NORTHWESTERN UNIVERSITY

Essays on the Sociological Analysis of Segregation and Natural Language

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Sociology

By

Antonio Nanni

EVANSTON, ILLINOIS

June 2022

ABSTRACT

Essays on the Sociological Analysis of Segregation and Natural Language

Antonio Nanni

This dissertation contributes to the theory of segregation and methodologies to measure it. The first two chapters focus on the traditional problem of quantifying segregation in traditional survey data through segregation indices. Segregation indices describe the segregation of an environment with one number – usually from 0 to 1. The last chapter focuses on a new form of data: unstructured textual data. It analyzes the issue of extracting stereotypical cultural schema from this kind of data using the increasingly-popular word embedding models.

In the first chapter, we show that segregation indices calculated from samples are biased and unreliable, especially in small samples. Often, researchers use segregation indices on samples to estimate the segregation in a population. Therefore, statistical inference on segregation indices is necessary, but methods to conduct this kind of inference are scarcely available and not generally used. To obviate the problem, the chapter formulates two new general techniques based on non-parametric Bayesian models. The new techniques are applicable to any segregation index or function of segregation indices. To demonstrate their capability, the chapter tests the Bayesian techniques on the D and Theil indices, and the decomposition of the Theil index. Extensive Monte Carlo simulations compare the performances of the new Bayesian techniques with the current

standard practice and currently-best available alternative, a bootstrap-based estimator. In all of the simulations, the new techniques provide more reliable inferences than previously achieved. Particularly, the Bayesian techniques appear remarkably more accurate on small samples and in the production of confidence intervals. We recommend using the new Bayesian techniques to conduct inference, especially in smaller samples.

The second chapter analyzes the issue of comparing segregation indices. Often, researchers use segregation indices to compare segregation in different environments. However, it is very difficult to interpret the differences in segregation indices between two environments, since traditional indices mixes different phenomena. The chapter formalizes the problem of interpreting change in segregation and builds a new family of indices that is interpretable from this perspective. One of its member, Q, is both interpretable and strongly decomposable, as is the Theil index. To formulate of Q, the paper also provides new results about margin-free indices (Charles and Grusky, 1995). It formulates the only way to build margin-free indices and provides a new solution to the zero problem afflicting these indices. As a result, the chapter also formulates the index Q^*, which is the first strongly-decomposable margin-free index.

The third chapter analyzes the use of word embedding models in the social sciences. Word embedding models represent each word from a textual corpus as a vector in a multi-dimensional space. They are increasingly popular in the social sciences for their ability to capture cultural schemas from readily-available textual corpora. Sociologists have used word embedding models to study a variety of different issues: from the association of obesity to gender, to the evolution of the concept of social class. A growing literature in computer science and linguistics examines how words become vectors, but fewer works analyze how to extract meaning from such vectors in order to draw social scientific conclusions. The chapter focuses on the theoretical and methodological assumptions governing the latter process. It shows that previous social scientific research relies on a simple model of meaning in word-vectors. Subsequently, it formulates a more general model linking meaning and vectors – the “simple algebra of meaning”. The simple algebra of meaning subsumes previous methodologies and paves the way for methodological innovation in the social scientific use of word embedding models. Finally, the chapter draws upon the new model to expand the current uses of word embedding models. It shows how to 1. accommodate non-binary oppositions, 2. analyze entire documents (as opposed to single words), 3. consider more than one concept at the same time, 4. decompose the meaning of documents into a function of the meaning of their words. As an example, the chapter tests the new methodologies on a corpus of 30,228 abstracts about climate change and estimates the Lovecraftian aura of words from publicly-available word embedding.

Acknowledgments

It takes a village to finish grad school. I want to thank the members of my committee, Lincoln, James, and Beth. In their unique ways they have all been intellectual guides and role models. I want to thank the colleagues and friends at NU Sociology for their support through the years. In particular, Bello, Yannick, Dylan and Devin. I want to thank the participants to the AQMW for providing me an intellectual community in these years. Thank you Sasha and Brendan for having been incredible office mates and friends.

Thank you to Scott, Amanda, Silvia and JC for making Oak Park closer to home. I am deeply indebted to Nellie for having welcomed us wholly. I don't know where I will travel to in my next chapters, but your kindness, affection, and joy will always follow me. I am happy I still managed to show you Oak Park's Igloo. Grazie.

Ma non posso ringraziare in inglese la mia famiglia in Italia: i miei genitori, mio fratello, Alga, la Dada ed il dado Mauro, ed i tanti amici. Marcello, Sotti, Giacomo, Marquez, Brogna, Sighi, il gruppo di semiotica, i nerd. Ricordo di esser partito per gli Stati Uniti pensando "Amici miei". L'ho pensato ogni giorno da allora.

Ringrazio Elena per i tanti anni insieme e per avermi accompagnato negli Stati Uniti. È stata un'avventura e sono felice di averla vissuta con te. Questa tesi, con il suo complesso sistema di sofferenza e gioia, è dedicata a te.

We've come a long, long way together

Through the hard times and the good

I have to celebrate you

I have to praise you like I should

Table of Contents

List of Tables

List of Figures

Introduction

What is segregation? How do we measure it? How do we draw robust conclusions about it from data? These are some of the most debated questions in the social sciences, where segregation often plays a central role in the study of inequality (Jahn et al., 1947; Hutchens, 1991; Frankel and Volij, 2011). This dissertation contributes to this long-standing debates with its three chapters.

The first two chapters are firmly traditional in their settings and goals. They both start with the issue of defining and measuring segregation in data from Census or similar data collection efforts. In this kind of data, individuals are often questioned about their group (for example, race) and unit (for example, neighborhood). The group-unit joint distribution constitutes the core of traditional analysis of segregation. This data form is discussed explicitly in Section 2.2, but it also underlies Chapter 1 as well as most previous work about segregation in the social sciences (Reardon and Firebaugh, 2002).

This traditional form of data has been the basis for the rich literature about segregation indices (Jahn et al., 1947; Duncan and Duncan, 1955; James and Taeuber, 1985; Reardon and Firebaugh, 2002; Frankel and Volij, 2011). Segregation indices quantify segregation in an environment starting from the joint distribution of the group and unit variables. They map any data of this kind to a real number – usually normalized between 0 and 1 – where a higher number indicates more segregation. With respect to the endeavour of quantifying segregation through indices, the contributes two new methodological and theoretical tools.

Chapter 1 discusses the issue of performing statistical inference about segregation indices. Previous literature has repeatedly noticed that the sample distribution of segregation indices appears positively biased: the segregation index of a sample is expected to be higher than the segregation index in the entire population (Cortese et al., 1976; Winship, 1977; Carrington and Troske, 1997; Allen et al., 2015). Appendix D proves that this empirical observation is based on theoretical properties of indices as mathematical (convex) functions. This is the case for the D (Duncan and Duncan, 1955), Mutual information (Theil and Finizza, 1971) and Atkinson (James and Taeuber, 1985) indices, some among the most popular indices. Chapter 1 provides a workaround to this problem: if we know *a priori* that the sample-level index is biased, cannot we systematically estimate an index value as lower than the level registered in the sample? Starting from this core intuition, the chapter develops two original, generally-applicable inferential methodologies for segregation indices: the DP and C-DP methods, which are based on non-parametric Bayesian statistics (Ferguson, 1973; Li et al., 2019). The chapter performs extensive analysis of the performances of these two new inferential approaches, which appear to be consistently more reliable than the current standard practice (plug-in estimator) and the best available alternative (bootstrap correction).

Chapter 2 discusses a core problem of the index literature: how we map a joint distribution of groups and units to a single value representing segregation (Duncan and Duncan, 1955; Coleman et al., 1982; James and Taeuber, 1985; Hutchens, 1991; Frankel and Volij, 2011). As the chapter details, there are different, incompatible approaches to this time-honored issue: the exposure approach, the evenness approach, and the association approach (Massey and Denton, 1988; Reardon and Firebaugh, 2002). These approaches are incompatible on a deep, theoretical level. Such incompatibility emerges when we consider the issue of change: how should an index change as the underlying environment changes? By construction, these approaches necessarily provide different answers (Coleman et al., 1982).

This incompatibility of the three main approaches to quantifying segregation immediately spurs a question: how do we interpret a difference in segregation indices between two environments? Notwithstanding a long, extended debate on the question (Blau and Hendricks, 1979; Semyonov and Scott, 1983; James and Taeuber, 1985; Charles and Grusky, 1995; Grusky and Levanon, 2006; Elbers, 2021), current indices and (more in general) methodologies do not provide a satisfactory answer: the difference between two indices is very difficult to interpret substantially. In fact, an analyst cannot distinguish what empirical phenomena are driving (and to what extent) the observed change in segregation levels. This is not a secondary issue, our ability to assess historical trajectories (Jacobs, 1989; Tomaskovic-Devey et al., 2006) and spatial differences (McCall, 2001) in segregation hinges on this crucial question.

The chapter reframes the issue of change and its interpretation in a broader theoretical and methodological framework. It uses recent results from mathematics (Osius, 2004) to characterize the class of margin-free segregation indices (Charles and Grusky, 1995; Elbers, 2021; Bouchet-Valat, 2022) – see Appendix F. Then, it builds on this result to create a new family of margin-free segregation indices – the Centered-Norm family. Among this family, the chapter zooms in on Q^*, which has the special property of being strongly-decomposable, like the Theil or Atkinson indices (Frankel and Volij, 2011) Q^* is the first margin-free index to have this property. Unfortunately, margin-free segregation indices may be counter-intuitive as they do not represent the experience of the individuals in an environment (Elbers, 2021). However, the chapter shows that we can modify the margin-free Q^* into Q, which is not margin-free and represents the experience of the (geometric-)average individual. Q is easy to interpret substantially, since it neatly separates the influence of the marginal distributions from the influence of the structural association element. This makes it possible to clearly distinguish the influence of the changing margins as opposed to different patterns within the units (Grusky and Levanon, 2006).

Chapter 3 leaves the traditional forms of segregation data to focus on unstructured textual data, which is increasingly abundant and an important source for the social sciences (Evans and Aceves, 2016). Unstructured textual data can be used to analyze segregation and in general the association of groups and units – for example, the association of occupations and gender (Garg et al., 2018) or the association of income and education (Kozlowski et al., 2019). In particular, the use of word embedding techniques (Mikolov et al., 2013) is growing in popularity in the social sciences (see for example Stoltz and Taylor, 2021; Arseniev-Koehler et al., 2022; Nanni and Fallin, 2021). This family of computational models maps each word of a corpus to points in space, preserving (some among) the semantic properties of the words. Importantly, the word-point maps that are produced capture implicit associations between social groups and concepts, in the roughly the same way that survey (Kozlowski et al., 2019; Joseph and Morgan, 2020; van Loon and Freese, 2022) or implicit association tests (Caliskan et al., 2017) capture them. The point is that it is often easier to obtain substantial amount of unstructed texts from groups than to survey them – especially when exploring the culture of groups from the past (Nelson, 2021).

Notwithstanding the growing interest and the great potential of word embedding techniques to explore social scientific topics, the methodological and theoretical basis for the use of these techniques in the social sciences remains fuzzy. The basic idea that word embedding captures semantic relations through the distributional hypothesis (words appearing in similar contexts are similar, see Harris, 1954) is under-specified. As Lenci (2008) notices, the distributional hypothesis is not even a semantic hypothesis primarily. In fact, word embedding can also retrieve morphological regularities such as verb inflection (see for example Levy and Goldberg, 2014a). Thus, the distributional hypothesis is not sufficient to justify word embedding efficacy to retrieve semantic relations – let alone to justify the specific methodologies that computer scientists and social scientists use to retrieve these relations. As as a result, the social sciences have developed a battery of methodologies that appear closely related and usually support the same conclusions (Joseph and Morgan, 2020).

Chapter 3 provides a geometric-semantic model that serves two purposes: on the one hand, it justifies the current methodological practices theoretically and, on the other hand, it suggests further methodological development based on the theory. As per the theoretical development, the chapter proposes a semantic-geometric model that regulates how semantic relations are translated into the mutual positions of point-words in space. This model legitimizes the current common practice in the analysis of embedding – that is, the use of cosine similarity – but also suggests new ways to use word embedding models in the social sciences. The new expansions encompasses: analysis of non-binary categories, multi-dimensional projections of words, analysis of documents, and the decomposition of document analysis over the length of the document. These are new methodologies. The chapter showcases their potential with an analysis of Lovecraftian terms from pre-trained Google News word embeddings (for analysis of non-binary dimension) and an analysis of 30,000 climate-change-related abstracts for all other methods (Nanni and Fallin, 2021).

The three chapters together advance our current theoretical and methodological understanding of segregation both for traditional data and, potentially, for new unstructured sources of data.

CHAPTER 1

Better Statistical Inference for Segregation Indices: Two General Non-parametric Bayesian Approaches

1.1. Introduction

Segregation is one of the fundamental concepts for our understanding of inequality. Segregation indices quantify the amount of segregation between two or more groups in a given environment (school district, city, job market, etc.) by analyzing how the groups are distributed in different organizational units (schools, census tracts, occupations, etc.).[1] In the long debate on segregation, much energy has been spent in the creation of new indices because different indices may support different conclusions, making the choice of an index a matter of contention (Duncan and Duncan, 1955; Massey and Denton, 1988; Hutchens, 2004).

However, previous research has mostly ignored the statistical properties of segregation indices, even if researchers often use indices on sample data (see for example Beller, 1984; Charles and Grusky, 1995; Cotter et al., 1997; McCall, 2001; Blau et al., 2013). Naturally, any index calculated from a sample is a sample statistic estimating the population index. Whereas social scientists routinely use statistical inference on sample statistics such as proportions or model coefficients, statistical inference on segregation indices is seldom performed because inferential techniques are often not available or simply not popular.

This is not a small omission: previous research shows that inference on segregation index is critical. First, early works (Cortese et al., 1976; Winship, 1977; Carrington and Troske, 1997) showed that a sample from a population will likely exhibit some segregation even when the groups are distributed uniformly at random in the organizations – a condition that intuitively coincides with the absence of segregation (Jahn et al., 1947; Williams, 1948). These works proposed ways to amend such undesirable characteristic of segregation indices, even if such amendments are rarely used in practical application. Second, recent research (Ransom, 2000; Rathelot, 2012; Allen et al.,

[1] Massey and Denton (1988) distinguish as many as five different dimensions of segregation that indices may pick up on: evenness, exposure, concentration, centralization, clustering. Following their classification, this paper considers exclusively evenness-based indices, but the techniques presented are extensible to the other dimensions as well.

2015; D'Haultfœuille and Rathelot, 2017; Logan et al., 2018; Reardon et al., 2018) has created techniques to conduct full statistical inference for specific segregation indices. Most of these works (Ransom, 2000; Rathelot, 2012; Allen et al., 2015; D'Haultfœuille and Rathelot, 2017) have focused on the D and Gini indices, while others have focused on the income segregation indices H^R and R^R (Logan et al., 2018; Reardon et al., 2018). Overall, these two lines of research clearly indicate that statistical inference on segregation indices is difficult because of the positive bias in the sample statistic. Unfortunately, the proposed inferential tools are mostly limited to few indices: the literature has not focused on techniques that can be applied smoothly to the vast array of situations that arise in applied research.

For this reason, this paper focuses on inferential techniques with general applicability. We use the word "general" in the following sense:

1: A general inferential technique applies to all kinds of indices: traditional two-groups indices (Duncan and Duncan, 1955; Massey and Denton, 1988), multigroup indices (Reardon and Firebaugh, 2002), and spatial indices (Reardon and O'Sullivan, 2004).

2: A general inferential technique applies to all indices. Applying it to a new index requires, at most, the calculation of a derivative.

3: A general inferential technique can produce inference for functions of indices, such as the decomposition of the Theil index (Lichter et al., 2015; Fiel, 2013; Ferguson and Koning, 2018).

The previous literature has formulated two techniques that are generalizable, if not fully general. First, Ransom (2000) proposed an asymptotic inferential procedure based on the delta method. Originally, he applied it to inference for the D and Gini indices, but it can be extended further: below, we apply this technique to the Theil index. More recently, Allen et al. (2015) formulated bootstrap techniques for inference on the D index. Like the delta method, these bootstrap techniques can be applied in a vast array of situations (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). Even if inferential techniques tailored to specific indices may be more efficient (see for example Allen et al., 2015), these general inferential techniques are extremely helpful since they cover the vast array of needs arising in applications.

Following this guidelines, this paper formulates two new general approaches to statistical inference for segregation indices based on non-parametric Bayesian modeling – specifically, the Dirichlet process mixture model. The paper pursues two distinct inferential objectives. First, it evaluates the different techniques based on their ability to provide a precise point estimate. Point estimates of segregation indices are useful because they often enter as variables in statistical models (see for example Cutler and Glaeser, 1997; Quillian, 2014): if the point estimates are systematically (not randomly) biased, the resulting model will also be biased. Second, the paper evaluates the ability of the estimators to perform interval estimation. This is mostly helpful when the analyst seeks to assess segregation trends from samples (see for example King, 1992) and needs an assessment of the variability of the point estimate. The precision of point estimates will be evaluated through their Root Mean Squared Error (RMSE); intervals will be evaluated based on their coverage properties. As its driving application, the paper will use gender segregation on the workplace (Hakim, 1979; Jacobs, 1989). Therefore, we will use "men" and "women" as groups and "occupations" as units for concreteness. The focus on gender segregation also entails that the paper will exclusively consider segregation between two groups, as opposed to three or more groups. This is convenient since it simplifies the testing procedure and the presentation of the results, but the techniques can be extended to multigroup and spatial settings.

The paper proceeds by first introducing some notation and two segregation indices: the D (Duncan and Duncan, 1955) and Theil index (Mora and Ruiz-Castillo, 2011). Possibly, these are the most popular segregation indices. Howevr, the scope of the discussion extend beyond these indices. Then, the paper formulates the problem of statistical inference for segregation indices. It empirically shows that the sampling distribution of the D and Theil indices are biased – a well known issue (Cortese et al., 1976; Paninski, 2003; Allen et al., 2015). Incidentally, Appendix D proves that the D sample values are positively biased and extends this results to other indices – that is, the mutual information index (Mora and Ruiz-Castillo, 2011) and the Atkinson index (James and Taeuber, 1985; Frankel and Volij, 2011). The fourth section formulates the four general inferential approaches that the paper compares: the (1) plug-in, (2) bootstrap, (3) Dirichlet process and (4) corrected Dirichlet process approaches. Finally, the next two sessions test these approaches. In the fifth section, the paper tests the behaviour of the different inferential techniques in small sample situations. In the fifth section, the paper tests the inferential techniques on the decomposition of the Theil index (Frankel and Volij, 2011; Mora and Ruiz-Castillo, 2011) by sampling from complete 1940 Chicago census data (Logan et al., 2018). As far as we know, this is the first paper to propose and test inferential techniques for the decomposition of a segregation index.

Alongside the evaluation of the Bayesian techniques, the paper evaluates the current standard practice of using the sample value of a segregation index as an estimate – referred to as the "plug-in" estimator. Both the bootstrap and Bayesian inferential approaches appear superior to the current practice in every test. Based on the simulation tests, on any previous test about this matter (compare Logan et al., 2018; Reardon et al., 2018; Rathelot, 2012; Allen et al., 2015), and on the mathematical results presented in Appendix D, the conclusion is that the current practice needs fixing because it is biased and can result in grossly-inaccurate results. This is the case for both indices examined here as well as, to the best of our knowledge, for all indices ever analyzed from this perspective.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	
W	4,027	6,427	4,182	2,701	6,145	106	95	102	1,390	486	3	25,664
M	5,856	5,691	3,570	3,607	2,012	357	3,754	2,367	3,227	3,262	8	33,711
	9,883	12,118	7,752	6,308	8,157	463	3,849	2,469	4,617	3,748	11	59,375

Table 1.1. A table representing the distribution of women (W) and men (M) in 11 different occupational categories in the US in 2003. The data is a representative sample from the Current Population Survey microdata.

Besides the meager performances of the plug-in estimator, the proposed Bayesian techniques appear generally superior to the current state of the art – i.e. the bootstrap approach by Allen et al. (2015). This is particularly the case in smaller samples, where previous techniques often fail egregiously. From this perspective, the newly-formulated Bayesian techniques allow researchers to perform reliable inference on segregation indices in cases where both the current standard practice (the plug-in estimator) as well as the current state of the art (the bootstrap technique) fail.

1.2. Notation and Indices Formulation

Before proceedings, it is helpful to introduce the notation to be used in the paper with the help of Table 1.1 as a reference:

G**:** The number of different groups whose segregation we want to assess. For example, in Table 1.1, $G = 2$ and the groups are men and women.

O**:** The number of different organizational units in an environment. For example, in Table 1.1, $O = 11$ and the organizational units correspond to occupational groups.

T**:** An environment whose segregation must be assessed. T is formally represented as a matrix in ${\mathbb{R}_+}^{G\times O}$, where $\mathbb{R}_+$ are the positive real numbers. For example, see Table 1.1.

N**:** The total number of individuals in Table T. For example, in Table 1.1, $N = 59,375$.

$T_{i,}$**:** The i^{th} row of an environment T, representing the distribution of a single group in different occupations. For example, in Table 1.1 the row for the second group (M) is $[5,856; 5,691; 3,570; 3,607; 2,012; 357, 3,754;\ 2,367; 3,227; 3,262; 8]$

$T_{,j}$**:** The j^{th} column of T, showing the group distribution within the j^{th} occupation. For example, in Table 1.1 the column for the third occupation (O_3) is $[3,570; 4,182]$

$T_{i,j}$**:** The number of individuals belonging to group i and organization j. For example, in Table 1.1, $T_{2,3} = 22,620$

π_G**:** The proportion of individuals belonging to any of the groups in the environment T. π_G is a vector in $\mathbb{R}^G$ whose sum is 1. For example, in Table 1.1: $\pi_G = \left[\frac{25,664}{59,375}, \frac{33,711}{59,375}\right] =$ $[0.432; 0.568]$.

$\pi_{g(i)}$**:** The proportion of individuals in the i^{th} group in the environment T. For example, in Table 1.1: $\pi_{g(2)} = \frac{33,711}{59,375} = 0.568$.

π_O**:** The proportion of individuals in any of the occupation in the environment T. π_O is a vector in $\mathbb{R}^O$ whose sum is 1. For example, in Table 1.1:

$$\pi_O = \left[\frac{9,883}{59,375}; \frac{12,118}{59,375}; \frac{7,752}{59,375}; \frac{6,308}{59,375}; \frac{8,157}{59,375}; \frac{463}{59,375}; \frac{3,849}{59,375}; \frac{24,69}{59,375}; \frac{4,617}{59,375}; \frac{3,748}{59,375}; \frac{11}{59,375}\right] =$$

$[0.166; 0.204; 0.131; 0.106; 0.137; 0.008; 0.065; 0.042; 0.078; 0.063; 0.000]$.

$\pi_{o(j)}$**:** The proportion of individuals in the j^{th} occupation in the environment T.For example, in Table 1.1: $\pi_{o(4)} = \frac{6,308}{59,375} = 0.106$.

$\pi_{O|g(i)}$**:** The proportion of individuals from any occupation among individuals belonging to group i. This is the normalization of row i by its row-margin. Formally, $\pi_{O|g(i)}$ is a vector in $\mathbb{R}^O$ whose sum is 1. For example, in Table 1.1:

$$\pi_{O|g(2)} = \left[\frac{5,856}{33,711}; \frac{5,691}{33,711}; \frac{3570}{33,711}; \frac{3,607}{33,711}; \frac{2,012}{33,711}; \frac{357}{33,711}; \frac{3,754}{33,711}; \frac{2,367}{33,711}; \frac{3,227}{33,711}; \frac{3,262}{33,711}; \frac{8}{33,711}\right]$$

$= [0.1737; 0.1688; 0.1059; 0.1070; 0.0597; 0.0106; 0.1114; 0.0702; 0.0957; 0.0968; 0.0002]$.

$\pi_{o(j)|g(i)}$: The proportion of individuals from occupation j among individuals belonging to group i. This is element j from $\pi_{O|g(i)}$. For example, in Table 1.1: $\pi_{o(3)|g(1)} = \frac{4{,}182}{25{,}664} = 0.163$.

$\pi_{G|o(j)}$: The proportion of individuals from any group among individuals in occupation j. This is the normalization of column j by its column-margin. Formally, $\pi_{G|o(j)}$ is a vector in $\mathbb{R}^G$ whose sum is 1. For example, in Table 1.1: $\pi_{G|o(2)} = \left[\frac{6{,}427}{12{,}118}; \frac{5{,}691}{12{,}118}\right] = [0.5304; 0.4696]$.

$\pi_{g(i)|o(j)}$: The proportions of individuals belonging to group i among individuals in occupation j. This is element i from $\pi_{G|o(j)}$. For example, in Table 1.1: $\pi_{g(1)|o(3)} = \frac{4{,}182}{7{,}752} = 0.5395$.

1.2.1. The D and Theil Indices

Based on the notation introduced above, a segregation index is a function $S(T)$ from a $G \times O$ matrix, T, to $\mathbb{R}$. Usually, $S(T)$ is normalized in such a way as to take value from 0 to 1, but this is not necessary. As mentioned in the introduction, the paper will consider only tables with two groups – that is, matrices with two rows. A long debate exists about the form of the function $S(T)$ and about its desirable properties (Duncan and Duncan, 1955; James and Taeuber, 1985; Hutchens, 1991; Frankel and Volij, 2011). Here, we will sidestep this debate by analyzing two among the most popular indices: the D and Theil indices.[2] Following the conceptual map of Massey and Denton (1988), these indices are evenness-based indices: they quantify segregation as the difference between the distributions of the two (or more) groups over occupations. Both indices are bounded between 1 and 0 and can be written in two forms: both as a function of $\pi_{O|g(i)}$ and as a function of $\pi_{G|o(j)}$. The two alternative formulations are helpful with different inferential techniques and both are presented below.

[2] In addition to these two indices, we also tested the four inferential techniques on the Gini (Hutchens, 1991) and Atkinson (James and Taeuber, 1985) indices. The complete results of the simulation tests are available as additional material. The substantial conclusions presented in the main text do not change when we consider these additional indices.

D: The D index was introduced by Jahn et al. (1947). It has been a very popular choice in applied research (see for example Jacobs, 1989; Stainback and Tomaskovic-Devey, 2012) and it is usually interpreted as the proportion of the minority group that needs to change occupation in order for the environment to be desegregated completely. Given a table T, it can be calculated as:

$$D(T) =$$

$$\frac{1}{2}\sum_{j=1}^{O} |\pi_{o(j)|g(1)} - \pi_{o(j)|g(2)}| = \tag{1.1}$$

$$\frac{1}{2}\frac{1}{\pi_{g(1)}\pi_{g(2)}}\sum_{j=1}^{O} \pi_{o(j)}|\pi_{g(1)|o(j)} - \pi_{g(1)}| \tag{1.2}$$

Theil: The Theil index is an entropy-based segregation index first proposed by Theil and Finizza (1971) drawing upon the concept of mutual information from information theory (Cover and Thomas, 2006). The index has grown in popularity (see for example Lichter et al., 2015; Ferguson and Koning, 2018) due to its important decompositional quality. In fact, its non-normalized version – named the ‘mutual information index’ (Theil, 1971) – is ‘strongly decomposable’ (Mora and Ruiz-Castillo, 2011; Frankel and Volij, 2011). However, even the Theil index analyzed here can be decomposed in a component representing segregation between sub-sets of units and a component representing segregation within the sub-sets – see Section 1.7 for an example. Now, let the entropy of a discrete distribution π be:

$$H[\pi] = -\sum_{i}^{K} \pi_i \log(\pi_i) \tag{1.3}$$

	O_1	O_2	O_3	π_G
W	0.2	0.1	0.1	0.4
M	0.1	0.2	0.3	0.6

Table 1.2. A table representing the relevant data for the calculation of a segregation index. The number in each cell represents the proportion of individual in a given group and unit.

where K is the number of categories in the support of π and π_i is the probability of π resulting in category i.[3] The Theil index can be formulated as:

$$Th(T) =$$

$$\frac{1}{H(\pi_G)} \sum_{i=1}^{G} \pi_{g(i)} \left(H[\pi_O] - H[\pi_{O|g(i)}] \right) = \tag{1.4}$$

$$\frac{1}{H(\pi_G)} \sum_{j=1}^{O} \pi_{o(i)} \left(H[\pi_G] - H[\pi_{G|o(j)}] \right) \tag{1.5}$$

1.3. The Statistical Behavior of the Plug-in Estimator

In bare terms, the goal of inference on segregation indices is to estimate the segregation index of an environment based on a sample from such environment. For this inferential task, we can consider both the environment and its sample as tables like Table 1.1, where the sample will contain a subset of the individuals from the sampled environment. However, not all information in Table 1.1 is relevant. Indeed, most segregation indices are scale invariant (Hutchens, 1991; Frankel and Volij, 2011): the value in each cell may be divided/multiplied by a positive constant without changing the value of the index, entailing that the table total (N) is irrelevant for the value of a segregation index. Therefore, to simplify exposition, both the environment tables and the sample tables will be represented as tables of proportions such as Table 1.2, where the content of every cell

[3] The base of the logarithm in the entropy equation is ultimately irrelevant in the present context. The paper uses the natural logarithm throughout.

represents the proportion of individuals in a given unit. We will indicate with τ the environment table and with T the sample table. In general, we will indicate with $\hat{S}(\tau)$ the (point) estimate of a segregation index and with $S(\tau)$ the value of the index in the full environment table.

It is default practice to calculate the index on the sample and use this as the final estimate without further elaboration – we will refer to this practice as the "plug-in" estimator. For example, if Table 1.2 was a sample table, we would apply equation (1.1) to the table to calculate the D index in the sample. In general, we will indicate with $S(T)$ the value an index in the sample; in this case, $S(T) = 0.33$. The plug-in estimator uses $S(T)$ as the final estimate of $S(\tau)$ – see equation (1.6). Thus, we would have $\hat{S}(\tau) = 0.33$ for Table 1.2 according to the plug-in estimator. As Ransom (2000) notices and we show in Appendix D, the plug-in estimator is consistent if the sampling schema is consistent. In addition, this estimator is undeniably convenient for its simplicity and generality: it is straightforward to calculate and will always apply. So, one may wonder if alternatives to the plug-in estimator are necessary at all.

There are reasons to suspect it is unreliable in small samples. Cortese et al. (1976) and Winship (1977) first noticed the positive bias of the D index when estimating the index of a non-segregated environment (see also Carrington and Troske, 1997). However the problem is broader: plug-in estimates appear to be biased in the entire range of an index. For example, Paninski (2003) shows implicitly that the plug-in estimator of the Theil index is biased in practically any situation (see also DeDeo et al., 2013). Allen et al. (2015) show that the plug-in estimates of the D index are almost always upwardly biased when the environment has 3 or 4 categories. Beyond the indices considered here, Logan et al. (2018) and Reardon et al. (2018) show that the plug-in estimations of the residential income segregation index H^R is upwardly biased when applied to data from a stratified sample – so much as to question previous conclusions about the rising of residential income segregation based on the American Community Survey (see for example Reardon and

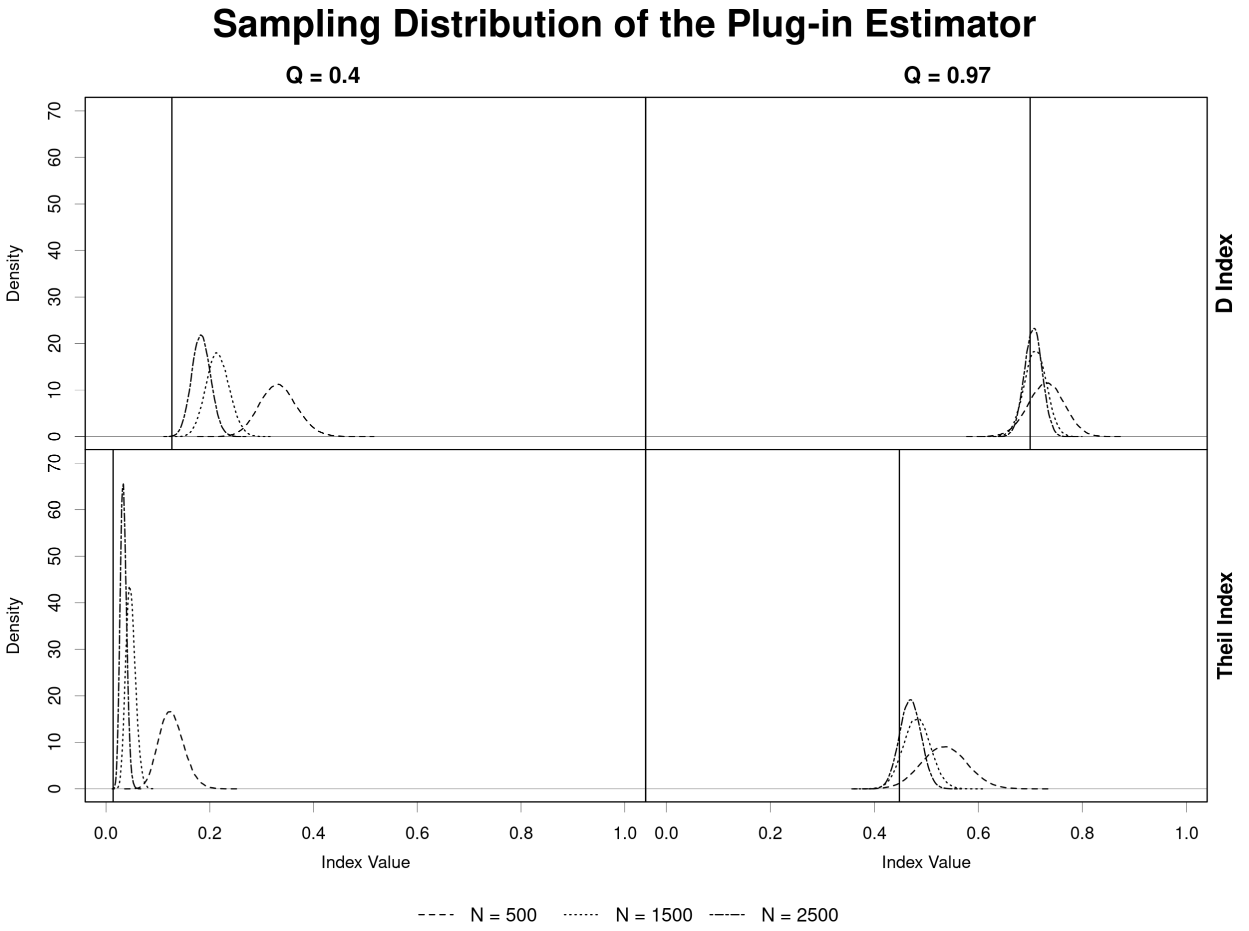
Bischoff, 2011). In fact, to the best of our knowledge, all the tests ever performed about this matter reach the same conclusion that the plug-in estimator is positively biased.

To further test this issue, we analyze the sample distribution of the plug-in estimator mathematically and empirically. Appendix D formally shows that the plug-in estimator is consistent, but positively biased: it is biased for any positive index when the environment's segregation is zero (generalizing Cortese et al., 1976; Winship, 1977) and it is almost always biased for the D (Atkinson, and Theil) indices. Unfortunately, these results do not reveal whether the positive bias is relevant from an empirical perspective. However, we can use simulations to analyze the issue empirically. Here, we use the simulation settings described in section 1.6 to study the sampling distribution of the plug-in estimator for the D and Theil indices.

Figure 1.1 compares the sampling distributions of the plug-in estimator for a table with 50 occupations under different sample sizes (500, 1,500, and 2,500) and different values of the Q parameter (0.4 and 0.97) when the minority proportion (P) is fixed at 0.2. As explained below, the Q parameter regulates the overall segregation of a simulated environment: a value of 0 corresponds to no segregation and a value of 1 corresponds to maximum segregation.[4] However, the value of the D and Theil indices do not increase linearly with Q. From this perspective, the two values of Q considered in the figure correspond to a regime of low (Q=0.4) and high (Q=0.97) segregation, but they are by no means extreme. The figure shows two important trends. First, the plug-in estimator is positively biased in all cases, even if the amount of bias changes depending on the index, sample size and value of Q. Second, sample size reduces bias and variance, but it is not sufficient to get an accurate estimate in the low-segregation regime. Even for the bigger sample sizes, the sampling distributions barely contain the population value for the two indices.

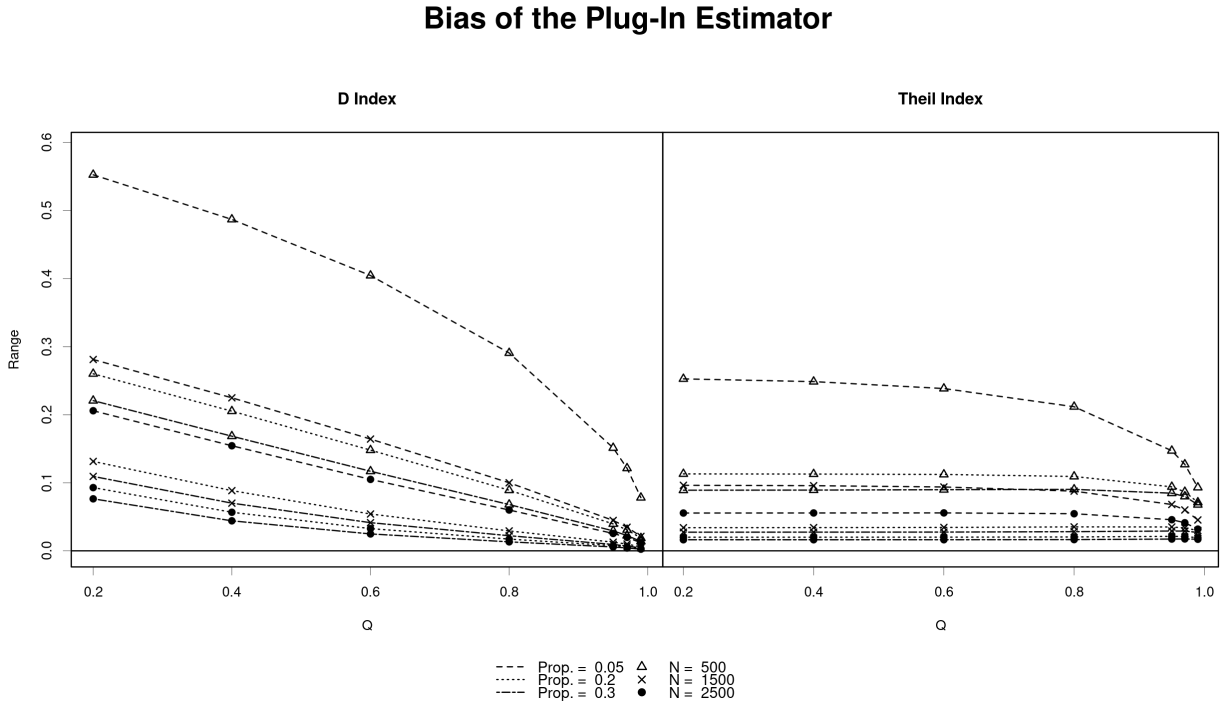
[4] The Q parameter is the only parameter of the hyperbola model (Duncan and Duncan, 1955), used below to create simulated environments. Its range spans the interval from 0 to 1 – see Section 1.6.

Figure 1.1 Examples of sampling distribution of the plug-in estimator for the D and Theil indices at different levels of Q and different sample sizes when the proportion minority (P) is fixed at 0.2. The dotted vertical line shows the true values of the index in the sampled environment.



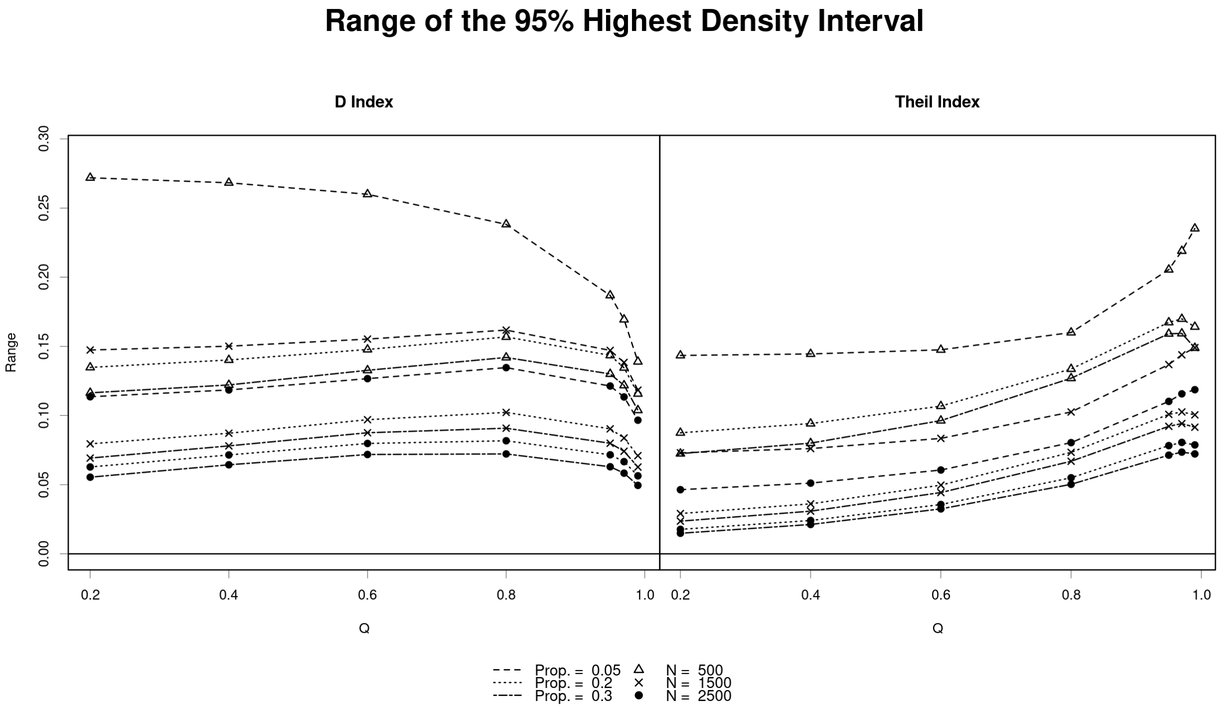
To analyze the issue further, Figure 1.2 shows the plug-in estimator's bias in all of the simulations analyzed in section 1.6: we examine all possible combinations of seven different Q values, three different sample sizes, and three P values. The figure shows that the plug-in estimator always presents a positive bias, which can be extreme in the least favorable situations. For example, the plug-in estimates overestimate the D index of roughly 0.55 on average in the most unfavourable situation. Consider that the D index ranges from 0 to 1: therefore the plug-in inflates the index for

Figure 1.2 Bias of the plug-in estimator of the D and Theil indices under different parametrizations of the Monte Carlo set-up described in section 1.6. “Prop.” is the proportion of individuals from the minority group in the environment. “N” is the sample size.



55% of its possible value. This means that it is substantially worthless to use the plug-in estimator for the D index in this case. Besides this extreme case, the bias of the plug-in estimator diminishes as sample size, index value in the environment (Q), and the proportion of the minority group increase. However, the bias remains substantial for lower value of the population indices in all tests. Finally, it is worth noticing the bias of the plug-in estimates does not follow the same pattern for the two indices considered here. Whereas the bias of the D index decreases consistently as *Q* increases, the Theil indices follows a different pattern altogether. As the right panel of Figure 1.2

Figure 1.3 Variability of the plug-in estimator of the D and Theil indices under different parametrizations of the simulation. "Prop." is the proportion of individuals from the minority group in the environment. Variability is quantified as the range of the 0.95 highest density of the sampling distribution of the plug-in estimator.



shows, the bias of the Theil index tends to remain constant up until very high values of the index in the population, where it finally diminishes.[5]

In addition to the expected bias issue, the plug-in estimator exhibits substantial variability even in relatively large sample. We quantify the variability of the plug-in estimator as the length of the 0.95 Highest Density Region (HDR) of the sampling distribution of the estimator. This interval represents the shortest achievable length for a 95% confidence interval once the plug-in estimator's bias has been corrected. Figure 1.3 shows the variability patterns of the plug-in estimator for the D

[5] If a statistic is positively biased (as is the case for segregation indices), the bias will necessarily diminish when the population value reaches the upper bound of the statistic's possible range. Intuitively, if a segregation index spans the interval from 0 to 1, its positive bias cannot exceed 0.01 when the population value is 0.99.

and Theil index in the same simulation settings as above. The figure shows that plug-in estimator can be substantially variable in the samples. This is especially the case for the D index, with the length of the HDR above 0.05 in all tested configuration except one. Even for the Theil index, however, roughly 75% of the configurations tested have an HDR above 0.05. Once again, the reader should consider that the indices range from 0 to 1, so that an interval of 0.05 covers 5% of its entire span. Like bias, variability depends on different factors. As expected, sample size diminishes it. Moreover, a higher minority proportion results in smaller variability, all else being equal. Finally, the effect of Q on the index examined: for the D index, variability diminishes for higher values; for the Theil index, variability generally augments for higher values.

As a whole, these results indicate that the plug-in suffers from practically-relevant positive bias and variability, even in relatively large samples. Given this is a consistent finding in the literature, we suspect these conclusions will hold beyond the present empirical test. If this is indeed the case, the plug-in estimator provides a distorted picture, where the differences between segregation regimes in different environments are smoothed out by the positive bias: everything looks more segregated than it actually is. Moreover, the plug-in estimator merely provides a point estimate: by itself, this estimate lacks any assessment of uncertainty. Curiously, it is currently accepted to not report any kind of interval estimate for segregation indices, even if this practice is strongly discouraged for virtually any other sample statistic. Yet, this is inconvenient considering the large variance of the sampling distributions.

For these reasons, it is important to create tools to perform reliable statistical inference in all those situations where the plug-in estimator is used. At the very least, we will need an interval estimator to assess the uncertainty of the plug-in point estimate.

1.4. General Tools for Inference

In this section, we illustrate four inferential approaches for segregation indices: (1) the plug-in estimator plus delta method, (2) the bootstrap-corrected method, (3) the Dirichlet process method (DP), and (4) the corrected Dirichlet process (C-DP) method. Importantly, we only consider approaches that are generally applicable to any segregation index (excluding e.g. Reardon et al., 2018; D'Haultfœuille and Rathelot, 2017) and to functions of segregation indices, preferably with minimal calculation required to adapt them to new indices or functions. For example, the inferential methodologies should be applicable to the decomposition of the Theil index or the difference between indices from different samples. Preferably, the proposed techniques should also require minimal calculations – only computational time to draw random samples. This way, practitioners will be able to get proper statistical inference in most situations where they currently only have the plug-in point estimates at their disposal.[6]

The former two methodologies were previously introduced by Ransom (2000) and Allen et al. (2015), respectively. They are generalized here to account for new indices and functions of indices. As discussed above, the plug-in estimator is the current standard as a point estimator, but it usually lacks any accompanying interval estimate. We use the asymptotic delta method for this purpose because it centers around the plug-in point estimate. As for the bootstrap method, it is worth mentioning that different bootstrap techniques may be generally applicable to segregation indices (see for example Allen et al., 2015). Here, we focus on the bias-correction technique because it promises to amend at least partially the bias issues plaguing the plug-in estimator, while being straightforward to implement. The Bayesian techniques are new. Like the bootstrap techniques, they use computer simulations to diminish the bias in the plug-in estimator and to calculate confidence intervals.

[6] Technically, we assume continuity of the index function $S(\tau)$ and we do not focus on the boundaries of τ.

1.4.1. The Plug-in Estimator and Delta Method

The plug-in estimator is simply the value of a segregation index as calculated from the sample. Asymptotically, this estimator is consistent and will eventually converge to the population value of a segregation index (Ransom, 2000). A little more formally, we can write:

$$\hat{S}_{PI}(\tau) = S(T) \tag{1.6}$$

where $\hat{S}_{PI}(\tau)$ represents the plug-in point estimate for a segregation index from a sample; $S(T)$ is the value of the segregation index in the sample.

Notice that equation (1.6) is merely a point estimate with no indication about uncertainty. Ransom (2000) proposed to use the delta method to assess the sampling uncertainty of the plug-in estimator in the case of the D and Gini indices. Indeed, the delta method provides an asymptotically-correct estimation of the standard error. We can use this standard error to calculate an asymptotic confidence interval centered around the plug-in point estimator. More in general, we can apply the delta method to any index: the relevant assumptions are that the index is a continuous function, has a continuous first-order derivative, and the derivative is different from zero. To the best of our knowledge, all popular indices satisfy these conditions.[7]

Therefore, we can generally write the confidence interval for the plug-in estimator at the α confidence level as:

$$CI_{PI}(\alpha)[S(\tau)] = \hat{S}_{PI}(\tau) \pm z_{1-\frac{\alpha}{2}} \cdot \sigma_{\Delta}[S(T)] \tag{1.7}$$

[7] It is worth mentioning we are making a more fundamental assumption about the sampling process: the sampled table T will converge to the population table τ. However, this assumption has nothing to do with segregation indices and their statistical properties. Therefore, we do not discuss it explicitly in the main text. We notice, however, that the assumption clearly holds for simple sampling design like simple random sampling – used in the simulations.

where $z_{1-\frac{\alpha}{2}}$ is the $1-\alpha$ percentile from a standard normal distribution and $\sigma_\Delta[S(T)]$ is the standard error as calculated from the delta method. The reader may notice that this confidence interval is very similar to the classic confidence interval for the mean. This is not by chance: equation (1.7) is ultimately an application of the central limit theorem (Agresti, 2013, chap. 16).

The actual calculation of $\sigma_\Delta[S(T)]$ requires some algebra based on the derivative of $S(T)$: complete formulas are presented in Appendix A for the case of the D and Theil indices. However, it is important to notice that the formula used in this paper are valid for simple random sampling only, which is the sampling strategy used in the simulation tests below.[8] Other sampling schemes require different calculations. For example, if the population is sampled through a group-stratified sampling – that is, the sample contains a fixed number of randomly chosen individual from each group – the standard error from the delta method would differ from the formulas presented in Appendix A. This detail is overlooked by Ransom (2000) when presenting this strategy for the first time. In our tests, we checked the consequences of mis-specifying the sampling scheme for the delta method and we empirically found that this mis-specification does not impact the performances of the estimator substantially.[9] However, we only considered three simple sampling schemes: simple random sampling and sampling stratified by group or occupation.[10] For more complex sampling strategies, the delta method calculation may be effectively impossible. We may opt to use the delta method calculated for a simple random sample schema – such as the formulas in Appendix A – hoping that the estimator will not dramatically underperform for this reason.

[8] We further assume an infinite population. This assumption is necessary for any asymptotic result to hold.

[9] Results available by the author.

[10] The `SISeg` R package backing this paper contains delta method calculation for simple random sampling and stratified sampling by group or occupation for 7 segregation indices.

Finally, the plug-in and delta estimator can easily be generalized to conduct inference on functions of an index or even on functions of indices from different samples. For example, the difference between indices from different environments. If we generally indicate with $T_1, T_2 \dots T_m$ the different samples from the environments $\tau_1, \tau_2 \dots \tau_m$ respectively and with $f\big(S(\tau_1), S(\tau_2) \dots S(\tau_m)\big)$ the function of interest in the population, then we can easily calculate $f\big(S(T_1), S(T_2) \dots S(T_m)\big)$ on the sample tables. Therefore, the plug-in estimator for this function:

(1.8)

$$\hat{f}_{PI}\Bigg(S(\tau_1), S(\tau_2) \dots S(T_m)\Bigg) = f\Bigg(S(T_1), S(T_2) \dots S(T_m)\Bigg)$$

(1.9)

$$CI_{PI}(\alpha)\Bigg[f\big(S(\tau_1), S(\tau_2), \cdots\big)\Bigg] = \hat{f}_{PI}\Bigg(S(\tau_1), S(\tau_2) \dots S(T_m)\Bigg) \pm z_{1-\frac{\alpha}{2}} \cdot \sigma_\Delta\Bigg[f\big(S(\tau_1), S(\tau_2), \cdots\big)\Bigg]$$

where $\sigma_\Delta\Bigg[f\big(S(\tau_1), S(\tau_2), \cdots\big)\Bigg]$ is the standard error for the function and sampling schema of interest as calculated through the delta method. For these equation to be asymptotically correct, it is sufficient that the function $f(\cdot)$ have continuous first derivatives different from zero in the range of interest. Moreover, we assume that the samples $T_1 \dots T_m$ are mutually independent – for example, independent samples of the same population or samples from different populations. [11]

To simplify exposition in what follows, we refer to all of the different estimates discussed in this section simply as the "plug-in estimator". As discussed above, we expect these estimators to be unreliable when the sample size is small. In this case, the point estimates will be grossly biased and, therefore, the confidence interval will be centered around an inaccurate estimate. However, even when the sample size is larger, we expect the other techniques to perform better because

[11] This assumption may not always be satisfied. For example, if two samples share a substantial number of sampled individuals because of their design, then the techniques illustrated in the main text are incorrect. This would be the case for some complex sample designs such as the Current Population Survey where different waves may share a substantial portion of sampled individuals.

the plug-in point estimates will remain the most biased among the point estimates considered; the other techniques try to compensate for the positive bias we expect to observe in the sample. At the same time, it is extremely important to examine the performances of the plug-in estimator because it is by far the most popular techniques in applied research – especially, the point estimate from equation (1.6) (see for example King, 1992; McCall, 2001; Quillian, 2014). Therefore, this estimator is a baseline for comparison.

1.4.2. Bias-Corrected Bootstrap Method

Since the plug-in estimate suffers from upward bias, it would be desirable to correct it by subtracting its bias: this is what the bootstrap bias-corrected confidence interval and point estimate do. Allen et al. (2015) introduce this method to produce inference for the D index. Here, we generalize it for the broader task of inference on segregation indices.

Bootstrap techniques are commonly used to produce inference in finite samples for biased estimators (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). In particular, the bias-correction method estimates the bias of an estimate through bootstrap repetitions and then subtracts it from the original estimate to create a new, supposedly-unbiased estimate. However, to create a truly unbiased estimate with this technique, the bias of original estimator should not change as a function of the quantity we are trying to estimate. This means that the value of the environment's segregation index (Q, in our simulations below) should not influence the bias of the plug-in estimator for this technique to work perfectly. Unfortunately, the bias of the plug-in estimator decreases as Q augments as shown in Figure 1.2. Therefore, the bootstrap bias-correction cannot create a genuinely de-biased estimate. Yet, as the tests show, this technique still improves over the plug-in point estimates substantially. The confidence interval is then constructed around the bias-corrected point estimate by using the appropriate percentile from the bootstrap distribution created for the bias estimation (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). For easiness of exposure, we refer to this method simply as the “bootstrap method”, even if we warn the reader that alternative bootstrap methodologies may be potentially applied to this inferential task – for example Allen et al. (2015) use the studentized bootstrap confidence interval.

To produce the bootstrap point estimate and confidence interval, the first step is the calculation of the plug-in point estimate – $\hat{S}_{PI}(T)$ from equation (1.6). The second step is the production of the bootstrap repetitions for the estimation of bias. This can be done efficiently by drawing from a multinomial distribution. First, we calculate the proportion of sampled individual in each cell (see for example Table 1.1):

$$\pi[i,j] = \frac{T_{i,j}}{N} \tag{1.10}$$

Then, we treat this proportion as the probability parameter in a multinomial distribution. From this distribution we draw N variates, where N is the sample size of the sample table T. This procedure is equivalent to resampling with replacement from the original sample – the staple of non-parametric boostrap (Efron and Tibshirani, 1994). By repeating this procedure B times, we can produce an arbitrary number of simulated tables T^*, on which we can calculate any segregation index. Let $S(T_i^*)$ be the value of the index calculated on the i^{th} simulated table.

Now consider the statistics:

$$b_i[S(T)] = \hat{S}_{PI}(T) - \left(S(T_i^*) - \hat{S}_{PI}(T) \right) = 2 \cdot S(T) - \hat{S}_{PI}(T_i^*)$$

The difference $S(T_i^*) - \hat{S}_{PI}(T)$ can be interpreted as an estimation of the bias of $\hat{S}_{PI}(T)$. Intuitively, $b_i[S(T)]$ represents a de-biased version of the plug-in estimator.

From here, we can use the sample distribution of $b_i[S(T)]$ (calculated from the bootstrap repetitions) to create new point and interval estimates. We can use the expected value of this statistic as the point estimate:

$$\hat{S}_B(\tau) = E\Big[b_i[S(T)]\Big] \approx \frac{1}{B}\sum_{i=1}^{B}\Big[\hat{S}_{PI}(\tau) - \big(S(T_i^*) - \hat{S}_{PI}(\tau)\big)\Big] = 2\cdot\hat{S}_{PI}(\tau) - \frac{1}{B}\sum_{i=1}^{B} S(T_i^*) \tag{1.11}$$

Finally, we can create a percentile confidence interval. Let $b[S(T)]_\alpha$ be the α percentile in the (simulated) sampling distribution of $b[S(T)]$. Then, we can construct a two-sided confidence interval at the α confidence level as:

$$CI_B(\alpha)[S(\tau)] = \big[b[S(T)]_{\alpha/2}\,;\, b[S(T)]_{1-\alpha/2}\big] \tag{1.12}$$

Like the plug-in estimator, the bootstrap technique can be generalized to scalar functions of indices from different samples, $f\big(S(\tau_1), S(\tau_2)\ldots S(\tau_m)\big)$. From each sample, can produce B bootstrap simulations independently. Then, we can match the simulated samples so as to obtain B independent bootstrap simulations, $S(T_{1,i}^*), S(T_{2,i}^*)\ldots S(T_{m,i}^*)$. From here, we can easily calculate $f\big(S(T_{1,i}^*), S(T_{2,i}^*)\ldots S(T_{m,i}^*)\big)$ on each simulation and use these values to construct a point and an interval estimate. As before, consider the statistic:

$$b_i\Big[f\big(S(\tau_1), S(\tau_2)\ldots S(\tau_m)\big)\Big] = 2\cdot\hat{f}_{PI}\Big(S(\tau_1), S(\tau_2)\ldots S(T_m)\Big) - f\Big(S(T_{1,i}^*), S(T_{2,i}^*)\ldots S(T_{m,i}^*)\Big)$$

Then, the point and interval estimates will be:

(1.13)

$$\hat{f}_B\Big(S(\tau_1),S(\tau_2)\dots S(\tau_m)\Big) = E\Big[\, b\big[f\big(S(\tau_1),S(\tau_2)\dots S(\tau_m)\big)\big]\,\Big]$$

(1.14)

$$CI_B(\alpha)\Big[f\big(S(\tau_1),S(\tau_2),\cdots\big)\Big] = \Big[b[f\big(S(T_1),S(T_2),\cdots\big)]_{\alpha/2}\,;\,b[f\big(S(T_1),S(T_2),\cdots\big)]_{1-\alpha/2}\Big]$$

In conclusion, the bootstrap method is a generally-applicable method providing an alternative to the plug-in estimator. Like the plug-in estimator discussed above, the bootstrap estimator is asymptotically correct for the most popular segregation indices, since they have a continuous first derivative (Shao and Tu, 1995). Substantially, the bootstrap method will be applicable in all cases where the plug-in estimator is applicable. Moreover, tests by Allen et al. (2015) and D'Haultfœuille and Rathelot (2017) show that this method substantially outperforms the plug-in estimator – mostly due to its ability to partially correct for the positive bias of the plug-in estimator. It is also able to easily provide confidence intervals and point estimates for functions of indices, where the delta method would be hard to calculate. Even if it relies on computationally-intensive simulations, the bootstrap estimator is actually quick to implement in modern-day computers: for commonly used indices, it will require few seconds to draw thousands of bootstrap simulations. For all of these reasons, we consider this method the current state of the art among general methodologies for inference on segregation indices.

1.5. A Non-parametric Bayesian Approach

We introduce two new methodologies based on the non-parametric Bayesian Dirichlet process mixture model (DPMM). The DPMM is a clustering model, which will be used to cluster together occupations based on their minority proportion – we discuss momentarily why clustering occupations might be any helpful. The DPMM is non-parametric in the sense that it does not limit the number of clusters in the data *a priori*. Technically, the model supposes that there are infinitely many clusters, but we can only observe a finite number of them since we observe a finite amount of data (Gershman and Blei, 2012). However, the DPMM still requires the analyst to fix one hyper-parameter, whose influence we analyze in Appendix C.

First, we provide an intuitive explanation of rationale behind these estimators. Second, we provide a more formal introduction to the DPMM using the Chinese restaurant process as a metaphor (Gershman and Blei, 2012; Li et al., 2019). Finally, we introduce two new estimators based on the DPMM: the Dirichlet process – DP(C) – estimator and the corrected Dirichlet process estimator – C-DP(C). Details about the Monte Carlo sampling procedure can be found in Appendix B.

1.5.1. The Core Intuition

Consider a table such as Table 1.2. Assume this table represents the observed proportion in each cell in a sampled table. We can create a statistical model of how Table 1.2 came to be sampled. In particular, we assume that the sample picks a fixed number of individuals from each occupation and that each occupation contains infinitely many individuals in the population. Under these assumptions, the sampling strategy treats each occupation column as an independent binomial distribution[12] parametrized by an occupation-specific probability vector, $\Pi_{G|O(i)}$. As usual, we are interested in estimating $S(\tau)$ starting from the observed data. Under this sampling strategy, $S(\tau)$ solely depends on the population-level probability vectors, $\Pi_{G|o(i)}$ (Rathelot, 2012; D'Haultfœuille and Rathelot, 2017). Therefore, the objective becomes to estimate these vectors.

[12] Occupations are modeled as multinomial distributions if there are more than two groups in the environment.

If the sample is small, it may be difficult to tell whether different occupations actually have different probability vectors. Indeed, even when the group proportions of two occupations are different in the observed sample ($\pi_{G|o(i)} \neq \pi_{G|o(j)}$), such difference may only be due to sampling variability, while at the population level the two occupations really share the same probability vector ($\Pi_{G|o(i)} = \Pi_{G|o(j)}$). At the extreme, if every occupation shares the same probability vector, then all the segregation we observe in the sample will be due to mere sampling variability – since any segregation index would be 0 in that case (Williams, 1948). Beyond this limit case where all observed segregation is due to sampling, we expect less segregation in an environment as the number of distinct probability vectors diminishes (while the number of occupations remain the same). This is immediately clear when we consider the decomposition of the Theil index: if a group of occupations share the same probability vector, the segregation within this group will be zero. Therefore, any within-group segregation we observe among these occupations would be a sampling artifact.

We can use this intuition in the inference process. We can assign the occupations of an environment into mutually exclusive clusters, with members of the same cluster supposedly sharing the same probability vector. This operation will generally result in an estimate of $S(\tau)$ that is lower than the plug-in estimator, since occupations within the same cluster will have zero within-cluster segregation. This is desirable given the positive bias we suspect the plug-in estimator to have.

This idea was first introduced by Rathelot (2012), albeit with a different statistical interpretation.[13] However, we improve substantially on the original proposal. First and foremost, Rathelot (2012) considers the number of clusters as a parameter that the analyst has to specify. This immediately implies that the point estimate of the original method is not consistent unless the analyst uses some statistically-grounded method to decide how many clusters there should be in an environment. This is what we do below using DPMM to re-establish the consistency of the point estimator. Second, Rathelot (2012) adopts a frequentist approach for estimation which appears to suffer from convergence issue (Allen et al., 2015, p. 58). We re-frame the model in a Bayesian framework using convenient conjugacy properties. This means that our model reliably converges in a relatively-fast time, which is what we empirically observe in the simulations. Finally, the Bayesian re-framing allow us to use the same estimation technique to draw inferences about functions of segregation indices – not simply their value in a population. This was a much harder problem in a frequentist framework.

1.5.2. DPMM for Segregation Indices

A Dirichlet process can be characterized as a probability distribution over probability distributions (Ferguson, 1973). We can use the Dirichlet process prior to cluster occupations from a sample table using DPMM (Escobar and West, 1995). Unlike other clustering techniques requiring the number of clusters as a hyper-parameter, this kind of mixture models will infer the number of clusters from

[13] Rathelot (2012) uses a different statistical framework to justify the model. Whereas we are interested in sampling, this author considers segregation as arising from a random process of recruitment of individuals into occupations. The objective is to draw inference about this process for the future, not about an original population(see also D'Haultfœuille and Rathelot, 2017). Moreover, the frequentist framework used in the original paper implies a different interpretation of the beta distribution in the mixture. Whereas we consider the beta distribution as a prior (and a posterior, given conjugacy) encoding our uncertainty about the probability vector $\Pi_{G|o(i)}$, Rathelot (2012) considers drawing from the beta-mixture as part of the random process itself. This implies that occupations in the same clusters do not share the same probability distribution, but the same beta-binomial distribution.

the data. We can describe the data-generating process behind the DPMM through the metaphor of the Chinese restaurant process (Li et al., 2019), which we will adapt to our specific inferential objective.

Metaphorically speaking, this process regulates how clients (representing an occupation O_i) are distributed in different tables (representing clusters c_j) in a Chinese restaurant (representing an environment τ). We assume the restaurant potentially has an infinite number of tables available. The first client entering the restaurant will always sit at the first table with no other clients sitting there – since no other client entered before her. The client will order some food, representing the minority proportion Π_1 in the occupation O_1. We assume the food at each table will be drawn from a prior Beta distribution. More in general, for any table (cluster c_j), the associated food (minority proportion $\Pi_{c(j)}$) will be drawn from the same prior distribution:

$$\Pi_{c(j)} \sim Beta(\alpha_0, \beta_0),\ \alpha_0, \beta_0 > 0 \tag{1.15}$$

where α_0 and β_0 are hyper-parameters to be specified. In all of the following tests, we fix them at $\alpha_0 = \beta_0 = 1/2$, which is the Jeffreys prior for this model (Bernardo and Smith, 2009).

After the first client, the second client arrives. She will either sit at the same table as the first client or she will sit alone at table number two. We will indicate the table of the i^{th} client with $c(i)$. The second client will sit at table one or two with the following probabilities:

$$P(c(2) = k) = \begin{cases} \frac{1}{1+C} & \text{for } k = 1 \\ \frac{C}{1+C} & \text{for } k = 2 \end{cases}$$

Here C is the concentration hyper-parameter regulating the likelihood that a new table (cluster) will be initiated by a client (occupation). As C grows, we will in general observe more clusters. If the second client joins the first at her table, she will consume the same food – that is, she will get the same minority proportion Π_1. Otherwise, she will order new food (proportion) from the prior distribution in equation (1.15).

More in general, the assignment of clients to tables will follow a rich get richer mechanism, where the probability of a new client sitting at a table k is proportional to the number of clients already sit at table k. We indicate with $O(k)$ the number of clients at table k, with $K(j_{-1})$ the number of different tables occupied before the arrival of client j, and with $c(j_{-1})$ the table assignment of all the clients arrived before j. Then, the probability of the j^{th} client sitting at any table is:

$$P(c(j)=k|c(j_{-1})) = \begin{cases} \frac{O(k)}{O-1+C} & \text{for } k \leq K(j_{-1}) \\ \frac{C}{O-1+C} & \text{for } k = K(j_{-1})+1 \\ 0 & \text{otherwise} \end{cases} \tag{1.16}$$

Leaving the metaphor aside, being assigned a cluster is equivalent to being assigned a minority proportion, $\Pi_{c(j)}$. Once an occupation j is assigned a cluster and a cluster-level proportion $\Pi_{c(j)}$, the number of minority (majority) individuals observed in the sample table T within occupation j will follow a binomial distribution:

$$\begin{aligned} T_{1,j}|c(j), \pi_{c(j)} &\sim Bin(T_{,j}, \Pi_{c(j)}) \\ T_{2,j}|T_{1,j} &= T_{,j} - T_{1,j} \end{aligned} \tag{1.17}$$

Here, the total number of individuals in an occupation ($T_{,j}$) are considered fixed – we can directly observe this parameter from the sample T.

Given that the Chinese restaurant metaphor is sequential in nature, it may be surprising to observe that the described process is actually exchangeable. This means that any order between occupations would result in the same final likelihood for table T (Gershman and Blei, 2012), entailing that we do not have to come up with any way to order occupations for the purpose of this technique. Based on exchangeability, we can create a Gibbs sampler to infer the cluster $c(j)$ of each occupation and the proportion $\Pi_c(j)$ for each cluster (Neal, 2000) – the sampler used in this paper is documented in Appendix B.

We can use this data generation process to produce a point and an interval estimate for any segregation index. From this perspective, the population table τ is the result of equations (1.15) and (1.16). Equation (1.17) models the act of sampling from the population to create the sample table T. For our present purposes, we can use the DPMM to produce statistical inference for segregation indices. We propose two different methods based on the DPMM: the Dirichlet process method – abbreviated in DP(C) – and the corrected Dirichlet process method – abbreviated in C-DP(C). Admittedly, the DPMM is not a realistic model of the way segregation arises in a social environment, but this is not the point. Given what we know about inference for segregation indices – namely, the positive bias of the plug-in estimator and the possible advantage of clustering occupations – the DPMM may provide a valuable, general alternative to the plug-in estimator and the bootstrap method. Indeed, the DPMM tries to cluster occupations, but it will not impose a clustering structure if the data clearly does not allow for it. The number of clusters is implicitly regulated by the concentration parameter C from equation (1.15): a higher value for C will result in more clusters, all else being equal. However, C works like a Bayesian prior in that its influence will vanish as the sample size augments. When the sample size augments, the number of clusters augments and occupations will mostly be grouped by themselves – see Appendix E. Therefore, the rich get richer dynamics of cluster assignment – equation (1.15) – generally results in a small number of clusters in small samples, entailing that the DPMM-based methods will produce lower estimates than the plug-in estimator; yet, in larger samples, these methods will produce estimates that are close to the asymptotically correct plug-in estimator. Generally, this is a desirable behaviour.

Notice that in a finite sample, the actual value of C will still bear some influence on the final estimate. To emphasize the dependence of the proposed methods on C, we explicitly indicate C in their abbreviation, $DP(C)$ and $C-DP(C)$. In sections 1.6 and 1.7, we fix $C=1$; thus, we actually test $DP(1)$ and $C-DP(1)$ in those experiments. In Appendix C, we test the effects of C on inference in the same Monte Carlo simulation setting. In the range of C tested, the results show that the effect of C are consistently small: as expected, C is practically irrelevant in the larger samples. However, the effects of C (and α_0 and β_0) are worth further exploration in future research.

1.5.3. The DP(C) Method

In the DPMM described in section 1.5.2, the cluster-specific $\Pi_c(j)$ from equation (1.15) represents the minority proportion in each occupation in the population. Such population-level $\Pi_{c(j)}$ are sufficient to calculate the population value of any index – $S(\tau)$ – if in addition we also know the proportion of the population in each occupation, which we will indicate with Π_O. As a matter of fact, we can approximate Π_O from the sample table T.[14] So, all we need from the DPMM are estimates of the different $\Pi_{c(j)}$, indicated with $\hat{\Pi}_{c(j)}$. For example, to get an estimate for the D and Theil indices given π_O and the inferred $\hat{\Pi}_{c(j)}$, it is sufficient to swap $\hat{\Pi}_{c(j)}$ for $\pi_{g(1)|o(j)}$ in equations (1.2) and (1.5).

Therefore, the objective is to infer the posterior distribution of $\hat{\Pi}_{c(j)}$ given T. We can do this inference through a Gibbs sampler on the DPMM – see Appendix B. The Gibbs sampler will produce B different samples from the posterior distribution of $\Pi_{c(j)}$ for all occupations, where B is

[14] While this may not be formally correct, Π_O is considerably easier to estimate than $S(\tau)$, so that its sampling error would not influence the final inference substantially. In other words, if π_O is not precise enough as an estimator of Π_O, estimating $S(\tau)$ with any sort of precision is probably impossible.

chosen by the analyst. As discussed in greater details in the Appendix, we aggregate these samples to create B tables, which are effectively samples from the posterior distribution of the environment τ. We indicate the i^{th} sampled environment with τ_i^*. From here, we can calculate $S(\tau_i^*)$ on the sampled tables to create a posterior distribution for $S(\tau)$, which we will indicate with $\mathscr{S}(\tau)$. This posterior can be used both to create a point estimate and an interval estimate.

As for the point point estimate, different reasonable Bayesian point estimates of $S(\tau)$ can be calculated from $\mathscr{S}(\tau)$. Here, we specifically use the expected value of τ because it minimizes the mean squared error (MSE) of the estimator (Bernardo and Smith, 2009, Proposition 5.2); we use the root mean squared error (RMSE) as a metric to judge the different estimators in the experiments below. Therefore, the point estimate we use is:

$$\hat{S}_{DP(C)}(\tau) = E\left[\mathscr{S}(\tau)\right] \approx \frac{1}{B}\sum_{i=1}^{B} S(\tau_i^*) \tag{1.18}$$

In our experiments, other possible point estimators – such as the median of $\mathscr{S}(\tau)$ – barely change the point estimates.

As for the interval estimate, a Bayesian confidence interval[15] at the α level for $S(\tau)$ can be estimated through the highest density region technique (Bernardo and Smith, 2009, Definition

[15] The use of the words 'confidence interval' for Bayesian models is technically incorrect as it bears clear reference to the hypothesis-test paradigm, which does not apply in a Bayesian framework. Many other names, such as 'credible interval' have been proposed. Here, we use this incorrect expression to emphasize the similarity of the goals of the DP methods and the other methods discussed. Technically, credible intervals should not be read as confidence intervals since their meaning is different: Bayesian credible intervals represent intervals that contain parameter with a certain probability, in light of the data, the model and the priors. Likely, credible interval provides a more intuitive interpretation with respect to confidence intervals, which is more aligned to the actual interpretation of interval estimation by practitioners (Morey et al., 2016).
However, we are still judging the Bayesian intervals for their frequentist properties. The possibility of using credible intervals as confidence intervals depends on the applicability of a Bernstein-von Mises theorem to the Dirichlet process mixture model, which is complex to establish and still an active area of research (Ghosal and van der Vaart, 2017). However, even if a Bernstein-von Mises theorem holds, this may not be relevant for the problem at hand since the Dirichlet process mixture model is hardly a realistic model of the data in the first place. Here, we sidestep these complex issues and we simply notice that empirically the DP method and the corrected DP method provide good coverage in the experiments below.

5.5). Let $[\mathscr{S}(\tau)]_\alpha$ be the α percentile of the posterior distribution $\mathscr{S}(\tau)$. The interval of interest is:[16]

$$CI_{DP(C)}(\alpha)[S(\tau)] = HDR(1-\alpha)\left[\mathscr{S}(\tau)\right] = \Big[[\mathscr{S}(\tau)]_a\,;\,[\mathscr{S}(\tau)]_b\Big] \tag{1.19}$$

$$where\ a,b : [\mathscr{S}(\tau)]_b - [\mathscr{S}(\tau)]_a = \min_{l-u=1-\alpha}([\mathscr{S}(\tau)]_u - [\mathscr{S}(\tau)]_l)$$

In practice we estimate the percentile of $\mathscr{S}(\tau)$ from the Gibbs samples $S(\tau_i^*)$: in symbols, $[\mathscr{S}(\tau)]_\alpha \approx [S(\tau*)]_\alpha$.

In addition to estimating a segregation index, the DP(C) method can be applied to estimate functions of the index, or even functions of indices from multiple sample tables. Assume we have m sample tables $T_1, T_2 \dots T_m$ from the environments $\tau_1, \tau_2 \dots \tau_m$. We can use Gibbs sampling to produce B variates from each of the posterior distributions of the segregation index, $\mathscr{S}(\tau_1),\ \mathscr{S}(\tau_2) \dots \mathscr{S}(\tau_m)$. We indicate with $\tau_{j,i}^*$ the i^{th} environment sampled from the posterior of the j^{th} environment. Inference for a function $f\big(S(\tau_1), S(\tau_2) \dots S(\tau_m)\big)$ can be based on its posterior distribution, $f\big(S(\hat{\tau}_1), S(\hat{\tau}_2) \dots S(\hat{\tau}_m)\big)$. Once again, we can use the expectation and highest density region of this distribution to create the point and the interval estimate, respectively. In symbols:

$$\hat{f}_{DP(C)}\Big(S(\tau_1), S(\tau_2) \dots S(\tau_m)\Big) = \sum_{i=1}^{B} \frac{1}{B} f\Big(S(\tau_{1,i}^*), S(\tau_{2,i}^*) \dots S(\tau_{m,i}^*)\Big) \tag{1.20}$$

$$CI_{DP(C)}(\alpha)\Big[f\big(S(\tau_1), S(\tau_2), \cdots\big)\Big] = HDR(1-\alpha)\Big[f\big(S(\hat{\tau}_1), S(\hat{\tau}_2) \dots S(\hat{\tau}_m)\big)\Big] \tag{1.21}$$

We approximate the percentile from this posterior with the percentile from the Gibbs samples, as we did for equation (1.19).

[16] We are assuming that the posterior is unimodal, which is the case in all simulations and model we have estimated for this paper. The definition would change slightly for non-unimodal posteriors.

Finally a note about computation time for the DP(C) method. While this method is certainly longer to compute than the plug-in and the bootstrap methods, its Gibbs sampler implementation is relatively quick. The computation time depends linearly on the number of columns in the sample table. In our experiment with the Chicago complete data in section 1.7, we tested this method on a sample table with more than 900 columns. In this rather extreme case, the method takes less than 3 minutes to sample the posterior 1,000 times on the cluster computer where the experiments were run – including a burn-in period of 500 iterations. We think this computational time is reasonable and it does not prevent the DP(C) method from being employed in practical applications.

1.5.4. The C-DP(C) Method

Suppose we have a known environment τ and an established sampling procedure. By applying the sampling procedure repeatedly to τ, we could create r different sample tables $T_1, T_2 \dots T_r$, each with its own plug-in point estimate $S(T_1), S(T_2) \dots S(T_r)$. These estimates would allow us to observe the sampling distribution of the plug-in estimator and calculate its bias. Assume we calculated the bias with no error and we observe a new sample table T_{r+1}. We could subtract the calculated bias from the plug-in point estimate $S(T_{r+1})$ to obtain a new unbiased point estimate with the same variance as the original plug-in estimate.

In this example, calculating the bias is unnecessary in the first place: we can calculate segregation directly on τ with no samples or bias involved. Yet, the core idea is still helpful in the usual setting, where we only observe one sample table T. In this setting, the issue becomes one of probability inversion: given the table T and a sampling procedure, which environment τ^* is likely to have produced T? From here, what is then the bias that likely lurks in the plug-in point estimate? Ultimately, this is the approach that the bootstrap method use to de-bias the plug-in estimator – see section 1.4.2. Inspired by the idea of Bayesian bootstrap (Rubin, 1981), we can suppose that the DPMM is the sampling procedure that created the observed sample table T. If that is the case, then the same Gibbs sampler that allows us to conduct inference for the DP(C) method will also allow us to create a posterior distribution of the plug-in bias. We can use this posterior to create new point and interval estimates. Importantly, even if the DPMM is not a realistic data generation process, this point estimator is consistent in the same case where the plug-in estimator is consistent – see Appendix E.

More precisely, let τ_i^* be the i^{th} sampled table from the posterior of distribution of τ. Then, using equation (1.17) we can create new, simulated samples from this environment, T_j^*. We can use T_j^* to estimate the bias of the plug-in estimator under the hypothesis that τ^* is the environment of interest. If we created R different simulated samples, then the bias would be calculated as

$$E_{T^*}\Big[\beta[S(T)|\tau^*]\Big] \approx \frac{1}{R}\sum_{j=1}^{R}\Big(S(T_j^*) - S(\tau^*)\Big)$$

Assume we sample B different posterior environments τ^* through the Gibbs sampler. Then, by the law of total expectation, the overall bias of the plug-in estimator can be estimated as:

$$E[\mathscr{B}(S)] = E_{\tau^*}\Big[\beta[S(T)]\Big] = E_{\tau^*}\Big[E_{T^*}\big[\beta[S(T)|\tau^*]\big]\Big] \approx \frac{1}{B}\sum_{i=1}^{B}\frac{1}{R}\sum_{j=1}^{R}\Big(S(T_j^*) - S(\tau^*)\Big) \tag{1.22}$$

where we explicitly signal in the sub-scripts the distribution over which the expectation is taken. Moreover, notice that the various $S(T_j^*) - S(\tau^*)$ are effectively samples from the posterior of the plug-in bias; for notational convenience, we indicate this distribution with $\mathscr{B}(S)$. Notice that $\mathscr{B}(S)$ is specific to the S index of interest and (more implicitly) to the observed sample T.

Overall, equation (1.22) calculates the bias we expect for a plug-in estimator given the data and the assumption that a Dirichlet process mixture model is the data-generating process. In principle, we can fix R as large as we like, but in the experiment that follows we simply use $R = 1$. Starting from this equation, we can produce a point estimate:

$$\hat{S}_{C-DP(C)} = E\left[\hat{S}_{PI}(\tau) - \mathscr{B}(S)\right] = \hat{S}_{PI}(\tau) - E\left[\mathscr{B}(S)\right] \tag{1.23}$$

Since we are sampling from the posterior $\mathscr{B}(S)$, we can use the same highest density region approach of equations (1.19) and (1.21) to create an interval estimate:

$$CI_{C-DP(C)}(\alpha)[S(\tau)] = HDR(1-\alpha)\left[\hat{S}_{PI}(\tau) - \mathscr{B}(S)\right] = \left[\hat{S}_{PI}(\tau) - [\mathscr{B}(S)]_a\, ;\, \hat{S}_{PI}(\tau) - [\mathscr{B}(S)]_b\right] \tag{1.24}$$

where $[\mathscr{B}(S)]_i$ indicates the i^{th} percentile from the plug-in bias posterior, which we approximate from the Gibbs samples.

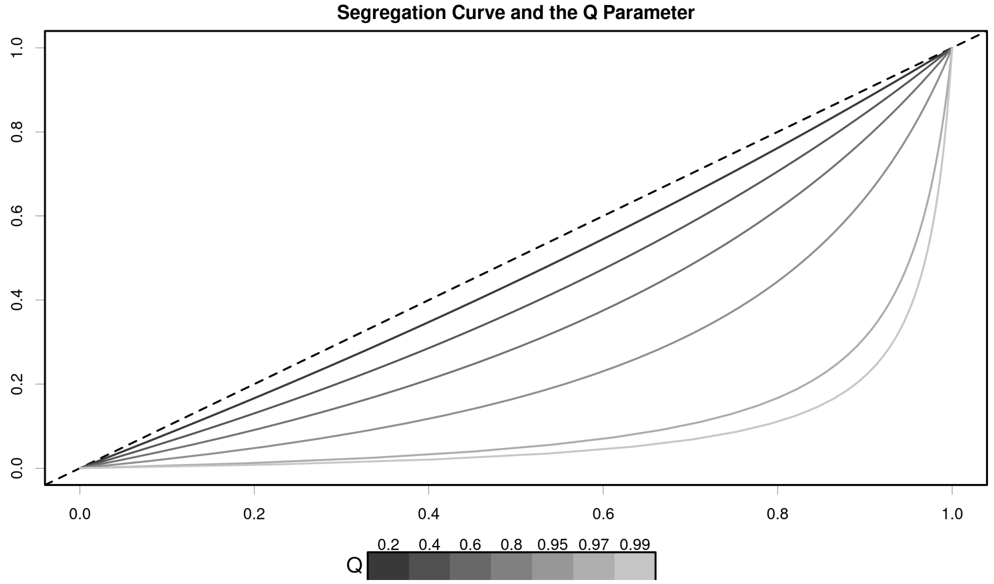
Moreover, it is straightforward to extend this method to estimate arbitrary functions of one or more segregation indices. Using the same strategy above, we can sample from the posterior distribution of the plug-in bias for the arbitrary function $f\big(S(\tau_1), S(\tau_2)\dots S(\tau_m)\big)$. Let $\mathscr{B}(f)$ indicate this posterior distribution. Then, the point and interval estimates will be

$$\hat{f}_{C-DP(C)}\left(S(\tau_1), S(\tau_2)\dots S(\tau_m)\right) = \hat{f}_{PI}\left(S(\tau_1), S(\tau_2)\dots S(\tau_m)\right) - E\left[\mathscr{B}(f)\right] \tag{1.25}$$

$$CI_{C-DP(C)}(\alpha)\left[f\big(S(\tau_1), S(\tau_2), \cdots\big)\right] = HDR(1-\alpha)\left[\hat{f}_{PI}\left(S(\tau_1), S(\tau_2)\dots S(\tau_m)\right) - \mathscr{B}(f)\right] \tag{1.26}$$

Finally, we notice that the computational time to get a point and interval estimate using the C-DP(C) method is substantially identical to the time necessary to get a DP(C) estimate.

Figure 1.4 Segregation curves of the synthetic environments sampled in the Monte-Carlo tests. Higher value of Q correspond to curves that are more distant from the non-segregation line – that is, the oblique, dashed line.



1.6. Simulation Tests

This section shows the results of extensive simulation-based tests on the methods presented in the section above. The test strategy follows (and extend) the strategy by Allen et al. (2015). The simulations test the amount of bias and the root mean squared error (RMSE) of the different point estimators as well as the coverage properties of the corresponding confidence intervals.

1.6.1. Simulation Setup

To simulate the sampling distribution of segregation indices, we first need environments' tables to be sampled. To create tables with an increasing amount of segregation, we use the hyperbola model by Duncan and Duncan (1955) to create segregation curves for the artificial environments. This model produces segregation curves following the trajectory of an hyperbole in the relevant portion of the Cartesian plane – i.e. in the square with vertices in $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. The hyperbola model has only one parameter Q between 0 and 1 regulating the amount of segregation. Specifically, we use the following values for Q: $0.2, 0.4, 0.6, 0.8,\ 0.95, 0.97, 0.99$. A lower value corresponds to less segregation. These values for Q create the segregation curves shown in Figure 1.4. In general, a higher value of Q corresponds to a higher segregation in the environment following the Lorentz criterion (Hutchens, 1991). Noticeably neither D nor the Theil strictly follow this criterion strictly. However, Duncan and Duncan (1955) show that the D index is the maximum vertical distance between the 45^o no-segregation line and the segregation curve of the population. Therefore, it is immediately clear from Figure 1.4 that D increases monotonically with Q. As for the Theil Index, its value is also influenced by the group proportion P, contradicting the Lorentz criterion. However, for fixed group proportions, the Theil index monotonically increases as Q increases.

The hyperbola model does not depend on the minority proportion in the environment, which is a free parameter. In the simulations, we consider three different values for the minority proportion P: $0.05 0.2, 0.3$. Therefore, we consider a total of 27 different environments to sample, resulting from the combination of the different values of Q and P. In all of the artificial environments, there are fifty occupations – i.e. $O = 50$ – and all occupations have the same expected number of individuals in them. After the creation of the synthetic environments, we simulated sampling procedures from the environments. The simulated samples uses a simple random sampling schema with three different sample sizes, N: $500, 1500, 2500$. This implies that we expect to observe 10, 30 or 50 individuals in each occupation, respectively.

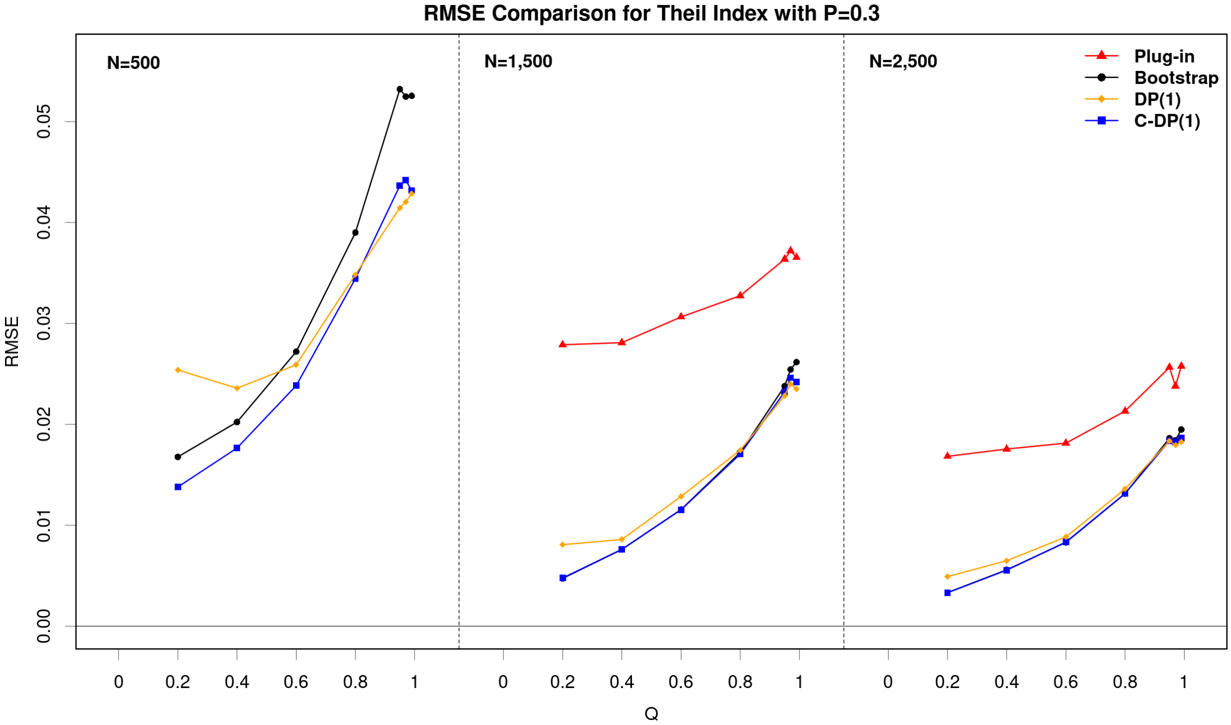
When presenting results below, we consider the Q parameter differently from the N and P parameters. Substantially, we consider the N and P parameters as known, whereas the Q value is considered unknown. Indeed, a practitioner estimating $S(\tau)$ from a sample will certainly know the size of her sample. Moreover, estimating P is substantially easier than estimating a segregation index. Therefore, the practitioner will know P with small uncertainty in comparison to $S(\tau)$. On the contrary, knowing Q or, more in general, the segregation curve of an environment amounts to knowing a segregation index. So,knowing Q would be equivalent to knowing the estimand in this setup. Therefore, we consider Q as unknown at the time of the estimation. This way, our conclusions will be best aligned to practical problems where a practitioner has to choose how to estimate a segregation index. Effectively, this means that we compare the performances of the estimators over 9 different parametrizations, resulting from the combination of the different values for N and P. In all of these 9 cases, we test the estimators on all of the 7 different Q values considered and aggregate the results.

More in detail, for each N-P-Q combination, we simulate 500 sample tables T and test all of the estimation techniques for the the D and Theil indices on each sample. In addition, we also tested the estimators for the Gini (Hutchens, 1991) and Atkinon indices (James and Taeuber, 1985). These indices do not change the general conclusion from the experiments on the D and Theil indices. For this reason, they are not discussed below, but the curious reader can consult the full results for these indices in the result file. Throughout the experiment, the number of bootstrap simulations and posterior sampling is fixed at $2,000$. The Gibbs samplers is also given a 500-samples burn-in period. All confidence intervals are calculated at a nominal 0.95 confidence level.

1.6.2. Results of the Simulation Experiments

Table 1.3 and Table 1.4 report the performances of the different estimators for the different parametrizations tested. For each N-P combinations, the tables report statics about the bias, root mean squared error (RMSE) and interval coverage of the inferential techniques over the differ values of Q. In section 1.3 above, we noticed that the bias of the plug-in point estimate is influenced by the Q value (see Figure 1.2); the simulations show that the parameter Q influences the behavior of all

Figure 1.5 A comparison of the RMSE of the different point estimates of the Theil index for $P = 0.3$. The y-axis measures the RMSE: ideally, we would like the estimators to follow the 0 line throughout the span of Q. Notice that we could not include the RMSE values for the plug-in estimator in the leftmost sub-plot because they are much higher than all other values. Including these values would have entailed rescaling the entire plot in a way that made the differences between all other estimators substantially unreadable.



other techniques as well. Even if this hardly surprising, it still creates some issues in the evaluation of the estimators' perfomances. For example, Figure 1.5 shows the RMSE values of the point estimators of the Theil index when $P = 0.3$. As the Figure shows, the RMSE of the different estimators follow the same overall pattern as the RMSE of the plug-in estimator. Yet, the rank of two estimators may be unclear. E.g. the C-DP(1) estimator has higher (worst) RMSEs for lower values of Q than the Bootstrap estimator, but it has lower (better) RMSEs for higher values of Q: which point estimator is better in this case? Leaving aside the present simulation settings, this means that an estimator may perform better or worst than another depending on the specific segregation curve in the sampled environment. Unfortunately, the segregation curve of an environment is unknown in practical applications, as discussed above. It follows that there is not sufficient information to always choose the best estimator. The same problem arises when comparing the coverage properties of the confidence intervals.

We report different statics (mean, median and range) for each metric (bias, RMSE, and coverage) because there is no immediate, natural way to evaluate the methods over the range of Q for fixed N and P – in addition, full results from the simulations are available as a file. However, when discussing the results we will refer to the min-max criterion to rank estimators. That is, we rank estimators based on their performances in the worst case scenario. For example, in the leftmost sub-plot of Figure 1.5, the best estimator according to the min-max criterion is DP(1) because it performs best in the most difficult situation, when $Q = 0.99$. Notice that our conclusions would be different if we used the mean as our criterion to judge estimators. The min-max performance of an estimator may be inferred from its range, reported in Table 1.3 and 1.4. We emphasize the min-max criterion since it is a conservative choice: it focuses on the robustness of an estimator in the worst cases. Since we do not know how the sampled environments will look like when these methods are applied to real world problems, the other criterion appear more arbitrary. For example, the mean of the RMSEs implicitly weight every value of Q equally, which is equivalent to supposing that the Q values follow a uniform distribution: this appears an arbitrary and unlikely assumption. Similarly, the median RMSE only shows what happens at a specific value of Q, disregarding the extreme values of the RMSE metric. However, this value may not be relevant in practical applications – where it may be common to encounter the values of Q that are associated with the most difficult (easy) estimation. By construction, the min-max criterion will not select the optimal estimator – which is impossible to know – but will strive to select the least bad estimator.[17] Notice that there are cases where one estimator is consistently better than another for all values of Q. In those cases, there is no question about which estimator is better among the two and we will say that an estimator dominates the other. In particular, we focus on the dominance over the bootstrap estimator, which we consider the current state of the art.

Two patterns immediately appear from the results in Table 1.3 and 1.4. First, it may not be straightforward to choose the best estimator, but the last position of the ranking is crystal-clear: the plug-in estimator is consistently the worst estimator by all metrics. The full results in the attached file show that the plug-in estimator is dominated by all the others in almost every instance. This applies to both the point and the interval estimates. This is not completely surprising (Allen et al., 2015; Rathelot, 2012), but it clearly shows that the current practice of using the plug-in point estimator is not warranted. Second, the DP(1) and C-DP(1) are the best estimators in small samples. In many instances, the bootstrap and the plug in intervals fail pathologically providing zero coverage with their intervals. This is particularly worrisome with the plug-in estimator, which provides a minimum coverage close to 0 even in the most favourable parametrization analyzed here – $N = 2,500$ and $P = 0.3$. In comparison, the Bayesian estimators provide a coverage around 0.8 even in the worst case scenario tested here. Leaving the most difficult case aside, the minimum coverage provided by the DP(1) and C-DP(1) methods are most often close to 0.9 for each N-P combination, making them the safer choices when the user wants to estimate an interval. For point estimates, the situation is less straightforward, but the DP(1) and C-DP(1) tend to dominate the bootstrap estimator when either N or P is low for both the D and the Theil index. To the

[17] Naturally, all of the conclusions are based on the limited tests conducted in the simulation exercise: we do not formally prove that any estimator is optimal from a min-max perspective. Therefore, it is impossible to state with certainty that an estimator will be min-max better than another in practical application based on the results presented. However, this is an intrinsic limitation of the simulation exercise not of the min-max criterion. The same criticism applies to any criterion we may choose to rank estimators.

		Bias				RMSE				Confidence Interval		
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
						N=500, P=0.05						
Plug In	0.295	0.286	0.078 - 0.548		0.298	0.290	0.085 - 0.550		0.015	0.000	0.000 - 0.084	
Bootstrap	0.211	0.190	0.050 - 0.424		0.221	0.200	0.066 - 0.428		0.162	0.054	0.000 - 0.416	
DP(1)	0.027	-0.025	-0.031 - 0.168	1	**0.103**	**0.085**	0.066 - 0.179	1	**0.959**	**0.968**	**0.888 - 0.984**	1
C-DP(1)	**0.023**	**-0.015**	**-0.031 - 0.142**	1	0.104	0.099	**0.063 - 0.165**	1	0.929	0.924	0.862 - 0.974	1
						N=500, P=0.2						
Plug In	0.113	0.089	0.017 - 0.259		0.122	0.098	0.033 - 0.261		0.445	0.526	0.000 - 0.880	
Bootstrap	0.040	0.012	-0.004 - 0.139		0.068	0.051	0.033 - 0.145		0.648	0.836	0.044 - 0.866	
DP(1)	0.005	-0.006	-0.018 - 0.069		**0.052**	**0.047**	**0.033 - 0.078**		**0.958**	**0.958**	**0.936 - 0.988**	1
C-DP(1)	**-0.003**	**-0.002**	**-0.031 - 0.037**		0.054	0.057	0.033 - 0.073		0.908	0.908	0.866 - 0.952	1
						N=500, P=0.3						
Plug In	0.092	0.069	0.014 - 0.220		0.101	0.079	0.031 - 0.222		0.499	0.664	0.000 - 0.916	
Bootstrap	0.031	0.006	-0.003 - 0.114		0.058	0.047	0.031 - 0.120		0.697	0.822	0.104 - 0.884	
DP(1)	**0.002**	-0.008	-0.019 - 0.053		0.046	0.043	**0.031 - 0.063**		**0.950**	**0.948**	**0.930 - 0.984**	1
C-DP(1)	-0.004	**-0.002**	**-0.025 - 0.024**		**0.048**	**0.053**	0.031 - 0.065		0.888	0.892	0.854 - 0.920	1
						N=1,500, P=0.05						
Plug In	0.124	0.099	0.022 - 0.281		0.132	0.107	0.037 - 0.283		0.405	0.400	0.000 - 0.828	
Bootstrap	0.045	0.015	-0.003 - 0.155		0.075	0.052	0.037 - 0.162		0.571	0.760	0.036 - 0.798	
DP(1)	0.008	-0.008	-0.016 - 0.080		**0.056**	**0.053**	0.035 - 0.089		**0.954**	**0.956**	**0.926 - 0.986**	1
C-DP(1)	**0.000**	**-0.002**	**-0.025 - 0.049**		0.058	0.060	**0.036 - 0.077**		0.906	0.904	0.874 - 0.944	1
						N=1,500, P=0.2						
Plug In	0.047	0.030	0.005 - 0.131		0.056	0.040	0.018 - 0.132		0.615	0.832	0.000 - 0.952	
Bootstrap	0.012	0.001	-0.001 - 0.059		0.033	0.031	0.019 - 0.065		0.797	0.886	0.310 - 0.938	
DP(1)	**0.000**	-0.001	**-0.013 - 0.020**		**0.030**	**0.032**	**0.018 - 0.041**		**0.951**	**0.948**	**0.930 - 0.982**	
C-DP(1)	-0.004	**0.000**	-0.021 - 0.003		0.031	0.033	0.019 - 0.044		0.899	0.902	0.848 - 0.944	
						N=1,500, P=0.3						
Plug In	0.037	0.021	0.005 - 0.108		0.046	0.032	0.017 - 0.110		0.656	0.894	0.000 - 0.944	
Bootstrap	0.009	0.000	-0.002 - 0.046		0.028	0.027	0.017 - 0.052		0.827	0.908	0.406 - 0.928	
DP(1)	**-0.001**	**-0.002**	**-0.014 - 0.012**		**0.026**	**0.024**	**0.016 - 0.035**		**0.947**	**0.940**	**0.928 - 0.990**	1
C-DP(1)	-0.004	**-0.002**	-0.019 - 0.000		0.027	0.028	0.017 - 0.040		0.902	0.910	0.836 - 0.944	
						N=2,500, P=0.05						
Plug In	0.083	0.058	0.011 - 0.206		0.092	0.067	0.026 - 0.208		0.513	0.696	0.000 - 0.904	
Bootstrap	0.026	**0.000**	-0.005 - 0.106		0.053	0.042	0.027 - 0.112		0.681	0.832	0.120 - 0.862	
DP(1)	**0.003**	-0.004	-0.018 - 0.050		**0.042**	**0.039**	**0.026 - 0.060**		**0.956**	**0.954**	**0.940 - 0.990**	1
C-DP(1)	-0.003	-0.001	**-0.024 - 0.025**		0.045	0.050	0.027 - 0.063		0.911	0.916	0.872 - 0.936	1
						N=2,500, P=0.2						
Plug In	0.031	0.015	0.004 - 0.093		0.038	0.026	0.014 - 0.094		0.686	0.906	0.000 - 0.940	
Bootstrap	0.007	0.000	-0.003 - 0.039		0.024	0.023	0.015 - 0.044		0.841	0.916	0.454 - 0.924	
DP(1)	**-0.001**	**-0.001**	**-0.012 - 0.008**		**0.022**	**0.022**	**0.014 - 0.032**		**0.950**	**0.940**	**0.930 - 0.994**	1
C-DP(1)	-0.003	-0.003	-0.015 - 0.002		0.024	0.025	0.015 - 0.036		0.909	0.918	0.850 - 0.932	
						N=2,500, P=0.3						
Plug In	0.025	0.014	0.003 - 0.077		0.032	0.023	0.013 - 0.078		0.686	0.906	0.000 - 0.940	
Bootstrap	0.005	**0.000**	-0.001 - 0.030		0.021	0.021	0.013 - 0.036		0.841	0.916	0.454 - 0.924	
DP(1)	**0.000**	0.001	**-0.006 - 0.003**		**0.019**	**0.019**	**0.013 - 0.028**		**0.950**	**0.940**	**0.930 - 0.994**	**1**
C-DP(1)	-0.002	**0.000**	-0.008 - 0.001		0.021	0.021	0.013 - 0.029		**0.909**	**0.918**	0.850 - 0.932	

Table 1.3. Monte Carlo Results for the D index. The ≥ signals whether an estimator dominates the Bootstrap estimator in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

	Bias				RMSE				Confidence Interval			
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
					N=500, P=0.05							
Plug In	0.187	0.208	0.096 - 0.250		0.193	0.212	0.113 - 0.252		0.129	0.000	0.000 - 0.536	
Bootstrap	0.102	0.112	0.054 - 0.139		0.117	0.122	0.086 - 0.143		0.362	0.288	0.044 - 0.752	
DP(1)	0.019	0.020	-0.023 - 0.057	1	0.061	0.062	**0.053 - 0.072**	1	**0.952**	0.960	**0.926 - 0.962**	1
C-DP(1)	**0.009**	**0.006**	**-0.010 - 0.031**	1	**0.058**	**0.057**	0.042 - 0.074	1	0.933	**0.942**	0.882 - 0.974	1
					N=500, P=0.2							
Plug In	0.100	0.109	0.070 - 0.112		0.106	0.114	0.081 - 0.115		0.170	0.032	0.000 - 0.502	
Bootstrap	0.027	0.027	0.019 - 0.038		0.045	0.047	0.027 - 0.059		0.853	0.852	0.770 - 0.956	
DP(1)	0.010	0.012	-0.014 - 0.027		0.037	**0.036**	**0.029 - 0.046**		0.920	**0.924**	0.856 - 0.956	
C-DP(1)	**0.004**	**0.004**	**-0.002 - 0.008**	1	**0.034**	0.037	0.017 - 0.047	1	**0.938**	0.922	**0.902 - 0.986**	
					N=500, P=0.3							
Plug In	0.085	0.088	0.068 - 0.091		0.091	0.090	0.078 - 0.097		0.188	0.144	0.000 - 0.442	
Bootstrap	0.018	0.014	0.005 - 0.030		0.037	0.039	0.017 - 0.053		0.889	0.914	0.768 - 0.994	
DP(1)	0.008	0.012	-0.011 - 0.023		0.034	0.035	**0.024 - 0.043**		0.910	0.930	0.808 - 0.936	
C-DP(1)	**0.003**	**0.003**	**0.001 - 0.007**	1	**0.032**	**0.034**	0.014 - 0.044	1	**0.928**	**0.928**	**0.894 - 0.980**	1
					N=1,500, P=0.05							
Plug In	0.078	0.087	0.047 - 0.095		0.084	0.090	0.061 - 0.097		0.231	0.024	0.000 - 0.664	
Bootstrap	0.027	0.026	0.023 - 0.030		0.041	0.042	0.031 - 0.050		0.786	0.784	0.728 - 0.838	
DP(1)	0.009	0.009	-0.005 - 0.023	1	0.032	**0.029**	**0.023 - 0.042**	1	0.933	**0.940**	0.888 - 0.960	1
C-DP(1)	**0.003**	**0.003**	**0.000 - 0.008**	1	**0.030**	**0.029**	0.016 - 0.043	1	**0.935**	0.926	**0.898 - 0.976**	1
					N=1,500, P=0.2							
Plug In	0.034	0.034	0.031 - 0.036		0.039	0.039	0.035 - 0.044		0.400	0.484	0.000 - 0.736	
Bootstrap	0.005	0.003	0.000 - 0.012		0.019	**0.019**	0.006 - 0.030		0.929	0.934	0.874 - 0.992	
DP(1)	0.004	0.006	-0.002 - 0.009		0.019	**0.019**	**0.010 - 0.027**		0.921	0.928	0.876 - 0.948	
C-DP(1)	**0.001**	**0.001**	**0.000 - 0.002**		**0.018**	**0.019**	0.006 - 0.028		**0.937**	**0.936**	**0.908 - 0.978**	
					N=1,500, P=0.3							
Plug In	0.028	0.028	0.027 - 0.028		0.033	0.033	0.028 - 0.037		0.459	0.624	0.000 - 0.750	
Bootstrap	0.003	0.001	-0.001 - 0.009		0.017	**0.017**	0.005 - 0.026		**0.949**	**0.948**	0.890 - 1.000	
DP(1)	0.003	0.004	-0.002 - 0.007		0.017	**0.017**	**0.008 - 0.024**		0.926	0.932	0.904 - 0.942	
C-DP(1)	**0.000**	**0.000**	**-0.001 - 0.002**		**0.016**	**0.017**	0.005 - 0.025		0.939	0.938	**0.914 - 0.978**	
					N=2,500, P=0.05							
Plug In	0.049	0.053	0.032 - 0.056		0.054	0.057	0.043 - 0.057		0.295	0.158	0.000 - 0.768	
Bootstrap	0.013	0.012	0.007 - 0.021		0.026	0.026	0.013 - 0.037		0.888	0.874	0.814 - 0.972	
DP(1)	0.007	0.006	-0.003 - 0.014		**0.023**	0.021	0.015 - 0.032		0.917	**0.942**	0.842 - 0.952	
C-DP(1)	**0.002**	**0.001**	**-0.001 - 0.004**	1	0.022	**0.022**	**0.009 - 0.033**	1	**0.943**	0.932	**0.918 - 0.978**	
					N=2,500, P=0.2							
Plug In	0.020	0.020	0.019 - 0.023		0.025	0.024	0.020 - 0.030		0.519	0.728	0.000 - 0.804	
Bootstrap	0.002	0.001	-0.001 - 0.008		0.014	**0.014**	0.004 - 0.022		**0.947**	0.936	0.894 - 1.000	
DP(1)	0.003	0.003	0.001 - 0.005		0.014	**0.014**	**0.006 - 0.021**		0.937	0.940	**0.926 - 0.946**	1
C-DP(1)	**0.000**	**0.000**	**-0.002 - 0.003**		**0.013**	**0.014**	**0.004 - 0.021**		0.943	**0.940**	0.918 - 0.976	1
					N=2,500, P=0.3							
Plug In	0.017	0.017	0.015 - 0.018		0.021	0.021	0.017 - 0.026		0.565	0.742	0.000 - 0.858	
Bootstrap	0.001	**0.000**	0.000 - 0.005		**0.012**	**0.013**	0.003 - 0.019		0.959	0.962	0.920 - 0.996	
DP(1)	0.002	0.003	-0.001 - 0.004		0.013	0.014	**0.005 - 0.018**		0.937	0.938	0.918 - 0.950	1
C-DP(1)	**0.000**	**0.000**	**-0.002 - 0.002**		**0.012**	**0.013**	0.003 - 0.019		**0.947**	**0.942**	**0.924 - 0.976**	

Table 1.4. Monte Carlo Results for the Theil index. The ≥ signals whether an estimator dominates the Bootstrap estimator in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

extent that it is possible to compare the present tests with previous benchmarks, it appears that the performances of the Bayesian inferential techniques are superior or very close to the performances of other techniques specifically-tuned for one or few indices (Allen et al., 2015; D'Haultfœuille and Rathelot, 2017).

Focusing on the specific indices, it appears that the DP(1) is the best method for the D index, where it generally fares better than the other estimators both in terms of point and interval estimates. In terms of RMSE, the DP(1) and C-DP(1) point estimate will generally be close, with the DP(1) estimates being slightly better in most cases. The bootstrap point estimates eventually catch up to the Bayesian estimates the for larger values of N or P. The difference between interval estimates is more pronounced. The DP(1) interval estimates never fall behind 0.92 coverage, showing over-coverage in some cases. This interval estimator consistently dominates the bootstrap estimator, which comes very short of nominal coverage even in the most favourable situations: in large sample sizes the bootstrap interval will provide a 0.45 coverage for the lowest values of Q. The C-DP(1) provides much better coverage than the bootstrap estimator – at worst, around 0.85 – but it still falls substantially short of the nominal 0.95 coverage.

In the Theil index, the C-DP(1) appear overall the best method. In fact, the C-DP(1) method consistently provides the best coverage among the interval estimates and the RMSE of its point estimates appear very close to the RMSE of the DP(1) estimates, which is the best point estimator. Specifically, the C-DP(1) provides the best bias correction, but its bias correction comes at the cost of a higher variance, making the RMSE of the C-DP(1) not always better than the RMSE of the bootstrap or DP(1) estimator. However, as an overall (i.e. point and interval) estimator, the C-DP(1) method appears to perform best on the Theil index. On the other hand, in applications that only need a point estimate, the DP(1) estimator may be marginally better, especially for smaller samples.

Overall, the new Bayesian methods appear to perform better than their best alternative, the bootstrap estimator. This is especially relevant for small sample sizes (or lower proportion minority). Indeed, notice that the performances of all point estimators grow closer for larger N or higher P: the rounded-up performance metrics of the bootstrap, DP(1), and C-DP(1) estimators are often very close in the largest samples. For example, compare the behavior of the C-DP(1) and bootstrap estimators the rightmost sub-plots of Figure 1.5: these methods are barely distinguishable in this large sample. Notice, again, that this does not apply to the plug-in estimator (the current standard), which will still lag behind all other estimators by some distance. Therefore, while the Bayesian methods appear to ameliorate the bootstrap point estimates in the great majority of cases, their progress is especially important in relatively small samples. This is also the case for the interval estimates: in smaller samples, the new methods provide ways to conduct interval estimates whereas the previous methods fail pathologically. Naturally, the question becomes what a relatively small sample is in real data. In fact, the definition of what counts as a "small sample" for segregation may be deceptive. For example, even large federal surveys such as the American Community Survey may be considered small for the purpose of estimating segregation (Logan et al., 2018; Reardon et al., 2018). We provide an example in the next section.

1.7. Inference for Functions of Indices. An Application to Real Data

This section pursues two goals. First, it studies the performances of the estimators in realistic data to gauge at what point a sample becomes "big" enough that all estimators are substantially equivalent. Second, it showcases the more general application of the previous techniques to functions of segregation indices. Specifically, we will conduct inference on the decomposition of the Theil index (Mora and Ruiz-Castillo, 2011; Frankel and Volij, 2011), which has become increasingly popular in recent applications (Ferguson and Koning, 2018; Fiel, 2013; Lichter et al., 2015).

The experiments in this section will start from real data: the complete population count from the 1940 Census in Chicago.[18] Following the same strategy as Logan et al. (2018), we will simulate simple random sampling of individuals from this data. In the samples, we will estimate racial segregation at tract level using the Theil index. As our racial groups, we will consider the black group as opposed to all others groups combined[19] . Moreover, we will divide Chicago into two parts: South Side and North Side[20]. Historically, southern Chicago has been a segregated part of the city with a high proportion of residents from racial or ethnic minorities (Sampson, 2012). Based on this partition of the city, we will decompose total segregation as the sum of segregation within and between the two parts. Importantly, this is an exercise to showcase how to conduct inference on the increasingly-popular Theil decomposition; this is not a substantial analysis of segregation patterns in 1940 Chicago – which should be based on a more thoughtful partition of the city.

More precisely, we consider only residential tracts: we select those tracts with more than 10 available housing units (occupied or not). This selection results in a total of 901 tracts out of the 935 total tracts. Therefore, our environment is a matrix with 2 rows and 901 columns. The environment has a total population of 3,396,068 individuals; the mean population per tract is 3,769 and it is highly variable, with a standard deviation of 2,833. Among this population, 8.1% of the individuals belong to the black racial group. Unsurprisingly, the environment is highly segregated: the Theil index is as high as 0.86. As a comparison, the D index for this environment is similarly very high, at 0.94. Based on the north-south partition mentioned above, we can decompose the total Theil index in a between component and a within component. The between component assesses the segregation between the northern and southern sides; the within component quantifies segregation

[18] Logan et al. (2018) made the data publicly available from the website: https://s4.ad.brown.edu/Projects/mapusa/index.html

[19] More than 99% of the non-black population is simply white. Therefore, considering white/black segregation instead of non-black/black segregation does not change any methodological or substantial conclusions.

[20] Any tract south of the South Branch of the Chicago River is considered South Side.

within the two parts of the city. The within component turns out to be 0.72 (83.3% of the total); the between component turns out to be 0.13 (16.7% of the total).

To test the different inferential techniques on this data, we repeatedly sample the 1940 Chicago environment with a simple random sample schema.[21] To compare performances at different sample sizes, we use samples of size 5,000, 10,000, and 15,000. These sample sizes result in an expected number of individuals sampled from each tract of 5.5, 11, and 16.5 respectively. As a comparison, the American Community Survey administered 65,000 interviews in the same Cook County area in 2019, but meanwhile the population of the area grew to roughly 5,275,000 individuals. We sample the environment 500 times per sample size, for a total of 1,500 samples. For each sample, we conduct inference both on the overall Theil index and on its between/within components using 1,000 bootstrap/posterior samples. For each sample, we produce all 4 point estimates – plug-in, bootstrap, DP(C) and C-DP(C) – but only 3 confidence intervals. Indeed, we exclude the plug-in (delta method) interval because it is immediately clear from the tests in section 1.6 as well Figure 1.6 that it will dramatically under-cover in these samples, while it is considerably more laborious to obtain than the other (better) interval estimates. For this reasons, we do not consider this interval estimator as a viable option and we only compare the other estimators. In all tests, intervals are calculated at the 0.95 coverage level.

Table 1.5 shows the RMSEs of the different point estimates,[22] while Figure 1.6 shows their sampling distribution. The table shows quite clearly that DP(1) and C-DP(1) are the best point estimator for the overall Theil index. However, it is surprisingly to notice that the plug-in estimator

[21] More precisely, we first sample one individual (uniformly at random) from each tract. Then, we complete the sample with a simple random sample of all remaining individuals from any tract. This procedure ensures that the sample table has the same dimension as the environment table, but it oversamples slightly from the least populated tracts. To keep the exercise simple, we did not re-weight the final samples.

[22] Notice that the RMSE of the Theil index is not the sum of the RMSEs of its components. The RMSE is not decomposable in this sense.

	N = 5,000						N = 10,000						N = 15,000					
	Total		Between		Within		Total		Between		Within		Total		Between		Within	
	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.	RMSE	Prop.
Plug-in	0.051	1	0.011	1	0.05	1	0.034	1	0.009	1	0.035	1	0.026	1	0.007	1	0.027	1
Bootstrap	0.035	0.695	0.011	0.999	0.035	0.703	0.022	0.638	0.009	1.001	0.023	0.65	0.015	0.595	0.007	1.001	0.016	0.617
DP(1)	0.019	0.368	0.015	1.285	**0.015**	0.305	0.013	0.368	0.007	0.846	**0.01**	0.294	0.01	0.388	0.005	0.778	**0.009**	0.319
C-DP(1)	**0.016**	0.322	**0.009**	0.766	**0.015**	0.305	**0.011**	0.322	**0.005**	0.611	0.011	0.305	**0.009**	0.342	**0.004**	0.599	**0.009**	0.33

Table 1.5. RMSE of the point estimates for different sample sizes in the 1940 Chicago experiment. For each sample size, the table shows the RMSE for the overall Theil index (*Total*) and its estimate between and within components. The *Prop.* columns show the RMSE of an estimator normalized by the RMSE of the corresponding plug-in estimator. Bolded values show the best RMSE in each situation.

performs better than DP(1) and as well as the bootstrap in the estimation of the between component. At the same time, it is also worth noticing that the RMSE for this component is quite small compared to the RMSE for the within component, entailing that the performances of all estimators are quite close in absolute terms. The usual hierarchy among estimators is re-established in the estimation of the within component. Overall, the C-DP(1) appears to be the best estimator in most cases considered here, confirming the conclusions from the simulations in section 1.6. If we ignore the between component, the RMSEs of the C-DP(1) estimator are consistently around 35% of the RMSEs of the plug-in estimator – compare this result with the bootstrap estimator, which achieves 60% of the plug-in RMSEs.

Figure 1.6 Sampling distribution of the plug-in, bootstrap, DP(1) and C-DP(1) point estimates for the overall Theil index and its decomposition from the 1940 Chicago data. The horizontal lines represents the true value in the population. The bars represent the highest density 0.95 region in the sampling distribution – the shortest interval containing 95% of the sampling distribution. All the plots on the same scale. Notice that these bars do *not* represent confidence intervals or any other sort of interval estimates.

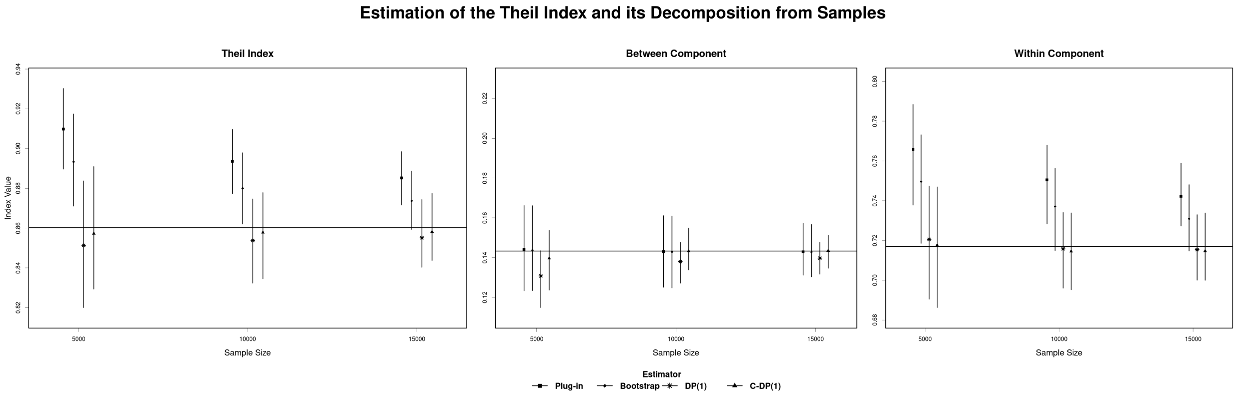


Figure 1.7 Confidence interval coverage of the bootstrap, DP(1) and C-DP(1) interval estimates for the overall Theil index and its decomposition from the 1940 Chicago data. The horizontal lines represents the nominal coverage level of 0.95. All the plots on the same scale.

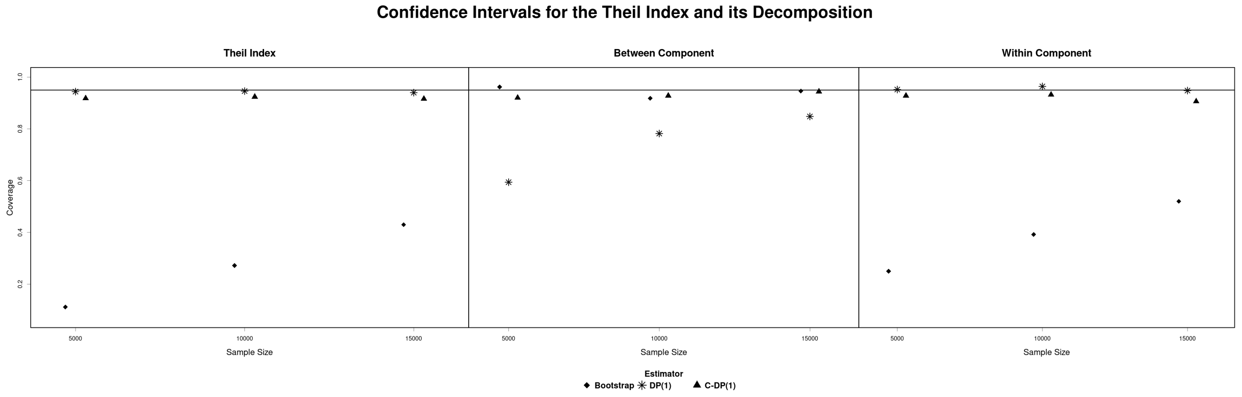


Figure 1.6 sheds further light on the estimators' behavior. For the overall Theil index, the plot shows that the bootstrap and plug-in estimators substantially fail to include the true value of the index in their distribution. The reason is that these estimators fail to estimate the within component of the index, which the Bayesian estimators can reliably get. As for the between component, all estimators appear to be close to the true value for all sample sizes – with the noticeable exception of the DP(1) estimator, who consistently show a small, but clear negative bias in this case. It is also surprising to notice that the C-DP(1) has shorter highest density intervals than the plug-in and bootstrap estimators for the between component, which is not the case generally for the other two statics estimated.

The coverage performances of the confidence intervals follow the same general pattern. The bootstrap estimator generally fails to provide nominal coverage for the overall Theil index, even for the largest sample size. Once again, this is due to the bootstrap failure to estimate the within component of the index, which has clear under-coverage issue with this estimator. On the contrary, the Bayesian estimators provide close-to-nominal coverage for the overall index and within component in all sample size. Yet, the between component shows another hierarchy among estimators, once again. For this component, the bootstrap and C-DP(1) estimators provide a close-to-nominal coverage, whereas the DP(1) estimator shows under-coverage issues. This means that, in general, the C-DP(1) is the best interval estimator because it remains close to the nominal coverage for all statistics. However, the C-DP(1) estimator still shows some minor issues. First, while its intervals are close to the nominal level, they show a small under-coverage in all cases: the provided coverage is always around 0.92. Perhaps more worryingly, the coverage does not ameliorate as the sample size augments. This is not completely surprising. Whereas the C-DP(1) point estimator is consistent, there is no theoretical guarantee that the asymptotic coverage of this interval estimate is 0.95; at the very least, we are not aware of such guarantee. This means that in applications with very big sample sizes, the user may be better served to use the bootstrap estimator – which should have converged to its theoretically-guaranteed coverage. However, this is likely a minor concern because any point estimate will be very precise in such large samples and the need for a confidence interval will be limited. In smaller samples, all of our tests show that the C-DP(1) – as well as the DP(1) – interval estimates are considerably more reliable than the bootstrap or delta confidence intervals. The present results are no exception.

Overall, this section showcases the use of statistical inference for index decomposition. To the best of our knowledge, this paper is the first to formulate and test inferential methods for this application. The experiments with real data confirm the conclusions we drew from the simulations in section 1.6: the C-DP(1) estimator is the most reliable estimator for the Theil index. The tests show that the Bayesian estimators are more precise than the plug-in and boostrap estimators in the decomposition of the Theil index – both in their point and in their interval estimates. More subtly, the results show that the sampling error is not equal among different components. In this case, most estimators could get a precise estimate of the between component, but the plug-in and bootstrap estimators over-estimated the within component (see Figure 1.6). As a result, the plug-in and bootstrap estimators under-estimate the proportion of segregation due to the between component. Most often, the between proportion and its trajectory over time are precisely the statistic of interest in empirical analyses (see for example Lichter et al., 2015). It is important to notice that the C-DP(1) estimator provides result more reliable in this sense as well.[23]

[23] In an additional analysis, we compared the sampling distributions and coverage of the different estimators for the ratio of the between component over the overall Theil index. As expected, C-DP(1) provides a better estimate and coverage, whereas the other estimators tend to under-estimate the between proportion of total segregation. Perhaps more surprisingly, the sampling distributions of the Bayesian estimators were more concentrated than the sampling distributions of the other samples. Therefore, the Bayesian estimators are not only less biased but also less variant in this case.

This section also leverages real data to assess the effect of sample size in realistic settings. From this perspective, it is interesting to notice that all point estimators appear to converge to the true value at the same rate: while all estimators get better as the sample size augments, their RMSEs' proportions are approximately stable across different sizes, see Table 1.5. This behavior differs from what we observed in section 1.6, where the bootstrap estimator eventually catches up with the Bayesian techniques in larger samples (see Figure 1.5). However, we believe that the sample sizes tested in the current experiments are small compared to the sizes tested above. In this section, we sample at most 16.5 individuals per tract in a population with a small minority proportion, whereas we sampled as many as 50 individuals per occupation above. Certainly, other factors such as the shape of the segregation curve or the distribution of the population over units (π_O) will affect the reliability of the estimates. Still, the two most glaring indicators of sample size – namely, size per unit and minority proportion – suggest that the effective sample sizes used in this experiment are small compared to the sample sizes above. Naturally, the reader will not fail to notice that the sample sizes in this section (from 5,000 to 15,000) are nominally larger than the sample sizes in the previous section (from 500 to 2,500), while effectively being smaller. Thus a question is natural: how do we measure sample size for this inferential task?

1.8. Practical Recommendations

After the the tests with artificial and real data, we can provide some guidance to the practitioners who wish to use segregation indices in empirical research. As discussed in section 1.7 the reliability of an index estimation will crucially depend on the number of occupations (units) considered and the proportion of minority individual in the sample. We can merge these two metrics into one: the most reliable metrics to judge sample size for this inferential task is the number of minority individuals per unit (MpU). When considering multi-group occupation, one may conservatively consider the smallest group in the calculation of MpU. Quite clearly, this is a crude metric. The final performance of an estimator will depend on other factors not considered by MpU. Case in point, the index used to quantify inequality appears to be relevant since, for example, the Theil index appears easier to estimate than the D index – as evident by comparing Table 1.3 with Table 1.4. However, one can use the MpU metrics to establish some rules of thumb.

When is the current practice of using the plug-in estimator warranted? For example, one may wonder whether large-scale, nation-wide surveys are large enough to make the plug-in estimates reliable and scarcely variant. When that is the case, the alternative inferential techniques examined in this paper are scarcely relevant. Based on the test above, one may argue that this actually hardly ever happen. For example, the largest MpU in the tests above is 15, which is achieved in section 1.6 when we test with a sample size of 2,500 individuals with 30% minority. Even in this case, the plug-in point estimator is considerably worst than all of the others. As a comparison, if we used the same tracts as in 1940 (see section 1.7), the 2019 American Community Survey would have a MpU of roughly 16.5 in Cook County for the black population, which makes up 20% of the entire population. Considering the patterns observed in all of the tests, it is unlikely that a 16.5 MpU justifies the use of the plug-in estimator: we expect the plug-in estimator to still under-perform in this situation with respect to the other estimators. For example, both Logan et al. (2018) and Reardon et al. (2018) conclude that the American Community Survey plug-in estimate need some form of correction for the H^R index of income segregation.

Indeed, the fundamental practical implication of this paper is that the current practice of using of plug-in point estimator without any interval estimator should be discouraged. Most analyses customarily use hundreds (Beller, 1984; Bianchi and Rytina, 1986; Weeden, 2004; Levanon and Grusky, 2016) if not thousands (Martin-Caughey, 2021) of units: the MpU metrics will hardly ever exceed 20. In these situations, the plug-in estimator is upwardly biased and may be grossly inaccurate, which is well understood at this point (Cortese et al., 1976; Allen et al., 2015; Gentzkow et al., 2019). Considering this issue, it is even more surprising that it is acceptable to report the plug-in point estimate of a segregation index without any assessment of uncertainty (see for example Martin-Caughey, 2021, Figure 4, but the practice is widespread). All of the alternative estimators amend these issues, at least partially. The proposed point estimators are undoubtedly more reliable than the plug-in estimator, while the interval estimators are generally applicable to most situations encountered in empirical analyses and appear to perform well overall. Furthermore, their computational cost is usually very low. The conclusion is that the plug-in estimator should not be used in the great majority of empirical applications and should certainly not be the first option considered.

What should then be the default estimator? Based on the experiments above, one may argue that the original Bayesian techniques formulated in this paper are the best (general) option for inference on segregation indices and their functions. Among the two Bayesian techniques proposed, we prefer the C-DP(C) method because of its higher degree of robustness to mispecification of the C parameter – see Appendix C. While this estimator does not always come on top in the all of the tests, its performances are consistently very close to the best. From what we can gauge, the only downsides of the C-DP(C) estimator are the difficult implementation of a Gibbs sampler – see Appendix B – and its computational cost. As for the implementation, we distribute our code in a publicly available `R` package, `SISeg`. This package allows the user to calculate the DP(C) and C-DP(C) estimator (as well as the bootstrap, plug-in and further estimators) with one line of code. As for the second downsides, we argue that, indeed, there are few cases where the use of the faster bootstrap estimator is well-grounded.

In general, the Bayesian estimators perform better than the others in small samples and in interval estimation. In our tests, the bootstrap point estimator provided performances close to the Bayesian estimators in larger samples – that is for McU greater or equal to 15. In these cases, the bootstrap point estimate and the Bayesian point estimates were substantially very close. Since the McU does not consider all of the factors influencing inference, we can be conservative and use the following conservative rule of thumb: if McU is greater of equal than 20 *and* the practitioner is not concerned with interval estimation, then one may prefer the bootstrap estimator over the Bayesian estimators, purely for computational expediency. This may be the case, for example, if the practitioner uses a segregation index as a variable in a statistical models (see for example Cutler and Glaeser, 1997; Quillian, 2014) estimated from a large sample. In this case, she will likely not be concerned with interval estimation and may as well use the bootstrap estimator if the sample is large enough. In all other cases, our recommendation is to create both a point and interval estimate using the C-DP(1) estimator, which proved to be reliable in all of the cases tested here.

1.9. Conclusions

In this paper, we examined the task of conducting statistical inference for segregation indices and their function. We sought to achieve generality: the proposed inferential techniques should be applicable to any segregation index and their functions. For this general task, we formulated four different estimators: plug-in (the current standard), bootstrap, DP(C) and C-DP(C). The latter two estimators are formulated here for the first time and draw upon Bayesian non-parametric models.

All of the considered estimators are consistent: they will converge to the true value in the population in large (technically, infinite) samples. However, we expect the plug-in estimator to be upwardly biased. This expectation is based on all other analysis about this matter (see for example Cortese et al., 1976; Winship, 1977; Carrington and Troske, 1997; Rathelot, 2012; Allen et al., 2015; D'Haultfœuille and Rathelot, 2017) and, more formally, on the mathematical arguments of Appendix D. In few words, this entails that segregation indices as measured from the sample – the plug-in estimator – are likely higher than segregation in the population. The bias can be very severe in smaller samples and it can lead to wrong conclusions, as shown in Figure 1.2.

This has important consequences for applied research. For example, when comparing two samples, the smaller sample will likely show higher segregation even if the segregation in the two sampled populations is perfectly equal. Similarly, a change in the minority proportion in the population will result in different plug-in estimates, all else being equal. To consider a concrete case, the bias in the plug-in estimator makes it likely to observe a decrease in the segregation of the (increasingly numerous) Latino/a group as measured from samples, even in the case that the actual segregation of this group had not diminished at all. For these reasons, it is important to employ methods that can correct the statistical issues of segregation indices.

To test whether the alternative estimators perform better than the plug-in estimator, we benchmarked the four examined estimators through extensive simulations. We applied the estimators to the popular D and Theil indices and we tested both the accuracy of their point estimates and the coverage of their interval estimates. The tests confirm that the plug-in estimator should generally be avoided. In general, the new Bayesian estimator appear to be more reliable than the others. This is especially the case for small samples and for interval estimation. We reached this conclusion in all of the tests: both with artificial data and with realistic data, both with the estimation of the indices' value (D and Theil) and with the estimation of the decomposition of the Theil index. Therefore, our recommendation is to use the C-DP(1) estimator, unless the practitioner is solely interested in a point estimate (as opposed to an interval estimate) and analyzes a large sample. In this case, the bootstrap estimator may be the most convenient estimator because of its computational speed. However, it is not straightforward to determine whether a sample is "large". As a crude rule of thumb, we propose to define “large” those samples that have at least 20 minority individuals per unit – see section 1.8 for a discussion. In all cases, the plug-in estimator should not be used.

The present work still presents a number of limitations. By its very nature, our analysis based on Monte Carlo simulation is not general and we could not test the effects of some factors that are likely to affect the quality of inference for segregation indices. For example, we did not test systematically how change in the marginal units’ distribution (π_O) affects inference. Second, we did not test how the estimators perform on different kinds of indices, such as multi-group (Reardon and Firebaugh, 2002) or spatial (Reardon and O’Sullivan, 2004) indices.[24] Most importantly, future studies should examine the influence of the C parameter on the Bayesian techniques and possibly propose a way to optimize it – see Appendix C.

Even after acknowledging these limitations, the proposed Bayesian estimators show important improvements over the current state of the art, i.e. the bootstrap estimator – not to mention the current standard practice, i.e. the plug-in estimator. In the tests, the new point estimators proved to ameliorate the bootstrap estimator substantially in small samples, where inference and bias correction is most critical; the new interval estimators showed close-to-nominal coverage in all cases. In other words, the new methods in this paper extend our ability to get reliable estimates for segregation indices (and their functions) to samples where the bootstrap interval estimates fail pathologically. For this reason, these methods can be used to improve our understanding of segregation and its mechanisms – especially, in relatively small samples where such estimation used to be impossible (see for example Bielby and Baron, 1986; Martin-Caughey, 2021). To help other researchers, all of the methods analyzed in the paper will be released in the R package

[24] In a separate analysis, we tested the perfomances of the estimators on the Gini and Atkinson indices. The general conclusions are unchanged.

`SISeg`. When an analysis uses sample data to calculate segregation indices, we believe the methods presented here will be useful if not fundamental.

CHAPTER 2

Interpreting Changes in Segregation: New Principles to Quantify Segregation

2.1. Quantifying Segregation: Theoretical Soundness and Interpretability

Segregation arises in a plethora of different contexts – for example, school (Reardon and Owens, 2014), place of residence (Lichter et al., 2015), work (Tomaskovic-Devey et al., 2006), voluntary associations (McPherson and Smith-Lovin, 1986), or place of worship (Dougherty, 2003) – and between many groups – for example, racial groups (Ferguson and Koning, 2018), different gender (Levanon and Grusky, 2016), and different sexual orientation (Tilcsik et al., 2015). It is a social scientific platitude that segregation in its multifarious forms causes and reproduces inequality, even if the exact mechanisms and the magnitude of their effects are harder to pinpoint (see for example Blau and Kahn, 2016; McCall, 2001; Faber, 2019) . However, to state that "segregation is causing inequality" it is necessary to notice that "environments where segregation is higher are generally more unequal" in some sense of the word "unequal". In turn, the latter statement depends on the ability to quantify segregation in an environment. Therefore, the quantification of segregation is a central task in the many strands of inequality research and has been a central concern of social scientific methodology for a long time (see for example Jahn et al., 1947): this task is pivotal to produce comparative statements such as "The lowest rate of [occupational-race] segregation was found in Riverside–San Bernandino and San Antonio" (Semyonov et al., 2000, p.181) or "[gender] desegregation [in field of study] has been substantial but has stalled for 20 or more years" (England et al., 2020, p. 6991).

For this reason, quantifying segregation has been a long standing methodological and theoretical challenge in the social sciences, as shown by the large number of proposed segregation indices (see for example Jahn et al., 1947; Theil, 1971; James and Taeuber, 1985; Hutchens, 2004). Starting from some data, a segregation index produces a real number describing the amount of segregation in an environment. Naturally, the quantification of segregation is a long standing issue precisely because it is not straightforward to build an index. Indices can follow different defensible approaches (Massey and Denton, 1988) that are often not compatible; not only the different approaches can result in contradictory findings (Karmel and Maclachlan, 1988; Coleman et al., 1982), but they also elicit a deeper debate about the definition of segregation itself.

As a consequence, segregation indices must pursue two goals. First, they should be based on explicit and sound theoretical principles about the meaning of segregation (Frankel and Volij, 2011; Hutchens, 1991, 2004; James and Taeuber, 1985): it is one of the tasks of applied researchers to ponder which index mostly align with their theoretical inclinations. At the same time, a segregation index should provide good interpretatibility – this is the second goal. With interpretability we do not mean that an index should have a formula that is easy to understand. Rather, we mean that the index should provide a clear guide to the different origins of segregation in the data. For example, the well-known index *D* is easy to compute and, in the case of residential segregation, it can be read as "the proportion of nonwhites who would have to change their tract of residence to make the distribution of the minority even throughout the city" (Duncan and Duncan, 1955, p. 211). However, *D* fails to provide good interpretatibility because it is impossible to pinpoint what characteristics of an environment causes *D* to be as large as it results. From this perspective, the Theil index (Mora and Ruiz-Castillo, 2011; Frankel and Volij, 2011) is more interpretable. Given a partition of an environment into smaller parts, the Theil index clearly indicates how much of the total segregation is due to each of the sub-parts (see for example Ferguson and Koning, 2018; Lichter et al., 2015).

Unfortunately, interpretability is particularly difficult when comparing segregation indices across environments. If the environment **A** is less segregated than **B** according to an index, what characteristics of **A** are driving segregation down? As formally discussed below, we can distinguish three kinds of change that influence indices: change in the group margin, change in the unit margin, and association (or structural) change. Thus, the difference in segregation between **A** and **B** may be due to any of these three types of change – and probably to a combination of all three. The literature on indices has extensively discussed how segregation indices should change as the environment changes (see for example Blau and Hendricks, 1979; Charles and Grusky, 1995; Watts, 1998); ultimately, this is the well-known issue of margin influence on segregation indices (Blau and Hendricks, 1979; James and Taeuber, 1985; Charles and Grusky, 1995; Watts, 1998; Jerby et al., 2005; Grusky and Levanon, 2006; Bouchet-Valat, 2022; Elbers, 2021).

Notice that the fulcrum of the issue is not about the theoretical foundations of segregation indices. Depending on "what is contained in the definition of segregation" (Coleman et al., 1982, p.178) according to the researcher, the three different sources of changes may (or not) affect the value of an index. However, even when the researchers eagerly accepts that different types of changes can influence segregation, it is important for analytical and policy reasons to understand what exactly is causing a difference in segregation between two environments. For example, assume that occupational segregation diminished in the US during the last 10 years: we may want to know if this change was due to firms hiring more equally (structural change) or to the demise of the most segregated firms (unit margin change) (Elbers, 2021).

Currently, no segregation index is fully adequate to the task of interpreting the difference of segregation between two environments. The great majority of indices confound at least two different sources of segregation. This means that when comparing two environments with different segregation levels, it is not possible to pinpoint what characteristics of the environments create such difference. Karmel and Maclachlan (1988), Mora and Ruiz-Castillo (2009) and Elbers (2021) proposes decomposition techniques that supposedly obviate the problems. The proposed techniques decompose the difference in segregation into three different components – i.e. the group margin difference, the unit margin difference, and the structural difference. However, these decompositions appear arbitrary because they can attribute to identical structural components disparate amount of segregation, as shown below. On the other hand, margin-free segregation indices (Charles and Grusky, 1995; Bouchet-Valat, 2022) are only sensible to structural change: from this perspective, the comparison of two margin-free indices is interpretable and unambiguous. Yet, these indices are necessarily non-representative of the experience of the population in the environment. Indeed, for any fixed values of these indices, the average individual in the population may personally experience any level of segregation (Elbers, 2021). Unfortunately, this is a counterintuitive property that make margin-free indices theoretically unsuitable for most researchers.

This paper formulates Q: an index whose overall value can be decomposed into a "structural" and a "marginal" component. The former component is attributable to the structural part of an environment, the latter component is an adjustment of the structural component exclusively due to the unit margin – notice that, like many other indices (James and Taeuber, 1985), Q is not sensible to the group margin. We will refer to the decomposition of segregation in the structural and marginal components as the "structural-decomposition". Unlike previous proposed techniques, the structural-decomposition of Q is not arbitrary: it will always attribute the same amount of segregation to an identical structural component. We will say that Q is unambiguously structural-decomposable. The paper shows that Q is by no means the only index with this feature, but it is especially interesting because it is also decomposable in another sense. Like the Theil index, Q can decompose segregation within and between different clusters of units in the environment (Mora and Ruiz-Castillo, 2011). For example, Q can decompose total residential segregation into segregation between different neighborhoods and within them. Decomposition of segregation in this sense will be referred to as "partition-decomposition". Frankel and Volij (2011) and Mora and

Ruiz-Castillo (2011) name this feature "strong unit decomposability" [1], but we prefer partition-decomposition as it emphasizes its difference from the structural-decomposition just introduced. Partition-decomposition is important to assess not only how much, but also where segregation changes (see for example Lichter et al., 2015; Fiel, 2013; Ferguson and Koning, 2018). In a few words, the structural-decomposition decomposes total segregation in the contribution of a structural element plus the contributions of the two marginal distributions. On the other hand, the partition-decomposition decomposes total segregation into segregation within clusters and between them. Using Q, the structural-decomposition can be nested within the partition-decomposfition. To an extent, even the reverse is true: the partition-decomposition can be nested within the structural-decomposition. This allows to check how the structural/marginal difference plays out within each cluster.

Along the way to build Q, the paper provides other theoretical results and methodological insights. On a theoretical level, the paper reframes segregation as a property of bi-variate statistical distribution. Often, segregation has been compared to inequality (James and Taeuber, 1985; Hutchens, 1991; Reardon and Firebaugh, 2002). Yet, the goal of quantifying inequality "is basically that of comparing two frequency distributions $f(y)$ of an attribute y" (Atkinson, 1970, p. 244) Thus, inequality is by definition a property of a univariate distribution. From here, comparing inequality and segregation misses the fundamental bivariate nature of segregation, even if it has certainly paid methodological dividends (James and Taeuber, 1985; Reardon and Firebaugh, 2002).

Based on this framework, the paper formally shows the only way to build margin-free segregation indices (see Appendix F). This result furthers a 40-years long discussion about quantifying

[1] More precisely, Mora and Ruiz-Castillo (2011) and Frankel and Volij (2011) use the expression "strong school decomposability", but this name does not translate well to the more general context of this paper.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	
W	4,027	6,427	4,182	2,701	6,145	106	95	102	1,390	486	3	25,664
M	5,856	5,691	3,570	3,607	2,012	357	3,754	2,367	3,227	3,262	8	33,711
	9,883	12,118	7,752	6,308	8,157	463	3,849	2,469	4,617	3,748	11	59,375

Table 2.1. A table representing the distribution of women (W) and men (M) in 11 different occupational categories in the US in 2003. The data is a representative sample from the Current Population Survey microdata.

segregation discounting any margin influences (Semyonov and Scott, 1983; Karmel and Maclachlan, 1988; Charles and Grusky, 1995; Elbers, 2021; Bouchet-Valat, 2022, see for example). Then, the paper formulates Q^*. Like the Theil and Q index, Q^* can decompose total segregation into a between and a within component. Thus, Q^* is also the first margin-free decomposable index. Beside Q^*, the paper formulates an entire family of new margin-free segregation indices: the family of centered-norm indices, to which Q^* belongs. These indices are based on a general property of bi-variate distributions discussed in Section 2.5.1 that, to the very best of our knowledge, is noticed here for the first time.

Overall, these methodological and theoretical results allow a more complete and robust analysis of trends in segregation. Q has very strong (and unambigous) interpretability. Q^* (and the centered-norm family) provide new methodological tools to researchers interested in quantifying segregation margin-free.

2.2. A Bivariate Framing of Segregation

As a first step to formulate segregation indices with desirable properties, we need to define segregation from a formal perspective. As a first step, we define an environment as a set of individuals belonging to mutually-exclusive groups and having a unit. For example, Table 2.1 shows the cross-tabulation of gender (group) and occupational category (unit) for a representative sample of 59,375 US individuals from 2003. The table shows the basic structure of segregation data: the row represents different groups; the columns represent the units. Moreover, notice the row (column) marginal distribution, which is simply the sum of the cells in each row (column) – that is, $\left(\frac{25{,}664}{59{,}375} = 0.432, \frac{33{,}711}{59{,}375} = 0.567\right)$ in Table 2.1. From a formal perspective, segregation is a property of the environment. More specifically, we can frame segregation as a property of the joint distribution of the unit and group variables in the environment – where "distribution" is to be intended in the statistical sense of the term. Since, we define segregation as a property of a distribution, we will only show normalized examples from now on – that is, tables summing to 1. The reader should consider that the first step in the application of the techniques discussed below is the normalization of the data – which we do not explicitly indicate in the equations.

Notice that the definition of segregation is more general than the structure shown by Table 2.1. First, Table 2.1 only has two groups, there may be arbitrarily many groups (Reardon and Firebaugh, 2002).[2] Even more important, the number of units may be (uncountably) infinite and span multiple dimensions. This is, for example, the case with spatial segregation(Reardon and O'Sullivan, 2004). Technically, Table 2.1 and a continuous map showing the density of the black/white population having the same data structure; Table 2.1 simply has 11 units in 1 dimension whereas the map would have an uncountable number of units (the points in the space) distributed in a 2-dimensional space. In any case, both datasets would show the group and unit of individuals. From this perspective, it is important to notice that we find the same group-unit structure even in less traditional data, such as real-time GPS data from smart phones (Athey et al., 2021) and vocabulary choice for congressional speeches (Gentzkow et al., 2019).[3]

[2] For some of the derivation, we assume that there are more units than groups. This assumption is verified in all practical applications the author has ever encountered.

[3] Admittedly, there are data structure whose segregation is interesting, but whose structure is not as simple as the group and units distribution we consider here. For example, Roberto (2018) incorporates in quantification of residential segregation the physical features of cities, while Echenique and Fryer (2007) and Ballester and Vorsatz (2014) quantify segregation in social networks.

Conceptualizing segregation as a property of the group-unit joint distribution is not wholly original. For example, Reardon and Firebaugh (2002) already mention that segregation can be framed "as the strength of association between nominal variables indexing group and organizational unit membership" (p. 33). However, previous works fail to notice that this framework markedly differentiates segregation from inequality – intended as in the expression "income inequality". For example, Reardon and Firebaugh (2002) also write that segregation is a disproportionality of group representation in the different units and that this "conceptualization links the measurement of segregation to the measurement of inequality" (p. 39), which is also a matter of disproportionality. In fact, the conceptual and methodological links between segregation and inequality date back at least to the work of Duncan and Duncan (1955) and emerge clearly in many fundamental works about segregation and its quantification (James and Taeuber, 1985; Hutchens, 1991, 2004; Mora and Ruiz-Castillo, 2011; Reardon and Firebaugh, 2002). These works unquestionably show that conceptualizing segregation as a kind of inequality is fruitful – at the very least because of the connection with the rich literature about inequality (see for example James and Taeuber, 1985; Hutchens, 1991). At the same time, it is important to acknowledge how segregation and inequality differ. By its very nature, inequality is a univariate issue: it regards the unequal distribution of one good.[4] As defined above, segregation is a property of a bivariate distribution – the joint distribution of the group and unit variables. While this may seem a petty claim, it has important consequences for the quantification of segregation and its properties – see the Appendix F.[5]

[4] This does not mean that inequality has only one source and may not be decomposed in factor (Shorrocks, 1982) or subgroups (Shorrocks, 1980)

[5] The evenness and association approaches precisely quantify segregation as some form of inequality in distribution between two (or more) groups. From this perspective, segregation is a form of inequality. Yet, these approaches are not the only (let alone the "right") approaches to segregation. Especially, for the present discussion it is simpler and more fruitful to separate segregation and inequality neatly.

Before proceeding, it is not possible to discuss the bivariate nature of segregation without properly introducing segregation indices. Segregation indices (or simply indices) are function mapping an environment to a real number, which is often normalized between 0 and 1. For indices, a higher number represents more segregation in an environment. However, this definition provides no guidance about building index. Thus, we can add three conditions indicating how indices should behave:

1: "no segregation exists if [the unit an individual belongs to] is uninfluenced by [group] factors" (Jahn et al., 1947, p.293) – that is, when the group and unit variables are independent, no segregation exists.

2: "complete segregation exists if [the different] groups are situated so that no members of one group reside in [units] in which there are members of the other group[s]" (Jahn et al., 1947, p.293) – when the unit an individual belongs to reveals her group, there is complete segregation.

3: Continuity: small change in an environment will correspond to small changes in the quantification of segregation.

The first two conditions were formulated in the Fourties (see also Williams, 1948), while the latter condition is necessary to prove many theoretical results (see for example Frankel and Volij, 2011) – including the results about margin-free indices given in the Appendix. The great majority of indices meet all of these three conditions.

Besides these common assumptions, existing indices draw upon different conceptualizations of segregation. Massey and Denton (1988) distinguish five ways to measure residential segregation;

	O_1	O_2	
W	0.23	0.07	0.3
M	0.37	0.33	0.7
	0.6	0.4	1

Table 2.2. A table representing the distribution of women (W) and men (M) in 2 different occupational categories.

	O_1	O_2	
W	0.35	0.11	0.46
M	0.29	0.25	0.54
	0.64	0.36	1

Table 2.3. A table representing the distribution of women (W) and men (M) in 2 different occupational categories. Derived from Table 2.2 by multiplying the first row by 1.5 and the second by $0,75$.

two of them (exposure and evenness) are generally applicable beside residential segregation. Reardon and Firebaugh (2002) further propose association as another approach to build indices. Exposure indices frames segregation as the amount of missing inter-group contact in the data (White, 1986); the isolation index is an example of an exposure index (Bell, 1954). On the other hand, evenness indices frames segregation as a difference in the unit-distribution of the groups; the D index (Duncan and Duncan, 1955) is a famous member of this class. Finally, association frames segregation as the distance between the current data and the expected data under the assumption of no dependence between the group and unit variables. For example, the Theil index (Mora and Ruiz-Castillo, 2011) belongs to the class of association indices. From the perspective of this classification, the Q and Q^* indices formulated in Sections 2.5 and 2.6 are association indices.

2.3. Comparing Segregation Indices: The Issue of Interpretability

Substantially, one of the major differences between evenness and the other approaches is the sensibility to change in the marginal distributions, which has been the focus of a long discussion (James and Taeuber, 1985; Charles and Grusky, 1995; Grusky and Levanon, 2006; Watts, 1998;

	O_1	O_2	
W	0.29	0.04	0.33
M	0.46	0.21	0.67
	0.75	0.25	1

Table 2.4. A table representing the distribution of women (W) and men (M) in 2 different occupational categories. Derived from Table 2.2 by multiplying the first column by 1.25 and the second by 0.625.

	O_1	O_2	
W	0.25	0.05	0.3
M	0.35	0.35	0.7
	0.6	0.4	1

Table 2.5. A table representing the distribution of women (W) and men (M) in 2 different occupational categories. Derived from Table 2.2 by changing the odds ratio of the table to 5

Coleman et al., 1982). First, we need to formally define changes between distributions. From this perspective, Blau and Hendricks (1979) noticed as early as 40 years ago that segregation (and indices with it) can be influenced by three kinds of phenomena: 1. a change in the group marginal distribution (for example, more women entering the labor force); 2. change in the unit marginal distribution (for example, more jobs in the service sector); 3. change in the way individuals are recruited (or end up) in each units (for example, change in the way companies hire). The former two kinds of change will be collectively referred to as "marginal" change. The latter change will be referred to as "structural" change.

Consider for example the distribution in Table 2.2, representing a fictional population of women (W) and men (M) in a two-occupations environment. We might multiply both rows of the table by two constants such that the table still sums to 1: this is the first kind of change discussed by Blau and Hendricks (1979). Table 2.3 is derived from Table 2.2 by multiplying the first and second rows by 1.5 and 0.75, respectively. Similarly, Table 2.4 has been obtained from Table 2.2 by multiplying the first column of 2.4 by 1.25 and the second column by 0.625. This is the second kind of transformation described by Blau and Hendricks (1979). Finally, compare Table 2.5 with Table 2.2. In this case, the marginal distributions of groups and units has not changed between the two tables, but probability mass has been moved to create what appears as a more imbalanced and segregated distribution with respect to Table 2.2 (Hutchens, 1991). This third kind of change is the structural change. In fact, Blau and Hendricks (1979) original point can be strengthen. Given an environment **A**, its difference from another environment **B** is a composition of the three changes just described. That is, these three kinds of changes are all we need to transform **A** into any **B**; we will formalize this observation momentarily.

A key point to notice is that the third kind of change appears to be more important and deeper than the others. The first kind of change leaves matter unaltered from the perspective of group members. Indeed, the unit distribution conditional on the group has not changed (up to rounding error) between Table 2.2 and 2.3. For instance, the unit distribution conditional on being in group W is $\left[\frac{0.23}{0.3} = 0.76, \frac{0.07}{0.3} = 0.23\right]$ in Table 2.2 and $\left[\frac{0.35}{0.46} = 0.76, \frac{0.11}{0.46} = 0.23\right]$ in Table 2.3. That is, a woman has 76% probability of being in occupation 1 in both distributions: from her perspective, nothing has changed. On the other hand, the second kind of change leaves matter unaltered from the unit perspective: as is easy to verify, the group distribution conditional on the units are identical in Table 2.2 and 2.4. However, the third kind of change alters the perspective of both group members and units. From this perspective, this is a more structural change. In fact, the marginal distributions of the group and unit variables has not changed from Table 2.2 to 2.5, but Table 2.5 appears to be more segregated nonetheless. As we shall see, this indicates that the association between the unit and group variables has changed. Such association is a core part of our intuition of segregation.

2.3.1. Sensibility to Different Kinds of Change

The empirical interpretations of the three changes outlined above depends on the kind of segregation analyzed – compare the discussion in Grusky and Levanon (2006) and Coleman et al. (1982). However, these changes generally correspond to different social mechanisms modifying an environment. For example, in the case of occupational segregation, the second kinds of change may correspond to organizations augmenting/diminishing their relative importance vis-a-vis other organization while retaining the same exact hiring practices. On the other hand, the third kinds of change can be interpreted as organizations changing their hiring patterns and potential employee adapting to them. Therefore, it is not surprising that the literature has been discussing what kinds of change should (not) influence a proper quantification of segregation. James and Taeuber (1985) and Hutchens (1991) argue that the first type of change should not influence segregation; both works derive this principle from convenient properties of the Lorenz curve. That is, the principle is mostly a consequence of the link between the inequality and the segregation literature – in fact, it is used to establish important theoretical results (Hutchens, 1991, 2004; Frankel and Volij, 2011). On the other hand, Charles and Grusky (1995) and Grusky and Levanon (2006) argue that indices should not be affected by any marginal change. For example, consider an environment where all occupations hire men and women more equally than usual, but where more segregated occupations (say teacher and computer programmer) are preponderant. Arguably, this environment would be less segregated than others, but an index that is sensitive to the second type of change would still quantify the segregation in this environment as high simply because the environment "was dealt a bad hand of cards (i.e., occupations)." (Grusky and Levanon, 2006, p. 560) These authors argue against the use of indices (in favour of parametric models) to quantify segregation. However, if an index is to be used, they defend margin-free indices – that is, indices that are only affected by the third kind of changes. Finally, Coleman et al. (1982) believe that indices should be affected by all three kinds of change. Specifically, they claim that the best way to quantify segregation in their application (school segregation) is by measuring exposure of different groups to each other: this metric will always be affected by the group and unit proportions in the environment.

While this discussion is certainly helpful to represent the different perspectives, it will never be conclusive. If we accept that exposure, evenness, and association are all reasonable approaches to quantify segregation (Massey and Denton, 1988; Reardon and Firebaugh, 2002), then it is impossible to conclude in favour of any argument. Exposure indices will always be affected by marginal change, since the expected exposure of groups to each others depend on both marginal distributions. On the other hand, evenness indices are not sensitive to type 1 change because, by definition, they quantify "the differential distribution of two social groups among" (Massey and Denton, 1988, p.283) units in an environment – naturally, a group distribution will always be normalized and does not depend on group proportion. Finally, association indices may be sensible to both type-1 and type-2 changes or neither of them. For example, the Theil index is sensible to all three changes, whereas margin-free indices are not sensible to marginal changes.

In conclusion, the sensibility of an index to different changes may be appropriate or not depending on the definition of segregation for the case-study, which should considers the empirical implications of each type of change (Coleman et al., 1982). For example, margin-free indices are likely more relevant for occupational segregation (since occupations are the same throughout the US) than for residential segregation (census tracts are not the same in different cities). If units are not the same, it is impossible to be dealt a bad hand of cads, since there is no one deck. However, even in residential segregation, it may be revealing to consider margin-free segregation indices in certain analyses – for instance, the city/suburbs division is ubiquitous in the US (Lichter et al., 2015).

2.3.2. Comparison Interpretability

Even if the decision about the theoretical soundness of the different kinds of indices is context-dependent, it still points to a broadly-general issue: interpretability. Interpretability concerns the ability to understand how segregative patterns change in different environments and how they influence segregation. Notice that interpretability regards the comparison of segregation between different environments.

From this perspective, the three kinds of change outlined above maps onto different empirical phenomena. Thus, a sentence like "environment **A** is less segregated than **B**" immediately spurs substantive questions about the differences between the two environments and how such differences have influenced the assessment of segregation. For example, Lichter et al. (2015) find that overall residential segregation has diminished in the US from 1990 to 2010. As they show using the decomposability properties of the Theil index, this contraction is due to segregation diminishing within-places, but it is countered by a between-places increase in segregation. In this case, the change of segregation is interpreted by partitioning space into mutually-exclusive Census places. However, partitioning is not always an option. Moreover, this kind of analysis does not addresses which kind of change among the three introduced above caused the observed difference in segregation within places. As a matter of fact, if an index is sensible to more than one type of change, it merges different substantive phenomena and makes it difficult to trace back what exactly is causing the change in segregation. This may be a major issues in some analyses.

There are two possible ways to improve the interpretability of a comparison. First, the researcher may use an index that is sensible only to structural change – that is, a margin-free index. To date, A (Charles and Grusky, 1995), R (Charles, 1992), and λ (τ) (Bouchet-Valat, 2022) are the only available margin-free indices. Notice that it would not make sense to use an index that is

	O_1	O_2	O_3
W	0.2	0.10	0.05
M	0.2	0.15	0.30
q	1	0.66	0.16

Table 2.6. A table representing the distribution of women (W) and men (M) in 3 different occupational categories. The bottom line is simply the result of dividing the values in the first row by the values in the bottom row.

only sensible to either one of the marginal changes because structural change is a core part of the intuition about segregation; in fact, such an index does not even exist. Unfortunately, margin-free indices often do not map onto the theoretical intuitions of the researcher because they can grossly misrepresent the experienced segregation of the individuals in an environment – as shown in Section 2.5. Second, decomposition techniques exist to assess the proportion of difference that is due to marginal changes as opposed to structural changes (Karmel and Maclachlan, 1988; Mora and Ruiz-Castillo, 2009; Elbers, 2021).[6] These techniques are generally applicable to any index and might appear as a possible solution to the interpretability problem. However, they are ultimately arbitrary as the decomposition they propose is not consistent. That is, these techniques attribute different amounts of segregation to the same structural components, as it will be apparent after the introduction of a more formal framework.

2.4. Interpreting Change

In this section, we formalize the long-standing discussion reported above. As a first step, it is important to start from a result by Osius (2004). From the present perspective, the result states that most environments can be represented as a triplet of independent elements: the group marginal

[6] Elbers (2021) technique is also concerned with the issue of appearing/disappearing units. That is, the ability to compare environments whose sets of units overlap only partially. This is a well-known issue in the analysis of occupational segregation (Blau et al., 2013), but it is not general across different kinds of segregation – there are no shared units between different cities. Thus, we will ignore the problem in the paper. However, it is easy to apply Elbers (2021) technique to account for (dis)appearing units with the proposed Q index.

	O_1	O_2	O_3
O_1	1	$\frac{3}{2}$	6
O_2	$\frac{2}{4}$	1	4
O_3	$\frac{1}{6}$	$\frac{1}{4}$	1

Table 2.7. A table representing the ϕ-matrix of Table 2.6. The value in each cell is the ratio of different qs from Table 2.6. For example, the value in the first row, second column is the first q (=1) divided by the second (=$\frac{2}{3}$).

	O_1	O_2	O_3
O_1	0	0.40	1.80
O_2	-0.40	0	1.39
O_3	-1.80	-1.39	0

Table 2.8. A table representing the logged ϕ-matrix of Table 2.6. The value in each cell is the natural logarithm of the ratio of different $q$ss from Table 2.6. For example, the value in the first-row, second-column is the logarithm of $\frac{3}{2}$. In turn, $\frac{2}{3}$ results from dividing the first q (=1) by the second (=$\frac{2}{3}$).

distribution, the unit marginal distribution, and an association element. The association element regulates the dependency structure between the group and the unit variables. That is, it is the element responsible to regulate whether the two variables are independent or dependent (and in what way). When the result applies, an environment can be represented as a triplet and different triplets correspond to different environments. The elements of the triplets are independent in the sense that they can freely vary within their domain, which does not depend on any other element in the triplet. This result is generally applicable, unless the distribution has 0s, as discussed below.

Now, what is the association element that enters in the triplet? In the case of discrete distributions on a finite domain – that is, tables – such association element is nothing but a vector of (log-)odds ratios; the same quantity used in logistic regression. For example Table 2.6 shows a fictional distribution of women and men (groups) into three different occupations (units). The last row of the table shows the ratio of women to men in each unit. If we take the ratio of such ratios, we create the odds ratios. Table 2.7 collects *all* the odds ratios in one matrix, which we call the

ϕ-matrix of the distribution. For example, the first-row second-column of the ϕ-matrix shows the result of dividing the first q of Table 2.6 (=1) by the second q (=$\frac{2}{3}$). Symmetrically, the second-row firt-column of the matrix shows the result of dividing the second q (=$\frac{2}{3}$) by the second q (=1).

For tables, Osius result is not new: it has been known since at least the Seventies that a table can be reconstructed from the marginal distributions and the odds ratios (Agresti, 2013, see for example). Yet, Osius (2004) generalizes this result considerably. From the perspective of quantifying segregation, this means that we can apply the indices proposed below to segregation in space (Reardon and O'Sullivan, 2004) or time Athey et al. (2021). Furthermore, the generality of the result allows us to make bolder theoretical statement – see Appendix F.

The result is important because we can re-interpret the different kinds of change discussed in Section 2.3 based on the triplet representation just introduced. Marginal changes will only affect marginal distributions[7], leaving the association element untouched (Elbers, 2021). On the other hand, structural change directly affects the association element of the distribution. Since these kinds of changes affect every element in a triplet, it means that we can use them to transform a distribution into any other distributions – exhausting the possible ways to change a distribution.

Limiting ourselves to environments that are representable as contingency tables[8], we can sum up the previous discussion in the equation:

$$\mathbf{B} = \mathbf{R} \cdot f\big(\phi[\mathbf{A}]\big) \cdot \mathbf{C} \tag{2.1}$$

[7] Notice the two kinds of marginal changes affect both marginal distributions at the same time. That is, the different kinds of marginal changes do not map on changing the margins one at the time.

[8] In the more general case of distributions that are not representable as finite contingency table, Osius (2004) shows how to create a general version of the proportional marginal fitting procedure (Deming and Stephan, 1940). This means that the same reasoning that we apply to simple contingency tables is generalizable to any joint distribution of groups and units (having a density).

where $\mathbf{A}$ is an initial environment, $\mathbf{B}$ is a any othe distribution on the same support, $\phi[\mathbf{A}]$ is the ϕ-matrix for $\mathbf{A}$, $f(\cdot)$ is a function mapping $\phi[\mathbf{A}]$ to another arbitrary ϕ-matrix, $\mathbf{R}$ and $\mathbf{C}$ are properly sized diagonal matrices. Notice that each element of equation (2.1) corresponds to a type of change: $\mathbf{R}$ regulates the first kind of (marginal) change, $\mathbf{C}$ corresponds to the second kind of (marginal) change (Watts, 1998), $f(\cdot)$ is structural change. In fact, the $\mathbf{R}$ and $\mathbf{C}$ matrices represent the operation of iterative marginal fitting, which can arbitrarily transform the margins of a joint distribution (Deming and Stephan, 1940; Fienberg, 1970).

Now suppose we compare the segregation in environment $\mathbf{A}$ and environment $\mathbf{B}$ from equation (2.1). It is natural to wonder how much of the difference in segregation is due to the three different kinds of change. That is, how much is due to $\mathbf{R}$, $\mathbf{C}$, and $f(\cdot)$. An even more straightforward question is how much of the overall difference is due to a. difference in the unit margins, b. difference in the group margin, or c. difference in the association elements. The difference in the two margins are due to $\mathbf{R}$ and $\mathbf{C}$, whereas the difference in the association elements are due to $f(\cdot)$.

To introduce some notation, we indicate with $S(\mathbf{A})$ the segregation in environment $\mathbf{A}$ according to an index $S(\cdot)$. We indicate with $p_{\mathbf{A}}(g)$ and $p_{\mathbf{A}}(u)$ the group and unit distribution of $\mathbf{A}$, respectively. We want to decompose the difference between $S(\mathbf{A})$ and $S(\mathbf{B})$ into three factors: the first one *purely* due to the structural difference $f(\cdot)$, the second one due to $p_{\mathbf{A}}(g)$ and $p_{\mathbf{B}}(g)$, and the third one due to $p_{\mathbf{A}}(u)$ and $p_{\mathbf{B}}(u)$. We will indicate these components with $\Delta_1(\phi[\mathbf{A}], \phi[\mathbf{B}])$, $\Delta_2(\phi[\mathbf{A}], \phi[\mathbf{B}], p_{\mathbf{A}}(g), p_{\mathbf{B}}(g))$, and $\Delta_3(\phi[\mathbf{A}], \phi[\mathbf{B},], p_{\mathbf{A}}(u), p_{\mathbf{B}}(u))$, respectively. The elements in parenthesis show that Δ_2 and Δ_3 depends on the respective marginal distributions, but also on the ϕ-matrices of $\mathbf{A}$ and $\mathbf{B}$ – this will be discussed momentarily.

From here, we can write the objective of creating an interpretable decomposition of the difference as the following exercise in decomposition:

$$\begin{aligned} S(\mathbf{A}) - S(\mathbf{B}) = & \\ & \Delta_1(\phi[\mathbf{A}], \phi[\mathbf{B}]) + \\ & \Delta_2(\phi[\mathbf{A}], \phi[\mathbf{B},], p_{\mathbf{A}}(g), p_{\mathbf{B}}(g)) + \\ & \Delta_3(\phi[\mathbf{A}], \phi[\mathbf{B},], p_{\mathbf{A}}(u), p_{\mathbf{B}}(u)) \end{aligned} \tag{2.2}$$

Thus, the difference in segregation between two distributions is the result of (a) the difference between their structural segregation (Δ_1) plus (b) the difference due the margins alongisde the structural component (Δ_2 and Δ_3).

The core intuition behind equation (2.2) is that segregation and its difference have a primitive structural component modified by the margins. The structural component is solely regulated by the association element – i.e. the ϕ-matrix. On the contrary, marginal distributions of groups and units do not tell us anything about segregation by themselves. For example, the same unit marginal distribution may concentrate individuals in very unbalanced or balanced units. Depending on this, the net unit-marginal contribution to segregation could be positive or negative, respectively. Therefore, it is impossible to discern a marginal contribution without checking the association element. Similarly, the quantification of the change in segregation due to the margins must also depend on the ϕ-matrices of $\mathbf{A}$ and $\mathbf{B}$. The equation follows this intuition since the effects of margins on segregation (Δ_2 and Δ_3) depend on the association structure. For this reason, the ϕ-matrices appear in every component of the decomposition – including Δ_2 and Δ_3.

The final objective of equation (2.2) is the same as the objective of the decompositions proposed by Elbers (2021) (Karmel and Maclachlan, 1988) and Mora and Ruiz-Castillo (2009). Yet, these decomposition techniques fail to satisfy equation (2.2) even if they are based on the same principles. Indeed, these techniques will provide a structural (corresponding to Δ_1) and two marginal components (corresponding to Δ_2 and Δ_3), but the proposed structural components depends on the marginal distributions as well as the ϕ-matrices. In fact, it is easy to verify that the structural component of these decomposition will change if we modify the marginal distribution of either **A** or **B** without altering the association structures of neither. This is clearly an issue, since the declared goal of these techniques is to decompose the difference of segregation in pure components.

A possible way to satisfy equation (2.2) is to use margin-free indices, as suggested for example by Charles and Grusky (1995). By construction, margin-free indices will only have structural segregation: the Δ_2 and Δ_3 will always be 0. However, margin-free indices come with their own issues, discussed below (see also Jerby et al., 2005; Elbers, 2021). As an alternative, we propose S-decomposable indices. We define an index to be S-decomposable if it can be written as the sum of a purely-structural element plus some modification due to the margins. The structural element represents the margin-free amount of segregation: the part of segregation purely due to the association element. The marginal distributions modify structural segregation. In formula:

$$S(\mathbf{A}) = g(\phi[\mathbf{A}]) + h_1(\phi[\mathbf{A}], p_{\mathbf{A}}(g)) + h_2(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) \tag{2.3}$$

where $g(\cdot)$, $h_1(\cdot)$, and $h_2(\cdot)$ are continuous[9] functions, with $g(\cdot)$ being also positive. An S-decomposable index will be interpretable in the sense of equation (2.2). Assume an index is S-decomposable, then it will immediately provide a decomposition of the difference:

$$\begin{aligned} S(\mathbf{A}) - S(\mathbf{B}) = \\ & g(\phi[\mathbf{A}]) - g(\phi[\mathbf{B}]) + \\ & h_1(\phi[\mathbf{A}], p_{\mathbf{A}}(g)) - h_2(\phi[\mathbf{B}], p_{\mathbf{B}}(g)) + \\ & h_2(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) - h_2(\phi[\mathbf{B}], p_{\mathbf{B}}(u)) = \\ & \Delta_1(\phi[\mathbf{A}], \phi[\mathbf{B}]) + \\ & \Delta_2(\phi[\mathbf{A}], \phi[\mathbf{B},], p_{\mathbf{A}}(g), p_{\mathbf{B}}(g)) + \\ & \Delta_3(\phi[\mathbf{A}], \phi[\mathbf{B},], p_{\mathbf{A}}(u), p_{\mathbf{B}}(u)) \end{aligned}$$

Beside the trivial case of pure margin-free indices (where h_1 and h_2 are constantly 0), the only example of an S-decomposable index is the marginal-weighted version of λ, formulated by Bouchet-Valat (2022). However, to maximize interpretability, we formulate the Q index, which is also decomposable in another sense while being easier to calculate and possibly more natural than λ. The first step to formulate a S-decomposable index is to create a margin-free index that can act as a basis for an S-decomposable index.

2.5. Building Margin-Free Indices

The formalization proposed in Section 2.4 allows us to clarify the meaning of margin-free indices. Consider Equation (2.1) again. The standard formulation of the margin-free property is

[9] Continuity of $g(\cdot)$, $h_1(\cdot)$, and $h_2(\cdot)$ immediately follows from the fact that $S(\mathbf{A})$ as a whole must be continuous.

	O_1	O_2	O_3
O_1	0	0	0
O_2	0	0	0
O_3	0	0	0

	O_1	O_2	O_3
O_1	0	∞	0
O_2	$-\infty$	0	∞
O_3	0	$-\infty$	0

Table 2.9. Logged ϕ-matrices for an independent distribution (left) and a completely segregated distribution (right).

that an index must not change for any **R** and **C**: a margin-free index is only sensible to $f(\cdot)$. Using Osius' result, we can reformulate the margin-free property: a margin-free index will assign the same segregation to all distributions having the same ϕ-matrix. In other words, an index is margin-free if and only if it is a function of the association object and only of the association object – see Appendix F. This is not completely surprising; Semyonov et al. (1984), Charles and Grusky (1995), and Elbers (2021) have already indicated the odds ratios as a way to build margin-free indices. However, this result also shows that the association element is the *only* way to build these indices – even beside the contingency table case considered in the literature so far. Therefore, the next step is to study the properties of the association element, that is the ϕ-matrix.

2.5.1. The Structure of the ϕ-Matrix

Rather than considering the ϕ-matrix it is actually more convenient to consider its logged version, the logged ϕ-matrix Quite clearly, the logged ϕ-matrix has a very strong structure, whic is apparent in Table 2.4. First, the logged ϕ-matrix is anti-symmetric: one can find opposite numbers on the opposite side of the diagonal, which in turn only contains zeros. Second, Table 2.9 show the logged ϕ-matrix for an independent distribution and a completely segregated distribution. These matrices show specific structural signatures: in the independent case, all elements are zero; in the completely-segregated case, the matrix only contains zero, infinity, and negative infinity.[10] The

[10] The completely segregated case is possibly even more difficult than Table 2.9 shows since we had to make some arbitrary judgment call for the completion of the table. One may argue that the ϕ-matrix is not well defined in this case

deeper issue is that in this case, Osius' decomposition does not apply: it is impossible to represent a completely segregated distribution as a triplet. Even more alarming, the same issues appear any time there is just one zero in a distribution (Osius, 2004, p. 265) – this is a well known issue for margin-free indices (Jerby et al., 2005; Grusky and Levanon, 2006). Unfortunately, zeros are very common in analyses of segregation, since it is frequent for a unit to only host members from one group. We will refer to this issue as the "zero problem", which we will discuss again in 2.5.5 . Notice that the zero problem prevents margin-free indices from assigning maximum segregation to completely-segregated distributions, since margin-free indices are undefined in such cases. That is, the zero problem effectively prevent margin-free indices to satisfy condition **2** from Section 2.2 and to be proper segregation indices. For the moment, we will assume that distributions contain no zeros.

However, the fundamental observation is that any row (or column) of a ϕ-matrix is sufficient to reconstruct the entire matrix, which indeed contains redundant information. For example, the second row of of Table 2.4 can be reconstructed from the first one by simply adding 0.4 to all cells (up to rounding errors). Similarly, the third row can be reconstructed by subtracting 1.8 to all cells in the first row. This property imply another interesting property of the ϕ-matrix (and association elements more in general) that has not been observed so far: the centered $p-$norm of every row (or column) of the logged ϕ-matrix is constant. In symbols, let $\log(\phi_{\mathbf{A}})[i,j]$ be the first row, j^{th} column of the logged-ϕ-matrix of an environment (and in general any joint distribution), then:

$$\sum_{j=1}^{U}\left|\log(\phi_{\mathbf{A}})[1,j]-\frac{1}{U}\sum_{j=1}^{U}\log(\phi_{\mathbf{A}})[1,j]\right|^{p}=\sum_{j=1}^{U}\left|\log(\phi_{\mathbf{A}})[i,j]-\frac{1}{U}\sum_{j=1}^{U}\log(\phi_{\mathbf{A}})[i,j]\right|^{p}\text{ for i} = 1\dots U \tag{2.4}$$

where U is the number of units in an environment and $p \geq 1$.

and, therefore, it does not exist. Table 2.9 still shows a logged ϕ-matrix in this case for argument's sake and because the shown matrix is in line with the results shown in Section 2.5.5.

	O_1	O_2	O_3
W	.003	.381	.016
M	.007	.114	.479
q	.336	3.358	.034

	O_1	O_2	O_3
W	.354	.043	.003
M	.546	.007	.047
q	.647	6.473	.065

Table 2.10. Two distributions sharing the same association element and group margin, but with different unit margins. The experience of the individuals in the two environments are radically different.

For example, for $p = 2$, equation (2.4) implies that the variance (standard deviation) of every row and column of a logged ϕ-matrix is identical. In the case of Table 2.4, for instance, we have $\frac{1}{U}\sum_{j=1}^{U}\log(\phi_{\mathbf{A}})[1,j] = \frac{1}{3}(0+0.4+1.8) = 0.73$ and the variance of the first row is $\frac{1}{3}[(0-0.73)^2 + (0.4-0.73)^2 + (1.8-0.73)^2 = 0.6$; it is easy to verify that the other rows/columns of the matrix have the same variance. Notice that this property is a consequence of the structure of the logged ϕ-matrix. If a row can be reconstructed from another by adding a constant to every cell, all rows will necessarily have the same variance because variance does not change when you add a constant to all of the observations.

Finally, it is important to notice how equation (2.4) behaves when the distribution is either independent or has a zero. In the former case, any p-norm will be 0 because all of the cell in a row will be 0. In the latter case, the p-norm will get to infinity as a cell in every row (column) will either be infinity or negative infinity. Therefore, as the underlying joint distribution of groups and units get closer to have a 0, the p-norms will get closer to infinity. In fact, this is a promising basis to build margin-free segregation indices.

2.5.2. Q^* and the Centered-Norm Family of Indices

We may use equation (2.4) as the basis to build an entire family of margin-free indices – and we can actually simplify the calculation. First, we indicate with $p_{\mathbf{A},ij}(g,u)$ the value of the probability density (mass) function for group i and unit j in an environment $\mathbf{A}$. For example in Table 2.11, we have $p_{\mathbf{A},12}(g,u) = 0.022$. Then, let us define the unit odds for the environment **A** as:

$$q_j[\mathbf{A}] = \frac{p_{\mathbf{A},1j}(g,u)}{p_{\mathbf{A},2j}(g,u)} \tag{2.5}$$

For legibility, we will omit indicating the environment explicitly in the notation unless it is necessary for clarity. The log odds ratio between the j^{th} and k^{th} column is simply $\log q_j - \log q_k$. It follows that all the elements in the i^{th} row of the ϕ-matrix are in the form $\log q_i - \log q_j$ for j from 1 to U. Now, adding/subtracting a constant to the row of a logged ϕ-matrix will not change the variance and, more in general, the centered p-norm shown in equation (2.4). Therefore, we can simplify the calculation of a p-norm by simply subtracting q_i from the i^{th} row of the logged ϕ-matrix. That is to say, we can simply consider the logged odds instead of the more cumbersome log odds ratios.[11] From here, we define the centered-norm family of indices as follows:

$$\begin{aligned} \bar{q} =& \frac{1}{U}\sum_{i=1}^{U} q_i \\ S_2^*(\mathbf{A},p) =& \frac{1}{U}\sum_{i=1}^{U} \left| q_i - \bar{q} \right|^p \end{aligned} \tag{2.6}$$

The subscript in S_2^* indicates that equation (2.7) only covers the two-groups case; we will generalize to the multi-group case in Section 2.5.4. For the moment, we single out the most important member of this family, which coincides with $p = 2$:

$$\begin{aligned} \bar{q} =& \frac{1}{U}\sum_{i=1}^{U} q_i \\ Q_2^*(\mathbf{A}) =& \frac{1}{U}\sum_{i=1}^{U} \left(q_i - \bar{q} \right)^2 \end{aligned} \tag{2.7}$$

[11] Technically, the odds are odds ratios with respect to a (non-existing) unit having an equal odd. What we are technically doing is to change the reference odd in the calculation of a sufficient vector of odds ratios so as to make the calculations more convenient.

Q_2^* is simply the variance of any row (column) of the ϕ-matrix of an environment $\mathbf{A}$. In other word, Q_2^* is the average squared distance of the log odds from the geometric mean of the odds. From this perspective, it is interesting to notice that the variance of the simple odds (as opposed to the log odds) is not margin-free: substituting the log odds for the odds (and the geometric mean for the simple mean) makes Q_2^* margin-free. Q_2^* is also a close relative of the A index proposed by Charles and Grusky (1995) and λ proposed by Bouchet-Valat (2022).[12]

Any member of the centered-norm family is a margin-free index in the sense that a. it will attribute 0 segregation to an environment where the group and the unit variables are independent, b. it is continuous and c. it is not sensible to any change that does not affect the association element of a distribution. However, these indices still suffer from the zero problem introduced in Section 2.5.1. As every other existing margin-free index, the centered-norm family is not well defined when the distribution presents zeros; this issue is discussed in Section 2.5.5. As a side note, the centered p-norm family is not well defined for environments with less than two units, but this is generally not a concern in real applications.

An important observation about the structure of Q_2^*, the centered-norm family, and, more in general, margin-free indices. Equation (2.7) attributes the same weight of $\frac{1}{U}$ to each unit in the environment. This means that units are not weighted by the amount of probability mass they have, as is usually the case (Reardon and Firebaugh, 2002). While this is natural given that margin-free indices are independent from the unit margin, it also implies that the margin-free unit cannot reflect the experience of the individual in the environment, but they are unit-centric. For example, Table 2.10 shows two distributions having the same ϕ-matrix and, thus, the same Q_2^* index (= 3.53). However, the individuals in the leftmost distribution are for the great majority distributed in units

[12] However, the multi-group version of Q^* differs substantially from the multi-group version of λ. Compare equation (2.10 in Section 2.5.4 with equation (18) from Bouchet-Valat (2022).

	O_1	O_2	O_3	O_4	O_5	O_6
W	.022	.022	.003	.008	.016	.03
M	.128	.128	.197	.192	.184	.07
Q	.169	.169	.017	.042	.084	.422

Table 2.11. An example of a partitioned table. The first part contains the first four units, the second part contains the last two units.

	O_1	O_2	O_3	O_4	O_5	O_6
W	.009	.009	.012	.012	.005	.002
M	.139	.139	.185	.185	.201	.1
q	.067	.067	.067	.067	.024	.024

Table 2.12. A partitioned table having zero within-parts segregation. Within each part, the qs are equal to the geometric mean of the qs within the same part of Table 2.11.

	O_1	O_2	O_3	O_4
W	.031	.031	.005	.012
M	.183	.183	.281	.274
q	.169	.169	.017	.042

	O_5	O_6
W	.052	.099
M	.615	.234
q	.084	.422

Table 2.13. Distribution derived from Table 2.11 conditioning on the parts. For example, the left distribution is a re-normalization of the first four units from Table 2.11 so that it sums to 1.

with a severe under-representation of the other groups. On the contrary, in the right distribution, the great majority of individuals belong to a unit whose group proportion is fairly close to the environment's overall group proportion. This means that the experience of the individuals is widely different in the environments – as shown by more traditional, non-margin-free indices such as D, whose values for the two distributions are 0.8 and 0.1, respectively. While this is a natural consequence of the property of being margin-free, it may not correspond to the theoretical intuition of a researcher (see for example Elbers, 2021).

2.5.3. Partition-Decomposition of Q_2^*

Not only Q^* is a familiar quantity (a variance!), but Q^* is partition-decomposable precisely because it is a variance. Consider a partition of the units in an environment into mutually exclusive clusters. For instance, Table 2.11 shows a distribution decomposed into two clusters – $\mathbf{P}_1$ and $\mathbf{P}_2$. Then, we can decompose the total segregation as quantified by Q^* in the following way:

$$Q_2^*(\mathbf{A}) = \sum_{i=1}^{2} \frac{|P_i|}{6} Q_2^*(\mathbf{P}_i) + Q_2^*(\bar{\mathbf{A}})$$

In this case, $\mathbf{P_1}$ and $\mathbf{P_2}$ are the (renormalized) clusters, shown in Table 2.13, while $\bar{\mathbf{A}}$ is a version of $\mathbf{A}$ having the odds within the clusters fixed at the clusters' geometric-average odds in $\mathbf{A}$, as shown in Table 2.12. This ensures that there is no within-cluster segregation in $\bar{\mathbf{A}}$, since the variance of a vector repeating the same element is 0 by construction.

More in general, let $\mathbf{P_1}, \mathbf{P_2} \dots \mathbf{P_{|P|}}$ be the partition of the units in an environment $\mathbf{A}$.[13] Then, we can write Q^* as

$$Q_2^*(\mathbf{A}) = \sum_{i=1}^{|P|} \frac{|P_i|}{U} Q_2^*(\mathbf{P}_i) + Q_2^*(\bar{\mathbf{A}}) \tag{2.8}$$

where $|P_i|$ is simply the number of units in a cluster. The distribution $\bar{\mathbf{A}}$ is constructed in such a way as to have as qs the geometric mean of the qs of $\mathbf{A}$ in the same cluster. Equation (2.8) represents the partition-decomposition of $Q_2^*(\mathbf{A})$. As Mora and Ruiz-Castillo (2011) exhaustively explains, this decomposition has intuitive meaning. The various $Q_2^*(\mathbf{P}_i)$ represents the amount by which overall segregation falls if the segregation within cluster i is eliminated; the sum of the different $Q_2^*(\mathbf{P}_i)$ represents the reduction in overall segregation that would arise if the segregation within

[13] For the partition-decomposition to work, the clusters within the unit partition must contain at least two units. This is a condition for the correct decomposition of both Q and Q^*

all cluster were eliminated; $Q_2^*(\bar{\mathbf{A}})$ represents the amount of segregation that can be attributed to the difference in group proportions across clusters (Mora and Ruiz-Castillo, 2011, p. 164-169). For this reason, the first sum in equation (2.8) is usually termed "within-segregation" while the last addend is usually termed "between segregation". Like structural-decomposition, partition-decomposition can help in the interpretation of differences (see for example Lichter et al., 2015; Ferguson and Koning, 2018). To the best of our knowledge, Q_2^* is the first margin-free index to be partition-decomposable, which results from the ANOVA decomposition of variance and, more in general, the law of total variance.

2.5.4. Extension to Multi-groups Setting

While we have introduced the centered-norm family for two-groups environment, we need to generalize it to a multi-group settings – that is, environments with more than 2 groups (Reardon and Firebaugh, 2002). The objective is to create a function that maintains the desirable properties of centered-norm indices – Q_2^* in particular – while being well defined in environments with multiple groups. Even if Osius theorem still applies to this more general case, the issue is that the ϕ-matrix is an object describing the association element for two groups.[14] Indeed, we have a different ϕ-matrix for every possible combination of groups. Thus, the solution is to consider each possible ϕ-matrix separately and calculate the overall segregation in an environment as the sum of S_2 applied to all possible ϕ-matrix. Intuitively, the total segregation of an environment is the sum of the segregation between each couple of groups. Say we have a total of G groups, we can write the

[14] The centered norms of all multi-group log odds-ratios taken together is not a constant, as it is in the two-groups case. However, the centered-norm of the log odds ratios for any two groups is still a constant.

	O_1	O_2	O_3	O_4
W	.1	.0	.3	.0
M	.0	.4	.0	.2
q	∞	.0	∞	.0
log q	∞	$-\infty$	∞	$-\infty$

Table 2.14. A completely-segregated distribution. The last two rows show the odds and the logged odds for each unit.

general family of center-norm indices:

$$S^*(\mathbf{A}, p) = \frac{1}{\binom{G}{2}} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} S_2^*(\mathbf{A}[i-j], p) \tag{2.9}$$

where $\mathbf{A}[i-j]$ is a re-normalized version of $\mathbf{A}$ only containing the i^{th} and j^{th} groups. Equation (2.9) is simply the sum of 2-groups indices per each couple of groups, normalized by the number of couples. The multi-group definition of Q^* immediately follows:

$$Q^*(\mathbf{A}) = \frac{1}{\binom{G}{2}} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} Q_2^*(\mathbf{A}[i-j]) \tag{2.10}$$

We notice that the general version of Q^* retains the partition-decomposability of its 2-groups version:

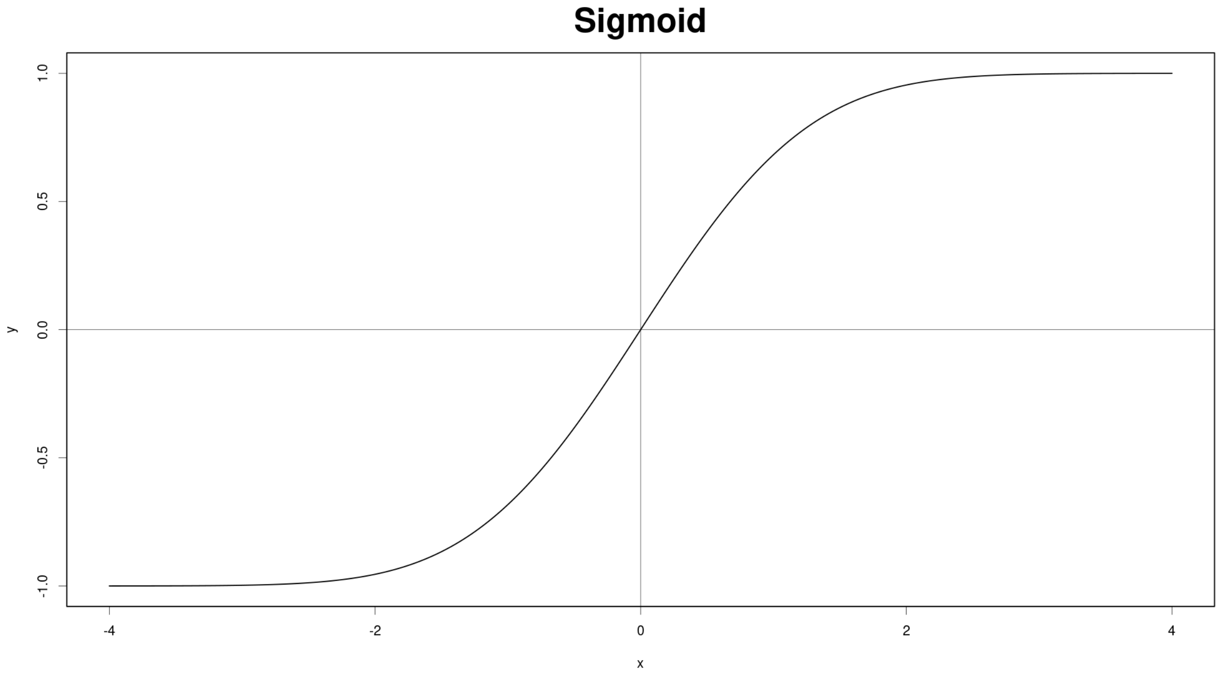
$$Q^*(\mathbf{A}) = \sum_{i=1}^{|P|} \frac{|P_i|}{U} Q^*(\mathbf{P}_i) + Q^*(\bar{\mathbf{A}}) \tag{2.11}$$

As before $\bar{\mathbf{A}}$ represents a distribution with no within cluster segregation and the various unit odds (for each couple of groups) fixed at their geometric average.

2.5.5. Solving the Zero Problem

The zero problem haunts the centered-norm indices and, in general, any margin-free index making them less than desirable according to the basic rules laid out in Section 2.2. As an example consider

Figure 2.1 Examples of a sigmoid function that may be used to solve the zero problem. In particular, this is $\sigma(x) = 2 \cdot \big(N(x) - 0.5\big)$, where $N(x)$ is the cumulative distribution of a standard normal distribution.



the completely-segregated distribution in Table 2.14, which reports both the odds (qs) and logged odds ($\log q$s) for each unit.

The most direct solution to the zero problem is to eliminate zeros altogether by adding something to the zero cells. This may be a simple smoothing value or may be the result of more complex imputation models (Grusky and Levanon, 2006). This solution is certainly convenient since it is very easy to apply in its basic form while it does not change in any way the calculation of an index. However, it should be pointed out that it will substantially bias an index because all odds will become closer to 1 and, especially, extreme odds will be relatively more affected by the change. Finally, the addition of some probability mass where there is none was may feel arbitrary at times

	O_1	O_2	O_3	O_4
O_1	0	2a	0	2a
O_2	-2a	0	-2a	0
O_3	0	2a	0	2a
O_4	-2a	0	-2a	0

Table 2.15. The sigmoid logged ϕ-matrix of the distribution in Table 2.14.

– even if it can be justified, for example, from the perspective of a Bayesian model. Nonetheless, this is the solution we adopt in Section 2.7 for simplicity's sake.

A less committal way to solve the zero problem is to map infinite to a finite value using a sigmoid function. For the present purposes, a useful sigmoid function $\sigma(x)$ will have the following characteristics:

Symmetric Domain: $\sigma(x) : \mathbb{R} \to (-a, a),\ a \in \mathbb{R}$

Continuity: $\sigma(x)$ is continuous

Odd: $\sigma(x) = -\sigma(-x)$. Notice, with the domain property, this implies that $\sigma(0) = 0$

Strictly Increasing: $\sigma(x) < \sigma(y) \Leftrightarrow x < y$

Furthermore, it is easy to justify with a limit argument that $\sigma(\infty) = a$ and, by the odd property, $\sigma(-\infty) = -a$.

If we apply a sigmoid function to the logged odds of a distribution, we can define a new sigmoid logged ϕ-matrix, such as the one shown in Table 2.14. The i^{th}-row, j^{th}-column cell of this new matrix is defined as $\sigma(\log q_i) - \sigma(\log q_j)$: this is well defined even for completely-segregated table. At the same time, such sigmoid logged ϕ-matrix appears to have all the structural properties discussed in Section 2.5.1. Therefore, we can define a modification of Q^* based on such sigmoid logged ϕ-matrix:

$$\begin{aligned}\bar{\sigma}(q) =& \frac{1}{U}\sum_{i=1}^{U}\sigma(q_i) \\ Q^*_\sigma(\mathbf{A}) =& \frac{1}{U}\sum_{i=1}^{U}\Big(\sigma(q_i)-\bar{\sigma}(q)\Big)^2\end{aligned} \tag{2.12}$$

Being a variance, Q^*_σ will still be partition-decomposable, as Q^* is. Furthermore, through Popoviciu inequality (Popoviciu, 1935) it is easy to show that the variance of each of each row/column is maximized precisely when the underlying distribution is completely-segregated. This means that Q^*_σ satisfies the second property of indices indicated in Section 2.2 and is, therefore, a proper index. Finally, the use of a sigmoid function makes it possible to normalize Q^*_σ between 0 and 1. Indeed, we can calculate its maximum value, whereas Q^* is unbounded. Yet, notice that normalizing Q^*_σ will partially hinder its partition-decomposability (Frankel and Volij, 2011; Mora and Ruiz-Castillo, 2011).

While this solution appears more natural than smoothing zeros by adding a (small) value, it introduces further elements of arbitrariness regarding the choice of the sigmoid function. For example, Figure 2.1 shows the function $\frac{1}{2}\cdot(N(x)-0.5)$, where $N(x)$ is the cumulative function of the standard normal distribution. This function substantially flattens all odds values beyond 3 to its maximum. Therefore, its use would tone down the difference between mildly-segregated and extremely-segregated environments according to Q^*. To avoid this inconvenience one may use the cumulative function of a non-standard normal distribution having a higher variance. Yet, any sigmoid function will flatten out extreme values because this is exactly what sigmoid functions do – it is only a matter of defining what “extreme” is. We leave the exploration of this issue to future research.

Finally, it is worth mentioning that both solutions can be applied to any margin-free index as well as Q, defined below. However, applying either of these solution to an index will not necessarily solve the zero problem completely. While the index may be well defined once we smooth the zeros or apply a sigmoid functions, it may still not be maximized for completely-segregated distribution – this is another reason why Q^* is pre-eminent among the centered-norm family.

2.6. The Q index and S-decomposability

The centered-norm family introduce in Section 2.5 can be used as a basis to create S-decomposable indices. The key observation is that the centered-norm family places an equal weight on each unit, but this is by no mean necessary (See also Bouchet-Valat, 2022). We can use the unit marginal distribution to weight the units and retrieve an index that is sensible to the individual experience in the environment:

$$\begin{aligned}
\bar{q} &= \sum_{i=1}^{U} p_i(u) q_i \\
S_2(\mathbf{A}, p, p(u)) &= \sum_{i=1}^{U} p_i(u) \left| q_i - \bar{q} \right|^p \\
S(\mathbf{A}, p, p_{\mathbf{A}}(u)) &= \frac{1}{\binom{G}{2}} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} S_2(\mathbf{A}[i-j], p, p_{\mathbf{A}}(u))
\end{aligned} \tag{2.13}$$

where $p_i(u)$ is the value of the density (probability mass) function of the unit margin at unit i. We will refer to this family of indices as the weighted centered-norm family.

This family has interesting properties for our purposes. First, members of the family have an intuitive structural-decomposition because of the obvious connection with the margin-free centered-norm family – as we will see momentarily for the special case of $p = 2$. Second, indices in the family are not margin-free because they depend on the unit marginal distribution. However, the indices in the family are still independent of the group margin. In the definition of S, all couples receive an equal weight of $\binom{G}{2}^{-1}$ regardless of the probability mass associated with each group. One may argue that the the contribution of each S_2 in the last equation should be weighted proportionally to the probability masses of the groups involved. However, adding a group-margin dependency will create path dependency in the structural-decomposition and complicate the partition-decomposition, whereas a sizeable part of the literature has emphasized that group proportions in an environment should not matter for the quantification of segregation (James and Taeuber, 1985). Therefore,we leave this topic for future research. Finally, an important remark: in the general multi-group case, the passed unit margin is the same as the overall unit margin, it does not mutate as $\mathbf{A}[i-j]$ changes, preserving the partition-decomposability of Q.

The next step is definition of the Q index, which is the member of the weighted centered-norm family having $p=2$. That is,

$$\begin{aligned}
\bar{q} &= \sum_{i=1}^{U} p_i(u) q_i \\
Q(\mathbf{A}, p(u)) &= \sum_{i=1}^{U} p_i(u) \left(q_i - \bar{q} \right)^2 \\
Q(\mathbf{A}, p(u)) &= \frac{1}{\binom{G}{2}} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} Q(\mathbf{A}[i-j], p(u))
\end{aligned} \tag{2.14}$$

As its margin-free counterpart Q^*, Q is partition-decomposable. Once again, consider a unit partition of the environment, where each unit is assigned to one of $|P|$ clusters, $P_1, P_2 \ldots P_{|P|}$. Each cluster has a weight, which is the sum of the probability mass of the units within it:

$$w_i = \sum_{j \in P_i} p_{\mathbf{A},j}(u) \text{ with } \sum_{i=1}^{|P|} w_i = 1$$

Then,

$$Q(\mathbf{A},p(u)) = \sum_{i=1}^{|P|} w_i Q(\mathbf{P}_i, p_{\mathbf{P}_i}(u)) + Q(\bar{\mathbf{A}}_w, p_{\bar{\mathbf{A}}_w}(u)) \tag{2.15}$$

As before, $\bar{\mathbf{A}}_w$ is distribution having no within-cluster segregation and every q set at its within-cluster geometric mean; the difference with respect to equation (2.11) is that the within-cluster mean is weighted by the unit-marginal probability/density function of $\mathbf{A}$.

As mentioned, Q has a natural structural-decomposition. The key observation is that evidently the difference between $Q(\mathbf{A}, p_{\mathbf{A}}(u))$ and $Q^*(\mathbf{A})$ is due to the unit margin, $p_{\mathbf{A}}(u)$. In fact, one may retrieve Q^* by passing a uniform distribution to Q as the unit margin. The intuitive meaning of the structural-decomposition will be clearer starting from the two-group case:

$$\begin{aligned} (2.16)\qquad & Q_2(\mathbf{A}, p_{\mathbf{A}}(u))) = \\ & Q_2(\mathbf{A}, p_{\mathbf{A}}(u)) + Q_2^*(\mathbf{A}) - Q_2^*(\mathbf{A}) = \\ & Q_2^*(\mathbf{A}) + \sum_{i=1}^{U} \left(w_i - \frac{1}{U}\right) \cdot \left(q_i - \bar{q}\right)^2 = \\ & Q_2^*(\mathbf{A}) + \Delta_{\mathbf{A}}(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) \end{aligned}$$

$$\begin{gathered} with \sum_{i=1}^{U} \left(w_i - \frac{1}{U}\right) \cdot \left(q_i - \bar{q}\right)^2 = \Delta_{\mathbf{A}}(\phi[\mathbf{A}], p_{\mathbf{A}}(u)), \\ \sum_{i=1}^{U} \left(w_i - \frac{1}{U}\right) = \sum_{i=1}^{U} w_i - \sum_{i=1}^{U} \frac{1}{U} = 1 - 1 = 0 \end{gathered}$$

$\Delta_{\mathbf{A}}$ represents the (unit) marginal contribution to the segregation of the two-groups environment $\mathbf{A}$. As the derivation shows, this contribution depends on the difference between the unit margin and the uniform distribution, which a is zero-sum difference. Therefore, the marginal contribution to segregation will be positive if the unit margin emphasizes those units that are less representative of the overall group proportion in the environment, negative otherwise. Going back to the two distributions of Table 2.10, their structural-decomposition would be $Q(\mathbf{T}_1, p_{\mathbf{T}_1}(u)) = 3.53 + 1.71 = 5.24$ (left distribution) and $Q(\mathbf{T}_2, p_{\mathbf{T}_2}(u)) = 3.53 - 3 = 0.53$ (right distribution). Such structural-decomposition of segregation follows the intuition that the unit distribution in the left (right) distribution emphasizes those units where group imbalance (balance) is higher, giving a positive (negative) contribution to overall segregation. We can easily generalize to the multigroup case:

$$
\begin{aligned}
(2.17) \qquad & Q(\mathbf{A}, p_{\mathbf{A}}(u))) = \\
& Q(\mathbf{A}, p_{\mathbf{A}}(u)) + Q^*(\mathbf{A}) - Q^*(\mathbf{A}) = \\
& Q^*(\mathbf{A}) + \Delta_{\mathbf{A}}(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) \\
\Delta_{\mathbf{A}}(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) &= \frac{1}{\binom{G}{2}} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} \Delta(\phi[\mathbf{A}[i-j]], p_{\mathbf{A}}(u))
\end{aligned}
$$

This means that Q is structural-decomposable: equation (2.17) fits the blueprint of equation (2.3), with $g(\phi[\mathbf{A}]) = Q^*(\mathbf{A})$, $h_1(\phi[\mathbf{A}], p_{\mathbf{A}}(g)) = 0$, and $h_2(\phi[\mathbf{A}], p_{\mathbf{A}}(u)) = \Delta_{\mathbf{A}}(\phi[\mathbf{A}], p_{\mathbf{A}}(u))$. Even if we will focus on Q, we notice that any member of the weighted centered-norm family will be similarly structural-decomposable.

Finally, we can nest structural-decomposition of equation (2.17) within the partition- decomposition of equation (2.15). Nesting allows us to reach a more complete understanding of segregation patterns by distinguishing the structural/marginal component within each cluster and between clusters. To improve legibility, we will simplify notation in what follows by removing the explicit reference to the unit-marginal distribution. Then:

$$Q(\mathbf{A}) = \sum_{i=1}^{P} w_i Q(\mathbf{P}_i) + Q(\bar{\mathbf{A}}_w)$$

$$\sum_{i=1}^{P} w_i \left[Q^*(\mathbf{P}_i) + \Delta_{\mathbf{P}_i} \right] + Q^*(\bar{\mathbf{A}}_w) + \Delta_{\bar{\mathbf{A}}_w} = \tag{2.18}$$

$$\left[\sum_{i=1}^{P} \frac{|P_i|}{U} Q^*(\mathbf{P}_i) + Q^*(\bar{\mathbf{A}}) \right] + \left[\sum_{i=1}^{P} w_i \Delta_{\mathbf{P}_i} + \Delta_{\bar{\mathbf{A}}_w} + r_{\mathbf{A}} \right] = \tag{2.19}$$

$$Q^*(\mathbf{A}) + \Delta_{\mathbf{A}}$$

$$\textit{with } r_{\mathbf{A}} = Q^*(\bar{\mathbf{A}}_w) - Q^*(\bar{\mathbf{A}}) + \sum_{i=1}^{|P|} \left(w_i - \frac{|p_i|}{U} \right) Q^*(\bar{\mathbf{P}}_i)$$

Equation (2.18) shows the nesting of the structural-decomposition within the partition-decomposition: it provides the structural-decomposition of each cluster and the between component. As the derivation above shows, any index that is structural-decomposable and partition-decomposable can be used to nest the two decompositions. However, Q is currently the only index with such properties. On the other hand, equation (2.19) nests the partition-decomposition within the structural-decompostion. This nesting is not so clear-cut, since there is a residual term ($r_{\mathbf{A}}$). Even if this term may be difficult to interpret in its single elements, it collects all terms that are dependent on the unit-margin, as we expect. Therefore, this nesting may be useful to interpret what part of the environment is driving the trend in the structural-decomposition, as we shall see in the next section.

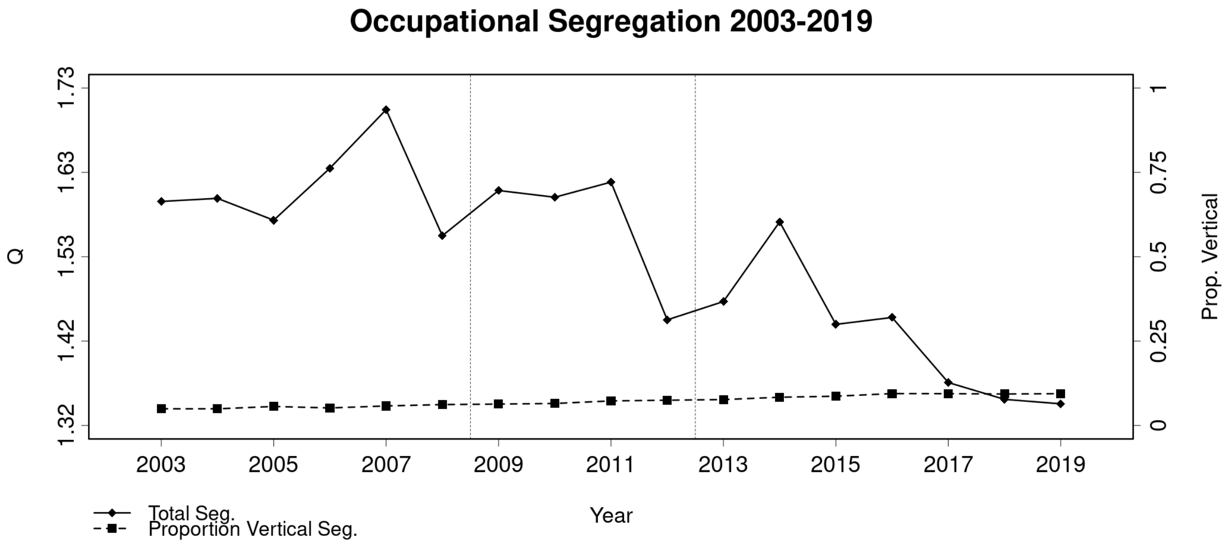
2.7. Application: Analyzing Occupational Segregation Over Time

As an application, this section tests the impact of the 2008 economic recession on gender segregation through an analysis of data from the Current Population Survey (CPS). Specifically, we will analyze gender segregation from 2003 to 2019 from CPS March microdata. The analysis will use an eleven-occupations coding of jobs, which appears to be unique of the CPS data. Even if coarse, this coding has the important quality of being consistent throughout the period; Table 2.1 shows an example of the data from 2003. We operationalize the recession period as the first Obama presidency, from 2009 to 2012. These years span the period in which the crisis happened and the time it took the US economy to reach pre-crisis levels in to many indicators – for example, gross product per capita. Even with just eleven occupational categories, the analyzed data still presents zeros in the least popular occupation (armed forces). To avoid the zero problem (see Section 2.5.5) we add 1 individual to all the cells in the data.

The analysis will focus on two questions. The first question regards the impact of the financial crisis on segregation. Reskin and Roos (1990) famously argue that rising unemployement rate augments gender segregation. Following their reasoning, men – who detain most of the decisional power on the workplace – prefer to dismiss female workers rather than fellow men and will likewise hire men over women when possible. The analyses by Oppenheimer (1970) and Fields and Wolff (1991) substantiate a positive version of this argument: a growing labour market is associated with desegregation. However, the effects of economic recession on segregation have not been thoroughly checked. Here, we use the techniques presented above for a test of the theoretical predictions by Reskin and Roos (1990).

The second question regards the potentially-different impact of the financial crisis on horizontal and vertical segregation. Previous literature has distinguished between two kinds of gender segregation (Hakim, 1979; Charles and Grusky, 2004): vertical segregation is based on the supposed inferiority of women, horizontal segregation is based on the gender-labels attached to various activities. According to Charles and Grusky (2004), current Western ideologies formally oppose vertical segregation because they have established parity between genders. Yet, the same Western ideologies still support the idea that women and men should hold different kinds of jobs, resulting

Figure 2.2 Trend in occupational segregation from CPS data, in 2003-2019. The dotted vertical lines represent the beginning and end of the economic crisis. The proportional vertical segregation is the overall segregation due to the between-segregation among occupational classes.

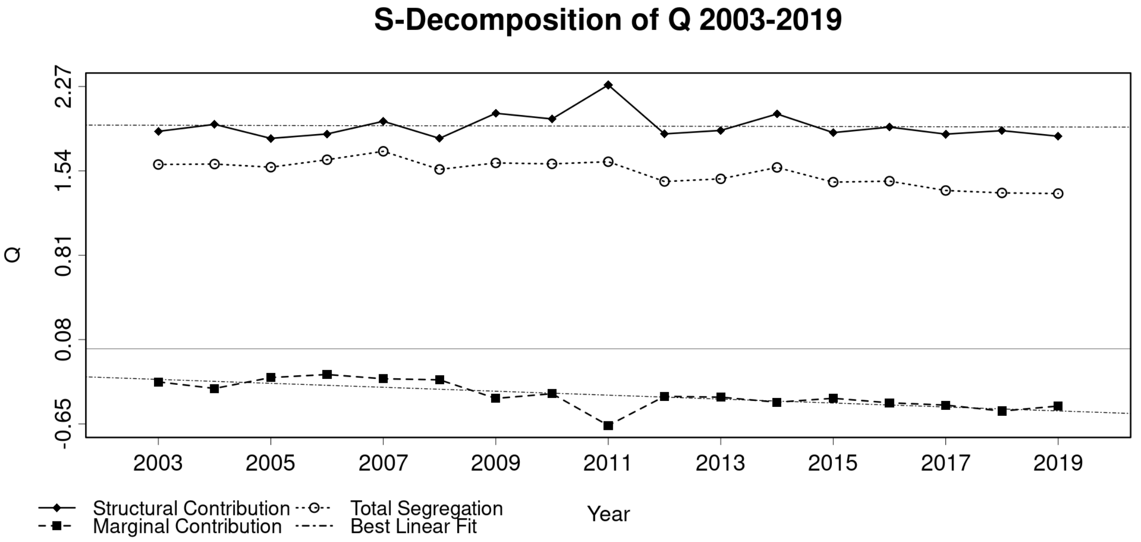


in rampant horizontal segregation. The prevalence of horizontal over vertical segregation entails that women and men are distributed among different occupations that are equally prestigious (Levanon and Grusky, 2016; Charles, 2003). However, considering the argument by Reskin and Roos (1990), one may wonder whether vertical and horizontal segregation follow similar trajectories during economic recession.

To analyze vertical and horizontal segregation, we will use the decomposition properties of Q. The eleven occupations are split into two occupational classes based on prestige and pay-off. Vertical segregation is operationalized as the segregation between the two occupational classes, lower prestige occupations and highly prestigious occupations[15] The division of occupations into

[15] Lower prestige occupations: Service occupations; Sales and related occupations; Office and administrative support occupations; Farming, fishing, and forestry occupations; Construction and extraction occupations; Installation, maintenance, and repair occupations; Production occupations; Transportation and material moving occupations; Armed Forces. Highly prestigious occupations: Management, business, and financial occupations; Professional and related occupations.

Figure 2.3 Structural-decomposition of overall occupation segregation from CPS data. The best linear fit shows the OLS line for the time trend in the marginal and structural component. Total segregation is simply the sum of the two component. Notice that the marginal component is negative for the whole period.

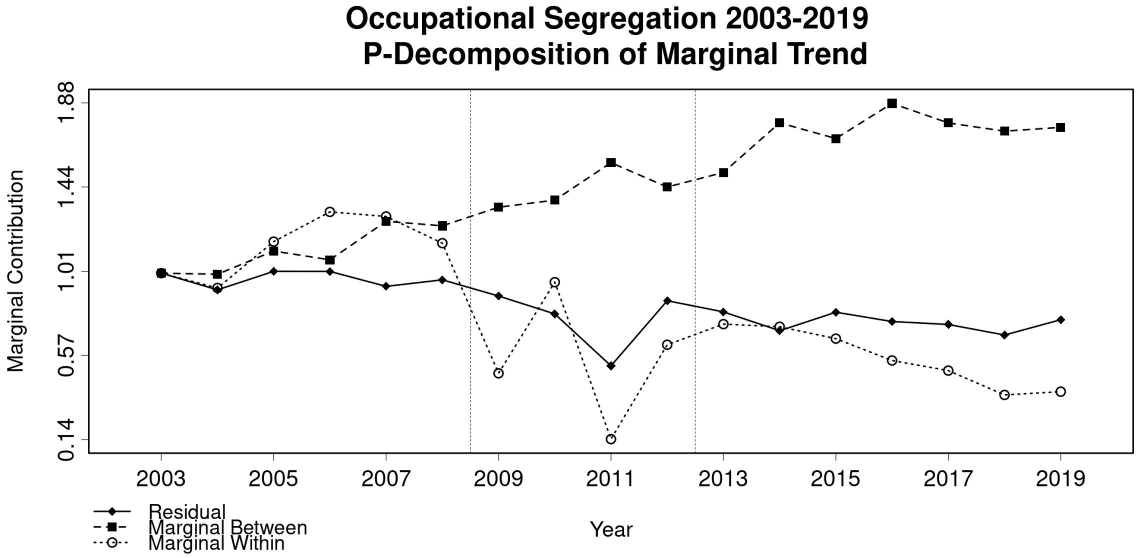


two groups is clearly confirmed by several indicators. For example, 26% of the workers in the first occupational class hold a Bachelor or a higher educational degree. Yet, among the second occupational class, 75% have a bachelor or a higher degree. Are women under-represented in highly prestigious occupations? Horizontal segregation is operationalized as the segregation among occupations with similar pay-off. Given the small number of occupational categories considered, this analysis is to be intended as an exercise to show the usefulness of partition- and structural-decomposition.

We start with the basic question: what is the overall segregation trend in the time considered? Figure 2.2 shows the overall trend in this time period. Overall, segregation has slightly decreased even if with occasional sparks in the data (especially in 2007 and 2014). From the plot, it appears that segregation has decreased especially during the crisis, with a major decrease in 2012. This is in opposition to the prediction by Reskin and Ross (1990). On the other hand, vertical segregation explains little of the overall segregation as predicted by Charles and Grusky (2004), even if there is a small, but clear increase of this kind of segregation from 5% to 9% of total segregation. Most of segregation, comes from within the two occupational classes. In fact, the decrease we see is mostly due to men becoming more common in the professional and office occupational class, which are occupations where women are generally over-represented. The change in these occupational classes is minor, but the two occupations have enough marginal weight to generally carry overall segregation slightly downward. The 2007 and 2014 sparks are mostly due to the construction occupational category, where women are considerably under-represented. In this occupational category a small increase in women's representation can change the unit odds quite substantially, so that the occupation is more prone to sampling error. At the same time, it is worth mentioning that the possible instability has a limited effects overall, since the increase is just around 7% from previous year.

If segregation is decreasing, what kind of change is driving the decrease? We can use structural-decomposition to assess whether the decrease is due to structural factors (occupation becoming more representative of the groups' odd) or to marginal factor (relatively-segregated occupations decreasing in importance). Figure 2.3 shows the structural-decomposition of Q in the time period. Before addressing any trend, notice that the marginal contribution to total segregation is negative for the whole period. Intuitively, this means that the more un-representative units are generally small – for example, the armed forces occupational category over-represents men heavily, however its marginal weight is very small. As for the downward trend of overall segregation, the figure shows that it almost exclusively driven by marginal change. Whereas, the structural component is substantially flat throughout the period, the marginal component is decreasing in a roughly linear fashion – this is shown by the simple (OLS) linear trend for the structural and marginal components

Figure 2.4 Partition-decomposition of the marginal component from the structural-decomposition. The residual in the plot is $r_{\mathbf{A}}$ in equation (2.19), the "Marginal Between" in the plot is $\Delta_{\mathbf{A}_w}$ in equation (2.19), and the"Marginal Within" in the plot is $\sum w_i \Delta_{\mathbf{P}_i}$ in equation (2.19).



of Q. Thus, more segregated occupational categories are becoming slightly less important over time. It is also interesting to notice that the marginal component partially compensates the more erratic behavior of the structural component. From a statistical perspective, Q should be more stable than Q^* as the odds of small units can be grossly mis-estimated in small units, but this estimation issues will have less impact on Q since small units are weighted less in Q.

The final step is the analysis of the within and between components of the marginal term – that is, nesting the partition-decomposition within the structural-decomposition. Figure 2.4 shows the partition-decomposition of the marginal component into the three terms of equation (2.19): $r_{\mathbf{A}}$ (residual), $\Delta_{\mathbf{A}_w}$ (marginal between component), $\sum w_i \Delta_{\mathbf{R}_i}$ (marginal within component). On a substantive level, the marginal between component corresponds to the marginal contribution to vertical segregation. Similarly, the marginal within component corresponds to the marginal contribution to horizontal segregation. The plot is normalized so as to show the trends regardless of the absolute value. The first noticeable trend is that the marginal component of vertical segregation is rising, as opposed to the trends of the remaining component. This means that the occupational cluster that is less representative of the market is rising in importance. Paradoxically, such cluster is the higher-prestigious cluster, where the ratio of women to men is close to parity in 2019. However, the overall women proportion in the whole job market is not so close to parity and it is better represented by the less prestigious cluster of occupations, where men are constantly prevalent. This means that women are ultimately over-represented in more prestigious occupations. Since more prestigious occupations are proportionally growing over time, the final result is that vertical segregation is augmenting because more equal, higher-prestige (but un-representative) occupations are rising – which is the opposite of the original understanding of vertical segregation (Charles and Grusky, 2004). Admittedly, the surprising result can be interpreted as an artifact of the very coarse occupational coding, which does not allow us to fine-tune the definition of ”prestigious” occupation; once again, this analysis is mostly an exercise in decomposition and comparison. The result is also a good reminder that a distribution that is representative of the job market is not the same as a distribution that is representative of the population – and indices are necessarily based on the former distribution. On another note, both the residual and the within marginal components show downward trends. Together, these components account for the overall decrease of segregation we see in Figure 2.2. In magnitude, the terms account for roughly the same amount of decrease. While the residual is difficult to interpret, the within marginal component is the weighted sum of the marginal components of the higher and the lower occupational cluster. Therefore, the downward trend in this component means that the two cluster becomes less segregated over time because their unit margins changes in such a way as to de-emphasize the most segregated unit. It is interesting that both negative trends appear to intensify (or even start) during the economic crisis.

Finally, it is interesting to notice that the marginal within-clusters contribution to segregation is positive (not shown in figure), as is the marginal contribution of vertical segregation (i.e. the between marginal component). On the other hand, notice that the overall marginal contribution is negative (see Figure 2.3). Thus, the marginal contribution to segregation within and between clusters is positive, whereas the overall marginal contribution is negative. At first, this result may seem contradictory, but it simply proves that the representativity of each occupation changes when considering an environment as a whole. In particular, low-prestige occupations that have a close-to-equal representation of women and men are not representative of the geometric mean of their cluster. However, they are much closer to being representative of the entire environment. Therefore, their overall marginal contribution to segregation ends up being negative, even if their within-cluster marginal contribution is actually positive.

In conclusion, the analysis shows segregation has decreased during the period analyzed and specifically during the economic crisis. The decrease is due to a slight change in the marginal distribution of the unit, while the structural component has remained largely untouched. In particular, it seems that lower-level occupations that are more equal are driving the change. On the other hand, vertical segregation is relatively small, but is facing a steady increase over time, which is driven by the diffusion of the more equal prestigious occupations.

2.8. Conclusions

Interpreting change in segregation is a fundamental issue in the social sciences (England et al., 2020; Tomaskovic-Devey et al., 2006; Massey et al., 2009, see for example). So far, the interpretation of segregation trends has been hindered by the segregation indices at our disposal, which confounds different sources of change (Blau and Hendricks, 1979; Semyonov et al., 1984) or are based on debatable theoretical principles (Charles and Grusky, 1995; Elbers, 2021). This paper formulated the interpretability task in a formal way – equation (2.2) – and built a family of index that guarantees interpretability, the weighted centered-norm family. The paper focused particularly on the Q index because of its decomposability properties: like the Theil index, Q is partition-decomposable, as shown in equation (2.15). That is, Q can distinguish the unit-clusters in the data originating more segregation (Mora and Ruiz-Castillo, 2011). Moreover, being a member of the weighted centered-norm family, Q is structural-decomposable. That is, it can unambiguously distinguish the segregation contribution of the association element and the marginal element of a distribution. Thanks to this, we can interpret the segregation difference between different environments in more detailed and unambiguous ways. Moreover, we can nest the two decompositions to get insights on the phenomena driving change.

As an example, Section 2.7 analyzed the change in occupational segregation between men and women. The section concluded that overall segregation is slightly decreasing and that the downward trend is mostly due to marginal change in lower-prestige occupations. On the other hand, vertical segregation between women and men is rising, even if in a surprising way: women are getting a foot in higher prestige occupations but are less successful in equalizing lower-prestige occupations, such as construction. Naturally, the analysis should be taken with some caution due to the very coarse occupational coding used. However, the analysis showcases the usefulness of nesting the partition-decomposition within the structural-decomposition and vice versa.

In order to formulate Q and its decomposability property, the paper formulated new margin-free indices and formalized their principles. Thanks to general results by Osius (2004), the paper showed that the only way to create margin-free indices is to use the association element of a distribution. Starting from the analysis of such element, the paper has formulated properties of the association element that have been so far gone unnoticed. Namely, that the centered p-norm of this element is a constant. Based on this, the paper formulated a new family of margin-free indices:

the centered-norm family. A member of this family (Q^*) has the property of being partition-decomposable and it is the first margin-free index to have such property. For this reasons, researcher interested in using margin-free indices may consider using it over available alternatives such as A (Charles and Grusky, 1995). Finally, the paper paved out a formal solution to the zero problem haunting all margin-free indices so far, as well as margin-free Q.

While the paper has introduced new concepts and methods, much remains to be explored. The paper formulated but did not explore the proposed solution to the zero problem based on sigmoid functions. Moreover, Q as formulated in equation (2.14) still suffers from the zero problem, even if it is not margin-free. Thus, an implementation of the proposed solution appears very relevant. In its multi-group form, Q can be reformulated to include influence from the group margins. Currently, Q weights all groups equally regardless of their prevalence in an environment, but this may be undesirable. At the same time, it is important to check how S- and partition-decomposability will change after the introduction of the group-margin influence. Finally, the paper comes short of characterizing the Q and Q^* indices – or more in general the centered-norm and weighted centered-norms families. Besides margin independence (or dependence), the formal properties of Q and Q^* remains to be explored. For example, Q^* (Q) will increase if a unit odds above the geometric-average (weighted geometric-average) of the units' odds further augments. However, much remains to be explored from this perspective.

Overall the paper advanced our current methodological and theoretical understanding of segregation and its quantification by providing a formal re-framing of a decade-long discussion and a possible solution to it. Thanks to their properties, Q and Q^* will help future research to analyze segregation and its patterns.

CHAPTER 3

A Simple Algebra of Meaning: Word Embedding Models in the Social Sciences beyond Cosine Similarity

3.1. Word Embedding Models in the Social Sciences: A Double Translation

The social sciences have a long standing interest in studying meaning and its production by groups and individuals (Mohr, 1998). Following this historical trend, word embedding models are becoming increasingly-popular among social scientists as a way to extract meaning from textual corpora (Taylor and Stoltz, 2020; Ash et al., 2020). These models have attracted much attention because of their ability to capture implicit and explicit cultural associations starting from widely-available raw textual data (Garg et al., 2018; Arseniev-Koehler and Foster, 2020; Jones et al., 2020).

From a technical perspective, word embedding models represent each word type from the textual corpus as a point in a highly-dimensional Euclidean space, following the principle that words appearing in similar contexts should be represented by points that are close (Turney and Pantel, 2010) – the exact mathematical meaning of "context" and "close" depends on the specific word embedding model used. This is an application of the distributional hypothesis (Harris, 1954), which states that two words have similar meaning if they tend to appear in similar contexts. Interestingly, embedding models often reproduce human judgement (Joseph, 2020) and their main advantage is their ability to extract implicit cultural schemas from unstructured corpora (Arseniev-Koehler and Foster, 2020; Boutyline and Soter, 2021). Indeed, it is often easier to obtain text from a group than to survey its members – in the study of past culture, word embedding models may be the best tool we have. Therefore, word embedding models opens up the possibility of studying cultural schemas and stereotypes using cheap and readily-available data.

As a consequence, the great promise of word embedding (and semantic computational methodologies more in general) is the ability to produce a cartography of the culture of a group – i.e. maps condensing complex meaning-making processes in a simplified representation (Lee and Martin, 2015; Stoltz and Taylor, 2021). From this perspective, we can think of word embedding models and their results as signs: “The sign stands for something, its object. It stands for that object, not in all respects, but in reference to a sort of idea, which I have sometimes called the ground of the representamen [sign].” (Peirce, 1931, p. 228) Like a sign, a map completely transforms the territory it represents, but it translates and retains those features that are essential for navigation, such as relative distances (Latour, 2013). Similarly, word embedding models dismiss most information in the analyzed corpus, but they create computational maps of a culture retaining important semantic aspects of the original corpus. In turn, these maps assist the informed interpreter in formulating and defending social scientific theories (Lee and Martin, 2015).

However, the output of word embedding models is still not a cultural map. Indeed, there are two distinct phases (or translations) in any social scientific use of word embedding models:

(1) The embedding model outputs a mapping of words to points in a highly-dimensional Euclidean space. This is the mere output of a computational word embedding model. Therefore, there is the translation of a corpus into a metric (normed) space populated with points, representing words.

(2) There is the translation of the geometric space above into semantic relationships between concepts – which, in turn, forms the basis for social scientific statements such as “the women’s movement in the 1960s and 1970s [...] had a systemic and drastic effect in women’s portrayals in literature and culture” (Garg et al., 2018, E3638)

A vast literature in computer science explores the first translation (see for example Mikolov et al., 2013; Pennington et al., 2014; Levy and Goldberg, 2014b; Hashimoto et al., 2016; Arora et al., 2016; Devlin et al., 2019; Dieng et al., 2019). This paper focuses on the second translation, which has received far less attention even if it is not trivial (Levy and Goldberg, 2014a; Arora et al., 2017; Peterson et al., 2020).

The first translation maps words to Euclidean points based on the distributional hypothesis (Harris, 1954), even if distinct embedding models differ in the way they operationalize this hypothesis into mathematical equations. However, the distributional hypothesis is not strictly a semantic hypothesis (Harris, 1954; Lenci, 2008). In fact, embedding models may be used to extract syntactic and morphological relations as well as semantic relations (Mikolov et al., 2013). Therefore, points representing words from a corpus are still not of great interest to the social sciences. Only after the second translation, word embedding models effectively produce cultural maps condensing socially-shared meaning (Arseniev-Koehler and Foster, 2020): it is only the latter translation that allows social scientists to study socially-shared culture and its evolution. Social sciences have already created methodologies for it (see for example Garg et al., 2018; Kozlowski et al., 2019; Stoltz and Taylor, 2019; Rodriguez et al., 2021), but a theoretical discussion about the semantic suppositions, limitations and models used in this process started only recently (Arseniev-Koehler and Foster, 2020; Stoltz and Taylor, 2021; Arseniev-Koehler, 2021; Boutyline and Soter, 2021). In turn, the lack of clear theoretical understanding impedes faster methodological advancement in the use of word embedding models in the social sciences.

As a starting point, the guiding principles of the second translation is specifically that the meaning of words is isomorphic to their positions as points in the Euclidean space (Arseniev-Koehler, 2021; Tversky, 1977; Kozlowski et al., 2019). Therefore, the second translation draws upon a semantic-geometric model – that is, a model of how positions of points maps onto semantic relationships and vice versa. It follows that we can explore meaning “by performing simple algebraic operations with the vector representation of words” (Mikolov et al., 2013, p. 5). That is, we can use simple geometrical operations on word embedding results to study meaning.

However, the exact algebraic/geometric operations performed on the word-vectors can differ. In fact, this is analogous to the fact that different word embedding models uses different mathematical implementations of the distributional semantics hypothesis, as mentioned above. Starting from the seminal paper of Mikolov and colleagues (2013), the parallelogram model of analogical reasoning has enjoyed the greatest popularity to extract semantic relationships from word embedding models (Peterson et al., 2020). This model was first proposed in the context of cognitive psychology (Rumelhart and Abrahamson, 1973) to explain analogical reasoning of the kind "*Paris* is to *France* what *Berlin* is to *Germany*". In turn, social scientists rely on a straightforward expansion of the parallelogram model: the trapezoidal model of analogical reasoning, discussed below.

As a consequence, social scientists continue to draw upon analogies to extract semantic relationships from word embedding models and they inherit the intrinsic limitations of analogical reasoning. In particular, social scientists inherit the focus on one binary category at the time and the interest in rather abstract concepts. It is not by chance that so many applications of embedding models in the social sciences have focused on gender (see in particular Garg et al., 2018; Lewis and Lupyan, 2020; Boutyline et al., 2020; Ash et al., 2020; Jones et al., 2020; Nelson, 2021): gender is a traditionally-binary concept with deep ramification in many cultural aspects (Bourdieu, 1979). From this perspective, gender is a textbook case-study for the application of analogical reasoning and its extension (see Bolukbasi et al., 2016). Even when not focused specifically on gender, most applications of word embedding models have analyzed one binary (or binarized) concepts, its association with other concepts, and the evolution over time of the associations (see for example Caliskan et al., 2017; Kozlowski et al., 2019; Arseniev-Koehler and Foster, 2020). While the current social scientific applications of word embedding models are insightful, culture is much more complex than analogies.

However, it is not necessary to limit ourselves to models of analogical reasoning. From this perspective, a growing literature shows that word embedding modeling can be used for semantic purposes well beyond the parallelogram model and its offshoots (Stoltz and Taylor, 2019; Arseniev-Koehler et al., 2022; Reimers and Gurevych, 2019; Rodriguez et al., 2021). Such methodologies expand the applications of word embedding for the study of culture and account for more complex cultural processes – such as textual (as opposed to word) meaning (Stoltz and Taylor, 2019; Arseniev-Koehler et al., 2022; Reimers and Gurevych, 2019) or group-specific cultural differences (Rodriguez et al., 2021). Effectively, this line of works explores new ways to translate word embedding results into semantic relationships – and eventually social scientific statements. This paper pursues the same goal of expanding the possible translations of word embedding results into semantic insights.

3.1.1. Original Contribution

This paper focuses on the projection task, which assesses the mutual engagement of two concepts (for example, the concept of *gender* and occupational titles, Garg et al., 2018; Jones et al., 2020) or the engagement of a whole document with a concept (for example, the engagement of a Shakesperean play with the concept of *life*, see Taylor and Stoltz, 2020). This task has been the main focus of the social scientific uses of word embedding models so far (see for example Bolukbasi et al., 2016; Garg et al., 2018; Kozlowski et al., 2019). As a result of this increasing interest, a sizeable number of closely-related, but distinct methodologies have been proposed to carry out the projection task (Joseph and Morgan, 2020), while most proposed methodologies unnecessarily limit the application of word embedding to binary concepts such as gender or class.

The reason for these limitations is the lack of discussion about the semantic assumptions regulating the projection task – that is, the kinds of semantic relation that it can unearth and the kinds it cannot. Drawing on work from cognitive science (Rumelhart and Abrahamson, 1973; Tversky, 1977; Peterson et al., 2020) and semiotics (Greimas and Courtés, 1982), this paper formulates a new semantic-geometric model that justifies the current standard methodologies as special cases, which can be written as Ordinary Least Square (OLS) regressions. We refer to this semantic-geometric model as the "simple algebra of meaning". From a theoretical perspective, the simple algebra of meaning justifies the projection task in the social sciences – e.g. the ubiquitous use of cosine similarity. Therefore, the paper formalizes, links and generalizes most uses of word embedding models in the social sciences so far.

In addition to formulating the simple algebra of meaning, the paper uses it to formulate new methodologies that expand the projection task beyond analogical reasoning. The paper proposes new methodologies to 1. analyze non-binary concepts, 2. more than one concept at the same time, 3. and entire documents. Finally, the paper shows 4. how to decompose the document-level analysis word-by-word. The new methodologies moves beyond analogical reasoning, but still draw upon the simple algebra of meaning. The paper shows that the new techniques can generate convincing results through the analysis of Lovecraftian words (that is, words related to the author H. P. Lovecraft) and a corpus 30,228 scientific abstracts about climate change (Nanni and Fallin, 2021).

3.1.2. Terminology and Notation

A quick note on the terminology and notation we will use through the paper. We refer to the results of a word embedding model with the general term "embeddings". When we refer to the representation of a word as a vector, we use the hyphened term "word-vector". The expression "embedding space" refers to the Euclidean space containing the estimated word-vectors and, possibly, vector representations of documents (named "document-vectors").

As for the notation, we will use specific notation to distinguish words from their respective word-vectors: words will be written in italics (e.g. *France*) whereas their vector representations will have an arrow on top of them (e.g. $\vec{France}$). The same notation will distinguish the concepts associated with a semantic dimension (e.g. *gender*) from their vector representations (e.g. $\vec{gender}$). More in general, all vectors will always be signaled by an arrow (e.g. $\vec{w}$) both in the text and in the equations. Finally, bold capital letters represent matrices (e.g. $\mathbf{W}$).

3.2. The Projection Task

Word embedding models have a long history (Landauer and Dumais, 1997; Turney and Pantel, 2010). Yet, their popularity surged in the last decade after Mikolov et al. (2013) showcased the ability of the CBOW and skipgram word embedding models to capture semantic relationships with a precision never before achieved (see also Baroni et al., 2014). In particular, Mikolov and colleagues showed that the skipgram and CBOW models achieved impressive results in the semantic analogy task. In this task, the model is asked to complete an analogy like "*Paris* is to *France* what *Berlin* is to … [*Germany*]". Even if the analogy task may appear distant from social scientific concerns, computer scientists soon realized that analogies of this kind could encode implicit cultural schemas about social groups and their activities. For example, the very same model trained by Mikolov and colleagues provides the analogy "*man* is to *woman* what *computer programmer* is to *homemaker*" (Bolukbasi et al., 2016). The point is that word embedding models are trained on large textual corpora where this kind of cultural associations are implicit – and sometimes explicit. Therefore, embedding models pick up the stereotypes of the culture producing the documents, as expected (Caliskan et al., 2017; Kozlowski et al., 2019).

The implicit cultural schemas encoded in embeddings are a problem to be solved when embeddings are used in downstream applications (Bolukbasi et al., 2016; Swinger et al., 2019; Ethayarajh et al., 2019), such as recommendation systems. On the other hand, the schemas in the embeddings are a valuable source of information for social scientists studying culture and its evolution. The versatility of the model is demonstrated by the variety of its applications in the social sciences. Indeed, social scientists have used word embedding models to study a variety of different issues: the association of *obesity* to *gender*, *socio-economic status*, *morality* and *health* (Arseniev-Koehler and Foster, 2020), the evolution of the concept of *social class* (Kozlowski et al., 2019), the evolution and geographical variation of the association of *gender* and *race* with *work* or *education* (Garg et al., 2018; Caliskan et al., 2017; Jones et al., 2020; Boutyline et al., 2020; Lewis and Lupyan, 2020), the association of the concept of *immigration* with *crime*, *family*, *education* and *jobs* in news outlets (Stoltz and Taylor, 2021), the association of *flowers*, *insects*, *weapons*, *musical instruments* and *race* with *pleasantness* (Caliskan et al., 2017). In fact, research has shown that word embedding models are often capable of reproducing survey results about beliefs and affective dimensions (Kozlowski et al., 2019; Joseph and Morgan, 2020; van Loon and Freese, 2022; Lewis and Lupyan, 2020) – even if debate exists about which semantic domains and questions are better suited to be studied with word embeddings (see for example Peterson et al., 2020; Arseniev-Koehler, 2021). For reasons clarified below, we will refer to the the task of extracting associations from embeddings as "the projection task". For example, which job title is mostly associated with femininity/masculinity (Garg et al., 2018; Jones et al., 2020; Lewis and Lupyan, 2020)?

To reiterate, most social scientific research using word embedding models relies on the projection task, which is therefore the methodological fulcrum of the social scientific interest in these models. Unfortunately, this task is not banal at all. Word embedding results are merely a mapping of words onto numerical vectors based on textual co-occurrences: they lack a natural interpretation to extract semantic information in order to perform social scientific inference. At their core, embeddings are simply a mathematical reduction of the enormous amount of information contained in a textual corpus. As Lee and Martin (2015) argue, information reduction is a fundamental step to create defensible inferences, but it must be coupled with an explicit and though-out interpretation of the reduced information. Thus, the projection task relies on a double process of translation: from a corpus to embeddings, from embeddings to semantic association.

Usually, the second translation relies on linear algebra operations on the embeddings – such as calculating the cosine between two words. However, the specific operations used to analyze meaning from embeddings outline different semantic-geometric models. We define a "semantic-geometric" model as a model that regulates the way meaning is translated into geometric constructions and vice versa. As the name suggests, a semantic-geometric models contains assumptions about semantics and geometry. We will discuss semantic assumptions below. As for geometric assumptions, we assume an Euclidean geometry for embedding spaces (but see Nickel and Kiela, 2017). The procedures that social scientists use for the projection task commit social scientists to a specific semantic-geometric model. Therefore, we start the discussion with a thorough analysis about the current way the projection task is usually accomplished.

3.2.1. A *de facto* Standard

To understand the assumptions behind the current way the projection task is performed, we examine the exemplary methodology by Kozlowski et al. (2019), which has been a model for subsequent works (Boutyline et al., 2020; Ash et al., 2020; Taylor and Stoltz, 2021) and is clearly explained. While this is not a sanctioned standard in the social sciences, many works that do not use Kozlowski et al. (2019) as a model can be shown to reduce to the same methodology or something very close (see for example Garg et al., 2018; Jones et al., 2020). Moreover, small variations of this methodology do not actually influence the final results in substantial ways (Jones et al., 2020;

Kozlowski et al., 2019; Garg et al., 2018). Therefore, we consider this methodology a *de facto* standard for the projection task.

Kozlowski and co-author are interested in the evolution of the concept of *social class* through the 20^{th} century in the US culture. They elaborate the methodology described below to answer questions such as: which is the most prestigious (unprestigious) job? Which is the most upper (lower) class sport? As a first step, they distinguish 7 different conceptualization of *social class* – for example, *social class* as a difference in *affluence*. For each conceptualization, the authors create a list of paired seed words that differ only with respect to the concept of interest – for example, the pair *rich* and *poor*. After estimating their embedding models on textual data from Google Books, they use the word pairs to calculate the vector representing the concept of interest in the embedding space – continuing with the previous example, this would be the vectors representing $\vec{affluence}$ in the embedding spaces. We will call such vector a "semantic dimension" of the embedding space. More specifically, they calculate a semantic dimension as a mean of the differences between paired seed words. If we indicate with $\mathscr{P}$ the set of word pairs, with $(w_{1,i}, w_{2,i})$ the i^{th} word pair, we can write:

$$\vec{D} = \sum_{(w_{1,i}, w_{2,i}) \in \mathscr{P}} \frac{\vec{w}_{1,i} - \vec{w}_{2,i}}{n} \tag{3.1}$$

where $\vec{D}$ is the semantic dimension of interest (for example, $\vec{morality}$) and n is the total number of word-pairs in $\mathscr{P}$. The result of this calculation is a direction in the embedding space where the two extrema are semantically associated with opposite concepts, such as *affluence* and *poverty* in the running example. We will refer to the two extrema of the semantic dimension as its "poles".

Afterwards, the authors calculate the cosine similarity between the semantic dimensions just calculated and selected words. This way, they assess which concepts are more (or less) engaged with the semantic dimensions. For example, they find that *camping* is the sport[1] most associated with *poverty* and *golf* is the sport most associated with *affluence* (Kozlowski et al., 2019, p. 912-913). In symbols, the engagement between a word ($\vec{w}$) and a semantic dimension ($\vec{D}$) is calculated as:

$$\beta(\vec{w},\vec{D}) = \cos(\vec{w},\vec{D}) = \frac{\vec{w}}{||\vec{w}||} \cdot \vec{D} \tag{3.2}$$

where $\beta(\vec{w},\vec{D})$ indicates the looked-for engagement, $||\vec{w}||$ is the Euclidean norm of the word-vector $\vec{w}$, $\frac{\vec{w}}{||\vec{w}||} \cdot \vec{D}$ is the standard dot product between the two vectors, and it is assumed that $||\vec{D}|| = 1$ to simplify the notation – as we shall see, the norm of $\vec{D}$ can be arbitrarily changed without changing the substantial conclusions from a projection task. A key property of equation (3.2) is that the use of cosine similarity bounds $\beta(\vec{w},\vec{D})$ between -1 and 1, which makes it easy to compare different associations.

Two observations are in order. First, this methodology is independent of the word embedding model used. The authors used a skipgram model, but they may have used the same methodology with other embedding models, such as GloVe (Pennington et al., 2014) or matrix factorization (Levy and Goldberg, 2014b). This shows that the projection task is truly a separate process from the creation of the embeddings. Second, Kozlowski et al. (2019) divide the projection task into two different sub-tasks. In the first sub-task, they use a list of seed words and a simple mean to calculate a semantic dimension – equation (3.1). In the second sub-task, they use cosine similarity to calculate the engagement of a word with the calculated semantic dimension – equation (3.2). The literature has used both sub-tasks frequently. In general, taking the mean of the differences between paired seed words is the most common way to estimate a semantic dimension (Garg et al.,

[1] Someone may argue that camping is not a sport. Here we are only reporting the definitions and results by (Kozlowski et al., 2019, p. 912-913).

2018; Arseniev-Koehler and Foster, 2020; Jones et al., 2020; Taylor and Stoltz, 2021; Boutyline et al., 2020).[2] As for the second sub-task, cosine similarity is substantially ubiquitous in social sciences' uses of embeddings (Bolukbasi et al., 2016; Stoltz and Taylor, 2019; Arseniev-Koehler and Foster, 2020; Joseph and Morgan, 2020; Ash et al., 2020; Boutyline et al., 2020; Jones et al., 2020).

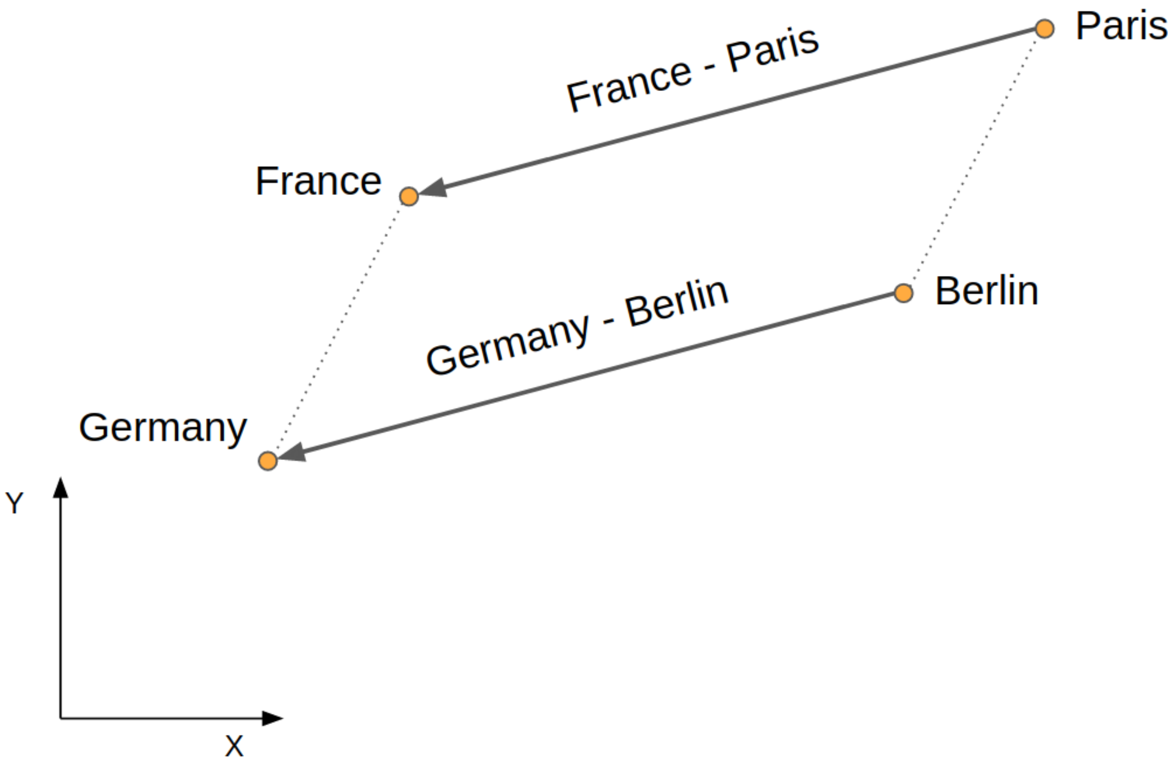
3.2.2. The semantic-geometric Model of the Projection Task

The strategy laid out by Kozlowsky and colleagues is a *de facto* standard for the use of embeddings in the social sciences. However, its popularity and achievements should not naturalize it. It implicitly relies on a semantic-geometric model – a model of how words' meaning is translated into the positions of word-vectors in the embedding space and vice versa. From this perspective, the projection task relies on two key assumptions, conveniently fleshed out in equations (3.1) and (3.2): semantic dimensions such as *gender* or *affluence* are represented as directions in the embedding space – equation (3.1); semantics is isomorphic to the relative positions of word-vectors in the embedding space – equation (3.2).

The first assumption comes from formal models of analogical reasoning through the seminal paper by Mikolov, Chen and colleagues (2013). These authors supposedly use the parallelogram model of analogical reasoning to solve the analogy task within embedding space. This model was first proposed by in the context of cognitive psychology to explain analogical reasoning of the kind "*Paris* is to *France* what *Berlin* is to *Germany*" (Rumelhart and Abrahamson, 1973; Peterson

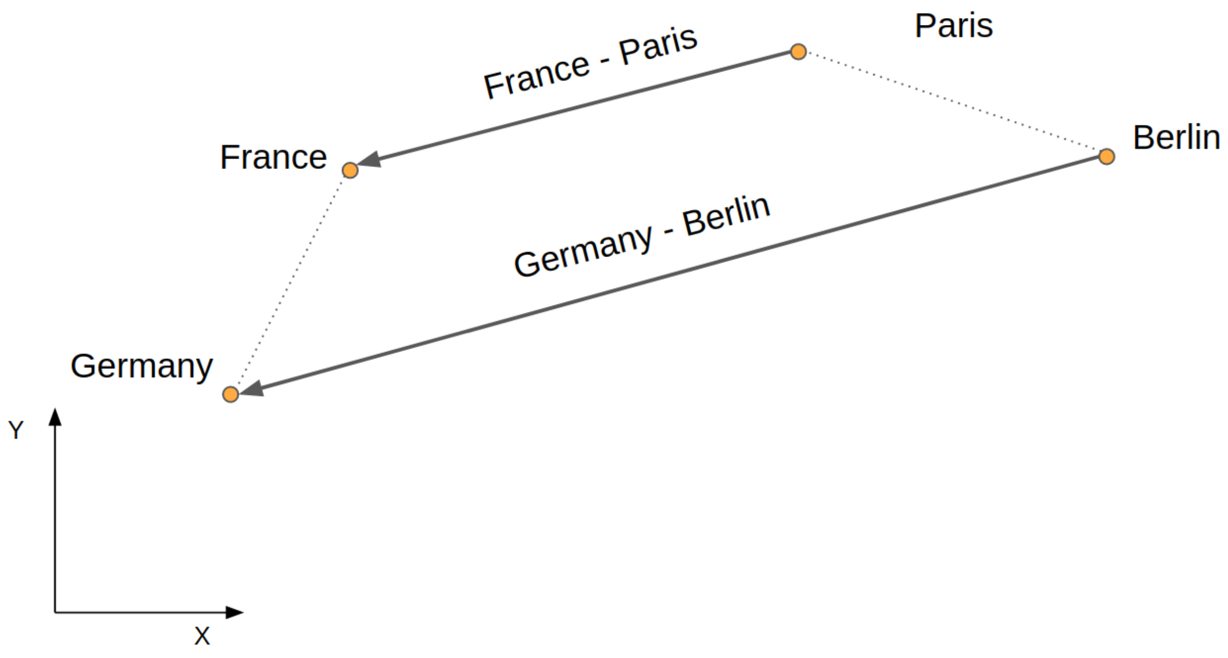
[2] It should be noticed that there is an alternative strategy for the calculation of semantic dimensions consisting in using the first principal component of the couples' differences as the semantic dimension (Bolukbasi et al., 2016; Ethayarajh et al., 2019). Under the model presented in section 3.3, it is possible to show that the mean of the differences and their first principal component converge to the same vector due to the interpretation of the first principal component as the fitted regression line minimizing the residual variance for a series of points (Pearson, 1901). However, it appears from our empirical test that the principal component strategy is substantially slower to converge. It is therefore ignored in the remaining of the paper.

Figure 3.1 A two-dimensional visualization of the analogy "*Paris* is to *France* what *Berlin* is to *Germany*" according to the parallelogram model (Rumelhart and Abrahamson, 1973).



et al., 2020) . It can be summed up in the simple equation: $\vec{France} - \vec{Paris} \approx \vec{Germany} - \vec{Berlin}$ (see Figure 3.1). Yet, Mikolov and co-authors do not actually use a parallelogram model to solve their analogies, as Levy and Goldberg (2014a) clarify. Rather, they use what we may name as a "trapezoidal" model of analogical reasoning. The difference with respect to the parallelogram model is that the parallelogram model assumes that semantic dimensions are represented by fixed-length line-segments in the embedding space. On the contrary, the trapezoidal model makes no reference to the length of the line-segments, only to a direction. In the analogy example above, the vectors $\vec{France} - \vec{Paris}$ and $\vec{Germany} - \vec{Berlin}$ would identify the direction representing the

Figure 3.2 A two dimensional visualization of the analogy "*Paris* is to *France* what *Berlin* is to *Germany*" according to the trapezoidal model (Mikolov et al., 2013; Levy and Goldberg, 2014a)



relationship *country-capital*. Therefore, $\vec{Germany}$ and $\vec{Berlin}$ will be the same distance as $\vec{France}$ and $\vec{Paris}$ in the parallelogram model (see Figure 3.1). However, in the trapezoidal model, the segment between $\vec{Paris}$ and $\vec{France}$ may be considerably shorter than the segment between $\vec{Berlin}$ and $\vec{Germany}$, despite the two segments being parallel (see Figure 3.2). The trapezoidal model can be summed up as $\vec{France} - \vec{Paris} \approx \beta(\vec{Germany} - \vec{Berlin})$, with the addition of a coefficient β regulating the length of the segment.

Notice that the trapezoidal and parallelogram models only imply that those words entering the same analogical relations would be joint by parallel line-segments[3]. For example, the vectors

[3] We are assuming analogies are transitive. Given two analogies such as (1) "*Paris* is to *France* what *Berlin* is to *Germany*" and (2) "*Berlin* is to *Germany* what *Rome* is to *Italy*", transitivity implies that (3) "*Paris* is to *France* what *Rome* is to *Italy*". It should be pointed out that some scholar believes that analogies are not necessarily transitive (Tversky, 1977).

$\vec{France}-\vec{Paris}$, $\vec{Germany}-\vec{Berlin}$, $\vec{Italy}-\vec{Rome}$, $\vec{China}-\vec{Beijing}$, etc. will all be parallel because they join two word-vectors entering the country-capital analogy. Therefore, semantic relations are represented as directions in space. Yet, beware of the orientation of the inference in analogical models: *if* two words enter in an analogy together, then we know their positions relative to one another. Crucially, the projection task reverses the inference direction: not only is it possible to infer embeddings' positions from the words' meaning, but it is also possible to infer words' meaning from their embeddings' positions. This is the second assumption of the projection task and it makes it possible to translate embeddings into semantics. More specifically, the use of cosine similarity (Equation 3.2) dictates the exact way in which semantic relations are encoded into embeddings' positions and vice versa. Effectively, the use of cosine similarity transforms analogical models into a full-fledged semantic-geometric model, where the meaning of words can be reconstructed from the decomposition of their word-vectors into basic semantic dimensions.

We can clarify this point with an example. Consider the question which sport is mostly associated with *affluence* among *camping* and *golf* – a question that Kozlowski et al. (2019) consider. Assume we have already calculated the direction representing $\vec{affluence}$ per equation (3.1); we indicate this direction with $\vec{A}$. Equation (3.3) answers the question by ranking the sports based on their cosine with the $\vec{A}$. From a geometric perspective, the cosine measures the (normalized) length of the projection of $\vec{camping}$ and $\vec{golf}$ onto the semantic dimension $\vec{affluence}$ – from here, the name "projection task". We can write this geometric intuition into two simple equations:

$$\vec{camping} = \beta(\vec{camping},\vec{A}) \cdot \vec{A} + \alpha_C \cdot \vec{\varepsilon}$$

$$\vec{golf} = \beta(\vec{golf},\vec{A}) \cdot \vec{A} + \alpha_G \cdot \vec{\varepsilon}$$

The vector $\vec{\varepsilon}$ represents the mixture of all semantic dimensions that are not $\vec{A}$, which we ignore at the moment. The cosines between the words and the dimension $\vec{A}$ are precisely (normalized) estimates of the coefficients $\beta(\vec{camping},\vec{A})$ and $\beta(\vec{Golf},\vec{A})$, as shown below. Therefore, the ranking between *golf* and *camping* depends on which coefficient β is the highest.

More in general, the projection task decomposes word-vectors into a weighted sum of a semantic dimension plus everything else. The projection task assumes that the magnitude and direction of the decomposition of each word-vectors represents the engagement of that word with the semantic dimension. In the example, a higher value of $\beta(\vec{golf},\vec{A})$ with respect to $\beta(\vec{camping},\vec{A})$ can justify the statement "*golf* is more associated with *affluence* than *camping*"[4] The semantic interpretation of the β coefficients (equivalent to the cosines) justifies the final result of the projection task.

The geometric-semantic model of the projection task supersedes analogical models[5], but the legacy of analogies is still present. The pair words used to estimate semantic dimensions – equation (3.1) – are those that may enter an analogy. For example, we can form the following analogy using seed words for the *affluence* dimension: "*rich* is to *poor* what *opulent* is to *indigent*." (Kozlowski et al., 2019, p. 935) Yet, analogical reasoning unnecessarily constricts the projection task. Since analogies are structured around pairs of words, they necessarily focus on binary concepts – that is, concepts with two poles. However, binarism limits the kind of meaning structure we can explore and it forces us to only analyze those kinds of concepts that have opposites. Second, analogies necessarily focus on one concept at the time – *affluence* in the previous example – whereas social scientists often focus on more than one concept and on their intersection (Nelson, 2021). Finally, analogies traditionally involves words, but social scientists may want to explore the meaning of

[4] The semantic interpretation of $\beta(\vec{camping},\vec{A})$ and $\beta(\vec{golf},\vec{A})$ will crucially depend on which pole of the $\vec{A}$ dimension is associated with the positive sign.

[5] The projection task implies the trapezoidal model (Figure 3.2). Indeed, $\vec{camping} - \vec{golf} = (\beta(\vec{camping},\vec{A}) - \beta(\vec{golf},\vec{A}))\cdot\vec{A} + (\alpha_C - \alpha_G)\vec{\varepsilon} = \beta_{C-G}\cdot\vec{A} + \alpha_{C-G}\cdot\vec{\varepsilon}_{C-G}$. When $\alpha_{C-G}\cdot\vec{\varepsilon}_{C-G} = \vec{0}$, this equation reduces to the trapezoidal model of analogical reasoning.

whole documents (Taylor and Stoltz, 2020). However, these limitations of analogical reasoning are not intrinsic to embedding models. Embedding models and their results can be used for more general analyses than the traditional projection task.

3.3. The Simple Algebra of Meaning

The remaining of the paper generalizes the semantic-geometric assumptions of the projection task, moving beyond analogical reasoning. As a theoretical justification of this generalization we formulate a new semantic-geometric model which we name the “simple algebra of meaning”. This model subsumes the model used so far in the projection task and, therefore, the trapezoidal model of analogies as well. Importantly, we are not claiming that this model universally regulates the geometry and meaning of word-vectors. At a minimum, co-appearences of words correlates with their syntactic, morphological and semantic similarities. Therefore, the positions of word-vectors cannot depend exclusively on their meaning: the embeddings of words must depend on a mixture of different linguistic features even in the unfeasible ideal case where embeddings perfectly encode word co-appearences (Levy and Goldberg, 2014b; Pennington et al., 2014). However, in those cases where the projection task actually captures meaning, the simple algebra of meaning (or similar models) must be effective in capturing the relation between meaning and geometry of word-vectors. Therefore, this model is a helpful tool to devise new methodologies, as shown below. Consequently, even if the simple algebra of meaning is certainly wrong, it is at least useful – and this is what we can ask to a model.

Like the projection task, the simple algebra of meaning assumes that the overall meaning of an embedded object (e.g. a word) is represented in the embedding space as a weighted sums of linear semantic dimensions. That is, a semantic dimension is a dimension (direction) of the Euclidean embedding space with a clear semantic meaning attached, such as the $\vec{affluence}$ dimension used in the example above. We can formalize this intuition into an equation:

$$\vec{w}_i = \sum_{j=1}^{N} \beta(\vec{w}_i, \vec{D}_j) \cdot \vec{D}_j + \vec{\varepsilon} \tag{3.3}$$

where $\vec{w}_i$ is the embedding of the object w_i, $\vec{D}_j$ is the j^{th} semantic dimension in the embedding space and the semantic dimensions $\vec{D}_1, \cdots, \vec{D}_N$ are linearly independent (but not necessarily orthogonal) semantic dimensions. Finally, $\vec{\varepsilon}$ is interpreted as a random error term, orthogonal to all semantic dimensions $\vec{D}_1, \cdots, \vec{D}_N$ – we discuss the random component of the model in a bit.

From a mathematical perspective, equation (3.3) is not an assumption: we can write any vector in an Euclidean space as a weighted sum of linearly independent bases. Similarly, the simple algebra of meaning is not a wholly new conception from a semantic perspective neither. The decomposition of words' semantics into semantic features is the linchpin of structural semantics (Hjelmslev, 1961; Greimas, 1983; Lyons, 1995). The new assumption is of a semantic-geometric nature: every human-readable semantic feature corresponds to a direction (base) in the semantic space and the meaning of embedded objects can be decomposed as a weighted sum of such semantic bases. The key point of Equation (3.3) is truly about the translation of embeddings into meaning and vice versa. It also worth noticing that equation (3.3) does *not* assume that there are primitive semantic dimensions exhausting the semantic need of the entire embedding space. That is, no specific set of semantic dimensions $(\vec{D}_1, \vec{D}_2 \ldots \vec{D}_n)$ is intrinsically more relevant than the others. Mathematically, there are infinitely many alternative sets of bases that can satisfy equation (3.3) and none of these sets is more noteworthy than the others. Therefore, the same exact embeddings can be analyzed from different, equally-valid perspectives.

Equation (3.3) provides theoretical ground to justify the current way that the task projection is carried out in the social sciences. Specifically, we can frame the projection task as an estimation problem – where the estimand is the meaning of the embedded object.

Let us consider the two sub-parts of the projection task orderly – equations (3.1) and (3.2). The goal of equation (3.1) is to pinpoint semantic dimensions as precisely as possible given the embeddings. Now, consider how two coupled seed words would differ under equation (3.3), say the words *rich* and *poor*. Whatever their full meaning, we can assume with Kozlowski and colleagues (2019) that these two words only differ along the semantic dimension of *affluence*. Therefore, their word-vectors will be in the following relationship:

$$\vec{rich} = \vec{poor} + \beta_A \vec{A} + \vec{\varepsilon}$$

where $\vec{A}$ represents the unknown *affluence* dimension, $\beta_A = \beta(\vec{rich}, \vec{A}) - \beta(\vec{poor}, \vec{A})$ is a simple scalar and $\vec{\varepsilon}$ is a random error from equation (3.3), which accounts for imprecision in the algorithm creating the embeddings. Therefore, $\vec{rich} - \vec{poor}$ is an (un-normalized) estimation of the semantic dimension $\vec{A}$. Moreover, if we can conjure two or more word-couples that we believe to differ only along the dimension of interest, we may want to combine the information derived from all word-couples to obtain a better estimate. For example, we may believe that the word-couple *opulent* and *indigent* only differ along the *affluence* dimension, exactly as the word-couple *rich* and *poor*. To combine the word-couples, we can assume that the random error in equation (3.3) has an expectation of 0:

$$E[\vec{\varepsilon}\,] = 0 \tag{3.4}$$

Under this assumption, the mean of the word-pairs differences – equation (3.1) – is an unbiased, non-normalized estimator of the semantic dimension of interest $\vec{A}$, justifying its use in empirical research.

Consider now the second part of the projection task, where cosine similarity quantifies the engagement of a concept/word with a semantic dimension $\vec{D}_i$, as in equation (3.2). According to equation (3.3), the word-vector meaning can be decomposed in a weighted sum of different semantic dimensions encompassing the semantic dimension of interest, $\vec{D}_i$. The key observation is that the cosine between a word $\vec{w}$ and the dimension $\vec{D}_i$ is an estimate of the weight $\beta(\vec{w},\vec{D}_i)$ associated with $\vec{D}_i$ in equation (3.3). Specifically, the cosine is proportional to the Ordinary Least Square (OLS) estimate of this weight, implying that the projection task is intrinsically linked to the estimation of linear models. This means that the projection task quantifies how much a given semantic dimension is relevant for the meaning of a word by estimating its decomposition from equation (3.3) with just one dimension $\vec{D}_i$ of interest.

To see this, consider again the word *camping* and its engagement with the concept of *affluence*. We indicate the *affluence* semantic dimension with $\vec{A}$. We assume we have already estimated a set of semantic bases containing the dimension $\vec{A}$ and that all bases have norm equal to 1. Now, consider the decomposition of the word-vector $\vec{camping}$:

$$\vec{camping} = \sum_{\vec{D}_i \in \mathscr{B}} \beta(\vec{camping},\vec{D}_i) \cdot \vec{D}_i + \beta(\vec{camping},\vec{A}) \cdot \vec{A} + \vec{\varepsilon}$$

where the set $\mathscr{B}$ contains all the semantic bases except for $\vec{A}$. If we make the key assumption that the semantic dimension of *affluence* is orthogonal to all other dimensions:

$$\vec{A} \perp \vec{D}_i,\ \forall \vec{D}_i \in \mathscr{B} \tag{3.5}$$

then the OLS estimate of β_A will be:

$$\hat{\beta}(\vec{camping},\vec{A}) = ||\vec{camping}|| \cdot \cos(\vec{camping},\vec{A}) \tag{3.6}$$

Therefore $\cos(\vec{camping},\vec{G})$ can be interpreted as a normalization between -1 and 1 of the OLS estimate. Even more clearly, if the word-vectors are normalized to have norm 1 (as is often the case, Garg et al., 2018; Kozlowski et al., 2019; Jones et al., 2020), then the cosine precisely coincides the OLS estimation of the coefficient. Importantly, even if the cosine is substantially an OLS estimate, the usual statistical framework of linear regression model does not apply here. For example, the use of OLS standard errors to calculate a confidence interval for $\beta(\vec{w},\vec{A})$ would be baseless – alternative procedures, such as bootstrap and subsampling, are appropriate (Antoniak and Mimno, 2018; Kozlowski et al., 2019).

Incidentally, equation (3.6) shows why we can assume that the norm of semantic dimension is 1. Changing the norm of $\vec{A}$ would proportionally change the estimated coefficient $\hat{\beta}(\vec{w},\vec{A})$ for all word-vector in the embedding space. As long as we are interested in knowing which concept is more (less) engaged with a certain pole of the dimension (see for example Garg et al., 2018; Caliskan et al., 2017; Kozlowski et al., 2019; Jones et al., 2020), changing the norm of $\vec{A}$ would not change any substantial conclusion.

3.4. Expanding the Projection Task

The previous section frames the projection task as an estimation of equation (3.3). In fact, we can change our estimation strategies to be less restrictive: we can consider new ways of conceptualizing opposition and we can consider multiple semantic dimensions at the same time. These goals can be achieved by changing how we estimate semantic dimensions – equation (3.7) – and how we estimate semantic coefficients – equation (3.6). Furthermore, if we accept the compositional semantic hypothesis (following Arora et al., 2017; Stoltz and Taylor, 2019), we will be able to apply the projection task to entire documents (as opposed to single words) and decompose the projection of a document word by word. The following paragraphs show how to achieve such expansions of the projection task and apply the new techniques to the estimation of a non-binary semantic dimensions (the *Lovecraftian* dimensions) and to the analysis of a corpus of 30,228 scientific abstracts (Nanni and Fallin, 2021).

3.4.1. Different Kinds of Oppositions

Semantic dimensions assign opposite concepts to opposite directions in the embedding space. For example, when Kozlowski et al. (2019) estimate the *gender* dimension through antonyms such as *he* and *she*, they assign one end of the *gender* dimension to the *masculine* pole and the other end to the *feminine* pole. Under this construction, the statement "*engineer* is more associated with the *masculine* pole than *nurse*" is equivalent to "*engineer* is less associated with the *feminine* pole than *nurse*", which implies a strictly-binary gender system. This kind of logical binarism is intrinsic to the estimation of semantic dimensions as the (mean) difference of two opposite poles – equation (3.1). More deeply, binary logic is intrinsic in the geometric shape of semantic dimensions, which are nothing but lines – and lines are defined as a connection of *two* opposites.

Now, it is difficult to abandon lines altogether because they are extremely convenient from a geometric perspective to decompose spaces: this is the essential intuition of linear algebra, linear models, and equation (3.3). However, it is still possible to escape this strict binarism by using lines along with a different concept of opposition. Specifically, we can use semiotic theory, which suggests that there are two different ways in which terms can be in opposition. On the one hand, there is the contrary relation, where two terms are perceived as opposite alternatives. This is the *masculine-feminine* relationship. On the other hand, there is the relation of contradiction, where a term directly negate another. In this case, the two opposites would be *masculine* and *anti-masculine*. Importantly, the equivalence of *anti-masculine* with *feminine* does not hold in general (see for example Trumbach, 1998).

Jointly, the two different relations of opposition form the semiotic square (Greimas and Courtés, 1982) and map onto different estimation strategies for semantic dimensions. The estimation of a semantic dimension through equation (3.1) corresponds to the contrary relation, where a semantic dimension connects two polarities with their own positive meaning. However, we may not want or not know how to specify a well-defined opposite for a concept of interest. Banally, what is the opposite of *red*? In those cases where we only have one positive polarity, we can still estimate a semantic dimension, but we are exploring a relation of contradiction.

As an example, consider the horror writer H. P. Lovecraft. In his most famous short stories, Lovecraft creates a typical ominous atmosphere populated with ill-intentioned ancient demigods such as Cthulhu. The narrative legacy of this author still permeates current popular culture (Campbell and Miéville, 2016) and one can wonder if word embedding models are able to capture the Lovecraftian aura of words. We have a clear idea about the positive side of the Lovecraftian semantic dimension, but there is no clear opposite. In this case, we can characterize the two polarities of the semantic dimension through contradiction as *Lovecraftian* and *anti-Lovecraftian*.

In practice, how can we estimate the *Lovecraftian* semantic dimension? We can start from few seed words whose Lovecraftian aura is clear, such as *Cthulhu*, and take their mean to estimate the Lovecraftian semantic dimension:

$$\vec{L} = \frac{\sum_{\vec{w}_i \in \mathscr{S}} \vec{w}_i}{n} \tag{3.7}$$

Most Lovecraftian	Coefficient	Least Lovecraftian	Coefficient
Cthulu	0.676	*boosted*	-0.135
HP Lovecraft	0.675	*reiterated*	-0.128
Lovecraft	0.663	*withdrew*	-0.119
Cthulhu Mythos	0.662	*cited*	-0.113
Drizzt	0.657	*highest*	-0.11
Xenogears	0.652	*helped*	-0.108
Planescape Torment	0.645	*consistent*	-0.108
Soul Reaver	0.634	*citing*	-0.107
Shadow Over Innsmouth	0.632	*consistently*	-0.107
Saint Seiya	0.631	*helping*	-0.106

Table 3.1. The most and least Lovecraftian words in the Google News 300-dimensional embedding model and their coefficients.

where $\mathscr{S}$ represents the set of seed words considered – see Appendix G for the complete list of seed words for this example.

Doing this, we are making the new assumption that non-Lovecraftian semantic dimensions are uniformly scattered around the Lovecraftian dimension in the seed words. For example, consider the decomposition of the word-vector $\vec{Cthulhu}$:

$$\vec{Cthulhu} = \beta(\vec{Cthulu}, \vec{L}) \cdot \vec{L} + \vec{\xi} + \vec{\varepsilon} \qquad (3.8)$$

$$E[\vec{\xi}] = 0;\; E[\vec{\varepsilon}] = 0$$

where $\vec{L}$ represents the Lovecraftian dimension and $\vec{\xi}$ represents the semantic contribution of all other dimensions. Assuming that $E[\vec{\xi}] = 0$ among the seed words $\mathscr{S}$, the mean of the seed words will be an unbiased estimate of $\vec{L}$. Even if this assumption may be difficult to meet (and check), it is practically sufficient that the Lovecraftian dimension is preponderant over all other dimensions in the seed words – that is, that the Lovecraftian aura is very clear in the seed words. In symbols, $||\beta(\vec{Cthulu}, \vec{L}) \cdot \vec{L}|| \approx ||\vec{Cthulu}||$. When this is the case, small deviations from the assumptions in equation (3.8) will still result in a reasonable estimate.

To validate the strategy of equation (3.7), we estimate the Lovecraftian dimension on the Google News 300-dimensional embedding model (Mikolov et al., 2013), which is a publicly-available skipgram model estimated on a bilion words from the news. We estimate the Lovecraftian dimension as the mean embedding of seven Lovecraftian seed words (see Appendix G). After estimating the Lovecraftian dimension, we can check if the words mostly engaged with the estimated dimension appear to be attuned to the Lovecraftian aura we are looking for. We are using cosine similarity for this check, following equation (3.2). Table 3.1 reports the ten words with highest projection on the Lovecraftian dimension, excluding seed words. These words are either directly related to Lovecraft and his narrative world or refer to modern-day cultural products (video-games or manga) with an ominous atmosphere.

After this encouraging result, one may wonder which words are least associated with the Lovecraftian dimension. Table 3.1 reports the 10 words least associated with the Lovecraftian semantic dimension, alongside their coefficients.[6] On an intuitive level, these words are neither especially opposed nor attuned to a Lovecraftian atmosphere. In fact, the table shows that the magnitude of the coefficients of the words are not particularly high compared to the magnitude of the coefficients for the positively associated words. We conclude that some words are clearly associated with the positive pole of the Lovecraftian dimensions (having a coefficient higher than 0.6), but no word is clearly associated with its negative pole, since the most negative coefficient is a modest -0.135.

This is not completely unexpected. On the one hand, if there had been a clear opposite of the *Lovecraftian* dimension, our intuition would probably have guided us there. From this perspective, the results reflect our intuition. On the other hand, Word2Vec and the other word embedding models are trained exclusively to find positive association. Thus, they are unable to find antonyms

[6] In Table 3.1, we only consider the most common 10,000 words by number of appearances in the corpus. Considering all words does not change our conclusions substantially, but produces more obscure words whose meaning is less clear.

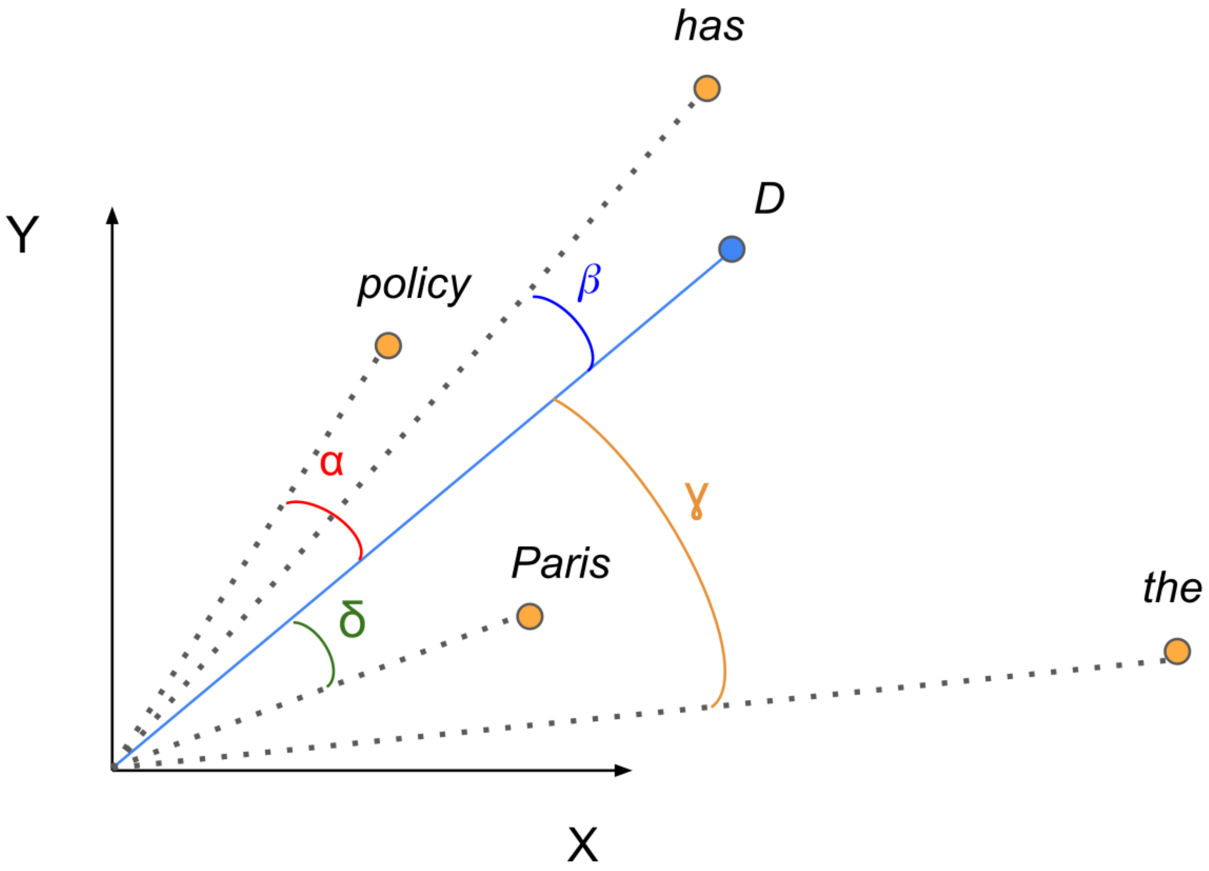
even when they clearly exist – for example, *cold* and *hot*. Rather, antonyms are likely to have positive cosine similarity since antonyms often appear in the same contexts. This is another reason why indicating both poles of a semantic dimension is necessary when the dimension is clearly binary.

Overall, these tests suggest that it is possible to estimate a semantic dimension using only a positive pole, based on the contradiction relation of opposition. This option is relevant when the semantic dimension of interest does not have a clear opposite – arguably, this is the case for most semantic dimensions. However, the projection task will produce different results with respect to the case where a semantic dimension has two contrary poles. Using only a positive pole to estimate a semantic dimension – as in Equation (3.7) – will result in meaningful positive association, but the negative pole of the dimension is unlikely to be particularly poignant. That is, the projection task appears to simply distinguish words that are relevant for a dimension from those that are not. On the other hand, using the relation of contrary as in Equation (3.1), the projection task can distinguish the association of words with either of the contrary poles.

3.4.2. Beyond Words: Representing Text Embedding

Social scientists often analyze collections of documents and are interested in extracting meaning from them (Mohr and Bogdanov, 2013) – as opposed to simple word meaning. In their original formulation, word embedding models focus on the meaning of single words, as the name suggests. However, the same models can also be used to analyze whole documents. In this paragraph, we generalize the algebra of meaning to documents – or, more in general, textual passages. To do this, the key assumption is the compositional semantic hypothesis. On a semantic level, this hypothesis suggests that the meaning of a text is a function of the meaning of the words composing it (Frege, 1948; Baroni et al., 2014). Translating compositionality in the embedding geometry, the

Figure 3.3 A two-dimensional example of the document generative model by (Arora et al., 2017). The vector $\vec{D}$ represents the document. The selection of the next word in the document depends on the cosines (i.e. angles) between the document-vector and the word-vectors. In this toy example, only the word-vectors $\vec{policy}$, $\vec{has}$, $\vec{Paris}$, $\vec{the}$ and their angles with $\vec{D}$ (α, β, γ and δ) are considered.



hypothesis suggests that we can represent a document (and consequently its meaning) as a function of the word-vectors in the document. The final result of this function of word-vectors will be a document-vector inhabiting the same embedding space where word-vectors live. Therefore, the simple algebra of meaning applies to the document-vectors as well: we can measure the engagement of a document with a semantic dimension exactly as we do for words (see for example Taylor and Stoltz, 2021; Stoltz and Taylor, 2021).

Some techniques already exist to calculate the embeddings of documents starting from their words. In particular, Arora et al. (2017) propose an interesting generative model of document composition. Each document is represented as a document-vector in the embedding space. Once the document-vector is placed in the space, it selects words sequentially and independently. The probability of a word to be selected is proportional to the norm of its corresponding word-vector and the cosine similarity of the word-vector with the document-vector (Figure 3.3) – see Arora et al. (2017) and Arseniev-Koehler et al. (2022) for more details on the model. To be more precise, Arora and colleagues also account for the fact that stop words are much more likely to be selected in comparison to all other words. They net-out the influence of such words from the document-vector representation by subtracting from document-vectors' matrix its first principal component. All in all, the document-vectors can be estimated as:

$$\vec{T_i} = \sum_{j=1}^{L_i} \frac{c_j}{L_i} \cdot \vec{w}_j = \mathbf{W}_i \cdot \frac{\vec{c}}{L_i} \tag{3.9}$$

$$\mathbf{T}_d = \mathbf{T} - \mathbf{T} \cdot (\vec{u} \cdot \vec{u}^T) \tag{3.10}$$

where $\vec{T}_{i,}$ is the representation of the i^{th} document, L_i is the number of words in the i^{th} document, $\vec{w}_j$ is the j^{th} word-vector in the i^{th} document, $\mathbf{W}_i$ is a matrix having all such word-vectors as its columns, c_j is a weight associated with $\vec{w}_j$, $\vec{c}$ is a vector containing such weights, $\mathbf{T}$ is the matrix containing $\vec{T_i}$ as its i^{th} row and $\vec{u}$ is the first principal component of $\mathbf{T}$. Notice that the final estimate for the documents' vectors are the rows of the $\mathbf{T}_d$ matrix, not $\mathbf{T}$.

The most important important part of the previous calculation is the weight vector $\vec{c}$, which represents "what is being talked about" (Arora et al., 2017, p. 3) as a vector in the embedding space. This vector contains the weight associated with each word in the document. Substantially, it represents how much each word is important for the overall meaning of the document. Based on their generative model, Arora and colleagues derive as the Maximum a Posteriori (MAP) estimate for this vector. For any word $\vec{w}_j$, its weight for the document is:

$$c_j = \frac{a}{a + p(\vec{w}_j)} \tag{3.11}$$

where c_j is the weight associated with word w_j, $p(\vec{w}_j)$ is the frequency of the word w_j in the corpus of documents and a is essentially a free parameter, which the authors originally fix at 0.0001. Importantly, the weight penalizes frequent words according to the principle that common words are less likely to convey important information (Manning et al., 2008).

However, this is not the only possible document representation. We can change the generative model slightly to obtain that the document-vector estimate is a simple mean of the word-vectors composing the document (see Appendix I). In this case, however, stop words are as important as more rare words, which may not be the most efficient strategy. Finally, it is worth noticing that we can retrieve the Concept Mover Distance (CMD) proposed by Stoltz and Taylor (2019) with another slight modification of the generative model (see Appendix I). The CMD implicitly uses a document-vector representation that is a weighted sum of the word-vectors in the document, where the document-specific weight for the word w_j is calculated as:

$$c_j = \frac{m_i}{L_i||\vec{w}_j||} \tag{3.12}$$

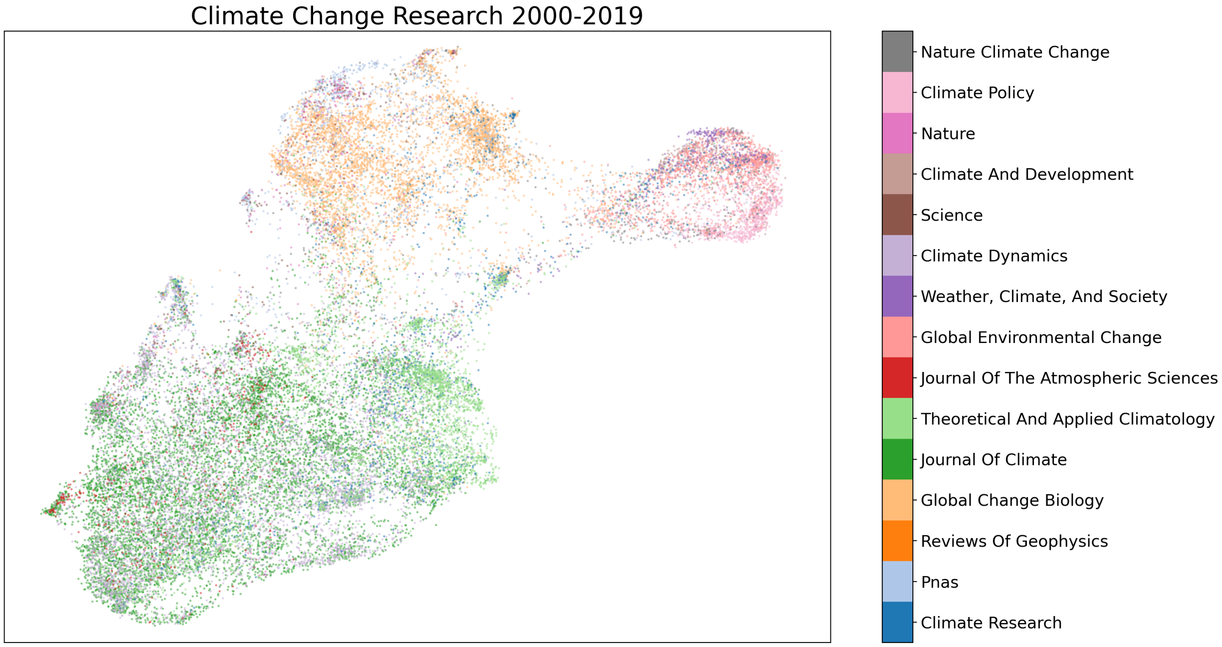
In the equation, L_i is the number of words in the document, m_i is the number of time a word appears in the document and $||\vec{w}_j||$ is the norm of the word-vector $\vec{w}_j$ (see Appendix H for the full derivation). Importantly, the norm of a word usually contains information about its frequency, with more frequent words having larger norms (Ethayarajh et al., 2019). Therefore, the CMD weighting scheme implicitly penalizes the weights of the most frequent words.

Overall, the generative models discussed above represents a document as a weighted sum of the word-vectors in the document.[7] Arora et al. (2017), Stoltz and Taylor (2019) and Taylor and Stoltz (2021) show that the weighting strategies in equations (3.11) and (3.12) produce highly effective representations of short documents, with the ability to uncover non-trivial meaning. It should be noticed, however, that alternative data-driven approaches to the creation of document-vectors exist. In particular, it is possible to mix word embedding with topic models to create data-driven estimation of the basic topics composing the document-vectors (Arora et al., 2018; Dieng et al., 2019). In this case, semantic dimensions are automatically estimated and documents are represented as a mixture of the automatically-generated dimensions (Arseniev-Koehler et al., 2022). This alternative data-driven strategies are powerful data-exploration tools, but their results do not generally align with the theoretical questions of the analyst (Chang et al., 2009). Since we focus on the theory-driven projection task, we ignore these data-driven techniques in this context, even if we emphasize that they are valid techniques for exploratory analysis of textual corpora.

As an application, we can calculate the document-vectors for a corpus of 30,228 abstracts collected from 15 journals and representing the last 20 years of research about climate science. The corpus has been collected and analyzed by Nanni and Fallin (2021), who estimate a 300-dimensional skip-gram embedding model from the corpus. Starting from the estimated word embeddings, we use equations (3.9) and (3.10) to produce a document-vector for each one of the abstracts. The result is visualized in Figure 3.4, where the embeddings of the abstracts are visualized in a 2-dimensional image through to the dimensionality-reduction technique UMAP (McInnes et al., 2018). As the figure shows, the abstracts tend to form three main clusters, which correspond

[7] Incidentally, the overall calculations in equation (3.9) and (3.10) can still be framed as a weighted sum of word-vectors. It is sufficient to subtract the principal component of the document-vectors matrix from the word-vectors matrix and then proceed with the weighted sum of the neat-out word-vectors. In other words, Arora et al. (2017) create another set of (word) embeddings and then proceed with the calculation of the document-vectors using their weighting schema with the newly-calculated embeddings.

Figure 3.4 Document-vectors for abstracts about climate change from 2000 to 2019 Nanni and Fallin (2021) The document-vectors are calculated following equations (3.9), (3.10) and (3.11) (Arora et al., 2017). The original embedding space is 300-dimensional and the image is produced by reducing the original dimensions to two through the uniform manifold approximation algorithm (McInnes et al., 2018). The rightmost cluster contains abstracts from the social sciences, the top cluster contains abstracts from biological sciences, the lower-left cluster contains abstract from meteorological sciences and climatology.

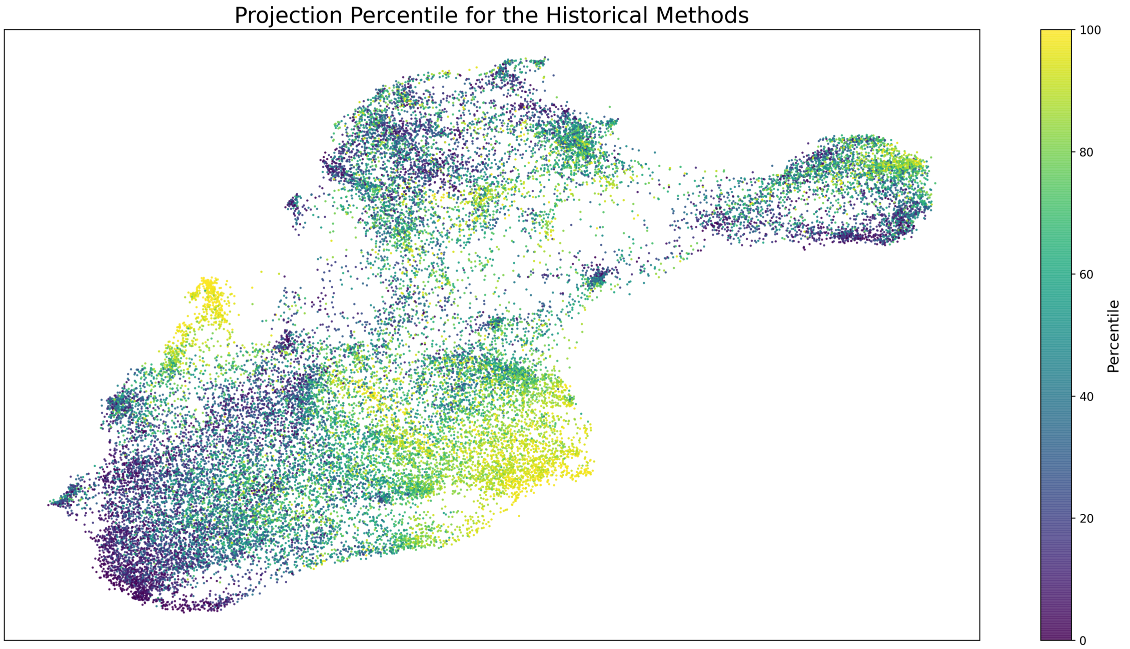


to meteorological sciences, biological sciences and social sciences. Moreover, abstracts from the same journal tend to cluster together. Since, journals are not part of the input for the embedding model, Figure 3.4 shows that the document-vectors are capturing actual semantic properties of the abstracts.

3.4.3. Multi-dimensional Projection

Through cosine similarity – equation 3.2 – we can estimate the engagement of a word (document) with a semantic dimension. Yet, social scientists are often interested in more than one dimension

Figure 3.5 Projection of each abstract onto the semantic dimension of the historical methods. The 100^{th} percentile contains the abstracts that are most engaged with historical methods. The rightmost cluster contains abstracts from the social sciences, the top cluster contains abstracts from biological sciences, the lower-left cluster contains abstract from meteorological sciences and climatology.



at the time, such as multiple conceptualizations of social class (Kozlowski et al., 2019) or intersectional identities (Nelson, 2021). The use of cosine similarity only allows the analysis of one dimension at the time, which is unsatisfactory at best and statistically improper at worst. However, the simple algebra of meaning easily shows how to move beyond this limitation.

Indeed, the connection between OLS regressions and the projection task clarifies when cosine similarity is appropriate as an estimator. If we have more than one semantic dimensions of interest, it is clear from equation (3.5) that cosine similarity assumes that all semantic dimensions are mutually independent (that is, orthogonal), as shown in equation (3.5). Under this assumption, the cosine will be an unbiased estimator of the coefficients from equation (3.3). However, if this assumption does not hold, the cosine estimator will be subject to an omitted variable bias.

It is important to discuss what omitted variables mean in the context of the projection task. In the usual linear model setting, there is a hypothetical model governing the process under analysis: the relevant variables and their relationships (observed or not) are a given. Starting from such variables and a causality structure, one may legitimately argue that all possibly-relevant variables were observed and accounted for or that experimental randomization guarantees that unobserved variables do not bias the estimation. For example, causality models may test the *fixed* relation between two variables – such as smoking and cancer. From an algebraic perspective, we have fixed coefficients corresponding to the real-world effects to be estimated.

The estimation of linear models in the context of the projection task works in an opposite fashion. Semantic dimensions of interest are not given *a priori*, but are chosen by the analyst depending on the research question. From an algebraic perspective, there are infinitely many sets of bases that can produce the observed embedding of words. As we mentioned above, none such set of bases is special in any way. Therefore, there is no underlying causality structure to be inquired – not even on a modeling level. This means that there is no set of primitive semantic dimensions defining all others – only a circle of infinitely many dimensions mutually defining each others (Eco, 1986). It is up to the analyst to pick (and estimate) those dimensions that are relevant for the research question. Effectively, this means that the existence of omitted variables in the context of semantic projection is a matter of debate: it depends on the specific research question. For example, if a research only focuses on gender, ignoring other social dimension (say, class) arguably does not introduce bias in the estimation (see for example Jones et al., 2020). In this case, the analyst assumes that all remaining semantic dimensions are orthogonal to the gender dimension. This is not entirely different from an experimental setup, where the treatment variable is orthogonal to all others by experimental construction.

Yet, there are situations where omitted variable bias is clearly introduced in the estimation of coefficients. For example, consider the work by Kozlwoski and coauthors (2019) once again. The authors contemplate the relation between the different class dimensions they estimate: "The angle between these dimensions can be calculated to capture the similarity between axes of cultural meaning, and it can be evaluated at multiple time points to trace shifts in categorical relations." (Kozlowski et al., 2019, p. 912) In this case, the semantic dimensions of interest are not assumed to be mutually orthogonal and the use of cosine as an estimator will introduce omitted variable bias.

Fortunately, the solution to this problem is clear. Rather than using cosine similarity to estimate only one of the weights of equation (3.3), we should estimate all weights at the same time. This is equivalent to estimating a linear regression model with more than one explanatory variable, which is a trivial task with modern-day computing power. Given a set of semantic dimensions of interest $\mathscr{D}$, we estimate them independently following either equation (3.1) or equation (3.7). Then, we collect their vector representations in the matrix $\mathbf{D}$, where the j^{th} column of $\mathbf{D}$ is the vector representing the j^{th} dimension. From here, we can estimate the projection of a word $\vec{w}$ onto the each of the semantic dimension through the standard OLS model:

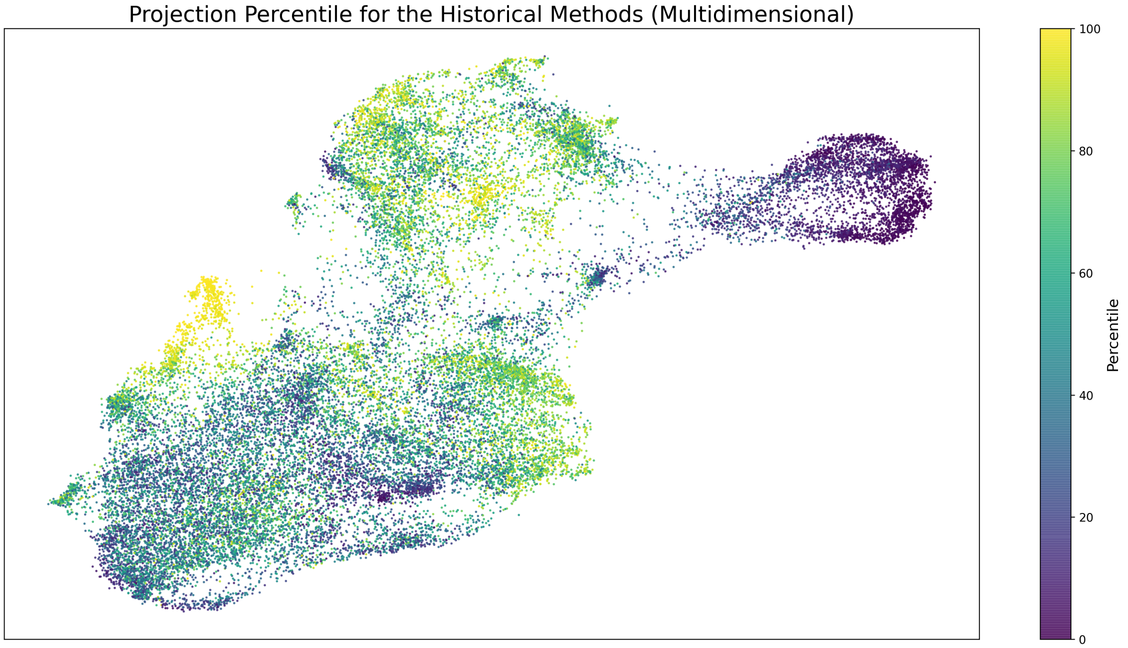
$$\vec{w} = \mathbf{D} \cdot \vec{\beta} + \vec{\varepsilon} \tag{3.13}$$

where the estimated vector $\vec{\beta}$ will contain the coefficients associated with each dimension of interest. Geometrically, this corresponds to projecting a word-vector (or a document-vector) onto the multi-dimensional space created by the semantic dimensions. Therefore, we refer to this OLS model as a “multi-dimensional projection”. This is a straightforward generalization of the projection task to multiple dimensions and it naturally deals with the presence of more than one semantic dimension of interest. Moreover, the multi-dimensional projection assigns to each dimension a unique coefficient. Previous solutions to deal with more than two polarities relied on the contrast of one poles against the grand-mean of all others. In this way, they implicitly assigned multiple coefficients to all poles for each words (see for example Garg et al., 2018; Swinger et al., 2019).

To test multi-dimensional projection, we can use the corpus of climate change abstracts introduced above. Nanni and Fallin (2021) specifically analyze the coupling of methodologies and substantive issues in climate change research over time. They distinguish four methodological macro-families: simulation (prediction of future outcomes based on computational models), historical methods (paleoclimatological methodologies looking at the distant pas, such as dendrochronology), document analysis (especially, regarding policies) and field methods (the collection and analysis of data on the field). Similarly, they distinguish four substantive issues: earth (earth and soil sciences), water (oceanography and marine sciences), air (atmospheric and meteorological sciences), and fire (fire science). Following the authors, we can estimate a semantic dimension per methodology/substantive issues using equation (3.7); the seed words for each dimension are listed in Appendices J.

As a test of the difference between multi-dimensional and uni-dimensiona projections, let us consider the historical methods family. We can use cosine similarity with the abstracts’ document-vectors to quantify the level of engagement with *Historical Methods*. In this case, the sheer magnitude of the cosine for each abstract is not relevant: we are interested in which abstracts are more (less) engaged with the methodology and their positions in the embedding space. Therefore, we concentrate on the percentile of the coefficient distribution. That is, we focus on the abstracts’ engagement with historical methods compared to all other abstracts. Figure 3.5 shows the percentile

Figure 3.6 Multi-dimensional projection of each abstract onto the semantic dimension of the *Historical Methods*. The 100^{th} percentile contains the abstracts that are most engaged with *Historical Methods*. The rightmost cluster contains abstracts from the social sciences, the top cluster contains abstracts from biological sciences, the lower-left cluster contains abstract from meteorological sciences and climatology.



of engagement of each abstract with historical methods through a 2-dimensional UMAP reduction of the embedding. To be clear, the analysis was conducted on the full 300-dimensional embedding space and the 2-dimensional reduction is only a way to visualize the results. From the figure, we can see that the abstracts engaging the most with historical methodologies tend to group around two different points within the metereology cluster. However, some abstracts in biology and the social sciences also appear to use historical methodologies – a surprising result further explored below.

Since we are explicitly considering four different macro-methodological families (*Simulation*, *Field Methods*, *Historical Methods* and *Document Analysis*), the cosine similarity will likely introduce omitted variable bias in our analysis, since the methodological dimensions are not mutually orthogonal. To get more reliable results, we can swap the cosine similarity with a full OLS model per each abstract, as in equation (3.13). In this way, we account for all methodological dimensions at the same time. The models can be estimated by any software that can estimate standard OLS regressions. As above, we are only interested in the historical dimension and in the relative engagement of abstracts with it. Thus, we construct the coefficient distribution for the historical dimension and transform it in the percentile distribution. Figure 3.6 shows the results of these operations and it is directly comparable to Figure 3.5.

ID	Title	DOI	Projection
1	The Continuum of hydroclimate variability in Western North America during the last millennium	10.1175/JCLI-D-11-00732.1	Multi-dimensional
2	Leap frog in slow motion	10.1111/gcb.13881	Multi-dimensional
3	Tree-ring estimates of Pacific decadal climate variability	10.1007/s003820100177	Multi-dimensional
4	Early Last Interglacial Greenland Ice Sheet melting	10.1007/s00382-013-1935-1	Multi-dimensional
5	Eigen analysis of tree-ring records	10.1007/s00704-011-0468-y	Multi-dimensional
6	Reconstructing the NH mean temperature	10.1175/2010JCLI3646.1	Uni-dimensiona
7	On the sensitivity of field reconstruction and prediction using empirical orthogonal functions	10.1175/JCLI-D-13-00089.1	Uni-dimensiona
8	Sensitivity of satellite-derived tropospheric temperature trends	10.1175/JCLI-D-15-0744.1	Uni-dimensiona
9	Intercomparison of machine learning methods for statistical downscaling	10.1007/s00704-018-2613-3	Uni-dimensiona
10	Updated precipitation reconstruction (AD 1482–2012) for Huashan, north-central China	10.1007/s00704-015-1387-0	Uni-dimensiona

Table 3.2. The 5 abstracts closest to the 99^{th} percentile for the uni-dimensiona and multi-dimensional projection models.

From the figure, it appears that the multi-dimensional projection of the abstracts onto the historical dimension is only partially consistent with the uni-dimensional projection shown in Figure 3.5. Indeed, the two projections have a Spearman rank correlation of roughly 0.5. This implies that the multi-dimensional projection use textual information differently. In particular, abstracts from the social science cluster do not engage with historical methods according to the multi-dimensional projection.

It is not easy to validate semantic projections in this case, since there is no clear ground-truth that can be used. A possible validation strategy is to manually assess whether high-percentile abstracts from either projection methods engage with historical methodologies. For example, Table 3.2 reports the titles of the five abstracts from the 99.008^{th} percentile to the 98.994^{th} percentile for both projections: these ten abstracts should be highly engaged with historical methodologies. At first sight, the multi-dimensional projection appear more reliable overall. To a human reader, only abstracts 6 and 10 appear relevant in this context among those singled out by the uni-dimensional projection. On the other hand, all abstracts found by the multi-dimensional projection appear relevant. Yet, the uni-dimensiona projection still classifies the five abstracts found by the multi-dimensional projection as highly relevant (from the 92^{th} percentile to the 99.5^{th} percentile), while the multi-dimensional projection agree with the uni-dimensiona projection in classifying the irrelevant abstracts 7 and 9 from the table as highly topical for historical methods (both abstracts are over the 90^{th} percentile in the multi-dimensional distribution). Therefore, multi-dimensional projections appear only marginally more precise than uni-dimensiona projections in this very small sample. In general, both methods appear to pick up on highly relevant abstracts, even if with some imprecisions.

However, it is worth exploring the main difference between the two results: the possible engagement of social sciences with historical methodologies. According to uni-dimensiona projections, a sizeable number of abstracts from the social science clusters appear to engage with such methodologies (Figure 3.5); multi-dimensional projections show no signs of such engagement (see Figure 3.6). We can examine this difference by manually checking those abstracts from the social sciences clusters that are above the 90^{th} percentile in the uni-dimensiona distribution. There are 178 such abstracts, amounting to roughly 5% of the abstracts in the social sciences cluster. After a manual check, very few of these abstracts appear relevant for paleoclimatological methods: most of them appear to study policy implications (for example, "Linguistic analysis of IPCC summaries for policymakers and associated coverage") or public perception (for example, "Public microblogging on climate change: One year of Twitter worldwide").[8] Noticeably, the two projection methods disagree on this set of abstracts. On the multi-dimensional distribution, such abstracts have a median percentile of 11.4: they are among the least relevant abstracts for historical methods. This analysis confirms that multi-dimensional projections are more precise in this case. Importantly, the multi-dimensional projection prevents the analyst from drawing the apparently-erroneous conclusion that the social sciences sizably engage with paleoclimate methods.

3.4.3.1. Application: The Institutionalization of Climate Change Research. On a more substantive level, we can use the multi-dimensional projections of each abstracts onto the methodological and substantial dimensions to analyze the change of climate change research over time. The coupling of methodology and substantial issues is the linchpin of disciplinary scientific identity

[8] The multi-dimensional projection correctly classifies two abstracts from the social science cluster as highly relevant for historical methods. The two abstracts are titled "Broadening the spatial applicability of paleoclimate information" and "Paleoclimate histories improve access and sustainability in index insurance programs". They appear to be the only abstracts relevant for paleoclimatolgy among those manually checked.

Figure 3.7 Spearman correlation coefficient for each methodology/substantial issue combination for the abstracts from the year 2000. Asterisks indicate statistical significance at the 0.9 alpha-level as estimated through the bootstrap procedure. A cross indicates that the coefficient is not robust in the bootstrap repetitions.

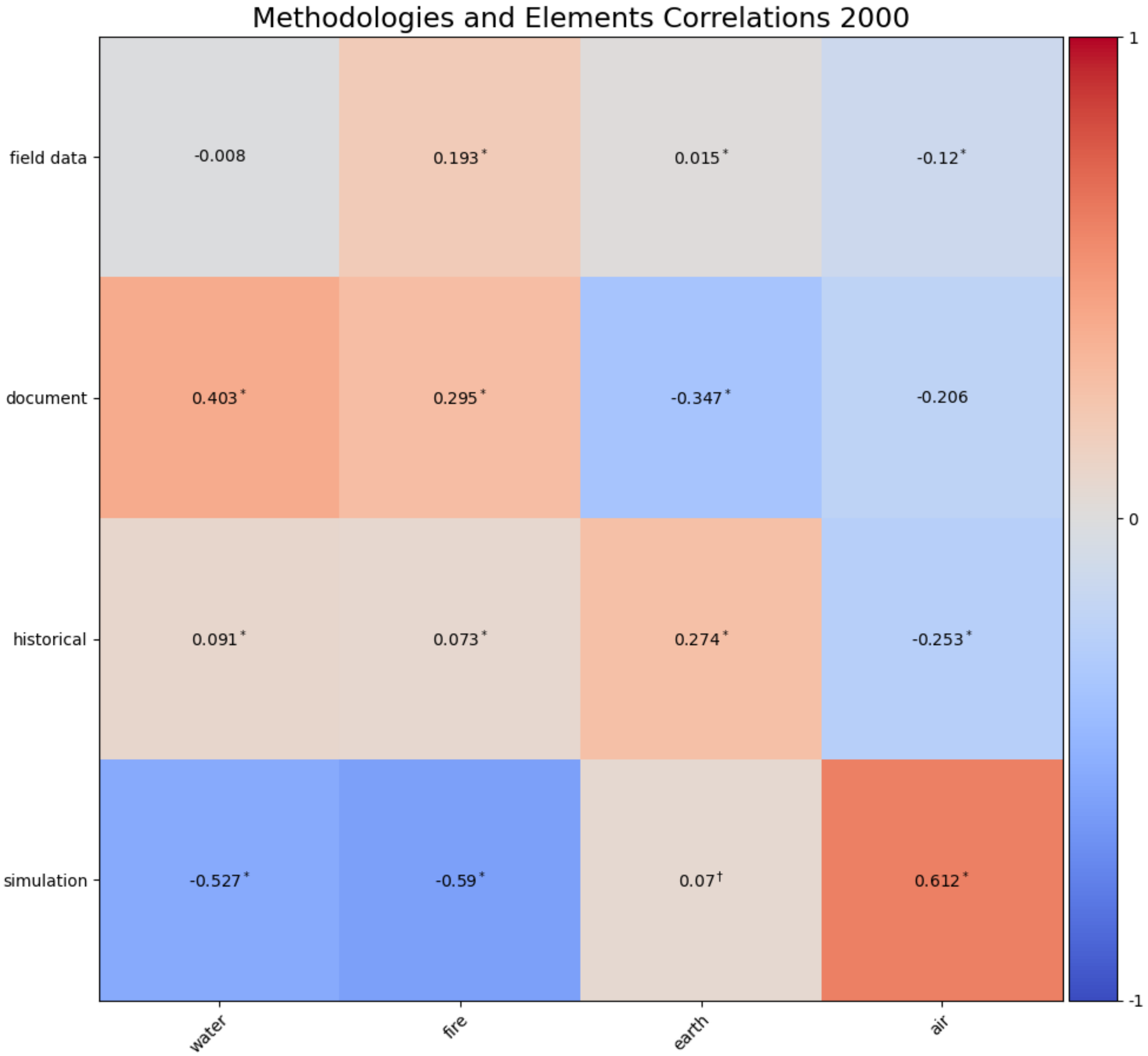
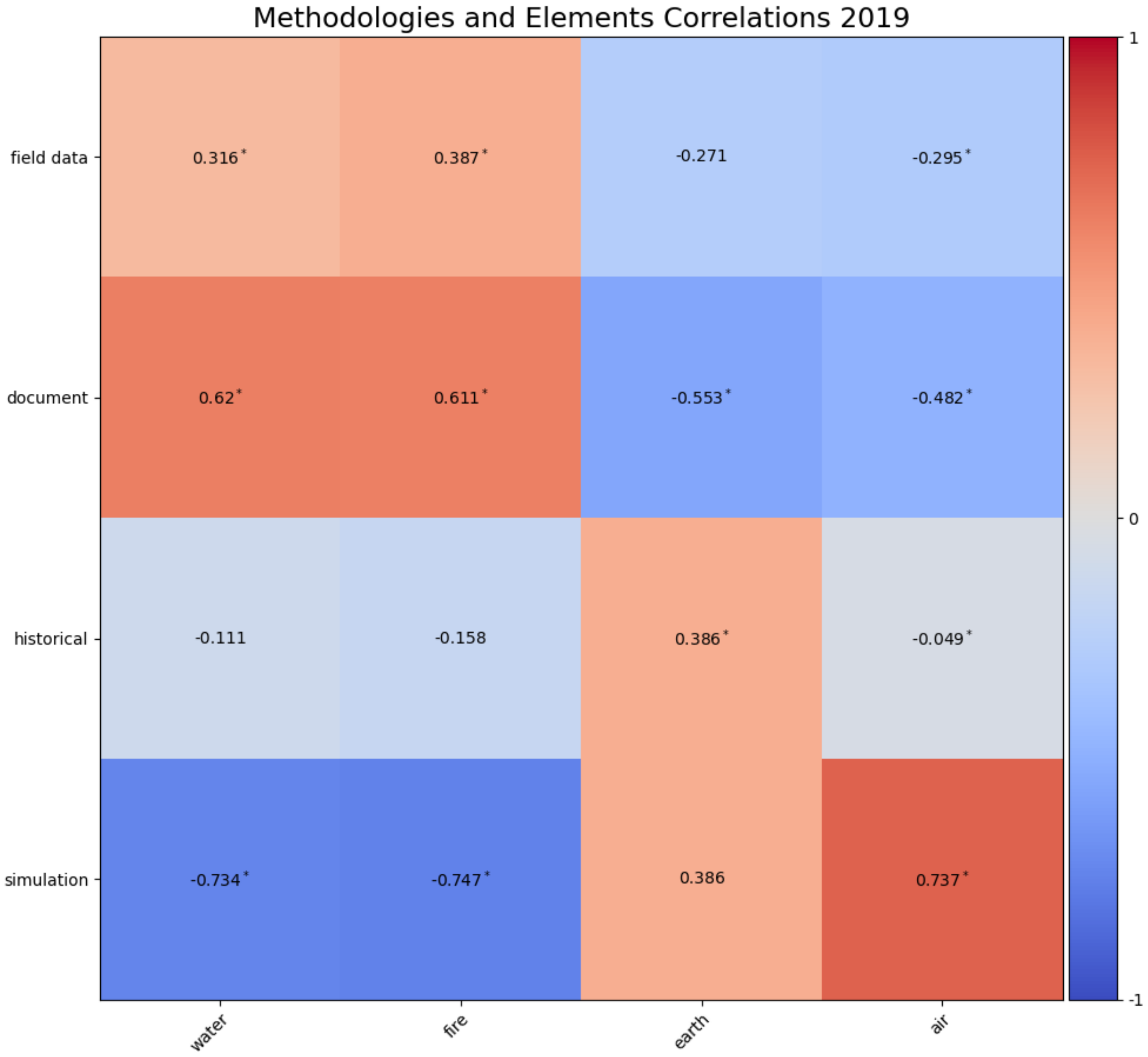


Figure 3.8 Spearman correlation coefficient for each methodology/substantial issue combination for the abstracts from the year 2019. Asterisks indicate statistical significance at the 0.9 alpha-level as estimated through the bootstrap procedure. A cross indicates that the coefficient may is not robust in the bootstrap repetitions.



(Breslau, 2003; Camic and Xie, 1994). Therefore, we can analyze the coupling of methodological and substantial dimensions in the abstracts collected by Nanni and Fallin (2021) to observe the formation of climate change scientific identity from 2000 to 2019. These were very important years for climate change research, with a remarkable increase of journals, conferences and funding opportunities centered around this topic (Haunschild et al., 2016; Stanhill, 2001). We suspect that alongside the increased interest in the topic, the field also experienced a substantial institutionalization of research practices, which limits the space of possible (and probable) research (Bourdieu, 1975).

To analyze this issue, we first project each abstract onto the methodological dimensions; then, we project the abstracts onto the substantive dimensions. We analyze the Spearman correlation of the methodological and substantive projections of each abstract. This way, we can quantitatively address how typical is the coupling of specific methodologies and issue – say, the coupling of *Fire* with *Simulation*. Figures 3.7 and 3.8 show the correlation of each method-issue in the year 2000 and 2019, respectively. In both years, we see similar patterns in research strategies: *Field* and *Document* methods are used to study *Water* and *Fire*, while *Air* is studied through *Simulation*, and *Earth* is studied through *Historical* methods. On the flip side, the analysis also shows which method-issue couplings are rare: *Document* is rarely paired with *Earth* or *Air*, while *Simulation* is rarely paired *Water* or *Fire*. As suspected, the coupling between methodologies and issues is becoming tighter during time. Correlations with magnitude above 0.1 (both positive and negative directions) in year 2000, increased in magnitude in 2019 – with one exception being the *Air-Historical* correlation. The coupling between methodologies and substantial issues became overall more predictable, at least at this high level, signaling an institutionalization of the field and a consequent restriction of probable research practices. This analysis benefits greatly from the greatest precision that multi-dimensional projection technique allows.

	Strength	Fine	Problem	Mathematical	Authority	Sociability	Coding	Housekeeping
Mean	0.055	0.014	-0.023	0.089	0.060	-0.022	-0.044	0.034
St. Dev.	0.529	0.330	0.377	0.431	0.516	0.371	0.511	0.497
Min	-6.774	-4.739	-5.005	-4.954	-5.032	-4.752	-4.633	-7.434
25%	-0.154	-0.141	-0.167	-0.138	-0.193	-0.198	-0.293	-0.210
50%	-0.022	-0.005	0.017	-0.005	0.018	-0.001	-0.118	-0.014
75%	0.150	0.143	0.165	0.241	0.304	0.143	0.059	0.206
Max	11.225	7.652	3.674	6.482	7.311	6.738	7.262	11.605

Table 3.3. Summary statistics for the projections of the 2010 Burning Glass job ads embeddings onto 8 skill-semantic dimensions

3.4.3.2. Application: Gendered Skills. To show the flexibility of the projection task, we will embed job ads data to track down the different kind of skills required in occupations that have disproportional amount of women (men). Stereotypically, women and men have different kinds of skills: women are supposedly more sociable and caring whereas men are more technically-gifted and authoritative (Ridgeway, 2011). These stereotypes emerge in occupational segregation: women are disproportionately represented in occupations that require workers to use stereotypical female skills, such as sociability or finger dexterity. At the same time, the presence of men in an occupation is associated with manual labor and technical skills (Levanon and Grusky, 2016). Besides the workplace, women and men frame themselves following such stereotypes (Cech, 2013) and are under-rewarded when going against the stereotypes (Correll et al., 2007).

In this application, we analyze a dataset of more than 6 million job openings advertisements in the U.S. from the year 2010. The data analyzed here have been collected and structured by Burning Glass Technologies (BG), which elaborate raw job post texts into structured data containing occupation, location, skill requirements, and more. The objective is to determine what kind of skills are required in the advertisements of those positions that have a disproportionate amount of female (male) workers. While Levanon and Grusky (2016) have analyzed the actual skills used on the job using O*NET data, we analyze the skills required for a position according to its own advertisement. Regardless of the actual contents of a job, the requirement of highly-gendered skills in the job advertisement can result into a gate-keeping effect against gender desegregation.

Technically, we examine each job openings as a "bag-of-skills", that is a collection of required skills. For example, a job opening may require "team-work, Python programming, statistical modeling". The skills have the traditional Zipf distribution of words in natural language: few skills are very popular and many appear only rarely. We substitute skills that appear 5 or less times with the token *RARE SKILL*, for a total of 12,965 unique skills. We use a 300-dimensional Word2Vec model (Mikolov et al., 2013) to embed each skill, considering every ads as one sentence.[9] Starting from the embedding of each skill, we use the technique by Arora et al. (2017) to create the embedding of each job advertisement in the data, where the job advertisement are considered as documents.

After producing the embedding for each job advertisement, we build 8 semantic/skill dimensions that we believe to be relevant for the gender composition of occupations based on the rich literature about stereotypically-female or stereotypically-male skills (Ridgeway, 2011). In particular, we draw upon Levanon and Grusky (2016) to build the *Upper-Body Strength*, *Fine Motor Skill*, *Problem Solving*, *Mathematical Skills*, *Authority*, *Sociability* and *Technical Coding Skill* dimensions. We also create a *Housekeeping* dimension that represents the extension of housekeeping duties in the workplace – for example, the skill of *Laundry*, or *Bathing*. All of the seed words are listed in Appendix K.

We project the job advertisements on the 8 skill dimensions using multi-dimensional projections. Table 3.3 provides the summary statistics for the projections of the job advertisements. Validating the results is easier than in the previous application because we can use federal US data

[9] We set the context window parameter of the Word2Vec algorithm to 100, bigger than any actual ads in the data. This way, we disregard the order in which the skills of each opening are presented in the data.

	Coefficient	**[0.025**	**0.975]**		
Upper-Body Strength	-0.9691	-0.971	-0.967		
Fine Motor Skill	0.6531	0.650	0.656		
Problem Solving	-0.5096	-0.513	-0.507	**Fixed Effects:**	No
Mathematical Skills	-0.4613	-0.464	-0.459	R^2**:**	0.320
Authority	-0.0191	-0.021	-0.017	**N:**	6,195,940
Sociability	0.3977	0.395	0.400		
Technical Coding Skills	-0.3434	-0.345	-0.341		
Housekeeping	0.6033	0.601	0.605		

Table 3.4. OLS Regression of the log-odds ratio of women to men on the projections of the 2010 Burning Glass job ads embeddings onto 8 skill-semantic dimensions. Each job ads is mapped onto a SOC occupational category, whose women-to-men proportion is taken from the 2010 American Community Survey data.

to assess the gender composition of the job openings. In particular, BG maps each advertisement to a 6-digit Standard Occupational Classification (SOC) occupations – a code used in federal surveys. Using this code, we can obtain the gender proportion for each advertised occupation from the 2010 American Community Survey (ACS) data. Naturally, ACS provides a national-level gender proportion for each SOC code, which will not correspond to the organization-level gender proportion for the advertised jobs. However, ACS data still provides strong guidance on the gender proportion we expect on average for each opening.

Therefore, as a validation, we assess the association of each dimensions' projections with women proportion in the advertised occupation (according to ACS). First, we transform women proportion into an unbounded variable by applying a log-odds ratio transformation to the proportion. Then, we estimate two simple OLS regressions where the dependent variable is the log-odds ratio of women proportion. In the first regression model, we simply regress the dependent variable on the skill-dimensions' projections of the advertisements (and an intercept):

$$LOR(w_i) = \alpha + \sum_{j=1}^{8} \beta_i \cdot Skill_{i,j}$$

	Coefficient	**[0.025**	**0.975]**		
Upper-Body Strength	-0.1288	-0.130	-0.127		
Fine Motor Skill	0.1921	0.190	0.194		
Problem Solving	-0.0470	-0.049	-0.045	**Fixed Effects:**	Yes
Mathematical Skills	-0.1387	-0.140	-0.137	R^2**:**	0.779
Authority	0.0383	-0.037	-0.039	**N:**	6,195,940
Sociability	0.1233	0.122	0.125		
Technical Coding Skills	-0.1114	-0.113	-0.110		
Housekeeping	0.1378	0.136	0.139		

Table 3.5. OLS Regression of the log-odds ratio of women to men on the projections of the 2010 Burning Glass job ads embeddings onto 8 skill-semantic dimensions. Each job ads is mapped onto a SOC occupational category, whose women-to-men proportion is taken from the 2010 American Community Survey data. The model contains fixed-effects controls for the effects of 22 major SOC occupational categories.

where $LOR(w_i)$ is the log-odds ratio of women proportion for the occupation of the i^{th} job advertisement, $Skill_{i,j}$ is the projection of the i^{th} job advertisement onto the j^{th} skill dimension.Table 3.4 shows the resulting coefficients for each skill dimension. As expected, *Upper-Body Strength*, *Problem Solving*, and *Mathematical Skills* are strongly associated with an under-representation of women. On the other hand, *Fine Motor Skill*, *Housekeeping*, and *Sociability* are strongly associated with an over-representation of women. Finally, *Authority* has a negative association with women's representation, but the magnitude of its effect is an order of magnitude smaller than any other effect. These results are in line with our expectations as well as the analysis by Levanon and Grusky (2016) using O*NET data.

In the second regression model, we add fixed effects to account for unobserved heterogeneity within macro-occupational categories. That is, the first two digits of the SOC codes represent the occupational macro-family of each 6-digit, detailed occupation. For example, the General Dentist detailed occupation is nested within the Healthcare Practitioner macro-occupation. In the second model, we add a fixed-effect control to account for the effect of 2-digit macro-occupations:

$$LOR(w_i) = \alpha_k + \sum_{j=1}^{8} \beta_i \cdot Skill_{i,j}$$

where α_k is an intercept specific to the k^{th} macro-occupation. Table 3.5 shows the coefficients for the 8 skill dimensions in this fixed-effects specification. The coefficients follow the same general pattern as the coefficients in the first model, even if the magnitude of the effects is generally hampered down. Therefore, within macro-occupational categories, those detailed occupations that require more steroretypically-male (-female) skills have a lower proportion of women (men). That is, the association of skills with gender is reproduced *within* each macro-occupation in a fractal manner (Levanon and Grusky, 2016). The only partial exception is the *Authority* dimension, which is now positively associated with women representation. However, the magnitude of this effect is still considerably smaller than all other effects.

Overall, this application of the projection task to the analysis of job advertisements shows the flexibility and robustness of embedding techniques to analyze social scientific questions. This application is simply a confirmation that the proposed techniques capture actual trends in the data. For this reason, we stress that the results substantially replicate our expectations and previous results obtained with totally different methods and data (Levanon and Grusky, 2016). Yet, the use of the projection task on this kind of data can produce new, interesting results. It can be used to explore the time- and spatial-variation in skill requirements for the same occupation and its implication for gender segregation. On a practical level, these techniques can also be used to set up experiments on skill perception and reward (Correll et al., 2007) as well as actually creating job advertisements that are perceived as more welcoming for groups that are under-represented in an occupation.

3.4.4. Decomposing Textual Projections

One of the assumption behind the document representations discussed in section 3.4.2 is that the represented document engages with semantic dimensions persistently throughout its length. Naturally, this assumption may be problematic for longer documents, which often engage with different topics through their length. When that is the case, a document is best analyzed by splitting it into sub-sections. Fortunately, it is possible check the engagement persistence of a document: it is possible to decompose a document's engagement with a dimension as a sum of the engagement of its words. This is possible because both the creation of a document-vector – equations (3.9) and (3.10) or (3.12) – and the OLS estimation of a linear model – equations (3.13) – are linear functions. Through this decomposition, it is possible to check the evolution of the engagement of a document with a semantic dimension as the document progresses from its beginning to its end.

Word	*The*	*Paris*	*Agreement*	*has*	*provided*	*a*	*new*	*framework*	*for*	*climate*	*policy*
Weight	0.0	0.004	0.002	0.0	0.003	0.0	0.001	0.002	0.0	0.0	0.001
Weight %	0%	1.41%	0.65%	0.12%	0.96%	0.01%	0.32%	0.59%	0.03%	0.04%	0.35%
Document Proj.	-0.01	1.79	0.48	0.08	0.07	0.1	-0.4	0.98	0.34	0.63	1.58
Historical Proj.	0.14	-1.15	-0.31	0.13	0.3	0.25	0.81	-0.43	0.3	-0.03	-0.9
Simulation Proj.	-0.02	-0.11	0.33	0.11	-0.26	-0.12	-0.14	-0.19	-0.06	0.25	-0.44
Field Method Proj.	-0.23	-0.13	-0.05	-0.27	0.28	-0.20	0.18	0.34	-0.57	-0.52	0.16

Table 3.6. The projection decomposition for the first sentence in the abstract *Climate Clubs and the Macro-Economic Benefits of International Cooperation on Climate Policy* (Paroussos et al., 2019).

Let us consider the document representation shown in equations (3.9). Let $\vec{T}_d$ be the vector representing a document; that is, a row in the matrix T_d from equations (3.10). We are interested in the projection of $\vec{T}_d$ on the semantic dimensions $\vec{D}_1, \vec{D}_2 \ldots \vec{D}_n$. Following equation (3.13), we estimate such projection through the OLS regression:

$$\vec{T}_d = \mathbf{D} \cdot \vec{\beta} + \vec{\varepsilon}$$

where $\mathbf{D}$ is a matrix containing the n semantic dimensions as columns while $\vec{\beta}$ is a $n \times 1$ vector containing the n projection coefficients as its elements. Following the standard OLS estimation, we estimate the coefficients of the model as:

$$\begin{aligned}
\hat{\vec{\beta}} =& (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D} \cdot \vec{T}_d = \mathbf{M} \cdot \vec{T}_d = \\
& \mathbf{M} \cdot (\mathbf{W}_i \cdot \vec{c} - (\vec{u} \cdot \vec{u}^T) \cdot \mathbf{W}_i \cdot \vec{c}) = \\
& \mathbf{M} \cdot (\mathbf{I} - \vec{u} \cdot \vec{u}^T) \cdot \mathbf{W}_i \cdot \vec{c} = \mathbf{M}' \cdot \mathbf{W}_i \cdot \vec{c}
\end{aligned} \tag{3.14}$$

where $\mathbf{M} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}$, $\mathbf{I}$ is the identity matrix, and $\mathbf{M}' = \mathbf{M} \cdot (\mathbf{I} - \vec{u} \cdot \vec{u}^T)$. Notice that in we use $\vec{T}_d = \mathbf{W}_i \cdot \vec{c} - (\vec{u} \cdot \vec{u}^T) \cdot \mathbf{W}_i \cdot \vec{c}$ from equations (3.11) and (3.10)[10], where $\mathbf{W}_i$ is a matrix containing the word-vectors of the words composing the document as columns and $\vec{c}$ is the vectors of weights associated with such word-vectors, from equation (3.11).

In short, equation (3.14) shows that the coefficients of a documents are the weighted sum of the coefficients of its words, $\mathbf{M}' \cdot \mathbf{W}_i$ in the equation. Specifically, the equation shows this is the case for the document representation proposed by Arora and colleagues (2017), but it is simple to show that a similar decomposition hold when the document representation is a simpler weighted

[10] Technically, we use the transpose of the matrix $\mathbf{T}_d$ from equation (3.10), since we represent $\vec{T}_d$ in equation (3.14) as a column vector, as usual when expressing linear models in matrix form.

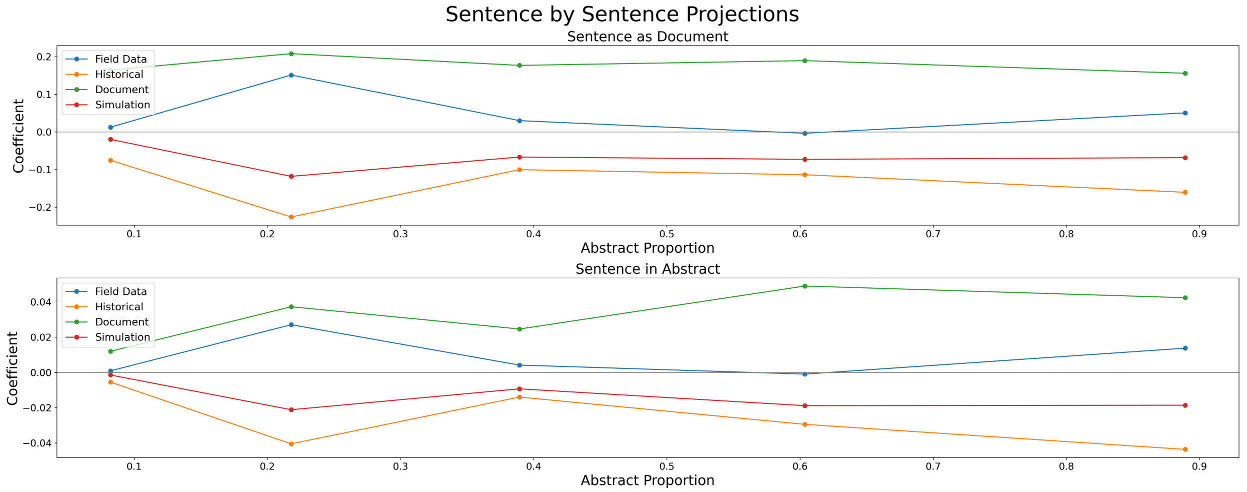
sum of the word-vectors composing it, such as the CMD representation proposed by (Stoltz and Taylor, 2019).[11] As a result, we can examine which words and sentences mostly influence the the overall projection of a document and how the projections on different semantic dimensions change through the document.

As an example, we will consider the abstract *Climate Clubs and the Macro-Economic Benefits of International Cooperation on Climate Policy* (Paroussos et al., 2019), from the corpus of abstracts introduced above. As a first step, we estimate model (3.13) for this abstract, where the semantic dimensions of interest are the four methodological dimensions discussed above (*Simulation*, *Historical Methods*, *Field Methods* and *Document Analysis*). Since this research analyzes the Paris agreement from a macro-economic perspective, it is a clear case of policy analysis and the multi-dimensional projection should indicate that the abstract mostly engage with the *Document* semantic dimension. This is indeed the case. The multi-dimensional projection of the abstract suggests that the abstract mostly engage with the *Document Analysis* methodology ($\beta = 0.17$). The second most relevant methodology is *Simulation* ($\beta = 0.04$), while the coefficients for the remaining methodological families (*Historical Methods* and *Field Methods*) are negative. Importantly, all semantic dimensions must have the same norm to make coefficient comparisons meaningful: the norms work as a scale factor and we must ensure that all compared coefficients are on the same scale. In this case, we fix the dimensions' norm to 1 for all dimensions.

As a first step in the decomposition of the projection, let us analyzes which words are most influential for the overall projection of the abstract. Consider for example, the first sentence of the abstract. This sentence contains 12 words, whose weights and projection are shown in Table 3.4.4. Following equation (3.14), we can calculate the contribution of this sentence to the overall abstract projections as a weighted sum of its words' projections on each dimension. Both the

[11] In that case $\vec{\beta} = \mathbf{M} \cdot \vec{T}_d = \mathbf{M} \cdot (\mathbf{W}_i \cdot \vec{c}) = (\mathbf{M} \cdot \mathbf{W}_i) \cdot \vec{c}$.

Figure 3.9 Sentence by sentence decomposition of the projections of the abstracts *Climate Clubs and the Macro-Economic Benefits of International Cooperation on Climate Policy* (Paroussos et al., 2019). The top panel shows the projections of each sentence treating the sentences as separate documents. The bottom panel shows the projections of each sentence within the abstract: the sum of these projections equals the overall projections of the entire abstract.



weights and the words' projections are shown in Table 3.4.4. From this calculation, the coefficient of this sentence for the *Document* dimension is the highest (0.01) and all other coefficients are either negative or negligible. Specifically, Table 3.4.4 shows that two words (*Paris* and *policy*) mainly determine this result: both words have relatively high weights and are strongly associated with the *Document* dimension, making the *Document* projection of this portion of text higher than the other projections.

In fact, equation (3.14) shows that we can decompose the overall coefficients of an abstract as a sum of the projections of its individual sentences. The bottom panel of Figure 3.9 shows the coefficients of each sentence on all semantic dimensions of interest. Once again, the overall coefficients of the abstract is simply a sum of the sentences' coefficients shown in the plot – so that this is a proper decomposition of the overall coefficients. However, using equations (3.11) and (3.10), the coefficients of a sentence implicitly depends on how long the sentence is vis-à-vis the length of the entire document. Ultimately, the coefficients we see on the bottom panel of Figure 3.9 are weighted by the length of the sentence.[12] We can avoid this issue by treating each sentence as a separate document and recalculating the coefficients – it is sufficient to rescale the weights to do such calculation. The top panel of Figure 3.9 shows the resulting coefficients for each sentence, independent of their length. Overall, we can see that the *Document* dimension is the most important dimension in each single sentence and that the relative importance of each dimensions is roughly constant through the document.

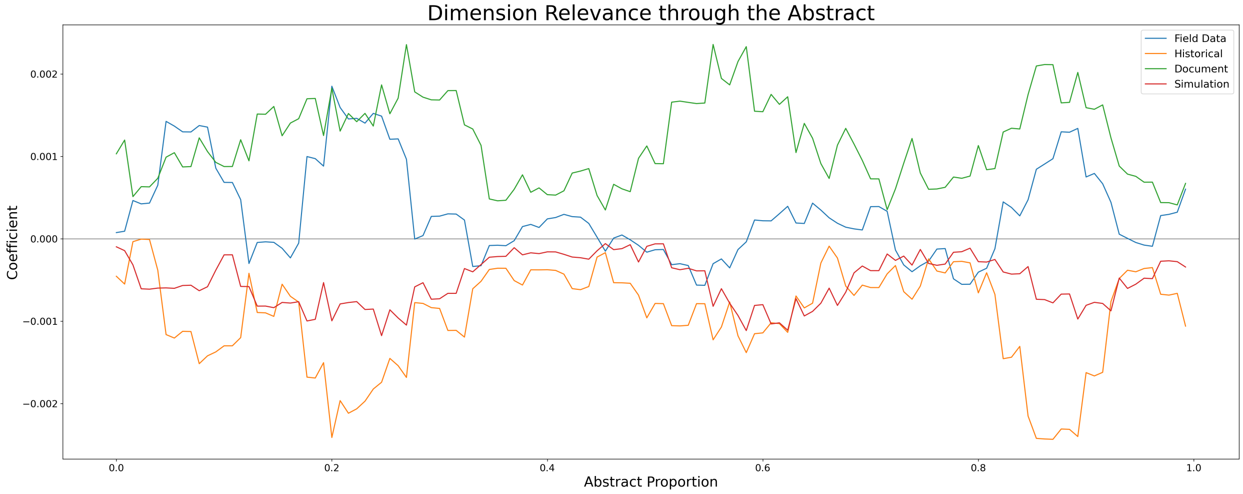
The decomposition can be even more granular by analyzing at the projections of a moving window of words. Figure 3.10 shows the coefficients of a 10-words-long moving window on the same abstract. As before, the overall coefficients of the abstract can be retrieved by summing the points along the plot[13]. The figure shows that the coefficients values for all dimensions of interest have peaks and depressions along the abstracts, revealing more variability than Figure 3.9. However, the relative importance of the coefficients still remains consistent through the abstract, with the *Document* dimension clearly being the dimension with which the abstract mostly engages.

In general, this decomposition exercise emphasizes some features of document representation and the projection task. From Table 3.4.4, we can appreciate the importance of down-weighting common words. For example, the word-vector $\vec{a}$ appears positively associated with *Historical* methodologies (0.25), but negatively associated with the *Field Data* methodologies (-0.2). To a human reader, this is hardly convincing since the article *a* is a stopword that is not associated with any specific topic. However, this is ultimately irrelevant in the current example due to the

[12] This is due to the fact that the equations (3.11) and (3.10) divide the word-vectors composing a document by the total number of words in a document, L_i in equation (3.11). Indeed, the same issue would appear if we represented a document with a simple mean of its word vectors.

[13] In this case, the decomposition is approximate because the sum of the coefficients in the plot under-weight the 5 words at the beginning and end of the abstracts. Yet, as the document gets longer the sum of the moving window coefficients will tend to the overall document coefficients. In the example in the text, the difference are substantially negligible.

Figure 3.10 Projections of a 10-words window through the abstract *Climate Clubs and the Macro-Economic Benefits of International Cooperation on Climate Policy* (Paroussos et al., 2019)



exiguous weight associated with *a*. Therefore, corpus-specific weighting schemes make document representation more robust (Arora et al., 2017). Second, Table 3.4.4 shows that the *Document* coefficients for $\vec{Paris}$ and $\vec{policy}$ are higher than 1. Unlike cosine projections, multi-dimensional projections are not normalized between −1 and 1. Indeed, we show in Appendix L that it is impossible to normalize the coefficients of a multi-dimensional projection. Finally, beyond locating the most important words or passages in a document, the decomposition of the overall projection can reveal change of topics within the document. In case the topic of a document changes through its length, it would be ultimately convenient for the analysis to split the document into separate, mono-thematic sub-documents.

3.5. Conclusions

This paper has examined the use of word embedding in the social sciences to extract cultural schemas and associations from textual corpora. The paper has examined the previous uses of word embedding models in the social sciences and shown that they follow a simple semantic-geometric model of meaning, which directly descends from previous formal models of analogical reasoning. In the second part, the paper has formulated a more general semantic-geometric model – the simple algebra of meaning. This new model represents the meaning of words as a weighted sum of elementary semantic dimensions in the embedding space. The simple algebra of meaning generalizes the current use of embeddings in the social sciences beyond the constraints imposed by analogical reasoning. First, the paper has shown how to create semantic dimensions with only one positive pole, reaching beyond the binary logic that previous works have mostly used. Second, the paper has shown how to represent documents in the embedding space, generalizing the models by Arora and colleagues (2017) and by Stoltz and Taylor (2019). Third, the paper has shown how to calculate the engagement of words and documents with more than one semantic dimension at the same time, which can be achieved using standard OLS estimation of a multiple linear regression. Finally, the paper has shown how to decompose the projection of a document using the projections of the words composing it. This decomposition can be used to check the semantic *ductus* of the document. Overall, the new techniques we propose expand the current practices greatly: from analyzing the engagement of a word with one binary semantic dimension to analyzing (and decomposing) the engagement of an entire document with an arbitrary number of non-binary dimensions. We believe this generalization will be helpful to social scientists who wish to use embedding models to study culture and its evolution.

On a deeper level, formulating the semantic-geometric model governing the projection task makes it possible to question it. First, we could simply use other ways to decompose meaning or other estimation strategies. The current semantic-geoemtric model used in the social sciences is a simple linear decomposition – see equation (3.3). There is ample room for innovating this model even in the context of linear decomposition. Indeed, the estimation of linear models is a well studied issue and we can take advantage of a century of study about the topic. For example, we could regularize coefficient estimation through LASSO regressions (Hastie et al., 2001), use an instrumental variable approach to estimation, or change altogether to a more general semi-parametric additive models (Hastie and Tibshirani, 1990). Second, we could adopt a more data-driven approach to modeling. Whereas theoretical priors drive the the estimation of semantic dimensions in the projection task, we could wonder which dimensions would better explain the variation we observe in the embedding space and how the word-vectors and document-vectors decompose along these dimensions (Arseniev-Koehler et al., 2022). This would amount to choose semantic dimensions starting from the data. Depending on the constraints we impose on such data-driven decomposition of meaning, we may be dealing with a principal component analysis (Osgood et al., 1957), a semi non-negative matrix factorization (Ding et al., 2010; Arora et al., 2018), or a non-negative matrix factorization (Lee and Seung, 1999). Finally, formalizing the projection task will help introducing more advanced embedding models in the social sciences. Indeed, computer science research on word embedding has moved from fixed to contextual embeddings (Devlin et al., 2019) and has incorporated non-Euclidean curved spaces as embedding spaces (Nickel and Kiela, 2017; Li et al., 2020). A precise formulation of the semantic-geometric model governing the projection task can help adapting it to the new embedding techniques – substantially, we are looking for projection techniques in a different geometric settings or with moving embeddings.

In conclusion, word embedding models are rising in popularity in the social sciences (Stoltz and Taylor, 2021). The theoretical and methodological discussion in this paper systematizes and generalizes the current use of these models with a formal model of meaning and how meaning is translated into word-vectors' positions. The present discussion will help social scientists to refine current practices and to find new uses of word embedding models.

Conclusion

This dissertation has engaged with decades-long questions about segregation and its quantification as well as more recent development of computational model to extract cultural schemas from unstructured corpora of textual data.In both cases, key questions regard the precise mathematical definition of the associative patterns in the data as well as concrete methodological strategies to measure these patterns robustly.

Social sciences have provided multiple, discrepant answers to these theoretical and methodological questions. Results presented here provide new answers and systematize previous works. All of the chapters have a methodological core centering around new techniques, but they also provide new theoretical frameworks for the issues raised in the literature. From this perspective, the dissertation advance both our practical and theoretical understanding of segregation.

Chapter 1 focuses on the issue of statistical inference for segregation indices. It formulates new estimators for segregation indices based on non-parametric Bayesian models. On a theoretical level, the chapter defines segregation indices formally and provides mild sufficient conditions for segregation indices to be positively biased. Doing this, it proves empirical observations starting from the Seventies (Winship, 1977; Cortese et al., 1976). Methodologically, the chapter benchmarks the performance of the new estimators for the popular D and Theil indices in extensive Monte Carlo simulations. The simulations show that the Bayesian estimators appear to be more reliable than the others, especially in small samples and interval estimation.

The results in the chapter show that the common practice of simply applying the segregation index to the sample (the plug-in estimator) is sub-optimal and possibly troublesome. It will lead to positive bias, skewed comparisons and wrong conclusions, especially in small samples. The

Bayesian techniques as well as the bootstrap estimator by Allen et al. (2015) amend part of the bias and provide much more reliable estimates. The techniques are available in the `SISeg` `R` package.

Chapter 2 analyzes the issue of interpreting change in segregation. It formulates the idea of interpretable segregation indices and creates new indices that are easier to interpret on a substantial level. The chapter proposes the structural-decomposition of an index: the decomposition of total segregation as the sum of structural and marginal effects on segregation. Using structural-decomposition, it is possible to understand what factors are driving the change in segregation: it provides a direct assessment of the marginal/structural contributions and their evolution over time. The chapter formulates a new index, Q, that is structural-decomposable, but also partition-decomposable – that is, decomposable in the traditional sense in which the Theil index is decomposable (Frankel and Volij, 2011). Doing this, the chapter provides a deeper discussion about the possible sources of segregation in the data, between marginal and structural segregation. This theoretical discussion also offers further clarification on so-called margin-free indices, whose properties are characterized in Appendix F. Based on the characterization, the chapter formulates an entire family of margin-free indices. Particularly, it formulates Q^*, which is the first partition-decomposable margin free index.

Finally, Chapter 3 provides a theoretical discussion about semantics for social scientific purposes. The chapter analyzes the theoretical foundation of the social scientific uses of word embedding models. It formulates a geometric-semantic model, "the simple algebra of meaning". This is a model of how meaning is translated into geometrical features of the embedding space. It justifies previous uses of word embeddings in the social sciences and implies further methodological development. Based on the simple algebra of meaning, the chapter shows how to create semantic dimensions with only one positive pole, how to represent documents in the embedding space, how to calculate the engagement of words and documents with more than one semantic dimension at the same time, and how to decompose the projection of a document using the projections of the words composing it. Overall, the new techniques allow the researcher to analyze (and decomposing) the engagement of an entire document with an arbitrary number of non-binary dimensions. On a more theoretical level, the chapter invites theoretical thinking about the kind of meaning that word embedding models can capture – and the kind of meaning that they cannot capture (Arseniev-Koehler et al., 2022). For example, the use of Euclidean metrics constrain the kind of semantic relationships that word embedding can grasp (Tversky, 1977; Peterson et al., 2020).

Future Research. Taken together, the findings in this dissertation pave the way for future methodological research and empirical applications. On a methodological note, all chapters invite further refinement. The first chapter has not fully explored the consequences of the hyper-parameter settings for the DP, C-DP methodologies. More in general, the chapter suggests that new inferential techniques for segregation indices can be devised that start from the assumption that segregation indices in sample will be positively biased and we should counter-balance such bias. The second chapter invites the formulation and analysis of new margin-free and structural-decomposable indices besides Q^* and Q, formulated in the chapter. The zero problem should also be further explored, as it is likely to appear in empirical applications (Jerby et al., 2005). Finally, the third chapter focused thoroughly on previous-generation word embedding models, based on fixed embeddings; current generation word embedding models do not map words to one point in space, but change the geometrical representation of words depending on the context. The re-formulation of the projection task (and the simple algebra of meaning) for this new generation of models should be a high priority for social scientists – since new generation models shows increased accuracy in capturing semantic relations (Devlin et al., 2019).

Empirically, the techniques in the first chapter allow robust inference about segregation in samples that were too small to analyze with previously available techniques (see for example Bielby and Baron, 1986; Martin-Caughey, 2021). The new segregation indices proposed in the second chapter invite a new analysis of historical macro-segregation trends with the objective of isolating the source of change in the trend. For example, is the decline in gender segregation observed from 1950 to 1980 (England et al., 2020; Jacobs, 1989) due to marginal or structural change? What specific occupational categories are driving it? Finally, the third chapter allows researchers to explore with word embedding questions that have been previously impossible to address, such as the salience of non-binary concepts in a document (see for example Nanni and Fallin, 2021).

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). New York, NY, USA: Wiley.

Allen, R., S. Burgess, R. Davidson, and F. Windmeijer (2015). More reliable inference for the dissimilarity index of segregation. *The Econometrics Journal 18*(1), 40–66.

Angrist, J. D. and J.-S. Pischke (2009, October). *Mostly Harmless Econometrics: An Empiricist's Companion*. Number 8769 in Economics Books. Princeton University Press.

Antoniak, M. and D. Mimno (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics 6*, 107–119.

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2016). A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics 4*, 385–399.

Arora, S., Y. Li, Y. Liang, T. Ma, and A. Risteski (2018). Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics 6*, 483–495.

Arora, S., Y. Liang, and T. Ma (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*.

Arseniev-Koehler, A. (2021). Theoretical foundations and limits of word embeddings: what types of meaning can they capture?

Arseniev-Koehler, A., S. D. Cochran, V. M. Mays, K.-W. Chang, and J. G. Foster (2022). Integrating topic modeling and word embedding to characterize violent deaths.

Proceedings of the National Academy of Sciences 119(10), e2108801119. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2108801119.

Arseniev-Koehler, A. and J. Foster (2020, March). Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What It Means to Be Fat.

Ash, E., D. L. Chen, and A. Ornaghi (2020). Stereotypes in High-Stakes Decisions : Evidence from U.S. Circuit Courts. The Warwick Economics Research Paper Series (TWERPS) 1256, University of Warwick, Department of Economics.

Athey, S., B. Ferguson, M. Gentzkow, and T. Schmidt (2021). Estimating experienced racial segregation in US cities using large-scale GPS data. *Proceedings of the National Academy of Sciences 118*(46).

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory 2*(3), 244 – 263.

Ballester, C. and M. Vorsatz (2014). Random walk-based segregation measures. *The Review of Economics and Statistics 96*(3), 383–401. Publisher: The MIT Press.

Baroni, M., R. Bernardi, and R. Zamparelli (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies 9*(6), 5–110.

Baroni, M., G. Dinu, and G. Kruszewski (2014, June). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 238–247. Association for Computational Linguistics.

Bell, W. (1954). A Probability Model for the Measurement of Ecological Segregation. *Social Forces 32*(4), 357–364. Publisher: Oxford University Press.

Beller, A. (1984). Trends in occupational segregation by Sex and Race. In B. Reskin (Ed.), *Sex Segregation in the Workplace: Trends, Explanations, Remedies*, pp. 11–26. Washington, DC:

The National Academies Press.

Bernardo, J. M. and A. F. M. Smith (2009). *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley.

Bianchi, S. M. and N. Rytina (1986, February). The decline in occupational sex segregation during the 1970s: census and cps comparisons. *Demography 23*(1), 79–86.

Bielby, W. T. and J. N. Baron (1986, January). Men and Women at Work: Sex Segregation and Statistical Discrimination. *American Journal of Sociology 91*(4), 759–799. Publisher: The University of Chicago Press.

Blau, F. and L. M. Kahn (2016). The Gender Wage Gap: Extent, Trends, and Explanations. Technical Report 21913, National Bureau of Economic Research, Cambridge, MA.

Blau, F. D., P. Brummund, and A. Y.-H. Liu (2013, April). Trends in Occupational Segregation by Gender 1970–2009: Adjusting for the Impact of Changes in the Occupational Coding System. *Demography 50*(2), 471–492.

Blau, F. D. and W. E. Hendricks (1979). Occupational Segregation by Sex: Trends and Prospects. *The Journal of Human Resources 14*(2), 197–210.

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai (2016). Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, USA, pp. 4356–4364. Curran Associates Inc.

Bouchet-Valat, M. (2022, February). General Marginal-free Association Indices for Contingency Tables: From the Altham Index to the Intrinsic Association Coefficient. *Sociological Methods & Research 51*(1), 203–236. Publisher: SAGE Publications Inc.

Bourdieu, P. (1975, December). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information 14*(6), 19–47.

Bourdieu, P. (1979). The Kabyle House or the World Reversed. In *Algeria 1960*, pp. 133–153. Cambridge, UK & New York: Cambridge University Press.

Boutyline, A., A. Arseniev-Koehler, and D. Cornell (2020, June). School, Studying, and Smarts: Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009.

Boutyline, A. and L. K. Soter (2021, August). Cultural Schemas: What They Are, How to Find Them, and What to Do Once You've Caught One. *American Sociological Review 86*(4), 728–758. Publisher: SAGE Publications Inc.

Breslau, D. (2003, June). Economics invents the economy: Mathematics, statistics, and models in the work of Irving Fisher and Wesley Mitchell. *Theory and Society 32*(3), 379–411.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science 356*(6334), 183–186.

Camic, C. and Y. Xie (1994). The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review 59*(5), 773–805.

Campbell, R. and C. Miéville (2016). *The Age of Lovecraft*. University of Minnesota Press.

Carrington, W. J. and K. R. Troske (1997). On Measuring Segregation in Samples with Small Units. *Journal of Business & Economic Statistics 15*(4), 402–409.

Cech, E. A. (2013). The Self-Expressive Edge of Occupational Sex Segregation. *American Journal of Sociology 119*(3), 747–789.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, pp. 288–296. Curran Associates, Inc.

Charles, M. (1992). Cross-National Variation in Occupational Sex Segregation. *American Sociological Review 57*(4), 483–502.

Charles, M. (2003, December). Deciphering Sex Segregation. *Acta Sociologica 46*(4), 267–287.

Charles, M. and D. B. Grusky (1995). Models for Describing the Underlying Structure of Sex Segregation. *American Journal of Sociology 100*(4), 931–971.

Charles, M. and D. B. Grusky (Eds.) (2004). *Occupational Ghettos: The Worldwide Segregation of Women and Men*. Stanford: Stanford University Press.

Coleman, J., T. Hoffer, and S. Kilgore (1982). Achievement and Segregation in Secondary Schools: A Further Look at Public and Private School Differences. *Sociology of Education 55*(2), 162–182. Publisher: [Sage Publications, Inc., American Sociological Association].

Correll, S. J., S. Benard, and I. Paik (2007). Getting a Job: Is There a Motherhood Penalty? *American Journal of Sociology 112*(5), 1297–1338.

Cortese, C. F., R. F. Falk, and J. K. Cohen (1976). Further Considerations on the Methodological Analysis of Segregation Indices. *American Sociological Review 41*(4), 630–637.

Cotter, D. A., J. DeFiore, J. M. Hermsen, B. M. Kowalewski, and R. Vanneman (1997). All Women Benefit: The Macro-Level Effect of Occupational Integration on Gender Earnings Equality. *American Sociological Review 62*(5), 714–734.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience.

Cutler, D. M. and E. L. Glaeser (1997). Are Ghettos Good or Bad? *The Quarterly Journal of Economics 112*(3), 827–872. Publisher: Oxford University Press.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and Their Application*. Number 1 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

DeDeo, S., R. X. D. Hawkins, S. Klingenstein, and T. Hitchcock (2013). Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems. *Entropy 15*(6), 2246–2276.

Deming, W. E. and F. F. Stephan (1940, December). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics 11*(4), 427–444.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

D'Haultfœuille, X. and R. Rathelot (2017). Measuring segregation on small units: A partial identification analysis. *Quantitative Economics 8*(1), 39–73.

Dieng, A. B., F. J. R. Ruiz, and D. M. Blei (2019). Topic Modeling in Embedding Spaces. _eprint: 1907.04907.

Ding, C. H., T. Li, and M. I. Jordan (2010). Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(1), 45–55.

Dougherty, K. D. (2003, March). How Monochromatic is Church Membership? Racial-Ethnic Diversity in Religious Community. *Sociology of Religion 64*(1), 65–85. _eprint: https://academic.oup.com/socrel/article-pdf/64/1/65/4830648/64-1-65.pdf.

Duncan, O. D. and B. Duncan (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review 20*(2), 210–217.

Echenique, F. and J. Fryer, Roland G. (2007). A Measure of Segregation Based on Social Interactions. *The Quarterly Journal of Economics 122*(2), 441–485. _eprint: https://academic.oup.com/qje/article-pdf/122/2/441/5470068/122-2-441.pdf.

Eco, U. (1986). *Semiotics and the Philosophy of Language*. Advances in semiotics. Bloomington, IN: Indiana University Press.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. Number 57 in Monograph on Statistics and Applied Probability. Washington D.C.: Champman & Hall/CRC.

Elbers, B. (2021, February). A Method for Studying Differences in Segregation Across Time and Space. *Sociological Methods & Research*, 0049124121986204. Publisher: SAGE Publications Inc.

England, P., A. Levine, and E. Mishel (2020). Progress toward gender equality in the United States has slowed or stalled. *Proceedings of the National Academy of Sciences 117*(13), 6990–6997. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/117/13/6990.full.pdf.

Escobar, M. D. and M. West (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association 90*(430), 577–588. Publisher: Taylor & Francis.

Ethayarajh, K., D. Duvenaud, and G. Hirst (2019, July). Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 1696–1705. Association for Computational Linguistics.

Evans, J. A. and P. Aceves (2016, July). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology 42*(1), 21–50.

Faber, J. W. (2019). Segregation and the Cost of Money: Race, Poverty, and the Prevalence of Alternative Financial Institutions. *Social Forces 98*(2), 819–848. _eprint: https://academic.oup.com/sf/article-pdf/98/2/819/31200442/soy129.pdf.

Ferguson, J.-P. and R. Koning (2018, April). Firm Turnover and the Return of Racial Establishment Segregation. *American Sociological Review 83*(3), 445–474.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics 1*(2), 209 – 230. Publisher: Institute of Mathematical Statistics.

Fiel, J. E. (2013, August). Decomposing School Resegregation: Social Closure, Racial Imbalance, and Racial Isolation. *American Sociological Review 78*(5), 828–848. Publisher: SAGE Publications Inc.

Fields, J. and E. N. Wolff (1991). The Decline of Sex Segregation and the Wage Gap, 1970-80. *The Journal of Human Resources 26*(4), 608–622.

Fienberg, S. E. (1970). An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics 41*(3), 907 – 917. Publisher: Institute of Mathematical Statistics.

Frankel, D. M. and O. Volij (2011). Measuring school segregation. *J. Economic Theory 146*(1), 1–38.

Frege, G. (1948). Sense and Reference. *The Philosophical Review 57*(3), 209–230.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.

Gelman, A. and D. B. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science 7*(4), 457 – 472. Publisher: Institute of Mathematical Statistics.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019, July). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica 87*(4), 1307–1340. Publisher: John Wiley & Sons, Ltd.

Gershman, S. J. and D. M. Blei (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology 56*(1), 1–12.

Ghosal, S. and A. van der Vaart (2017). Infinite-Dimensional Bernstein–von Mises Theorem. In *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, pp. 361–390. Cambridge University Press.

Greimas, A. and J. Courtés (1982). *Semiotics and Language: An Analytical Dictionary*. Advances in semiotics. Bloomington, IN: Indiana University Press.

Greimas, A. J. (1983). *Structural semantics : an attempt at a method.* Lincoln: University of Nebraska Press.

Grusky, D. B. and A. Levanon (2006). Describing Occupational Segregation in Sparse and Incomplete Arrays. *Sociological Methods & Research 34*(4), 554–572.

Hakim, C. (1979, November). Occupational Segregation: A comparative study of the degree and pattern of the differentiation between men and women's work in Britain, the United States. Technical Report 9, Department of Employment.

Harris, Z. S. (1954, August). Distributional Structure. *WORD 10*(2-3), 146–162.

Hashimoto, T. B., D. Alvarez-Melis, and T. S. Jaakkola (2016). Word Embeddings as Metric Recovery in Semantic Spaces. *Transactions of the Association for Computational Linguistics 4*, 273–286.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Champman & Hall/CRC.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer Science + Business Media.

Haunschild, R., L. Bornmann, and W. Marx (2016, July). Climate Change Research in View of Bibliometrics. *PLOS ONE 11*(7), e0160393. Publisher: Public Library of Science.

Hjelmslev, L. (1961). *Prolegomena to a theory of language.* Madison: University of Wisconsin Press.

Hutchens, R. (1991). Segregation curves, Lorenz curves, and inequality in the distribution of people across occupations. *Mathematical Social Sciences 21*(1), 31 – 51.

Hutchens, R. (2004). One Measure of Segregation. *International Economic Review 45*(2), 555–578.

Jacobs, J. A. (1989). Long-Term Trends in Occupational Segregation by Sex. *American Journal of Sociology 95*(1), 160–173.

Jahn, J., C. F. Schmid, and C. Schrag (1947). The Measurement of Ecological Segregation. *American Sociological Review 12*(3), 293–303.

James, D. R. and K. E. Taeuber (1985). Measures of Segregation. *Sociological Methodology 15*, 1–32.

Jerby, I., M. Semyonov, and N. Lewin-Epstein (2005). Capturing Gender-Based Microsegregation: A Modified Ratio Index for Comparative Analyses. *Sociological Methods & Research 34*(1), 122–136.

Jones, J. J., M. R. Amin, J. Kim, and S. Skiena (2020). Stereotypical Gender Associations in Language Have Decreased Over Time. *Sociological Science 7*(1), 1–35.

Joseph, J. E. (2020, March). The agency of habitus: Bourdieu and language at the conjunction of Marxism, phenomenology and structuralism. *Language & Communication 71*, 108–122.

Joseph, K. and J. Morgan (2020, July). When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 4392–4415. Association for Computational Linguistics.

Karmel, T. and M. Maclachlan (1988). Occupational Sex Segregation —Increasing or Decreasing?*. *Economic Record 64*(3), 187–195.

King, M. (1992, April). Occupational Segregation by Race and Sex, 1940- 88. *Montly Labor Review 115*(4), 30.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019, October). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review 84*(5), 905–949. Publisher: SAGE Publications Inc.

Kusner, M. J., Y. Sun, N. I. Kolkin, and K. Q. Weinberger (2015). From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 957–966. JMLR.org. event-place:

Lille, France.

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Latour, B. (2013). *An Inquiry into Modes of Existence: An Anthropology of the Moderns*. Cambridge, MA: Harvard University Press.

Lee, D. D. and H. S. Seung (1999, October). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Lee, M. and J. L. Martin (2015). Coding, counting and cultural cartography. *American Journal of Cultural Sociology 3*(1), 1–33.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics 20*(1), 1–31.

Levanon, A. and D. B. Grusky (2016, September). The Persistence of Extreme Gender Segregation in the Twenty-first Century. *American Journal of Sociology 122*(2), 573–619.

Levy, O. and Y. Goldberg (2014a, June). Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Ann Arbor, Michigan, pp. 171–180. Association for Computational Linguistics.

Levy, O. and Y. Goldberg (2014b). Neural Word Embedding as Implicit Matrix Factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 2177–2185. Curran Associates, Inc.

Lewis, M. and G. Lupyan (2020, October). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour 4*(10), 1021–1028.

Li, L., L. Wu, and J. Evans (2020). Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. *Poetics 78*, 101428.

Li, Y., E. Schofield, and M. Gönen (2019). A tutorial on Dirichlet process mixture modeling. *Journal of Mathematical Psychology 91*, 128–144.

Lichter, D. T., D. Parisi, and M. C. Taquino (2015). Toward a New Macro-Segregation? Decomposing Segregation within and between Metropolitan Cities and Suburbs. *American Sociological Review 80*(4), 843–873.

Logan, J. R., A. Foster, J. Ke, and F. Li (2018). The Uptick in Income Segregation: Real Trend or Random Sampling Variation? *American Journal of Sociology 124*(1), 185–222.

Lyons, J. (1995). *Linguistic Semantics. An Introduction.* Cambridge University Press.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Martin-Caughey, A. (2021, October). What's in an Occupation? Investigating Within-Occupation Variation and Gender Segregation Using Job Titles and Task Descriptions. *American Sociological Review 86*(5), 960–999. Publisher: SAGE Publications Inc.

Massey, D. S. and N. A. Denton (1988). The Dimensions of Residential Segregation*. *Social Forces 67*(2), 281–315.

Massey, D. S., J. Rothwell, and T. Domina (2009). The Changing Bases of Segregation in the United States. *The Annals of the American Academy of Political and Social Science 626*, 74–90. Publisher: [Sage Publications, Inc., American Academy of Political and Social Science].

McCall, L. (2001). *Complex Inequality.* New York, NY: Routledge.

McInnes, L., J. Healy, and J. Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* _eprint: 1802.03426.

McPherson, J. M. and L. Smith-Lovin (1986). Sex Segregation in Voluntary Associations. *American Sociological Review 51*(1), 61–79.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR abs/1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, USA, pp. 3111–3119. Curran Associates Inc.

Mohr, J. W. (1998, August). Measuring Meaning Structures. *Annual Review of Sociology 24*(1), 345–370. Publisher: Annual Reviews.

Mohr, J. W. and P. Bogdanov (2013, December). Introduction—Topic models: What they are and why they matter. *Poetics 41*(6), 545–569.

Mora, R. and J. Ruiz-Castillo (2009, January). The invariance properties of the mutual information index of multigroup segregation. In Y. Flückiger, S. F. Reardon, and J. Silber (Eds.), *Occupational and Residential Segregation*, Volume 17 of *Research on Economic Inequality*, pp. 33–53. Emerald Group Publishing Limited.

Mora, R. and J. Ruiz-Castillo (2011, June). Entropy-Based Segregation Indices. *Sociological Methodology 41*(1), 159–194.

Morey, R. D., R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers (2016, February). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review 23*(1), 103–123.

Nanni, A. and M. Fallin (2021). Earth, wind, (water), and fire: Measuring epistemic boundaries in climate change research. *Poetics 88*, 101573.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Nelson, L. K. (2021, March). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics*, 101539.

Nickel, M. and D. Kiela (2017). Poincaré Embeddings for Learning Hierarchical Representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Oppenheimer, V. (1970). *The Female Labor Force in the United States*. Berkeley: University of California Press.

Osgood, C., G. Suci, and P. Tannenbaum (1957). *The Measurement of Meaning*. Urbana, Chicago and London: University of Illinois Press.

Osius, G. (2004, November). The association between two random elements: A complete characterization and odds ratio models. *Metrika 60*(3), 261–277.

Osius, G. (2009). Asymptotic Inference for Semiparametric Association Models. *The Annals of Statistics 37*(1), 459–489.

Paninski, L. (2003, June). Estimation of Entropy and Mutual Information. *Neural Comput. 15*(6), 1191–1253.

Paroussos, L., A. Mandel, K. Fragkiadakis, P. Fragkos, J. Hinkel, and Z. Vrontisi (2019, July). Climate clubs and the macro-economic benefits of international cooperation on climate policy. *Nature Climate Change 9*(7), 542–546.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572. Publisher: Taylor & Francis.

Peirce, C. S. (1931). *Collected Papers of Charles Sander Peirce*, Volume 2. Cambridge, MA: Belknap.

Pennington, J., R. Socher, and C. Manning (2014, October). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics.

Peterson, J. C., D. Chen, and T. L. Griffiths (2020, December). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition 205*, 104440.

Popoviciu, T. (1935). Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica (Cluj) 9*, 129–145.

Quillian, L. (2014). Does Segregation Create Winners and Losers? Residential Segregation and Inequality in Educational Attainment. *Social Problems 61*(3), 402–426. Publisher: [Oxford University Press, Society for the Study of Social Problems].

Ransom, M. R. (2000). Sampling Distributions of Segregation Indexes. *Sociological Methods & Research 28*(4), 454–475.

Rathelot, R. (2012). Measuring Segregation When Units are Small: A Parametric Approach. *Journal of Business & Economic Statistics 30*(4), 546–553.

Reardon, S. F. and K. Bischoff (2011). Income Inequality and Income Segregation. *American Journal of Sociology 116*(4), 1092–1153.

Reardon, S. F., K. Bischoff, A. Owens, and J. B. Townsend (2018, December). Has Income Segregation Really Increased? Bias and Bias Correction in Sample-Based Segregation Estimates. *Demography 55*(6), 2129–2160.

Reardon, S. F. and G. Firebaugh (2002, August). Measures of Multigroup Segregation. *Sociological Methodology 32*(1), 33–67.

Reardon, S. F. and A. Owens (2014). 60 Years After Brown: Trends and Consequences of School Segregation. *Annual Review of Sociology 40*(1), 199–218.

Reardon, S. F. and D. O'Sullivan (2004). Measures of Spatial Segregation. *Sociological Methodology 34*(1), 121–162.

Reimers, N. and I. Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR abs/1908.10084*. _eprint: 1908.10084.

Reskin, B. and P. Roos (1990). *Job Queues, Gender Queues*. Philadelphia: Temple University Press.

Ridgeway, C. L. (2011). *Framed by Gender*. New York, NY: Oxford University Press.

Roberto, E. (2018, August). The Spatial Proximity and Connectivity Method for Measuring and Analyzing Residential Segregation. *Sociological Methodology 48*(1), 182–224. Publisher: SAGE Publications Inc.

Rodriguez, P., A. Spirling, and B. Stewart (2021). Embedding Regression: Models for Context-Specific Description and Inference in Political Science.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics 9*(1), 130–134.

Rumelhart, D. E. and A. A. Abrahamson (1973, July). A model for analogical reasoning. *Cognitive Psychology 5*(1), 1–28.

Sampson, R. (2012). *Great American city : Chicago and the enduring neighborhood effect*. Chicago; London: The University of Chicago Press.

Semyonov, M., Y. Haberfeld, Y. Cohen, and N. Lewin-Epstein (2000). Racial Composition and Occupational Segregation and Inequality across American Cities. *Social Science Research 29*(2), 175–187.

Semyonov, M., D. R. Hoyt, and R. I. Scott (1984, November). The place of odds ratios in the study of place, race and differential occupational opportunities. *Demography 21*(4), 667–671.

Semyonov, M. and R. I. Scott (1983, May). Industrial shifts, female employment, and occupational differentiation: a dynamic model for American cities, 1960–1970. *Demography 20*(2), 163–176.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics. New York, NY: Springer.

Shorrocks, A. F. (1980). The Class of Additively Decomposable Inequality Measures. *Econometrica 48*(3), 613–625.

Shorrocks, A. F. (1982). Inequality Decomposition by Factor Components. *Econometrica 50*(1), 193–211. Publisher: [Wiley, Econometric Society].

Stainback, K. and D. Tomaskovic-Devey (2012). *Documenting Desegregation: Racial and Gender Segregation in Private Sector Employment Since the Civil Rights Act*. New York: Russell Sage Foundation.

Stanhill, G. (2001, February). The Growth of Climate Change Science: A Scientometric Study. *Climatic Change 48*(2), 515–524.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Stoltz, D. S. and M. A. Taylor (2019, July). Concept Mover's Distance: measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science 2*(2), 293–313.

Stoltz, D. S. and M. A. Taylor (2021). Cultural cartography with word embeddings. *Poetics*, 101567.

Swinger, N., M. De-Arteaga, N. T. Heffernan IV, M. D. Leiserson, and A. T. Kalai (2019). What Are the Biases in My Word Embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, New York, NY, USA, pp. 305–311. Association for Computing Machinery. event-place: Honolulu, HI, USA.

Taylor, M. A. and D. S. Stoltz (2020). Concept Class Analysis: A Method for Identifying Cultural Schemas in Texts. *Sociological Science 7*(23), 544–569.

Taylor, M. A. and D. S. Stoltz (2021, May). Integrating semantic directions with concept mover's distance to measure binary concept engagement. *Journal of Computational Social Science 4*(1), 231–242.

Theil, H. (1971). *Principles of Econometric*. New York, NY: Wiley.

Theil, H. and A. J. Finizza (1971). A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology 1*(2), 187–193.

Tilcsik, A., M. Anteby, and C. R. Knight (2015). Concealable Stigma and Occupational Segregation: Toward a Theory of Gay and Lesbian Occupations. *Administrative Science Quarterly 60*(3), 446–481.

Tomaskovic-Devey, D., C. Zimmer, K. Stainback, C. Robinson, T. Taylor, and T. McTague (2006). Documenting Desegregation: Segregation in American Workplaces by Race, Ethnicity, and Sex, 1966–2003. *American Sociological Review 71*(4), 565–588.

Trumbach, R. (1998). *Sex and the Gender Revolution, Volume 1: Heterosexuality and the Third Gender in Enlightenment London*. Sex & the Gender Revolution Volume One. University of Chicago Press.

Turney, P. D. and P. Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence 37*, 144–188.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352. Place: US Publisher: American Psychological Association.

van Loon, A. and J. Freese (2022, February). Word Embeddings Reveal How Fundamental Sentiments Structure Natural Language. *American Behavioral Scientist*, 00027642211066046. Publisher: SAGE Publications Inc.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehatR$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis 16*(2), 667 – 718. Publisher: International Society for Bayesian Analysis.

Watts, M. (1998, November). Occupational gender segregation: Index measurement and econometric modeling. *Demography 35*(4), 489–496.

Weeden, K. (2004). Profile of Change: Sex Segregation in the United States, 1910-2000. In M. Charles and D. B. Grusky (Eds.), *Occupational Ghettos: The Worldwide Segregation of Women and Men*, pp. 131–178. Stanford: Stanford University Press.

White, M. J. (1986). Segregation and Diversity Measures in Population Distribution. *Population Index 52*(2), 198–221.

Williams, J. J. (1948). Another Commentary on So-Called Segregation Indices. *American Sociological Review 13*(3), 298–303. Publisher: [American Sociological Association, Sage Publications, Inc.].

Winship, C. (1977). A Revaluation of Indexes of Residential Segregation. *Social Forces 55*(4), 1058–1066.

Appendix to Chapter 1

A. Delta Method

The plug-in interval estimator uses the delta method to calculate the standard deviation. This paragraph shows how to construct delta-method-based variances for the indices considered in the main text.

In general, every variance from the delta method is the result of a quadratic form:

$$\sigma^2\big[S(T)\big] = \vec{\lambda}^T \cdot N \cdot \Omega \cdot \vec{\lambda} \tag{A1}$$

where $\sigma^2\big[S(T)\big]$ is the approximate variance of $S(T)$, $\vec{\lambda}$ is a $(G \cdot O) \times 1$ vector of first order partial derivative of the index of interest, Ω is a $(G \cdot O) \times (G \cdot O)$ symmetric variance-covariance matrix and N is a $(G \cdot O) \times (G \cdot O)$ diagonal matrix representing the effect of sample size on variance – so that variance diminishes as sample size increases. The contents of Ω and N depend exclusively on the sampling schema or statistical model used, whereas the contents of the $\vec{\lambda}$ vector are unique for each index. As usual in these cases, Ω and $\vec{\lambda}$ can only be estimated from the data. Therefore, this section should distinguish between the theoretical quantities (based on the parameters of the different frameworks) and the estimates of such quantities (derived from the data). For the sake of brevity and simplicity, we will only describe the data-based estimates (see Agresti, 2013, chap. 16), which are the only relevant estimates in this case.

In all that follows, let $T_{g(i)}$ $(T_{o(j)})$ be the sum of the i^{th} (j^{th}) row (column) of table T. We also "vectorize" the table T so that it transforms from a matrix in $\mathbb{R}^{G \times O}$ to a vector in $\mathbb{R}^{v}$ where $v = G \cdot O$. We use the notation $T(i)$ to refer to the i^{th} position in this vectorized form of T. Specifically, we vectorize by row. So, $T(1)$ corresponds to the first row, first column element of the original table; $T(2)$ corresponds to the first row, second column element of the original table; etc. We indicate with $\pi_{T(i)}$ the $T(i)^{th}$ proportion in the original table. That is, if $\sum T$ is the total number of individual in the sample table T, then $\pi_{T(i)} = \frac{T(i)}{\sum T}$.

A.1. Variance-Covariance Matrices, Ω and Sample Size

The variance-covariance matrix Ω can be derived by considering the independence assumptions of the sampling schema. Here we consider the general case for any number of different groups G. The case of $G = 2$ – used in the simulations – is a special case of the following equations.

For the simple random sample used in the main text:

$$\Omega = \begin{bmatrix} \pi_{T(1)}(1-\pi_{T(1)}) & -\pi_{T(1)}\pi_{T(2)} & -\pi_{T(1)}\pi_{T(3)} & \cdots & -\pi_{T(1)}\pi_{T(O\cdot G)} \\ -\pi_{T(1)}\pi_{T(2)} & \pi_{T(1)}(1-\pi_{T(1)}) & -\pi_{T(1)}\pi_{T(2)} & \cdots & -\pi_{T(2)}\pi_{T(O\cdot G)} \\ \vdots & & \ddots & & \vdots \\ -\pi_{T(1)}\pi_{T(O\cdot G)} & -\pi_{T(2)}\pi_{T(O\cdot G)} & \cdots & & -\pi_{T(O\cdot G)}(1-\pi_{T(O\cdot G)}) \end{bmatrix} \tag{A2}$$

In this case, the N matrix is an appropriately-sized diagonal matrix in the form:

$$N = \begin{bmatrix} \frac{1}{N} & 0 & \cdots & 0 \\ 0 & \frac{1}{N} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{N} \end{bmatrix} \tag{A3}$$

A.2. Derivative Vectors, $\vec{\lambda}$

The vector $\vec{\lambda}$ depends both on the sampling schema and the segregation index. Unlike the previous paragraph, here we will focus on the $G = 2$ special case.

We provide the derivative for the Theil index and its non-normalized version, the mutual information index (Mora and Ruiz-Castillo, 2011) – which we will indicate with MI – see equation

(A15). Providing both expression is convenient. First, the derivative of the mutual information index is used in the expression of the derivative of the Theil index. Second, the derivative vectors of the mutual information index can be used to produce inference on this index, which some scholars prefer (Frankel and Volij, 2011; Mora and Ruiz-Castillo, 2011).

For simple random sampling, the derivatives are taken with respect to all possible table proportion, $\pi_{T(i)}$. In the equations below, let q be the group (row) of $\pi_{T(i)}$ – for example, $q = 1$ if $\pi_{T(i)}$ is from the first group (row) of T; let r be the occupation (column) of $\pi_{T(i)}$ – for example, $r = 3$ if $\pi_{T(i)}$ is in the third occupation (column) of T. Since we only deal with with two groups (rows) in the following equations, we will also use $\not{q}$ to indicate the row to which $\pi_{T(i)}$ does not belong.

The $\vec{\lambda}$ vector for the D, mutual information, and Theil indices

$$\begin{aligned}\frac{\partial D(T)}{\partial \pi_{T(i)}} =& \frac{1}{2}\frac{1}{\pi_{g(q)}}\Bigg[(-1)^{q+1} sgn\big(\pi_{o(r)|g(q)} - \pi_{o(r)|g(\not{q})}\big) \\ &+ (-1)^{q}\sum_{j=1}^{O} \pi_{o(j)|g(q)}\, sgn\big(\pi_{o(j)|g(1)} - \pi_{o(j)|g(2)}\big)\Bigg]\end{aligned} \tag{A4}$$

$$\frac{\partial MI(T)}{\partial \pi_{T(i)}} = \log\left(\frac{\pi_{T(i)}}{\pi_{o(r)}\pi_{g(q)}}\right) - 1 \tag{A5}$$

$$\frac{\partial Th(T)}{\partial \pi_{T(i)}} = \frac{1}{H(\pi_G)}\left[(\log \pi_{g(q)} + 1)\cdot Th(T) + \frac{\partial MI(T)}{\partial \pi_{T(i)}}\right] \tag{A6}$$

where $sgn(x)$ is the sign function, which is 1 when x is positive, -1 when x is negative and 0 when $x = 0$; $H(\pi_G)$ is the entropy function defined in equation (1.3). Equation (A4) coincides with Equations (4a) and (4b) in Ransom (2000)

B. Gibbs Sampler for the DP(C) and C-DP(C) Method

In this Appendix, we document the Gibbs sampler for the Dirichlet Process Mixture Model used both in the DP(C) and the C-DP(C) methods. In general, we will follow Algorithm 2 by Neal (2000) for the clustering part, adapting it to our specific mixture model. Notice that we are drawing upon the conjugacy of the beta-binomial model.

The hyper-parameters C, α_0 and β_0 are positive known scalar. Let the Markov state be represented as

1: A vector $\vec{c}$, where $c(i)$ is the cluster assigned to the i^{th} occupation. The length of this vector is fixed at O.

2: A vector $\vec{\Phi}$, where Φ_j is the minority/majority proportion associated with each cluster. This vector has a variable length, since the number of clusters may change from one iteration of the loop to the next. In general, $\vec{\Phi}$ will have length from 1 to O.

3: A vector $\vec{\Pi}$ where $\Pi_i = \Phi_{c(i)}$. That is, $\vec{\Pi}$ maps each occupation with the proportion in its cluster. This vector has fixed length equal to O. Quite clearly, $\vec{\Pi}$ does not add any new information to $\vec{c}$ and $\vec{\Phi}$, but it is notationally convenient to single out this vector.

We use the following notation: $O(k_{-i})$ is the number of occupations in cluster k excluding the i^{th} occupation from consideration, K is the number of clusters containing at least one occupations (notice K may change across iterations), $K(-i)$ is the number of clusters containing at least one occupation not considering the i^{th} occupation (also changing during iterations), $G(k)$ is the total number of individuals belonging to group 1 among the occupations in cluster k, similarly $N(k)$ is the total number of individuals among the occupations in cluster k. $T_{\sum,j}$ is the j^{th} element of the row-sum of T; that is, $T_{\sum,j} = \sum_{i=1}^{G} T_{i,j}$. For convenience, we assume that clusters are labelled with integers starting from 1 to K with no gaps between them – they can be re-labelled after each iteration to enforce this is true.

We will calculate the posterior for an arbitrary function of m independent segregation indices, $f\big(S(\tau_1)\ldots S(\tau_m)\big)$, with $m \geq 1$. For each environment$\tau_{[r]}$, with $r = 1\ldots m$, we observe an independent sample $T_{[r]}$ from the environment. Each environment will have its own parameters and Markov chain. One may also assign different hyper-parameters to the different environments, but we do not consider this possibility here. We will change notation accordingly to indicate which environment we are referring to. For example, O_r indicates the number of occupations in the r^{th} environment.

I: For each occupation $i = 1\ldots O_r$, assign the occupation to a cluster based on the following probability:

$$\text{(A7)}\qquad P(c(i) = k|\vec{c}_r, \vec{\Phi}_r, \vec{\Pi}_r) \propto \begin{cases} \frac{O(k_{-i})}{O_r - 1 + C} \cdot dBin(T_{1,i,[r]}; \Phi_{k,[r]}, T_{\Sigma,i,[r]}) & \text{for } k \leq K_r(-i) \\ \frac{C}{O_r - 1 + C} \cdot dBBin(T_{1,i,[r]}; \alpha_0, \beta_0, T_{\Sigma,i,[r]}) & \text{for } k = K_r(-i) + 1 \end{cases}$$

where $dBin(T_{1,i,[r]}; \Phi_{k,[r]}, T_{\Sigma,i,[r]})$ $[dBBin(T_{1,i,[r]}; \alpha_0, \beta_0, T_{\Sigma,i,[r]})]$ is the probability (density) distribution of the binomial (beta-binomial) distribution at $T_{1,i}$ with parameters $P = \Phi_{k,[r]}, n = T_{\Sigma,i,[r]}$ $(\alpha = \alpha_0, \beta = \beta_0, n = T_{\Sigma,i,[r]})$. If the occupation is assigned to a new cluster, immediately draw a proportion from the posterior, as described in the next passage. Otherwise, continue with the assignment loop.

II: For each cluster, $j = 1\ldots K$ draw a proportion from the posterior distribution:

$$\text{(A8)}\qquad \phi_{k,r}|\vec{c}_r, \vec{\Phi}_r, \vec{\Pi}_r \sim Beta\Big(\alpha_0 + G_r\big(k\big),\ \beta_0 + N_r\big(k\big) - G_r\big(k\big)\Big)$$

III: Build a variate from the posterior environment τ_r^*. In this variate, the i^{th} occupation column is represented as $\pi_{o(i)} \cdot \left[\Pi_{c(i),r}\ , (1-\Pi_{c(i),r})\right]$.

IV: For each occupation $i = 1 \dots O$, sample $T_{1,i,[r]}^*$ and $T_{2,i,[r]}^*$ as:

$$T_{1,i,[r]}^* | \vec{c}_r, \vec{\Phi}_r, \vec{\Pi}_r \sim Bin(\Pi_{c(i),r}, T_{\sum,i,[r]}) \tag{A9}$$

$$T_{2,i,[r]}^* = T_{\sum,i,[r]} - T_{1,i,[r]}^* \tag{A10}$$

V: Build a new sample table T_r^* by assembling the columns $T_{,i,[r]}^* = [T_{1,i,[r]}^*, T_{2,i,[r]}^*]$ in one marix.

After these passages are complete for all m environments:

VI: Calculate a variate from the posterior of $f\Bigg(S(\tau_1), S(\tau_2) \dots S(\tau_m)\Bigg)$:

$$f\Bigg(S(\tau_1), S(\tau_2) \dots S(\tau_m)\Bigg) | \vec{c}_1, \vec{\Phi}_1, \vec{\Pi}_1 \dots \vec{c}_m, \vec{\Phi}_m, \vec{\Pi}_m = f\Bigg(S(\tau_1^*), S(\tau_2^*) \dots S(\tau_m^*)\Bigg) \tag{A11}$$

VII: Calculate a variate from the posterior of $f\Bigg(S(T_1), S(T_2) \dots S(T_m)\Bigg)$:

$$f\Bigg(S(T_1), S(T_2) \dots S(T_m)\Bigg) | \vec{c}_1, \vec{\Phi}_1, \vec{\Pi}_1 \dots \vec{c}_m, \vec{\Phi}_m, \vec{\Pi}_m = f\Bigg(S(T_1^*), S(T_2^*) \dots S(T_m^*)\Bigg) \tag{A12}$$

This iteration can be repeated B times to produce B samples from the posterior distributions of $f\Bigg(S(T_1), S(T_2) \dots S(T_m)\Bigg)$ and $f\Bigg(S(\tau_1), S(\tau_2) \dots S(\tau_m)\Bigg)$. Notice that this sampling procedure can be used to produce inference for the value of a segregation index in one environment; in that case, $m = 1$ and $f\big(S(\tau_1)\big)$ is simply the identity function. As is always the case, the sampler needs time to mix and start sampling from the posterior distribution of interest. In the experiment of section 1.6, we generally fix this time by taking extra 500 burn-in samples at the beginning. In applied cases, the user can utilize any of the available tests for convergence (Gelman and Rubin, 1992; Vehtari et al., 2021). Finally, the sampler above does not focus on the clustering structure of the data as a final product. While the clustering structure is certainly used in the estimation, inferring clusters is not the real goal of this sampler. Therefore, this procedure does not suffer from the label-switching problems afflicting the use of mixture models for clustering (Stephens, 2000).

C. The Influence of the *C* Parameter

The results from the Monte Carlo simulations in section 1.6 and section 1.7 show that the Bayesian estimators generally perform better than the others, especially in those small-sample situations where it is most critical to have an alternative to the plug-in estimator. However, the Bayesian techniques depend on the choice of a hyper-parameter, C: we may have been particularly lucky in getting a C value that was particularly convenient for the tests we performed. In fact, if the performances of the Bayesian techniques are very sensitive to the choice of a good C parameter, the Bayesian techniques may actually perform worst than the alternatives besides a short interval of the parameter's value. Therefore, it is important to assess the influence of the C parameter on the inferential performances of the DP(C) and C-DP(C). For this purpose, we replicated the tests of section 1.6 and section 1.7 with two further values for C, $C = 0.5$ and $C = 2$.

To set expectation, notice that equation (A7) shows specifically how C influences the final inferential results. In general, a lower value of C will result in fewer clusters, which in turn should result in a reduced estimate of a segregation index. That is, the estimated segregation monotonically decreases with C. For this reasons, we expect the performance metrics (e.g. RMSE or coverage) of DP(C) and C-DP(C) to generally be convex in C. In other words, there must be one optimal amount of correction given the sample strategy and population characteristics: a reasonable performance metrics will get monotonically worst as we get farther away from that optimal value. In addition, the influence of C should decrease as the the sample size augments. In fact, C becomes irrelevant when the sample sizes grows to infinity. Finally, DP(C) will be more sensible to C than C-DP(C). Consider equation (1.22). The value of C will influence both the $S(T_j^*)$ and $S(\tau^*)$ terms in the equation, but the influence will conceivably be in the same direction. Therefore, any change in $S(T_j^*)$ will be partially compensated by the change in $S(\tau^*)$ and vice versa. On the contrary, the DP(C) will be fully influenced by any change in $S(\tau^*)$.

Tables .7 to .10 show the results of these Monte Carlo tests. In practice, the tables mostly confirm the expectation. The influence of C diminishes substantially as soon as we leave the small-sample regime. In small samples, the influence of C is still not so great as to change any of the conclusions from the main text: the Bayesian estimators are substantially more reliable than the others for C in the interval from 0.5 to 2. Moreover, changes in the C parameter do not appear to produce estimators that dominate the default estimators used in the main text – i.e. DP(1) and C-DP(1). That is to say, a value of C is more or less convenient depending on the underlying segregation of an environment, which is not something we can address in actual samples. This indicates that there is no always-best choice for C in the range tested – even if, to be precise, there are few bigger samples where a different choice of C would actually produce a dominant estimator for the D index (see Table .7). For both DP(C) and C-DP(C), the D index appears more sensible to the influence of C. As for the Theil index, both estimators appear to be quite robust to the choice of C. Finally, we anticipated that the C-DP(C) estimator will be less responsive to C than DP(C). This is correct: for both indices, the C-DP(C) estimator shows little sensitivity as soon as we leave the smallest sample size considered in the experiments. For the Theil index in particular, different values of C barely influence the C-DP(C) estimator, even in the smallest sample.

We conclude that both DP(C) and C-DP(C) appear robust to the influence of C, with the C-DP(C) appearing especially robust. Given that C-DP(C) usually shows better or similar performances with respect to the DP(C) estimator, its greater robustness to C is possibly a good reason to prefer it over DP(C).

The results of this section are certainly re-assuring, but they do not provide any definitive answer about how to set C. It is clear that we can chooses values of C that will produce very bad

	Bias				RMSE				Confidence Interval			
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
N=500, P=0.05												
C-DP(0.5)	**-0.005**	-0.019	**-0.066 - 0.089**		**0.103**	**0.095**	**0.066 - 0.145**		**0.941**	0.918	**0.912 - 0.986**	
C-DP(1)	0.023	-0.015	-0.031 - 0.142		0.104	0.099	0.063 - 0.165		0.929	0.924	0.862 - 0.974	
C-DP(2)	0.051	**0.004**	-0.016 - 0.197		0.112	0.097	0.059 - 0.212		0.877	**0.926**	0.626 - 0.956	
N=500, P=0.2												
C-DP(0.5)	-0.010	-0.002	-0.044 - 0.015		0.056	**0.054**	0.033 - 0.085		0.902	**0.912**	0.824 - 0.960	
C-DP(1)	**-0.003**	-0.002	**-0.031 - 0.037**		0.054	0.057	**0.033 - 0.073**		**0.908**	0.908	**0.866 - 0.952**	
C-DP(2)	0.005	-0.002	-0.016 - 0.060		**0.053**	0.055	0.033 - 0.077		0.897	0.908	0.826 - 0.934	
N=500, P=0.3												
C-DP(0.5)	-0.009	**-0.001**	-0.034 - 0.007		0.050	**0.047**	0.031 - 0.074		0.884	0.884	0.838 - 0.962	
C-DP(1)	-0.004	-0.002	**-0.025 - 0.024**		0.048	0.053	0.031 - 0.065		**0.888**	**0.892**	**0.854 - 0.920**	
C-DP(2)	**0.002**	**-0.001**	-0.016 - 0.044		**0.047**	0.051	**0.031 - 0.063**		0.882	0.886	0.844 - 0.910	
N=1,500, P=0.05												
C-DP(0.5)	-0.009	-0.003	**-0.042 - 0.025**		0.060	0.063	0.037 - 0.090		**0.909**	**0.914**	0.838 - 0.956	
C-DP(1)	**0.000**	**-0.002**	-0.025 - 0.049		**0.058**	**0.060**	**0.036 - 0.077**		0.906	0.904	**0.874 - 0.944**	
C-DP(2)	0.009	**-0.002**	-0.015 - 0.074		**0.058**	**0.060**	0.036 - 0.092		0.887	0.902	0.764 - 0.934	
N=1,500, P=0.2												
C-DP(0.5)	-0.006	-0.001	-0.029 - 0.003		0.033	0.033	0.019 - 0.052		0.894	**0.904**	0.806 - 0.950	
C-DP(1)	-0.004	**0.000**	-0.021 - 0.003		0.031	0.033	0.019 - 0.044		0.899	0.902	0.848 - 0.944	
C-DP(2)	**-0.001**	**0.000**	**-0.012 - 0.015**		**0.030**	**0.032**	**0.019 - 0.039**		**0.905**	0.902	**0.870 - 0.942**	
N=1,500, P=0.3												
C-DP(0.5)	-0.006	**-0.001**	-0.024 - 0.000		0.029	0.028	0.017 - 0.046		0.892	0.904	0.804 - 0.926	
C-DP(1)	-0.004	-0.002	-0.019 - 0.000		0.027	0.028	0.017 - 0.040		0.902	**0.910**	0.836 - 0.944	
C-DP(2)	**-0.002**	**-0.001**	**-0.012 - 0.008**		**0.026**	0.028	**0.017 - 0.034**	1	**0.904**	0.906	**0.874 - 0.944**	
N=2,500, P=0.05												
C-DP(0.5)	-0.008	-0.001	-0.031 - 0.008		0.047	0.046	0.027 - 0.071		0.907	0.910	0.840 - 0.962	
C-DP(1)	**-0.003**	-0.001	**-0.024 - 0.025**		0.045	0.050	0.027 - 0.063		**0.911**	**0.916**	**0.872 - 0.936**	
C-DP(2)	**0.003**	-0.001	-0.016 - 0.043		**0.044**	**0.045**	**0.026 - 0.061**		0.902	0.914	0.850 - 0.946	
N=2,500, P=0.2												
C-DP(0.5)	-0.005	-0.001	-0.018 - 0.002		**0.025**	**0.025**	0.015 - 0.040		0.900	0.918	0.826 - 0.924	
C-DP(1)	-0.003	-0.003	-0.015 - 0.002		0.024	**0.025**	0.015 - 0.036		0.909	0.918	0.850 - 0.932	
C-DP(2)	**-0.001**	**0.000**	**-0.011 - 0.006**		0.023	0.024	**0.015 - 0.031**	1	**0.914**	0.918	**0.878 - 0.932**	
N=2,500, P=0.3												
C-DP(0.5)	-0.003	0.000	-0.013 - 0.001		0.022	0.022	0.013 - 0.032		0.901	**0.918**	0.844 - 0.920	
C-DP(1)	-0.002	0.000	-0.008 - 0.001		0.021	**0.021**	0.013 - 0.029		0.911	0.914	0.860 - 0.938	
C-DP(2)	**-0.001**	0.000	**-0.006 - 0.003**		**0.020**	**0.021**	**0.013 - 0.026**	1	**0.912**	0.912	**0.866 - 0.948**	

Table .7. Monte Carlo comparison of different concentration values of the C-DP(C) method applied to the D index. The ≥ signals whether an estimator dominates C-DP(1) in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

	Bias				RMSE				Confidence Interval			
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
N=500, P=0.05												
DP(0.5)	**-0.004**	-0.028	**-0.071 - 0.108**		**0.098**	0.090	0.068 - 0.138		0.964	**0.966**	**0.910 - 0.992**	
DP(1)	0.027	-0.025	-0.031 - 0.168		0.103	0.085	0.066 - 0.179		**0.959**	0.968	0.888 - 0.984	
DP(2)	0.058	**0.013**	-0.039 - 0.229		0.116	**0.082**	**0.064 - 0.236**		0.847	0.970	0.280 - 0.984	
N=500, P=0.2												
DP(0.5)	-0.008	-0.009	**-0.043 - 0.037**		**0.052**	0.048	0.033 - 0.079		0.953	0.954	0.896 - 0.998	
DP(1)	**0.005**	**-0.006**	-0.018 - 0.069		**0.052**	**0.047**	**0.033 - 0.078**		**0.958**	**0.958**	**0.936 - 0.988**	
DP(2)	0.020	**-0.006**	-0.015 - 0.106		0.057	0.050	0.033 - 0.111		0.909	0.956	0.638 - 0.974	
N=500, P=0.3												
DP(0.5)	-0.009	-0.007	-0.037 - 0.025		0.047	**0.042**	0.031 - 0.071		0.947	0.946	0.902 - 0.992	
DP(1)	**0.002**	-0.008	-0.019 - 0.053		**0.046**	0.043	**0.031 - 0.063**		**0.950**	0.948	**0.930 - 0.984**	
DP(2)	0.014	**-0.003**	-0.011 - 0.086		0.049	0.045	0.031 - 0.091		0.919	**0.950**	0.712 - 0.986	
N=1,500, P=0.05												
DP(0.5)	**-0.007**	-0.009	**-0.040 - 0.045**		**0.056**	**0.050**	**0.036 - 0.082**		**0.949**	**0.950**	0.882 - 0.988	
DP(1)	0.008	-0.008	-0.016 - 0.080		**0.056**	0.053	0.035 - 0.089		0.954	0.956	**0.926 - 0.986**	
DP(2)	0.024	**-0.005**	-0.012 - 0.119		0.062	0.055	0.035 - 0.125		0.889	0.956	0.518 - 0.966	
N=1,500, P=0.2												
DP(0.5)	-0.006	-0.001	-0.029 - 0.001		0.031	0.026	0.018 - 0.048		0.937	0.930	0.884 - 0.990	
DP(1)	**0.000**	-0.001	**-0.013 - 0.020**		**0.030**	0.032	**0.018 - 0.041**		**0.951**	**0.948**	**0.930 - 0.982**	
DP(2)	0.007	**0.000**	-0.004 - 0.042		**0.030**	**0.030**	0.018 - 0.048		0.943	0.944	0.906 - 0.974	
N=1,500, P=0.3												
DP(0.5)	-0.006	-0.003	-0.026 - -0.001		0.027	**0.023**	0.017 - 0.045		0.937	0.934	0.856 - 0.996	
DP(1)	**-0.001**	-0.002	**-0.014 - 0.012**		**0.026**	0.024	**0.016 - 0.035**		**0.947**	**0.940**	0.928 - 0.990	
DP(2)	0.004	**0.001**	-0.003 - 0.031		**0.026**	0.027	0.016 - 0.037		0.945	0.938	**0.930 - 0.980**	
N=2,500, P=0.05												
DP(0.5)	-0.007	-0.004	**-0.033 - 0.024**		0.043	0.041	0.026 - 0.069		**0.948**	**0.948**	0.894 - 0.982	
DP(1)	**0.003**	-0.004	-0.018 - 0.050		**0.042**	**0.039**	**0.026 - 0.060**		0.956	**0.954**	**0.940 - 0.990**	
DP(2)	0.015	**0.001**	-0.007 - 0.081		0.045	0.044	0.026 - 0.086		0.919	**0.954**	0.700 - 0.966	
N=2,500, P=0.2												
DP(0.5)	-0.005	-0.002	-0.020 - 0.002		0.024	0.023	0.014 - 0.040		0.942	0.942	0.874 - 0.990	
DP(1)	**-0.001**	**-0.001**	**-0.012 - 0.008**		**0.022**	**0.022**	0.014 - 0.032		**0.950**	0.940	0.930 - 0.994	
DP(2)	0.004	0.002	-0.001 - 0.025		**0.022**	0.023	**0.014 - 0.031**		0.949	**0.946**	**0.934 - 0.974**	
N=2,500, P=0.3												
DP(0.5)	-0.003	-0.001	-0.011 - 0.001		0.021	0.021	0.013 - 0.032		0.947	**0.946**	0.916 - 0.982	
DP(1)	**0.000**	**0.001**	**-0.006 - 0.003**		**0.019**	**0.019**	0.013 - 0.028		**0.951**	**0.946**	0.932 - 0.992	
DP(2)	0.004	**0.001**	-0.002 - 0.018		**0.019**	0.021	**0.013 - 0.025**		0.948	0.942	**0.932 - 0.974**	

Table .8. Monte Carlo comparison of different concentration values of the DP(C) method applied to the D index. The $\geq$ signals whether an estimator dominates DP(1) in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

	Bias				RMSE				Confidence Interval			
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
						N=500, P=0.05						
C-DP(0.5)	**0.003**	**-0.003**	**-0.012 - 0.023**		**0.056**	0.058	0.036 - 0.076		0.933	0.914	0.880 - 0.984	
C-DP(1)	0.009	0.006	-0.010 - 0.031		0.058	**0.057**	0.042 - 0.074		**0.933**	**0.942**	0.882 - 0.974	
C-DP(2)	0.016	0.016	-0.010 - 0.042		0.060	**0.057**	**0.051 - 0.072**		0.919	0.930	**0.888 - 0.940**	
						N=500, P=0.2						
C-DP(0.5)	**0.003**	**0.003**	**-0.003 - 0.007**		0.034	**0.036**	0.016 - 0.047		0.933	**0.922**	0.894 - 0.984	
C-DP(1)	0.004	0.004	-0.002 - 0.008		0.034	0.037	0.017 - 0.047		**0.938**	**0.922**	**0.902 - 0.986**	
C-DP(2)	0.005	0.006	-0.001 - 0.008		0.034	0.037	0.017 - 0.047		0.936	0.916	0.890 - 0.986	
						N=500, P=0.3						
C-DP(0.5)	**0.002**	**0.002**	**0.000 - 0.006**	1	**0.031**	**0.034**	0.014 - 0.044		**0.930**	0.920	**0.898 - 0.984**	
C-DP(1)	0.003	0.003	0.001 - 0.007		0.032	**0.034**	0.014 - 0.044		0.928	0.928	0.894 - 0.980	
C-DP(2)	0.004	0.003	0.002 - 0.008		0.032	0.035	0.014 - 0.044		0.929	**0.934**	0.884 - 0.984	
						N=1,500, P=0.05						
C-DP(0.5)	**0.002**	**0.001**	**-0.002 - 0.007**		0.030	0.029	0.015 - 0.043		0.935	0.930	0.892 - 0.974	
C-DP(1)	0.003	0.003	0.000 - 0.008		0.030	0.029	0.016 - 0.043		0.935	0.926	0.898 - 0.976	
C-DP(2)	0.005	0.005	0.000 - 0.010		0.030	0.029	0.017 - 0.043		**0.939**	**0.932**	**0.900 - 0.976**	
						N=1,500, P=0.2						
C-DP(0.5)	0.001	0.001	0.000 - 0.002		0.018	0.019	0.006 - 0.028		0.934	0.934	0.906 - 0.976	
C-DP(1)	0.001	0.001	0.000 - 0.002		0.018	0.019	0.006 - 0.028		0.937	**0.936**	**0.908 - 0.978**	
C-DP(2)	0.001	0.001	0.000 - 0.002		0.018	0.019	0.006 - 0.028		**0.938**	0.930	0.902 - 0.986	
						N=1,500, P=0.3						
C-DP(0.5)	0.000	0.000	-0.001 - 0.002		0.016	0.017	0.005 - 0.025		0.937	0.934	**0.918 - 0.968**	
C-DP(1)	0.000	0.000	-0.001 - 0.002		0.016	0.017	0.005 - 0.025		0.939	**0.938**	0.914 - 0.978	
C-DP(2)	0.000	0.000	-0.001 - 0.002		0.016	0.017	0.005 - 0.025		**0.942**	0.936	0.914 - 0.978	
						N=2,500, P=0.05						
C-DP(0.5)	**0.001**	**0.001**	**-0.002 - 0.004**		0.022	0.022	0.009 - 0.033		0.940	0.930	0.918 - 0.980	
C-DP(1)	0.002	**0.001**	**-0.001 - 0.004**		0.022	0.022	0.009 - 0.033		**0.943**	**0.932**	0.918 - 0.978	
C-DP(2)	0.003	0.002	0.000 - 0.005		0.022	0.022	**0.010 - 0.032**		**0.943**	0.930	0.918 - 0.980	
						N=2,500, P=0.2						
C-DP(0.5)	**0.000**	0.000	**-0.002 - 0.002**		0.013	0.014	0.004 - 0.021		0.939	0.934	**0.918 - 0.966**	
C-DP(1)	**0.000**	0.000	-0.002 - 0.003		0.013	0.014	0.004 - 0.021		**0.943**	0.940	**0.918 - 0.976**	
C-DP(2)	0.001	0.000	-0.002 - 0.003		0.013	0.014	0.004 - 0.021		**0.943**	**0.942**	0.906 - 0.982	
						N=2,500, P=0.3						
C-DP(0.5)	0.000	0.000	-0.002 - 0.002		0.012	0.013	0.003 - 0.019		0.945	**0.944**	**0.932 - 0.968**	
C-DP(1)	0.000	0.000	-0.002 - 0.002		0.012	0.013	0.003 - 0.019		**0.947**	0.942	0.924 - 0.976	
C-DP(2)	0.000	0.000	-0.002 - 0.002		0.012	0.013	0.003 - 0.019		**0.947**	**0.944**	0.922 - 0.984	

Table .9. Monte Carlo comparison of different concentration values of the C-DP(C) method applied to the Theil index. The ≥ signals whether an estimator dominates C-DP(1) in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

	Bias				RMSE				Confidence Interval			
Method	Mean	Median	Range	≥	Mean	Median	Range	≥	Mean	Median	Range	≥
					N=500, P=0.05							
DP(0.5)	**0.006**	**0.001**	**-0.025 - 0.038**		**0.056**	**0.053**	0.039 - 0.075		0.962	**0.952**	0.928 - 0.990	
DP(1)	0.019	0.020	-0.023 - 0.057		0.061	0.062	**0.053 - 0.072**		**0.952**	0.960	**0.926 - 0.962**	
DP(2)	0.034	0.040	-0.025 - 0.080		0.069	0.069	0.060 - 0.084		0.804	0.944	0.286 - 0.972	
					N=500, P=0.2							
DP(0.5)	**0.003**	**0.004**	**-0.016 - 0.018**		**0.034**	**0.035**	0.021 - 0.047		**0.953**	**0.948**	**0.934 - 0.980**	
DP(1)	0.010	0.012	-0.014 - 0.027		0.037	0.036	0.029 - 0.046		0.920	0.924	0.856 - 0.956	
DP(2)	0.018	0.023	-0.014 - 0.039		0.042	0.042	**0.038 - 0.044**		0.783	0.886	0.274 - 0.952	
					N=500, P=0.3							
DP(0.5)	**0.002**	**0.005**	**-0.013 - 0.015**		**0.032**	**0.033**	0.018 - 0.044		**0.937**	**0.932**	**0.920 - 0.966**	
DP(1)	0.008	0.012	-0.011 - 0.023		0.034	0.035	0.024 - 0.043		0.910	0.930	0.808 - 0.936	
DP(2)	0.015	0.021	-0.012 - 0.033		0.038	0.039	**0.033 - 0.042**		0.794	0.878	0.352 - 0.934	
					N=1,500, P=0.05							
DP(0.5)	**0.003**	**0.002**	**-0.006 - 0.015**		**0.030**	**0.028**	0.017 - 0.043		**0.953**	**0.944**	**0.938 - 0.978**	
DP(1)	0.009	0.009	-0.005 - 0.023		0.032	0.029	0.023 - 0.042		0.933	0.940	0.888 - 0.960	
DP(2)	0.015	0.018	-0.006 - 0.033		0.035	0.035	**0.031 - 0.041**		0.791	0.910	0.274 - 0.952	
					N=1,500, P=0.2							
DP(0.5)	**0.001**	**0.002**	**-0.003 - 0.005**		**0.018**	**0.018**	0.007 - 0.028		**0.940**	**0.946**	**0.910 - 0.966**	
DP(1)	0.004	0.006	-0.002 - 0.009		0.019	0.019	0.010 - 0.027		0.921	0.928	0.876 - 0.948	
DP(2)	0.008	0.010	-0.001 - 0.013		0.021	0.021	**0.014 - 0.027**		0.857	0.898	0.596 - 0.950	
					N=1,500, P=0.3							
DP(0.5)	**0.000**	**0.001**	**-0.003 - 0.004**		**0.016**	**0.017**	**0.006 - 0.024**		**0.935**	**0.936**	**0.914 - 0.968**	
DP(1)	0.003	0.004	**-0.002 - 0.007**		0.017	**0.017**	**0.008 - 0.024**		0.926	0.932	0.904 - 0.942	
DP(2)	0.006	0.009	-0.002 - 0.010		0.018	0.019	**0.011 - 0.024**		0.873	0.908	0.654 - 0.936	
					N=2,500, P=0.05							
DP(0.5)	**0.003**	**0.003**	**-0.004 - 0.009**		**0.022**	**0.021**	0.011 - 0.033		**0.946**	**0.946**	**0.908 - 0.972**	
DP(1)	0.007	0.006	-0.003 - 0.014		0.023	**0.021**	0.015 - 0.032		0.917	0.942	0.842 - 0.952	
DP(2)	0.011	0.012	-0.004 - 0.021		0.026	0.023	**0.021 - 0.031**		0.814	0.918	0.392 - 0.952	
					N=2,500, P=0.2							
DP(0.5)	**0.001**	**0.001**	**-0.001 - 0.003**		**0.013**	**0.014**	0.004 - 0.021		**0.942**	**0.944**	0.910 - 0.972	
DP(1)	0.003	0.003	0.001 - 0.005		**0.014**	0.014	0.006 - 0.021		0.937	0.940	**0.926 - 0.946**	
DP(2)	0.005	0.006	0.001 - 0.008		**0.015**	0.015	0.008 - 0.020		0.880	0.912	0.664 - 0.936	
					N=2,500, P=0.3							
DP(0.5)	**0.000**	**0.001**	**-0.003 - 0.002**		**0.012**	**0.013**	**0.004 - 0.018**		**0.944**	**0.944**	**0.926 - 0.972**	
DP(1)	0.002	0.003	-0.001 - 0.004		0.013	0.014	**0.005 - 0.018**		0.937	0.938	0.918 - 0.950	
DP(2)	0.004	0.006	0.000 - 0.006		0.013	0.015	0.007 - 0.019		0.888	0.916	0.732 - 0.948	

Table .10. Monte Carlo comparison of different concentration values of the DP(C) method applied to the Theil index. The ≥ signals whether an estimator dominates DP(1) in a simulation setting. Bolded values represent the best metric in each situation; estimators have their ranges bolded when they have the best metrics from a min-max perspective.

estimates. Banally, for $C = 0$, the DP(0) point estimate will always be 0, regardless of the data – notice that we do not consider this boundary case when we prove consistency. Yet, even in this boundary case, the C-DP(C) technique will produce a better estimate than the DP(C) estimator because it will still change with the data. What we can say for sure is that the influence of C results mild in the previous tests. Moreover, the default value of $C = 1$ appears to provide very good results in all tests.

Concluding, further research about the C parameter is warranted. For example, by equation (A7) the actual influence of C depends on the number of units (O) in an environment: as O augments, the same value of C will result in a smaller probability of a new cluster. Perhaps, the best way to choose C would be to calculate it proportionally to O – whatever O may be. Empirically, for fixed values of $T_{,i}$, $T_{\sum,i}$, α_0, and β_0, setting $C = 1$ will result in a much higher probability of a new cluster in the artificial data of section 1.6 (where $O = 50$) than in the Chicago data of section 1.7 (where $O = 901$). Likely, this means that the tests in the two sections implicitly used quite different settings for the Bayesian techniques, even if the C values were formally identical. Considering that in both sections the Bayesian techniques performed better than the others, this is a further reassurance about the robustness of these techniques. From this perspective, a meaningful diagnostic is the mean probability of being assigned to a new cluster during the Gibbs sampling – the second case in equation (A7). This is ultimately the relevant quantity that C influences. How does this probability changes as C changes? If it is quite robust to changes in C, Bayesian estimates will not be much influenced by this parameter. However, studying C is not only a way to insure robustness for the Bayesian estimators, but also a potential way to get even better estimates. This is a promising venue for future research to further increases the performances of the Bayesian estimators.

D. The Statistical Behavior of the Plug-in Estimator

Previous studies (Cortese et al., 1976; Winship, 1977; Carrington and Troske, 1997) have repeatedly observed bias issues with the plug-in estimator of segregation indices, or used its asymptotic normality to build confidence intervals (Ransom, 2000). In this section we generalize previous results about the statistical behavior of the plug-in estimator, expanding empirical observations through theoretical results.

As a highlight, we show that the plug-in estimator is consistent and we provide sufficient conditions for the plug-in estimators of the D, mutual information, Theil and Atkinson indices to be positively biased. These conditions are very mild and should be most often verified in actual applications, with the possible exception of the conditions for the Theil index. This means that the plug-in estimator is most likely upwardly biased for these indices. To the best of our knowledge, these are the the most general results about the plug-in estimators of segregation indices up-to-date.

D.1. Definition of an Environment, a Sample and a Segregation Index

We start with a statistical definition of an environment whose segregation is to be measured:

Definition D.1. *An environment $\tau \in \mathbb{R}_+^{G \times O}$ is a matrix such that*

1: $\sum_{i=1}^{G} \sum_{j=1}^{O} \tau_{i,j} = 1$

2: $O \geq 2$ *and* $G \geq 2$

3: $\tau_{i,\Sigma} := \sum_{j=1}^{O} \tau_{i,j} > 0$ *for* $i = 1 \dots G$ *and* $\tau_{\Sigma,j} := \sum_{i=1}^{G} \tau_{i,j} > 0$ *for* $j = 1 \dots O$

4: $\tau_{i,j} \geq 0$ *for* $j = 1 \dots O$ *and* $i = 1 \dots G$

We indicate with $\mathbb{R}_\tau$ the subset of $\mathbb{R}_+^{G \times O}$ such $\tau \in \mathbb{R}_\tau$ iff τ satisfies all of the properties above for any $O \geq 2$ and any $G \geq 2$

Notice that we consider only the case where $\tau_{i,j} \geq 0$, in line with Hutchens (1991). We also assume the dimension G and O as fixed and given. This is not a restrictive assumption. When analyzing segregation from a sample, we most often know what groups we will consider and what the occupations (units) are present in the environment. We use $\tau_{i,\sum}$ ($\tau_{\sum,j}$) to indicate the i^{th} (j^{th}) element of the column-sum (row-sum) of τ.

Moreover, notice that we do not consider the case of τ being a random process, like Rathelot (2012) or D'Haultfœuille and Rathelot (2017) do. Considering τ a constant appears more natural in the context considered here, where the objective is to estimate a population (fixed) value from a sample. On a substantial level, considering τ a process entails the assumption that the social process we are examining is constant (stationary) over time. This may be plausible when, for example, we examine school segregation. The student distribution we see today is determined by the students' assignment to schools in the last few years: students usually stay in the same educational institution for less than five years. Therefore, we may assume that the process creating school segregation has been stationary in the last five years. However, the same assumption is questionable in other cases. The occupational segregation we observe from nationally representative surveys is the result of the occupational segregation of many cohorts of workers: it is rather unlikely that the process assigning workers to occupation has remained unchanged in the last 30 years (England et al., 2020). Therefore, while the estimation of a fixed quantity from a population is always a concern as long as we have a sample, estimating the parameters of a random process may be the wrong target in some cases. Thus, we consider τ a constant matrix. As a bonus, this also simplifies (slightly) some of the proofs in Appendix E.

We define two sets of environments that are of interest in the study of segregation indices:

Definition D.2. *A matrix τ is said to be minimally-segregated if $\tau_{i,j} = \tau_{i,\sum} \cdot \tau_{\sum,j}$ for $i = 1 \dots G$, $j = 1 \dots O$. We indicate the set of minimally-segregated matrices with $\mathbb{R}_{\tau_0}$*

Definition D.3. *A matrix τ is said to be maximally-segregated if $\tau_{i,j} > 0 \Rightarrow \tau_{k,j} = 0,\ k \neq i$ for $i = 1 \dots G,\ j = 1 \dots O$. We indicate the set of maximally-segregated environments with $\mathbb{R}_{\tau_1}$*

Simply put, in a minimally-segregated matrix knowing the group to which an individual belongs does not tell us anything about its occupation (unit). On the other hand, in a maximally-segregated matrix, knowing the occupation (unit) of an individual completely reveals her group because no couple of individuals from different groups share the same occupation (unit).

Definition D.4. *A consistent sample is a sequence of random variables $T_1(\tau), T_2(\tau) \dots T_N(\tau)$ with support in $\mathbb{R}_+^{G \times O}$ such that:*

1: $\sum_{i=1}^{G} \sum_{j=1}^{O} T_{i,j,n}(\tau) = n$ *for* $n = 1 \dots N$

2: $\tau_{i,j} = 0 \Rightarrow T_{i,j,n}(\tau) = 0$ *for* $i = 1 \dots G,\ j = 1 \dots O,\ n = 1 \dots N$

3: $\Pi_n(\tau) := \frac{T_n(\tau)}{n} \xrightarrow{P} \tau$

Property 2. simply states that we cannot sample non-existing individuals.

To simplify notation, we write T_n and Π_n instead of $T_n(\tau)$ resp. $\Pi_n(\tau)$ when the reference environment is clear from the context. All of the notation introduced in Section 1.2 will still be used to refer to various components of T_n and Π_n.

In some of the results below, we focus our attention on a sub-set of consistent samples. We explicitly define them:

Definition D.5. *An unbiased sample is a consistent sample such that:*

5: $E[T_n(\tau)] = \tau$ *for* $n = 1 \dots N$

Algorithm 1 Row Reduction

```
$R(T) = 0$
**for** $i = 1 \dots G$ **do**
  **if** $T_{i,\Sigma} \neq 0$ **then**
    **if** $R(T) = 0$ **then**
      $R(T) \leftarrow T_{i,}$
    **else**
      Append $T_{i,}$ to $R(T)$
    **end if**
  **end if**
**end for**
**return** $R(T)$
```

6: $E[\Pi_{o(j)|g(i)}(\tau)] = \frac{\tau_{i,j}}{\tau_{i,\Sigma}}$ *for all* $i = 1 \dots G$, $j = 1 \dots O$, $n = 1 \dots N$

Less formally, a consistent sample is a sampling strategy that eventually converges to the environment being sampled. An unbiased sample further ensures that the sample being drawn is an unbiased sample of the environment for any finite sample size. In particular, the sample in unbiased in the following two senses: the proportion of individuals within each cell of the sample is unbiased; considering only individuals from a specific group, the proportion of individuals in any occupation is unbiased. Most of the sampling schemas used in practice are unbiased (or strive to be) in the sense above. For example, the simple random sampling schema used in the Monte Carlo simulations of sections 1.6 is an unbiased sample.

Algorithm 2 Column Reduction

```
$C(T) = 0$
for $j = 1 \ldots O$ do
    if $T_{,\sum j} \neq 0$ then
        if $R(T) = 0$ then
            $R(T) \leftarrow T_{,j}$
        else
            Append $T_{,j}$ to $C(T)$
        end if
    end if
end for
return $C(T)$
```

However, we need to be careful here. There is no guarantee that $\Pi_n(\tau) \in \mathbb{R}_\tau$ because $\Pi_n(\tau)$ may not satisfy property 2. of the definition of an environment. That is, we may only sample one group or one occupation from the population. Banally, if a sample contains one individual, we will never observe more than one group – assuming an individual belongs to one and only one group. This has implication for the behavior of maximally and minimally segregated environment. Notice that an environment cannot be both maximally and minimally segregated by property 3. of the definition of environment. Yet, outside $\mathbb{R}_\tau$, a maximally segregated matrix can also be a minimally segregated matrix – think, for example, of a $2 \times O$ table where the second row only contains 0s. We need to handle this issue with some care, since we need to define segregation indices for the different T_n that may be observed in real applications. Now, for a matrix τ, let $\tau_{i,}$ be its i^{th} row and $\tau_{,j}$ be its j^{th} column. We define the following:

Definition D.6. *A row reduction is a function $\mathbb{R}_+^{G\times O} \to \mathbb{R}_+^{r\times O}$, $r \leq G$ described in the Row Reduction algorithm*

Definition D.7. *A column reduction is a function $\mathbb{R}_+^{G\times O} \to \mathbb{R}_+^{r\times O}$, with $l \leq O$, described in the Column Reduction algorithm*

Definition D.8. *A matrix to which the Row Reduction and Column Reduction algorithms have been applied will be referred to as a "reduced matrix". We will indicate such matrix with $\rho(T_n)$. We say that a reduced matrix is "degenerate" if $\rho(T_n) \in \mathbb{R}_+^{r\times l}$, with $r = 1$ or $l = 1$, possibly both. If $\rho(T_n)$ is degenerate, we write $\rho(T_n) \in \mathbb{R}^D$*

A reduced matrix has no column or row whose sum is zero: in few words, it ignores those rows and columns. Notice an important property of the reduction algorithms above: $\rho(k \cdot T) = k \cdot \rho(T)$ for $k \neq 0$, implying $\rho(T_n) = n\rho(\Pi_n)$. Therefore, $\rho(T_n) \in \mathbb{R}^D$ iff $\rho(\Pi_n) \in \mathbb{R}^D$.

A reduced table $\rho(T_n)$ is not guaranteed to be in $\mathbb{R}_\tau$ since there is no guarantee that it will have at least 2 rows and 2 columns. However, it is quite clear how to define a segregation index on a matrix with only one row or column: there is no possible segregation in such cases. Moreover, notice that by property 3. of its definition, the probability that a consistent sample draws a degenerate table must diminish for larger samples, since the sample get closer to the (non-degenerate) environment. In other words, we will eventually sample individuals from more than one group and one occupation in an environment with more than one group/occupation.

We are now ready to define a segregation index:

Definition D.9. *A segregation index $S(\tau)$ is a function $\mathbb{R}_+^{G\times O} \to \mathbb{R}$ such that:*

1: $S(\tau)$ *is continuous*

2: $m \leq S(\tau) \leq M$ *for some* $m < M \in \mathbb{R}$

3: $S(\tau) = m$ *iff* $\tau \in \mathbb{R}_{\tau_0}$ *or* $\rho(\tau) \in \mathbb{R}^D$

4: *If* $S(\tau) = M$ *then* $\tau \in \mathbb{R}_{\tau_1}$

5: $S(k \cdot \tau) = S(\tau)$, $k > 0$

6: $S\big(\rho(\tau)\big) = S(\tau)$ *unless* $\rho(\tau) \in \mathbb{R}^D$

Property 5. implies that it is sufficient to examine what happens in $\mathbb{R}_\tau$ to understand the behavior of a segregation index. These properties follow a long-established tradition (Williams, 1948; Jahn et al., 1947; James and Taeuber, 1985; Frankel and Volij, 2011) and, to the best of our knowledge, all popular segregation indices follow this definition. Certainly, the D and Theil indices considered in the main text (as well as the Mutual Information and Atkinson indices, defined below) have the properties in the definition.

D.2. Consistency and Bias of the Plug-in Estimator

Lemma D.1. *The plug-in estimator of a segregation index from a consistent sample,* $\hat{S}_{PI}(\tau) := S(T_n(\tau))$*, is consistent in the sense that* $\hat{S}_{PI}(\tau) \xrightarrow{P} S(\tau)$

Proof. Since $\Pi_n \xrightarrow{P} \tau$, the continuous mapping theorem implies $S(\Pi_n) \xrightarrow{P} S(\tau)$. From property 5. of the definition of $S(\cdot)$ and property 3. of the definition of a consistent sample, $S(T_n) = S(\Pi_n)$. □

While the consistency of the plug-in estimator is important, in any finite sample the plug-in estimator appears to be biased, as discussed in the main text. For the D, Mutual information indices and Atkinson indices, we will prove results that show that, under very mild conditions regarding the sampling schema, this is indeed the case. To specify those conditions, we need a technical results about Jensen inequality.

Theorem D.2. *Let X be a random variable with support within the convex set $\mathscr{X}$. Let $f(x)$ be a convex function defined on $\mathscr{X}$. If a subset $\mathscr{X}_s$ of $\mathscr{X}$ exists where Jensen inequality applies strictly – that is, $f(E[X|X \in \mathscr{X}_s]) > E[f(X|X \in \mathscr{X}_s)]$ – and $P(X \in \mathscr{X}_s) > 0$, then Jensen inequality is strict for X: $f\big(E[X]\big) < E\big[f(X)\big]$*

PROOF.

$$\begin{aligned}
&f(E[X]) = && \text{L.T.E.}\\
&f\Bigg(P(X \notin \mathscr{X}) \cdot E\big[X|X \notin \mathscr{X}_s\big] + P(X \in \mathscr{X}_s) \cdot E[X|X \in \mathscr{X}_s]\Bigg) \leq && \text{Convexity of f(x)}\\
&P(X \notin \mathscr{X}_s) \cdot f\Bigg(E\big[X|X \notin \mathscr{X}_s\big]\Bigg) + P(X \in \mathscr{X}_s) \cdot f\Bigg(E[X|X \in \mathscr{X}_s]\Bigg) < && \text{Hypothesis}\\
&P(X \notin \mathscr{X}_s) \cdot f\Bigg(E\big[X|X \notin \mathscr{X}_s\big]\Bigg) + P(X \in \mathscr{X}_s) \cdot E\Bigg[f\big(X\big)|X \in \mathscr{X}_s\Bigg] \leq && \text{Jensen ineq.}\\
&P(X \notin \mathscr{X}_s) \cdot E\Bigg[f\big(X\big)|X \notin \mathscr{X}_s\Bigg] + P(X \in \mathscr{X}_s) \cdot E\Bigg[f\big(X\big)|X \in \mathscr{X}_s\Bigg] = && \text{L.T.E.}\\
&E\Bigg[f\big(X\big)\Bigg]
\end{aligned}$$

L.T.E. is the law of total expectation; “Jensen ineq.” stands for Jensen inequality. □

When the previous result applies, we will say that “Jensen inequality is anywhere strict” under a certain distribution X. Now we can formulate the main theorem of this section:

Theorem D.3. *Let $S(\tau)$ be a segregation index in the form $S(\tau) = c + \sum_{j=1}^{O} f_j(\Pi_{o(i)|g(1)} \ldots \Pi_{o(i)|g(G)})$, where c is a constant and $f_j(\cdot)$ are convex functions for $j = 1 \ldots O$. Let $T_n(\tau)$ be a* *finite sample from an unbiased sampling, with n fixed. If Jensen inequality is anywhere strict for any of the summands $f_j(\Pi_{o(j)|g(1)}\dots\Pi_{o(i)|g(G)})$, then $\hat{S}_{PI}(\tau)$ is positively biased.*

PROOF. We have to prove $E\big[\hat{S}_{PI}(\tau)\big] > S(\tau)$. This is equivalent to $E\big[S(T_n)\big] > S\big(\tau\big)$ by the definition of the plug-in estimator.

Now:

$$
\begin{aligned}
&E\big[S(T_n)\big] = && \text{Form of } S(T)\\
&E\Bigg[c + \sum_{j=1}^{O} f_j(\Pi_{o(j)|g(1)}\dots\Pi_{o(i)|g(G)})\Bigg] = && \text{Linearity}\\
&c + \sum_{i=1}^{O} E\Bigg[f_j(\Pi_{o(j)|g(1)}\dots\Pi_{o(i)|g(G)})\Bigg] > && \text{Strict Jensen ineq.}\\
&c + \sum_{i=1}^{O} f_j\Bigg(E[\Pi_{o(j)|g(1)}\dots\Pi_{o(i)|g(G)}]\Bigg) = && T_n \text{ is unbiased}\\
&S(\tau)
\end{aligned}
$$

□

Corollary D.3.1. *Let $S(\tau)$ be a segregation index in the form $S(\tau) = c + \sum_{j=1}^{O} f_j(\Pi_{o(j)|g(1)}\dots\Pi_{o(j)|g(G)})$, where c is a constant and $f_j(\cdot)$ is a convex function for $j = 1\dots O$. Let $T_n(\tau)$ be a finite sample from an unbiased sampling, with n fixed. The plug-in estimator of $S(\tau)$ will not be negatively biased*

PROOF. This immediately follows from the proof of Claim D.3 once we relax the condition that Jensen inequality is anywhere strict. □

Given these results, it is now natural to wonder which segregation indices fit the functional form given in Claim D.3. To prove convexity (and, as a result, that an index fits the desired functional form) we will use the following technical lemma. This is necessary because we need to demonstrate the convexity of the function $f_j(\Pi_{o(j)|g(1)} \dots \Pi_{o(j)|g(G)})$ in all variables, even when such function only depends on a sub-set of them.

Lemma D.4. *Let $f(\cdot)$ be a function $\mathbb{R}^{m+n} \to \mathbb{R}$ such that $f(x_1, \dots, x_m, q_1, \dots, q_n) =$ $f(x_1 \dots x_m, r_1 \dots, r_n)$ for all $q_1 \dots q_n, r_1 \dots r_n \in \mathbb{R}$. If $F(\cdot)$ is convex in the first m variables, then it is convex for all variables.*

Proof. The Lemma follows from the definition of convexity. Let $\vec{x}_1$ and $\vec{x}_2$ be two $m \times 1$ vector in $\mathbb{R}^m$, let $\vec{q}_1$ and $\vec{q}_2$ be two $n \times 1$ vectors in $\mathbb{R}^n$, and let $t \in (0,1)$. Then, $f\big(t \cdot \vec{x}_1 + (1-t) \cdot \vec{x}_2,\, t \cdot \vec{q}_1 + (1-t) \cdot \vec{q}_2\big) = f\big(t \cdot \vec{x}_1 + (1-t) \cdot \vec{x}_2, \vec{q}_1\big) \leq t \cdot f\big(\vec{x}_1, \vec{q}_1\big) + (1-t) \cdot f\big(\vec{x}_2, \vec{q}_1\big) = t \cdot f\big(\vec{x}_1, \vec{q}_1\big) + (1-t) \cdot f\big(\vec{x}_2, \vec{q}_2\big)$, where the inequality is due to convexity in $\vec{x}$. □

As in the main text, we will focus on indices that are evenness-based indices following the taxonomy by Massey and Denton (1988). For this purpose, we will introduce two new indices not considered in the main text, but still relevant according to previous works (see for example Frankel and Volij, 2011; Mora and Ruiz-Castillo, 2011): the Atkinson index and the Mutual Information index.

Definition D.10. *The Atkinson index* [14] *is*

$$Atk(\tau) =$$

$$1 - \sum_{j=1}^{O} \big(\prod_{i=1}^{G} \Pi_{o(j)|g(i)}\big)^{\frac{1}{G}} = \tag{A13}$$

$$1 - \sum_{j=1}^{O} \Pi_{\Sigma,j} \left(\prod_{i=1}^{G} \frac{\Pi_{g(i)|o(j)}}{\Pi_{g(i)}} \right)^{\frac{1}{G}} \tag{A14}$$

Definition D.11. *The Mutual Information index is*

$$MI(\tau) =$$

$$\sum_{i=1}^{G} \Pi_{i,\Sigma} \left(H[\Pi_{\Sigma,}] - H\left[\frac{\Pi_{i,}}{\Pi_{i,\Sigma}}\right] \right) = \tag{A15}$$

$$\sum_{j=1}^{O} \Pi_{\Sigma,j} \left(H[\Pi_{,\Sigma}] - H\left[\frac{\Pi_{,j}}{\Pi_{\Sigma,j}}\right] \right) = \tag{A16}$$

$$H(\Pi_{,\Sigma}) Th(\tau)$$

where the $H(\cdot)$ is the entropy function defined in equation (1.3) of the main text; $\Pi_{,\Sigma}$ *(*$\Pi_{\Sigma,}$*) is the vector having as its* i^{th} *(*j^{th}*) element* $\Pi_{\Sigma,i}$ *(*$\Pi_{j,\Sigma}$*), and* $\Pi_{i,}$ *(*$\Pi_{,j}$*) is simply the* i^{th} *(*j^{th}*) row (column) of* Π

Notice that we do not limit our discussion to two groups, but consider the Atkinson and Mutual Information indices as multi-group indices. Given its strict relation with the Mutual Information index, we also consider the Theil index as a multi-group index (Reardon and O'Sullivan, 2004) –

[14] We consider only one possible parametrization of the Atkinson index. The index has a tunable parameter whose effects we are not considering in this context. For a discussion of its properties and parameters, see Frankel and Volij (2011).

this is a generalization of the simulations in the main text. We still consider the D index merely a two groups index, since it is not clear how to generalize it to more than two groups.

Lemma D.5. *Let $T_n(\tau)$ be an unbiased sampling such that Jensen inequality is anywhere strict for $f_j(\cdot) = |\Pi_{o(j)|g(1)} - \Pi_{o(j)|g(2)}| + 0 \cdot \sum_{i=1}^{2} \sum_{k \neq j} \Pi_{o(k)|g(i)}$. Then, $\hat{D}_{PI}(\tau)$ is positively biased.*

Proof. Consider the function $g_j\big(\Pi_{o(j)|g(1)}, \Pi_{o(j)|g(2)}\big) = |\Pi_{o(j)|g(1)} - \Pi_{o(j)|g(2)}|$. All the elements of its Hessian $\mathbf{H}_{g_j}$ are zero. Therefore, $\mathbf{H}_{g_j}$ is positive semi-definite and g_j is convex. By Lemma D.4 this means that $f_j(\cdot) = |\Pi_{o(j)|g(1)} - \Pi_{o(j)|g(2)}| + 0 \cdot \sum_{i=1}^{2} \sum_{k \neq j} \Pi_{o(k)|g(i)}$ is convex.

From here, the D index fits the mold of Claim D.3 with $c = 0$, $f_j(\cdot) = |\Pi_{o(j)|g(1)} - \Pi_{o(j)|g(2)}| + 0 \cdot \sum_{i=1}^{2} \sum_{k \neq j} \Pi_{o(k)|g(i)}$. □

Lemma D.6. *Let $T_n(\tau)$ be an unbiased sample such that $P\big(Atk(T_n) = Atk(\tau)\big) \neq 1$, then $\hat{Atk}_{PI}(\tau)$ is positively biased.*

Proof. Consider the function, $-f_j(\cdot) = \big(\prod_{i=1}^{G} \Pi_{o(j)|g(i)}\big)^{\frac{1}{G}}$. For its Hessian, $\mathbf{H}_{f_j}$ we have:

$$
\begin{aligned}
&\vec{a}^T \mathbf{H}_{f_j} \vec{a} = \\
&\frac{1}{\prod_{i=1}^{G} (\Pi_{o(j)|g(i)})^{2-\frac{1}{G}}} \cdot \Bigg(-(G-1) \sum_{i=1}^{G} a_i^2 \prod_{k \neq i} \Pi^2_{o(j)|g(k)} \\
&+ 2 \sum_{i=1}^{G-1} \sum_{k=i+1}^{G} a_i a_k \Pi_{o(j)|g(i)} \Pi_{o(j)|g(k)} \prod_{l \neq i,k} \big(\Pi_{o(j)|g(l)}\big)^2 \Bigg) = \\
&\frac{1}{\prod_{i=1}^{G} (\Pi_{o(j)|g(i)})^{2-\frac{1}{G}}} \cdot \Bigg(-\sum_{i=1}^{G-1} \sum_{k=i+1}^{G} \prod_{l \neq i,k} \Pi^2_{o(j)|g(l)} (a_i \Pi_{o(j)|g(k)} - a_k \Pi_{o(j)|g(i)})^2 \Bigg) < 0
\end{aligned}
$$

Notice that the last inequality is strict because at least one $\Pi_{o(j)|g(i)}$ will be greater than 0. Therefore, $-f_j(\Pi_{o(j)|g(G)} \ldots \Pi_{o(j)|g(G)})$ is strictly concave in the relevant domain and $f_j(\cdot)$ is strictly convex in the same region. From here, the Atkinson index fits the mold of Claim $D.2$ with $c = 1$ and $f_j(\Pi_{o(j)|g(1)} \dots \Pi_{o(j)|g(G)}) = -\left(\prod_{i=1}^{G} \Pi_{o(j)|g(i)}\right)^{\frac{1}{G}}$.

Since $f_j(\Pi_{o(j)|g(G)} \dots \Pi_{o(j)|g(G)})$ is strictly convex, Jensen inequality will always be strict as long as there will be variation in any of the f_j functions. If there is no variation, then Jensen inequality will transform into an equality and $Atk(T_n(\tau)) = Atk(\tau)$ with probability 1. □

The bias of the plug-in estimator of the Mutual Information index can also be proved using Jensen inequality as well, but we can take a more direct approach using results from information theory.

Lemma D.7. *Let $T_n(\tau)$ be an unbiased sample such that Jensen inequality is anywhere strict for the Mutual Information between the "group" and the "occupation" variables, then $\hat{MI}_{PI}(\tau)$ is positively biased.*

Proof. The Mutual Information index is a direct transposition of Mutual Information from information theory (Theil and Finizza, 1971). Specifically, the Mutual Information index measures the information between the $\Pi_{\Sigma,}$ and $\Pi_{,\Sigma}$, which we will indicate with $\Pi_G(\tau)$ and $\Pi_O(\tau)$, respectively. These are the "occupation" and "group" variables. We indicate the Mutual Information between the "group" and "occupation" variable from a sample with $I(\Pi_G(T_n);\Pi_O(T_n))$, where $\Pi_G(T_n)$ and $\Pi_O(T_n)$ will be measured on the sample T_n. Thus, $\hat{MI}_{PI}(\tau) = MI(T_n) = I(\Pi_G(T_n);\Pi_O(T_n))$.

Now, let us indicate with $\tau_0\big(T_n\big)$ the table such that $\tau_{0,i,j}(T_n) = \Pi_{g(i)}(T_n) \cdot \Pi_{o(j)}(T_n)$. Notice that this table is minimally-segregated: $MI\big(\tau_0(T_n)\big) = 0$. It is the case that $MI(T_n) = I(\Pi_G(T_n);\Pi_O(T_n))$ = $KL\big(T_n||\tau_0(T_n)\big)$, where $KL(\cdot)$ is the Kullback-Leibner divergence between two distributions (Cover and Thomas, 2006, Paragraph 2.3).

It is well known (Cover and Thomas, 2006, Theorem 2.7.2) that $KL(\cdot)$ is convex in the following sense: $KL(\lambda \cdot P_1 + (1-\lambda)P_2 || \lambda \cdot Q_1 + (1-\lambda)Q_2) \leq \lambda \cdot KL(P_1||Q_1) + (1-\lambda)KL(P_2||Q_2)$ with $0 \leq \lambda \leq 1$, where P_1, P_2, Q_1 and Q_2 are probability mass functions on identical support. Let us indicate with t_1 and t_2 two matrices with the same dimensions as T_n summing to 1, then the mutual information index is convex in the support of T_n:

$$
\begin{aligned}
& MI\Big(\lambda \cdot t_1 + (1-\lambda) \cdot t_2\Big) = \\
& KL\Big(\lambda \cdot t_1 + (1-\lambda) \cdot t_2 || \lambda \cdot \tau_0(t_1) + (1-\lambda) \cdot \tau_0(t_2)\Big) \leq \\
& \lambda \cdot KL\big(t_1 || \tau_0(t_1)\big) + (1-\lambda) \cdot KL\big(t_2 || \tau_0(t_2)\big) = \\
& \lambda \cdot MI(t_1) + (1-\lambda) \cdot MI(t_2)
\end{aligned}
$$

From here, we can apply Jensen inequality:

$$E\big[\hat{M}I_{PI}(T_n)\big] = E\big[MI(T_n)\big] \geq MI\big(E[T_n]\big) = MI(\tau)$$

The inequality will be tight when Jensen inequality is anywhere tight for $I(\Pi_G(T_n);\Pi_O(T_n))$. □

Lemma D.8. *Let $T_n(\tau)$ be an unbiased sample such that a. Jensen inequality is anywhere strict for the Mutual Information between the "group" and the "occupation" variables and b. $Cov\Bigg(\frac{1}{H[\Pi_G(T_n)]}, I\big(\Pi_G(T_n), \Pi_O(T_n)\big)\Bigg) > 0$, then $\hat{T}h_{PI}(\tau)$ is positively biased.*

Proof. Notice that $Th(\tau) = \frac{1}{H[\tau_{i,}]}MI(\tau)$. We also notice that $\frac{1}{H(\cdot)}$ is a convex function since it is the composition of a concave function ($H(\cdot)$) with a convex non-increasing function, $(\frac{1}{x})$. From here:

$$E\left[Th(T_n)\right] = E\left[\frac{1}{H[\Pi_G(T_n)]}MI(T_n)\right] > Th(\tau) \Leftrightarrow$$
$$E\left[\frac{1}{H[\Pi_G(T_n)]}MI(T_n)\right] - E\left[\frac{1}{H[\Pi_G(T_n)]}\right]E\left[MI(T_n)\right] > Th(\tau) - E\left[\frac{1}{H[\Pi_G(T_n)]}\right]E\left[MI(T_n)\right] \Leftrightarrow$$
$$Cov\left(\frac{1}{H[\Pi_G(T_n)]}, MI(T_n)\right) > Th(\tau) - E\left[\frac{1}{H[\Pi_G(T_n)]}\right]E\left[MI(T_n)\right]$$

From convexity and the hypotheses, we have

$$E\left[\frac{1}{H[\Pi_G(T_n)]}\right] \geq \frac{1}{H\left[E[\Pi_G(T_n)]\right]} = \frac{1}{H\left[\tau_{i,}\right]}$$
$$E\left[MI(T_n)\right] > MI\big(E[T_n]\big) = MI\big(\tau\big)$$

Indeed, for unbiased samples $E[\Pi_{g(i)}(T_n)] = \sum_{j=1}^{O} E[\Pi_{g(i),o(j)}] = \tau_{i,}$. This implies

$$E\left[\frac{1}{H[\Pi_G(T_n)]}\right]E\left[MI(T_n)\right] > \frac{1}{H\left[\tau_{i,}\right]}MI(\tau) = Th(\tau)$$

From the hypotheses

$$Cov\left(\frac{1}{H[\Pi_G(T_n)]}, MI(T_n)\right) > 0 > Th(\tau) - E\left[\frac{1}{H[\Pi_G(T_n)]}\right]E\left[MI(T_n)\right]$$

□

The previous Lemmas specify sufficient conditions for the plug-in estimator to be biased for the D, Theil, Atkinson and Mutual Information index. The next question is whether these sufficient conditions will be encountered in practice. The first condition is that the sample should be unbiased. This seems straightforward. Most complex and simple design schemas are designed to be unbiased. The condition that Jensen inequality is anywhere strict for the relevant function is very mild. Consider for the example the D index. This condition is satisfied if it is possible to sample two tables such, for any of $j = 1 \ldots O$, in the first table $\Pi_{o(j)|g(1)} > \Pi_{o(j)|g(2)}$, while in the second table for the same j $\Pi_{o(j)|g(1)} < \Pi_{o(j)|g(2)}$. Most likely, this condition will be met if there is an occupation anywhere close to have an equal representation of both groups in the population. For the Mutual Information and Atkinson index, the possibility of sampling any two different tables will satisfy the condition.

Therefore, the sufficient conditions is very mild for the D, Atkinson and Mutual Information. However, this may not be the case for the Theil index, which requires a specific covariance to be positive – see Lemma D.8. In the simulations, we empirically observed that this covariance is very close to 0, but still negative. This is not enough to make the Theil index unbiased, as should be clear from the empirical results (see Figure 1.2). However, it is an indication that the conditions given above to justify the bias of the Theil index may need further generalization and they appear to not cover the empirical results in the main text.

After these results about specific indices, let us consider what happens for any segregation index at the boundaries – that is, the estimation of $S(\tau)$ when $\tau \in \mathbb{R}_{\tau_0}$ or $\tau \in \mathbb{R}_{\tau_1}$. The next claim proves and generalizes the empirical observations of Cortese et al. (1976) and Winship (1977) about the positive bias of the plug-in estimator when there is no segregation in the environment. The claim states that the plug-in estimator of any index is positively biased when there is no segregation in the population, $\tau \in \mathbb{R}_{\tau_0}$, unless the segregation index in the sample is not really random – that is, unless $P(S(T_n) = m) = 1$.

Theorem D.9. *Let $T_n(\tau)$ be a consistent sample such that $P(\rho(T_n) \in \mathbb{R}_{\tau_0} \cup \mathbb{R}^D) < 1$. If $\tau \in \mathbb{R}\tau_0$, the plug-in estimator is biased: $E[\hat{S}_{PI}(\tau)] > S(\tau)$*

Proof. We define $\mathbb{R}_m = \mathbb{R}_{\tau_0} \cup \mathbb{R}^D$. Notice $S(T_n) = m$ iff $\rho(T_n) \in \mathbb{R}_m$ by property 6. and 3. of the definition of a segregation index.

Let $\mathbb{R}_{\tau_>} = \cup_{r=1}^{G} \cup_{l=1}^{O} [\mathbb{R}_+^{r\times l} - \mathbb{R}_m^{r\times l}]$. Let $I(x,\ Q)$ be the indicator function such that $I(x,\ Q) = 1$ if $x \in Q$ and $I(x,\ Q) = 0$ otherwise. Now, consider the two random variables $S(\Pi_{n,>}) := S\big(\rho(\Pi_n)\big) \cdot I(\Pi_n, \mathbb{R}_{\tau_>})$ and $S(\Pi_{n,0}) := S\big(\rho(\Pi_n)\big) \cdot I(\Pi_n, \mathbb{R}_{\tau_m})$. Notice that $S(\Pi_{n,>}) > m$ when $\Pi_n \in \mathbb{R}_{\tau_>}$.

From here, we can write

$$
\begin{aligned}
&E[\hat{S}_{PI}(\tau)] = E[S(\Pi_n)] = \\
&E[S(\Pi_{n,0})] + E[S(\Pi_{n,>})] = \\
&P(\Pi_n \in \mathbb{R}_{\tau_m}) \cdot \Big(E[S(\Pi_{n,0})|\Pi_n \in \mathbb{R}_{\tau_m}] + E[S(\Pi_{n,>})|\Pi_n \in \mathbb{R}_{\tau_m}]\Big) + \\
&P(\Pi_n \in \mathbb{R}_{\tau_>}) \cdot \Big(E[\Pi_{n,0}|\Pi_n \in \mathbb{R}_{\tau_>}] + E[S(\Pi_{n,>})|\Pi_n \in \mathbb{R}_{\tau_>}]\Big) = \\
&P(\Pi_n \in \mathbb{R}_{\tau_m}) \cdot (m+0) + P(\Pi_n \in \mathbb{R}_{\tau_>}) \cdot \Big(0 + E[S(\Pi_{n,>})|\Pi_n \in \mathbb{R}_{\tau_>}]\Big) > \\
&P(\Pi_n \in \mathbb{R}_{\tau_m}) \cdot m + P(\Pi_n \in \mathbb{R}_{\tau_>}) \cdot m = m = S(\tau_0)
\end{aligned}
$$

Therefore, $E[\hat{S}_{PI}(\tau)] > S(\tau_0)$. □

Second, the following claim provides provides a counter-result when the sampled environment is maximally-segregated. For most indices, the plug-in estimate is likely to have no error for maximally-segregated environments – leaving aside matrices in $\mathbb{R}^D$.

Theorem D.10. *Let $S(\cdot)$ be a segregation index such that $S(\tau) = M$ iff $\tau \in \mathbb{R}_{\tau_1}$. If $S(\tau) = M$ and $\rho\big(T_n(\tau)\big) \notin \mathbb{R}^D$, then $S\big(T_n(\tau)\big) = S(\tau)$*

Proof. Property 2. of the definition of a consistent sample implies that $\tau \in \mathbb{R}_{\tau_1} \Rightarrow \rho\big(T_n(\tau)\big) \in \mathbb{R}_{\tau_1}$. Since $\rho(T_n) \notin \mathbb{R}^D$, the claim follows from the hypothesis. □

When the lemma applies, it implies that the plug-in estimator will always be equal to M if the environment is maximally segregated – naturally, the reverse is not true. However, the lemma does not always apply because not every segregation index is such that $\tau \in \mathbb{R}_{\tau_1}$ iff $S(\tau) = M$. Most noticeably, the lemma does not apply to the Mutual Information and Theil indices (Mora and Ruiz-Castillo, 2011) where not every $\tau \in \mathbb{R}_{\tau_1}$ achieves the maximum the indices are sensible to change in the group margins. Yet, the lemma applies to other popular indices such as the D, Atkinson and Gini indices.

E. Consistency of the DP(C) and C-DP(C) Methods

In this section, we will show that the $DP(C)$ and the $C-DP(C)$ are consistent estimator of a segregation index if the plug-in index is consistent. That is, if T_n is a consistent sample following Definition D.4, DP(C) and C-DP(C) will be consistent – like the plug-in estimator.

To do this, we will show that the Gibbs sampler of Appendix B produces consistent posterior draws. We consider the case of two-groups tables that we analyze in the main text. We signal dependency on sample size n by adding the subscript n to notation when the domain of a variable depends on it – for example, $T_{h,k,n}$ is the h^{th} row, k^{th} column of the sample table T, dependent on sample size n. To simplify exposition, we assume that in the environment table τ no cell equals 0 – this is only relevant in the evaluation of the first limit. From property 2. of the definition of consistent sample, one can show that in case an occupation has a cell equal to 0, its probability of being assigned to a new cluster is 1. Therefore, this assumption is not relevant for the final conclusion.

E.1. Consistency of DP(C)

We will start by limit behavior of equation (A7), specifically for the creation of a new cluster, $k = K(-i)+1$:

$$\lim_{N\to\infty} P(c(i) = K(-i)+1|\vec{c},\vec{\Phi},\vec{\Pi}) =$$

$$\lim_{N\to\infty} \frac{\frac{C}{O-1+C}\cdot dBBin(T_{1,i};\alpha_0,\beta_0,T_{\sum,i})}{\frac{C}{O-1+C}\cdot dBBin(T_{1,i};\alpha_0,\beta_0,T_{\sum,i}) + \sum_{h=1}^{K(-1)}\frac{O(k_{-i})}{O-1+C}\cdot dBin(T_{1,i};\Phi_h,T_{\sum,i})} =$$

$$\lim_{N\to\infty} \frac{1}{1+\sum_{h=1}^{K(-1)}\frac{O(k_{-i})\cdot dBin(T_{1,i};\Phi_h,T_{\sum,i})}{C\cdot dBBin(T_{1,i};\vec{\alpha}_0,T_{\sum,i})}}$$

Let us focus on the addends in the denominator:

$$\lim_{n\to\infty} \frac{O(k_{-i}) \cdot dBin(T_{,i}; \Phi_h, T_{\sum,i})}{C \cdot dBBin(T_{,i}; \alpha_0, \beta_0, T_{\sum,i})} \propto$$

$$\lim_{n\to\infty} \Gamma(T_{\Sigma,i,n} + \alpha_0 + \beta_0) \frac{(\Phi_h^{T_{1,j,n}})(1-\Phi_h)^{T_{2,j,n}}}{\Gamma(T_{1,i,n} + \alpha_0)\Gamma(T_{2,i,n} + \beta_0)} =$$

$$\lim_{n\to\infty} (T_{\Sigma,i,n} + \alpha_0 + \beta_0)^{T_{\Sigma,i,n} + \alpha_0 + \beta_0 - \frac{1}{2}} \prod_{g=1}^{G} \frac{(\Phi_h^{T_{1,j,n}})(1-\Phi_h)^{T_{2,i,n}}}{(T_{1,i,n} + \alpha_0)^{T_{1,i,n} + \alpha_0 - \frac{1}{2}} \cdot (T_{2,i,n} + \beta_0)^{T_{2,i,n} + \beta_0 - \frac{1}{2}}}$$

In the last passage, we apply the Stirling's formula to the Gamma function: $\Gamma(x) = \sqrt{2\pi} x^{x-\frac{1}{2}} exp(-x)$.

By the continuous mapping theorem, we can evaluate the previous limit using the τ table to which $\frac{T_n}{n}(\tau)$ is converging in probability – see Definition D.4. It is convenient to write $\tau_{1,i} = v_{1,i} \cdot \tau_{\sum,i}$ and $\tau_{2,i} = (1 - v_{1,i}) \cdot \tau_{\sum,i}$, where $0 \leq v_{1,i} \leq 1$. In words, $v_{1,i}$ is the proportion of individuals from group 1 among individuals in occupation i in the sampled environment, τ. From here, we may write $T_{\sum,i,n} \xrightarrow{P} n \cdot \tau_{\sum,i}$, $T_{1,i,n} \xrightarrow{P} n\tau_{1,i} = n \cdot v_{1,i} \cdot \tau_{\sum,i}$ and $T_{2,i,n} \xrightarrow{P} n \cdot (1 - v_{1,i}) \cdot \tau_{\sum,i}$.

Therefore, we can evaluate:

$$\lim_{n\to\infty}(n\cdot\tau_{\Sigma,i}+\alpha_0+\beta_0)^{n\cdot\tau_{\Sigma,i}+\alpha_0+\beta_0-\frac{1}{2}}\cdot$$

$$\cdot\frac{\Phi_h^{n\cdot v_{1,i}\cdot\tau_{\Sigma,i}}(1-\Phi_h)^{n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}}}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)^{n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0-\frac{1}{2}}\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)^{n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0-\frac{1}{2}}}=$$

$$\lim_{n\to\infty}\left(\frac{n\tau_{\Sigma,i}+\alpha_0+\beta_0}{n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0}\right)^{\alpha_0}\left(\frac{n\tau_{\Sigma,i}+\alpha_0+\beta_0}{n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0}\right)^{\beta_0}\cdot$$

$$\left(\frac{n\tau_{\Sigma,i}+\alpha_0+\beta_0}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)}\right)^{-\frac{1}{2}}\cdot$$

$$\left((n\cdot\tau_{\Sigma,i}+\alpha_0+\beta_0)\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)^{v_{1,i}}\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)^{1-v_{1,i}}}\right)^{n\cdot\tau_{\Sigma,i}}\propto$$

$$\lim_{n\to\infty}\left(\frac{n\tau_{\Sigma,i}+\alpha_0+\beta_0}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)}\right)^{-\frac{1}{2}}\cdot$$

$$\left((n\cdot\tau_{\Sigma,i}+\alpha_0+\beta_0)\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)^{v_{1,i}}\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)^{1-v_{1,i}}}\right)^{n\cdot\tau_{\Sigma,i}}=$$

$$\lim_{n\to\infty}\left(\frac{n\tau_{\Sigma,i}+\alpha_0+\beta_0}{(n\cdot v_{1,i}\cdot\tau_{\Sigma,i}+\alpha_0)\cdot(n\cdot(1-v_{1,i})\cdot\tau_{\Sigma,i}+\beta_0)}\right)^{-\frac{1}{2}}\cdot\left(\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{v_{1,i}^{v_{1,i}}(1-v_{1,i})^{1-v_{1,i}}}\right)^{n\cdot\tau_{\Sigma,i}}$$

The last passage comes from repeatedly applying L'Hôpital's rule to the factor within the parenthesis.

Now, The factor $\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{v_{1,i}^{v_{1,i}}(1-v_{1,i})^{1-v_{1,i}}}$ is maximized when $\Phi_{1,h}=v_{1,i}$, in which case this factor is clearly 1. In practice, this will happen when the h^{th} existing cluster have a minority proportion that is exactly equal to the minority proportion in the i^{th} column of τ. This is an important distinction for what follows and it is worth creating an indicator function.

$$I_{h_i}=\begin{cases}1 & \text{if } \Phi_h=v_{1,i}\text{, for any } h\\ 0 & \text{otherwise}\end{cases}\tag{A17}$$

If we assume $I_{h_i} = 0$, we may write $\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{v_{1,i}^{v_{1,i}}(1-v_{1,i})^{1-v_{1,i}}} = r_i$, where $0 < r_i < 1$. From here, we can evaluate the logarithm of the limit, given that all factors are strictly positive:

$$\lim_{n\to\infty}\Bigg(-\frac{1}{2}\log\big(n\cdot\tau_{\Sigma,i}+\alpha_0+\beta_0\big)+\frac{1}{2}\log(n\cdot v_{g,i}\cdot\tau_{\Sigma,i}+\alpha_0)$$

$$+\frac{1}{2}\log(n\cdot(1-v_{g,i})\cdot\tau_{\Sigma,i}+\beta 0)+n\cdot\tau_{\Sigma,i}\log r_i\Bigg) =$$

$$-O\big(\log n\big)+O\big(\log n\big)+O\big(\log n\big)-O\big(n\big) =$$

$$-O\big(n\big) = -\infty$$

Thus, the limit is 0 when $\Phi_{1,h} \neq v_{1,i}$. On the other hand, when $\Phi_{1,h} = v_{1,i}$, the limit will go to infinity since $\frac{\Phi_h^{v_{1,i}}(1-\Phi_h)^{1-v_{1,i}}}{v_{1,i}^{v_{1,i}}(1-v_{1,i})^{1-v_{1,i}}} = 1$.

This implies that the probability of creating a new cluster behaves as follows:

$$P(c(i) = K(-i)+1|\vec{c},\vec{\Phi},\vec{\Pi}) \xrightarrow{P} \begin{cases} 0 & \text{if } I_{h_i} = 1 \\ 1 & \text{otherwise} \end{cases} \tag{A18}$$

With a very similar derivation, one can show that

$$P(c(i) = h|\vec{c},\vec{\Phi},\vec{\Pi}) \xrightarrow{P} \begin{cases} r_h & \text{if } I_{h_i} = 1 \text{ and } \Phi_h = v_{1,i} \\ 0 & \text{otherwise} \end{cases} \tag{A19}$$

where $0 < r_h \leq 1$ is a scalar between 0 and 1 dependent on the number on the population table τ and the cluster assignment $\vec{\Phi}$. To simplify the proof, we will assume that at most one cluster h exists such that $\phi_{1,h} = v_{1,i}$; under this assumption $r_h = 1$ since probabilities must sum to 1 and the probability of occupation i being assigned to any other cluster tends to 0.

We indicate with V_i the event that for the i^{th} occupation (a) $c(i) = K(-1)+1$ when $I_{h_i} = 0$ or (b) $c(i) = h$ when $I_{h_i} = 1$, where $\Phi_h = v_{1,k}$ for all occupations in cluster h. We create another indicator variable I_{V_i}. $I_V(i)$ is one if V_i is verified. Notice the distribution of V_i:

$$\begin{aligned} I_{V_i} \sim & I_{h_i} Ber\Big(P\big(c(i) = h|\vec{c}, \vec{\Phi}, \vec{P}i, I_{h_i} = 1\big)\Big) + \\ & (1 - I_{h_i}) Ber\Big(P\big(c(i) = K(-i) + 1|\vec{c}, \vec{\Phi}, \vec{P}i, I_{h_i} = 0\big)\Big) \xrightarrow{P} \\ & I_{h_i} Ber\big(1\big) + (1 - I_{h_i}) Ber\big(1\big) = Ber\big(1\big) \end{aligned}$$

where $Ber(\pi)$ is a Bernoulli variable with probability π. The variables in this series are uniformly bounded between 0 and 1, given the support of any Bernoulli distribution. Moreover, the square of the variables ($I_{V_i}^2 = |I_{V_i}|^2$) are bounded as well. Therefore, the variables and their squares are uniformly integrable. By the uniform integrability theorem, this implies that

$$E[I_{V_i}] = E[|I_{V_i}|] \to E[Ber\big(1\big)] = 1$$

$$E[I_{V_i}^2] = E[|I_{V_i}|^2] \to E[Ber\big(1\big)^2] = 1$$

From here, the variance

$$Var[I_{V_i}] = E[I_{V_i}^2] - E[I_{V_i}]^2 \to 0$$

We can apply the plug-in the previous results in Chebyshev inequality to directly obtain:

(A20) $$I_{V_i} \xrightarrow{P} 1$$

Now consider the case where $I_{V_i} = 1$ and $I_{h_i} = 0$. The new cluster will draw a group proportion following equation (A8):

$$\Phi_{K(-i)+1} \sim Beta(\alpha_0 + T_{1,i,n}, \beta_0 + T_{\sum,i,n} - T_{1,i,n}) \xrightarrow{P}$$
$$Beta\left(n \cdot \tau_{\sum,i,n} \cdot (\frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i}), n \cdot \tau_{\sum,i} \cdot (\frac{\beta_0}{n \cdot \tau_{\sum,i}} + 1 - v_{1,i})\right)$$

Once again the variables in the series and their squares are uniformly bounded. Notice:

$$E\left[|\Phi_{K(-i)+1}|\right] = E\left[\Phi_{K(-i)+1}\right] \to E\left[Beta\left(n \cdot \tau_{\sum,i} \cdot \ (\frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i}), n \cdot \tau_{\sum,i} \cdot (\frac{\beta_0}{n \cdot \tau_{\sum,i}} + 1 - v_{1,i})\right)\right] =$$
$$\lim_{n\to\infty} \frac{\frac{\alpha}{n \cdot \tau_{\sum,i}} + v_{1,i}}{\frac{\beta_0}{n \cdot \tau_{\sum,i}} + 1 - v_{1,i} + \frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i}} = v_{1,i}$$

$$E\left[|\Phi_{K(-i)+1}|^2\right] = E\left[\Phi^2_{K(-i)+1}\right] \to E\left[Beta\left(n \cdot \tau_{\sum,i} \cdot (\frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i}), n \cdot \tau_{\sum,i} \cdot (\frac{\beta_0}{n \cdot \tau_{\sum,i}} + 1 - v_{1,i})\right)^2\right] =$$
$$\lim_{n\to\infty} \frac{(\frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i} + \frac{1}{n \cdot \tau_{\sum,i}})(\frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i})}{(\frac{\beta_0}{n} + 1 - v_{1,i} + \frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i})(\frac{\beta_0}{n \cdot \tau_{\sum,i,n}} + 1 - v_{1,i} + \frac{\alpha_0}{n \cdot \tau_{\sum,i}} + v_{1,i} + \frac{1}{n \cdot \tau_{\sum,i}})} = v^2_{1,i}$$

The two results together imply that:

$$Var\left[\Phi_{K(-i)+1}\right] = E[\Phi^2_{K(-i)+1}] - E[\Phi_{K(-i)+1}]^2 \to 0$$

From here, we can apply Chebyshev inequality to obtain:

$$\Phi_{K(-i)+1} | I_{h_i} = 0, I_{v_i} = 1 \xrightarrow{P} v_{1,i} \Rightarrow$$

(A21) $$\Pi_i | I_{h_i} = 0, I_{v_i} = 1 \xrightarrow{P} v_{1,i}$$

where Π_i is the group proportion associated with the i^{th} cluster, following the same notation as Appendix B.

Consider the case where $I_{V_i} = 1$ and $I_{h_i} = 1$. More than one occupation can be assigned to this cluster by the end of part I from Appendix B. Consider the case that $v_{1,k} = \Phi_h$ for all of the assigned occupation, implying that they all have identical $v_{1,k}$ in the population table τ. This cluster will draw a new group proportion according to equation (A8). Let $\mathscr{V}$ be the sets of such occupations being clustered together in cluster h. Following the same logic as above, we have that $N(h) = \sum_{i \in \mathscr{V}} T_{\sum,i} \xrightarrow{P} n \cdot \sum_{i \in \mathscr{V}} \tau_{\sum,i} := n \cdot \tau_{\sum,\mathscr{V}}$ and $G(h) = \sum_{i \in \mathscr{V}} T_{1,i,n} \xrightarrow{P} n \cdot \sum_{i \in \mathscr{V}} v_{1,i} \tau_{\sum,i} = v_{1,i} \tau_{\mathscr{V},i}$. Now, we have that the group proportion for cluster h will be

$$\begin{aligned}\Phi_h \sim & Beta(\alpha_0 + G(h), \beta_0 + N(h) - G(h)) \xrightarrow{P} \\ & Beta\left(n \cdot \tau_{\sum,\mathscr{V}} \cdot (\frac{\alpha_0}{n \cdot \tau_{\sum,\mathscr{V}}} + v_{1,k}), n \cdot \tau_{\sum,\mathscr{V}} \cdot (\frac{\beta_0}{n \cdot \tau_{\sum,\mathscr{V}}} + 1 - v_{1,k}) \right)\end{aligned}$$

A similar argument to the one above shows

$$\begin{aligned}& \Phi_h | I_{h_i} = 1, I_{v_i} = 1 \xrightarrow{P} v_{1,i} \Rightarrow \\ \text{(A22)} \qquad & \Pi_i | I_{h_i} = 1, I_{v_i} = 1 \xrightarrow{P} v_{1,i}\end{aligned}$$

Putting equations (A20), (A21), (A22) together:

(A23)
$$
\begin{aligned}
\Pi_i =& \big(\Pi_i | I_{h_i} = 1, I_{V_i} = 1\big) \cdot I_{h_i} \cdot I_{V_i} + \\
& \big(\Pi_i | I_{h_i} = 0, I_{V_i} = 1\big) \cdot (1 - I_{h_i}) \cdot I_{V_i} + \\
& \big(\Pi_i | I_{h_i} = 1, I_{V_i} = 0\big) \cdot I_{h_i} \cdot (1 - I_{V_i}) + \\
& \big(\Pi_i | I_{h_i} = 0, I_{V_i} = 0\big) \cdot (1 - I_{h_i}) \cdot (1 - I_{V_i}) = \\
& I_{V_i} \cdot \Bigg(\big(\Pi_i | I_{h_i} = 0, I_{V_i} = 1\big) \cdot (1 - I_{h_i}) + \big(\Pi_i | I_{h_i} = 1, I_{V_i} = 1\big) \cdot I_{h_i} \Bigg) + \\
& (1 - I_{V_i}) \cdot \Bigg(\big(\Pi_i | I_{h_i} = 0, I_{V_i} = 0\big) \cdot (1 - I_{h_i}) + \big(\Pi_i | I_{h_i} = 1, I_{V_i} = 0\big) \cdot I_{h_i} \Bigg) \xrightarrow{P} \\
& 1 \cdot \Bigg(v_{1,i} \cdot (1 - I_{h_i}) + v_{1,i} \cdot I_{h_i} \Bigg) + 0 \cdot \Bigg(\not{v}_{1,i} \cdot (1 - I_{h_i}) + \not{v}_{1,i} \cdot I_{h_i} \Bigg) = v_{1,i}
\end{aligned}
$$

where we are assuming $\Pi_i | I_{h_i} = 1, I_{V_i} = 0$ and $\Pi_i | I_{h_i} = 0, I_{V_i} = 0$ will converge at least weakly to something, which we indicate with $\not{v}_{1,i}$. Notice that if $\not{v}_{1,i}$ exists, it will be tight because it is bounded between 0 and 1.

Therefore, for every occupation i, Π_i will converge in probability to the respective $v_{1,i}$. Notice that:

$$\tau^*_{1,i} := T_{\Sigma,i,n} \cdot \Pi_i \xrightarrow{P} \tau_{\Sigma,i} \cdot v_{1,i} = \tau_{1,i}$$

and consequently

$$\tau^*_{2,i} := T_{\Sigma,i,n} \cdot (1 - \Pi_i) \xrightarrow{P} \tau_{\Sigma,i} \cdot (1 - v_{1,i}) = \tau_{2,i}$$

As discussed in point III of Appendix *B*, we assemble (that is, stack) all of the and $\tau^*_{1,i}$ and $\tau^*_{2,i}$ in one table, τ^*. Once we do, we arrive at the main conclusion of this section:

(A24)
$$\tau^* \xrightarrow{P} \tau$$

By the continous mapping theorem:

(A25)
$$S(\tau^*) \xrightarrow{P} S(\tau)$$

E.2. Consistency of C-DP(C)

The C-DP(C) constructs a new simulated table T^* starting from τ^*, according to equations (A9) and (A10). Consider each column of T^* separately. We assume that Π_i converges in probability to some constant $s_{1,i}$. Following equation (A23), we have $s_{1,i} = v_{1,i}$, but this is not necessary for the consistency of C-DP(C). Under this relaxed assumption, we will have $\tau \xrightarrow{P} \tau^s$, where τ^s is some constant environment matrix not necessarily equal to τ. Under this new hypothesis, $S(\tau^*) \xrightarrow{P} S(\tau^s)$ by the continuous mapping theorem.

From equations (A9) and (A23) we have:

$$\Pi_i \xrightarrow{P} s_{1,i}$$

$$T^*_{1,i,n}|\Pi_i \sim Bin(\Pi_i, T_{\sum,i,n}) \Rightarrow$$

$$\frac{T^*_{1,i,n}}{T_{\sum,i,n}}|\Pi_i \sim \frac{Bin(\Pi_i, T_{\sum,i,n})}{T_{\sum,i,n}} = \frac{\sum_{j=1}^{T_{\sum,i,n}} Ber(\Pi_i)}{T_{\sum,i,n}} \xrightarrow{P} \Pi_i$$

where *Ber* is a Bernoulli random variable and the last passage is an application of the weak law of large number.

From here, consider an arbitrarily small $\varepsilon > 0$. For notational convenience, let us use the indicator $I_{s_{1,i}}$ to indicate the event $|\Pi_i - s_{1,i}| > \varepsilon$. By the definition of convergence in probability, we have that $lim_{n\to\infty}P(I_{s_{1,i}}) = 0$ and $lim_{n\to\infty}P(|\frac{\sum_{j=1}^{T_{\sum,i,n}} Ber(\Pi_i)}{T_{\sum,i,n}} - \Pi_i| > \varepsilon) = 0$ Then

$$
\begin{aligned}
&\lim_{n\to\infty} P(|\frac{T^*_{1,j,n}}{T_{\sum,i,n}} - s_{1,i}| > \varepsilon) = \\
&\lim_{n\to\infty} \bigg[P(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 1\big) - s_{1,i}| > \varepsilon)P(I_{s_{1,i}} = 1) + \\
&P\bigg(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 0\big) - s_{1,i}| > \varepsilon\bigg)P(I_{s_{1,i}} = 0)\bigg] = \\
&\lim_{n\to\infty} \bigg[P(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 1\big) - s_{1,i}| > \varepsilon)P(I_{s_{1,i}} = 1) + \\
&P\bigg(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 0\big) - s_{1,i}| > \varepsilon\bigg)\big(1 - P(I_{s_{1,i}} = 1)\big)\bigg] = \\
&\lim_{n\to\infty} \bigg[P(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 1\big) - s_{1,i}| > \varepsilon)P(I_{s_{1,i}} = 1) - \\
&P\bigg(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 0\big) - s_{1,i}| > \varepsilon\bigg)P(I_{s_{1,i}} = 1) + \\
&+P\bigg(|\big(\frac{Ber(\Pi_i)}{T_{\sum,i}}|I_{s_{1,i}} = 0\big) - s_{1,i}| > \varepsilon\bigg)\bigg] = \\
&1 \cdot 0 - 0 \cdot 0 + 0 = \\
&0
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\text{(A26)} \qquad &\frac{T^*_{1,i,n}}{T_{\sum,i,n}} \xrightarrow{P} s_{1,i} \Rightarrow \\
&T^*_{1,i,n} = T_{\sum,i,n} \cdot \left(\frac{T^*_{1,i,n}}{T_{\sum,i,n}}\right) \xrightarrow{P} n \cdot \big(\tau^s_{\sum,i} \cdot s_{1,i}\big) = n \cdot \tau^s_{1,i}
\end{aligned}
$$

Similarly, using equation (A10)

$$T^*_{2,i,n} = T_{\sum,i,n} - T^*_{1,i,n} = T_{\sum,i,n} \cdot \left(1 - \frac{T^*_{1,i,n}}{T_{\sum,i,n}}\right) \xrightarrow{P} n \cdot \tau^s_{\sum,i} \cdot (1 - s_{1,i}) = n \cdot \tau^s_{2,i} \tag{A27}$$

Once we stack all the variables $T^*_{1,i,n}$ and $T^*_{2,i,n}$ variables into a table T^* we get:

$$T^*_n \xrightarrow{P} n \cdot \tau^s \tag{A28}$$

As a reminder, from the definition of $S(\tau)$ in Appendix D, we have that $S(a \cdot \tau) = S(\tau)$ where $a \in \mathbb{R}_+$. Thus, by the continuous mapping theorem, equation (A25) and consistency of the plug-in estimator:

$$\begin{aligned} & S(T^*) \xrightarrow{P} S(n \cdot \tau^s) = S(\tau^s) \Rightarrow \\ & \hat{S}_{PI}(\tau) - \big(S(T^*) - S(\tau^*)\big) \xrightarrow{P} S(\tau) - \big(S(\tau^s) - S(\tau^s)\big) = S(\tau) \end{aligned} \tag{A29}$$

where we applied linearity of the convergence in probability.

Appendix to Chapter 2

F. A Characterization of Margin-Free Indices

As discussed in the main text, we define an environment as a joint distribution of the "group" (g) and unit (u) variables, $p(g,u)$. The joint distribution represents the likelihood of observing an individual with a given group-unit couple.

Segregation Index

A segregation indices are usually defined as functions $S : \mathbb{R}_{+}^{n\times m} \to \mathbb{R}$. That is, from a table to a real number (see for example Hutchens, 2004; Frankel and Volij, 2011). In this framework, three properties that are generally shared by all popular segregation indices:

Non-triviality: Two tables T and T' exist such that $S\big(T\big) \neq S\big(T'\big)$

Continuity: $S\big(T\big)$ is continuous (Frankel and Volij, 2011)

Homogeneity: $S\big(k \cdot T\big) = S\big(T\big)$ for $k > 0$ (Hutchens, 1991)

Segregation Index from a Bivariate perspective

In a more general way, we will define a segregation index as a function $S : p(g,u) \to \mathbb{R}$. We will still require continuity (defined using a statistical distance) and non-triviality:

Non-triviality: Two distributions $p(g,u)$ and $p'(g,u)$ exist such that $S\big(p(g,u)\big) \neq S\big(p'(g,u)\big)$

Continuity: $S\big(p(g,u)\big)$ is continuous (Frankel and Volij, 2011)

However, homogeneity is enforced by the fact that we normalize the data in order to interpret it as a distribution: if T represents the data, we may write that $p(g,u) := \frac{T}{\sum T}$.

Theorem F.1 (Osius Theorem)**.** *Let $p(x,y)$ be a joint density. Let $p(x) = \int p(x,y)dy$, $p(y) = \int p(x,y)dy$, $\phi(x,y) = \log \frac{p(x,y)p(x^*,y^*)}{p(x^*,y)p(x,y^*)}$ with x^*, y^* arbitrary reference points in the support of $p(x)$ and $p(y)$:*

1: *If $p(x,y)$ is finite and strictly positive (> 0) everywhere in its support, then $p(x,y)$ is uniquely determined by the triplet $< p(x), p(y), \phi(x,y) >$*

2: *For $p(x)$, $p(y)$, $\phi(x,y)$ finite everywhere, a joint density $p(x,y)$ exists having the triplet $< p(x), p(y), \phi(x,y) >$ as its component.*

This is an abridged and less general version of the Uniqueness Theorem and Existence Theorem by (Osius, 2009, p.462-463). It is sufficient for our purposes. The interested reader may find the proof in the original paper.

The theorem substantially means that a joint distribution may be decomposed into its marginal distribution ($p(x)$, $p(y)$) and a function describing the association between the two variables, $\phi(x,y)$. We refer to $\phi(x,y)$ as "log odds ratios with respect to a reference" or, simply, "log odds ratios". Notice, we de-emphasize the dependence of $\phi(x,y)$ on the chosen reference point since this is irrelevant for Osius theorem and, ultimately, for measuring segregation with the Q index. Therefore, we may represent a joint distribution with the notation $< p(x), p(y), \phi(x,y) >$, which we will refer to as "a triplet".

For our purposes, Osius theorem means that we can decompose any strictly positive joint distribution of groups and units into three elements: (i) group proportion, (ii) unit marginal distribution, and (iii) a third object regulating the association of unit and groups. As we shall see momentarily, this provides clear guidance about creating margin-free segregation indices.

However, the theorem does not always apply: we may be interested in measuring segregation of joint distributions that are not strictly positive. Indeed, it is far from uncommon that one group is not present within a unit. In those cases, the joint density $p(g,u)$ will be 0 for that group-unit combination. For non-strictly-positive density, the triplet $< p(g), p(u), \phi(u,g) >$ does not uniquely determine $p(g,u)$ because there are different $p(g,u)$ corresponding to a triplet. We define a boundary joint distribution as a distribution such that for some non-zero-probability region h of the support of $p(u)$ we have $P(u \in h | g = i) = 0$ and $P(u \in h | g = j) > 0$ for any two different groups i and j. In the h region, $\phi(u,g)$ is not finite for the i^{th} group and the distribution cannot be represented solely as the triplet $< p(g), p(u), \phi(u,g) >$. However, continuity of segregation indices still enforces that margin-free indices must solely depends on $\phi(g,u)$, even in the boundary distributions.

Now, we need to re-formulate the 'margin-free' property in a way that is more amenable to technical treatment.

F.1. Margin-free Indices on Finite Support

The property of being margin-free is defined by Watts (1998) as follows. Consider an environment T represented as a table in $\mathbb{R}_{+}^{n\times m}$, with n,m finite and greater or equal to 2. A segregation index is a function $S : \mathbb{R}_{+}^{n\times m} \to \mathbb{R}$.

Now, consider the two diagonal matrices with strictly positive diagonal entries $\Gamma \in \mathbb{R}_{+}^{n\times n}$ and $\Lambda \in \mathbb{R}_{+}^{m\times m}$.

Definition F.1 (Definition 1 of Margin-free Index)**.** *Given a positive table T, a segregation index is margin free if and only if*

$$S(T) = S(\Gamma \cdot T \cdot \Lambda)$$

for any valid T, Γ and Λ.

Starting from this definition, we can arrive at a second equivalent definition. For the moment, we will focus on a subset of matrices where all elements are strictly positive – that is $T_{i,j} > 0$. For this sub-set of environments, we have the following equivalent definition:

Theorem F.2 (Triplet Definition of Margin-free index)**.** *Let a distribution T' be represented as the triplet $< p(g), p(u), \phi(g,u) >$, where $p(g)$ and $p(u)$ are the group marginal distribution* *and unit marginal distribution, respectively, while $\phi(g,u)$ is a valid log odds ratio function for the distribution. A segregation index on a finite support is margin-free if and only if:*

$$S(< p(g), p(u), \phi(g,u) >) = S(< p'(g), p'(u), \phi(g,u) >)$$

for any valid marginal distribution $p(g), p'(g), p(u)$ and $p'(u)$.

PROOF. We need to prove that Definition 1 and the triplet definition are equivalent in the relevant sub-set. As a first step we build a probability distribution from a matrix T representing the data. We can do this by simply defining:

$$T' = \frac{T}{\sum T} :=< p(g), p(u), \phi(u,g) >$$

Notice that T' may be interpreted as a probability distribution since it is positive and sums to one. From here, we may as well represent T' as a the triplet $< p(g), p(u), \phi(g,u) >$ following Osius theorem – which applies given that we are considering strictly positive matrices.

Now, consider the iterative proportional fitting process described by Deming and Stephan (1940) . This is a well-known iterative algorithm to transform any finite-support distribution $< p(g), p(u), \phi(u,g) >$ into another distribution $< p'(g), p'(u), \phi(u,g) >$ with arbitrary marginal distributions $p'(g), p'(u)$ and the same log odds ratios. We will use $IPF(A \rightarrow B)$ to indicate that iterative proportional fitting is used to transform distribution A into B, with arbitrary marginal distributions.

Now, the t^{th} passage of the algorithm can be represented as a matrix multiplication:

$$A(t) = \big(\Gamma(t) \cdot A(t-1)\big) \cdot \Lambda(t) = \Gamma(t) \cdot \big(A(t-1) \cdot \Lambda(t)\big) \tag{.15}$$

where $A(t-1)$ is the result of the algorithm after the t^{th} passage and $\Gamma(t)$, $\Lambda(t)$ are the diagonal matrices. The way these matrices are constructed at each iteration is described, for example, in Deming and Stephan (1940) at pages 439-442, but it is irrelevant in this context. It follows that the iterative proportional fitting $IPF(A \to B)$ after t iterations may be represented as

$$B_t = \Gamma(t) \cdot \big(\ldots \big(\Gamma(2)\big(\Gamma(1)\big(A(0) \cdot \Lambda(1)\big)\big) \cdot \Lambda(2)\big)\big) \ldots \cdot \Lambda(t)\big) = \Gamma_t \cdot A \cdot \Lambda_t$$

where

$$\Gamma_t = \Pi_{i=1}^{t} \Gamma(i)$$

$$\Lambda_t = \Pi_{i=1}^{t} \Lambda(i)$$

Moreover, notice that Γ_t and Λ_t will still be properly-dimensioned diagonal matrices with strictly positive entries on the diagonal because strictly positive diagonal matrices are closed under multiplication. Fienberg (1970) shows that $IPF(A \to B)$ will provide asymptotically correct results – that is a table with the wanted marginal distributions. Thus, the form $\Gamma_t \cdot T' \cdot \Lambda_t$ may be used to change the marginal distributions arbitrarily as t approaches infinity: $lim_{t\to\infty} B_t = B$

Following, definition 1, an index is margin-free iff $S(T') = S(\Gamma_t \cdot T' \cdot \Lambda_t) := S(T'')$. Now, $T'' :=< p'(g), p'(u), \phi(g,u) >$, with arbitrary marginal distributions $p'(g), p'(u)$. Using homogeneity of segregation indices, we have:

$$S(T) = S(\Gamma_t \cdot T \cdot \Lambda_t) \Leftrightarrow S\left(\frac{T}{\sum_T}\right) = S(T') = S\left(\Gamma_t \cdot \frac{T}{\sum_T} \cdot \Lambda_t\right) = S(T'') \Leftrightarrow$$

$$S\big(< p(g), p(u), \phi(g,u) > \big) = S\big(< p'(g), p'(u), \phi(g,u) > \big)$$

$\square$

From here, we can generalize the definition of a margin-free segregation index.

Definition F.2 (Generalized Definition of a Margin-free Index)**.** *A segregation index is margin-free if and only if:*

$$S\big(< p(u), p(g), \phi(g,u) > \big) = S\big(< p'(u), p'(g), \phi(g,u) > \big)$$

for any strictly positive marginal distribution $p(u), p(g), p'(u), p'(g)$.

This definition is useful since it defines the margin-free property unequivocally on a wider array of cases than the original definition.

Theorem F.3 (Main Theorem)**.** *A segregation index is margin free if and only if it depends exclusively on the log odds ratios,* $\phi(x,y)$

Proof. Osius theorem implies that any joint distribution outside the boundaries can be written as a triplet $p(g,u) =< p(g), p(u), \phi(g,u) >$. First, consider the region where this theorem applies. If an index only depends on $\phi(g,u)$, then $S(< p'(g), p'(u), \phi(g,u) >) =$ $S(< p''(g), p''(u), \phi(u,g) >)$ for any $p'(g), p'(u), p''(g), p''(u)$. Thus the index is margin-free by the general definition of margin-free index.

Conversely, assume the S index is margin-free in the region where Osius theorem applies. By hypothesis, S is non-trivial. Therefore, consider any two joint distributions, $p'(g,u)$ and $p''(g,u)$, such that $S\big(p'(g,u)\big) \neq S\big(p''(g,u)\big)$. By Osius theorem we may write these joint distribution as $p'(g,u) =< p'(g), p'(u), \phi'(g,u) >$ and $p''(g,u) =< p''(g), p''(u), \phi''(g,u) >$. Further, consider a new joint distributions, $p''_a(g,u) =< p'(g), p'(u), \phi''(g,u) >$ – notice that $p''_a(g,u)$ may be obtained from $p''(g,u)$ by changing only the margins, using for example the IPF procedure described by Deming and Stephan (1940). Since S is margin free, $S(p''_a(g,u)) = S(p''(g,u)) \neq S(p'(g,u))$. By

Osius theorem, we must have $\vec{\phi}' \neq \vec{\phi}''$, otherwise $p'(g,u)$ would be identified by the same triples as $p''_a(g,u)$ and the two distribution would be identical, implying $S(p''_a(g,u)) = S(p'(g,u))$. Therefore, $S(p'(g,u)) \neq S(p''(g,u))$ implies $\phi'(g,u) \neq \phi''(g,u)$ for margin-free indices.

Now, consider the boundaries, where Osius theorem does not apply. We have discussed in Section 2.5.5 above that a sigmoid function of $\vec{\phi}$ is sufficient to define a margin-free segregation index on the boundaries. However, the inverse must be treated with more caution.

As discussed above, a joint distribution is not uniquely determined by any triplet $< p(g), p(u), \phi(g,u) >$, therefore, we must have some extra elements determining the joint distribution in addition to the usual triplet. Assume that a distribution on the boundary is uniquely determined by the triplet $< p(g), p(u), \phi(g,u) >$ alongside some other information $v(g,u)$. For example, in the case of a 2 by 2 table, $v(g,u)$ could be the value of the non-zero cells of the table. We will show that a margin-free index cannot depend on v. For any $v'(g,u) \neq v''(g,u)$, let $p'_B =< p'(g), p'(u), \sigma(\phi_B(g,u)), v'(g,u) >$ and $p''_B =< p''(g), p''(u), \sigma(\phi_B(g,u)), v''(g,u) >$, where $\sigma(x)$ is a sigmoid function. That is, p'_B and p''_B are two different boundary distributions with identical sigmoid log odds ratios. We want to show that for any margin-free segregation index $S(p'_B) = S(p''_B)$. Therefore, S cannot depend on v even at the boundary.

Assume S is margin-free. We will assume we can build the sequence of joint distributions $p'_n =< p'_n(g), p'_n(u), \sigma(\phi'_n(g,u)), v'_n(g,u) >$ where $\lim_{n\to\infty} p'_n = p'_B$ and p'_n is not a boundary joint distributions for any n. Similarly, we will assume we can build the alternative sequence $p''_n =< p''_n(g), p''_n(u), \sigma(\phi''_n(g,u)), v''_n(g,u) >$ having the similar properties that $\lim_{n\to\infty} p''_n = p''_B$ and p''_n is never a boundary joint distribution. Notice that $\lim_{n\to\infty} \sigma(\phi'_n(g,u)) = \lim_{n\to\infty} \sigma(\phi''_n(g,u)) = \sigma(\phi_B(g,u))$. Therefore, for any properly defined metric $d()$ on the space of $\sigma(\phi)$ functions, $\lim_{n\to\infty} d\big(\sigma(\phi'_n(g,u)), \sigma(\phi''(g,u))\big) = 0$. Since the elements in the sequences p'_n and p''_n are not boundary distributions, we have that $S(p'_n)$ and $S(p''_n)$ will only depend on $\phi'_n(g,u)$ and $\phi''_n(g,u)$, respectively. By continuity of S and σ, $\lim_{n\to\infty} d\big(\sigma(\phi_n'(g,u)),\sigma(\phi''(g,u))\big)=0$ implies $\lim_{n\to\infty} d\big(\phi_n'(g,u),\phi''(g,u)\big)=0$, which implies $\lim_{n\to\infty}|S(p_n')-S(p_n'')|=0$. Thus, $\lim_{n\to\infty} S(p_n')=\lim_{n\to\infty} S(p_n'')$. Applying continuity of S once again:

$$S(p_B')=S\big(\lim_{n\to\infty} p_n'\big)=\lim_{n\to\infty} S\big(p_n'\big)=\lim_{n\to\infty} S\big(p_n''\big)=S\big(\lim_{n\to\infty} p_n''\big)=S(p_B'')$$

Therefore, S will not vary with $v(g,u)$ even for boundary joint distributions. With an identical argument, one can show that S will vary with $\sigma(\phi(g,u))$, even at the boundaries. Thus, S only depends on $\sigma(\phi(g,u))$ and ultimately on $\phi(g,u)$. □

Appendix to Chapter 3

G. Seed Words for the Lovecraft Example

Seed words for the $\overrightarrow{Lovecraftian}$ **semantic dimension:** *Cthulhu*, *Dunwich*, *Innsmouth*, *Lovecraftian*, *Miskatonic*, *Mythos*, *Necronomicon*

Lovecraftian Words: *abnormal*, *accursed*, *amorphous*, *antediluvian*, *antique*, *Arkham*, *blasphemous*, *charnel*, *cyclopean*, *decadent*, *demoniac*, *eldritch*, *fainted*, *foetid*, *fungus*, *furtive*, *hideous*, *immemorial*, *indescribable*, *Kingsport*, *Leng*, *loathsome*, *madness*, *manuscript*, *nameless*, *noisome*, *shunned*, *singularly*, *spectral*, *stench*, *tentacles*

Anti-Lovecraftian words: *adorable*, *appealing*, *beautiful*, *bright*, *charming*, *cheerful*, *cheery*, *cute*, *dainty*, *delightful*, *fluffy*, *foxy*, *glitter*, *handsome*, *kitten*, *lovely*, *pleasant*, *pooch*, *pretty*, *puppy*, *soft*, *sunny*, *whep*

H. Concept Mover Distance

In the present framework, The concept mover distance (CMD) introduced by Stoltz and Taylor (2019) and later refined in Taylor and Stoltz (2021) can be interpreted as the representation of a document through a weighted sum of its word-vectors and subsequent projection of the document on a dimension of interest. In this section, we will show we this interpretation holds.

Let s be the sequence of words forming a document and let n be the length of s. We indicate with $\mathscr{D}$ the set of word-vectors in the document. We associated each word-vectors $\vec{w}_i$ in $\mathscr{D}$ with a weight c_i. If the word w_i appears m_i times in s, its weight will simply be $c_i = \frac{m_i}{n}$ – that is, c_i is the relative frequency of the word w in s. We collect all such relative frequencies in the vector $\vec{c}$. Finally, let $\vec{D}$ be a semantic dimension. We indicate with $\vec{D}$ the semantic dimension of interest – as usual, $||\vec{D}|| = 1$.

The CMD is introduced as the sum of the elements in a matrix measuring the distance of document s from document s', where the document s' is simply $\vec{D}$. In this case, the to-be-summed matrix results from an element-wise multiplication (or Hadamard product) of a $n \times 1$ transition matrix T and an equally-dimensioned similarity matrix S. The element $S_{i,j}$ of S contains the cosine between $\vec{w}_i$ and $\vec{D}$. As mentioned in the main text, we respect the original vocabulary in the paper, but it is important to realize that the CMD is a similarity measure (as opposed to a distance measure) because S contains cosines, which are a measure of similarity (not distance). In fact, a higher value of $S_{i,j}$ indicates that $\vec{w}_i$ and $\vec{D}$ are closer in the embedding space. As for T, the authors consider two options (denoted with T^* and T') in order to pick the one giving the minimum sum, which, in turn, is an upper-bound approximation of the harder-to-find earth moving proximity (Kusner et al., 2015).[15] Leaving aside the theoretical reasoning, it is sufficient to notice

[15] There is some confusion in the presentation of CMD in the original paper (Stoltz and Taylor, 2019). The authors declare in their equation (5) to look for a maximum (not a minimum) between T^* and T', but in their code they look for a minimum. Likely, the confusion is due to change from Euclidean distance to cosine similarity for matrix S with

that “the $T^*_{i,j}$ [...] will always be the relative frequency of $\vec{w}_i$ [that is, c_i] in s and the $T^{'}_{ij}$ weighting will always be 1.” (Stoltz and Taylor, 2019) (p. 11, our notation). Quite clearly, $c_i \leq 1$ for any i. Therefore, T^* will always give the minimum sum we are looking for and it coincides with the vector $\vec{c}$ introduced above. Overall, we can write the CMD as:

$$CMD(s,\vec{D}) = \sum \left(S \odot T^*\right)$$

where we indicate the Hadamard pruduct with $\odot$ and $\sum \left(M\right)$ represents the operation of summing the elements of a matrix M

For our purposes, we should notice that $S = W_n^T \cdot \vec{D}$, where W_n is the matrix having $\frac{\vec{w}_i}{||\vec{w}_i||}$ as its i^{th} column. Therefore, we can write the CMD as:

$$\begin{aligned}
CMD(s,\vec{D}) =& \sum \left(S \odot T^*\right) = \\
& S^T \cdot T^* = S^T \cdot \vec{c} = \\
& \vec{D}^T \cdot W_n \cdot \vec{c} = \\
& \sum_{w_i \in \mathscr{D}} \vec{D}^T \cdot \frac{\vec{w}_i}{||\vec{w}_i||} \cdot c_i = \\
& \vec{D}^T \cdot \sum_{w_i \in \mathscr{D}} \frac{\vec{w}_i}{||\vec{w}_i||} \cdot c_i = \vec{D}^T \cdot \sum_{w_i \in \mathscr{D}} \vec{w}_i \cdot \frac{m_i}{||\vec{w}_i|| n} = \\
& \vec{D}^T \cdot W \cdot \vec{c}_n
\end{aligned}$$

where W is the matrix having $\vec{w}_i$ as its i^{th} column, and $\vec{c}_n$ is a vector of weights having $\frac{c_i}{||\vec{w}_i||}$ as its i^{th} element.

respect to Kusner et al. (2015). This change transform CMD from a lower bound approximation of the earth moving *distance* to an upper bound approximation of a earth moving *similarity*. We stick to the code implementation of CMD.

Notice, $W_n \cdot \vec{c}$ is simply the mean of the normalized $\vec{w}_i \in D$ – which we can interpret as a representation of the document s in the embedding space. In turn, $W_n \cdot \vec{c}$ is equivalent to $W \cdot \vec{c}_n$, which is a weighted sum of the word-vectors in $\mathscr{D}$. Overall, the CMD is the projection of the document s onto $\vec{D}$, where s is represented as a document-vector in the embedding space. Therefore, the techniques discussed in the main text about weighted sums applies to CMD as well.

It is interesting to notice that the CMD weighting scheme ($\vec{c}_n$) implicitly discounts words that are more frequent. In fact, the weight associated to a word-vector will be inversely proportion to its norm. In turn, larger norms are associated with more frequent words (Ethayarajh et al., 2019). Compare the CMD with a representation of a document as a simple mean of its word-vectors. To obtain the projection of this representation, we can slightly change the CMD from $\vec{D} \cdot W_n \cdot \vec{c}$ to $\vec{D} \cdot W \cdot \vec{c}$. Clearly, this representation is closely related to the CMD. However, unlike the CMD, it does not change the weight of word-vectors based on their norms.

I. Words Mean Representation of Documents

The main text mentions the geometric interpretation by Arora et al. (2017) of document composition. The main idea is that a document can be represented as a central focus in the embedding space selecting words based on the proximity of their word-vector in the embedding space. Following this main intuition, we show that, under assumptions similar to those made in Arora et al. (2017), this geometrical representation can justify document representation as in the CMD (see H) or as the mean of the word-vectors forming the document.

Let s be the sequence of words forming a document, we indicate with $\mathscr{D}$ the set of word-vectors in the document. We indicate with n the number of words in the document, that is the number of words in s. As in section H, we associate each word-vectors $\vec{w}_i$ in $\mathscr{D}$ with a weight $c_i = \frac{m_i}{n}$, where m_i is the number of times word w appears in s. Further, let $\mathscr{W}$ be the set of all word-vectors in the embedding space.

As mentioned in the main text, we assume that there is a central focus of the document, which can be represented by the document-vector $\vec{d}$ in the same embedding space as the word-vectors $\vec{w} \in \mathscr{W}$. Further, we use the following model:

Word selection: Given $\vec{d}$, the words of a document are selected sequentially and independently. Specifically, the probability that the next word selected will be w_j is $P(\vec{w}_j|\vec{d}) = \frac{\exp(\vec{d} \cdot \vec{w}_j)}{Z_d}$, where $Z_d = \sum_{\vec{w}_j \in \mathscr{W}} \exp(\vec{d}_n \cdot \vec{w}_j)$ is a normalizing constant. Given this assumption, the overall likelihood of a document becomes $P(\vec{s}|\vec{d}_n) = \prod_{\vec{w}_i \in \mathscr{D}} \frac{\exp(\vec{d}_n \cdot m_i \cdot \vec{w}_i)}{Z_d}$.

Notice that the model is unidentified because the probability of selecting any word does not change as the norm of $\vec{d}$ changes. In other words, we must input some information about $||\vec{d}||$ to be able to estimate $\vec{d}$. Under the same assumptions as (Arora et al., 2017), we can approximate the normalizing factor $Z(\vec{d}) \approx Z \exp(||\vec{d}||^2)$, where Z is a scalar constant not changing with $\vec{d}$ (Arora et al., 2018). This approximation is doubly useful. It identifies the model by posing a constraint on $||\vec{d}||^2$ and makes model estimation much easier (see below). In practice, we can reframe $\exp(||\vec{d}||^2)$ as a regularization for the likelihood. Following usual regularization procedures, we can add a parameter λ to the regularization, so that we rewrite $Z(\vec{d}) \approx Z\lambda \exp(||\vec{d}||^2)$. In what follows, we fix $\lambda = \frac{1}{2}$. Then, the MAP estimator for $\vec{d}$:

$$\begin{aligned}
P(\vec{s}|\vec{d},\lambda = \frac{1}{2}) = \prod_{\vec{w}_i \in \mathscr{D}} \frac{\exp(\vec{d}\cdot m_i \cdot \vec{w}_i)}{Z_d} \qquad & \approx \prod_{\vec{w}_i \in \mathscr{D}} \frac{\exp(\vec{d}\cdot m_i \cdot \vec{w}_i - \frac{1}{2}||\vec{d}||^2)}{Z} \\
\log P(\vec{s}|\vec{d},\lambda = \frac{1}{2}) = \mathscr{L}(\vec{s}|\vec{d},\lambda = \frac{1}{2}) \qquad & \approx \sum_{\vec{w}_i \in \mathscr{D}} \vec{d}\cdot m_i \cdot \vec{w}_i - \frac{n}{2}||\vec{d}||^2 - n\log Z \\
& = \vec{d}\cdot \sum_{\vec{w}_i \in \mathscr{D}}^{n} m_i \cdot \vec{w}_i - \frac{n}{2}||\vec{d}||^2 - n\log Z
\end{aligned}$$

$$\hat{\vec{d}} = \arg\max_{\vec{d}} \mathscr{L}(\vec{s}|\vec{d},\lambda = \frac{1}{2}) \approx \sum_{\vec{w}_i \in \mathscr{D}} \frac{m_i}{n} \cdot \vec{w}_i$$

Altogether, these assumptions imply that the approximate MAP estimator of $\vec{d}$ is

$$\hat{\vec{d}} = ||\hat{\vec{d}}|| \cdot \hat{\vec{d}}_n = \sum_{\vec{w} \in \mathscr{D}} \frac{m_i}{n} \vec{w}_i = \sum_{\vec{w} \in \mathscr{D}} c_i \vec{w}_i$$

This is simply the mean of the word-vectors forming the document s.

We can change the modeling slightly to obtain the CMD representation of a document. In particular, let $w_{j,n} = \frac{\vec{w}_j}{||\vec{w}_j||}$. Then, we can reframe the previous model as:

Word selection: Given $\vec{d}$, the words of a document are selected sequentially and independently. The probability that the next word selected will be $w_j \in \mathscr{W}$ is $P(\vec{w}_j|\vec{d}) = \frac{\exp(\vec{d}\cdot \vec{w}_{j,n})}{Z_d}$, where $Z_d = \sum_{\vec{w}_j \in \mathscr{W}} \exp(\vec{d} \cdot \vec{w}_{j,n})$ is a normalizing constant. Thus, the overall likelihood of a document is $P(\vec{s}|\vec{d}) = \prod_{\vec{w}_i \in \mathscr{D}} \frac{\exp(\vec{d}\cdot m_i \cdot \vec{w}_{i,n})}{Z_d}$.

Once again, the model is unidentified because word selection does not depend on the the norm of the document-vector, $||\vec{d}||$. As before, we approximate Z to identify the model: $Z_d \approx Z\exp(\lambda||\vec{d}||^2)$ (Arora et al., 2017, 2018)). We fix $\lambda = \frac{1}{2}$ and follow the same exact steps as above to obtain the new MAP estimate for $\vec{d}$:

$$\hat{\vec{d}} = \sum_{\vec{w} \in \mathscr{D}} \frac{m_i}{n} \vec{w}_{i,n} = \sum_{\vec{w} \in \mathscr{D}} \frac{m_i}{n||\vec{w}_i||} \vec{w}_i$$

As shown in section H, this is the document-vector representation used in the CMD.

J. Seed Words for Climate Change Methodologies

Seed words for $\vec{Field\ Methods}$**:** *interview*, *ethnographic*, *questionnaire*, *elicitation*, *workshop*, *discussion*, *archival*, *observational*, *survey*, *fieldwork*, *experiment*

Seed words for $\vec{Simulation\ Methods}$**:** *racmo*, *om1*, *cgcm*, *gcms*, *simulation*, *rca3*, *hirham5*, *hirham*, *echam5*, *echam4*, *echam3*, *echam*, *tbm*, *hadam3h*, *aogcm*, *opyc3*, *simulation*, *regression*, *aquaplanet*, *pca*, *som*, *multimodel*, *vsvd*, *downscaling*, *eof*, *ccsm4*, *ccsm3*, *ccsm2*, *ccsm*, *csm1*, *ccm2*, *ccm3*, *cam2*, *cam3*, *cam4*, *cam5*, *lm2*, *agcm4*, *agcm3*

Seed words for $\vec{Document\ Methods}$**:** *discourse*, *review*, *sna*, *assessment*, *evaluation*, *report*, *qca*, *qualitative*, *text*

Seed words for $\vec{Historical\ Methods}$**:** *reconstruction*, *pbm dendrochronology*, *paico*, *dendrochronological*, *dendroclimatological*, *palynological*, *paleoecological*, *stratigraphy*, *multiproxy*, *paleoceanographic*, *paleoclimate*

K. Seed Words for the Burning Glass Data Application

Seed words for $\vec{Upper-Body\ Strength}$**:** *Natural Gas Extraction*, *Road Construction*, *Carpentry*, *Furniture Moving*

Seed words for $\vec{Fine\ Motor\ Skill}$**:** *Manual Dexterity*

Seed words for $\vec{Problem\ Solving}$**:** *Problem Solving*, *Detail-Oriented*, *Troubleshooting*, *Analytical Skills*

Seed words for $\vec{Mathematical\ Skills}$**:** *Financial Modeling*, *Simulation*, *Clustering*, *Physics*, *Statistics*, *Statistical Analysis*, *Geometry*

Seed words for $\vec{Authority}$**:** *Staff Management*, *Supervisory Skills*, *People Management*, *Business Administration*, *Strategic Planning*, *Business Planning*, *Business Management*

Seed words for $\vec{Sociability}$**:** *Communication Skills*, *Teamwork / Collaboration*, *Verbal / Oral Communication*, *Teaching*, *Listening*, *Persuasion*

Seed words for $\vec{Technical\ Coding\ Skills}$**:** *Software Development*, *Software Engineering*, *Relational Databases*, *Oracle*

Seed words for $\vec{Housekeeping}$**:** *Food Preparation*, *Housekeeping*, *Senior Care*, *Laundry, Sorting*, *Toileting*, *Bathing*

L. Magnitude of Coefficients in Multidimensional Projections

In this section we show that the coefficients in a multidimensional projection cannot be normalized between -1 and 1 if all semantic dimensions have the same norm, as recommended in the main text.

Consider a word $\vec{w}$ and a set of semantic dimensions $\vec{D}_1, \vec{D}_2 \dots \vec{D}_n$ such that $||\vec{D}_i|| = r > 0$ for $i = 1 \dots n$. By equation (3.3):

$$\vec{w} = \sum_{i=1}^{n} \beta_i \vec{D}_i + \vec{\varepsilon}$$

We partition the semantic dimensions into two sets. The first only contains $\vec{D}_n$, the second contains all other dimension. We will show that β_n cannot be normalized between -1 and 1 if $||\vec{D}_i|| = ||\vec{D}_j||$ for $j, i = 1 \dots n$ and $i \neq j$.

Without loss of generality, we pick $r = 1$. We can rewrite previous equation as

$$\vec{w} = \sum_{i=1}^{n-1} \beta_i \vec{D}_i + \beta_n \vec{D}_n + \vec{\varepsilon} = \beta_k \vec{D}_k + \beta_n \vec{D}_n + \vec{\varepsilon} \tag{.16}$$

where $\beta_k = ||\sum_{i=1}^{n} \beta_i \vec{D}_i||$ and $\vec{D}_k = \frac{\sum_{i=1}^{n} \beta_i \vec{D}_i}{\beta_k}$. Therefore, $||\vec{D}_k|| = 1$ and $\beta_k \vec{D}_k = \sum_{i=1}^{n} \beta_i \vec{D}_i$.

Now, we regress $\vec{D}_n$ on $\vec{D}_k$:

$$\vec{D}_n = \delta_n \vec{D}_k + \vec{D}_{n,\perp}$$

where $\vec{D}_{n,\perp}$ is the vector of residuals from such regression and δ_n is equal to the cosine between $\vec{D}_n$ and $\vec{D}_k$, since $||\vec{D}_n|| = ||\vec{D}_k|| = 1$. Notice that $\vec{D}_k \cdot \vec{D}_{n,\perp} = 0$ since $\vec{D}_{n,\perp} \perp \vec{D}_k$ by construction. Therefore

$$||\vec{D}_{n,\perp}||^2 = ||\vec{D}_n||^2 - \delta_n^2 ||\vec{D}_k||^2 = 1 - \delta_n^2$$

By the Frisch-Waugh-Lovell theorem we can obtain an estimate for $\hat{\beta}_n$ from equation (.16) by regressing a partialled-out version of $\vec{w}$ on $\vec{D}_{n,\perp}$. However, the same coefficient can also be estimated by directly regressing $\vec{w}$ on $\vec{D}_{n,\perp}$ (Angrist and Pischke, 2009):

$$\vec{w} = \beta_n \vec{D}_{n,\perp} + \vec{\varepsilon}_n$$

Using OLS on this model, we obtain:

$$\hat{\beta}_n = \frac{||\vec{w}||}{\sqrt{1-\delta_n^2}} \cos(\vec{D}_{n,\perp}, \vec{w})$$

In general $\frac{||\vec{w}||}{\sqrt{1-\delta_k^2}} \neq 1$ and the magnitude of $\hat{\beta}_n$ will not be normalized between -1 and 1. In fact, given a set of basis to form $\vec{D}_k$, it is always possible to pick the next basis $\vec{D}_n$ such that the estimated coefficient $\hat{\beta}_n$ will be arbitrary big – it is sufficient to let $\delta_n \to 1$.

Notice that we could re-normalize $\vec{D}_n$ so that the previous equation reduces to a simple cosine. Indeed, for $\vec{D}'_n = \vec{D}_n \cdot \frac{||\vec{w}||}{\sqrt{1-\delta_k^2}}$, the previous equation reduces to $\hat{\beta}'_n = \cos(\vec{D}'_{n,\perp}, \vec{w})$, where $1 \leq \cos(\vec{D}'_{n,\perp}, \vec{w}) \leq 1$. However, if we proceeded with such re-normalization on all dimensions, we would have that $||\vec{D}'_i|| \neq ||\vec{D}'_j||$ for $i \neq j$ because $\delta_i \neq \delta_j$ in general. Therefore, the coefficients of the different dimensions would be on different scales and not directly comparable.