FACULTY OF
BIOSCIENCE ENGINEERING

A COMPARATIVE STUDY OF DIVERSITY INDICES FOR ASSESSING FOOD BIODIVERSITY AND ITS RELATION WITH MORTALITY IN EUROPE

Jill Deygers
Student ID: 01810042

Promotors: Dr. Aisling Daly, Prof. Dr. Carl Lachat
Tutor: Dr. Giles Hanley-Cook

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of master in The Master of Science in Bioscience Engineering: Food Science and Nutrition.
Academic year: 2022 - 2023

GHENT
UNIVERSITY

ACKNOWLEDGEMENTS

First of all, I would like to thank everyone who contributed to make this thesis a great success.

I am grateful to prof. Carl Lachat and Dr. Giles Hanley-Cook for introducing me to the fascinating concept of food biodiversity. Additionally, I would like to extend my thanks to Dr. Aisling Daly for providing a mathematical perspective, which further enhanced the depth of my research, and for helping me to polish the written content. I would also like to acknowledge the International Agency for Research on Cancer for providing the data that formed the basis of my investigation. The availability of such a large database allowed me to explore various indices.

I also want to express my appreciation to the developers of Chat GPT, whose tool gave me prompt feedback and made me think even more critically about all the things I wrote.

A special thanks goes out to my friends, whose company and support provided enjoyable breaks during this thesis year.

In this section, I would also like to thank my grandfather for his unwavering support throughout my studies:
Lieve opa, dank je voor alle tips, extra uitleg en mnemotechnische middeltjes die je mij aanleerde. Als ik ergens vastliep in een cursus, ging jij er desnoods dag en nacht op zoeken tot je het antwoord gevonden had. Deze thesis ging misschien wel je petje te boven, maar zonder al jouw hulp in wetenschappelijke vakken, ging ik de leerstof nooit zo goed beheersen.

Last but certainly not least, I want to express my deepest gratitude to my family for all their support, not only throughout this year but throughout my entire life. Your constant encouragement and belief in me have been the driving force behind my accomplishments.

Thank you all for being an integral part of this incredible journey.

Jill

CONTENTS

LIST OF ABBREVIATIONS

BMI	Body mass index
Corr	Correlation
CS	Common share
De	Denmark
DQ	Dietary questionnaire
EPIC	The European Prospective Investigation into Cancer and Nutrition study
FAO	Food and Agriculture Organization of the United Nations
FFQ	Food frequency questionnaire
Fr	France
FTF	Face-to-face interviews
Ge	Germany
IARC	International Agency for Research on Cancer
IPBES	Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services
It	Italy
NCLASS	Detailed EPIC food classification system
Ne	The Netherlands
No	Norway
SA	Self-administered
SDG	Sustainable Development Goals
Sp	Spain
Sw	Sweden
UK	United Kingdom
US	Unique share

SAMENVATTING

Hoewel biodiversiteit belangrijk is voor duurzame en gezonde voeding, is het nog onduidelijk welke biodiversiteitsindex het meest geschikt is voor voeding en hoe deze index samenhangt met gezondheid. De Hill-getallen zijn eenvoudige indices, die vergelijking tussen groepen mogelijk maken en nog niet zijn toegepast op voedselinnamegegevens. $Hill_0$, $Hill_1$, $Hill_2$ en $Hill_\infty$ werden berekend op basis van voedselinnamegegevens (ingedeeld op soortniveau) van de *European Prospective Investigation into Cancer and Nutrition* studie (n=476 768; 9 landen). De indices werden op zowel energiebasis als gewichtsbasis bepaald en correlaties werden onderzocht. De associaties van biodiversiteitsindices met totale sterfte werden beoordeeld met behulp van *Cox proportional hazards* modellen. Hoewel $Hill_\infty$ alleen gebaseerd is op het aandeel van de meest geconsumeerde soort, terwijl $Hill_2$ de proporties van alle soorten omvat, waren zij sterk gecorreleerd: de Pearson-correlatiecoëfficiënt was 0.89 op energiebasis en 0.93 op gewichtsbasis. Het gebruik van gewicht of energie als eenheid leidde tot significant verschillende waarden voor $Hill_1$, $Hill_2$, en $Hill_\infty$ bij dezelfde persoon. De *hazard ratio's* wezen op een omgekeerd verband tussen de indices en de totale sterfte. Wanneer de laagste (referentie) en hoogste kwintielen vergeleken werden, waren de *hazard ratio's* 0.66 en 0.65 voor $Hill_0$, 0.82 en 0.80 voor $Hill_1$, 0.84 en 0.84 voor $Hill_2$, en 0.86 en 0.90 voor $Hill_\infty$ wanneer de indexen respectievelijk op gewichts- of energiegebaseerde gegevens berekend waren. In het algemeen tonen de bekomen resultaten aan dat $Hill_\infty$ een eenvoudige en relevante index is voor de beoordeling van biodiversiteit en gezondheid.

SUMMARY

Although food biodiversity is important for sustainable healthy diets, it is still unclear which biodiversity index is most appropriate to measure food biodiversity based on food intake data, and how this index might be associated with health outcomes. The Hill numbers are simple diversity indices that allow comparison between different groups, but they have not yet been applied to diets. The $Hill_0$, $Hill_1$, $Hill_2$, and $Hill_\infty$ indices were calculated using food intake data from the European Prospective Investigation into Cancer and Nutrition (n=476,768; 9 countries). The indices were described (using both weight-based and energy-based species proportions) and their correlations were analysed. The associations of biodiversity indices with total mortality were assessed using Cox proportional hazards models. Although $Hill_\infty$ is based only on the proportion of the most abundant species in the diet while $Hill_2$ includes all proportions, they were strongly correlated: Pearson correlation coefficient of 0.89 using energy-based data and 0.93 using weight-based data. Using weight or energy as a unit resulted in significantly different values of $Hill_1$, $Hill_2$, and $Hill_\infty$ for the same diet. Hazard ratios indicated an inverse relationship between these indices and total mortality for this dataset. Comparing the lowest (reference) and highest quintiles, the hazard ratios were 0.66 and 0.65 for $Hill_0$, 0.82 and 0.80 for $Hill_1$, 0.84 and 0.84 for $Hill_2$, and 0.86 and 0.90 for $Hill_\infty$ when the indices were based on weight or energy units, respectively. Overall, the results show that $Hill_\infty$ is a simple and relevant index for assessing biodiversity and health.

1. INTRODUCTION

The Origin of Species by Charles Darwin contains only one illustration: the Tree of Life. It illustrates how all living species evolved from the same ancestor to different species over time. Millions of years of evolution have created rich biodiversity. Unfortunately, the rate of biodiversity loss, one of the nine planetary boundaries (Figure 1.1), is alarming, and humans are responsible for accelerating extinction rates by a thousand times their natural background rates (Barnosky et al., 2012; Pimm et al., 2014; Rockström et al., 2009).

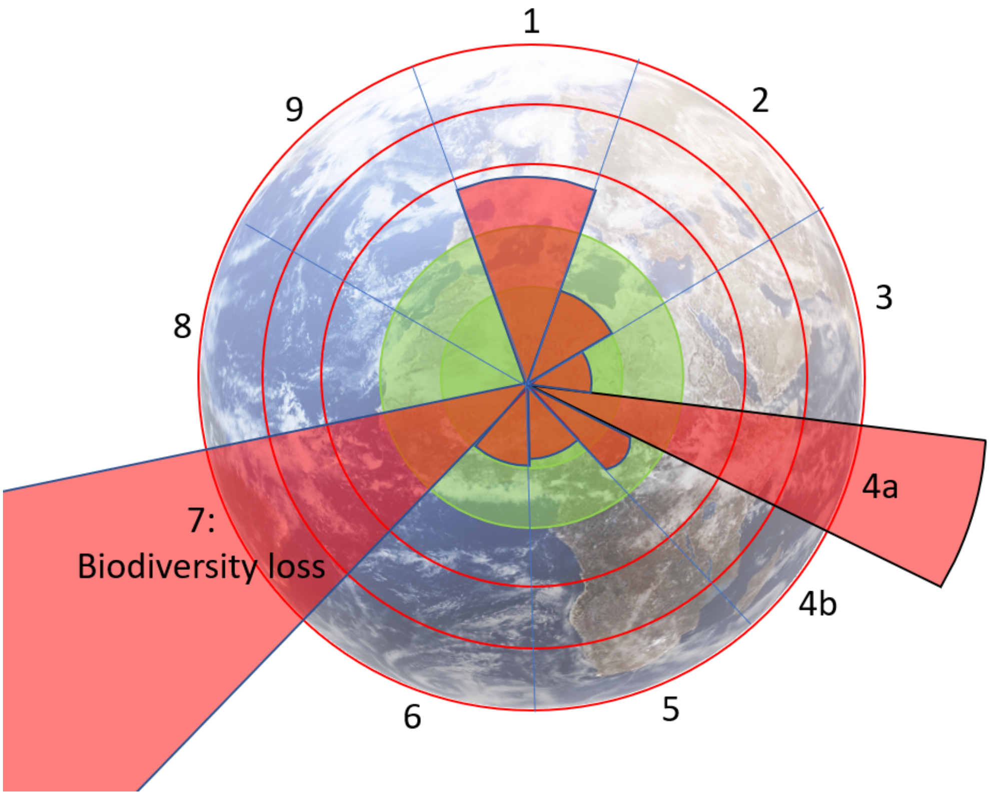

Figure 1.1: The nine planetary boundaries defined by Rockström et al. (2009), indicating biodiversity loss is occurring at an alarming rate. 1: climate change, 2: ocean acidification, 3: stratospheric ozone depletion, 4a: nitrogen cycle, 4b: phosphorus cycle, 5: freshwater use, 6: change in land use, 7: biodiversity loss, 8: atmospheric aerosol loading (not yet quantified) and 9: chemical pollution (not yet quantified). Adapted from Rockström et al. (2009).

Massive agricultural land expansions and technical evolutions made it possible to produce more calories than ever before (Ramankutty et al., 2018). But this remarkable achievement has come at a steep price. Intensification negatively impacts the diversity of cultivated crops (Antonelli et al., 2020) and domesticated farm animal breeds (Commission on Genetic Resources for Food and Agriculture [CGRFA], 2015a), which plays a crucial role in ensuring food security.

The International Union for the Conservation of Nature Red List of Threatened Species serves as the most reliable indicator of the conservation status of global biodiversity. At the agricultural level, the biodiversity of the world is rapidly declining: 17% of the world's farm animal breeds (CGRFA, 2015a) and 11% of edible plants (Antonelli et al., 2020) are now threatened with extinction.

Declining food biodiversity, which is the variety of consumed species, implies a loss of genetic diversity in species (Food & Agriculture Organization of the United Nations and Bioversity International [FAO] & Bioversity International, 2017). As a result, there is a reduction in the adaptability of species to different environmental conditions, making them more susceptible to pests, diseases, and a changing climate (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services [IPBES], 2019). Therefore, this genetic diversity is crucial to guarantee food security in the future, which for more than 2.3 billion people is not guaranteed at this time (FAO et al., 2022).

Food biodiversity is crucial not only for future food security but also for meeting current nutrient requirements. Approximately 1.6 billion people, including toddlers and non-pregnant women of childbearing age, experience deficiencies in at least one of the three essential micronutrients (iron, zinc, and vitamin A for children; iron, zinc, and folic acid for women), with the actual global population affected by these deficiencies likely exceeding this number (Stevens et al., 2022). Improving dietary diversity (i.e., eating more food groups) has a positive association with micronutrient intake (Arimond et al., 2010; Gómez et al., 2020; Kennedy et al., 2007).

To monitor and improve food biodiversity in the future, it is important to be able to quantify it. This can be achieved by using diversity indices. Despite being well-established in ecology, and their clear potential for application, species diversity indices have not yet been used widely in the context of nutrition and its associations with health consequences (Hanley-Cook et al., 2022). In this thesis, the goal is to quantify food biodiversity using ecological diversity indices.

The following research questions will be considered:

- What is the best suitable index to quantify food biodiversity in a European population in a simple way?
- How does a higher food biodiversity score of an individual in a European population correlate with overall mortality rates?
- What is the influence of using weight-based compared to energy-based species proportions on the index?

Quantifying food biodiversity has previously been studied by Hanley-Cook et al. (2021) and Lachat et al. (2018). Together with the narrative review of different food biodiversity indices by Hanley-Cook et al. (2022), these studies formed the starting point for this thesis.

The indices are calculated based on the dietary questionnaires of 476,768 people from of nine countries enrolled in the European Prospective Investigation into Cancer and Nutrition study (EPIC). EPIC is a prospective cohort study that is coordinated by the International Agency for Research on Cancer (IARC) and includes information on mortality and disease risk in Europe. (IARC, 2023)[1]

The thesis begins by providing an introduction to various definitions of food biodiversity in the literature review (Chapter 2). Subsequently, the literature review covers the benefits, current status, and reasons for the decline in food biodiversity. Additionally, it discusses current policies related to food biodiversity, along with suggestions for policy improvement. The thesis then moves on to explain indices and their application, accompanied by examples of their use by other researchers. In Chapter 3, the methodology and dataset used for the study are described in detail. Chapter 4 presents the study results, followed by a detailed discussion of the findings. In Chapter 5, recommendations for further research are given. The thesis concludes with a summary of the entire work in Chapter 6.

[1] Citing sources after the period is employed to signify that the source covers all the preceding sentences in the paragraph (Pollefliet, 2022, p.71).

2. LITERATURE REVIEW

To start the literature review it is important to explain some definitions of food biodiversity (Section 2.1). Understanding the significance of food biodiversity is essential, as it plays a vital role in sustainable healthy diets (Section 2.2.1). As its current state is worrisome (Section 2.2.2), addressing the causes of its decline is critical for improvement (Section 2.2.3). To achieve this, policy measures can be crucial (Section 2.2.4).

2.1 Definitions

Food biodiversity is a component of agricultural biodiversity, which, in turn, is encompassed within the broader concept of biodiversity (Figure 2.1). Biodiversity is defined by the Convention on Biological Diversity as the variety of all living organisms found in all ecosystems (Convention on Biological Diversity, 2006). Agricultural biodiversity (i.e., agrobiodiversity) is a subset of it. According to FAO and Bioversity International (2017), it encompasses the diversity of organisms (animals, plants, and microorganisms) that are used either directly or indirectly (i.e., soil microorganisms, predators and pollinators) in agriculture or food. Food biodiversity focuses only on the plants, animals, and other species that are actually consumed (FAO & Bioversity International, 2017).

Within food biodiversity, scientists can focus on different levels: food groups, species or variety. Food group diversity (indicated by a dietary diversity score) leads to a higher probability of meeting dietary nutrient requirements (Arimond et al., 2010; Gómez et al., 2020; Kennedy et al., 2007). However, within food groups (e.g., fruits and vegetables), there are also large differences in nutritional composition between species or varieties. For example, the vitamin A content of carrots is almost seven times higher than for mango (Aremu & Nweze, 2017), while depending on the variety of banana the vitamin A level varied by a factor of 10,000 (Burlingame et al., 2009).

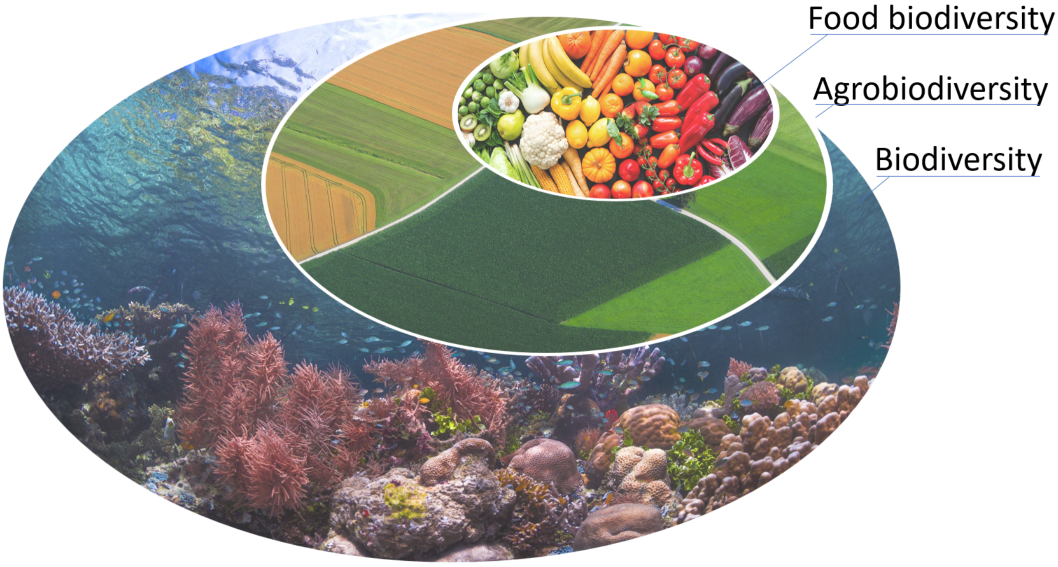

Figure 2.1: Different levels of biodiversity with respect to human consumption of food. Images provided via PowerPoint.

2.2 Food biodiversity at a glance

2.2.1 The benefits of food biodiversity

As stated in the introduction, improving food biodiversity is a new approach to help combat biodiversity loss. Food biodiversity plays a role in achieving sustainable healthy diets, which are both beneficial for the environment and human health (FAO & World Health Organization, 2019).

Each crop plays a unique role in the agroecosystem and combining them can have ecological as well as nutritional benefits. To explain this, the 'three sisters', which are corn, squash, and beans, are given as an example. By combining these three plants and taking advantage of their mutual support during growth, land use can be optimised, leading to increased yields (Risch & Hansen, 1982). Nutritionally, corn contains high proportions of carbohydrates and some amino acids, beans provide additional essential amino acids, and squash is rich in vitamin A, so that none of these plants alone can meet all nutritional requirements (Declerck et al., 2011). So, this example highlights that food biodiversity can lead to synergistic effects.

Environmental benefits

Food biodiversity boosts resilience of agrifood systems by reducing the pressure on a single species, increasing food security. In fact, the more diverse the plants and animals in an agrifood system, the more likely they are to survive challenging conditions. Agrobiodiversity enhances crop breeding possibilities by providing access to a diverse range of genes (Thrupp, 2000). Additionally, it contributes to the functionality of agricultural systems, since different crops can have different functions in the field, such as weed suppression (Bilalis et al., 2010). Unfortunately, with current production methods combined with soil depletion, declining resources, and climate change, food security cannot be guaranteed in the future (Khoury et al., 2014).

Genetic variability underpins the resilience of food systems, as stated by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, 2019). Unfortunately, the selection of the best-adapted varieties by farmers and plant breeders to maximise yield has resulted in a declining trend in genetic variability. As these selected species are so adapted to local conditions, they are unlikely to survive climate change. Therefore, enhanced genetic variability by food biodiversity ensures greater food security. (IPBES, 2019)

Conservation of diverse plants can be positively impacted by improving agricultural biodiversity. The opposite approach of targeting agricultural biodiversity is monocropping, which involves growing only one crop over a large area. However, pollinators find it harder to obtain their required nutrients in a less diverse agricultural environment (Aizen et al., 2019). Given that three out of four crops depend on pollination, these pollinators play a crucial role in the continued growth of these plants in the future (IPBES, 2019). Pollination services are important for food security and a 50% reduction could lead to 700,000 excess deaths annually from micronutrient deficiencies (Global Panel on Agriculture and Food Systems for Nutrition, 2020). Consequently, agricultural intensification, which involves a greater reliance on monocultures, can have both direct (due to the cultivation of only one crop) and indirect negative effects (the destruction of pollinators) on food biodiversity.

Nutritional and health benefits

Two effects describe the relationship between food biodiversity and nutritional benefits: the complementarity effect and the sampling effect. The complementarity effect suggests that combinations of foods can provide a more nutrient-sufficient diet due to the interactions of nutrients (Declerck et al., 2011). For example, vitamin C (e.g., in broccoli) enhances the absorption of nonheme iron from another food product (e.g., lentils) (Lynch & Cook, 1980). The sampling effect suggests that as the number of foods consumed by a person increases, the probability that one of these foods will be high in a particular ingredient also increases. Therefore, due to the sampling effect, the chances of consuming sufficient essential nutrients are greater when different types of food are consumed (Declerck et al., 2011). In addition, alternating different food items reduces the risk of eating food products with high concentrations of certain toxic substances.

Dietary diversity is associated with several health benefits. For example, diets that excluded multiple food groups were associated with higher mortality risk, according to Kant et al. (1993). Moreover, this dietary diversity score is a significant predictor of micronutrient intake in specific groups like non-breastfed children (Kennedy et al., 2007), and also in women (Arimond et al., 2010; Gómez et al., 2020). In addition to the dietary diversity score, Lachat et al. (2018) showed that species richness (total number of species consumed) showed positive associations with micronutrient adequacy. Lastly, the Women's Dietary Diversity Project developed a new indicator: Minimum Dietary Diversity for Women indicator, which focuses on the count of consumed food groups (FAO & FHI 360, 2017). According to them, the consumption of at least five out of the ten specified food groups serves as a useful proxy for ensuring micronutrient adequacy among women of reproductive age.

On the other hand, some studies have shown that dietary diversity can also have negative effects on health status. Zhang et al. (2017) concluded that a higher dietary diversity score is associated with excessive energy consumption and obesity. In addition, Bezerra and Sichieri (2011) reported a positive correlation between dietary diversity and the consumption of unhealthy food groups.

Other benefits

In addition to the environmental and nutritional benefits, food biodiversity offers a range of other advantages. These include income security for farmers and enhanced culinary experiences.

As agrobiodiversity ensures more resilience and food security, diversification also provides more certainty of income for farmers. In case of bad weather conditions, low market prices for certain products and outbreaks of diseases, farmers can rely on the sales of other crops (Zimmerer, 2015).

Furthermore, food biodiversity is linked to cultural eating habits. Traditional dishes are often made with species typically for a region. So, experimenting with different cuisines can increase the diversity of the consumed species. Moreover, diversity on a plate increases food experience. Antoni Aduriz, recognised by two Michelin stars, even has a recipe with more than one hundred ingredients (Spence et al., 2017).

2.2.2 The status of food biodiversity

An increasing number of people consume the same few species. This is because the food supply between countries is becoming more similar over time. Of 52 crops investigated in a study, all crops except cotton seeds increased in geographical spread between 1961 and 2009 (Khoury et al., 2014). Moreover, there is increasingly less variation between the diets of different people, as most of their energy intake now originates from only a few species. Worldwide, 60% of the energy intake from food comes from rice, wheat, and maize, as reported by Loftas (1995). In Europe, 45% of our energy intake comes from wheat and potatoes, combined with beef and pork (Hanley-Cook et al., 2021).

Status of plants

The number of species consumed is not necessarily driven by species loss. There are many more edible plants than are currently consumed. Willett et al. (2019) mention 14,000 edible plants, of which only 150-200 of them are eaten. Antonelli et al. (2020) provide smaller estimates but highlight a similar phenomenon: they suggest there are 7039 edible plants, of which 417 are consumed. In particular, 30% of these 7039 plants examined are on the Red List of Threatened Species of The International Union for Conservation of Nature.

Status of animals

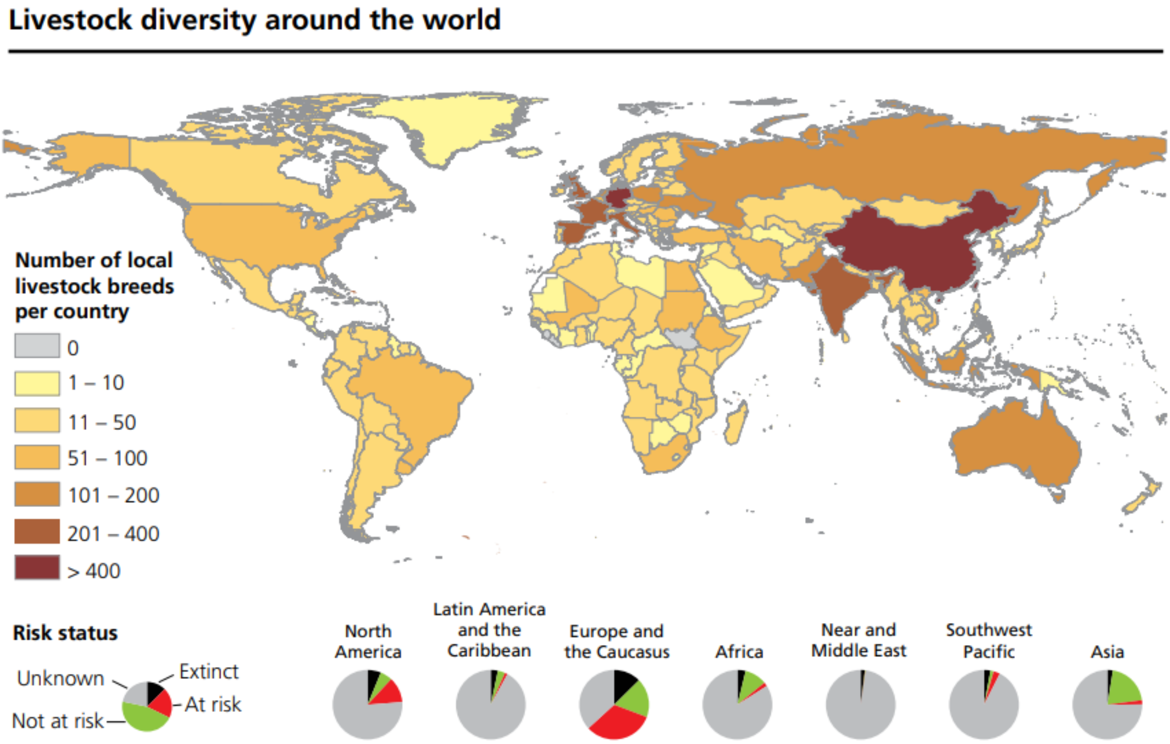

Figure 2.2: Global map with threat status of livestock breeds. From CGRFA (2015b).

Figure 2.2 shows a map of the number of animal breeds per country and risk status per continent (CGRFA, 2015b). In general, few terrestrial animal species, namely 38 domesticated birds and mammals, are used in agriculture and food production (CGRFA, 2015a). According to them, 17% of farm breeds worldwide are at risk of extinction, but this could be an underestimation as 58% of them had an unknown risk status. However, their study also mentions that there are many more animal species than those domesticated for food today: only 10% of animal species are currently raised for human consumption, and an even lower proportion of avian species, only 0.1%.

In addition to plants and terrestrial animals, fish are also frequently consumed in human diets. But only 10 of the 600 farmed species in aquaculture account for more than 50% of overall productivity (FAO, 2020). Due to increasing demand, fishing has had the greatest impact on the biodiversity of marine ecosystems in the last 50 years, according to IPBES (2019). To meet this growing demand, an increasing area of the ocean has been fished, and an increasing number of fish have been or are being overfished. IPBES mentions that, currently, more than 55% of the oceans are fished on a large scale. In addition, of all marine fish stocks, 7% are underfished, 33% are overexploited, and the rest are maximally sustainably fished. Next to the loss of biodiversity, overfishing also has an impact on the fish itself: they mature much faster to increase their reproduction rates (IPBES, 2019). Furthermore, global warming has caused local marine biodiversity to change, as fish move to colder areas (Hannesson, 2007), also affecting food security.

2.2.3 The causes of declining food biodiversity

There are several causes of declining food biodiversity, but most are linked to the growing world population. As the population grows exponentially (Figure 2.3), the demand for food also increases. Increasing the supply leads to more specialisation and deforestation. Both factors drive deterioration in food biodiversity, and are discussed in more detail below.

First, government subsidies have stimulated higher production yields. On the one hand, this has helped to meet the increased demand. On the other hand, this has also forced farmers to specialise in a few crops and to stop the production of more diverse and resilient local varieties. (Antonelli et al., 2020) Moreover, this has discouraged the production of foods that do not receive the same level of support, like fruits and vegetables (FAO et al., 2022). Therefore, some species became more affordable, compared to others, resulting in a nutrition gap (Ramankutty et al., 2018).

Second, to meet rising demand, the biodiversity of the environment is being sacrificed. Almost 90% of global deforestation is due to agriculture: almost 50% is for planting crops and 38.5% is for grazing (FAO, 2022). Deforestation has a major irreversible impact on biodiversity, because it enormously reduces both wildlife and suitable habitat, but also on the available wild foods which are important for several communities and for increasing resilience (Antonelli et al., 2020).

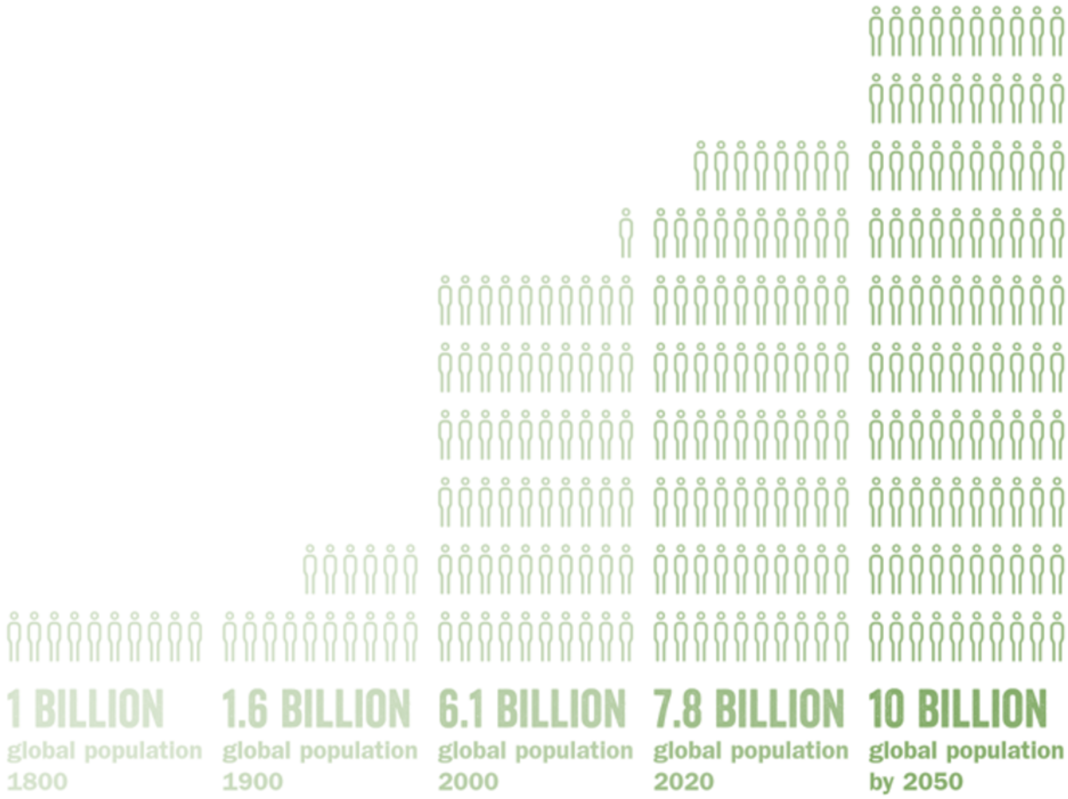

Figure 2.3: The trend of the exponentially growing world population. From Antonelli et al. (2020).

2.2.4 Food biodiversity policies

Although food consumption and production play a central role in the climate and biodiversity crises, little attention has been paid to them in the global policy agenda, according to EAT-Lancet Commission (2020). This report states that “food has so far not been considered central to global policy agendas such as the Paris Agreement, Sustainable Development Goals (SDGs), or Convention on Biological Diversity” (EAT-Lancet Commission, 2020, p.7), which means there is still room for improvement. In the following paragraphs, the role of food in the SDGs and the Convention on Biological Diversity is explained.

Current policy

With the introduction of 17 SDGs, the United Nations has attempted to set targets by 2030 to improve sustainable living conditions on Earth. The targets relevant to improving food biodiversity are as follows:

- SDG 2 aims for improvement in food security and sustainable agriculture;
- SDG 12 aims for sustainable consumption and production and less food loss;
- SDG 13 aims to strengthen resilience to combat climate change;
- SDG 14 aims to end overfishing and restore fish stocks;
- SDG 15 aims to limit deforestation due to land conversion. (General Assembly of the United Nations, 2015)

Unfortunately, according to IPBES (2019), many of the SDGs related to biodiversity will not be reached by 2030 if we continue to live as we currently do.

The Convention on Biological Diversity set Aichi biodiversity targets that align well with the SDGs. These 20 targets were planned to be met by 2020. Unfortunately, none of the targets was fully met. In their vision by 2050, the Convention of Biological Diversity wants to commit to a shift in food production so that food can be provided in a sustainable way for all people. Unfortunately, we are also not on track to achieve the goals by 2050 either. (Secretariat of the Convention on Biological Diversity, 2020)

Improvements in policy

Governments and other organisations can help slow down food biodiversity loss. For example, subsidies can be shifted from focusing on mass production to sustainable and healthy food production (Secretariat of the Convention on Biological Diversity, 2020). To improve food biodiversity and food security, farmers must increase the genetic diversity of their crops and animals (IPBES, 2019). Furthermore, more attention can be paid to both wild and locally grown crops which are now underutilised, such as bulbous chervil

(*Chaerophyllum bulbosum*) (Antonelli et al., 2020). Finally, efforts to prevent food losses (e.g., improving technology) can be improved, as this reduces the pressure on production (Secretariat of the Convention on Biological Diversity, 2020).

Reducing food waste can be a solution to meet this rising demand without sacrificing biodiversity. According to United Nations Environment Programme (2021), approximately 931 million tons of food (17% of total production) is lost globally. Interestingly, this study reveals that waste per person within a household is similar between different country income groups. This challenges previous narratives that emphasised the significance of consumer food waste mainly in high-income countries. The results of United Nations Environment Programme suggest that actions to reduce consumer food waste are relevant in all countries, regardless of their income level.

2.3 Quantifying food biodiversity

Indices are essential to monitor and evaluate the state of food biodiversity, and to guide policies. They quantify food biodiversity and allow us to determine which diets are most biodiverse. This quantification is useful to make comparisons between people, as well as to measure change over time.

2.3.1 Components of food biodiversity

Biodiversity indices are based on one or more components of food biodiversity (Section 2.3.1). Food biodiversity is subdivided into three components: richness, evenness, and disparity. In Figure 2.4 the concept is illustrated.

Richness is defined as the total number of elements (e.g., species) in a diet, so it is independent of the proportion of each element. To make determination possible, it is important to know what to consider as a single element. For instance, counting can be performed on the basis of food groups like for the Minimum Dietary Diversity for Women index (Food & Agriculture Organization of the United Nations & FHI 360, 2016), but also more specifically on the species level. It should be noted that this statement is not always straightforward: for example, ecologically, beef and milk belong to the same species, but in nutrition belong to different functional food groups (i.e., flesh foods vs dairy) (Hanley-Cook et al., 2022).

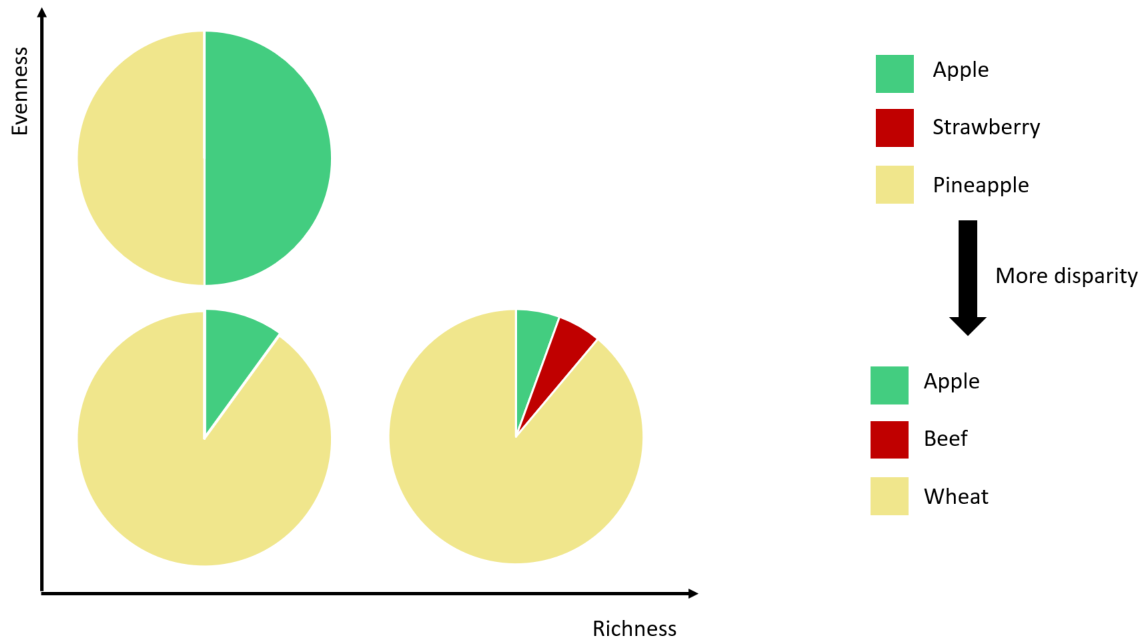

Figure 2.4: Three components of food biodiversity: richness, evenness, and disparity.

In contrast to richness, evenness is based on the proportion of elements in a diet. Evenness reflects the distribution of these proportions in the diet: the more food items that are consumed in equal amounts, the more even the diet will be. However, from a nutritional perspective, it is not always advantageous to consider equal amounts as a better diet. For example, looking at food groups as elements, the Flemish Food Triangle points out that meat has to be consumed in smaller amounts than fruits and vegetables.

Disparity compares different foods according to their composition, while in the other components nutritional differences in the other components are not explicitly considered. Depending on the research objectives, different functional traits of foods (e.g., iron content) can be used to compare different food products.

There exist many indices that measure one or more of these components in an ecological setting. However, there are important differences between ecological and nutritional contexts, as mentioned in the preceding paragraphs. Therefore, attention should be paid to the fact that these indices were not originally developed with nutrition in mind.

2.3.2 Notation

S, $\mathbf{p}$ and α are the parameters commonly used to calculate food biodiversity.

S represents the total number of species. In food data, composite products (e.g., pancakes) are traced back to their ingredients (e.g., flour, milk, and eggs) to determine the total number of species.

The vector $\mathbf{p}$ combines the proportions of S food elements. Each element, denoted as p_i, takes a value between 0 and 1. Therefore, the elements of $\mathbf{p}$ always sum up to 1. These elements can be calculated, for example, energy-based and weight-based. On an energy basis, p_i is determined as the energy of species i relative to the total energy of all consumed species, whereas p_i on a weight basis is equal to the weight of species i divided by the total weight consumed.

S and $\mathbf{p}$ are common to all indices, but some indices have additional parameters, for example, a sensitivity parameter α in the Hill numbers, which defines if the emphasis is placed on species with high or with low abundances (Section 2.3.3).

2.3.3 Hill numbers

The Hill numbers are a family of indices that contain (transformations of) most classical diversity indices. These classical indices do not include disparity, which means that they consider each food item to be equally different from another food item. Hill numbers include richness indices and heterogeneity indices. The latter method combines richness and evenness. These indices are discussed in Section 2.3.4.

The general formula to calculate the Hill numbers is shown in Equation 2.1 (Hill, 1973).

$$Hill_\alpha = \left(\sum_{i=1}^{S} p_i^\alpha\right)^{\frac{1}{1-\alpha}} \tag{2.1}$$

The sensitivity parameter α is adjusted according to the index that is calculated. The higher the value of α, the more emphasis is placed on the food items present in large proportions in the diet and vice versa. (Hill, 1973)

The output of H_α is a number of effective species S', which represents the number of different species consumed if the diet were to be perfectly uniform (Hill, 1973). If a person's diet has a Hill number equal to x, then this would mean that their diet is as diverse as another person's diet who consumed x different species in exactly the same proportions. The number of effective species can therefore be calculated by substituting every proportion with $1/S'$ (Heip et al., 1998).

Conversion to a number of effective species means that it is possible to compare Hill numbers as they have the same unit, regardless of the value of α. An ecology forum held by Ellison (2010) suggests that even if the aim is to describe the diversity of a single person's diet, using number of effective species is still the best diversity measure.

2.3.4 Classical cardinal indices

Suppose two people each eat two species per day. One eats everything in equal amounts (i.e., the vector of relative abundances is given by $\mathbf{p} = (0.5, 0.5)$), while for another person the relative abundances of the species in the diet are given by $\mathbf{p} = (0.9, 0.1)$. Are these then equally diverse? No, because both evenness and richness are important parameters for assessing diversity. (Peet, 2003)

Heterogeneity indices take proportions as well as species number into account and therefore represent both richness and evenness. According to Ricotta (2005), there is a 'jungle' of these indices. In the following paragraphs, only three of them are discussed: the Shannon, Simpson, and Berger-Parker indices. These indices are chosen based on their previous applications to quantify food biodiversity (Hanley-Cook et al., 2022).

Hill number with $\alpha = 0$ and the richness index

Richness is a cardinal index that represents the number of objects in a dataset. In the case of dietary species richness, it counts the number of unique species in a diet. Hanley-Cook et al. (2021) showed that dietary species richness is inversely associated with the total risk of mortality rates in nine European countries.

Richness can be derived from the formula for the Hill numbers by setting α equal to 0, but the value of the index is of course just equal to S by definition (Hill, 1973):

$$R = Hill_0 = \left(\sum_{i=1}^{s} p_i^0\right)^{\frac{1}{1-0}} = S \tag{2.2}$$

Hill number with $\alpha = 1$ and the Shannon index

The Hill number with $\alpha = 1$ is equivalent to the exponent of the Shannon index H_{Sh} (Equation 2.3) (Hill, 1973). Expressing this index in effective species makes it easier to interpret, as the Shannon index has nats or bits as a unit (Heip et al., 1998).

$$\exp(H_{Sh}) = Hill_1 = \lim_{\alpha \to 1}\left(\left(\sum_{i=1}^{S} p_i^{\alpha}\right)^{\frac{1}{1-\alpha}}\right) = \exp\left(-\sum_{i=1}^{S} p_i \ln(p_i)\right) \tag{2.3}$$

The Shannon index was originally created to estimate how many yes/no questions are needed to guess the next letter in a word when the frequency of a letter is known. It measures entropy and is expressed in bits if the base of the logarithm is 2 and in nats if the natural log is calculated. (Shannon, 1948) In dietary data, it could be described as an index that determines how many yes/no questions are needed to reconstruct the order of a person's diet if it is known how much of each species is consumed. Currently, the Shannon index is the most widely used diversity index in food supply studies (Hanley-Cook et al., 2022).

Remans et al. (2014) used the Shannon diversity index to measure food production and supply diversity at the national level, finding significant regional differences in food production diversity and supply diversity. Additionally, they observed an inverse relationship between the Shannon index based on food supply and the prevalence of child stunting, wasting, and being underweight. Gustafson et al. (2016) developed a new sustainable nutrition security assessment methodology that incorporates the Shannon index based on national food production data to holistically evaluate food system performance across multiple domains, including nutrition, environment, economic, social, resilience, safety, and waste. Tian et al. (2017)

used individual food consumption data to compare the diversity of food consumed by men and women and found that dietary diversity was positively associated with overweight in men.

Hill number with $\alpha = 2$ and the Simpson index

Another way to define biodiversity is using the Simpson index (H_{Si}). The indicator is based on the probability of randomly selecting two items from a diet that are of the same food species, which equals p_i^2 (Equation 2.4). These probabilities are calculated and summed for all species. The explanation above shows that a more diverse diet, which has a lower probability of selecting the same species twice in a row, causes a lower index.

$$H_{Si} = \sum_{i=1}^{S} p_i^2 \tag{2.4}$$

As the Simpson index is lower for a more diverse diet, the Gini-Simpson index (= $1 - H_{Si}$) or the Simpson-dominance index (= $1/H_{Si}$, Equation 2.5) are more logical to interpret and therefore are often used (Daly et al., 2018; Heip et al., 1998).

$$\frac{1}{H_{Si}} = H_{Simpson\ dominance} = Hill_2 = \left(\sum_{i=1}^{S} p_i^2 \right)^{\frac{1}{1-2}} = \frac{1}{\sum_{i=1}^{S} p_i^2} \tag{2.5}$$

The Gini-Simpson index is used by de Oliviera Otto et al. (2015) to calculate the diversity of food items using food intake data based on food frequency questionnaires. In their study, higher diversity was weakly positively associated with dietary quality. According to Lachat et al. (2018), richness has a stronger positive association micronutrient adequacy compared to the Gini-Simpson index, using food intake data of several African and Asian countries.

Katanoda et al. (2006) proposed a new index for food biodiversity called Quantitative Index for Dietary Diversity ($H_{QUANTIDD}$) (Equation 2.6). It is based on the complement of the Simpson index, but the index is calculated for n food groups rather than S species. Therefore, p_i is equal to the relative abundance of a food group. Furthermore, it is normalised by dividing by the maximum value that can be reached, so when p_i is equal to $1/n$. The authors used the index on Japanese consumption data with the proportions based on energy intake as well as on the amount of food intake and found that the value for $H_{QUANTIDD}$ increased from the 1960s to the 1970s, which was a period of rapid economic growth in Japan.

$$H_{QUANTIDD} = \frac{1 - \sum_{i=1}^{n} p_i^2}{1 - 1/n} \tag{2.6}$$

Hill number with $a = \infty$ and the Berger-Parker index

The final index to be discussed is the Berger-Parker index (H_{BP}), which equals the Hill number with $a = \infty$ (Equation 2.7). Along with the Simpson index and Shannon index, this index is also widely used to describe food biodiversity (Hanley-Cook et al., 2022). This simple index only takes into account the prevalence of the most abundant species. Thus, a considerable amount of information is omitted from the diversity estimation, but data collection is easier as it can be restricted to a sample of regularly consumed species.

$$H_{BP} = Hill_{\infty} = \lim_{a \to +\infty} \left(\left(\sum_{i=1}^{S} p_i^a \right)^{\frac{1}{1-a}} \right) = \frac{1}{\max(\mathbf{p})} \tag{2.7}$$

The Berger-Parker index was used to define diversity based on national food supply data by Khoury et al. (2014). Their findings indicated a global decline in this index from 1961 to 2009.

2.3.5 Drawbacks of classical indices

Unfortunately, there are still problems with quantification. First, there is still no consensus on the definition of species diversity (Moreno & Rodríguez, 2010). In addition, while indices are translated from ecology to food, they were not originally created to describe food biodiversity. Two examples of clear differences are the units (i.e., number of species in ecology versus weight or energy of the species contributing to the diet) and data collection (i.e., investigation of species in a certain geographical area in ecology versus dietary questionnaires for food). Third, there exist many biodiversity indices, each having different calculation methods and potentially having other units. As a result, researchers frequently choose different indices for similar research objectives, making it difficult or even impossible to compare studies in the existing literature (Hanley-Cook et al., 2022).

Consider the following example to illustrate the drawbacks of using the most common food biodiversity indices: richness, Shannon index, and Gini-Simpson index (Hanley-Cook et al., 2022). Suppose that person A eats 25 species in equal amounts, person B eats 50 species in equal amounts and person C eats also 25 species, but unevenly divided:

$$\mathbf{p} = (0.4, 0.2, 0.1, 0.05, 0.05, 0.01, 0.01, \cdots, 0.01).$$

The values of the indices for these three diets are shown in Table 2.1.

Table 2.1: Example of richness, Shannon and Gini-Simpson index using two diets with evenly divided species for person A and B and unevenly divided species for person C. The units of the Shannon index are nats, and the Gini-Simpson index is a probability.

	Richness	**Shannon**	**Gini-Simpson**
Person A	25	3.22	0.96
Person B	50	3.91	0.98
Person C	25	2.14	0.78

The first conclusion is that it is difficult to compare different indices as different units are used. Moreover, there is clearly an issue due to the nonlinearity of the indices: although intuitively person A has a diet twice as diverse as person B, this is not reflected by the Shannon or Gini-Simpson indices (Daly et al., 2018). Another problem with the Shannon index is that it is difficult to estimate if 3.91 or any value is ‘good’ in terms of diversity, since its units do not have an intuitive interpretation.

Calculating food biodiversity in terms of Hill numbers can solve these problems. In Table 2.2, the Hill numbers ($Hill_0$, $Hill_1$, $Hill_2$, and $Hill_\infty$) are shown.

Table 2.2: Example of Hill numbers using two diets with evenly divided species for person A and B and unevenly divided species for person C. The units of the Hill indices are effective number of species.

	$Hill_0$	$Hill_1$	$Hill_2$	$Hill_\infty$
Person A	25	25	25	25
Person B	50	50	50	50
Person C	25	8.49	4.61	2.5

According to the data presented in Table 2.2, it is evident that person A's diet is twice as diverse as person B's, regardless of the chosen Hill number. It is worth noting that the Hill indices all have identical values because the species in the diet are perfectly evenly distributed, although this is rarely the case in practical scenarios. Person C, on the other hand, shares the same $Hill_0$ value as person A, but their values diverge for other Hill numbers. As the value of α increases, the corresponding Hill indices of person C decrease. This can be attributed to the presence of numerous species in small proportions: with lower values of α, greater emphasis is placed on rare species, and vice versa. This explains why the values are larger for $Hill_0$ than for $Hill_\infty$.

2.3.6 Leinster-Cobbold index

To take into account all three components of biodiversity (richness, evenness, and disparity), the Hill numbers do not suffice, since they do not incorporate disparity. In the Leinster-Cobbold index this extra element, called a similarity matrix, is included. (Leinster & Cobbold, 2012)

$$D_\alpha = \left(\sum_{i=1}^{S} p_i \left(\mathbf{Zp}\right)_i^{\alpha-1} \right)^{\frac{1}{1-\alpha}}, \tag{2.8}$$

$$(\mathbf{Zp})_i = \sum_{j=1}^{S} Z_{ij} p_j \tag{2.9}$$

The similarity matrix **Z** is a SxS matrix that combines all similarities between different species. An element of the matrix, Z_{ij}, is a normalised distance that is based on the parameters chosen according to the research objectives (e.g., fibres, vitamin A, carbohydrates and iron). Z_{ij} is equal to 1 if species i is completely similar to species j and equal to 0 if they are completely different. In the formula for the Hill numbers, a similarity value of zero is assigned to each pair of different species, so the similarity matrix is equal to the identity matrix. (Leinster & Cobbold, 2012)

The significance of $(\mathbf{Zp})_i$ is the probability that a random species is similar to the i^{th} species (Equation 2.9). Therefore, it measures the ordinariness of a species, which is higher when there are more species similar to the i^{th}

species. This ordinariness is at least equal to the abundance of the species. (Leinster & Cobbold, 2012)

As $(\mathbf{Zp})_i$ signifies ordinariness, the average ordinariness is given $\sum_{i=1}^{S} p_i (\mathbf{Zp})_i$. This measure, referred to as concentration, has a high value when the majority of the diet is concentrated in a small number of closely related species. Although the average ordinariness is inversely related to diversity, the negative exponent in the Leinster-Cobbold index allows it to serve as a measure of diversity. (Leinster & Cobbold, 2012)

2.3.7 Distances for calculating the Leinster-Cobbold index

Based on some selected food traits, it is possible to incorporate a similarity matrix into diversity indices. In this way, the Leinster-Cobbold index, can take into account that, for example, lettuce and cucumber are nutritionally less diverse than chicken and cucumber.

In order to define a similarity matrix, the distance metric must be determined. There are different options for this, and two of them will be discussed below: the Euclidean distance and the Jaccard distance.

Note that to determine the similarity matrix, the distances have to be converted from dissimilarity to similarity measures. This can be done by taking the complement of the dissimilarity value.

First, an example is introduced to illustrate (Table 2.3). Suppose iron and fibre are used as characteristics, so that the iron value is plotted on the x-axis and fibre on the y-axis. For each species in the diet, the relative amount of iron and fibre is calculated by dividing by the maximum amount of iron in one of the species. These are thus all relative amounts of a nutrient and are now between 0 and 1 and unitless.

Table 2.3: Example of iron and fibre content of two theoretical species.

species	**iron (mg/100g)**	**fibre (mg/100g)**	**relative iron**	**relative fibre**
i	0.5	0	1	0
j	0.25	2.5	0.5	1

The Euclidean distance is a simple measure to compare two different food products based on their characteristics. Each characteristic is plotted on a different axis. Then, the food products are represented by dots based on their characteristics. The Euclidean distance between the two points represents the dissimilarity.

The Euclidean distance d_{ij} is given by the following formula:

$$d_{ij} = \sqrt{(i_{iron} - j_{iron})^2 + \left(i_{fibre} - j_{fibre}\right)^2} = 1.11 \tag{2.10}$$

For n characteristics, the traits of all species are plotted in an n-dimensional space and the distance between two species would be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (i_k - j_k)^2} \tag{2.11}$$

Another measure of distances that can be used with food data is the Jaccard distance (Hanley-Cook et al., 2022; Wang et al., 2021). The Jaccard distance is calculated by dividing the unique share (US) of two food products by the sum of the unique and the common share (CS) (Wang et al., 2021). The 'common share' refers to the amount of the nutrient or component that is present in both food products, while the 'unique share' is the (excess) amount present in one of the food products compared to the other one.

Suppose the same example as above:

$$CS = \max(i,j)_{iron} + \max(i,j)_{fibre} = 2 \tag{2.12}$$

$$US = \max(i,j)_{iron} - \min(i,j)_{iron} + \max(i,j)_{fibre} - \min(i,j)_{fibre} = 1.5 \tag{2.13}$$

$$d_{ij} = \frac{US}{CS} = 0.75 \tag{2.14}$$

For n characteristics, for all species, the minimum and maximum of all n characteristics would be determined and used to calculate the Jaccard distance (Wang et al., 2021):

$$d_{ij} = 1 - \frac{\sum_{k=1}^{n} min(i,j)_k}{\sum_{k=1}^{n} max(i,j)_k} \tag{2.15}$$

2.3.8 Similarity-sensitive indices

A similarity-sensitive index used in food data is called Rao's quadratic diversity index (*Rao*) (Hanley-Cook et al., 2022). To write this diversity index in terms of the Leinster-Cobbold index, the reciprocal of the complement of it is used (Equation 2.16) (Leinster & Cobbold, 2012).

$$\frac{1}{1-Rao} = D_2 = \left(\sum_{i=1}^{S} p_i\,(\mathbf{Zp})_i^{2-1}\right)^{\frac{1}{1-2}} = \frac{1}{\left(\sum_{i=1}^{S} p_i\,(\mathbf{Zp})_i\right)} \tag{2.16}$$

Green et al. (2021) used the Rao's quadratic diversity index to determine food biodiversity based on worldwide food supply data of FAO of 2020 using the Euclidean distance. They found that the values of this index were lower in Europe than for other regions with similar incomes. Furthermore, Wang et al. (2021) defined nutritional redundancy (*NR*) in an individual's dietary intake assessment as the portion of their food diversity, calculated by the Gini-Simpson index ($1-H_{Si}$), that cannot be explained by their nutrient diversity, calculated by *Rao* using the Jaccard distance. This results in the equation $NR = 1 - H_{Si} - Rao$. They observed that nutritional redundancy and nutrient diversity tend to be similar across different dietary studies.

2.3.9 Drawbacks of similarity-sensitive indices

The major drawback of using the Leinster-Cobbold index is the inconsistency of the similarity measure between different studies. This similarity matrix can namely be different in every study because the researcher determines the traits and the distance measure. On the other hand, this also means that the index is versatile and can focus on different objectives (e.g., functional or phylogenetic differences). Another drawback is the computational load. To compare the diets of two persons based on the species they consumed, the distances between every two species must be calculated.

2.4 Conclusion

Food biodiversity, which focuses on the diversity of the crops, animals, and other species that are consumed, is crucial for attaining a sustainable healthy diet. Food biodiversity enhances food security by promoting adequate nutrient intake now and guaranteeing food supply in the future.

Although there are many edible species, an increasing number of people consume the same few species. Communication about the advantages and implementing food biodiversity in the global policy agenda will help to tackle food biodiversity loss.

Quantifying food biodiversity is an important step in addressing biodiversity loss and guiding policies. While several indices have been developed to measure biodiversity, most of them have been designed for ecological settings, and it may be worth using them for food data as well. As shown in Table 2.4, these indices vary significantly in terms of the biodiversity components they measure, their data needs and their mathematical behaviour. Hence, careful interpretation is needed when applying them to food data and the impact of diverse diets on human health can be used for validation.

In this thesis, the focus will be on determining the best-suited index for quantifying the diversity of diets in a simple way using food intake data. One approach is the use of the Hill numbers that can take into account both species richness and evenness. Some researchers have already investigated the performance of other indices, such as the Shannon or Simpson indices, using food intake, production or supply data, but to our knowledge, the Hill numbers have never been used to quantify the diversity of diets using food intake data before. The Hill numbers are particularly well-suited for quantifying food biodiversity as the data needed, calculation method and interpretation are quite simple compared to other indices. Although these biodiversity indices offer a few conceptual advantages, empirical evaluation on real diets remains necessary, which presents research challenge that will be tackled in the next chapters.

Table 2.4: Overview of the characteristics of possible food biodiversity indices.

	Shannon	Simpson	$Hill_0$	$Hill_1$	$Hill_2$	$Hill_\infty$	Leinster-Cobbold
Food biodiversity components							
Richness	+	+	+	+	+	+	+
Evenness	+	+	-	+	+	+	+
Disparity	-	-	-	-	-	-	+
Comparability with other research							
Frequently used in literature	+	+	+	-	-	-	-
Uniformity of calculation	+	+	+	+	+	+	-
Use of index							
Easy interpretation of unit	-	-	+	+	+	+	+
Avoids nonlinearity issues	-	-	+	+	+	+	+
Easy interpretation of calculation	-	+	+	-	+	+	-
Use of data							
Simplicity of data collection	-	-	+	-	-	+	- -
Independence of details questioned	+	+	-	+	+	+	-
Biased towards . . .							
Common species	+	+	-	-	+	+	depends
Rare species	-	-	+	-	-	-	depends

3. METHODOLOGY

3.1 Research goals

As outlined in the literature review, the Hill numbers show promise for quantifying food biodiversity. This thesis aims to evaluate the applicability of four Hill numbers to food intake data. Specifically, $Hill_0$ (richness), $Hill_1$ (the exponential of the Shannon index), $Hill_2$ (the reciprocal of the Simpson index), and $Hill_\infty$ (the Berger-Parker index) will be examined. Additionally, a newly created index, $Jill_x$, will be tested on the data.

The food intake data used in this research is sourced from the European Prospective Investigation into Cancer and Nutrition study, provided by the International Agency for Research on Cancer (IARC, 2023). This dataset comprises food intake information collected between 1992 and 2000 as part of the EPIC study. Furthermore, by using this dataset, it is possible to explore the correlation between indices and health, specifically by examining overall mortality rates.

By using real food intake data, deeper insights can be gained into the practical suitability of the indices that have shown promise in the literature.

3.2 Data collection within the European Prospective Investigation into Cancer and Nutrition study

The European Prospective Investigation into Cancer and Nutrition study is a continuing prospective cohort study (IARC, 2023). The EPIC study examines the associations between lifestyle, dietary, environmental, and metabolic variables and cancer and other chronic diseases. Local ethics committees and the Internal Review Board of the International Agency for Research on

Cancer approved the study, and all participants gave their written informed consent.

The EPIC study includes more than 500,000 volunteers (between 25 and 70 years old) enrolled between 1992 and 2000 in 23 administrative centres in 10 different countries: Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom. Most of the participants are chosen from the general population in a particular region, but the exceptions were the cohorts from:

- France: female participants in a health insurance programme for school personnel;
- Utrecht in the Netherlands: attendees at breast cancer screenings;
- Ragusa in Italy: blood donors and their spouses;
- Oxford in the United Kingdom: mainly vegetarian and healthy eaters.

Data collection, recruitment, and study design details have previously been published (Ferrari et al., 2008; Riboli et al., 2002; Riboli & Kaaks, 1997).

3.2.1 Baseline data collection

Each participant was subjected to a detailed characterisation before participation. Sociodemographic data, educational level, personal and family medical history, lifestyle factors (such as smoking, alcohol consumption, and physical activity), and menstrual and reproductive histories of women were collected using questionnaires. In each centre, anthropometric measurements, such as height, weight, and waist and hip circumferences, were taken, but in France, Oxford, and Norway self-reported data were used (Haftenberger et al., 2002). When self-reported measurements were not available, average values based on centre, age, and gender were used.

3.2.2 Dietary data collection

Validated dietary questionnaires (DQ) were used to assess the diet over the preceding year of each participant at enrolment. The method was based on food frequency questionnaires (FFQs) and varied per study centre with an estimation of individual average portions or with the same standard portion assigned to all participants, as well as questionnaires combining an FFQ and 7-day dietary records (Table 3.1). Extensive quantitative dietary questionnaires were used in Germany, northern Italy (including Florence, Turin, and Varese), and the Netherlands to estimate individual average portions. In France, Ragusa (Italy), and Spain, these quantitative dietary questionnaires were structured by meals. Semiquantitative FFQs were used in Denmark, Naples (Italy), Norway, Umeå (Sweden), and the UK. FFQs were combined with a 7-day dietary record in the UK and with a 14-day record on hot meals in Malmö (Sweden). With the exception of Ragusa and Naples (Italy), and Spain, where face-to-face interviews (FTF) were conducted, the DQs were self-administered (SA) in most locations. (Riboli et al., 2002) Using established techniques (such as breaking down recipes into their component parts), post-harmonisation of DQ data was carried out to produce a standardised food list whose degree of information is comparable between centres.

The foods registered in the FFQ (e.g. recipes, drinks, composite food) were converted into species using the FoodEx2 classification of the European Food Safety Authority (Hanley-Cook et al., 2021). Food that could not be classified into a single species was classified with the detailed EPIC food classification system (NCLASS). Minced meat is an example of an NCLASS group and could not be classified because it can theoretically consist of, for example, chicken, beef, pig or even a mixture of these.

3.2.3 Vital status and cause of death follow-up

Most countries used connections to health boards and population-based cancer, health insurance, pathology, and mortality registers to acquire data on vital status and death. In contrast, Germany used unattended follow-up mailings and subsequent enquiries to municipality registries, regional health departments, physicians, or hospitals to identify deceased individuals. In France, the national death index and the school employee health insurance database were used to retrieve data on people who had passed away. Depending on the country, the study's end of follow-up/closure dates ranged from 2009 to 2014.

Table 3.1: Data collection of EPIC participants (Riboli et al., 2002). FFQ: food frequency questionnaire, FTF: face-to-face and SA: self-administered.

Country	**Centre**	**Method**	
Denmark	Aarhus	FFQ: semi-quantitative	SA
Denmark	Copenhagen	FFQ: semi-quantitative	SA
France	France	FFQ: structured by meals	SA
Germany	Heidelberg	FFQ: individual average portions	SA
Germany	Potsdam	FFQ: individual average portions	SA
Italy	Ragusa	FFQ: structured by meals	FTF
Italy	Florence	FFQ: individual average portions	SA
Italy	Turin	FFQ: individual average portions	SA
Italy	Varese	FFQ: individual average portions	SA
Italy	Naples	FFQ: semi-quantitative	FTF
Netherlands	Bilthoven	FFQ: individual average portions	SA
Netherlands	Utrecht	FFQ: individual average portions	SA
Norway	Tromsø	FFQ: semi-quantitative	SA
Spain	Granada	FFQ: structured by meals	FTF
Spain	Murcia	FFQ: structured by meals	FTF
Spain	Navarra	FFQ: structured by meals	FTF
Spain	San Sebastian	FFQ: structured by meals	FTF
Spain	Asturias	FFQ: structured by meals	FTF
Sweden	Malmö	FFQ: non-quantitative + 14-day record on hot meals	SA
Sweden	Umeå	FFQ: semi-quantitative	SA
United Kingdom	Cambridge	FFQ: semi-quantitative + 7-day record	SA
United Kingdom	Oxford	FFQ: semi-quantitative + 7-day record	SA
Greece	Excluded		

3.3 Data processing

The analysis was performed using *R* (R Core Team, 2022), *RStudio* (Posit team, 2023), and several packages. All scripts are made available via https://github.com/jidygers/ThesisFoodBiodiversity.

3.3.1 Exclusion of data

In this thesis, the dietary data of 476,768 participants is used to calculate the selected biodiversity indices. The analyses included 451,390 participants (with 46,627 recorded deaths between 1992 and 2014) to determine mortality risk ratios. The exclusion criteria are summarised in Figure 3.1.

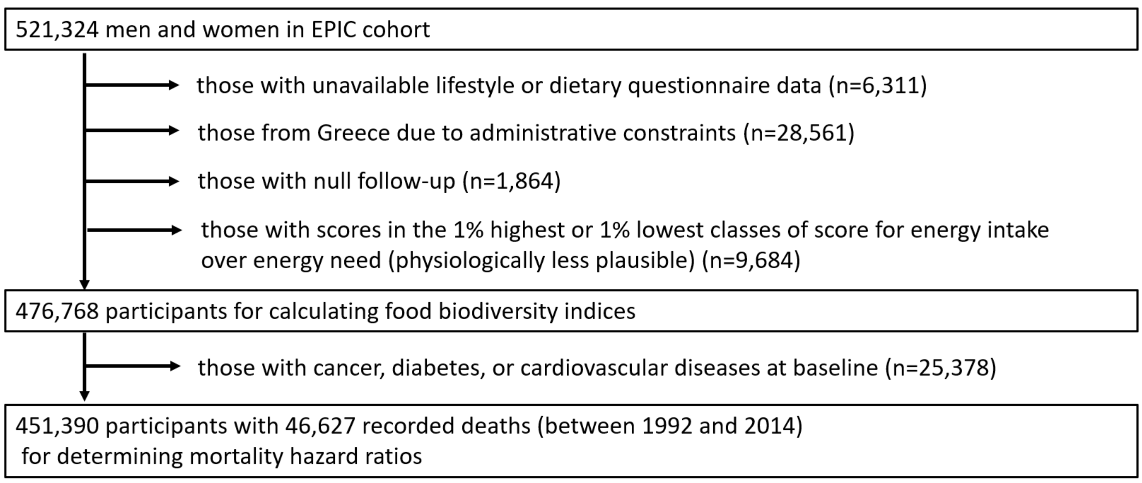

Figure 3.1: Flowchart of exclusion of the data obtained from the EPIC study.

To facilitate communication, the term "full dataset" will be used to refer to the group of the 476,768 individuals used for calculating the food biodiversity indices. On the other hand, the term "overall mortality dataset" will be used to denote the subset of 451,390 participants specifically used for determining the hazard ratios related to overall mortality.

3.3.2 Collaboration with IARC

Detailed food intake data of the EPIC study of IARC was used to calculate the selected biodiversity indices. Due to data protection regulations, the dietary intake data of all individuals could not be shared outside of IARC and its data servers. Instead, I was permitted to write scripts that were then sent to IARC for their staff to perform the analysis using the protected dataset.

As a first step, a small test dataset was created and shared with me, compromising the dietary intake of 93 randomly selected people, for the purpose of allowing me to understand the data structure and content. The term "test dataset" will be used to refer to this. Based on these data, I wrote scripts in *R* to calculate food biodiversity indices and calculate the average amount of each species consumed.

Then, these scripts were sent to IARC where they were executed by the data analysts using the full dataset (n=476,768), and the results were returned to me for further analysis. These outputs included the calculated indices along with the country of origin, age at recruitment, gender, and body mass index (BMI) for each individual. Furthermore, multivariable-adjusted Cox proportional hazards regression models were run by IARC data scientists, and the results, including hazard ratios, were returned to me for further analysis. Only for these models, I did not personally write scripts. Additionally, I created graphs based on the provided outputs and data to visualise the results and support further analysis and interpretation.

3.3.3 Food biodiversity calculation

The food biodiversity indices chosen for this thesis were based on the most suitable indices identified in the literature (Section 2.3). Specifically, Hill numbers with $\alpha = \{0, 1, 2, \infty\}$ were calculated using dietary intake data of the EPIC study to estimate the number of effective species in individuals' diets.

$$Hill_0 = S \tag{3.1}$$

$$Hill_1 = \exp\left(-\sum_{i=1}^{S} p_i \ln(p_i)\right) \tag{3.2}$$

$$Hill_2 = \frac{1}{\sum_{i=1}^{S} p_i^2} \tag{3.3}$$

$$Hill_\infty = \frac{1}{\max(\mathbf{p})} \tag{3.4}$$

In the equations above, S represents the total number of species in an individual's diet, and p_i is the proportion of species i in that person's diet. p_i is calculated in two different ways: based on both the diet expressed in terms of weight (g/day) and energy (kcal/day), allowing comparison between these two units. On an energy basis, p_i is determined as the energy contribution of species i relative to the total energy of all species consumed. On a weight basis, p_i is calculated as the weight of species i divided by the total weight consumed by the individual. All values of the Hill numbers are expressed in effective species.

Moreover, a new index named $Jill_x$ was proposed, which is based on $Hill_\infty$. The purpose of $Jill_x$ is to incorporate multiple data points while minimising the requirement for detailed diet proportions. Specifically, $Jill_x$ is calculated as the inverse of the average abundance of the x most abundant species in the diet. It can therefore be thought of as an extension of $Hill_\infty$, by considering not one but x species. By varying the value of x and comparing the results with other indices, insights can be gained into the optimal number of species to include in the index calculations.

The index is also expressed in terms of effective species, which can be demonstrated by considering a scenario where S' species are divided equally ($p_i = 1/S'$). In this case, the resulting value of the index would be equivalent to the number of species present in the diet:

$$Jill_x = \frac{1}{\frac{\sum_{i=1}^{x} 1/S'}{x}} = S' \tag{3.5}$$

The indices are calculated using the *hillR* package (Li, 2018) and *R* base functions.

3.3.4 Comparison and visualisation of indices

Pearson and Spearman correlation coefficients were calculated to compare the different indices and to compare the values based on weight and energy.

Furthermore, the associations between the indices based on quintiles were examined.

Quintiles, which divide a population distribution into five equal parts, are commonly used in large data sets to facilitate the drawing of conclusions. Given the noise in the data collected, percentiles between different indices are unlikely to match. Working with a less detailed categorisation, such as quintiles, allows indices to be compared while taking account of natural variability.

In cases where strong correlations exist between two indices, preference can be given to the index that has a simpler calculation method, since this thesis prioritises the practical application of biodiversity indices to dietary intake data, rather than the indices' mathematical behaviour.

The mortality hazard ratios calculated by IARC (by generating multivariable-adjusted Cox proportional hazards regression models) were used to validate the relation between the indices and health outcomes. One criterion for the food biodiversity index is that it should exhibit a negative correlation with overall mortality. For these models, the data were stratified by gender, age and study centre and adjusted for smoking status, educational level, marital status, physical activity, alcohol intake and total energy intake, Mediterranean diet score, red and processed meat intake, and fibre intake. The stratification and adjustment variables were chosen according to the article of Hanley-Cook et al. (2021). Moreover, the absolute number of deaths per 10,000 person-years without adjustment and stratification was calculated to compare the indices. The absolute number of deaths per 10,000 person-years can be understood as the average number of individuals expected to die within a one-year observation period if a population of 10,000 people is monitored (Tenny & Boktor, 2022).

Based on all outcomes, various graphs were created using the *ggplot2* package (Wickham, 2016) and the *GGally* package (Schloerke et al., 2021) for analysis and visualisation.

3.4 Polishing text and improving word variation

To improve the readability of the written content of this thesis, ChatGPT, a language model developed by OpenAI, was used. The goal of using ChatGPT was to improve the self-written text by refining sentence structure, improving grammar and introducing broader vocabulary.

The use of ChatGPT in refining the text encouraged critical thinking, since ChatGPT also introduced alternative perspectives and rephrased sentences that deviated from the original intention. This required careful review and evaluation of the revised text to ensure that the intended message was conveyed correctly.

4. RESULTS AND DISCUSSION

In this chapter the food biodiversity indices, calculated using the EPIC dataset described in the previous chapter, are visualised and discussed. The indices considered are as follows: $Hill_0$ (richness), $Hill_1$ (exponential of the Shannon index), $Hill_2$ (reciprocal of the Simpson index), $Hill_\infty$ (Berger-Parker index), and $Jill_x$. All these indices are expressed in effective species. A detailed explanation of the formulas for these indices was provided in Section 2.3 and Section 3.3.3.

4.1 Overview of index distributions

4.1.1 Overview

Figure 4.1 provides an overview of the distributions of the four calculated indices based on the data obtained from the EPIC dietary questionnaires.

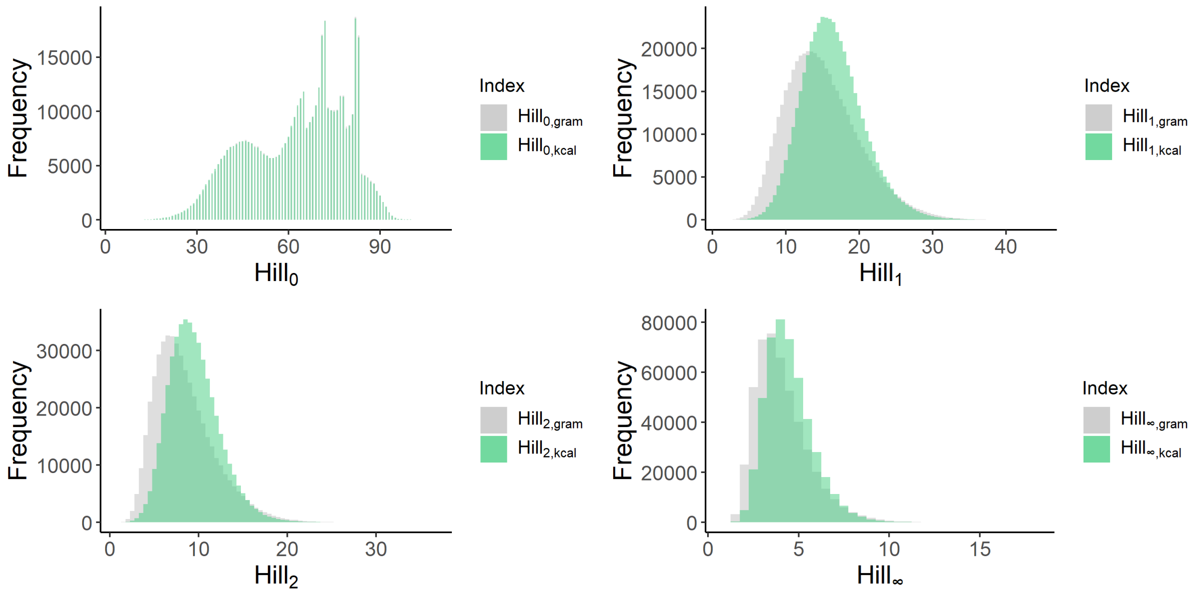

Figure 4.1: Overview of the distribution of the weight- and energy-based indices using the full dataset (n=476,768). The graphs of $Hill_{0,gram}$ and $Hill_{0,kcal}$ coincide.

The graphs of $Hill_{0,gram}$ and $Hill_{0,kcal}$ coincide. This equivalence between the distributions of $Hill_{0,gram}$ and $Hill_{0,kcal}$ is due to the fact that $Hill_0$ is not affected by the proportions of the foods, making it unit independent (Section 4.5). The distributions of $Hill_1$, $Hill_2$ and $Hill_\infty$ are remarkably smoother and closer to a skewed normal distribution than $Hill_0$, for both weight-based and energy-based indices.

In addition, the distributions of the indices using energy-based data are shifted to the right, with the exception of $Hill_0$. This implies that the values of the food biodiversity indices are on average higher when calculated using energy-based data than when using weight-based data. The observed rightward shift in the distribution of weight-based indices compared to energy-based indices can be attributed to the more even distribution of the latter.

Moreover, the number of effective species between the different indices is significantly different. The absolute values of a food biodiversity index cannot be interpreted in isolation; reference values for determining what is considered 'good' or 'bad' are necessary to properly interpret the values. Instead, the focus of this section is on identifying differences between indices of the same type, assessing the correlation between different indices, and comparing the classification of individuals based on different indices. It is important to note that comparisons are only valid when made between individuals using the same index or when comparing different indices for a single person, but not when comparing two different indices for two different individuals.

4.1.2 Respondent characteristics

The four indices were each plotted against four distinct respondent characteristics: country of residence, gender, age, and BMI. Analysis of the plots revealed noticeable variations in $Hill_0$ values across different countries, as demonstrated in Figures 4.2 and 4.3. On the contrary, no apparent differences were observed for the other characteristics and indices of the respondent, as depicted in the Appendix (Figures A1, A2, A3, A4, A5 and A6). This means, for example, that at the population level, comparisons of $Hill_\infty$ values between individuals of different weight classes can be made without the need to take gender differences into account.

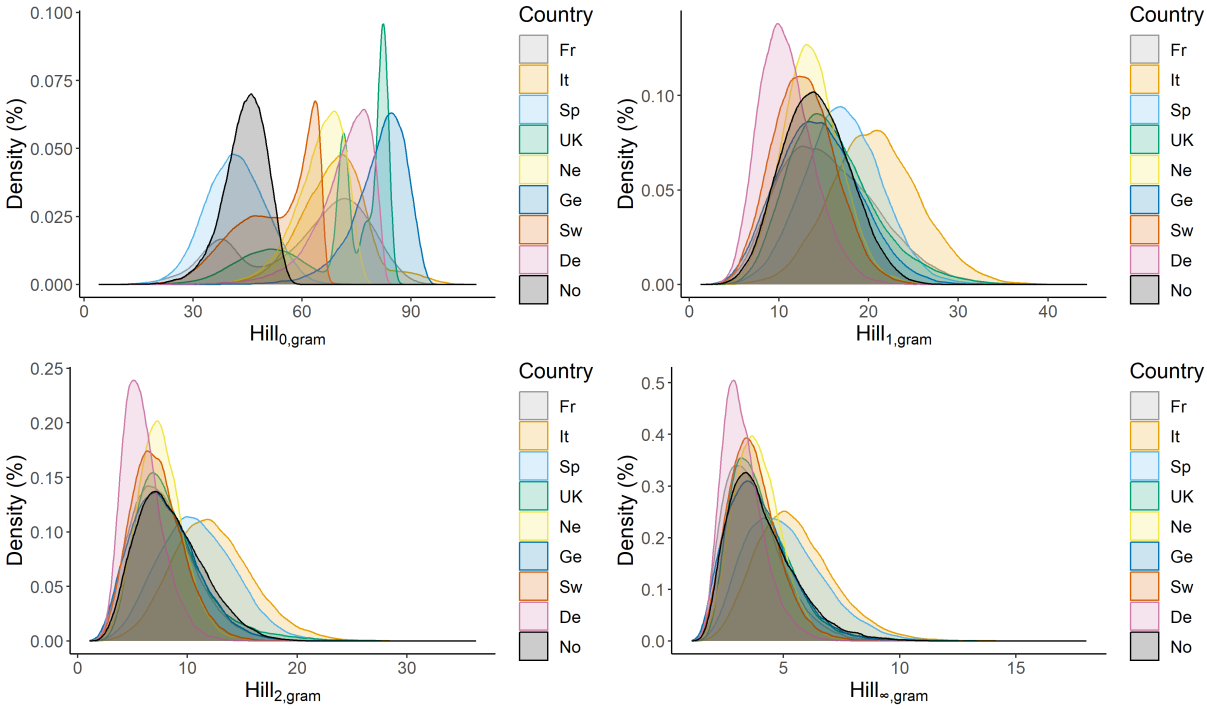

Figure 4.2: Association between weight-based indices and respondent's country of residence using the full dataset (n=476,768). Fr: France, It: Italy, Sp: Spain, UK: United Kingdom, Ne: The Netherlands, Ge: Germany, Sw: Sweden, De: Denmark and No: Norway.

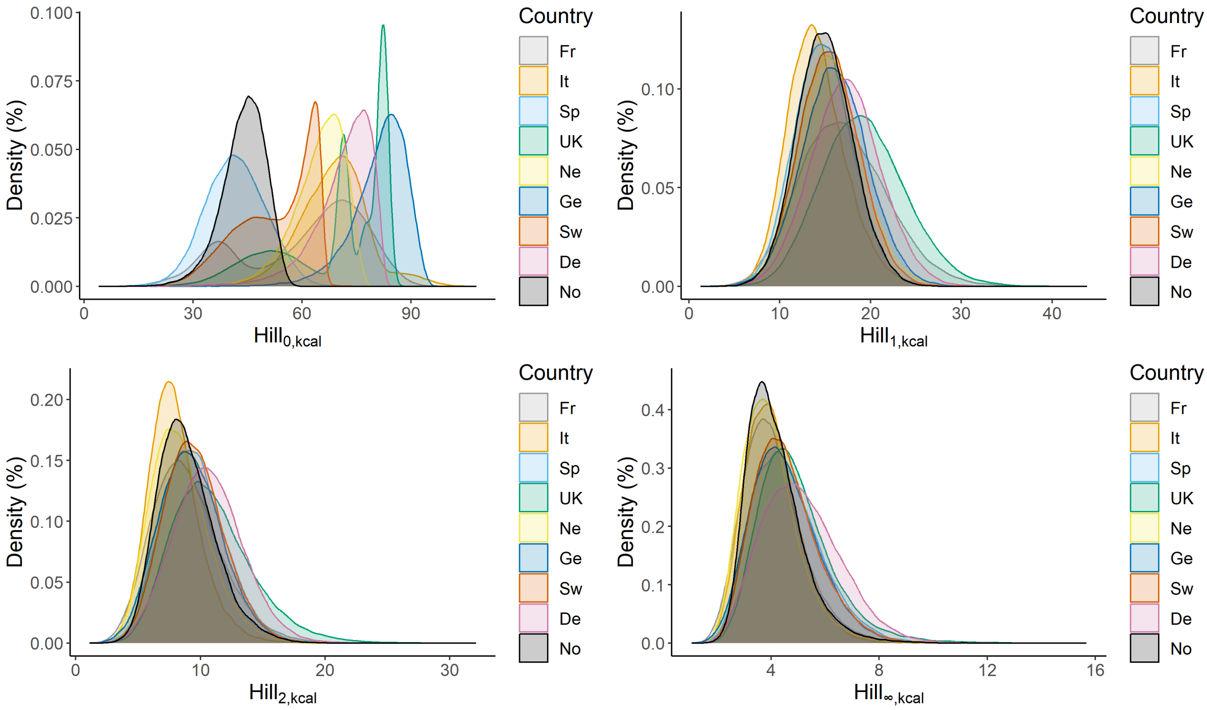

Figure 4.3: Association between energy-based indices and respondent's country of residence using the full dataset (n=476,768). Fr: France, It: Italy, Sp: Spain, UK: United Kingdom, Ne: The Netherlands, Ge: Germany, Sw: Sweden, De: Denmark and No: Norway.

$Hill_0$ could be heavily influenced by the level of detail included in the food frequency questionnaire: the higher the level of detail, the higher the corresponding richness. As the food frequency questionnaires in EPIC differed between the participating countries (Riboli et al., 2002), this could explain the differences (Section 4.6). On the other hand, it can also be assumed that the consumption profiles varied between countries, which is why the distribution of $Hill_0$ varied. However, if this statement is true, this is not visible in other indices.

4.2 Comparison of the Hill numbers

One index can provide an identical conclusion or an interpretation of the data as another index, although its calculation may be simpler or requires fewer data. Comparing of different Hill numbers for the same input data provides information on whether using different indices to compare people's diets in terms of food biodiversity would give the same results.

4.2.1 Correlation between the indices

Figures 4.4 and 4.5 give an overview of the correlation between the calculated indices. The lower triangle of the figures shows two-dimensional density plots that visualise the relationships between each index pair, while the upper triangle displays the corresponding Pearson correlation coefficients. The diagonal of the figures illustrates the density plots of each individual index. Based on these figures, several observations can be made.

First, the two-dimensional density plots in the lower triangle representing $Hill_0$ as a function of other Hill numbers exhibit clear groupings. This is due to the expression of richness in natural numbers while the other Hill numbers have continuous-valued outputs.

Secondly, in the case of weight-based calculations, the correlation between $Hill_0$ and the other indices is notably low: ranging from 0.039 to 0.239. On the other hand, the correlation between the other indices is substantially higher, with values ranging from 0.788 to 0.931.

For the energy proportion-based indices, the correlation between $Hill_0$ and the other indices is slightly higher: ranging from 0.146 to 0.404. In con-

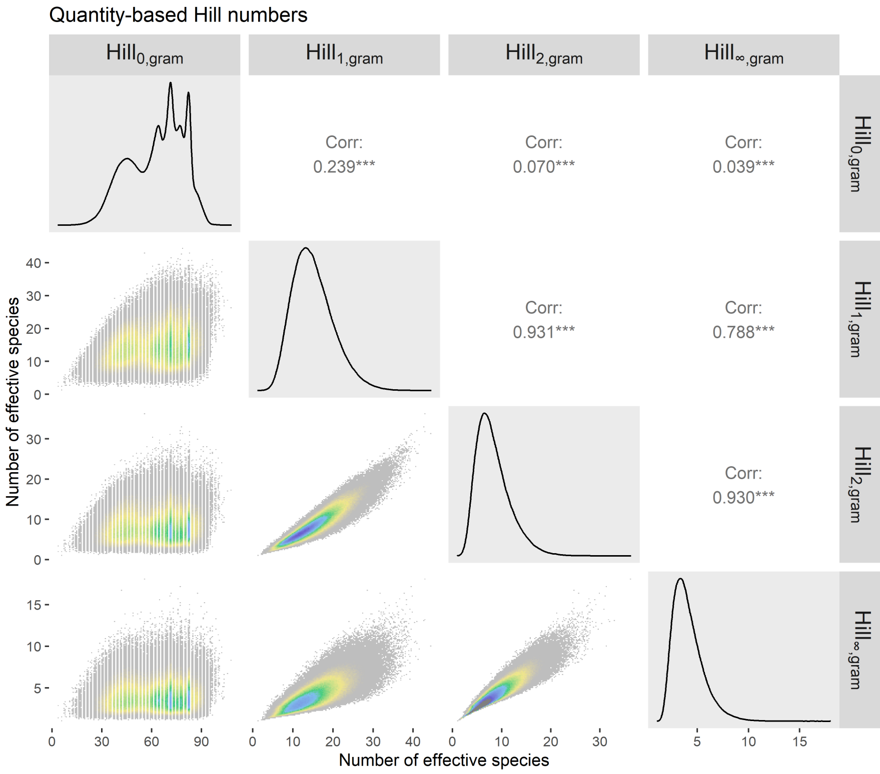

Figure 4.4: Relationships between weight-based Hill numbers and Pearson correlation coefficient (Corr) using the full dataset (n=476,768).

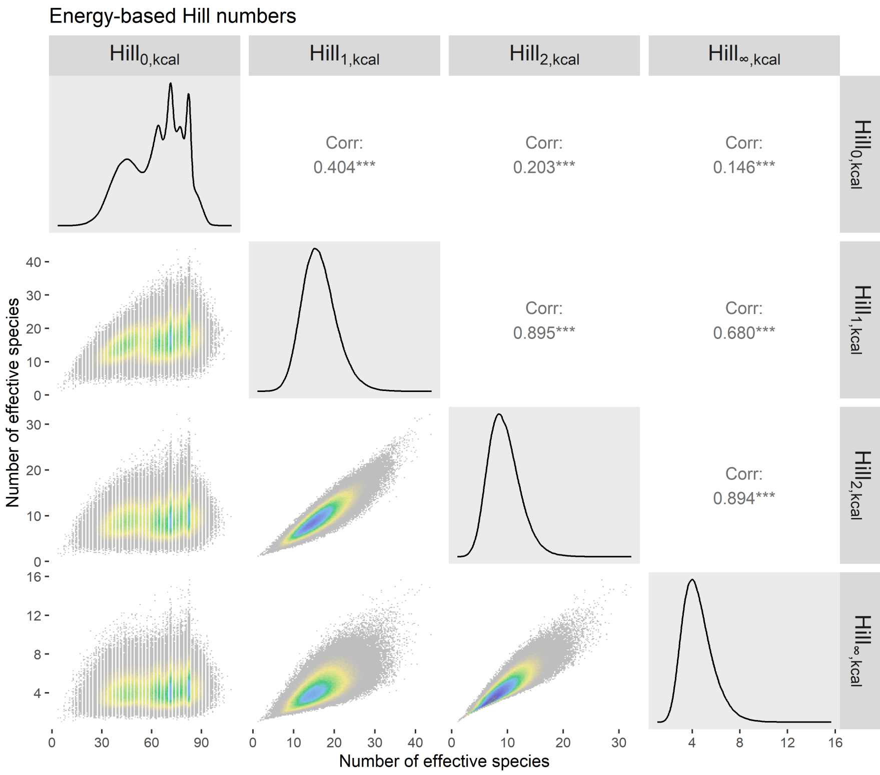

Figure 4.5: Relationships between weight-based Hill numbers and Pearson correlation coefficient (Corr) using the full dataset (n=476,768).

trast, the correlation between $Hill_1$ and $Hill_2$ is slightly lower for the energy proportion-based indices: 0.680. Meanwhile, the correlations between $Hill_2$ and $Hill_1$, and $Hill_2$ and $Hill_\infty$ are similar to the weight-based indices: about 0.9.

It is worth noting that, based on the results, the Spearman correlation coefficients are similar to the Pearson correlation coefficients for each pair of indices (Appendix Table A1). This implies a strong monotonic relationship where there is a strong linear relationship.

In addition to Figures 4.4 and 4.5, the associations between the indices based on quintiles were examined. In this case, people were classified into quintiles and each pair of indices was compared to determine whether someone is classified in the same quintile based on different indices (Figure 4.6).

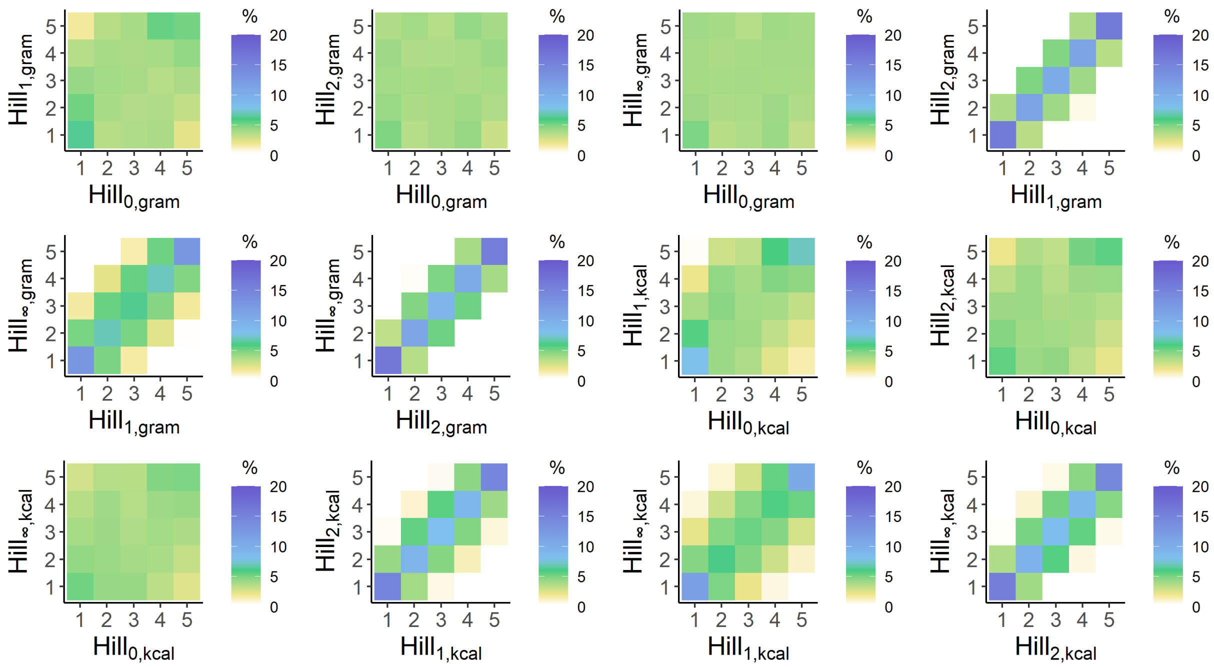

Figure 4.6: Comparative analysis of quintile divisions for different indices using the full dataset (n=476,768). The colour code represents the percentage of the whole cohort classified in a given quintile for the index on the x-axis and a given quintile for the index on the y-axis.

In this context, if a person's food biodiversity score is classified within the same quintile or differs by only one quintile for both calculation methods, it is considered acceptable. However, if the classification of their scores differs by more than one quintile, they are considered to be mismatched. A summary of the percentage of people mismatched is given in Table 4.1, where the upper triangle represents the weight-based indices, and the lower triangle represents the energy-based indices.

Comparison of $Hill_0$ with the other indices revealed a higher percentage of mismatches for the weight-based indices, ranging from 42% to 47%, compared to energy-based indices which had a relatively lower range of 34% to 43%. On the contrary, for the other indices, the energy-based indices exhibited higher percentages of mismatches than the weight-based indices. The indices that displayed the most consistent classification were $Hill_1$ compared to $Hill_2$ and $Hill_2$ compared to $Hill_\infty$, with percentages of mismatches ranging between 1.4% and 5.5%.

Table 4.1: Total percentages of mismatches in quintiles based on the different Hill numbers using the full dataset (n=476,768). A mismatch is defined as the classification of a person's Hill numbers deviating by more than one quintile. Upper triangle represents the weight-based indices and lower triangle represents the energy-based indices.

	$Hill_0$	$Hill_1$	$Hill_2$	$Hill_\infty$
$Hill_0$		42%	47%	47%
$Hill_1$	34%		2.2%	13%
$Hill_2$	42%	5.5%		1.4%
$Hill_\infty$	43%	20%	4.0%	

Feasibility of division in quintiles

Since Figures 4.2 and 4.3 show that there is a difference in the distribution between European countries, it is helpful to examine whether the boundaries of each quintile are sufficiently far apart to divide the population into quintiles. If not, it is possible that small changes in diet, or small changes in estimates of the amount of food consumed could have large effects on a person's quintile classification. To examine this possibility, Figure 4.7 visualises the quintile cut-off points per person and index.

The absolute distance between the two points indicates the feasibility of classifying the population into quintiles: the greater the absolute distance, the easier it will be to categorise. From this graph, it can be deduced that the range between different cut-offs in the same country is quite similar for the indices and there are small variations between the countries. Overall, the boundaries for each index and each country are far enough apart to allow a meaningful quintile-based classification of the data.

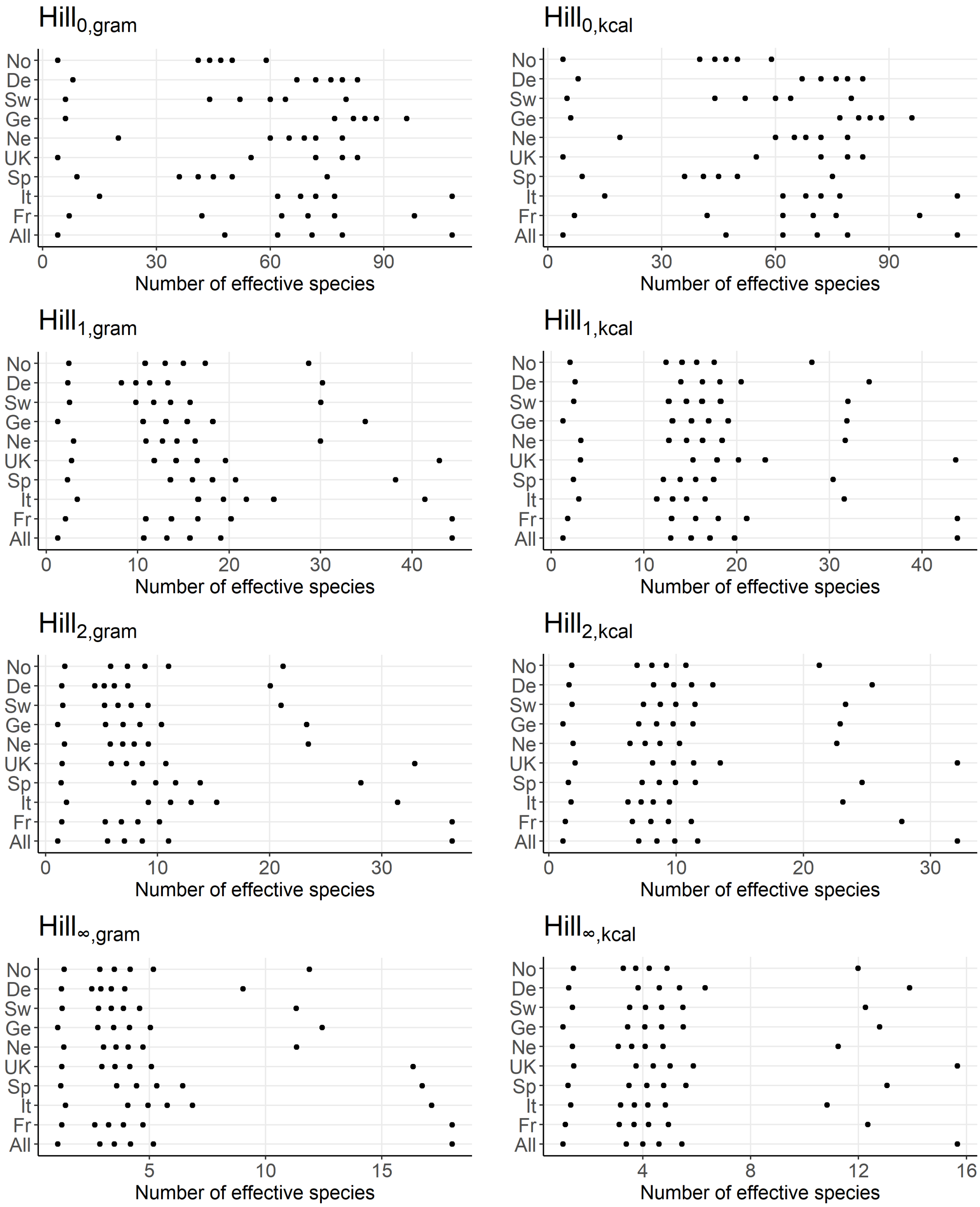

Figure 4.7: Boundaries of the quintiles per index and per country using the full dataset (n=476,768). Fr: France, It: Italy, Sp: Spain, UK: United Kingdom, Ne: The Netherlands, Ge: Germany, Sw: Sweden, De: Denmark and No: Norway.

4.2.2 $Hill_1$ compared to the other Hill numbers

Due to the strong correlation demonstrated between $Hill_1$ and $Hill_2$, along with the low number of mismatches, the results shown here indicate that these indices can be used almost interchangeably to compare different diets in terms of their biodiversity scores. On one hand, $Hill_1$ is widely used because of its exceptional capacity to weigh items accurately according to their frequency without favouring either rare or common elements (Jost, 2006). On the other hand, higher-order Hill numbers can typically be estimated with greater confidence, since lower-order Hill numbers tend to overemphasise the presence of rare species, where the sampling process is more uncertain (Daly et al., 2018). Furthermore, within the Hill numbers calculated in this thesis, the calculation of $Hill_1$ is more complex, as it is based on logarithmic and exponential factors, while $Hill_2$ is determined by a simpler function (Section 3.3.3). Therefore, $Hill_2$ is preferred over $Hill_1$ for practical applications to dietary intake data, as there is less uncertainty and it is easier to interpret the calculation (Daly et al., 2018; Heip et al., 1998).

Whereas $Hill_1$ does not favour common or rare species, both $Hill_0$ and $Hill_\infty$ are heavily influenced by rare and common species, respectively. However, assigning more weight to rare or common species is not necessarily negative, but this bias should be kept in mind when using this index and align with the specific research objectives. Due to ease of calculation and simple data collection of $Hill_0$ and $Hill_\infty$, they are preferred over $Hill_1$ in this thesis.

4.2.3 $Hill_2$ compared $Hill_0$ and $Hill_\infty$

Figures 4.4, 4.5 and 4.6 provide valuable insights into the relationships between different Hill numbers as measures of food biodiversity. Analysis revealed similar classification profiles and a strong correlation between $Hill_2$ and $Hill_\infty$, with a Pearson correlation coefficient of approximately 0.9 for both weight- and energy-based data. These results suggest that the two indices can be used almost interchangeably for comparing different diets in terms of their biodiversity scores, since the conclusions drawn from one index would be similar to those drawn from the other. Consequently, if a simple and efficient measure of food biodiversity is desired, $Hill_\infty$ may be preferred over $Hill_2$, as it only requires the abundance of the most common species to be determined, whereas $Hill_2$ needs the proportions of all species in the diet (Hill, 1973).

4.2.4 Comparison $Hill_0$ and $Hill_\infty$

Despite the strong correlation between $Hill_2$ and $Hill_\infty$, the analysis shown here also indicated a lack of correlation between $Hill_0$ and either of the other two indices. This suggests that $Hill_0$ may not be used interchangeably with the other indices. Logically, the correlation is less pronounced, as this index does not capture the underlying patterns of species evenness in the diet (Hanley-Cook et al., 2022). As both $Hill_0$ and $Hill_\infty$ are simple and efficient measures of food biodiversity, the advantages and disadvantages are discussed in the next section.

4.3 Trade-offs of $Hill_0$ and $Hill_\infty$

4.3.1 Need for a common unit

Calculating species richness does not require the selection of a unit for quantifying the portions in the diet, as it is not based on abundance of species. This is an advantage of $Hill_0$ over $Hill_\infty$, as $Hill_\infty$ has substantial differences in the outcome based on the selected unit. This will be discussed later in Section 4.5.

4.3.2 Data requirements

On one hand, $Hill_0$ is more advantageous than $Hill_\infty$ because it does not require the collection of proportions of species consumed, which is challenging to obtain (Rathje & Murphy, 2001; Smith et al., 1991; Thiébaut et al., 2007). Although proportions of species in the diet are not necessary, an extensive food frequency questionnaire is crucial to accurately predict the number of species consumed.

On the other hand, the most significant advantage of $Hill_\infty$ is that it requires only one datapoint, which is the proportion of the most abundant species in a person's diet. However, the prediction of this is difficult, which means that more than one data point will be required to determine $Hill_0$ more accurately. Nevertheless, unlike $Hill_0$, the questionnaire could be limited to species that are likely to be commonly consumed by a given population.

However, this requires a preparatory step to identify the most commonly consumed species.

Note that for both $Hill_0$ and $Hill_\infty$, decomposition using standard recipes is crucial to determine the outcome more accurately. For more information on the challenges regarding data collection, see Section 4.6.

4.3.3 Robustness to errors in the data

When calculating $Hill_\infty$, errors in the collection of data of the most abundant species have clear and significant effects on the values. However, the chance of a participant in the EPIC study being classified in the lowest quintile instead of the highest quintile is rather small, since the consumption of the one or two most consumed species needs to be reduced by about 540 grams or 370 kcal on average to go from the first quintile to the fifth quintile (Table 4.2). Similarly, the most consumed species in the diet should be reduced by 360 grams or 230 kcal to classify someone in the first quintile instead of the fifth. Therefore, it is unlikely that one is mismatched in the first instead of in the fifth quintile or vice versa.

Table 4.2: Difference between first and fifth quintile for $Hill_\infty$ investigated on the test dataset (n=93).

	$Q_1 \longrightarrow Q_5$	$Q_5 \longrightarrow Q_1$
Proportion of p_{max}	Decrease	Increase
Average change [standard deviation] (grams)	539 [289]	372 [120]
Average change [standard deviation] (kcal)	357 [145]	232 [93]

In contrast to $Hill_\infty$, errors in the recorded data of abundant species have less impact on the values of $Hill_0$. With an average of 63 and a standard deviation of 16 species, adding one extra species has minimal impact.

4.3.4 Respondent characteristics

As stated in Section 4.1.2, the differences between countries are only visible for $Hill_0$, so if the goal is to compare diets between countries, this index is preferred. However, differences in the food frequency questionnaires or in the cohorts (e.g., a higher proportion of vegetarian and healthy eaters in Oxford compared to other cities) could also explain the observed differences in the distributions of the values of $Hill_0$. Therefore, further research is required to identify the drivers of these differences between the distributions.

4.3.5 Validation based on mortality rates

Hazard ratios are used to validate the relationship between the indices and health outcomes. It is expected that a higher food biodiversity score would correspond to lower overall mortality rates. In other words, an increase in the food biodiversity index should be associated with a reduced risk of mortality. Figure 4.8 shows graphs that visually represent the relationship between mortality hazard ratios and quintiles for both $Hill_0$ and $Hill_\infty$ using the overall mortality dataset (n=451,390). For each quintile, the mortality hazard ratio is compared to a reference, which is the quintile with the highest probability of mortality. Thereby, the first quintile is equal to the reference in this case. The horizontal error bars show the 95% confidence intervals. Similar graphs for $Hill_1$ and $Hill_2$ are given in Appendix (Figure A7).

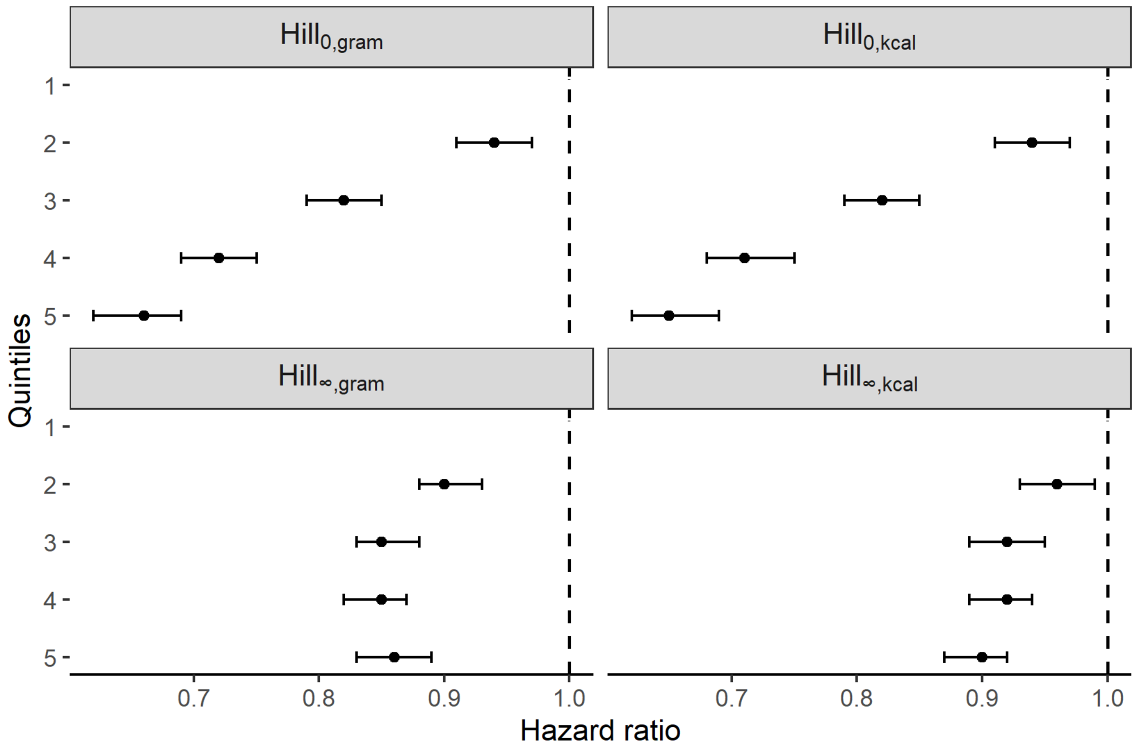

Figure 4.8: Mortality hazard ratios for $Hill_0$ and $Hill_\infty$ using the overall mortality dataset (n=451,390).The horizontal error bars show the 95% confidence intervals.

There is a clear inverse relationship for $Hill_0$: the higher the quintile, the lower the hazard ratio, i.e., the lower the risk of mortality. On the contrary, this trend is less clear for $Hill_\infty$. However, there is still a significant difference between the first quintile and the others.

Apparently, the hazard ratios for $Hill_0$ are lower than for $Hill_\infty$ in the highest quintile, reflecting people with the highest score for food biodiversity. Therefore, if the focus is on reducing the population's probability of dying, $Hill_0$ is the best index according to these results.

The absolute number of deaths per 10,000 person-years without adjustment or stratification for $Hill_{0,gram}$, $Hill_{0,kcal}$, $Hill_{\infty,gram}$, and $Hill_{\infty,kcal}$ is given in the Appendix (Table A2). The outcomes of the first/fifth quintiles are 76/49, 76/50, 79/44, and 63/65 deaths per 10,000 person-years, respectively. The fifth quintile has a reduced mortality compared to the first quintile for all indices, except for $Hill_{\infty,kcal}$, where the outcomes are similar to those for the first quintile.

Interestingly, when considering weight-based indices, conclusions regarding their association with mortality differ between relative hazard ratios and absolute number of deaths. When comparing the first and fifth quintile, it is observed that $Hill_{0,gram}$ exhibits a stronger reduction in hazard ratios compared to $Hill_\infty$. Conversely, $Hill_\infty$ shows a greater reduction in the absolute number of deaths. This discrepancy can be attributed to the corrections applied in the hazard ratio model, such as adjustment for physical activity (Section 3.3.4).

Unfortunately, the absolute number of cases of death per 10,000 person-years is not yet clear, as there are some discrepancies in the results for hazard ratios in this thesis and those in the article by Hanley-Cook et al. (2021). Further research in collaboration with IARC should provide insight into why these analyses do not yield the same results.

4.4 Creating a new index: $Jill_x$

The observed correlation between $Hill_2$ and $Hill_\infty$ is notable, but there is room for improvement. Evidence suggests that individuals are consuming a significant amount of the same species, as reported by Loftas (1995) and Hanley-Cook et al. (2021).

In the small test dataset (n=93), an interesting finding emerged when analysing the dietary composition. Specifically, when the proportions of the 20%

most consumed species (on average 13 species) in an individual's diet are summed up, it was found that these species accounted for a significant portion, approximately 82%, of the energy or weight distribution within the diet.

The pattern, where a relatively small number of species contribute substantially to the overall diet, aligns with the well-known Pareto principle, often referred to as the 80:20 rule. According to this principle, it is hypothesised that approximately 80% of outcomes (e.g. weight or energy in the diet) can be attributed to just 20% of the components (e.g. species). Numerous other examples of the Pareto principle can be found across different scientific disciplines, and here a similar effect in the EPIC dataset has been demonstrated.

Furthermore, the distribution of the ten most abundant species in the diet was investigated using the small test dataset (n=93) (Figure 4.9).

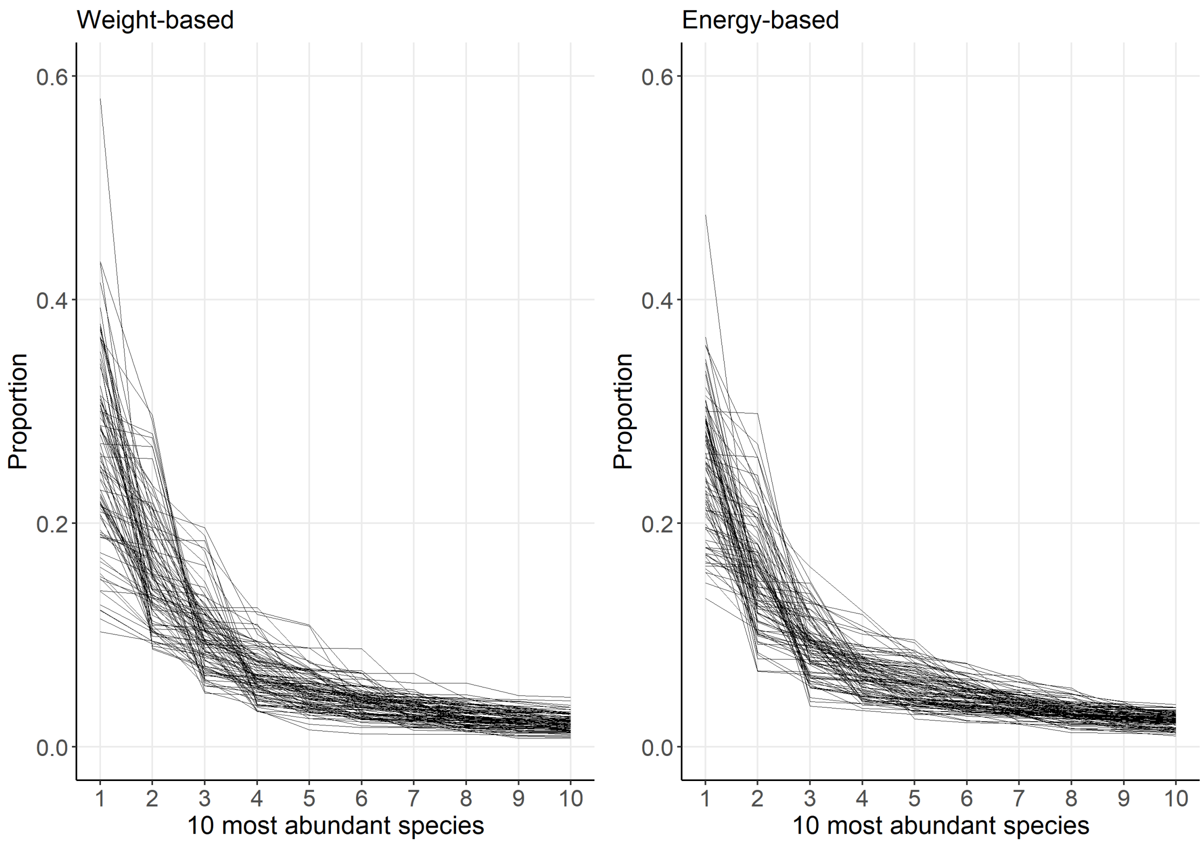

Figure 4.9: Abundance of the ten most abundant species per person using the test dataset (n=93).

According to Figure 4.9, the two or three most abundant species play a major role in the diet, and the other species are present in much smaller proportions (smaller than around 10%). Therefore, the calculation of a new index, $Jill_x$, was tested by extending $Hill_\infty$ with maximum four extra species and considering their average proportion. Specifically, $Jill_x$ is equal to the inverse of the average of the x most abundant species in the diet. Following promising results on the test dataset, the $Jill_x$ index was applied to the full dataset (n=476,768). Subsequently, the Pearson correlation coefficient between $Jill_x$ and $Hill_2$ was calculated for x ranging from one (equalling $Hill_\infty$) to five ($Jill_5$) (Figure 4.10). Note that the Spearman correlation coefficients are similar (Appendix A1), indicating a strong monotonic and linear relationship.

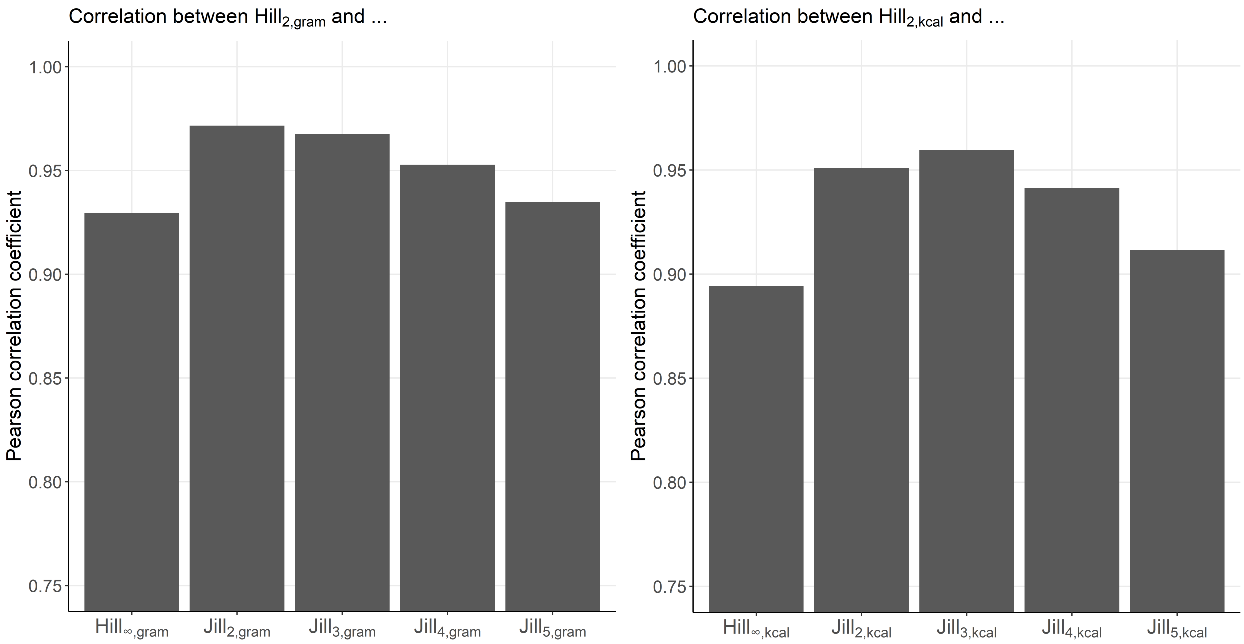

Figure 4.10: Pearson correlation coefficients between $Hill_2$ and $Hill_\infty$ (reference) and between $Hill_2$ and a newly created index $Jill_x$ using the full dataset (n=476,768).

Figure 4.10 shows that the correlation coefficients of the values for $Hill_2$ with the outputs of both $Jill_{2,gram}$ and $Jill_{3,kcal}$ are the highest. So, if the goal is to align the index output with $Hill_2$, the inverse of the two (weight-based) or three (energy-based) most abundant species is the best option. However, a consensus based on other research will be needed on the value of x, so that the indices between different studies can be compared.

In conclusion, the application of the Pareto principle, along with the observation that two or three species contribute significantly to the overall diet, explain why $Hill_\infty$ and $Jill_x$ can provide good estimates of food biodiversity even though they only take the few most abundant species as input.

4.5 Impact of selected dietary unit

To perform calculations for most food biodiversity indices, a unit for food data must be selected, such as weight or energy. In the present study, $Hill_0$ is the exception, as it is not based on the proportions of food in the diet.

Figure 4.11 compares the indices according to the different units used for the calculation. The Pearson correlation coefficients between weight-based and energy-based food biodiversity scores for $Hill_0$, $Hill_1$, $Hill_2$ and $Hill_\infty$ are 1.00, 0.42, 0.23, and 0.13, respectively. In this case, the Spearman correlation coefficients are similar to the Pearson correlation coefficients (Appendix Table A1). Therefore, a strong monotonic relationship exists, where there is a strong linear relationship.

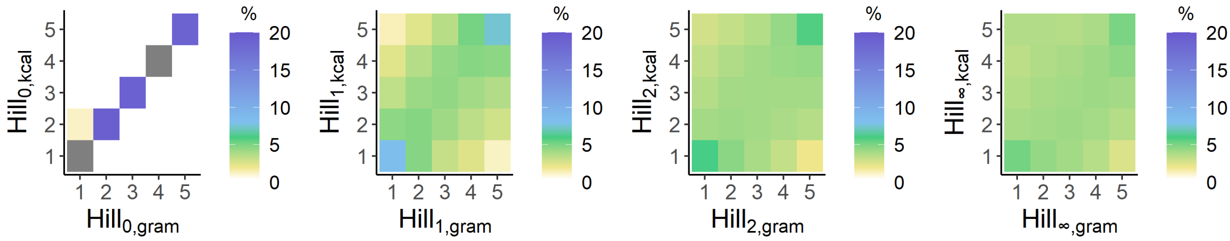

Figure 4.11: Energy-based compared to weight-based Hill numbers, represented by quintiles using the full dataset (n=476,768). The colour code represents the percentage of the whole cohort classified in a given quintile for the index on the x-axis and a given quintile for the index on the y-axis.

The classification in quintiles based on $Hill_{0,gram}$ and $Hill_{0,kcal}$ is nearly identical because the proportions of species are not taken into account. While a perfect correlation is expected, there is a small difference between the values for $Hill_{0,gram}$ and $Hill_{0,kcal}$ due to the data coding. There is one NCLASS food group, which includes water and additives, where the weight is counted but the energy is set to zero. As a result, a significant number of people scored higher for $Hill_{0,gram}$ than for $Hill_{0,kcal}$. However, some individuals may still be classified in a lower quintile for $Hill_{0,gram}$ because the upper boundary of the first quintile is higher for this index (48) compared to $Hill_{0,kcal}$ (47).

In contrast, for the other indices, the choice of unit is found to strongly influence the outcome of the index. This means that some individuals are in the highest quintile for the weight-based data and are in the lowest quintile for the energy-based data, and vice versa. Similarly, from the food supply data used by Khoury et al. (2014), it can be derived that there is a noticeable difference between $Hill_{\infty,gram}$ and $Hill_{\infty,kcal}$.

Consequently, the choice of the unit has a clear impact on the conclusions drawn regarding food biodiversity. Therefore, it is essential for researchers to keep in mind the mismatch between indices based on different units and to select the appropriate unit that aligns with their research objectives.

The standard deviations, as computed for each index, are presented in Table 4.3. Standard deviations for $Hill_0$ are similar, while for the other indices, the weight-based food biodiversity scores tend to have greater variability compared to the energy-based scores. This discrepancy suggests that to achieve an equivalent standard error, a larger sample size will be required for the weight-based food biodiversity indices in comparison to the energy-based indices. This implies differences in data collection protocols will be required, depending on the dietary unit that will be used.

Table 4.3: Standard deviations of the Hill numbers using the full dataset (n=476,768).

Index	Weight-based	Energy-based
$Hill_0$	16.04	16.11
$Hill_1$	5.05	4.21
$Hill_2$	3.39	2.84
$Hill_{\infty,kcal}$	1.47	1.29

4.5.1 Using energy-based data

Energy-based data can be advantageous for studies focused on nutrition since energy intake is regulated by the human body: a person will consume about as much energy as he uses (Hall et al., 2012). Using energy-based data for food biodiversity indices was also proposed by Jones et al. (2021) and used by de Oliviera Otto et al. (2015). However, it is difficult to determine through FFQs the exact amount of energy consumed due to factors such as food variety (Hazo & Yirgalem, 2022; Lasek et al., 2020), ripeness (Phillips et al., 2021), and processing (Barr & Wright, 2010). Moreover, some researchers disagree with the currently used Atwater factors that calculate energy values for foods, as they may not accurately reflect the energy available for the human body. For example, almonds actually have lower energy values than predicted with Atwater factors due to the fact that significant amounts of the macronutrients cannot be digested (Novotny et al., 2012).

4.5.2 Using weight-based data

Weight-based data and energy-based data can diverge significantly from one another. For example, sweet potatoes and cucumbers have almost the same weight, but the energy-density (kcal/100g) of the sweet potatoes is almost eight times higher (Nubel, 2023). To calculate food biodiversity based on consumption, food supply or food production data, researchers often use weight-based data (Borkotoky et al., 2018; Gustafson et al., 2016; Lachat et al., 2018; Remans et al., 2014; Tian et al., 2017). Therefore, using weight-based data makes it easier to compare with such established studies. Furthermore, quantities are easier to measure than the energy content of foods.

In the present study, drink consumption, such as tea and wine, had a significant impact on the calculation of weight-based food biodiversity indices. However, it is important to note that not all drinks necessarily reflect high food biodiversity. While tea is made from only a few tea leaves and does not contribute significantly to environmental diversity, wine is made from grapes and a significant number of grape plants are required to produce it. Therefore, it is important to consider factors such as juice yield and dilution when referring to environmental diversity.

In conclusion, further research is needed on the selection of the right dietary unit in different contexts.

4.6 Challenges of collecting food data

Collecting food data is often challenging and no dietary intake questionnaire is capable of capturing the complete variability of the species consumed. Similarly, the food intake data provided by IARC has some important limitations in quantifying food biodiversity.

4.6.1 Overview of food data in the EPIC study

An overview of the food data is provided in Figures 4.12 and 4.13 by showing the most commonly consumed species and their average contribution to the diets of the persons in the full dataset (n=476,768). It is evident that some species make a significant contribution, while the majority are present in smaller quantities. This pattern was also observed when investigating the people's individual diet (Section 4.4).

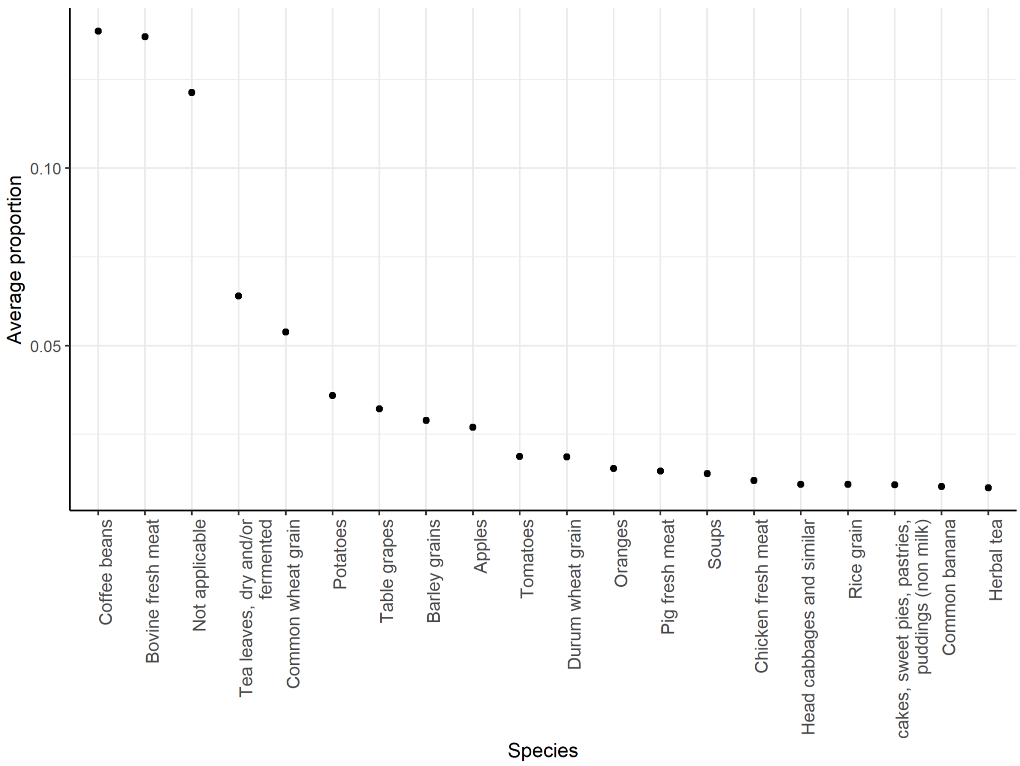

Figure 4.12: Abundance curve of average amount that the twenty most abundant species contributed to the diet of a person using the full dataset (n=476,768).

When examining the abundance graph based on weight (Figure 4.12), the most abundant species are coffee beans, bovine meat, 'not applicable' (i.e., water consumption included), tea leaves, and common wheat. Conversely, for energy-based proportions (Figure 4.13), the most abundant species are bovine meat, wheat, pig meat, and potatoes, as already described by (Hanley-Cook et al., 2021). Notably, the decline in the abundance of species is steeper in the energy-based graph.

It should be noted that foods were classified according to species, which implies that milk and beef belong to the same species (i.e., Bovine fresh meat), despite their nutritional differences. In this way, diversity indices are

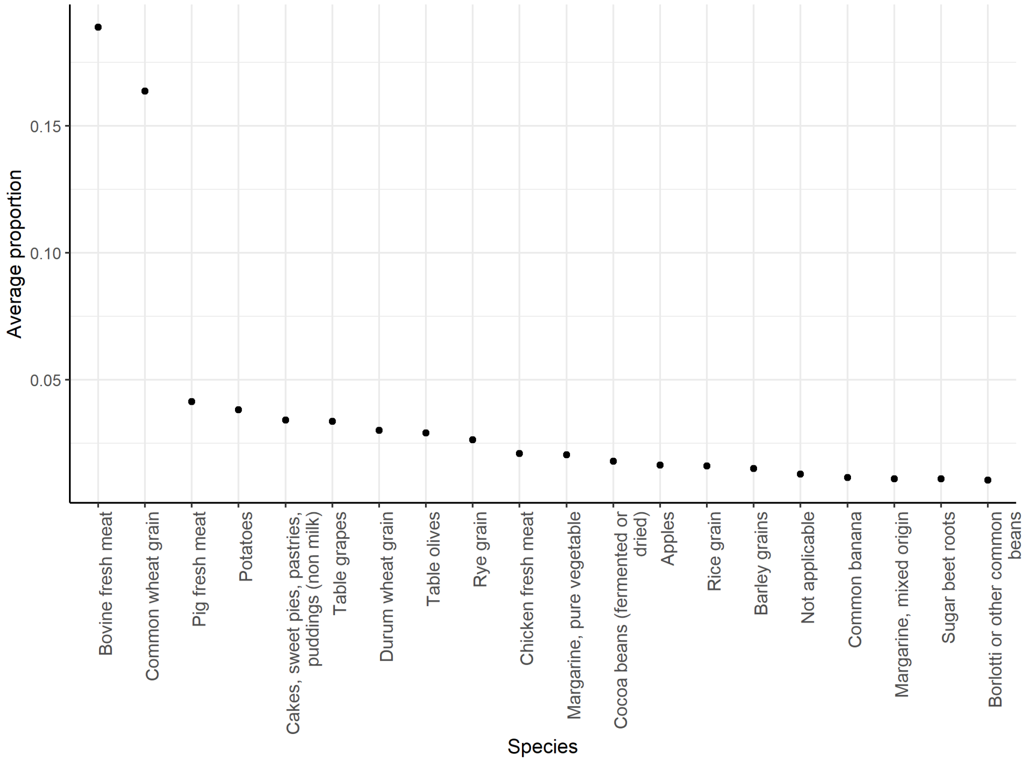

Figure 4.13: Abundance curve of the average energy that the twenty most abundant species contributed to the diet of a person using the full dataset (n=476,768).

calculated through the lens of 'biodiversity conservation'. (Hanley-Cook et al., 2021)

4.6.2 Remarks on food data of EPIC

There are some limitations in the dataset used that affect the accuracy of the calculation of the food biodiversity index. Firstly, the EPIC data are not fully broken down by species and still include NCLASS food groups, such as "pastries, sweet cakes, cakes and puddings". However, these could be further divided into individual ingredients according to standard recipes: flour, sugar, butter, and eggs (i.e., species: wheat, sugar beet, cow, and chicken). Using the small dataset, on average, 20% of the total species consumed, based on energy content, fall into these NCLASS food groups, while this proportion is 11% when calculated based on weight. On the other hand, by decomposition of the recipes, assumptions have to be made: it is unknown whether some ingredients (e.g., herbs) are actually used by the participant, which also has an impact on the accuracy of the calculated indices.

Second, drinks are included in the dataset (e.g., wine consumption represented by 'table grapes'), whereas Jones et al. (2021) removed all alcoholic beverages. Also, non-alcoholic drinks, for instance, coffee, tea and water (represented by 'not applicable') are in the top five most consumed species in the data by weight (Figure 4.12). Therefore, drinks can have a major impact on food biodiversity indices, such as the $Hill_{\infty}$ index, as stated in Section 4.5.

Moreover, the EPIC study collected the data using food frequency questionnaires, meaning a questionnaire was prepared in advance to obtain information on intake. FFQs are widely used in research to record dietary habits due to their ability to capture seasonal changes in food consumption over long periods (Carroll et al., 2012; van der Toorn et al., 2020). However, according to Smith et al. (1991), episodic memory of the diet deteriorates as time passes and people tend to rely on general knowledge about foods to reconstruct their past diet. They also state that these general knowledge-based reconstructions of past diets are likely based on individuals' beliefs, and possibly even their hopes, about their habitual diet. Therefore, FFQs are biased, affecting the accuracy of the data. For example, people tend to overestimate their consumption of healthy foods and underestimate their consumption of unhealthy foods, which is known as the "lean cuisine syndrome" (Rathje & Murphy, 2001). Additionally, people often report consuming less of the foods that they actually eat frequently and vice versa, referred to as the "flattened slope phenomenon" (Thiébaut et al., 2007). These limit the usefulness of FFQs for accurate measurements of dietary intake.

The Commission on Genetic Resources for Food and Agriculture recognises the need for more information on composition and intake, such as those of wild and under-used animal breeds and species, to assess the value of food biodiversity for nutrition and food security (FAO, 2016). Multiple 24-hour recalls can provide more accurate and detailed information (e.g., variety) on the specific products consumed (Carroll et al., 2012). However, the cost in terms of time and money of collecting 24-hour recalls is high, making FFQs preferred for large cohort studies (Kolodziejczyk et al., 2012).

Lastly, the data are outdated: FFQs were conducted around 1990 and consumption patterns have already changed over the years (Dokova et al., 2022). Consequently, obtaining more up-to-date information on food consumption is essential to achieve a more accurate understanding of the current state of food biodiversity.

However, the EPIC data also have clear strengths. With a vast and diverse sample size encompassing multiple European countries, EPIC provides a robust foundation for research and analysis. The validated food frequency questionnaires encompassed the assessment of usual diet intake, which makes them suitable for various studies. Furthermore, the longitudinal design of EPIC allows investigation of dietary factors and their impact on health outcomes over time.

4.6.3 Suggestions for food data collection

Several key considerations are essential to obtain accurate food biodiversity measurements. First, it is important to ensure that the data used are up to date, as the diets of people constantly evolve (Dokova et al., 2022). One useful approach for collecting data from large populations is to use mobile applications to gather multiple 24-hour recalls (Cade, 2017). Alternatively, if FFQs are preferred because they reflect longer time periods, the number of questions can be reduced when using $Hill_\infty$ or $Jill_x$ indices. This is because these indices only require information on the most abundant species, making it sufficient to query the most commonly consumed species in a particular population over a certain period.

Secondly, it is critical to ensure that food intake data are converted to the species level as accurately and as specifically as possible. Emerging technologies such as artificial intelligence can help convert food intake data into species data (Lee et al., 2022). Additionally, it is important to clearly distinguish between alcoholic, non-alcoholic drinks and solid foods in the data collection to ensure accurate analysis.

Third, the number of participants in the study can be reduced to achieve a good estimate of food biodiversity. This will depend on the population being studied, the objectives of the research and the index chosen, as explained by the standard deviations in Section 4.5.

4.7 Ecology versus food

Most biodiversity indices were originally created to quantify biodiversity in ecology. While there are clear overlaps between ecological and food biodiversity, there are also distinct differences. These similarities and dissimilarities will be discussed based on the three components of biodiversity: richness, evenness, and disparity (Section 2.3.1).

4.7.1 Richness

Richness is important both in food and ecology because a higher number of species in a certain area or a diet result in a higher diversity (Hanley-Cook et al., 2021). However, there exist indices in ecology based on richness, such as those correcting for sample size, which cannot be applied to food data. In ecology, larger samples of organisms in a certain area increase the probability of having a higher number of species (a feature known as the sampling effect). Similarly, a hypothesis regarding food data could suggest that higher daily energy or weight intake may be associated with greater richness. This hypothesised relationship is investigated in Figure 4.14 using the full dataset of EPIC.

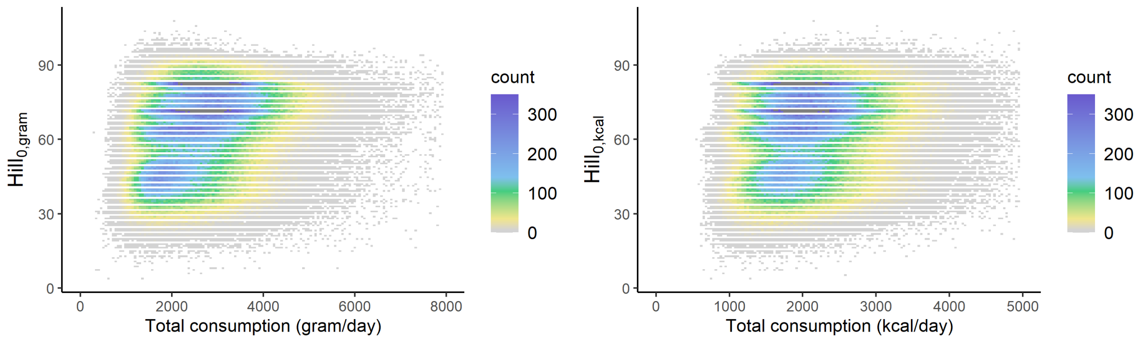

Figure 4.14: $Hill_{0,gram}$ and $Hill_{0,kcal}$ based on total consumption using the full dataset (n=476,768).

However, from these graphs it can be deduced that there is no clear relationship between richness and total food consumed in this case. This can be illustrated with the following example: a low-calorie salad may have high species richness, while a high-calorie snack like chips may have low species richness. So, indices based on this principle, like the Menhinick index (Menhinick, 1964), Margalef index (Margalef, 1957/1973) and Odum, Cantlon, and Kornicker index (Odum et al., 1960), cannot be used for food data.

It should also be noted that there is no clear relationship between $Hill_1$, $Hill_2$, and $Hill_\infty$ and the total consumption (Appendix Figure A8).

4.7.2 Evenness

While evenness is not taken into account in $Hill_0$, it is incorporated into the calculation of the other calculated indices. Focusing on the EPIC data, energy-based abundances lead to higher evenness values than weight-based abundances (Section 4.1.1).

Although evenness is important in the context of ecology, the connection between evenness in a diet and environmental biodiversity is unclear. On one hand, higher evenness in the consumption of worldwide species improves biodiversity in the environment, by ensuring that a few species are not consumed to an excessive degree (Jones et al., 2021). On the other hand, evaluating food biodiversity at the individual level, as done in this thesis, may not accurately reflect agrobiodiversity. For example, a group of ten individuals may have a high food biodiversity score, but their diet may still be predominantly based on the same species. Conversely, a group of ten individuals with a low biodiversity score, but consuming all distinct species, may enhance genetic diversity and resilience.

Furthermore, the relationship between diet evenness and health is unclear. The dietary guidelines recommend different amounts for each food group, such as less meat compared to vegetables (Cámara et al., 2021). Although food group evenness is not recommended, the uniformity of species consumption may be associated with better health outcomes. For instance, since the species diversity of meat (CGRFA, 2015a) is lower compared to fruits and vegetables (Antonelli et al., 2020; Willett et al., 2019), consuming equal amounts of different species may result in higher consumption of fruits and vegetables and align with the recommended food group intake. However, in this thesis, Hill numbers with more emphasis on evenness (i.e., $Hill_2$ and $Hill_\infty$) are associated with higher hazard ratios compared to lower Hill numbers (i.e., $Hill_0$), which focus on richness (Section 4.3.5). Whether this relationship is due to evenness, or some other factor needs further investigation.

4.7.3 Disparity

Lastly, incorporating disparity in a uniform index is challenging in the context of both food and ecology. For example, there are infinite options to compare foods using a combination of different food traits, such as vitamin A content, fibre content, or energy content. This also complicates the comparison between studies. Therefore, the analysis of indices that include disparity was beyond the scope of this thesis (Section 4.8.2).

4.8 Other possible indices

As stated in the literature review, there is a wide variety of options for indices to quantify biodiversity. This thesis focused on the Hill numbers, as the interpretation of the outcome and the calculation method was feasible on the large dataset provided by IARC. In the next paragraphs, other possible indices are discussed and briefly compared to the Hill numbers. Figure 4.15 shows all indices that have been proposed for use with dietary data in the context of this thesis. The reasons why the Margalef index, the Menhinick index and the Odum, Cantlon and Kornicker index were not suitable for calculating food biodiversity are discussed in Section 4.7.2.

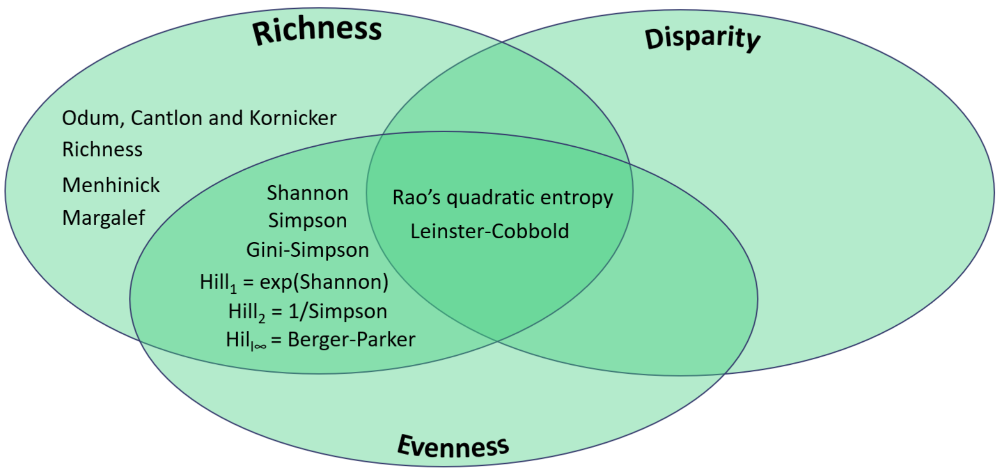

Figure 4.15: All possible indices in this thesis represented by Venn diagram based on the components of biodiversity they incorporate.

4.8.1 Interpretation of the outcome

For a measurement to be used on large scales by people of different backgrounds, the interpretation of the result must be easy. All Hill numbers and the Leinster-Cobbold indices are expressed in effective species, representing the number of different species consumed, if the diet were perfectly uniform (Hill, 1973; Leinster & Cobbold, 2012). Since all Hill numbers have the same unit, they can be easily compared. Even if the goal is to describe the diversity of a single person's diet, equivalent species are still the best diversity measure, according to an ecology forum held by Ellison (2010).

Currently, the Shannon index is the most widely used diversity index in food supply studies (Hanley-Cook et al., 2022). This index was used, for example, by Remans et al. (2014) to measure food production and supply diversity at the national level. However, this index is expressed as nats or bits of information (Shannon, 1948) and nonlinearity makes interpretation and comparison difficult (Daly et al., 2018). The Gini-Simpson index is also often used as a diversity index (Heip et al., 1998), but the problem of nonlinearity is even more pronounced for this index (Daly et al., 2018).

4.8.2 The ratio of difficulty of the index versus useful information

As the goal of this research is to provide a simple index, the complexity of the index is an important parameter. Of all the indices considered as possible options in this thesis, the Leinster-Cobbold index is the most complex. However, unlike most other indices, the Leinster-Cobbold index can reflect disparity, along with richness and evenness (Leinster & Cobbold, 2012).

This means that food items can be compared on the basis of their nutritional traits. On one hand, nutritional diversity, which is an important element in food biodiversity, can be incorporated in the analysis. On the other hand, this implies that this index requires a substantial amount of data and a high computational effort to determine the result, because the proportion and selected nutritional characteristics of all species are required. Furthermore, voluntary selection of characteristics and the calculation method to compare food products is a second bottleneck, because they can differ significantly between studies (Green et al., 2021; Wang et al., 2021). Consequently, a standardised approach is lacking, and comparisons between studies are difficult to establish.

4.9 Feasibility in practice

In this section, the feasibility of using $Hill_\infty$ and $Jill_x$ indices in practical applications will be discussed, in particular several important decision points that researchers must consider in the context of their own specific study.

First, while the calculations and data processing of the indices themselves are straightforward and do not require specialised software, there is one contextual choice that the researcher needs to make: the value of x for the $Jill_x$ index. Further research is needed to determine the optimal value of x for different data and study settings. Based on the specific findings of this thesis, it is recommended to use $x = 2$ for weight-based data and $x = 3$ for energy-based data.

Second, the process of collecting dietary intake data and converting it into a species-based format can pose significant design challenges. Quantitative food frequency questionnaires are commonly used to assess dietary intake. Several factors need to be determined when designing these questionnaires, including the choices of the dietary unit, the time frame, the formulation of questions, and the number of questions. For example, the choice of dietary unit, such as grams or kilocalories, can be tailored to the specific research objectives. In addition, selecting a suitable time frame is crucial, as a longer duration (e.g., 1 year versus 1 week) increases the likelihood of obtaining a higher food biodiversity index but also increases the chances of biases in the responses (Smith et al., 1991).

Third, to determine the value of $Hill_\infty$ or $Jill_x$, it is important to focus on the most frequently consumed species. However, it should be noted that the most consumed species may not always correspond directly to the most consumed *foods*, due to processing. For example, wheat grain is used in cookies, bread, pasta, and various other food products. The focus on the most frequently consumed species allows for a limited number of questions based only on these particular species, compared to the number of questions needed to determine all species that were consumed, as other indices require. For the formulation of the questions, it is necessary to identify the species that are most commonly consumed in a particular region, which can be obtained through literature research, questionnaires, or analysis of food supply data.

Finally, proportions, which are required to calculate the indices, need to be estimated using information on a person's total energy or weight intake. This estimation can be achieved using average values, such as 2000 kcal for women and 2500 kcal for men, but this average can vary substantially from the true value for different individuals (Lam & Ravussin, 2017; Schoeller, 1995).

4.10 Reflection on the sustainability of increasing food biodiversity in Europe

Although promoting food biodiversity can be beneficial for human health and the planet, this also influences the entire food chain, as it may require other farming methods and processing steps. In this section, some critical remarks are made on the sustainability of improving food biodiversity scores of people's diet in terms of their impact on people, planet, and profit.

4.10.1 People

This thesis revealed an inverse correlation between high values of various food biodiversity indices and overall mortality, indicating the significance of improved food biodiversity for human health (Section 4.3.5). In addition to health, food biodiversity is also important from a cultural and gastronomical perspective (Spence et al., 2017). Increasing consumer awareness of the benefits of food biodiversity can contribute to higher food biodiversity scores (FAO, 2016). However, food choices are also influenced by other factors such as market availability and individual preferences for certain foods (Bioversity International, 2017). Food preferences play a decisive role, as individuals tend to consume what they like the most and avoid what they dislike (Ahern et al., 2013). Furthermore, market availability affects food biodiversity, as limited diversity in stores reduces the chances of consuming a wide range of species (Kalaitzis et al., 2007, April 23–25; Siegel et al., 2014).

4.10.2 Planet

Considering market availability, certain regions may not be suitable for cultivating specific crops due to soil nutrients (Dhakal & Lange, 2021) and temperature (Sheehy et al., 2006). In such cases, importing food crops becomes a solution to enhance food biodiversity. Unfortunately, increased importation has an adverse impact on greenhouse gas emissions (Neira et al., 2016). When promoting food biodiversity, other crucial factors to consider are avoiding either deforestation (FAO, 2022) or excessive pressure on certain species, such as overfished fish (IPBES, 2019). On the other hand, emphasising forgotten and underutilised species like bulbous chervil (*Chaerophyllum bulbosum*) can be a viable solution (Antonelli et al., 2020). In conclusion, while food biodiversity improves food security and resilience to climate change, it should be implemented with caution to prevent problem shifting.

4.10.3 Profit

Improving food biodiversity requires making different foods affordable for consumers and viable for producers. Limited affordability of certain species negatively impacts biodiversity scores, as demonstrated by the focus on energy-rich crops during the Green Revolution, which widened the nutrition gap by making them more affordable while leaving others unattainable (Ramankutty et al., 2018). Similarly, exclusive delicacies like truffles are only accessible to the wealthy, posing a challenge in diversifying affordable food options.

Currently, profit-driven investments tend to favour specific, non-diverse value chains. However, diverse food systems are crucial for risk diversification, mitigating potential hazards like infections and conflicts. Promoting the inclusion of other species to enhance food biodiversity has wide-ranging impacts on the entire food chain and the individuals involved in its production. Therefore, it is crucial to emphasise fair trade practices and provide financial support to facilitate this transition.

To conclude, increasing food biodiversity has significant effects on people, the planet, and profit. However, it is essential to conduct further research to comprehensively understand and document all the advantages and disadvantages of this transition in practice. It is important to note that there are important trade-offs between these dimensions, and a thorough risk-benefit analysis is necessary to make informed decisions and ensure the sustainability of such initiatives.

5. SUGGESTIONS FOR FURTHER RESEARCH

This thesis provides a first insight into quantifying food biodiversity, but more research is needed. Conducting more research would allow the results to be compared between studies in various regions.

Therefore, more advanced technology could improve the collection of dietary information for a large number of people, but to enjoy all the benefits of emerging technology, high-quality, verified systems will be required (Cade, 2017). A study by Lee et al. (2022) created an artificial intelligence model that allows users to enter the name of a meal and serving size to assess the nutrients consumed. They also highlighted that this model may be used in all nations and circumstances, thanks to the user's capacity to modify the recipes. Similarly, an artificial intelligence model can be developed to assess information on the species consumed.

As mentioned above, further research is also needed on the indices themselves, for example:

- Is there a clear correlation between $Hill_{\infty}$ and $Hill_2$ in other studies (e.g., low-income countries, more recent data) based on FFQs, dietary recall data or food production data?
- Is there a difference between food biodiversity indices based on datasets with foods and drinks treated as separate groups, compared to the ones where food and drinks are combined? What is the best option to take?
- What is the actual relationship to health? This must be further explored, as there is discordance between the results presented in this thesis and those presented in the paper of Hanley-Cook et al. (2021), although the same data are used.
- What is the relationship between the Hill numbers and other indices, such as the Minimum Dietary Diversity for Women (FAO & FHI 360,

2016), the Mediterranean diet score (Buckland et al., 2009), the Healthy Eating Index, and the Dietary Approaches to Stop Hypertension score (de Oliviera Otto et al., 2015)?

- What is the link between food biodiversity and other environmental aspects, such as biodiversity, greenhouse gas emissions, and water and land use?
- Is there an association between food biodiversity and socio-economic status?

Food biodiversity is an interesting parameter as it reflects the diversity of the species consumed in a specific region, which may differ from the species produced locally. For example, pears are exported from Belgium and bananas are imported. Nicholson et al. (2021) state that Europe exhibits the largest disparity between crop imports and production compared to other regions. However, according to a meta-analysis conducted by Powell et al. (2015), low-income countries often demonstrate a positive association between dietary diversity and agrobiodiversity. Consequently, comparing food biodiversity and agrobiodiversity in a certain region can give interesting results.

Finally, the absolute values of a food biodiversity index still cannot be interpreted in isolation. More research is needed to establish reference values for the food biodiversity indices that can define what is considered 'good' or 'bad' in terms of a healthy sustainable diet. In this thesis, the indices can only be compared, which means that people can be grouped according to their food biodiversity score. In the future, it will be necessary to know which value has to be reached, for example, on a country level, to have a sufficient level of biodiversity in order to achieve a specific goal or purpose. Jones et al. (2021) has already attempted to assess food biodiversity using the Shannon index. Reference values would be helpful to governments in implementing measures to improve biodiversity.

6. CONCLUSION

6.1 Summary

Food biodiversity plays a crucial role in addressing current and future food security challenges by meeting nutrient requirements and increasing resilience to environmental changes. However, global policy agendas do not yet prioritise food biodiversity. To set clear targets therein, food biodiversity needs to be easily quantified, and research on these methods is still limited.

In this study, the potential of using Hill numbers, derived from ecological diversity indices, to quantify food biodiversity was investigated based on food intake data of 476,768 people from nine different European countries. Of the four Hill numbers examined ($Hill_0$, $Hill_1$, $Hill_2$, and $Hill_\infty$), $Hill_1$ and $Hill_2$ theoretically emerged as the most appropriate indices because they take into account the proportion of all species in an individual's diet. Unlike $Hill_1$ and $Hill_2$, the $Hill_\infty$ index requires only one data point: the proportion of the most consumed species in the diet. Due to a strong correlation between $Hill_\infty$ and $Hill_2$ in the present study (Pearson correlation coefficient of 0.93 for $Hill_{\infty,gram}$ and 0.89 for $Hill_{\infty,kcal}$), $Hill_\infty$ may be an equivalent indicator as $Hill_2$ measuring food biodiversity.

Furthermore, no significant variations were observed in the values of $Hill_\infty$ across different European countries, age groups, genders, body mass index classes, and quantities consumed. This implies that, for example, comparisons of $Hill_\infty$ values between individuals of different weight classes can be made without the need to account for gender differences in the overall population analysis. A notable, inverse relationship was observed between $Hill_\infty$ and hazard ratios for total mortality stratified by gender, age, and study centre and adjusted for smoking status, educational level, marital status, physical activity, alcohol intake, and total energy intake, Mediterranean diet score, red and processed meat intake, and fibre intake. When individuals were divided into quintiles, the relative risk was reduced to 0.86 for $Hill_{\infty,gram}$ and 0.9 for $Hill_{\infty,kcal}$ in the fifth quintile compared to the first quintile. In the model without corrections, it resulted in an absolute overall mortality of 79/44 for $Hill_{\infty,gram}$, and 63/65 for $Hill_{\infty,kcal}$ deaths per 10,000 person-years in the first/fifth quintile.

In addition, this study examined the relationships between energy-based and weight-based indices, highlighting the influence of the chosen unit of measurement on outcomes. Energy-based indices are suitable for assessing nutritional aspects, whereas weight-based data is generally easier to obtain compared to energy-based data. Overall, it is essential for researchers to keep in mind the mismatch between indices based on different units and to select the appropriate unit that aligns with their research objectives.

To address the limitation of relying solely on the proportion of one species in the diet, a new indicator, $Jill_x$, was proposed. $Jill_x$ represents the inverse of the mean abundances of the x most commonly consumed species and this showed an even stronger correlation with $Hill_2$, specifically a Pearson correlation coefficient of 0.96 for $Jill_{2,gram}$ and 0.96 for $Jill_{3,kcal}$.

6.2 Limitations

The study encountered several limitations. Firstly, the dietary intake data were collected using a single food frequency questionnaire that covered a one-year period. While this approach takes into account seasonal variations in dietary habits, it may not capture long-term changes in dietary patterns that can occur over several years. Additionally, self-reported dietary intake is prone to biases.

Secondly, the data was not fully categorised at the species-level, and the impact of subspecies on food biodiversity could not be analysed with the available data.

Thirdly, due to data protection regulations, sharing the data beyond IARC was restricted, resulting in limited analysis performed by the data scientists of IARC, based on the scripts I provided. Moreover, discrepancies with the findings of (Hanley-Cook et al., 2021) raise the need for further collaborative research with IARC to clarify associations with mortality. Additionally, since the study participants exhibited generally more health-conscious behaviours, the obtained values and their relationship with overall mortality may not be representative of the entire European population.

Furthermore, it should be recognised that neither $Hill_\infty$ nor $Jill_x$ can capture the full complexity of food diversity. Nevertheless, these are practical and simple indices that can be used to good effect. For example, they can provide an estimate of an individual's food biodiversity score based on a food frequency questionnaire with a limited number of questions on the most consumed species in a given region. Future research should then focus on determining the optimal number of questions and further checking and validating the choice of index, and its behaviour across different datasets.

6.3 Overall conclusion

Overall, this study contributes to understanding and providing a framework for assessing food biodiversity scores. By including $Hill_\infty$ or $Jill_x$ in research initiatives and later in policy agendas, we can work to promote sustainable food systems and safeguard the well-being of current and future generations.

REFERENCES

Ahern, S. M., Caton, S. J., Bouhlal, S., Hausner, H., Olsen, A., Nicklaus, S., Møller, P., & Hetherington, M. M. (2013). Eating a rainbow. introducing vegetables in the first years of life in 3 European countries. *Appetite*, *71*, 48–56. https://doi.org/10.1016/j.appet.2013.07.005

Aizen, M. A., Aguiar, S., Biesmeijer, J. C., Garibaldi, L. A., Inouye, D. W., Jung, C., Martins, D. J., Medel, R., Morales, C. L., Ngo, H., Pauw, A., Paxton, R. J., Sáez, A., & Seymour, C. L. (2019). Global agricultural productivity is threatened by increasing pollinator dependence without a parallel increase in crop diversification. *Global Change Biology*, *25*(10), 3516–3527. https://doi.org/10.1111/gcb.14736

Antonelli, A., Fry, C., Smith, R., Simmonds, M., Kersey, P., Pritchard, H., Abbo, M., Acedo, C., Adams, J., Ainsworth, A., Allkin, B., Annecke, W., Bachman, S., Bacon, K., Bárrios, S., Barstow, C., Battison, A., Bell, E., Bensusan, K., . . . Zhang, B. (2020). *State of the world's plants and fungi 2020*. Kew, Royal Botanic Gardens. https://doi.org/10.34885/172

Aremu, S. O., & Nweze, C. C. (2017). Determination of vitamin A content from selected Nigerian fruits using spectrophotometric method. *Bangladesh Journal of Scientific and Industrial Research*, *52*(2), 153–158. https://doi.org/10.3329/bjsir.v52i2.32940

Arimond, M., Wiesmann, D., Becquey, E., Carriquiry, A., Daniels, M. C., Deitchler, M., Fanou-Fogny, N., Joseph, M. L., Kennedy, G., Martin-Prevel, Y., & Torheim, L. E. (2010). Simple food group diversity indicators predict micronutrient adequacy of women's diets in 5 diverse, resource-poor settings. *The Journal of Nutrition*, *140*(11, Suppl.), S2059–S2069. https://doi.org/10.3945/jn.110.123414

Barnosky, A. D., Hadly, E. A., Bascompte, J., Berlow, E. L., Brown, J. H., Fortelius, M., Getz, W. M., Harte, J., Hastings, A., Marquet, P. A., Martinez, N. D., Mooers, A., Roopnarine, P., Vermeij, G., Williams, J. W., Gillespie, R., Kitzes, J., Marshall, C., Matzke, N., . . . Smith, A. B. (2012). Approaching a state shift in Earth's biosphere. *Nature*, *486*(7401), 52–58. https://doi.org/10.1038/nature11018

Barr, S. B., & Wright, J. C. (2010). Postprandial energy expenditure in whole-food and processed-food meals: Implications for daily energy expenditure. *Food & Nutrition Research*, *54*(1), Article 5144. https://doi.org/10.3402/fnr.v54i0.5144

Bezerra, I. N., & Sichieri, R. (2011). Household food diversity and nutritional status among adults in Brazil. *The International Journal of Behavioral Nutrition and Physical Activity*, *8*(1), Article 22. https://doi.org/10.1186/1479-5868-8-22

Bilalis, D., Papastylianou, P., Konstantas, A., Patsiali, S., Karkanis, A., & Efthimiadou, A. (2010). Weed-suppressive effects of maize–legume intercropping in organic farming. *International Journal of Pest Management*, *56*(2), 173–181. https://doi.org/10.1080/09670870903304471

Bioversity International. (2017). *Mainstreaming agrobiodiversity in sustainable food systems: Scientific foundations for an agrobiodiversity index*. https://hdl.handle.net/10568/89049

Borkotoky, K., Unisa, S., & Gupta, A. K. (2018). State-level dietary diversity as a contextual determinant of nutritional status of children in India: A multilevel approach. *Journal of Biosocial Science*, *50*(1), 26–52. https://doi.org/10.1017/S0021932017000013

Buckland, G., Gonzalez, C. A., Agudo, A., Vilardell, M., Berenguer, A., Amiano, P., Ardanaz, E., Arriola, L., Barricarte, A., Basterretxea, M., Chirlaque, M. D., Cirera, L., Dorronsoro, M., Egues, N., Huerta, J. M., Larranaga, N., Marin, P., Martinez, C., Molina, E., . . . Moreno-Iribas, C. (2009). Adherence to the Mediterranean diet and risk of coronary heart disease in the Spanish EPIC cohort study. *American Journal of Epidemiology*, *170*(12), 1518–1529. https://doi.org/10.1093/aje/kwp282

Burlingame, B., Charrondiere, R., & Mouille, B. (2009). Food composition is fundamental to the cross-cutting initiative on biodiversity for food and nutrition. *Journal of Food Composition and Analysis*, *22*(5), 361–365. https://doi.org/10.1016/j.jfca.2009.05.003

Cade, J. E. (2017). Measuring diet in the 21st century: Use of new technologies. *Proceedings of the Nutrition Society*, *76*(3), 276–282. https://doi.org/10.1017/S0029665116002883

Cámara, M., Giner, R. M., González-Fandos, E., López-García, E., Mañes, J., Portillo, M. P., Rafecas, M., Domínguez, L., & Martínez, J. A. (2021). Food-based dietary guidelines around the world: A comparative analysis to update AESAN Scientific Committee dietary recommendations. *Nutrients*, *13*(9), Article 3131. https://doi.org/10.3390/nu13093131

Carroll, R. J., Midthune, D., Subar, A. F., Shumakovich, M., Freedman, L. S., Thompson, F. E., & Kipnis, V. (2012). Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *American Journal of Epidemiology*, *175*(4), 340–347. https://doi.org/10.1093/aje/kwr317

Commission on Genetic Resources for Food and Agriculture. (2015a). *The second report on the state of the world's animal genetic resources for food and agriculture*. Food & Agriculture Organization of the United Nations. https://www.fao.org/3/i4787e/i4787e.pdf

Commission on Genetic Resources for Food and Agriculture. (2015b). *The second report on the state of the world's animal genetic resoures for food and agriculture - in brief*. Food & Agriculture Organization of the United Nations. https://www.fao.org/3/a1260e/a1260e.pdf

Convention on Biological Diversity. (2006). *Convention text - Article 2: Use of terms*. https://www.cbd.int/convention/text/

Daly, A. J., Baetens, J. M., & De Baets, B. (2018). Ecological diversity: Measuring the unmeasurable. *Mathematics*, *6*(7), Article 119. https://doi.org/10.3390/math6070119

de Oliviera Otto, M. C. d. O., Padhye, N. S., Bertoni, A. G., Jr, D. R. J., & Mozaffarian, D. (2015). Everything in moderation - dietary Diversity and quality, central obesity and risk of diabetes. *PLOS ONE*, *10*(10), Article e0141341. https://doi.org/10.1371/journal.pone.0141341

Declerck, F. A. J., Fanzo, J., Palm, C., & Remans, R. (2011). Ecological approaches to human nutrition. *Food and Nutrition Bulletin*, *32*(1, Suppl.), S41–S50. https://doi.org/10.1177/15648265110321S106

Dhakal, C., & Lange, K. (2021). Crop yield response functions in nutrient application: A review. *Agronomy Journal*, *113*(6), 5222–5234. https://doi.org/10.1002/agj2.20863

Dokova, K. G., Pancheva, R. Z., Usheva, N. V., Haralanova, G. A., Nikolova, S. P., Kostadinova, T. I., Egea Rodrigues, C., Singh, J., Illner, A.-K., & Aleksandrova, K. (2022). Nutrition transition in Europe: East-West dimensions in the last 30 years - A narrative review. *Frontiers in Nutrition*, *9*. https://doi.org/10.3389/fnut.2022.919112

EAT-Lancet Commission. (2020). *Diets for a better future*. https://eatforum.org/knowledge/diets-for-a-better-future/

Ellison, A. M. (2010). Partitioning diversity. *Ecology*, *91*(7), 1962–1963. https://doi.org/10.1890/09-1692.1

Ferrari, P., Day, N. E., Boshuizen, H. C., Roddam, A., Hoffmann, K., Thiébaut, A., Pera, G., Overvad, K., Lund, E., Trichopoulou, A., Tumino, R., Gullberg, B., Norat, T., Slimani, N., Kaaks, R., & Riboli, E. (2008). The evaluation of the diet/disease relation in the EPIC study: Considerations for the calibration and the disease models. *International Journal of Epidemiology*, *37*(2), 368–378. https://doi.org/10.1093/ije/dym242

Food & Agriculture Organization of the United Nations. (2016). *Voluntary guidelines for mainstreaming biodiversity into policies, programmes and national and regional plans of action on nutrition*. Commission on Genetic Resources for Food and Agriculture. https://www.fao.org/3/i5248e/i5248e.pdf

Food & Agriculture Organization of the United Nations. (2020). *The state of world fisheries and aquaculture 2020. Sustainability in action*. https://www.fao.org/3/ca9229en/ca9229en.pdf

Food & Agriculture Organization of the United Nations. (2022). *Fra 2020 remote sensing survey*. https://doi.org/10.4060/cb9970en

Food & Agriculture Organization of the United Nations, & Bioversity International. (2017). *Guidelines on assessing biodiverse foods in dietary intake surveys*. https://www.fao.org/3/i6717e/i6717e.pdf

Food & Agriculture Organization of the United Nations, & FHI 360. (2016). *Minimum dietary diversity for women - A guide to measurement*. https://www.fao.org/3/i5486e/i5486e.pdf

Food & Agriculture Organization of the United Nations, International Fund for Agricultural Development, United Nations Children's Fund, World Food Programme, & World Health Organization. (2022). *The state of food security and nutrition in the world 2022*. Food & Agriculture Organization of the United Nations. https://doi.org/10.4060/cc0639en

Food & Agriculture Organization of the United Nations, & World Health Organization. (2019). *Sustainable healthy diets: Guiding principles*. https://www.fao.org/3/ca6640en/ca6640en.pdf

General Assembly of the United Nations. (2015). A/RES/70/1 - Transforming our world: The 2030 Agenda for Sustainable Development. https://undocs.org/en/A/RES/70/1

Global Panel on Agriculture and Food Systems for Nutrition. (2020). *Future food systems: For people, our planet, and prosperity*. https://www.glopan.org/wp-content/uploads/2020/09/Foresight-2.0_Future-Food-Systems_For-people-our-planet-and-prosperity.pdf

Gómez, G., Nogueira Previdelli, Á., Fisberg, R. M., Kovalskys, I., Fisberg, M., Herrera-Cuenca, M., Cortés Sanabria, L. Y., Yépez García, M. C., Rigotti, A., Liria-Domínguez, M. R., Guajardo, V., Quesada, D., Murillo, A. G., & Brenes, J. C. (2020). Dietary diversity and micronutrients adequacy in women of childbearing age: Results from ELANS study. *Nutrients*, *12*(7), Article 1994. https://doi.org/10.3390/nu12071994

Green, A., Nemecek, T., Smetana, S., & Mathys, A. (2021). Reconciling regionally-explicit nutritional needs with environmental protection by means of nutritional life cycle assessment. *Journal of Cleaner Production*, *312*, Article 127696. https://doi.org/10.1016/j.jclepro.2021.127696

Gustafson, D., Gutman, A., Leet, W., Drewnowski, A., Fanzo, J., & Ingram, J. (2016). Seven food system metrics of sustainable nutrition security. *Sustainability*, *8*(3), Article 196. https://doi.org/10.3390/su8030196

Haftenberger, M., Lahmann, P. H., Panico, S., Gonzalez, C. A., Seidell, J. C., Boeing, H., Giurdanella, M. C., Krogh, V., Bueno-de-Mesquita, H. B., Peeters, P. H. M., Skeie, G., Hjartåker, A., Rodriguez, M., Quirós, J. R., Berglund, G., Janlert, U., Khaw, K. T., Spencer, E. A., Overvad, K., . . . Slimani, N. (2002). Overweight, obesity and fat distribution in 50- to 64-year-old participants in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Public Health Nutrition*, *5*(6b), 1147–1162. https://doi.org/10.1079/PHN2002396

Hall, K. D., Heymsfield, S. B., Kemnitz, J. W., Klein, S., Schoeller, D. A., & Speakman, J. R. (2012). Energy balance and its components: Implications for body weight regulation. *The American Journal of Clinical Nutrition*, *95*(4), 989–994. https://doi.org/10.3945/ajcn.112.036350

Hanley-Cook, G. T., Daly, A. J., Remans, R., Jones, A. D., Murray, K. A., Huybrechts, I., De Baets, B., & Lachat, C. (2022). Food biodiversity: Quantifying the unquantifiable in human diets. *Critical Reviews in Food Science and Nutrition*, 1–15. https://doi.org/10.1080/10408398.2022.2051163

Hanley-Cook, G. T., Huybrechts, I., Biessy, C., Remans, R., Kennedy, G., Deschasaux-Tanguy, M., Murray, K. A., Touvier, M., Skeie, G., Kesse-Guyot, E., Argaw, A., Casagrande, C., Nicolas, G., Vineis, P., Millett, C. J., Weiderpass, E., Ferrari, P., Dahm, C. C., Bueno-de-Mesquita, H. B., . . . Lachat, C. (2021). Food biodiversity and total and cause-specific mortality in 9 European countries: An analysis of a prospective cohort study. *PLOS Medicine*, *18*(10), Article e1003834. https://doi.org/10.1371/journal.pmed.1003834

Hannesson, R. (2007). Global warming and fish migrations. *Natural Resource Modeling*, *20*(2), 301–319. https://doi.org/10.1111/j.1939-7445.2007.tb00209.x

Hazo, H., & Yirgalem, A. (2022). Comparison of different sweet potato (Ipomoea batatas L.) varieties in terms of nutritional value. *Journal of*

Nutrition and Food Sciences, *11*(6), Article 1000047. https://www.longdom.org/open-access-pdfs/comparison-of-different-sweet-potato-ipomoea-batatas-l-varieties-in-terms-of-nutritional-value.pdf

Heip, C., Herman, P., & Soetaert, K. (1998). Indices of diversity and evenness. *Océanis*, *24*(4), 61–87. https://www.vliz.be/imisdocs/publications/221019.pdf

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. (2019). *The global assessment report on biodiversity and ecosystem services: Summary for policymakers.* https://doi.org/10.5281/zenodo.3553579

International Agency for Research on Cancer. (2023). EPIC - European Prospective Investigation into Cancer and Nutrition. https://epic.iarc.fr/

Jones, S. K., Estrada-Carmona, N., Juventia, S. D., Dulloo, M. E., Laporte, M.-A., Villani, C., & Remans, R. (2021). Agrobiodiversity Index scores show agrobiodiversity is underutilized in national food systems. *Nature Food*, *2*(9), 712–723. https://doi.org/10.1038/s43016-021-00344-3

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kalaitzis, P., van Dijk, G., & Baourakis, G. (2007, April 23–25). *Euro-Mediterranean supply chain developments and trends in trade structures, in the fresh fruit and vegetable sector* [Paper presentation]. European Association of Agricultural Economists 103rd Seminar, Barcelo- na, Spain. https://doi.org/10.22004/ag.econ.9428

Kant, A. K., Schatzkin, A., Harris, T. B., Ziegler, R. G., & Block, G. (1993). Dietary diversity and subsequent mortality in the First National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *The American Journal of Clinical Nutrition*, *57*(3), 434–440. https://doi.org/10.1093/ajcn/57.3.434

Katanoda, K., Kim, H.-S., & Matsumura, Y. (2006). New Quantitative Index for Dietary Diversity (QUANTIDD) and its annual changes in the Japanese. *Nutrition*, *22*(3), 283–287. https://doi.org/10.1016/j.nut.2005.06.014

Kennedy, G. L., Pedro, M. R., Seghieri, C., Nantel, G., & Brouwer, I. (2007). Dietary diversity score is a useful indicator of micronutrient intake in non-breast-feeding Filipino children. *The Journal of Nutrition*, *137*(2), 472–477. https://doi.org/10.1093/jn/137.2.472

Khoury, C. K., Bjorkman, A. D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., Rieseberg, L. H., & Struik, P. C. (2014). Increasing homogeneity in global food supplies and the implications for food security. *Proceedings of the National Academy of Sciences*, *111*(11), 4001–4006. https://doi.org/10.1073/pnas.1313490111

Kolodziejczyk, J. K., Merchant, G., & Norman, G. J. (2012). Reliability and validity of child/adolescent food frequency questionnaires that assess foods and/or food groups. *Journal of Pediatric Gastroenterology and Nutrition*, *55*(1), 4–13. https://doi.org/10.1097/MPG.0b013e318251550e

Lachat, C., Raneri, J. E., Smith, K. W., Kolsteren, P., Van Damme, P., Verzelen, K., Penafiel, D., Vanhove, W., Kennedy, G., Hunter, D., Odhiambo, F. O., Ntandou-Bouzitou, G., De Baets, B., Ratnasekera, D., Ky, H. T., Remans, R., & Termote, C. (2018). Dietary species richness as a measure of food biodiversity and nutritional quality of diets. *Proceedings of the National Academy of Sciences*, *115*(1), 127–132. https://doi.org/10.1073/pnas.1709194115

Lam, Y. Y., & Ravussin, E. (2017). Variations in energy intake: It is more complicated than we think. *The American Journal of Clinical Nutrition*, *106*(5), 1169–1170. https://doi.org/10.3945/ajcn.117.167742

Lasek, O., Barteczko, J., Barć, J., & Micek, P. (2020). Nutrient content of different wheat and maize varieties and their impact on metabolizable energy content and nitrogen utilization by broilers. *Animals*, *10*(5), Article 907. https://doi.org/10.3390/ani10050907

Lee, H.-A., Huang, T.-T., Yen, L.-H., Wu, P.-H., Chen, K.-W., Kung, H.-H., Liu, C.-Y., & Hsu, C.-Y. (2022). Precision nutrient management using artificial intelligence based on digital data collection framework. *Applied Sciences*, *12*(9), Article 4167. https://doi.org/10.3390/app12094167

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: The importance of species similarity. *Ecology*, *93*(3), 477–489. https://doi.org/10.1890/10-2402.1

Li, D. (2018). hillR: Taxonomic, functional, and phylogenetic diversity and similarity through Hill Numbers [Computer software]. *Journal of Open Source Software*, *3*(31), Article 1041. https://doi.org/10.21105/joss.01041

Loftas, T. (Ed.). (1995). *Dimensions of need*. Food & Agriculture Organization of the United Nations. https://www.fao.org/3/U8480E/U8480E00.htm

Lynch, S. R., & Cook, J. D. (1980). Interaction of vitamin C and iron. *Annals of the New York Academy of Sciences*, *355*(1), 32–44. https://doi.org/10.1111/j.1749-6632.1980.tb21325.x

Margalef, R. (1973). Information theory in ecology (Translation Bureau (NJB) Foreign Languages Division, Department of the Secretary of State of Canada, Trans.). *Traslation Series*, *2609*, 1–140. https://digital.csic.es/bitstream/10261/284346/1/Margalef_1973.pdf (Original work published in 1957)

Menhinick, E. F. (1964). A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, *45*(4), 859–861. https://doi.org/10.2307/1934933

Moreno, C., & Rodríguez, P. (2010). A consistent terminology for quantifying species diversity? *Oecologia*, *163*(2), 279–282. https://doi.org/10.1007/s00442-010-1591-7

Neira, D. P., Fernández, X. S., Rodríguez, D. C., Montiel, M. S., & Cabeza, M. D. (2016). Analysis of the transport of imported food in Spain and its contribution to global warming. *Renewable Agriculture and Food Systems*, *31*(1), 37–48. https://doi.org/10.1017/S1742170514000428

Nicholson, C. C., Emery, B. F., & Niles, M. T. (2021). Global relationships between crop diversity and nutritional stability. *Nature Communications*, *12*, Article 5310. https://doi.org/10.1038/s41467-021-25615-2

Novotny, J. A., Gebauer, S. K., & Baer, D. J. (2012). Discrepancy between the Atwater factor predicted and empirically measured energy values of almonds in human diets. *The American Journal of Clinical Nutrition*, *96*(2), 296–301. https://doi.org/10.3945/ajcn.112.035782

Nubel. (2023). *Internubel: Database of tradenames*. Retrieved May 20, 2023, from https://www.internubel.be/

Odum, H., Cantlon, J., & Kornicker, L. (1960). An organizational hierarchy postulate for the interpretation of species-individual distributions, species entropy, ecosystem evolution, and the meaning of a species-variety index. *Ecology*, *41*(2), 395–399. https://eurekamag.com/research/024/200/024200366.php

Peet, R. (2003). The measurement of species diversity. *Annual Review of Ecology and Systematics*, *5*(1), 285–307. https://doi.org/10.1146/annurev.es.05.110174.001441

Phillips, K. M., McGinty, R. C., Couture, G., Pehrsson, P. R., McKillop, K., & Fukagawa, N. K. (2021). Dietary fiber, starch, and sugars in bananas at different stages of ripeness in the retail market. *PLOS ONE*, *16*(7), Article e0253366. https://doi.org/10.1371/journal.pone.0253366

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection.

Science, *344*(6187), Article 1246752. https://doi.org/10.1126/science.1246752

Pollefliet, L. (2022). *Scoren met je scriptie [Scoring with your thesis]*. Owl Press.

Posit team. (2023). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. http://www.posit.co/

Powell, B., Thilsted, S. H., Ickowitz, A., Termote, C., Sunderland, T., & Herforth, A. (2015). Improving diets with wild and cultivated biodiversity from across the landscape. *Food Security*, *7*(3), 535–554. https://doi.org/10.1007/s12571-015-0466-5

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramankutty, N., Mehrabi, Z., Waha, K., Jarvis, L., Kremen, C., Herrero, M., & Rieseberg, L. (2018). Trends in global agricultural land use: Implications for environmental health and food security. *Annual Review of Plant Biology*, *69*(1), 789–815. https://doi.org/10.1146/annurev-arplant-042817-040256

Rathje, W. L., & Murphy, C. (2001). *Rubbish!: The Archaeology of garbage*. University of Arizona Press. https://books.google.be/books/about/Rubbish.html?id=1i_B1c573OQC&redir_esc=y

Remans, R., Wood, S. A., Saha, N., Anderman, T. L., & DeFries, R. S. (2014). Measuring nutritional diversity of national food supplies. *Global Food Security*, *3*(3), 174–182. https://doi.org/10.1016/j.gfs.2014.07.001

Riboli, E., Hunt, K. J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U. R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., . . . Saracci, R. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutrition*, *5*(6b), 1113–1124. https://doi.org/10.1079/PHN2002394

Riboli, E., & Kaaks, R. (1997). The EPIC Project: Rationale and study design. European Prospective Investigation into Cancer and Nutrition. *International Journal of Epidemiology*, *26*(1, Suppl.), S6–S14. https://doi.org/10.1093/ije/26.suppl_1.S6

Ricotta, C. (2005). Through the jungle of biological diversity. *Acta Biotheoretica*, *53*(1), 29–38. https://doi.org/10.1007/s10441-005-7001-6

Risch, S. J., & Hansen, M. K. (1982). Plant growth, flowering phenologies, and yields of corn, beans and squash grown in pure stands and mixtures in Costa Rica. *Journal of Applied Ecology*, *19*(3), 901–916. https://doi.org/10.2307/2403292

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S. I., Lambin, E., Lenton, T., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P., Costanza, R., Svedin, U., . . . Foley, J. (2009). Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*, *14*(2), Article 32. https://doi.org/10.5751/ES-03180-140232

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2021). *Ggally: Extension to 'ggplot2' (R package version 2.1.2)* [Computer software]. https://CRAN.R-project.org/package=GGally

Schoeller, D. A. (1995). Limitations in the assessment of dietary energy intake by self-report. *Metabolism*, *44*, 18–22. https://doi.org/10.1016/0026-0495(95)90204-X

Secretariat of the Convention on Biological Diversity. (2020). *Global Diversity Outlook 5*. https://www.cbd.int/gbo/gbo5/publication/gbo-5-en.pdf

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sheehy, J. E., Mitchell, P. L., Allen, L. H., & Ferrer, A. B. (2006). Mathematical consequences of using various empirical expressions of crop yield as a function of temperature. *Field Crops Research*, *98*(2), 216–221. https://doi.org/10.1016/j.fcr.2006.02.008

Siegel, K. R., Ali, M. K., Srinivasiah, A., Nugent, R. A., & Narayan, K. M. V. (2014). Do we produce enough fruits and vegetables to meet global health need? *PLOS ONE*, *9*(8), Article e104059. https://doi.org/10.1371/journal.pone.0104059

Smith, A. F., Jobe, J. B., & Mingay, D. J. (1991). Retrieval from memory of dietary information. *Applied Cognitive Psychology*, *5*(3), 269–296. https://doi.org/10.1002/acp.2350050308

Spence, C., Wang, Q. J., & Youssef, J. (2017). Pairing flavours and the temporal order of tasting. *Flavour*, *6*(1), Article 4. https://doi.org/10.1186/s13411-017-0053-0

Stevens, G. A., Beal, T., Mbuya, M. N. N., Luo, H., Neufeld, L. M., Addo, O. Y., Adu-Afarwuah, S., Alayón, S., Bhutta, Z., Brown, K. H., Jefferds, M. E., Engle-Stone, R., Fawzi, W., Hess, S. Y., Johnston, R., Katz, J., Krasevec, J., McDonald, C. M., Mei, Z., . . . Young, M. F. (2022). Micronutrient deficiencies among preschool-aged children and women of reproductive age worldwide: A pooled analysis of individual-level data from population-representative surveys. *The Lancet Global Health*, *10*(11), e1590–e1599. https://doi.org/10.1016/S2214-109X(22)00367-9

Tenny, S., & Boktor, S. W. (2022). Incidence. In *StatPearls: Content is king*. StatPearls Publishing. Retrieved May 20, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK430746/

Thiébaut, A. C., Freedman, L. S., Carroll, R. J., & Kipnis, V. (2007). Is it necessary to correct for measurement error in nutritional epidemiology? *Annals of Internal Medicine*, *146*(1), 65–67. https://doi.org/10.7326/0003-4819-146-1-200701020-00012

Thrupp, L. A. (2000). Linking agricultural biodiversity and food security: The valuable role of agrobiodiversity for sustainable agriculture. *International Affairs*, *76*(2), 265–281. https://doi.org/10.1111/1468-2346.00133

Tian, X., Wu, M., Zang, J., Zhu, Y., & Wang, H. (2017). Dietary diversity and adiposity in Chinese men and women: An analysis of four waves of cross-sectional survey data. *European Journal of Clinical Nutrition*, *71*(4), 506–511. https://doi.org/10.1038/ejcn.2016.212

United Nations Environment Programme. (2021). *Food Waste Index Report 2021*. https://www.unep.org/resources/report/unep-food-waste-index-report-2021

van der Toorn, J. E., Cepeda, M., Kiefte-de Jong, J. C., Franco, O. H., Voortman, T., & Schoufour, J. D. (2020). Seasonal variation of diet quality in a large middle-aged and elderly Dutch population-based cohort. *European Journal of Nutrition*, *59*(2), 493–504. https://doi.org/10.1007/s00394-019-01918-5

Wang, X.-W., Grodstein, F., Bhupathiraju, S. N., Sun, Q., Zhang, X., Hu, F., Weiss, S. T., & Liu, Y.-Y. (2021). Nutritional redundancy in human diet. *bioRxiv*. https://doi.org/10.1101/2021.02.04.429776

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* [Computer software]. Springer-Verlag. https://link.springer.com/book/10.1007/978-3-319-24277-4

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L. J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J. A., De Vries, W., Majele Sibanda, L., . . . Murray, C. J. L. (2019). Food in the Anthropocene: The EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet*, *393*(10170), 447–492. https://doi.org/10.1016/S0140-6736(18)31788-4

Women’s Dietary Diversity Project Study Group. (2017). Development of a dichotomous indicator for population-level assessment of dietary diversity in women of reproductive age. *Current Developments in Nutrition*, *1*(12), Article e001701. https://doi.org/10.3945/cdn.117.001701

Zhang, Q., Chen, X., Liu, Z., Varma, D. S., Wan, R., & Zhao, S. (2017). Diet diversity and nutritional status among adults in southwest China. *PLOS ONE*, *12*(2), Article e0172406. https://doi.org/10.1371/journal.pone.0172406

Zimmerer, K. S. (2015). Understanding agrobiodiversity and the rise of resilience: Analytic category, conceptual boundary object or meta-level transition? *Resilience*, *3*(3), 183–198. https://doi.org/10.1080/21693293.2015.1072311

APPENDIX

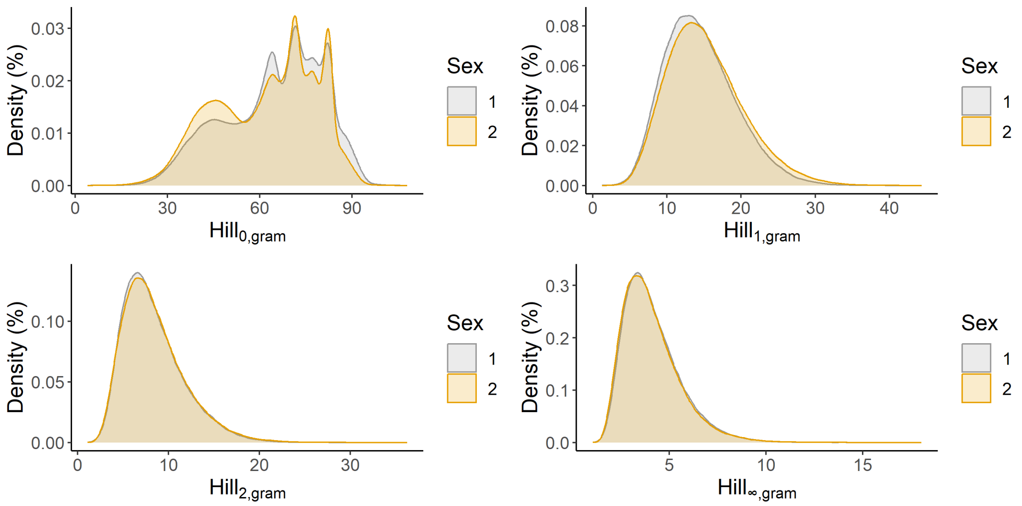

Figure A1: Association between weight-based indices and gender using the full dataset (n=476,768).

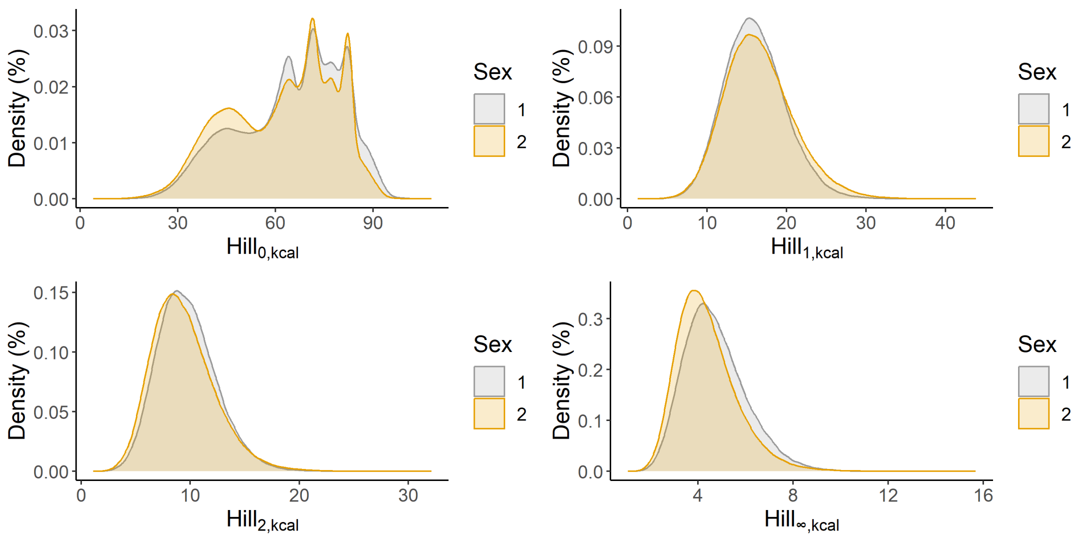

Figure A2: Association between energy-based indices and gender using the full dataset (n=476,768).

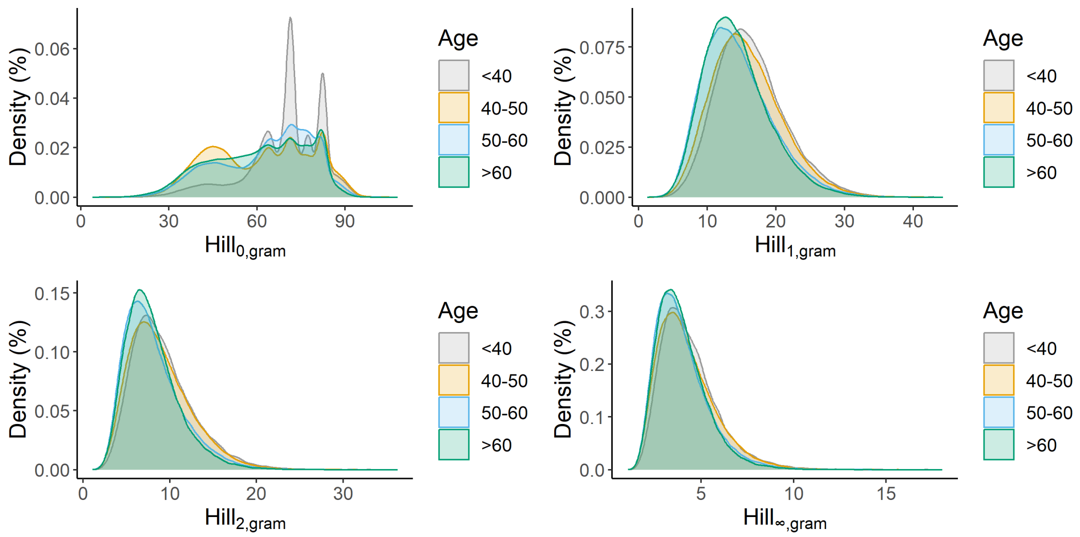

Figure A3: Association between weight-based indices and age at recruitment using the full dataset (n=476,768).

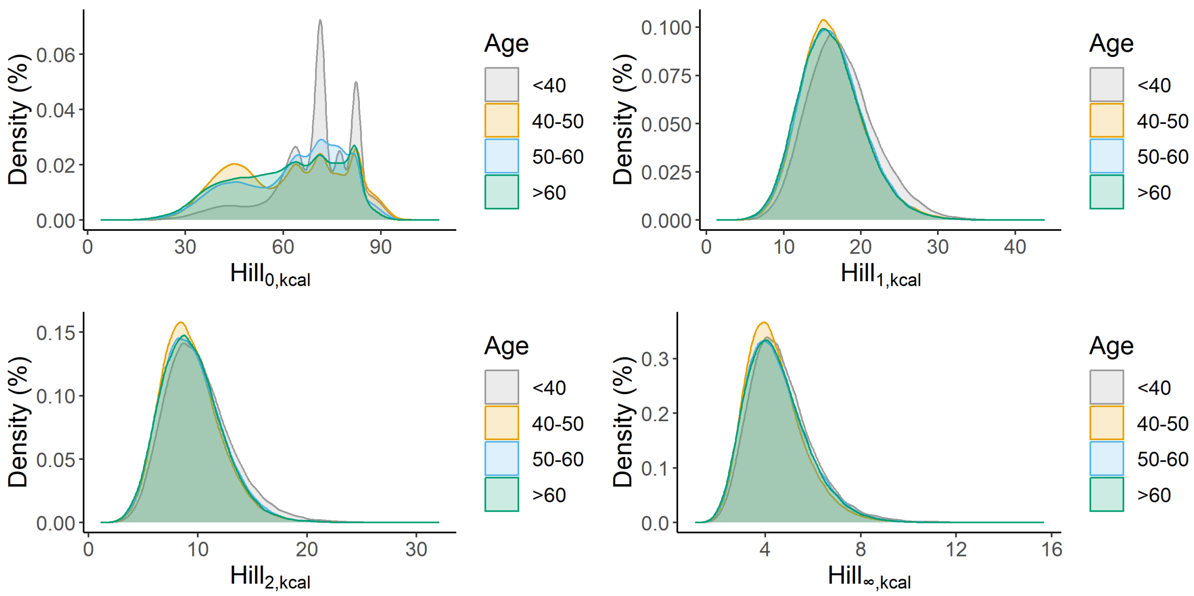

Figure A4: Association between energy-based indices and age at recruitment using the full dataset (n=476,768).

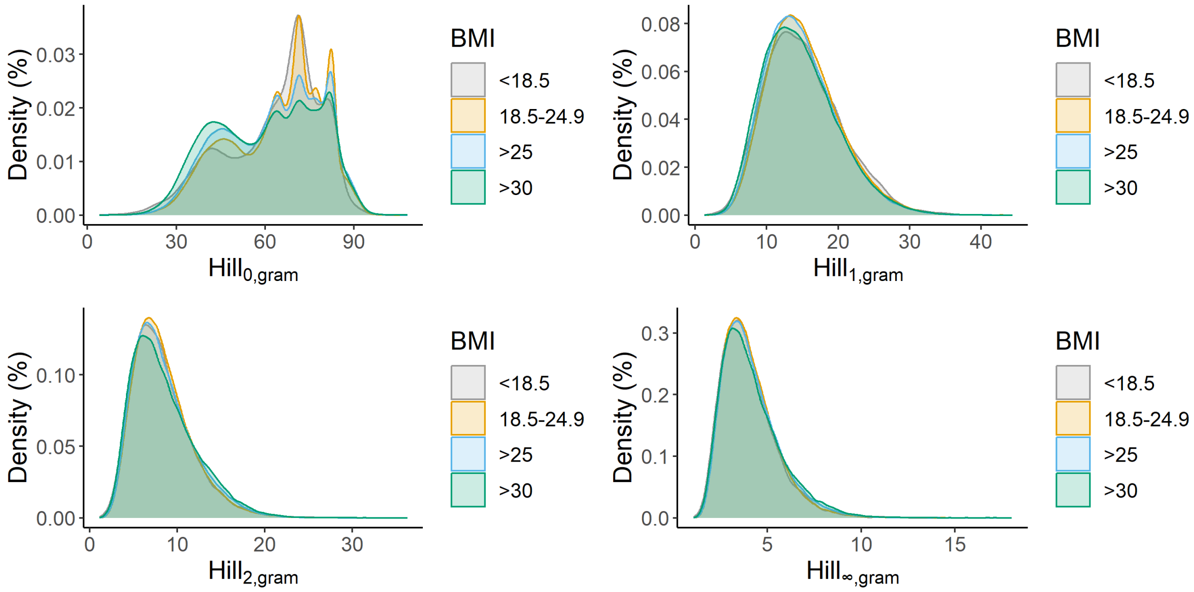

Figure A5: Association between weight-based indices and body mass index (BMI) using the full dataset (n=476,768).

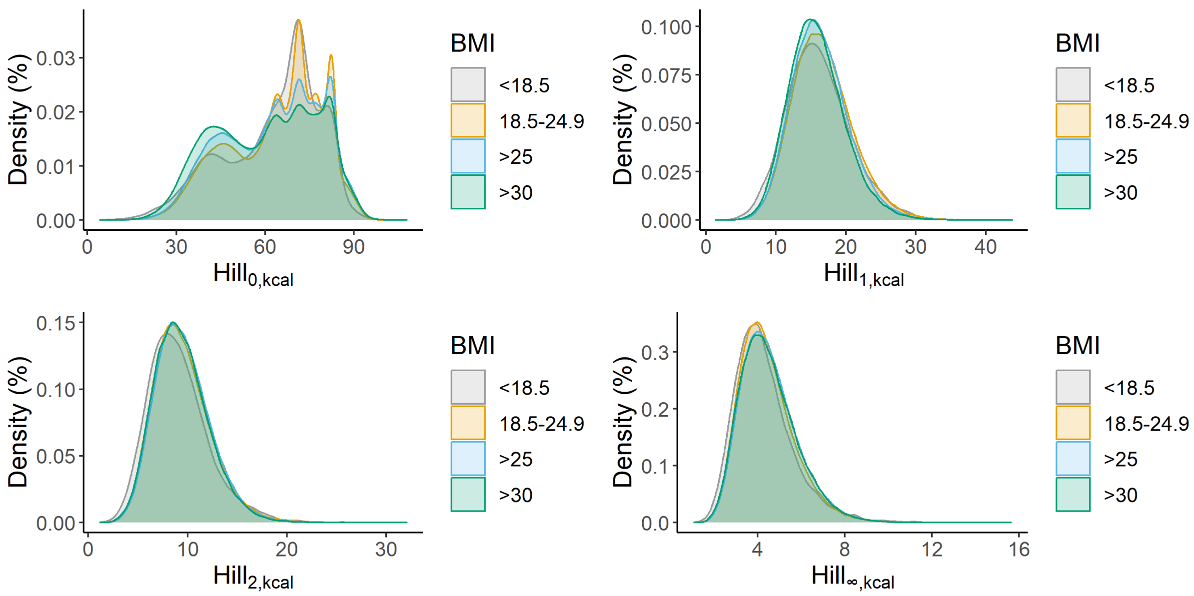

Figure A6: Association between energy-based indices and body mass index (BMI) using the full dataset (n=476,768).

Table A1: Pearson and Spearman correlation coefficients between different Hill numbers and $Jill_x$ indices using the full dataset (n=476,768).

Correlation coefficient between			**Pearson**	**Spearman**
$Hill_{0,kcal}$	&	$Hill_{0,gram}$	1.00	1.00
$Hill_{1,kcal}$	&	$Hill_{1,gram}$	0.42	0.39
$Hill_{2,kcal}$	&	$Hill_{2,gram}$	0.23	0.21
$Hill_{\infty,kcal}$	&	$Hill_{\infty,gram}$	0.13	0.12
$Hill_{0,gram}$	&	$Hill_{1,gram}$	0.24	0.22
$Hill_{0,gram}$	&	$Hill_{2,gram}$	0.07	0.06
$Hill_{0,gram}$	&	$Hill_{\infty,gram}$	0.04	0.04
$Hill_{1,gram}$	&	$Hill_{2,gram}$	0.93	0.93
$Hill_{1,gram}$	&	$Hill_{\infty,gram}$	0.79	0.79
$Hill_{2,gram}$	&	$Hill_{\infty,gram}$	0.93	0.93
$Hill_{0,kcal}$	&	$Hill_{1,kcal}$	0.40	0.40
$Hill_{0,kcal}$	&	$Hill_{2,kcal}$	0.20	0.20
$Hill_{0,kcal}$	&	$Hill_{\infty,kcal}$	0.15	0.15
$Hill_{1,kcal}$	&	$Hill_{2,kcal}$	0.9	0.89
$Hill_{1,kcal}$	&	$Hill_{\infty,kcal}$	0.68	0.68
$Hill_{2,kcal}$	&	$Hill_{\infty,kcal}$	0.89	0.89
$Hill_{2,gram}$	&	$Jill_{2,gram}$	0.97	0.98
$Hill_{2,gram}$	&	$Jill_{3,gram}$	0.97	0.96
$Hill_{2,gram}$	&	$Jill_{4,gram}$	0.95	0.93
$Hill_{2,gram}$	&	$Jill_{5,gram}$	0.93	0.90
$Hill_{2,kcal}$	&	$Jill_{2,kcal}$	0.95	0.96
$Hill_{2,kcal}$	&	$Jill_{3,kcal}$	0.96	0.96
$Hill_{2,kcal}$	&	$Jill_{4,kcal}$	0.94	0.93
$Hill_{2,kcal}$	&	$Jill_{5,kcal}$	0.91	0.89

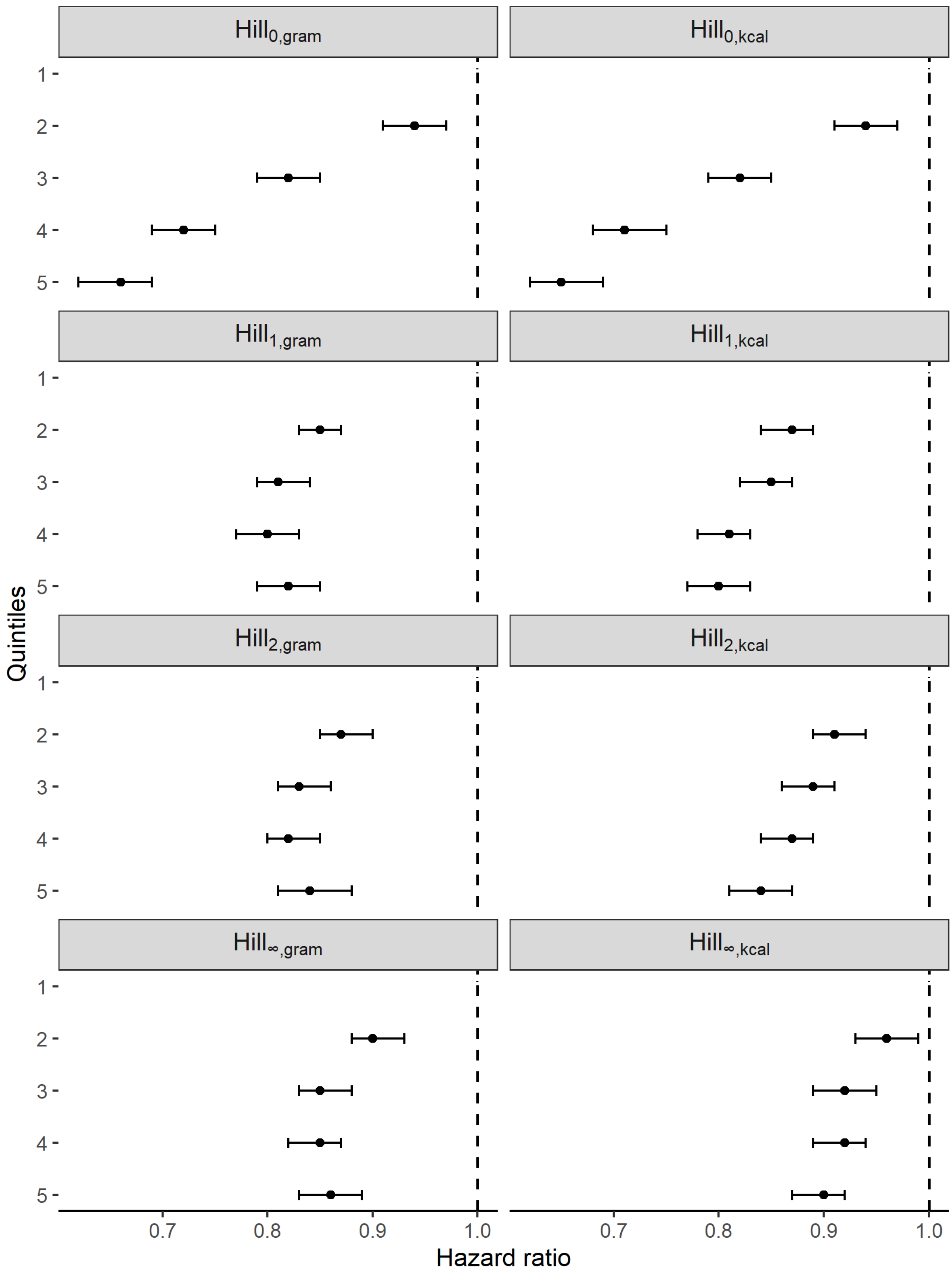

Figure A7: Hazard ratios for the Hill numbers using the overall mortality dataset (n=451,390).

Table A2: Overview of data provided by IARC on overall mortality: hazard ratios and absolute number of deaths per 10,000 person-years using the overall mortality dataset (n=451,390).

Index	**Lower limit (<=)**	**Upper limit (<)**	**Quintile**	**Number observed**	**Cases of death**	**Years in study**	**Hazard ratio**	**CI (lower)**	**CI (upper)**	**Cases per 10,000 years**	**difference with reference**
$Hill_{0,gram}$		47	1	95677	12301	1615508	1	Ref.		76.14	
	47	62	2	95191	12014	1616287	0.94	0.91	0.97	74.33	1.81
	62	71	3	95439	8939	1619193	0.82	0.79	0.85	55.21	20.94
	71	78	4	78361	6676	1298408	0.72	0.69	0.75	51.42	24.73
	78		5	86722	6697	1357086	0.66	0.62	0.69	49.35	26.79
$Hill_{1,gram}$		10.86	1	95970	15017	1573521	1	Ref.		95.44	
	10.86	13.41	2	91353	10400	1520687	0.85	0.83	0.87	68.39	27.05
	13.41	15.95	3	89519	8567	1493858	0.81	0.79	0.84	57.35	38.09
	15.95	19.29	4	87975	7022	1469874	0.8	0.77	0.83	47.77	47.66
	19.29		5	86573	5621	1448541	0.82	0.79	0.85	38.80	56.63
$Hill_{2,gram}$		5.62	1	94495	13542	1561214	1	Ref.		86.74	
	5.62	7.13	2	91630	10677	1529943	0.87	0.85	0.9	69.79	16.95
	7.13	8.77	3	89924	8972	1506644	0.83	0.81	0.86	59.55	27.19
	8.77	11.12	4	88515	7562	1474049	0.82	0.8	0.85	51.30	35.44
	11.12		5	86826	5874	1434632	0.84	0.81	0.88	40.94	45.80
$Hill_{\infty,gram}$		2.88	1	93213	12260	1547922	1	Ref.		79.20	
	2.88	3.53	2	91547	10594	1527817	0.9	0.88	0.93	69.34	9.86
	3.53	4.22	3	90187	9235	1508386	0.85	0.83	0.88	61.22	17.98
	4.22	5.22	4	89036	8083	1480385	0.85	0.82	0.87	54.60	24.60
	5.22		5	87407	6455	1441971	0.86	0.83	0.89	44.77	34.44
$Hill_{0,kcal}$		47.00	1	97040	12422	1637246	1	Ref.		75.87	
	47.00	62.00	2	95270	12012	1621284	0.94	0.91	0.97	74.09	1.78
	62.00	71.00	3	95674	8948	1624058	0.82	0.79	0.85	55.10	20.77
	71.00	78.00	4	77825	6613	1288278	0.71	0.68	0.75	51.33	24.54
	78.00		5	85581	6632	1335616	0.65	0.62	0.69	49.65	26.22
$Hill_{1,kcal}$		12.96	1	92978	12025	1514052	1	Ref.		79.42	
	12.96	15.14	2	90556	9603	1492328	0.87	0.84	0.89	64.35	15.07
	15.14	17.20	3	90093	9141	1495204	0.85	0.82	0.87	61.14	18.29
	17.20	19.82	4	89351	8398	1495920	0.81	0.78	0.83	56.14	23.28
	19.82		5	88412	7460	1508978	0.8	0.77	0.83	49.44	29.99
$Hill_{2,kcal}$		7.09	1	91016	10063	1497795	1	Ref.		67.19	
	7.09	8.51	2	90148	9195	1489471	0.91	0.89	0.94	61.73	5.45
	8.51	9.91	3	90275	9323	1501457	0.89	0.86	0.91	62.09	5.09
	9.91	11.72	4	90348	9395	1509543	0.87	0.84	0.89	62.24	4.95
	11.72		5	89603	8651	1508215	0.84	0.81	0.87	57.36	9.83
$Hill_{\infty,kcal}$		3.40	1	90471	9518	1504312	1	Ref.		63.27	
	3.40	4.00	2	89895	8942	1492200	0.96	0.93	0.99	59.92	3.35
	4.00	4.62	3	89947	8995	1496719	0.92	0.89	0.95	60.10	3.17
	4.62	5.46	4	90361	9408	1503060	0.92	0.89	0.94	62.59	0.68
	5.46		5	90716	9764	1510191	0.9	0.87	0.92	64.65	-1.38

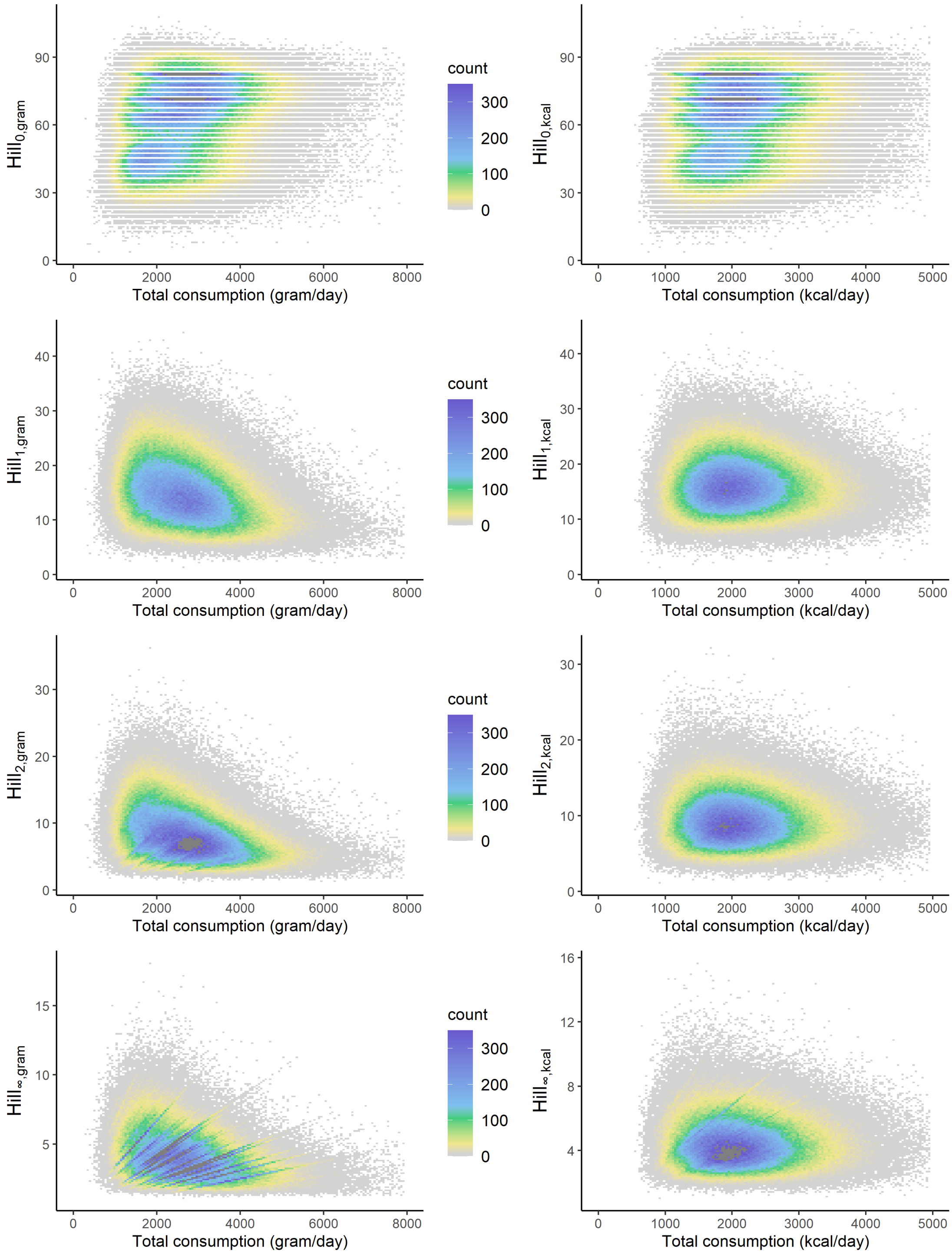

Figure A8: Weight-based and energy-based Hill numbers based on total consumption per day (n=476,768).