distribution function For a given value, the area under the probability function up to that value (i.e. cdf(x) = Pr[X<=x]). When calculated as part of deriving the relative index of inequality, the cumulative distribution function of an area-based socioeconomic measure (ordered from most affluent to most deprived) for a given value can be interpreted as the proportion of the population who are more affluent.

denominator There are two definitions of denominator that depend on the measure being calculated. For calculating rates, the denominator is the amount of person-time observed during which time cases were eligible to occur. For calculating ABSMs, the denominator is the total number of persons in an area for which the ABSM was measured.

deprivation "Deprivation can be conceptualized and measured, at both the individual and area level, in relation to: material deprivation, referring to 'dietary, clothing, housing, home facilities, environment, location and work (paid and unpaid), and social deprivation, referring to rights in relation to 'employment, family activities, integration into the community, formal participation in social institutions, recreation and education' "(from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

direct age standardization A method for adjusting a population rate for age, yielding the hypothetical rate that would have been observed if the population being studied had the same age distribution as an externally defined standard population. In direct standardization, stratum specific rates are multiplied by weights derived from a standard reference population, and summed to yield a summary rate. Rates standardized to the same external standard may be meaningfully compared to examine differences that are not due to age.

ecosocial theory A theory that seeks to "integrate social and biological reasoning and a dynamic, historical and ecological perspective to develop new insights into determinants of population distributions of disease and social inequalities in health." The core concepts for ecosocial theory include 1. embodiment, 2. pathways to embodiment, 3. cumulative interplay between exposure, susceptibility, and resistance, and 4. accountability and agency. (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

etiologic period The duration of time over which a disease develops, referring to the time from an initial exposure to the time at which the outcome caused by this exposure occurs.

exact confidence limits Exact confidence limits that do not rely on a normal approximation. We used exact confidence limits to calculate confidence intervals when the rate was zero.

gamma confidence intervals Confidence intervals for the direct standardized rate based on the gamma distribution. A practical consequence of using gamma confidence intervals is that confidence intervals for rates will not cross zero. For more details see Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Statistics in Medicine 1997;16:791-801

gender "A social construct regarding culture-bound conventions, roles and behaviors for, as well as relationships between and among, women and men and boys and girls." (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.) – with this definition focused on social divisions predicated on dominant social structures and norms shaped by both both sexism and

gender binarism (see: Krieger N. Measures of Racism, Sexism, Heterosexism, and Gender Binarism for Health Equity Research: From Structural Injustice to Embodied Harm-An Ecosocial Analysis. Annu Rev Public Health. 2020 Apr 2;41:37-62. doi: 10.1146/annurev-publhealth-040119-094017. Epub 2019 Nov 25.).

geocoding The assignment of a numeric code to a geographical location

geographical information systems Technology based systems that combine layers of geographic data to offer a greater understanding of the characteristics of places.

Gini A measurement of inequality that ranges between 0 and 1, which is the ratio of the area under the Lorenz curve to the area under the diagonal on a graph of the Lorenz curve. A value of one would indicate complete inequality of distribution, while a 0 indicates no inequality.

GIS see geographical information systems

incidence rate The number of events divided by the person-time at risk.

incidence rate difference The absolute difference between two incidence rates. The incidence rate among the exposed proportion of the population, minus by the incidence rate in the unexposed portion of the population, gives an absolute measure of the effect of a given exposure.

incidence rate ratio The ratio of two incidence rates. The incidence rate among the exposed proportion of the population, divided by the incidence rate in the unexposed portion of the population, gives a relative measure of the effect of a given exposure.

index of concentration at the extremes (ICE) a measure of spatial social polarization quantifying the concentrations, within a specified area, of social groups at what are defined to be the extremes of deprivation and privilege for the specific metric chosen; examples of ICE measures can be in relation to income (quantifying the concentration of high vs. low income households, capturing high vs low economic privilege), racialized segregation (e.g., using the social groups White non-Hispanic vs. Black non-Hispanic persons, thereby capturing high vs low racialized privilege), or racialized economic segregation (e.g., using the social groups white non-Hispanic high income households vs. Black non-Hispanic low-income households, thereby capturing the joint exposure of racialized and economic residential segregation).

indirect age standardization A method for adjusting a population rate for age, yielding the hypothetical rate that would have been observed if the population being studied had the same age distribution as an externally defined standard population. Indirect standardization is based on deriving an expected number of events using an externally defined standard population, and contrasting this value to the observed number of events in the population being studied. The expected number of events is derived by multiplying the stratum-specific counts in the study population by stratum-specific rates from a standard population. The ratio of total observed events to the number expected is the standardized mortality (or morbidity) ratio (SMR). The indirect standardized rate is calculated by multiplying the SMR by the crude rate from the standard population.

lifecourse perspective "Refers to how health status at any given age, for a given birth cohort, reflects not only contemporary conditions but embodiment of prior living circumstances, in utero onwards" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.), with analyses taking into account age at exposure, duration of exposure, etiologic period, and whether there are critical or sensitive periods (by age group) in which exposures are most likely to increase risk (or protect from) the specified health outcomes.

material deprivation see deprivation

multilevel analysis Analyses that conceptualize and analyze associations at multiple levels, e.g., employ individual- and area-based data in relation to a specified outcome. These analyses typically entail the use of variance components models to partition the variance at multiple levels, and to examine the contribution of factors measured at these different levels to the overall variation in the outcome.

numerator There are two definitions of numerator that depend on the measure calculated. For calculating rates, the numerator is the number of events observed. For calculating ABSMs, the numerator is the number of persons or households in an area with the socioeconomic characteristic of interest.

occupational class A measurement of socioeconomic position based upon job characteristics. One example is the original British Registrar General's Social Class scheme, created in the early 20th c CE, which was based on skill. This was replaced in 2001 by the National Statistics Socio-Economic Classification system (NS-SEC), an occupational metric based on "employment relations and conditions of occupations" (see: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010). In a US context, the NS-SEC can be adapted to create an ABSM "working class" measure, comprised of occupations in which those employed are primarily non-supervisory employees (see: Krieger N, Barbeau EM, Soobader MJ. Class matters: U.S. versus U.K. measures of occupational disparities in access to health services and health status in the 2000 U.S. National Health Interview Survey. Int J Health Serv. 2005;35(2):213-36. doi: 10.2190/JKRE-AH92-EDV8-VHYC.)

operational definition A description of a variable in terms of how the variable is actually measured.

person-time The sum of the time at risk for all persons in a population.

Poisson model A regression model used for count data.

population attributable fraction The theoretical reduction of incidence that would be expected if the entire population had the same level of exposure as a specified referent group (which could be a group with low or no exposure).

poverty "To be impoverished is to lack or be denied adequate resources to participate meaningfully in society" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

poverty area In the US, the federal criteria for being a "poverty area" is to be an area with a 20% or more of the population below the poverty line (see: https://www.census.gov/library/publications/1995/demo/sb95-13.html).

poverty line A poverty threshold that takes into account household size and age composition and intended to indicate an income level below which subsistence needs are not met. The poverty line in the US is based on a value of three times the cost of the economy food basket in 1963, adjusted for inflation. See: "How the Census Bureau Measures Poverty (Official Measure)" at: http://www.census.gov/hhes/poverty/ povdef.html

public health surveillance system A structure that facilitates the continuous and systematic collection of descriptive information for monitoring the health of populations (from Buehler, Chapter 22: Surveillance, in Rothman and Greenland, Modern Epidemiology, 2nd edition, 1998, p 435-457).

racialized group and US categories of "race/ethnicity" "A social, not biological, category, referring to social groups, often sharing cultural heritage and ancestry, that are forged by oppressive systems of race relations, justified by ideology, in which one group benefits from dominating other groups, and defines itself and others through this domination and the possession of selective and arbitrary physical characteristics (for example, skin color)" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.). In the US, all federal data, including US census data, must conform to the 1997 Office of Management and Budget Revisions to the Standards of Classification of Federal Data on Race and Ethnicity, which require classification into categories of "races" and "ethnicity" (defined solely as "Hispanic" vs. "non-Hispanic"); see: https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf . Work is underway at the US Census to shift from asking 2 separate questions (one about "race," the other about "ethnicity") to one question that includes both (along with the option to tick as many boxes as relevant); see: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html

rate difference see incidence rate difference

rate ratio see incidence rate ratio

relative index of inequality A summary measure of "total population impact" that takes into account both the socioeconomic gradient in the outcome, as well as the population distribution of the socioeconomic variable. The RII is interpretable as the ratio of the rate in the theoretically most deprived segment of the population, compared to the rate in the theoretically least deprived segment.

RII see relative index of inequality

SEP see socioeconomic position

sex "A biological construct premised upon biological characteristics enabling sexual reproduction" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

social class "Refers to social groups arising from interdependent economic relationships among people" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

social deprivation see deprivation

socioeconomic position "An aggregate concept that includes both resource-based and prestige-based measures, as linked to both childhood and adult social class position" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

socioeconomic status A term referring to prestige-based measures of socioeconomic position, as determined by rankings in a social hierarchy (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

spatiotemporal Of, relating to, or existing in both space and time.

spatiotemporal mismatch A mismatch of data derived from different sources that arises because of (1) inconsistency of boundaries between data sources and/or (2) inconsistency of timeframe between data sources.

transpose To reverse the orientation of a matrix, so that the values across the rows become the values down the columns, and the values of the columns become the values across the rows.

wealth Conceptually, wealth refers to accumulated assets. An ABSM to capture wealth is operationalized from census data as percent of owner-occupied homes worth more than 400% of the median value of owned homes.

ZCTA see "Zip code tabulation area"

ZIPcode "Administrative units established by the United States Postal Service ... for the most efficient delivery of mail, and therefore generally do not respect political or census statistical area boundaries" (from Appendix A. Census 2000 Geographic Terms and Concepts).

ZIPcode tabulation area A statistical geographic area that approximates the delivery area for a U.S. Postal service Zip code. This approximation replaces the Zip code areas used by the Census Bureau in conjunction with the 1990 and earlier censuses.(from Appendix A. Census 2000 Geographic Terms and Concepts.)

Z-score Also referred to as Z-ratio or Z-value, it is equal to a value of X minus the mean of X, divided by the standard deviation.