

Decoding Segregation:

The topological landscape of segregation forms and a bottom-up ontology

Supplementary information: Codebook for identifying segregation forms

The interpretive procedure ('coding') of identifying n-grams as representations ('labels') of segregation forms involved the work of a group of researchers ('coders'). We deployed inductive coding, a method of data analysis where codes are developed directly from the data rather than being predetermined or based on existing frameworks. This approach is often used in exploratory research to allow new themes and patterns to emerge from the data. Researchers develop codes and categories directly from the raw data through close reading and iterative review (O'Reilly, 2012). We coded an n-gram either as (a) a valid representation of a segregation form, (b) a valid segregation form that an existing label should represent, or (c) not a representation of a segregation form. We opted for an iterative thematic analysis to achieve inter-coder transparency, consistency and reliability in consensual interpretations (Hemmler et al., 2020). The sequential procedure of individual coding allows the repetitive evaluation of the quality of the coding process, building coherence and stability of interpretations in coding decisions across the research team (Cascio et al., 2019). Here is the sequence of protocols collectively developed to support and guide the coding process.

- Protocol 1 Definitions of meaningful and typological n-grams
- Protocol 2 Trigrams
- Protocol 3 Synonymous or interchangeable terms
- Protocol 4 Race, ethnicity, nationality, religion and culture
- Protocol 5 Segregation in the work realm
- Protocol 6 Segregation in the educational realm
- Protocol 7 Social groups and urban areas
- Protocol 8 Multiplicity and intensity of segregation forms
- Protocol 9 Regional and place-specific situations of segregation
- Protocol 10 Scales of segregation
- Protocol 11 Temporality
- Protocol 12 Age and segregation
- Protocol 13 Voluntary and involuntary segregation
- Protocol 14 Negative terms in n-grams
- Protocol 15 Fourgrams
- Protocol 16 Hyphen in compound terms
- Protocol 17 Spelling style and typographical errors

The complete list of coded n-grams can be viewed here: Table ST03
Each code's definitions and references are available in an associated collaborative project, [Segregation Wiki](#).

Protocol 1: Definitions of meaningful and typological n-grams

- A meaningful n-gram represents a segregation form as an act, practice or process of separation or restriction on contact between individuals or social groups based on characteristics or distinguishing attributes.
- In line with an inductive, bottom-up approach, this definition was not determined a priori. Instead, it emerged from the coding process by carefully considering concrete candidates for valid segregation forms in an exploratory and open-ended manner. The authors discussed all cases where disagreement occurred and strove for consensus when possible.
- An n-gram is a sequence of n units of text in a particular order – in our case, having 'segregation' as the last unit. In the English language syntax, identifying such forms is facilitated by, even though not limited to, the structure '<adjective> segregation' including one or two preceding words: '<x> segregation' (bigram) or '<x y> segregation' (trigram). 'X' and 'y' may represent words that qualify or define segregation forms (e.g. 'social segregation', 'occupational gender segregation'). We are aware that such construction may limit the identification of segregation-related n-grams that do not use such a syntactic construction (e.g. 'segregation in urban trajectories'). However, we estimate their occurrence is relatively rare compared to the more common structure outlined above.
- Coders must review every n-gram potentially representing an SF identified by the algorithmic procedure (see manuscript, Method).
- Coders must assess terms denoting potential segregation forms by accessing the corpus's titles, abstracts, keywords, and full texts of documents.
- N-grams representing potential SFs may or may not appear repeatedly in different publications in the sampled literature.
- Coders will assess how the n-gram is used in different publications until they know whether it identifies and represents an SF.
 - Most segregation forms explored in the sampled literature are likely to have low frequencies. Nearly half of the validated n-grams (412 or 51.2 % of the sample) appeared only in one document within the sample, whereas 15.2% appeared in two documents.
 - N-grams representing SFs may appear only in abstracts, titles or keywords and not in the main text of manuscripts.
- When coders identify an n-gram potentially representative of a segregation form but whose use was not confirmed in the Scopus-derived dataset, they may use the literature outside the dataset by running searches in Google Scholar as a final resource.
- Coders should not code the following:
 - Non-meaningful n-grams, e.g. "without segregation".
 - Meaningful but non-typological n-grams:
 - Terms defining intensities, e.g. "increasing segregation", "less segregation", "robust segregation", "strong segregation", "accentuate class segregation";
 - Generic or incidental adjectives, i.e. terms used in a mere qualitative descriptive form, e.g. "extraordinary segregation", "systematic segregation", "persistent segregation", "pervasive segregation", etc.
 - Actions or processes, e.g. "measuring segregation", "achieved class segregation", etc.

Protocol 2: Trigrams

- Combinatorial by nature, trigrams have the potential to express emerging concepts and subtle segregation forms. Trigram validation requires attention due to the exponential nature of combinations that the research field can produce.

- Trigrams may be frequently composed of valid bigrams. There is considerable potential for virtually any pair of bigrams to be combined into trigrams, e.g. 'housing segregation' and 'ethnic segregation' combined into 'ethnic housing segregation'.
- Trigrams addressing segregation forms can also emerge without combinatorial work with existing bigrams and should be coded, e.g. 'activity-space segregation' or 'segregation within segregation'.
- Trigrams addressing aspects of segregation seemingly identified as bigrams should be carefully considered and, if they address conceptually or empirically valid segregation forms, should be coded as trigrams, e.g. 'personal network segregation', 'friend network segregation' and 'ethnic network segregation' as variations of 'network segregation' (itself a frequently used shorter representation of 'social network segregation').
- This procedure should support the identification of seemingly meaningful bigrams that are, in fact, trigrams, e.g. 'Crow segregation' should be coded as 'Jim Crow segregation'; 'jure segregation' as 'de jure segregation', as 'facto segregation' as 'de facto segregation'.
 - In these cases, appropriate measures should be taken to preserve the accuracy of counting bigrams and trigrams in the analysis of segregation research. For instance, the number of 'personal network segregation' occurrences should be subtracted from the total 'network segregation' count.

Protocol 3: Synonymous or interchangeable terms

- Authors in the field may address the same segregation form through different terms or n-grams. On the other hand, subtle empirical differences may be addressed through a range of terms, including similar ones (e.g., 'racial segregation' and 'racist segregation'; 'ethno-national segregation' and 'ethno-nationalist segregation'). Such differences, similarities, and complementarities may cause difficulties in interpreting whether terms are synonymous or not. Therefore, coders must consider what authors identify as segregation forms in their works.
- From an initial perspective, aggregating similar n-grams under a single label could lead to a more easily knowledgeable set of identified segregation forms. Furthermore, keeping synonymous terms as distinct SFs implies that their frequencies in the sampled literature would be counted separately, dispersing their potential representativity measured by frequency and centrality in networks of associations within the corpus.
- However, nuances in terminology matter and may have conceptual and empirical consequences. Coding different n-grams under the same label representing a single SF poses risks of oversimplification. Coders should not avoid the multiplication of terminology and SFs when there are nuances potentially intended by the authors. Aggregating similar, complementary or frequently associated terms will be performed in the taxonomic stage of this work through a semantic, network and cluster-based classification of segregation forms into segregation types (see manuscript, ontology section).
- Coders should only consider two or more n-grams synonymous when they can be used interchangeably without reducing their meaning, intelligibility or empirical implications. Only in these cases can synonymous n-grams be represented by a single label.
- In the event of identifying *two or more n-grams used in the sampled literature as synonymous*, coders should generally choose the n-gram as a label for the segregation form based on the year of first publication and, as a second criterion, the total number of publications in case of n-grams having the same year of first publication. Examples:
 - As both terms relate to segregation made officially obligatory or required by law or rule, 'mandatory segregation' (first year 2009 in Scopus) should be coded as 'mandated segregation' (2006) and added to a single count (i.e. amounting to 7 publications);
 - Segregation forms 'legalized segregation' (2001) and 'legislated segregation' (2011, 4 publications) should be coded as 'legal segregation' (1985).

- 'multiscalar segregation' (2016) should be coded as 'multiscale segregation' (2015).
- Coders must pay attention to *similar but not interchangeable terms*. Examples:
 - 'female segregation' and 'women segregation': women are frequently interpreted as female adults, so not all females are women, and these terms cannot be used interchangeably;
 - 'friend segregation' and 'friendship segregation' cannot be used interchangeably, as the former could be interpreted as the separation of individuals based on their friendships while the latter encompasses patterns in which friendships are segregated based on certain criteria, such as race, ethnicity, gender, or other social dimensions;
 - 'forced segregation' is not synonymous with 'enforced segregation': the former implies coercion or the use of physical or psychological pressure; the latter implies a rule, law or policy upheld to ensure compliance, involving authority, but not necessarily coercion or force;
 - 'communal segregation', 'communitarian segregation' and 'community segregation' are not synonymous;
 - 'poor segregation' and 'poverty segregation' are not synonymous.
- In general, verbs' *past participle and past tense* should be coded as adjectives qualifying a segregation form when the adjectivized n-gram has been identified in the sampled literature. For example, 'legalized segregation', as the participle of 'to legalize', should be coded as 'legal segregation'.
 - Exceptions should be made when the participle adds relevant differences in meaning, e.g. 'institutional segregation' and 'institutionalized segregation': the former refers to segregation practices and policies that occur within specific institutions, such as schools, workplaces, prisons, or other formal organizations, denoting the separation of groups within these institutions based on various characteristics, such as race, gender, or socioeconomic status. The latter refers to segregation forms that become entrenched and normalized within society's structures, policies, and practices, a deeper level of systemic segregation upheld by laws, cultural norms, and institutional practices over time.
 - When the n-gram with the past participle form is found in the sampled literature but not its adjectivized version, coders should keep the participle original version, e.g. 'formalized gender segregation'.
- Trigrams may use a combination of terms in which the first is a specification of the second, thus introducing redundancy, which should be avoided. Examples:
 - Segregation form (SF) 'household residential segregation' should be coded as 'residential segregation' as the term 'residential' is assumed to address household location; SF 'income economic segregation' should be coded as 'income segregation', as income falls within the economic realm; 'social class segregation' should be coded as the bigram 'class segregation', 'Schelling residential segregation' should be coded as 'Schelling segregation', given that Schelling's (1969) classic model is a dynamic residential segregation model.
- Trigrams with *modifiers*:
 - When strictly possible with no loss of meaning, trigrams with modifiers 'based', 'level', 'specific', 'wise', etc. should be coded under corresponding existing bigrams.
 - Unigram 'based': e.g. 'spatial-based segregation' should be coded as 'spatial segregation'; 'asylum-based segregation' as 'asylum segregation'; 'gender-based segregation' as 'gender segregation'; 'income-based segregation' as 'income segregation'.
 - Unigram 'level': e.g. 'school-level segregation' should be coded as 'school segregation'; 'tract level segregation' as 'tract segregation'; 'establishment-level segregation' as 'establishment segregation'.

- Other modifiers: code 'gender-specific segregation' and 'gender-wise segregation' as 'gender segregation'; 'region-wide segregation' as 'regional segregation'; etc.
- Trigrams with modifiers should be coded as such when no corresponding bigram exists in the sampled literature or when modifiers are necessary to specify the nature and scope of segregation. Removing them would result in a loss of meaning.
 - Examples: SFs 'color-based segregation', 'people-based segregation', 'place-based segregation', 'trajectory-based segregation', 'destination-based segregation', 'distance-based segregation', 'street-level segregation'.
 - Given their use in the analysed literature, both terms 'individual-level segregation' and 'individual segregation' should be coded as themselves, i.e. under separate labels.
 - 'Horizontal occupational segregation' should be coded as itself: although both unigrams 'horizontal' and 'occupational' are work-related terms, 'horizontal' entails a hierarchical work condition.

Protocol 4: Race, ethnicity, nationality, religion and culture

- Race and ethnicity are central subjects in disciplines associated with segregation research, such as sociology, anthropology, cultural studies, urban studies, demography, biology, and philosophy. They involve complex definitions and self-identification based on cultural origins, nationality, colour, minority status, tribe, language, religion, historical elements or combinations of these concepts and others. Race and ethnicity are dimensions that intersect differently in different contexts. There is no universally accepted or standardized taxonomy of ethnicities because ethnicity is a complex and culturally specific concept that can vary significantly from one region or society to another (UN, 2017). Given the degree of contestation of race and ethnicity classifications, an approach that draws upon a standardized and unitary set of knowledge categories can hardly be sustained (Aspinall, 2005).
- Coders must limit themselves to recognizing segregation forms related to racial and ethnic groups as authors in SR identify them. Examples of related valid segregation terms are:
 - Segregation of social groups based on race and ethnicity:
 - 'African-American segregation', 'black-nonblack segregation', 'black racial segregation', 'black immigrant segregation' and 'black self-segregation' cannot be represented by the bigram 'black segregation'.
 - 'Chinese immigrant segregation' cannot be represented by 'Chinese segregation'.
 - 'Mexican-American segregation' and 'Mexican immigrant segregation' cannot be represented by 'Mexican segregation'; 'Chicano segregation' cannot be represented by 'Mexican segregation', as it addresses Mexican descendants.
 - 'Latinx school segregation' should be coded as 'Latino school segregation' as the latter emerged earlier and is more frequent in the sample. However, "Latinx school segregation" is used in more recent works, thus indicating a tendency to employ more inclusive language.
 - Regional groups and context-dependent definitions of ethnic groups: e.g., Maori (the indigenous Polynesian people of New Zealand), and Uyghur (a Turkic ethnic group primarily living in the Xinjiang Uyghur Autonomous Region of the People's Republic of China and dispersed in several Central Asian countries, including Kazakhstan, Kyrgyzstan, and Uzbekistan).
 - Nationality-related segregation forms, e.g. 'Mexican residential segregation', 'Puerto Rican segregation', 'American segregation', etc.
 - Segregation between racial or ethnic groups, e.g. Asian-white segregation, black-white segregation, Asian-American segregation, and Mexican-American segregation.

- Unigrams 'ethnic' and 'ethno' identify a rich diversity of forms, e.g. ethnic segregation, ethnic social segregation, ethnic minority segregation, ethnically oriented segregation, ethnic group segregation; and social groups, including intersectional social dimensions and material forms, e.g. ethnic physical segregation, ethnic residential segregation, ethnic school segregation, ethnic workplace segregation, ethnic employment segregation, ethnic vertical segregation, ethnic friendship segregation, ethnic housing segregation, ethnic network segregation, ethnospatial segregation, ethnogeographic segregation, ethnoregional segregation, ethno-ideological segregation, ethnonational segregation, ethnopolitical segregation.
 - Ethnic-religious segregation forms: e.g. ethnosectarian segregation, ethnic congregation segregation, ethnic faith segregation.
 - Ethnoreligious groups: e.g. Muslim segregation, Islamic segregation, New Christian segregation, Jewish residential segregation
- Segregation related to migration, e.g. migrant segregation, migration segregation, immigrant segregation, immigrant-native segregation, etc.
 - Coders should consider that 'migrant segregation' focuses on the segregation experienced by migrants within a host society; in turn, 'migration segregation' refers to segregation patterns related to migration processes.

Protocol 5: Segregation in the work realm

- Segregation in the occupational realm and workplace environment is a large research field with various components and aspects. 'Occupational segregation' is the most frequent bigram (1,292 hits in our sample, plus 19 for 'occupation segregation').
- Many terms overlap to address segregation in the *occupational realm* (e.g., employment segregation, occupation segregation, job segregation, work segregation, labour segregation, position segregation, horizontal segregation and vertical segregation) and the *workplace environment* (e.g. workplace segregation, establishment segregation, intra-firm segregation).
- Coders should validate such terms in recognition of such richness and subtlety.
 - Given analytical specificities, 'inter-occupational' and 'intra-occupational segregation' should not be classified as 'occupational segregation'; interfirm segregation should not be classified as 'firm segregation'. 'Interfirm racial' and 'interfirm gender segregation' should be validated as trigrams, along with 'firm racial segregation'.
 - The terms 'intra-firm segregation' and 'within-firm segregation' were not found in the Scopus corpus, even though they can be found in the Google Scholar database.
 - 'Interbuilding vertical segregation' should be validated, even though there was no bigram for 'interbuilding segregation' in the Scopus corpus (a term found in the Google Scholar database).

Protocol 6: Segregation in the educational realm

- One of the earliest SFs in the Scopus database is 'school segregation' (Gault, 1914). Educational segregation forms have since unfolded into considerable diversity. Coders should pay attention to terms addressing:
 - Subtle educational dynamics, e.g. ability segregation, learning segregation, learning child segregation, slow learner segregation;
 - Stages and ages in education, e.g. childhood school segregation, 'elementary school segregation', 'primary school segregation', 'preschool segregation', 'secondary school segregation', higher educational segregation;
 - Social and economic conditions, e.g. public school segregation, educational socioeconomic segregation, economic school segregation;

- Racial and ethnic dimensions, e.g. multiracial school segregation, educational racial segregation, ethnic school segregation, Roma school segregation, immigrant school segregation;
- Legal aspects, e.g. de jure school segregation, official school segregation;
- Intersectional forms, e.g. school gender segregation, college racial segregation, educational sex segregation;
- Spatial areas and scale: metropolitan school segregation, neighborhood educational segregation, suburban school segregation, within-school segregation, inter-school segregation, etc.

Protocol 7: Social groups and urban areas

- Coders should identify terms broadly addressing social groups, such as 'group segregation' and 'community segregation,' and terms addressing the complexities of groups and their relations, e.g., multigroup segregation, community racial segregation, community socioeconomic segregation, ethnic group segregation, racial group segregation, and cultural group segregation.
- Coders must consider terms denoting internal or external relations as segregation forms of social groups. For instance:
 - 'intergroup segregation' refers to separating different social, ethnic, or cultural groups in different contexts, such as residential areas, schools, workplaces, or public spaces.
 - 'intercommunity segregation' relates to the separation or isolation of different communities within a society, often based on race, ethnicity, religion, social class, or geography. It can occur voluntarily, where individuals or communities choose to live and associate with others who share the same characteristics, or it can result from systemic inequalities and discrimination;
 - 'intracommunity segregation' refers to the separation or division of people within a specific community or locality rather than between different communities. This segregation form can be based on social, economic, or cultural factors and manifests in different forms, including racial or ethnic segregation, where certain racial or ethnic groups within a community are separated or isolated from others. It can also refer to economic segregation, where individuals of different socioeconomic statuses live in separate areas or neighborhoods within the same community;
 - 'interclass segregation' addresses income inequality and class structures even within specific social groups, e.g. the poor and nonpoor within an ethnic group;
 - 'intrasocial segregation' addresses the separation or division of individuals within a particular racial group, i.e. when people of the same racial background are segregated or separated based on factors such as socioeconomic status, residential patterns, educational opportunities, and social interactions.
 - 'intra-african segregation' refers to the discrimination and division that exists within African societies based on factors such as ethnicity, language, religion, or socio-economic status.
 - 'intergenerational segregation' emerges when different age groups, such as children, young adults, and elderly individuals, are separated from one another and do not interact within the same spaces or communities;
 - 'Intraschool segregation' includes a specific dimension that refers to the separation or division of students within a single school or educational institution (Hernández, 2019), also addressed as 'within-school segregation'. Note that 'school segregation' is more broadly defined as segregation within the education system, from the division of students into different schools or educational institutions, with multiple connotations (e.g. segregation within schools, segregation between schools, segregation of schools from other institutions, urban areas or features, etc.).

Protocol 8: Multiplicity and intensity of segregation forms

- Coders should consider terms denoting multiplicity or intensity as features explored by authors to identify specific segregation forms, contexts or scales. For instance:
 - N-grams that recognize the complexity of segregation frequently use the prefix 'multi' to denote the condition of the diversity of contexts or scales, e.g. 'multi-contextual segregation', 'multi-dimensional segregation', 'multi-faceted segregation', 'multi-level segregation', and 'multi-scale segregation'.
 - N-grams that refer to social, cultural, or demographic groups experiencing segregation in complex relationships, e.g. 'multigroup segregation', 'multi-ethnic segregation', 'multiracial segregation'.
 - N-grams that address social relations, work and education, e.g. 'multi-occupational segregation' as it addresses diverse aspects of the working population, 'multigroup school segregation', and 'multiracial friendship segregation'.
 - 'Hypersegregation' denotes an extreme form of segregation manifested across different dimensions (Massey and Denton, 1989).

Protocol 9: Regional and place-specific situations of segregation

- Coders should recognize regional forms of segregation as they may address context-specific societal, institutional and material characteristics and historical trajectories, e.g. 'American segregation', 'American racial segregation', 'South African segregation', 'Brazilian segregation', 'European segregation'; 'Irish residential segregation', 'Chinese segregation', etc.
- Different regional contexts may present specific SFs, e.g. Apartheid segregation, a government-implemented system of racial segregation enforced in South Africa from 1948 to 1994, with strict laws that classified individuals into different racial groups and dictated where they could live, work, and socialize; 'Hukou segregation', a state-enforced form of residential segregation related to household registration in rural or urban territories in China.
- Contextual forms of segregation may include place-specific situations, such as:
 - 'swimming pool segregation', the practice of separating individuals of different races or ethnicities into separate swimming facilities or designated areas within the same pool;
 - 'hospital segregation', the practice of segregating patients based on race, ethnicity, socioeconomic status, or other characteristics in healthcare settings;
 - 'bus segregation', the practice of segregating people based on their race or ethnicity on buses, is a discriminatory practice prevalent in many parts of the world, particularly during the era of racial segregation in the United States from the late 19th century to the mid-20th century (e.g. 'local bus segregation', 'racial bus segregation').
- Structural geographical forms of segregation, such as 'north-south segregation,' should be coded as valid trigrams.
- Certain geographical delimitations can be too vague to qualify a segregation form and thus should not be coded as valid n-grams, e.g. 'Southern segregation' or 'Northern racial segregation'.

Protocol 10: Scales of segregation

- Coders recognize terms that address the active role of scale in specific forms of segregation, such as:
 - General terms addressing instances of scale, e.g. local, global, macro, micro, metropolitan, small scale, large scale, and so on;
 - Localized units or areas, e.g. 'district segregation', 'within-district segregation', 'between-district segregation';

- Trigrams combining scale, social dimensions or media of segregation, e.g. 'local gender segregation', 'local residential segregation', 'black local segregation', 'metropolitan racial segregation', 'metropolitan economic segregation', 'metropolitan residential segregation'.
- Following Protocol 4 on synonymous and interchangeable terms, n-grams containing past participle unigrams and modifiers 'based', 'level', 'specific', 'wise' should be coded under corresponding existing bigrams, e.g. 'localized segregation' coded as 'local segregation'; 'metropolitan-level segregation' as 'metropolitan segregation', 'block-level segregation' as 'block segregation', 'macro-level segregation' as 'macro-segregation'.

Protocol 11: Temporality

- Coders should note terms denoting temporal forms, situations, conditions, or periods of segregation.
- Code as valid n-grams denoting historical periods such as 'colonial', 'postcolonial', 'neocolonial', 'neo-feudal', 'modernist', 'postmodern', e.g. 'modernist segregation' as in modernist urbanism and the functional segregation of land uses and zoning in the 20th Century; colonial segregation, colonial-era segregation, intraracial postcolonial segregation; post-soviet segregation, post-socialist segregation.
 - Particular time-related n-grams denoting temporally non-specific or vague periods should not be coded, e.g. 'postwar segregation'.
- N-grams posed as adjectives addressing segregation situations that recently came into existence or are known for a short time, such as 'new', 'recent', 'novel', 'unfamiliar', 'contemporary', or not denoting historical periods, should not be coded as valid, e.g. 'new segregation', 'early segregation', 'contemporary urban segregation'.
- Terms indicating temporal regression or renewal of segregation (e.g., desegregation or resegregation) are not considered segregation forms but changes to or reversion of segregation processes.

Protocol 12: Age and segregation

- Coders should consider valid:
 - Segregation forms based on age groups such as 'adult', 'adolescent', 'youth' and 'childhood' (encompassing 'children'), 'elderly', 'aged', 'older adult', and 'age segregation'.
 - The segregation between age groups, e.g. adolescent-adult segregation'.
 - Age-based segregation forms combined with other features of segregation: 'age grade segregation', 'cultural age segregation', 'occupational age segregation', 'societal age segregation', 'residential age segregation', 'age neighborhood segregation', 'elderly residential segregation', or 'spatial age segregation' (the physical separation of age groups, i.e. "when individuals of different ages do not occupy the same space and, thereby, lack mutual interactions" – Miliadis and Psyllidis, 2022).

Protocol 13: Voluntary and involuntary segregation

- Coders should consider segregation forms denoting a personal or group drive to self-segregate based on specific characteristics or identities, as well as those denoting imposed and involuntary segregation. Several bigrams and trigrams refer to these types of segregation and their associations with other social and material segregation forms. Examples:

- 'voluntary segregation', 'self-segregation', 'self-imposed segregation', 'institutional self-segregation', 'racial self-segregation', 'Muslim self-segregation', 'self-territorial segregation'.
- 'Imposed segregation'; 'Involuntary segregation', 'involuntary spatial segregation'.
- 'Legal segregation': 'legally-enforced', 'legally accepted', 'legally entrenched', 'legally mandated', 'legally sanctioned', 'legalized state segregation', etc.; 'extralegal segregation', 'illegal segregation', 'socio-legal segregation', etc.; combinatorial terms 'legal racial segregation', 'legal residential segregation', etc.
- 'Enforced segregation': 'enforced gender', 'enforced physical', 'enforced racial', 'enforced residential' segregation, etc.
- 'Mandated segregation': 'mandated racial segregation', etc.
- 'De jure segregation' is a form mandated by law or government policy typically based on race, ethnicity, or other characteristics. It was used during the era of racial segregation in the United States, known as 'Jim Crow segregation' (also to be coded as valid).
 - Coders should also note related trigrams, such as 'de jure racial segregation' and 'de jure school segregation'.
 - 'De facto segregation' refers to segregation as a result of societal patterns and individual choices rather than being legally mandated, and should be coded as an opposed form to 'de jure segregation'.
- 'Coercive segregation', 'coercive gender segregation'.
- 'Formal segregation': 'formal job segregation', 'formal racial segregation', 'formal residential segregation'; as opposed to 'informal segregation' and its combinatorial forms, e.g. 'informal racial segregation', 'informal residential segregation', etc.
- 'Compulsory segregation': 'compulsory racial segregation', 'compulsory residential segregation', etc.
- 'Unintended segregation', 'intentional segregation' (also to be coded as valid).

Protocol 14: Negative terms in n-grams

- Even though they can certainly be helpful under specific contexts and purposes, negative terms such as 'non-economic segregation', 'non-employment segregation', 'non-racial', 'non-spatial segregation' or 'non-residential segregation' should not be coded as segregation forms as the prefix 'non' addresses the reverse of segregation forms, i.e. everything else surrounding a specific segregation form, implying virtually endless definitions.
- Exceptions are to be considered in contexts of oppositions between social groups, such as 'black-nonblack segregation' and 'poor-nonpoor segregation'.

Protocol 15: Fourgrams

- Fourgrams were omitted from the analysis because the estimated potential to add new valid segregation forms was deemed negligible in relation to the interpretive effort required to discern SFs from all four-word phrases ending in 'segregation'.

Protocol 16: Hyphen in compound terms

- N-grams referring to segregation forms may frequently use hyphens in compound terms. Indeed, the use of hyphens is inconsistent across the literature, as there is a general trend toward dropping them. Many compound terms that were once hyphenated are now written as single words.
- Therefore, as a general rule, we also leaned towards dropping the hyphen. This had the additional benefit of being more inclusive of trigrams that could otherwise be classified as fourgrams and, thus, be excluded from our sample.

- Coders adopted general guidelines to determine whether an n-gram with or without a hyphen should represent an SF as a label under which variations should be coded:
 - Relational prefixes 'intra' and 'inter': code as non-hyphenated labels, e.g.:
 - code 'intra-racial segregation' as 'intraracial segregation'
 - code 'intra-urban segregation' as 'intraurban segregation'.
 - Exception: use a hyphen when a prefix ending in a vowel is followed by a word starting with the same vowel, e.g., 'intra-African segregation'.
 - Multiplicity prefix: code as non-hyphenated labels, e.g.:
 - code 'multi-ethnic segregation' as 'multiethnic segregation'
 - code 'multi-group segregation' as 'multigroup segregation'
 - code 'multi-scale segregation' as 'multiscale segregation'.
 - Positional prefixes 'within' and 'between': code as hyphenated labels, e.g.:
 - 'within-school segregation', 'within-district segregation', 'within-group segregation'.
 - Scale prefixes 'micro' and 'macro': code as hyphenated labels, e.g.:
 - 'micro-segregation', 'macro-segregation'.
 - Prefixes addressing social and material conditions of segregation: 'socio', 'ethno', 'spatio', 'techno': code as non-hyphenated labels, e.g.:
 - code 'socio-economic segregation' as 'socioeconomic segregation'
 - code 'socio-cultural segregation' as 'sociocultural segregation'
 - Exception: use a hyphen when a prefix ending in a vowel is followed by a word starting with the same vowel, e.g., 'socio-occupational segregation'.
 - code 'ethno-cultural segregation' as 'ethnocultural segregation'
 - code 'ethno-racial segregation' as 'ethnoracial segregation'
 - code 'spatio-temporal segregation' as 'spatiotemporal segregation'
 - code 'spatio-perceptual segregation' as 'spatioperceptual segregation'
 - Temporal prefixes 'post' and 'pre': code as non-hyphenated labels, e.g.:
 - 'post-colonial segregation' as 'postcolonial segregation'
 - Suffixes 'level', 'based': in cases where such suffixes are kept, code as the hyphenated form, e.g.:
 - 'individual-level segregation', 'street-level segregation'.
 - 'people-based segregation', 'place-based segregation', 'stereotype-based segregation', 'subject-based segregation'.
- Exceptions to these guidelines still might be found in the identified SFs and their labels.

Protocol 17: Spelling style and typographical errors

- Coders should adopt the American English style of spelling to represent n-grams:
 - Use "-or" (American) for words ending in "-our" (British), e.g. code "colour" and "neighbourhood" as "color" and "neighborhood".
 - Use "-ize" (American) for words ending in "-ise" (British), e.g. code "racialised" as "racialized".
 - Follow the American English style consistently for other cases of differences in spelling.
- Eventually, segregation forms might be spelt with typos, e.g. 'socioespacial segregation' instead of 'sociospatial segregation'. Coders should consider such eventual instances and code terms under the most approximate and linguistically correct labels.

See the complete list of bigrams and trigrams in the sampled literature, their year and country of first appearance, and frequency in Table ST03.

See centrality values and the co-occurrence network derived from the sample in Table ST11.

See the definition of each code in the open collaborative project [Segregation Wiki](#).

References

Aspinall, P. J. (2005). The operationalization of race and ethnicity concepts in medical classification systems: issues of validity and utility. *Health Informatics Journal*, 11(4), 259-274.

Cascio, M. A., Lee, E., Vaudrin, N., & Freedman, D. A. (2019). A Team-based Approach to Open Coding: Considerations for Creating Intercoder Consensus. *Field Methods*, 31(2), 116–130.
<https://doi.org/10.1177/1525822X19838237>

Gault R.H. (1914). Preventives of delinquency. *Journal of Educational Psychology*, 5(1), 32-36.
<https://doi.org/10.1037/h0066312>

Hemmler, V. L., Kenney, A. W., Langley, S. D., Callahan, C. M., Gubbins, E. J., & Holder, S. (2022). Beyond a coefficient: an interactive process for achieving inter-rater consistency in qualitative coding. *Qualitative Research*, 22(2), 194–219. <https://doi.org/10.1177/1468794120976072>

Hernández, M. (2019). White middle-class families and sociocultural diversity in schools: A literature review. *The Urban Review*, 51(2), 270-300.

Massey, D. S., & Denton, N. A. (1989). Hypersegregation in US metropolitan areas: Black and Hispanic segregation along five dimensions. *Demography*, 26, 373-391.

Milias, V., & Psyllidis, A. (2022). Measuring spatial age segregation through the lens of co-accessibility to urban activities. *Computers, Environment and Urban Systems*, 95, 101829.

O'Reilly, K. (2012). *Ethnographic methods*, 2nd edition. New York: Routledge.

Schelling, T. C. (1969). Models of Segregation. *The American Economic Review*, 59(2), 488–493.
<http://www.jstor.org/stable/1823701>

United Nations. Statistical Office. (2017). *Principles and recommendations for population and housing censuses*. New York: United Nations, Department of International Economic and Social Affairs, Statistical Office.