UNIVERSITY OF TURKU

Compositional data analysis applied to a study of movement behaviours of recent retirees

Jesse Pasanen

MSc Thesis
March 2021

DEPARTMENT OF MATHEMATICS AND STATISTICS

UNIVERSITY OF TURKU
Department of Mathematics and Statistics

PASANEN, JESSE: Compositional data analysis applied to a study of movement behaviours of recent retirees
MSc Thesis, 36 pages
Statistics
March 2021

Studies of movement behaviours have numerous applications. A recent approach involves studying the time spent on different activity types using compositional data analysis. In compositional data analysis, several variables are constrained to an arbitrary sum and the primary interest is their proportions of the whole.

This thesis explores the mathematical foundations of the study of compositional data and their practical applications. First, mathematical operations are defined for compositions using Aitchison geometry. Methods are presented for transforming compositions into real-valued coordinates and back. Various statistical methods are also defined for compositions and compositional data.

Some of the techniques presented are demonstrated by applying them to a study of movement behaviours. REACT is a randomized controlled trial study focusing on whether commercial activity trackers affect movement behaviours among the recently retired. By using compositional data analysis, the proportions of time spent on different activity types can be studied. Based on the results, it would appear that those who used activity trackers spent a slightly higher portion of their day on physical activity than those who did not.

Keywords: Compositional data analysis, Aitchison geometry, Isometric logratio coordinates, Physical activity, Movement behaviours, REACT.

Contents

1 Introduction

In recent years, activity trackers worn on the wrist have become available on consumer markets. These devices are used to track physical activity throughout the day, set activity goals and remind users to move regularly. Their main purpose is to help users maintain regular activity. If such devices prove to be effective, they can be an affordable method of improving health at a national level. REACT, a randomized controlled trial study, was funded by the Academy of Finland and the Finnish Ministry of Education and Culture to study the effect the activity trackers have on the physical activity of the recently retired (Leskinen et al. 2021). This study is part of FIREA (Finnish Retirement and Aging), a broader study on the health effects of retirement and aging.

The study was performed as a controlled trial with two groups. The intervention group was given an activity tracker, while the control group was not. Their physical activity was then measured at certain points over the period of a year. For measuring physical activity, each subject was equipped with a separate wrist-mounted accelerometer, worn for a week.

Historically, different movement behaviours such as physical activity, sedentary behaviour and sleep have been studied separately from each other. However, these studies often neglected the fact that increasing one type of activity causes the other types to decrease, which may have led to flawed results. Later studies have accounted for this by treating movement behaviours as compositional data, which represent the different types of activity as proportions of time spent on them daily.

Compositional data constitute a type of multivariate data where the only information variables carry are their relative proportions of the overall observation. This means that each variable is constrained by the values of the other variables in the data. Due to these constraints, common multivariate techniques are unsuitable for compositional data. Various analysis methods have been developed to take into account the nature of compositional data. These methods are typically grouped under the term compositional data analysis (CoDA).

This thesis examines various mathematical and statistical tools that have been developed for handling compositional data, and applies them to analyse the data provided by REACT. Our primary interests are determining whether using activity trackers affects daily movement behaviour, how this change occurs over time, and how changing the amount of one type of activity affects the others. Chapter 2 introduces Aitchison geometry, the vector space used to represent compositional data, as well as coordinate representations and transformations used for handling compositional data. Data preprocessing and visualization are also discussed. Chapter 3 covers statistical methods that can be used to analyse compositional data. Chapter 4 provides an overview of the REACT data used in the analysis, and Chapter 5 covers the analysis itself. Finally, Chapter 6 presents a summary of the findings and discusses possibilities for further studies and analysis.

1.1 Compositional data analysis in the literature

Physical activity has a major effect on many aspects of health and life in general, and as such has been the subject of numerous studies in the past. Many of these either considered different movement behaviours in isolation or did not properly adjust their analysis to account for the effects of the other behaviours. (Pedišić 2014). Previous applications of compositional data analysis on the other hand have mainly focused on subjects such as geology or chemistry, where compositions tend to have large numbers of parts and can be measured accurately (Buccianti, Mateu-Figueras, and Pawlowsky-Glahn 2006). In recent years, a growing number of studies have applied CoDA to movement behaviours, as can be seen in a 2020 review of various CoDA studies (Janssen et al. 2020).

Compositional data analysis is a relatively recent development in statistics, although its roots date to the late nineteenth century (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015, 5-7). Many foundational concepts were developed by John Aitchison, whose logratio-based approach is still the basis for many concepts. His major theoretical work, *The Statistical Analysis of Compositional Data*, was released in 1986 (Aitchison 1986). Many fundamental concepts are also named after him, such as the Aitchison geometry discussed in section 2.1. He remained active in the development of CoDA into the early twenty-first century. (Pawlowsky-Glahn and Buccianti 2011, 3-9).

The main source for the theoretical sections of this thesis is a textbook on the subject of compositional data analysis, *Modeling and Analysis of Compositional Data* (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015), which provides a comprehesive coverage of the theory of compositional data analysis, including advanced subjects that are not covered in this thesis. Additionally, the book *Applied Compositional Data Analysis* is used to supplement theoretical sections as well as provide a foundation for the analysis done in R (Filzmoser, Hron, and Templ 2018).

2 Representing Compositional Data

Compositional data are defined to be observations for which the only relevant information is carried in the proportions of the components. That is, one is not interested in the specific values of the components, but in their proportions of the whole. The sample space of compositional data is the simplex. A simplex is defined as the collection of vectors whose components sum up to some pre-specified constant. The choice of the constant κ is essentially arbitrary and is usually determined by the type of data being analysed.

Definition 1 (Simplex)**.** *The D-dimensional simplex S^D is defined as*

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) | x_i > 0, i = 1, 2, \dots, D; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Any vector can be mapped onto a simplex by scaling the components by their sum, and then normalizing the result with a constant. This operation is called a closure.

Definition 2 (Closure). *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_D)$*,* $x_i > 0, i = 1, 2, \ldots, D$*, be a strictly positive vector. Its closure to* $\kappa > 0$ *is defined as*

$$C(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

The result of the closure is on the simplex, with $C(\mathbf{x}) \in S^D$*.*

It can be noted that since the sum of elements is constant for every $\mathbf{x} \in S^D$, the closure essentially multiplies the vector by an arbitrary constant, that is $C(\mathbf{x}) = a\mathbf{x}$, where $a > 0$. With this in mind, we can show that the closure satisfies a property called *scale invariance*, a feature that will be relevant later.

Lemma 1. $C(\mathbf{x}) = C(a\mathbf{x})$*, where* $a > 0$*.*

Proof.

$$\begin{aligned} C(a\mathbf{x}) &= \left(\frac{a\kappa x_1}{\sum_{i=1}^{D} a x_i}, \frac{a\kappa x_2}{\sum_{i=1}^{D} a x_i}, \ldots, \frac{a\kappa x_D}{\sum_{i=1}^{D} a x_i} \right) \\ &= \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right) = C(\mathbf{x}). \end{aligned}$$

□

Instead of considering a full composition, it is often desireable to inspect only certain compositional parts. This can be required, for example, because certain components are not relevant to the analysis, or to simplify analysis and visualization. A subset of a composition is called a subcomposition. Any subcomposition is also a composition and can be analysed with the same methods.

Definition 3 (Subcomposition). *Let* $\mathbf{x} \in S^D$ *be a D-dimensional composition, and let* $\mathrm{Sub}(\mathbf{x})$ *be a function which selects some proper subset of the components from* $\mathbf{x}$*. The subcomposition* $\mathbf{y}$ *is defined as the closure of this function,*

$$\mathbf{y} = C(\mathrm{Sub}(\mathbf{x})).$$

The subcomposition has a smaller dimension than the original composition, with $\mathbf{y} \in S^K$*,* $1 < K < D$*. The* $\mathrm{Sub}()$ *function should select same parts on any particular composition.*

There are three principles that should be fulfilled by statistical methods that are to be used on compositional data. These are called the principles of compositional analysis. The first principle is that of *scale invariance*. Since compositional data carries only relative information, results obtained from statistical methods should not depend on any scaling factors on the data. In practice, this means that compositional analysis methods disregard scale and consider only relative sizes.

Definition 4 (Scale invariance). *Let* $\mathbf{x} \in S^D$ *be a D-dimensional composition,* $a \in \mathbb{R}_+$ *a positive real valued constant, and* $f(\cdot)$ *a function on* $\mathbb{R}^D_+$*. The function* f *is scale invariant if* $f(\mathbf{x}) = f(a\mathbf{x})$ *for every* $\mathbf{x}$ *and* a*.*

In Lemma 1 we demonstrated that the closure operation is scale invariant. This means that by Definition 3 all subcompositions are also scale invariant, since simply selecting a subset of the scaled parts does not change their relative scaling.

The second principle is that of *permutation invariance.* This means that analysis does not consider the order of the compositional parts. In general, most methods fulfill this principle as long as they do not depend on the order the variables are stored in the dataset. The third principle is that of *subcompositional coherence.* In basic terms, conclusions made from subcompositions should not contradict conclusions made from the same components in the full composition. This means that subcompositions should behave similarly to orthogonal projections in real space. In practice, this means that scale invariance is preserved in arbitrary subcompositions, and that the distance between two compositions is equal to or greater than the distance between their subcompositions. This guarantees that inferences made about the relations of components in a subcomposition also hold for the full composition.

Definition 5 (Subcompositional coherence)**.** *In order for subcompositional coherence to be in effect, two conditions must be fulfilled. First, if $\Delta_p(\mathbf{x}, \mathbf{y})$ is a measure of distance between two p-dimensional compositions, then*

$$\Delta_D(\mathbf{x}, \mathbf{y}) \geq \Delta_K(\mathbf{x}_K, \mathbf{y}_K)$$

must apply for all D-part compositions $\mathbf{x}$, $\mathbf{y}$ and their K-part subcompositions $\mathbf{x}_K$, $\mathbf{y}_K$. Second, scale invariance should be preserved in arbitrary subcompositions, meaning that the ratios of parts in the subcomposition should be the same as the corresponding ratios in the original composition.

2.1 Aitchison geometry

Aitchison geometry is named after John Aitchison, who contributed greatly to research of the properties of compositional data. Aitchison geometry defines a vector space for the simplex by defining several operations for compositions, analogous to transformation and scaling operations in other geometries.

Definition 6 (Basic operations)**.** *The perturbation of compositions $\mathbf{x}, \mathbf{y} \in S^D$ is defined as*

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \ldots, x_D y_D),$$
$$\mathbf{x} \ominus \mathbf{y} = C(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \ldots, \frac{x_D}{y_D}).$$

Additionally, for compositions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in S^D$,

$$\bigoplus_{i=1}^{k} \mathbf{x}_i = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_k.$$

The powering of a vector $\mathbf{x} \in S^D$ by a constant $a \in \mathbb{R}$ is defined as

$$a \odot \mathbf{x} = C(x_1^a, x_2^a, \ldots, x_D^a).$$

With these operations, the simplex S^D can be shown to be a vector space. There are several rules which need to be fulfilled.

Definition 7 (Vector space)**.** *A vector space is a set V for which two operations, vector addition $(+)$ and scalar multiplication $(\cdot)$, are defined. The vector addition is an operation between two vectors in V, and the scalar multiplication is an operation between a real number and a vector in V. These operations must fulfill several requirements.*

The vector addition operation must fulfill the following conditions:

1. *Closure: For all vectors $\mathbf{u}, \mathbf{v} \in V$, $(\mathbf{u} + \mathbf{v}) \in V$.*

2. *Commutative law: For all vectors $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.*

3. *Associative law: For all vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.*

4. *Additive identity: There exists a neutral element $\mathbf{n} \in V$ for which $\mathbf{n} + \mathbf{v} = \mathbf{v} + \mathbf{n} = \mathbf{v}$ for all vectors $\mathbf{v} \in V$.*

5. *Additive inverses: For all vectors $\mathbf{v} \in V$ there exists a vector $\mathbf{v}^{-1} \in V$ for which $\mathbf{v}^{-1} + \mathbf{v} = \mathbf{v} + \mathbf{v}^{-1} = \mathbf{n}$.*

The scalar multiplication operation must fulfill the following conditions:

1. *Closure: For all vectors $\mathbf{v} \in V$ and scalars $c \in \mathbb{R}$, $(c \cdot \mathbf{v}) \in V$.*

2. *Distributive law: For all vectors $\mathbf{u}, \mathbf{v} \in V$ and scalars $c \in \mathbb{R}$, $c \cdot (\mathbf{u} + \mathbf{v}) = c \cdot \mathbf{u} + c \cdot \mathbf{v}$.*

3. *Distributive law: For all vectors $\mathbf{v} \in V$ and scalars $c, d \in \mathbb{R}$, $(c + d) \cdot \mathbf{v} = c \cdot \mathbf{v} + d \cdot \mathbf{v}$.*

4. *Associative law: For all vectors $\mathbf{v} \in V$ and scalars $c, d \in \mathbb{R}$, $c \cdot (d \cdot \mathbf{v}) = (cd) \cdot \mathbf{v}$.*

5. *Unitary law: For all vectors $\mathbf{v} \in V$, $1 \cdot \mathbf{v} = \mathbf{v}$.*

Lemma 2 (Simplex as a vector space)**.** *The simplex S^D is a vector space. The perturbation of compositions $\mathbf{x}, \mathbf{y} \in S^D$ is the vector addition operator, and the powering of $\mathbf{x}$ by a constant $a \in \mathbb{R}$ is the scalar multiplication operator.*

Proof. We may first note that chaining perturbations and powerings introduces multiple closures to the equation. Per Defintion 2, we know that $C(\mathbf{x}) = a\mathbf{x}$ for some arbitrary constant $a > 0$, and per Lemma 1 we know that $C(a\mathbf{x}) = C(\mathbf{x})$, which simplifies the proofs somewhat. We will use the constants k, l and m to represent the effects of inner closures where necessary. We will next verify the conditions of Definition 7 one-by-one.

1. Per Definition 2, the product of perturbation is on the simplex, $\mathbf{x} \oplus \mathbf{y} \in S^D$.

2. $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$, since $x_i y_i = y_i x_i$.

3. $\mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z}) = (\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z}$, since $C(x_i(l y_i z_i)) = C((m x_i y_i) z_i)$.

4. The neutral element is $\mathbf{n} = (a, a, \ldots, a) = (\frac{\kappa}{D}, \frac{\kappa}{D}, \ldots, \frac{\kappa}{D}) \in S^D$, the D-part composition whose elements all have the same value. Since pertubation with $\mathbf{n}$ is equivalent to scaling with a constant, per Lemma 1 we have $\mathbf{x} \oplus \mathbf{n} = \mathbf{n} \oplus \mathbf{x} = \mathbf{x}$.

5. Since $x_i \frac{1}{x_i} = 1$, the inverse element of $\mathbf{x}$ is the composition $\mathbf{x}^{-1} = (1/x_1, 1/x_2, \ldots, 1/x_D)$, and $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{x}^{-1} \oplus \mathbf{x} = C(1, 1, \ldots, 1) = \mathbf{n}$.

Since the five conditions are fulfilled, we have shown that perturbation is a vector addition operation.

1. Per Definition 2, the product of powering is on the simplex, $a \odot \mathbf{x} \in S^D$.

2. $a \odot (\mathbf{x} \oplus \mathbf{y}) = a \odot \mathbf{x} \oplus a \odot \mathbf{y}$, since $C((kx_i y_i)^a) = C(l y_i^a m x_i^a)$.

3. $(a + b) \odot \mathbf{x} = a \odot \mathbf{x} \oplus b \odot \mathbf{x}$, since $C((kx_i)^{a+b}) = C(l x_i^a m x_i^b)$.

4. $a \odot (b \odot \mathbf{x}) = (ab) \odot \mathbf{x}$, since $C((l x_i^b)^a) = C(x_i^{ab})$.

5. $1 \odot \mathbf{x} = \mathbf{x}$, since $x_i^1 = x_i$.

Since the five conditions are fulfilled, we have also shown that powering is a scalar multiplication operation. Since the two operations are defined and fulfill the conditions, the simplex S^D is a vector space.

□

In addition to the basic operations, various other useful vector operations can be defined for compositional data.

Definition 8 (Aitchison inner product)**.** *The inner product of compositions* $\mathbf{x}, \mathbf{y} \in S^D$ *is defined as*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Definition 9 (Aitchison norm)**.** *The norm of a composition* $\mathbf{x} \in S^D$ *is defined as*

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}.$$

Definition 10 (Aitchison distance)**.** *The distance between compositions* $\mathbf{x}, \mathbf{y} \in S^D$ *is defined as*

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}.$$

As an example of the utility of the Aitchison geometry, consider two two-part compositions: (500, 100) and (1, 5). The two differ quite significantly in their proportions, which are 5:1 and 1:5, respectively. If we were to sum the two together as regular vectors, we would get (501, 105), whose ratio is far closer to 5:1 than 1:5. The significance of the second composition disappears almost entirely due to the scale of the first. This is precisely what we seek to avoid in compositional data analysis.

If we perturb the compositions instead, our result (without closure) is (500, 500), with a ratio of 1:1. The two compositions, which have inverse proportions, have canceled each other out. Perturbation has preserved the relevant information of proportions and ignored scale as irrelevant. This also serves to illustrate that perturbing serves to shift the proportions of its inputs towards each other.

As another example, the distance between the compositions (1,3) and (2, 6) is $\sqrt{(2-1)^2 + (6-3)^2} = \sqrt{10}$ with Euclidean methods, but 0 with the Aitchison distance in Definition 10. Since the two have the same ratio of 1:3, the Aitchison distance fulfills our goal of only considering relative proportions.

2.2 Coordinate representation

While the Aitchison geometry provides a mathematical framework for manipulating compositional data, common statistical procedures cannot be directly applied to data on the simplex. Instead, special transformations can be used to express compositions in real space. Since compositions are constrained to a sum, we can even reduce the dimensionality of the data without losing any information. Specifically, a D-part composition can be represented with $D-1$ real values, which act as coordinates expressing the composition in real space. As a simple example, the ratio between two values can be written as a single number, expressing a two-part composition with a single real-valued coordinate.

The most common type of composition to real space transformation are called logratio coordinates, which as the name suggests, are based on logarithmic transformations of ratios of the compositional parts. There are several types of logratio coordinates, including the *additive* (alr), *centered* (clr), and *isometric* (ilr) logratio transformations.

The alr transformation is a mapping of the simplex S^D to the real space $\mathbb{R}^{D-1}$. It is based on selecting one compositional part against which the others are compared, leading to the following transformation:

$$\mathrm{alr}_j(\mathbf{x}) = \left(\ln\frac{x_1}{x_j}, \ldots, \ln\frac{x_{j-1}}{x_j}, \ln\frac{x_{j+1}}{x_j}, \ldots, \ln\frac{x_D}{x_j}\right).$$

The alr transformation is not isometric, that is, it does not preserve distances, and produces a non-orthogonal coordinate system. Interpretation is also hindered, since one component's information is spread across all coordinates, and the others only show their relation to the part being compared. Due to these reasons, alr coordinates are rarely used.

The clr transformation is an isometry from the simplex S^D to a subset of real space $\mathbb{R}^D$. Since the result has D dimensions rather than the minimum $D-1$

necessary for representing D-part compositions, the results are called coefficients rather than coordinates. The coefficients are based on the geometric mean of the composition,

$$\operatorname{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g_m(\mathbf{x})}, \ldots, \ln \frac{x_D}{g_m(\mathbf{x})}\right), \tag{1}$$

where $g_m(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$. Despite being isometric there are still problems with using clr as coordinates. First, we lose the quality of dimension reduction. The sum of the coordinates is zero, restricting them to a $(D-1)$-dimesional hyperplane on $\mathbb{R}^D$. Furthermore, this makes the covariance matrix of the coordinates singular. Nevertheless, the clr transformation does also have useful properties, such as

$$\operatorname{clr}(a \odot \mathbf{x}_1 \oplus b \odot \mathbf{x}_2) = a \cdot \operatorname{clr}(\mathbf{x}_1) + b \cdot \operatorname{clr}(\mathbf{x}_2). \tag{2}$$

which can easily be verified by recalling Definition 6 and that $\ln(x^a y^b) = a \ln x + b \ln y$.

To avoid the problems presented by the alr or clr transformations we have to build an orthogonal coordinate system, which requires defining an orthonormal basis for the simplex. Coordinates based on such a basis are called ilr coordinates. We can build the basis by using the hyperplane defined by the clr coefficients. The ilr coordinates mapping to such a basis will belong to the $(D-1)$-dimensional real space $\mathbb{R}^{D-1}$, and avoid singularity in their covariance matrix.

The canonical basis of $\mathbb{R}^D$ is $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\}$, where $\mathbf{e}_i \in \mathbb{R}^D$ is defined as a vector of zeros whose ith element is 1. Any vector in $\mathbb{R}^D$ can be expressed as a linear combination of these basis vectors, $\mathbf{v} = \sum_{i=1}^{D} v_i \mathbf{e}_i$. The values $v_i \in \mathbb{R}$ are coordinates that express the real-valued vector in terms of the basis. We aim to find something equivalent for the simplex S^D. We cannot use the canonical basis as-is, since the basis vectors are not on the simplex, which by Definition 1 disallows zeros. Additionally, due to the sum constraint, the basis of the simplex will have $D-1$ coordinates.

We can transform the canonical basis to the simplex by taking the closure of their element-wise exponential, resulting in the generating system $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$, where

$$\mathbf{w}_i = C(\exp(\mathbf{e}_i)) = C(1, 1, \ldots, e, \ldots, 1).$$

Recalling Definition 6, we can now write any composition $\mathbf{x} \in S^D$ as $\mathbf{x} = \bigoplus_{i=1}^{D} \ln x_i \odot \mathbf{w}_i$. This is a generating system, rather than a basis, which means that the coefficients x_i are not unique. We can obtain an orthonormal basis from the generating system by omitting one generating composition $\mathbf{w}_i$ and using the Gram–Schmidt procedure on the resulting basis (Egozcue et al. 2003). Once a suitable basis has been obtained, ilr coordinates can be defined straightforwardly.

Definition 11 (Isometric logratio coordinates). *Let* $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})$ *be an orthonormal basis for the simplex* S^D *and* $\mathbf{x} \in S^D$ *a composition. The ilr transformation of* $\mathbf{x}$ *is the function* $\operatorname{ilr}(\mathbf{x}) = \mathbf{z}$*,* $\mathbf{z} \in \mathbb{R}^{D-1}$*, which satisfies the properties*

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} z_i \odot \mathbf{e}_i, \qquad z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a.$$

The vector $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})$ *holds the ilr coordinates of the composition.*

Transforming ilr coordinates back into compositions can be done via the contrast matrix of the basis used to define the ilr transformation.

Definition 12 (Contrast matrix)**.** *Let* $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})$ *be an orthonormal basis for the simplex* S^D *and let* $\mathbf{V}$ *be a* $(D-1) \times D$ *matrix with the rows* $\mathbf{V}_i = \mathrm{clr}(\mathbf{e}_i)$*. This matrix is called the contrast matrix of the basis, and its rows are called contrasts or logcontrasts.*

$\mathbf{V}$ has the properties that $\mathbf{V}\mathbf{V}' = \mathbf{I}_{D-1}$, where $\mathbf{I}_D$ is a $D \times D$ identity matrix, and $\mathbf{V}\mathbf{V}' = \mathbf{I}_D - \frac{1}{D}\mathbf{1}'_D\mathbf{1}_D$, where $\mathbf{1}_D$ is a row vector of D ones.

Based on Definitions 11, 12 and equation (2) we can define a relation between the clr and ilr transforms so that

$$\begin{aligned} \mathrm{clr}(\mathbf{x}) &= \mathrm{clr}(\bigoplus_{i=1}^{D-1} z_i \odot \mathbf{e}_i) \\ &= \sum_{i=1}^{D-1} z_i \cdot \mathrm{clr}(\mathbf{e}_i) \\ &= \mathbf{z}\mathbf{V}. \end{aligned} \tag{3}$$

Since the clr transform simply scales the original compositional parts and takes their logarithm, we can now define an inverse transformation for recovering compositions from their ilr coordinates.

Definition 13 (Inverse ilr transform)**.** *Let* $\mathbf{V}$ *be the contrast matrix of an orthonormal basis* $\mathbf{e}$ *for the simplex* S^D *as defined in Definition 12, and let* $\mathbf{z} = \mathrm{ilr}(\mathbf{x})$ *be the ilr coordinates of a composition* $\mathbf{x} \in S^D$ *obtained with a transform using the basis* $\mathbf{e}$*. The inverse ilr transformation is defined as*

$$\mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{z}) = C(\exp(\mathbf{z}\mathbf{V})).$$

There are an infinite number of orthonormal bases that can be generated for any particular simplex S^D. As such, the choice of basis should be done according to the needs of the analysis at hand.

2.2.1 Pivot coordinates

A specific type of ilr coordinates are called pivot coordinates. They are obtained by choosing a specific basis for the ilr transformation through selecting one part of the composition to act as a pivot which only appears in one coordinate.

Definition 14 (Pivot coordinates)**.** *Let the basis of the ilr transformation be composed of vectors* $\mathbf{e}_j$, $j = 1, \ldots, D-1$, *defined as*

$$\mathbf{e}_j = \sqrt{\frac{D-j}{D-j+1}} \left(0, \ldots, 0, 1, -\frac{1}{D-j}, \ldots, -\frac{1}{D-j}\right)',$$

with the jth element of the vector being 1*. The coordinates produced by this transformation are called pivot coordinates and take the form*

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \right). \tag{4}$$

From equation (4) it can be seen that the first component, x_1, appears only in the first coordinate z_1, which means that all relative information about this component is contained in the first coordinate. Pivot coordinates are therefore useful when specifically modeling the influence of a single part in a composition.

2.2.2 Balances

Balances are another specific type of ilr coordinates, which are used to compare groups of compositional parts. Each coordinate, called a balance, corresponds to two groups and reflects which of them is proportionally more prevalent in the whole composition. This is akin to balancing them on a scale, hence the name.

Balances are constructed with a sequential binary partition of the composition. A sequential binary partition is a method of dividing a vector into several groups of variables. As the name implies, it is based on repeatedly dividing the parts into two groups. For a D-dimensional composition, the partitioning will take $(D-1)$ steps, resulting in $(D-1)$ sets of positive and negative component groups. Each of these pairs of groups will then be used to construct a coordinate.

In the first step, all parts are sorted into two groups, positive (represented by +) and negative (represented by -). The groups can be freely determined, preferably so that comparing them is of interest in the analysis. At the subsequent steps the groups continue to be divided, one group per step. This is continued until further division is no longer possible. In steps after the first one, the parts that are not part of the group currently being divided are represented by zeros. The choice of which groups are positive and which negative is arbitrary, and mainly affects how the balances are interpreted, with positive groups corresponding to larger values in the coordinates.

Step	x_1	x_2	x_3	x_4	x_5
1	+	+	-	+	-
2	+	+	0	-	0
3	+	-	0	0	0
4	0	0	+	0	-

Table 1: Example of a binary partitioning of five variables x_1-x_5. In the first step, the variables are divided into two groups, positive and negative. The positive group is further divided in step two, and the positive group of the second step is divided in step three. The original negative group is finally divided in step four. At each step, variables not being partitioned are represented with 0.

An example of a binary partitioning is provided in Table 1. We have a five-part composition, and we wish to compare parts x_1, x_2 and x_4 against parts x_3 and x_5. To do this, we first assign a sign to each part. We decide to assign the first group as positive and the second as negative. This choice is largely arbitrary.

Having partitioned all the parts in the first step, we move to partitioning the positive group. Of these, we decide that comparing x_4 to x_1 and x_2 is the best choice. We therefore set the first two parts as positive, and x_4 as negative. Parts x_3 and x_5 are not considered in this step. In step three, we partition x_1 as positive and x_2 as negative. As the partitioning progresses, the groups are divided more and more finely, which makes the choice of partition become more and more arbitrary. Again, we ignore parts x_3-x_5.

At the final step, we move back to the original negative group, and set x_3 as positive and x_5 as negative. Part x_4 was partitioned into a group of one in step two, meaning there was no further need to divide it. The final partition is composed of the partitions made at all $D-1$ steps. Each balance corresponds only to those parts which are positive or negative at the corresponding step. Furthermore, a balance does not tell how large a portion the two groups together comprise. If we wanted to know the proportion of x_1 in the whole composition, for example, we would first have to check balance 1 to find out the relative proportion of x_1, x_2 and x_4; then look at balance 2 for the proportion of x_1 and x_2; and finally check the proportion of x_1 relative to x_2.

Definition 15 (Balances). *Given a D-dimensional composition, a sequential binary partition can be constructed in $D-1$ steps. In the kth step, the components are divided into a positive group represented by the indices $\mathbf{i}^k = (i_1^k, \ldots, i_{p_k}^k)$, and a negative group represented by the indices $\mathbf{j}^k = (j_1^k, \ldots, j_{m_k}^k)$, as well as some components which do not belong in either group. The balance, or coordinate, corresponding to the kth step is*

$$z_k = \sqrt{\frac{p_k m_k}{p_k + m_k}} \ln \left(\frac{(x_{i_1^k} x_{i_2^k} \cdots x_{i_{p_k}^k})^{1/p_k}}{(x_{j_1^k} x_{j_2^k} \cdots x_{j_{m_k}^k})^{1/m_k}} \right).$$

The balances $z_1, \ldots, z_{D-1}$ map to an orthonormal basis in the simplex S^D. The elements of the basis vector $\mathbf{e}_k = (e_1^k, \ldots, e_D^k)$ corresponding to the balance z_k are calculated as

$$e_l^k = \begin{cases} \frac{1}{p_k}\sqrt{\frac{p_k m_k}{p_k+m_k}} & \textit{for } l \in \mathbf{i}^k, \\ -\frac{1}{m_k}\sqrt{\frac{p_k m_k}{p_k+m_k}} & \textit{for } l \in \mathbf{j}^k, \\ 0 & \textit{otherwise.} \end{cases}$$

The interpretation of balances depends on the groupings made during the binary partitioning. For the kth coordinate, the sign of the coordinate indicates which of the two groups in step k is more dominant in the overall composition, with positive values corresponding to the positive group and vice versa. Values close to zero indicate that the groups are close in size. The first balance, which contains information from all the components, is the one most useful for analysis.

2.2.3 Data processing

The logratio approach to coordinates comes with a notable drawback: The composition cannot contain any zeros. This is due to the nature of the logratio: A zero value will either result in a division by zero or the logarithm of a zero, both undefined. Any zeros in compositions must therefore be removed before any coordinate transforms are applied, preferably in the preprocessing stage.

Zeros may be introduced to compositional data through various means. These include variables that may naturally have the value zero (structural zeros), zeros caused by inaccuracy in the measurement stage (rounded zeros), or zero counts caused by insufficient sampling (count zeros). Additionally, missing data due to measurement errors or other reasons can also cause similar issues in the analysis.

There are several ways of handling the problem of zeros. A straightforward approach would be to use only those observations which do not have zeroes or missing data. This will lead to a loss of information, however, which is often not desirable. The alternative is imputing zeros or missing values with some acceptable value. A simple way of imputing missing values is the k-nearest neighbour method. In this approach, we select the k observations which have the smallest Aitchison distance from the observation being imputed, and the median of their corresponding parts is imputed in place of the missing value. The size of the observations needs to be adjusted before this is done, since they may differ from each other in magnitude even though their proportions are similar. More advanced model-based iterative approaches may also be used (Filzmoser, Hron, and Templ 2018). Imputing rounded and count zeroes is somewhat more complex, since they deal with a detection limit on the measuring equipment used. In these cases models based on linear regression may be used. The R-package robCompositions provides multiple methods for imputing missing values and zeros, mainly based on iterative regression models (Templ, Hron, and Filzmoser 2011).

2.3 Visual representation

It is useful to study data visually, but compositions are difficult to plot using ordinary techniques. In addition, it quickly becomes difficult to represent any data of more than three dimensions as a two-dimensional image. Since inferences made from subcompositions can be extended to the full composition, compositions are often plotted using several plots of three-part subcompositions.

Ternary plots can be used to visualize three-part compositions. They are similar to two-dimensional scatterplots, except that possible values are restricted to the simplex in order to represent three variables on two dimensions simultaneously. The basic ternary plot is an equilateral triangle, with each side acting as a scale for one of the three compositional parts. Each side has an arrow which indicates a corner for each part. The closer an observation is to a corner, the higher is the corresponding part's proportion of the whole composition. Likewise, observations close to the center have roughly equal proportions of the three parts. The scales on each side can be used to assess the proportions visually.

Figure 1 shows an example of a ternary plot. It is configured for compositions with three parts named x, y and z. Each part has been assigned a corner: x

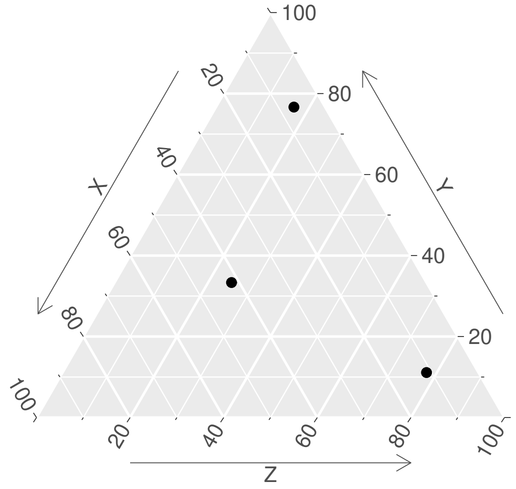

Figure 1: Example of a ternary plot, with three three-part compositions (x, y, z) represented by dots: (1,1,7) (Lower right), (5,4,3) (Center) and (20, 230, 50). (Top).

at the left, y at the top and z at the right. Each part also pertains to one of the sides as indicated by the arrows. Three (x, y, z) compositions have also been plotted on the graph: (1,1,7), (5,4,3) and (20, 230, 50). Composition (1, 1, 7) has a disproportionately large amount of z, and thus lies close to the right corner. In comparison, (5,4,3) has a more balanced ratio and sits near the center while (20, 230, 50) sits near y at the top. Despite a great difference in absolute values, (20, 230, 50) and (1, 1, 7) are approximately equally close to their respective corners. This illustrates the scale invariant nature of compositions.

Using the scales may seem slightly unintuitive at first. The compositional part that a particular scale belongs to can be determined by the corresponding labeled arrow. The orientation of the scale's numbers corresponds to a set of lines on the plot. Any composition on a particular line has the same proportion of the corresponding part, which is the value where the line intersects the scale. A good rule of thumb is that the lines get shorter as one moves up a scale. As an example of using the scale, we can read the scales to see that (20, 230, 50) is approximately 75% y (scale on the right, horizontal lines), 5% x (scale on the left, lines run downwards), and 15% z (scale at the bottom, lines run upwards). It can also be noted that each scale's 33% lines meet at exactly the center of the graph.

Based on the Aitchison geometry defined in Section 2.1, various geometric constructions can be added on the plot. Among these are straight lines, called compositional lines.

Definition 16 (Compositional lines)**.** *Compositional lines in the simplex S^D are* *defined as*

$$\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x}),$$

where $\mathbf{x}_0$ *is the starting point,* $\mathbf{x}$ *is the leading vector, and* α *is a real component.*

Two compositional lines $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *are parallel if they have the same leading vector, that is,*

$$\mathbf{y}_1 = \mathbf{x}_1 \oplus (\alpha_1 \odot \mathbf{x}),$$
$$\mathbf{y}_2 = \mathbf{x}_2 \oplus (\alpha_2 \odot \mathbf{x}).$$

Two compositional lines $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *are orthogonal if the Aitchison inner product of their leading vectors is zero, that is, for lines*

$$\mathbf{y}_1 = \mathbf{x}_0 \oplus (\alpha_1 \odot \mathbf{x}_1),$$
$$\mathbf{y}_2 = \mathbf{x}_0 \oplus (\alpha_2 \odot \mathbf{x}_2),$$

where $\mathbf{x}_0$ *is their intersection and* $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \mathbf{0}$*.*

Using parallel and orthogonal lines, grids can be drawn on ternary plots to aid in visualising compositional data. Parallel lines on the ternary plot do not appear parallel visually, which can make interpretation difficult.

In addition to straight lines, circles and ellipses can also be plotted on ternary plots. These can be used to visualize, for example, confidence intervals. It is often simpler to parameterise these figures using the coordinate system introduced in Section 2.2. Figure 2 shows an example of a confidence region on the simplex S^3 and real space $\mathbb{R}^2$, accompanied by observations and their coordinate transforms. Both regions are based on normal distributions: a regular multinormal distribution for the coordinates, and an equivalent normal distribution on the simplex as defined in Section 3.2 for the compositions. The non-regular shapes of the ternary plot are clearly visible.

Several operations can be applied to ternary plots to make interpretation easier. Centering shifts the observations by their mean so that they are centered around the center of the plot. Scaling shifts them so that unit variance is achieved, and observations are distributed more evenly along the entire plot. It should be noted that these operations can significantly transform the plot, in that the scales of the variables are no longer regularly spaced, and are no longer parallel to the sides of the triangle.

3 Statistical methods

3.1 Descriptive statistics

In order to analyse compositional data, it is useful to define descriptive statistics analogous to the mean and variance of regular random variables.

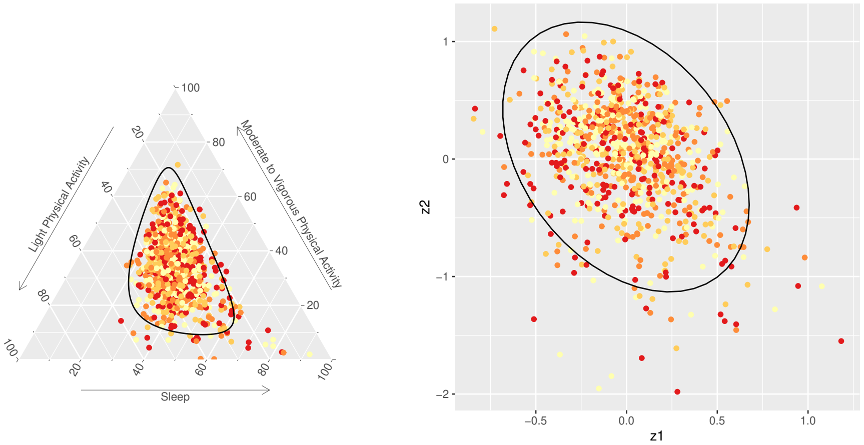

Figure 2: Example of a level 0.95 confidence region for the center of normally distributed compositional data, plotted on a ternary plot in S^3 with the compositional observations (left) and a scatterplot in $\mathbb{R}^2$ with transformed pivot coordinates (right). The coloring corresponds to the groupings used later in Section 4.

Definition 17 (Sample centre)**.** *The central tendency of a compositional data set* $\mathbf{X}$*, which has n observations of D-part compositions* $\mathbf{X}_i \in S^D$*, is defined as*

$$\operatorname{cen}(\mathbf{X}) = \hat{\mathbf{g}} = C[\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_D],$$

where $\hat{g}_j = (\prod_{i=1}^{n} X_{ij})^{1/n}$.

The sample centre corresponds to the equation $\operatorname{cen}(\mathbf{X}) = (1/n) \odot \bigoplus_{i=1}^{n} \mathbf{X}_i$ in the simplex space, which illustrates that the centre is the compositional equivalent of the arithmetic mean. The value of the centre will depend on the closure constant, which can be chosen based on the nature of the data.

Variability in the data set can be described with a variation matrix.

Definition 18 (Variation matrix)**.** *The variation matrix of a compositional data set* $\mathbf{X}$*, which has n observations of D-part compositions* $\mathbf{X}_i \in S^D$*, is defined as*

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1D} \\ t_{21} & t_{22} & \cdots & t_{2D} \\ \vdots & \vdots & & \vdots \\ t_{D1} & t_{D2} & \cdots & t_{DD} \end{bmatrix},$$

where t_{jk}, $j, k = 1, \ldots, D$ are the sample variances of pairwise logratios between compositional parts. That is,

$$t_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{jk}^i - \bar{z}_{jk})^2,$$

where $z_{jk}^i = \ln \frac{X_{ij}}{X_{ik}}$ *and* $\bar{z}_{jk} = \frac{1}{n} \sum_{i=1}^{n} z_{jk}^i$.

The variation matrix can be viewed as a measure of how consistent the proportion between two parts is over the entire dataset. It can also be seen to explain how variance is distributed among different logratios.

Definition 19 (Sample total variance)**.** *The global dispersion of a compositional data set* $\mathbf{X}$*, which has* n *observations of* D*-part compositions* $\mathbf{X}_i \in S^D$*, can be measured with its total variance, defined as*

$$\operatorname{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{j,k=1}^{D} t_{jk},$$

where t_{jk} *are elements of the variation matrix as defined in Definition 18.*

It is usually not sensible to calculate the total variance of non-compositional datasets, since they often contain variables which are measured in completely different scales, if they are comparable in the first place. However, since compositional data only carries relative information, all the components are measured on the same scale, i.e. the proportion of a single composition. This means that total variance has a natural interpretation as the average variance of a logratio. It can also be interpreted as the average squared Aitchison distance from the sample centre.

It can be noted that unlike the centre, neither the variation matrix nor the total variance depend on the closure constant, which is canceled out in the ratio.

3.2 Random compositions

There are two distributions which are commonly used to model random compositions, the normal distribution on the simplex, and the Dirichlet distribution. Of these, the normal distribution is more commonly used. The Dirichlet distribution is also popular but has more restrictions. Other existing distributions are generally modifications of the two distributions mentioned. Only the normal distribution will be used in this thesis.

The sample space of random compositions is the simplex. Since the simplex restricts the values of its vectors, some special rules are needed to describe distributions. There are two common approaches. In the conventional approach, the simplex is considered as a subspace of real space. In the compositional approach, the simplex is considered as an Euclidean space, for which coordinates are assigned from probability distributions. This is the approach used in this thesis.

The use of normal distribution on the simplex is based on representing the random composition in ilr coordinates, which are assumed to follow a multivariate normal distribution.

Definition 20 (Normal distribution on the simplex)**.** *If* $\mathbf{X}$ *is a random composition with sample space* S^D*, and the random orthonormal coordinates* $\mathbf{z} = \operatorname{ilr}(\mathbf{X})$ *follow a multivariate normal distribution on* $\mathbb{R}^{D-1}$ *where* $\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, then* $\mathbf{X}$ *is said to follow a normal distribution on* S^D*, denoted as* $\mathbf{X} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*.*

The parameters of the coordinate distribution, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, are also taken to be the parameters of the normal distribution on the simplex. While these parameters depend on the specific transformation used, this does not cause problems for estimation or analysis as long as the results are interpreted based on the specific transformation used. The parameter $\boldsymbol{\mu}$ is the mean of the coordinates, and the center of the random composition $\mathbf{X}$ is calculated as

$$\operatorname{cen}[\mathbf{X}] = \operatorname{ilr}^{-1}(\mathrm{E}[\operatorname{ilr}(\mathbf{X})]) = \operatorname{ilr}^{-1}(\boldsymbol{\mu}). \tag{5}$$

Since the distribution is parameterised with the parameters of the coordinates, several properties of the normal distribution translate as-is to the normal distribution on the simplex.

Lemma 3 (Transforming normally distributed compositions)**.** *Let* $\mathbf{X}$ *be a normally distributed random composition* $\mathbf{X} \in S^D$*,* $\mathbf{X} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Performing powering and pertubation on* $\mathbf{X}$ *with the real-valued constant* $b \in \mathbb{R}$ *and the composition* $\mathbf{a} \in S^D$ *produces the random composition* $\mathbf{Y} \in S^D$*, which is also normally distributed, i.e.*

$$\mathbf{Y} = \mathbf{a} \oplus (b \odot \mathbf{X}) \sim \mathcal{N}_S(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y).$$

The parameters of $\mathbf{Y}$ *can be derived from the original parameters, with*

$$\boldsymbol{\mu}_Y = \operatorname{ilr}(\mathbf{a}) + b\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}_Y = b^2\boldsymbol{\Sigma}.$$

Lemma 4 (Summation of normally distributed compositions)**.** *Let* $\mathbf{X}_k \in S^D$*,* $k = 1, \dots, n$*, be independently normally distributed compositions,* $\mathbf{X}_k \sim \mathcal{N}_S(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$*. The sum of these compositions is also a normally distributed composition* $\mathbf{Y} \in S^D$*,*

$$\mathbf{Y} = \bigoplus_{k=1}^{n} \mathbf{X}_k \sim \mathcal{N}_S(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y),$$

$$\boldsymbol{\mu}_Y = \sum_{k=1}^{n} \boldsymbol{\mu}_k, \qquad \boldsymbol{\Sigma}_Y = \sum_{k=1}^{n} \boldsymbol{\Sigma}_k.$$

One useful property relating to normal and multinormal distributions is the central limit theorem, which states that as the number of independent, identically-distributed random observations increases, their average approaches normality regardless of the true distribution of the samples. It can be shown that this theorem can also be applied to compositions through the coordinate transformation.

Theorem 1 (Central limit theorem for compositions)**.** *Let* $\mathbf{X}_i$*,* $i = 1, 2, \dots, n$*, be a sequence of independent random compositions that share the same center, that is* $\operatorname{cen}(\mathbf{X}_i) = \boldsymbol{\tau}$*. Furthermore, assume that the random ilr-coordinates* $\mathbf{z}_i = \operatorname{ilr}(\mathbf{X}_i)$ *share the same covariance matrix,* $\operatorname{cov}(\mathbf{z}_i) = \boldsymbol{\Sigma}$*. Let* $\bar{\mathbf{X}}$ *be the average of these compositions on the simplex,* $\bar{\mathbf{X}} = (1/n) \odot \bigoplus_{i=1}^{n} \mathbf{X}_i$*. The central limit theorem states that as* $n \to \infty$*, the distribution of the random composition* $\sqrt{n} \odot (\bar{\mathbf{X}} \ominus \boldsymbol{\tau})$ *converges to the normal distribution* $\mathcal{N}_S(\mathbf{0}, \boldsymbol{\Sigma})$*.*

3.3 Testing for normality

Testing the normality of compositional data can be based on the singular-value decomposition (SVD) of the ilr coordinates.

Definition 21 (Singular-value decomposition)**.** *Let* $\mathbf{Z}$ *be an* $n \times (D-1)$ *matrix of mean-centered ilr coordinates for compositional data. SVD decomposes the matrix into three parts,*

$$\mathbf{Z} = \mathbf{UDW}',$$

where $\mathbf{U}$ *is an* $n \times p$ *matrix with orthonormal columns holding the left singular vectors,* $\mathbf{D}$ *is a* $p \times p$ *diagonal matrix holding the positive singular values, and* $\mathbf{W}$ *is a* $(D-1) \times p$ *matrix with orthonormal columns holding the right singular vectors. The matrix* $\mathbf{U} = \mathbf{ZWD}^{-1}$ *contains normed and uncorrelated ilr coordinates.*

If $\mathbf{Z}$ is normally distributed, $\mathbf{U}$ follows (approximately) a $(D-1)$-variate standard multinormal distribution with independent components. Testing $\mathbf{Z}$ for normality can then be performed in two parts. First, the columns of $\mathbf{U}$ are tested for marginal normality with univariate tests. Any standard normality tests can be used. The second test focuses on the squared norms of the rows of $\mathbf{U}$,

$$\|\mathbf{u}_i\|^2 = \sum_{j=1}^{D-1} u_{ij}^2.$$

Since u_{ij} are independent standard normal variables under the assumption of normality, $\|\mathbf{u}_i\|^2$ follows a χ^2-distribution with $D-1$ degrees of freedom. Standard distributional tests, such as the Kolmogorov-Smirnov test, can again be used to test this. If the distributional tests indicate non-conformance, the normality of the data can be called into question.

3.4 Linear regression

Linear regression using compositions is in many ways similar to regular linear regression. The aim is to model linear relationships between a variable to be explained, or response, and one or more explanatory variables, or covariates. In the case of compositional regression, we can use the Aitchison geometry and normal distribution on the simplex to define regression models. The models may feature compositional responses, compositional covariates, or both. The different cases each require slightly different approaches. In this thesis we will focus on models with a compositional response and real-valued covariates. Other types of compositional regression models are discussed by Filzmoser et al. (Filzmoser, Hron, and Templ 2018, Chapter 10).

3.4.1 Regression with compositional response

Like in the real-valued case, compositional regression is based on treating the response variable as a normally distributed random variable. The mean of the response is assumed to be a linear combination of the covariates. The model can then be partitioned into a non-random systematic part and a random error part.

Definition 22 (Regression model with compositional response)**.** *Let* $\mathbf{x}$ *be a normally distributed composition on the simplex* S^D, *and let* $t_0, \cdots, t_r \in \mathbb{R}$ *be real-valued covariates. The mean of* $\mathbf{x}$ *is assumed to be linearly dependent on the covariates* t_k*.* $\mathbf{x}$ *can be written as*

$$\begin{aligned}\mathbf{x} &= (t_0 \odot \boldsymbol{\beta}_0) \oplus (t_1 \odot \boldsymbol{\beta}_1) \oplus \ldots \oplus (t_r \odot \boldsymbol{\beta}_r) \oplus \mathbf{e} \\ &= \bigoplus_{j=0}^{r} (t_j \odot \boldsymbol{\beta}_j) \oplus \mathbf{e},\end{aligned}$$

where $\boldsymbol{\beta}_j \in S^D$ *are compositional coefficients and* $\mathbf{e} \sim \mathcal{N}_S(\mathbf{0}, \boldsymbol{\Sigma})$ *is a normally distributed error term. Typically,* t_0 *is chosen to be 1, which means that* $\boldsymbol{\beta}_0$ *serves as an intercept term. The model is then written as*

$$\mathbf{x} = \boldsymbol{\beta}_0 \oplus \bigoplus_{j=1}^{r} (t_j \odot \boldsymbol{\beta}_j) \oplus \mathbf{e}. \tag{6}$$

Categorical covariates can be modeled by choosing one level of the variable as a reference level and giving the other levels their own coefficients. The corresponding covariates are dummy variables which have the value 1 *if the observation has the appropriate level of covariate and* 0 *otherwise.*

The model can equivalently be expressed in orthonormal coordinates as

$$\mathbf{x}^* = \boldsymbol{\beta}_0^* + \sum_{j=1}^{r} t_j \boldsymbol{\beta}_j^* + \mathbf{e}^*,$$

where $\mathbf{x}^*$, $\boldsymbol{\beta}_j^*$ and $\mathbf{e}^*$ are coordinates corresponding to the compositions in equation (6). To show this, we can recall equation (2), that is,

$$\operatorname{clr}(a \odot \mathbf{x}_1 \oplus b \odot \mathbf{x}_2) = a \cdot \operatorname{clr}(\mathbf{x}_1) + b \cdot \operatorname{clr}(\mathbf{x}_2).$$

Now, based on Definition 12 and equation (3) we can define the relation $\operatorname{ilr}(\mathbf{x}) = \operatorname{clr}(\mathbf{x})\mathbf{V}'$, and applying an ilr transform to the model equation

$$\begin{aligned}\operatorname{ilr}(\mathbf{x}) &= \operatorname{ilr}(\boldsymbol{\beta}_0 \oplus \bigoplus_{j=1}^{r} (t_j \odot \boldsymbol{\beta}_j) \oplus \mathbf{e}) \\ &= \operatorname{clr}(\boldsymbol{\beta}_0 \oplus \bigoplus_{j=1}^{r} (t_j \odot \boldsymbol{\beta}_j) \oplus \mathbf{e}) \cdot \mathbf{V}' \\ &= \left(\operatorname{clr}(\boldsymbol{\beta}_0) + \sum_{j=1}^{r} \operatorname{clr}(\boldsymbol{\beta}_j) t_j + \operatorname{clr}(\mathbf{e}) \right) \cdot \mathbf{V}' \\ &= \operatorname{ilr}(\boldsymbol{\beta}_0) + \sum_{j=1}^{r} \operatorname{ilr}(\boldsymbol{\beta}_j) t_j + \operatorname{ilr}(\mathbf{e}).\end{aligned}$$

This means that the compositional regression model directly corresponds to a multivariate regression model of orthonormal coordinates.

The aim of linear regression is to find estimates $\hat{\boldsymbol{\beta}}_k$ for the compositional coefficients $\boldsymbol{\beta}_k$ which best fit the data available. Based on the estimates, we can calculate fitted values $\hat{\mathbf{x}} = \hat{\boldsymbol{\beta}}_0 \oplus \bigoplus_{j=1}^{r}(t_j \odot \hat{\boldsymbol{\beta}}_k)$, as well as residuals $\hat{\mathbf{e}} = \mathbf{x} \ominus \hat{\mathbf{x}}$. Fitting the model is typically done by finding values which minimize the sum of the squared norms of the residuals, called the residual sum of squares (RSS),

$$\mathrm{RSS} = \sum_{i=1}^{n} \|\hat{\mathbf{e}}_i\|_a^2.$$

This is known as the least-squares method, corresponding to least-squares regression in the real-valued case. The actual estimation is performed in the coordinate space, where most common regression methods can be applied as-is.

Transformed into coordinate space, the RSS becomes

$$\mathrm{RSS} = \sum_{i=1}^{n} \|\hat{\mathbf{e}}_i^*\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{D-1}(\hat{e}_{ij}^*)^2,$$

with the Aitchison norms becoming Euclidian norms. The order of sums in the equation can be reversed, and since the terms are all non-negative, maximizing the sum can be reduced to maximizing the $D-1$ sums

$$\mathrm{RSS}_j = \sum_{i=1}^{n}(\hat{e}_{ij}^*)^2,$$

where RSS_j is the sum corresponding to the jth coordinate. These problems can be estimated independently, which means that the compositional regression model is solved by fitting $D-1$ real regression models, one for each coordinate. For the jth coordinate, the model to be estimated is

$$x_{ij}^* = \sum_{k=0}^{r} t_{ik}\beta_{kj}^* + e_{ij}^*,$$

with x_{ij}^*, β_{kj}^* and e_{ij}^* are the jth coordinates of the corresponding coordinate vectors. Statistics such as the t- and F-statistics are usable as-is with coordinate estimates.

The fitted values and residuals produced by the model do not depend on the selected coordinate basis, as long as they are transformed back into the simplex space with the appropriate inverse transformation. However, using established methods for statistical inference requires using the coordinate values. It is important to choose interpretable coordinate transformation to make analysis easier.

3.4.2 Within-composition regression

Frequently, in addition to using compositions as a response it may be desirable to study how the various parts of a composition affect each other. For example, one might wish to study how increasing spending in food expenses affects spending in other areas. As with normal regression, coordinate representations are used to study the compositions in real space. The basic form of within-composition regression involves choosing one part as the response variable, and then constructing pivot coordinates with the response as a pivot. The regression model is then built with the first coordinate as the response, and the rest as covariates.

Compared to regular regression, using coordinates in this manner is somewhat tricky. With pivot coordinates, all the relative information about the pivot variable is contained in the first coordinate. This makes the first coordinate a suitable choice for the response variable. The remaining coordinates, however, contain overlapping information about the covariates. This means that one cannot consider several coordinates as covariates in the same model. If one wishes to analyse the effect of several parts, a separate regression model can be formed for each covariate. For each model, the coordinates are formed so that the response part corresponds to the first coordinate, and the covariate part to the second coordinate. The other coordinates are also included in the model, since they contain information about the whole composition, but they are not of primary interest.

Definition 23 (Within-composition regression)**.** *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ *be a D-part composition,* $\mathbf{x} \in S^D$. *We can now choose two parts which are of interest, the response* x_l *and a covariate* x_k, *and define a reordered composition* $\mathbf{x}^{(lk)} = (x_1^{(lk)}, x_2^{(lk)}, \ldots, x_D^{(lk)}) = (x_l, x_k, \ldots, x_i, \ldots, x_D)$, *with* $i \neq l, k$. *The first two parts of the reordered composition are therefore the two parts of interest, i.e* $x_1^{(lk)} = x_l$ *and* $x_2^{(lk)} = x_k$ *. Slightly modifying the notation from Section 2.2.1, we can now define pivot coordinates based on this composition,*

$$z_j^{(lk)} = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_j^{(lk)}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k^{(lk)}}} \right).$$

$z_1^{(lk)}$ *once again contains all relative information about* x_l, *and* $z_2^{(lk)}$ *contains the relative information of* x_k *compared to all other non-response parts. Now, for each response-covariate pair we wish to examine, we can define a linear model of the form*

$$z_1^{(lk)} = b_1^{(lk)} + b_2^{(lk)} z_2^{(lk)} + \cdots + b_{D-1}^{(lk)} z_{D-1}^{(lk)} + \epsilon.$$

Each model is used to analyse a single explanatory part, meaning that only the intercept $b_1^{(lk)}$ *and coefficient* $b_2^{(lk)}$ *are used for statistics, p-values and such. Interpreting the coefficients is again not as straightforward as with regular variables, since they represent the proportional strengths of each part in the composition. The response variable is interpreted as the dominance of the part in question, or as its proportion of the entire composition. The explanatory variable is then interpreted as the proportion of the explanatory part compared to the rest of the non-response parts.*

Due to the nature of compositions, the response and the covariates are related in a manner that violates the assumptions of a normal regression model. Specifically, the covariates are usually assumed to be errorless, whereas with compositions any error in the response part is also reflected in the other parts. This problem can be rectified by using orthogonal regression models, or total least squares regression. In this case, however, strict distributional assumptions are required to make statistical inferences. This can be avoided by using resampling methods such as bootstrapping to estimate the sampling distribution of the statistics without making assumptions about the distribution of the data.

4 REACT dataset

The REACT controlled randomized trial is run by the public health division department of the Department of Clinical Medicine of the University of Turku (Leskinen et al. 2021). The REACT dataset comprises activity data for 231 participants, gathered over a period of twelve months. The participants were chosen on the basis of their having retired between the years 2016 and 2019. The average age of the participants was 65, and their ages ranged from 62 to 67 years. The participants were randomly divided into two groups, a control group and an intervention group. Participants in the intervention group were given activity trackers to be worn for one year, while the control group was left to live their life as normal. Activity data from the participants were gathered 0, 3, 6 and 12 months after they had started the study. Data were gathered by having each participant wear a wrist-mounted accelerometer for one week during each measurement point.

The raw acceleration data produced by the measurements are not directly usable in research. The data were preprocessed using several algorithms to produce measurements of daily activity levels and sleep times for each subject. The data were also supplemented with diaries kept by the subjects to gain accurate information about sleeping times. The daily movement behaviour of each participant was then partitioned into four categories: Sleep, sedentary behaviour, light physical activity (LPA) and moderate or vigorous physical activity (MVPA).

People typically maintain a certain level of activity for some time, during which the precise level of movement can vary significantly. This means that the data have to be divided into periods of activity based on the average amount of movement during the period. The accelerometer data was processed to find these periods, or bouts, using the open source R-package GGIR, version 1.7-1 (Migueles et al. 2019).

The daily activity bouts gathered over the week were combined to form an average representation of daily activity for each subject. The final dataset consisted of 231 entries, each containing four observations on the amount of minutes spent on sedentary time, light, moderate and vigorous activity, or sleep over a day. These observations form a set of compositional data. The dataset contained several zeroes and missing values which were imputed using methods provided by the R-package robCompositions, mentioned in Section 2.2.3.

The average duration of each measurement was approximately 1420 minutes, which is slightly less than 24 hours. The differences in length were caused by differing sleeping patterns, the time the accelerometer was started or removed, or other reasons. Measurements also differed somewhat in the number of days that data were collected. Aside from being averaged out, these considerations do not factor into compositional data analysis.

Before beginning the actual analysis, it is useful to study the overall compositional nature of the REACT data. Figure 3 shows the data as four ternary plots, one for each possible three-dimensional subcomposition. The observations are colored based on the measurement time. The graphs show that observations from all measurement times are grouped closely together, suggesting that there are no large differences between the different measurement times. The way the data are distributed on the graphs tells about the relative dominance of each activity type.

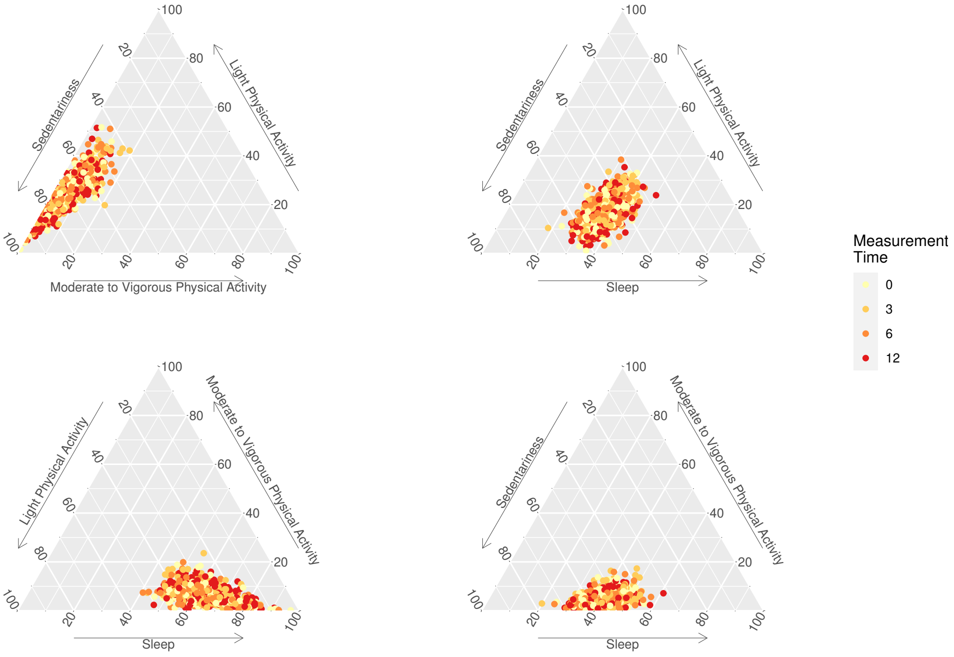

Figure 3: Ternary plots for different combinations of activity types, showing all subjects, colored by time.

Sleep and sedentariness dominate, with the measurements being concentrated near their corners on the left-hand graphs. On the right it can be seen that they take up roughly an equal amount of the entire composition, since the measurements lay roughly halfway between them. Light activity is more common than moderate-to-vigorous activity.

Figure 4 shows the same graph configuration as Figure 3. Now, the individual observations have been replaced with confidence intervals of the centers of the different groups, which are separated between time as well as bracelet use. The top right plot shows neatly overlapping regions, indicating that sedentariness, sleep or light activity do not greatly differ due to intervention. On the other plots, however, we can see that the confidence regions for the intervention group are skewed towards moderate-to-vigorous activity compared to non-users. This seems to indicate that activity trackers have the effect of increasing exercise.

5 Analysis

Table 2 shows the compositional means of each measurement, as defined in Definition 17, for both the control and intervention groups. The closure constant was chosen to be 1440, which is equal to the number of minutes in a 24-hour day. The values indicate the same general proportions as the earlier figures, with sleep and sedentary behaviour making up the bulk of the compositions, with smaller proportions of light and very small proportions of moderate or vigorous activity. On

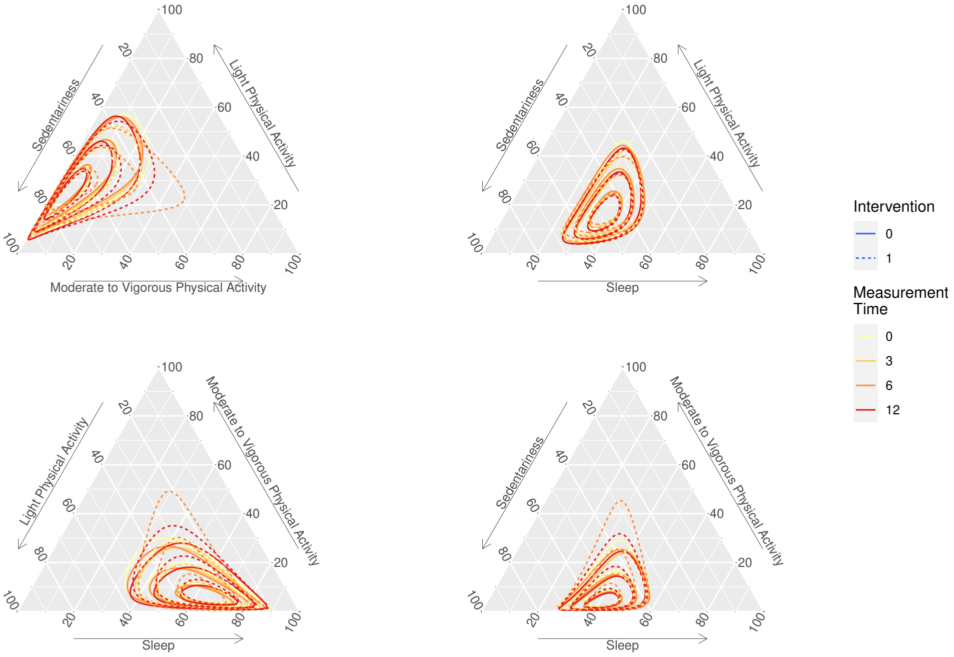

Figure 4: The same plots as in Figure 3, with 99%, 90% and 50% confidence regions based on the assumption of normally distributed data. Bracelet and non-bracelet subjects are differentiated by linetype.

Table 2: Compositional means of the average measurements over a week. The closure constant is 1440, representing minutes in a 24-hour day. The Time-column indicates how many months after starting the study the measurements took place.

	No tracker				Tracker			
Time	Sed.	LPA	MVPA	Sleep	Sed.	LPA	MVPA	Sleep
0	674.69	218.68	42.44	504.20	673.50	222.97	50.30	493.23
3	659.10	234.44	44.03	502.44	669.04	231.01	50.62	489.33
6	664.30	234.96	42.62	498.12	656.26	245.76	45.67	492.31
12	684.45	216.30	39.87	499.39	680.28	210.79	46.48	502.45

Table 3: Compositional variation matrices for different measurements. Sed.: Sedentariness. LPA: Light Physical Activity. MVPA: Moderate to Vigorous Physical Activity.

	Sed.	LPA	MVPA	Sleep
T0				
Sedentariness	0.00	0.14	0.34	0.01
Light Physical Activity	0.14	0.00	0.24	0.14
Moderate to Vigorous Physical Activity	0.34	0.24	0.00	0.36
Sleep	0.01	0.14	0.36	0.00
T3				
Sedentariness	0.00	0.17	0.39	0.01
Light Physical Activity	0.17	0.00	0.25	0.17
Moderate to Vigorous Physical Activity	0.39	0.25	0.00	0.40
Sleep	0.01	0.17	0.40	0.00
T6				
Sedentariness	0.00	0.13	0.32	0.01
Light Physical Activity	0.13	0.00	0.19	0.12
Moderate to Vigorous Physical Activity	0.32	0.19	0.00	0.33
Sleep	0.01	0.12	0.33	0.00
T12				
Sedentariness	0.00	0.18	0.39	0.01
Light Physical Activity	0.18	0.00	0.26	0.18
Moderate to Vigorous Physical Activity	0.39	0.26	0.00	0.41
Sleep	0.01	0.18	0.41	0.00

average, the participants spent around 11 hours per day sedentary, around 8 hours sleeping, around 4 hours on light activity and less than an hour on more demanding activity. While no major patterns are apparent, after 12 months of follow-up in both groups of subjects, sedentariness appears to have increased and LPA decreased for both groups of subjects, and the intervention group again appears to generally have a higher proportion of MVPA than the control group.

Table 3 shows the compositional variation matrices for different measurement times. Again, the change over time seems negligible. As seen in Definition 18, the variation matrix is based on the variances of logratios instead of the more common definitions of covariance. MVPA tends to have the highest covariances, which suggests that it varies the most among the different activities. This is consistent with the confidence regions seen in Figure 4.

5.1 Regression with a compositional response

Since the REACT study was designed as an intervention study, it is appropriate to begin the analysis with a model that contains only the intervention variable. This will allow us to study the overall effect of the intervention. As described in Section 3.4, we will be fitting models using different coordinate transformations

Table 4: Estimated regression models for coordinates from a pivot transform with moderate-to-heavy physical activity as the pivot variable. The only explanatory variable is the intervention. Coordinate z1 contains moderate-to-heavy activity contrasted against light activity, sleep and sedentariness. Coordinate z2 contains light activity, sleep and sedentariness, and coordinate z3 contains light activity and sleep.

Coordinate	Term	Estimate	Std. Error	t-statistic	p-value
z1	(Intercept)	-2.00	0.03	-70.50	0.00
	Intervention : Yes	0.12	0.04	2.93	0.00
z2	(Intercept)	0.56	0.01	43.42	0.00
	Intervention : Yes	0.00	0.02	0.11	0.91
z3	(Intercept)	-0.56	0.01	-41.96	0.00
	Intervention : Yes	0.01	0.02	0.72	0.47

of the compositional data. The model for the transformed coordinates of a single observation is

$$\mathbf{z} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{int} + \boldsymbol{\epsilon},$$

where $\mathbf{z}$ is a vector of transformed coordinates, x_{int} is the categorical variable representing intervention, and $\boldsymbol{\epsilon}$ is a normally distributed error term.

The choice of the coordinate transformation should be determined by the goals of the analysis. In this case, we are only interested in the effect of the intervention. We can refer to Figure 4, noting that MVPA is the only compositional part whose proportion seems to visibly differ based on the intervention. Therefore, we will start by fitting a model for pivot coordinates obtained by selecting MVPA as the pivot variable.

Table 4 shows the models fitted for each coordinate of the pivot transformation. As expected, the intervention has a significant effect on the first coordinate. As described in Section 2.2.1, the first pivot coordinate contains all the relative information of the part used as the pivot, and as such reflects the part's proportion of the whole composition. This indicates that there is a difference between the proportions of MVPA between the control and intervention groups.

In addition to the intervention, there are several explanatory variables available, such as age and sex. To begin, we can study how these variables affect the composition of activities by generating four sets of pivot coordinates, one for each part of the composition. Each set of coordinates has a different compositional part as the pivot. By fitting a separate regression model for the first coordinate of each coordinate transform, we can determine which explanatory variables affect which parts. Since the models are based on different transformations, they cannot be directly compared with each other, but they may be informative on how the explanatory variables affect the composition of daily activities.

Table 5: Separate regression models for the first coordinate of four different sets of pivot coordinates.

Compositional part	Term	Estimate	Std. Error	t-statistic	p-value
Sedentariness	(Intercept)	1.25	0.04	29.24	0.00
	Intervention : Yes	-0.04	0.03	-1.44	0.15
	Measurement time: 3	-0.03	0.04	-0.88	0.38
	Measurement time: 6	-0.03	0.04	-0.73	0.47
	Measurement time: 12	0.04	0.04	1.06	0.29
	Sex: Female	-0.10	0.03	-3.04	0.00
	Age: Over 65	0.07	0.03	2.54	0.01
Light physical activity	(Intercept)	-0.22	0.03	-6.99	0.00
	Intervention : Yes	-0.03	0.02	-1.31	0.19
	Measurement time: 3	0.05	0.03	1.64	0.10
	Measurement time: 6	0.09	0.03	3.42	0.00
	Measurement time: 12	-0.01	0.03	-0.48	0.63
	Sex: Female	0.13	0.03	4.94	0.00
	Age: Over 65	0.04	0.02	2.14	0.03
Moderate physical activity	(Intercept)	-1.88	0.06	-28.97	0.00
	Intervention : Yes	0.11	0.04	2.87	0.00
	Measurement time: 3	0.01	0.06	0.17	0.87
	Measurement time: 6	-0.06	0.06	-0.99	0.32
	Measurement time: 12	-0.06	0.06	-1.00	0.32
	Sex: Female	-0.03	0.05	-0.63	0.53
	Age: Over 65	-0.11	0.04	-2.62	0.01
Sleep	(Intercept)	0.85	0.03	27.54	0.00
	Intervention : Yes	-0.05	0.02	-2.68	0.01
	Measurement time: 3	-0.02	0.03	-0.81	0.42
	Measurement time: 6	-0.01	0.03	-0.42	0.68
	Measurement time: 12	0.03	0.03	1.12	0.26
	Sex: Female	0.01	0.03	0.45	0.65
	Age: Over 65	0.00	0.02	-0.19	0.85

The initial model for the first coordinate of each transformation is

$$z_1 = \beta_0 + \beta_1 x_{int} + \beta_2 x_{sex} + \beta_3 x_{time} + \beta_4 x_{age} + \epsilon,$$

where the explanatory variables x_i are all categorical, and ϵ is a normally distributed error term. Table 5 shows the results of the four regression models fitted on the first coordinates of the pivot transformations. Each model gives a general impression of how the compositional part in question is affected by the specific explanatory variables. Disregarding the intercepts, we can first note that the intervention, i.e. the presence of an activity tracker, has a positive effect on moderate to vigorous physical activity, and a negative effect on other types of activities. This indicates that the intervention group is more inclined towards moderate and heavy activity than the control group, at the expense of other activity types. The effects of intervention on MVPA and sleep are also statistically significant based on the F-test, with $p < 0.05$.

The time since starting the study, measured in months, does not appear to have an effect on most activities. The exception is light activity, where there seems to be a significant increase compared to the beginning at the six month mark. Sex has a significant effect on sedentariness and light activity, with women being more active. Age significantly affects all parts but sleep, with those over 65 showing less moderate activity and more lighter activity and sedentariness. Based on these preliminary findings it would seem appropriate to include all variables, aside from time, as the analysis continues. We will still include time for the sake of completeness, as well as the possibility that it has interesting interactions with the other variables.

Next, we will fit a model based on balances, as defined in Section 2.2.2. In this case, since we are interested in changes in activity, the first binary partition can be based on contrasting light and moderate-to-heavy activity with sleep and sedentary behaviour. The partitions corresponding to the balance coordinates are depicted in Table 6. Coordinate 1 contrasts sleep and sedentary behaviour with physical activity, and the second and third coordinates contrast the members of these two pairs with each other. As with the pivot coordinates, a model is fitted for each coordinate separately.

Coordinate	Sed.	LPA	MVPA	Sleep
1	-	+	+	-
2	0	+	-	0
3	+	0	0	-

Table 6: Binary partitioning for the REACT dataset. Partition 1: Physical activity vs. Sleep and Sedentary behaviour. Partition 2: LPA vs. MVPA. Partition 3: Sedentary behaviour vs. Sleep.

The model again uses all explanatory variables, that is

$$\mathbf{z} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{int} + \boldsymbol{\beta}_2 x_{sex} + \boldsymbol{\beta}_3 x_{time} + \boldsymbol{\beta}_4 x_{age} + \boldsymbol{\epsilon},$$

where the explanatory variables are categorical and $\boldsymbol{\epsilon}$ is a multinormal error term. The estimation results can be seen in Table 7.

Table 7: Regression models for three balance coordinates. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

Coordinate	Term	Estimate	Std. Error	t-statistic	p-value
z1	(Intercept)	-1.82	0.06	-30.38	0.00
	Intervention : Yes	0.08	0.04	2.09	0.04
	Measurement time: 3	0.05	0.05	0.91	0.36
	Measurement time: 6	0.03	0.05	0.63	0.53
	Measurement time: 12	-0.06	0.05	-1.16	0.25
	Sex: Female	0.08	0.05	1.67	0.09
	Age: Over 65	-0.06	0.04	-1.48	0.14
z2	(Intercept)	1.02	0.05	22.08	0.00
	Intervention : Yes	-0.09	0.03	-3.03	0.00
	Measurement time: 3	0.02	0.04	0.55	0.59
	Measurement time: 6	0.09	0.04	2.30	0.02
	Measurement time: 12	0.03	0.04	0.66	0.51
	Sex: Female	0.10	0.04	2.62	0.01
	Age: Over 65	0.09	0.03	3.17	0.00
z3	(Intercept)	0.24	0.02	14.13	0.00
	Intervention : Yes	0.01	0.01	0.78	0.44
	Measurement time: 3	-0.01	0.01	-0.44	0.66
	Measurement time: 6	-0.01	0.01	-0.65	0.52
	Measurement time: 12	0.01	0.01	0.37	0.71
	Sex: Female	-0.07	0.01	-5.15	0.00
	Age: Over 65	0.04	0.01	4.09	0.00

Table 8: Statistical significance of different terms in three separate regression models for balance coordinates. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

	Coordinates		
Term	z1	z2	z3
(Intercept)	0.00	0.00	0.00
Intervention: Yes	0.00	0.69	0.36
Age: Over 65	0.72	0.06	0.91
Measurement time: 3	0.29	0.23	0.39
Measurement time: 6	0.39	0.17	0.27
Measurement time: 12	0.87	0.42	0.90
Sex: Female	0.85	0.00	0.04
Intervention : Age	0.00	0.80	0.03
Intervention : Time 3	0.69	0.95	0.60
Intervention : Time 6	0.70	0.28	0.66
Intervention : Time 12	0.67	0.79	0.46
Intervention : Sex	0.23	0.38	0.50
Age : Sex	0.05	0.07	0.43
Age : Time 3	0.19	0.47	0.44
Age : Time 6	0.08	0.73	0.09
Age : Time 12	0.11	0.37	0.05
Sex : Time 3	0.99	0.10	0.83
Sex : Time 6	0.72	0.16	0.70
Sex : Time 12	0.44	0.27	0.46

Based on the balance-based regression models, it can again be seen that the intervention has a statistically significant effect on activity. Those with activity trackers spent on average more time on physical activitities than the control group, and additionally had a larger percentage of their physical activity be moderate or vigorous. This agrees with the earlier models and visual examinations. Out of the other explanatory variables, being female seems to increase the proportion of physical activity comparered to sedentary behaviour, the proportion of light activity compared to heavier activity, and the proportion of sleep compared to sedentariness. Being over 65 on the other hand indicates an increased proportion of sedentary behaviour compared to physical activity, increased proportion of light activity over heavier activities, and an increased proportion of sedentariness compared to sleep. It should be noted that since the dataset only includes subjects approximately between 60 and 70 years of age, the results do not necessarily translate directly to subjects younger or older than those.

With several categorical variables, it is possible that there are multiple interaction terms that could improve our models. Table 8 shows the p-values for the terms in models that have been saturated with every possible two-way interaction. The coordinates are the same balance coordinates as in the previous model. A few of

Table 9: Fitted regression models for three balance coordinates with interaction terms. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

Coordinate	Term	Estimate	Std. Error	t-statistic	p-value
z1	(Intercept)	-1.82	0.07	-27.15	0.00
	Intervention: Yes	0.23	0.06	3.95	0.00
	Age: Over 65	-0.09	0.10	-0.95	0.34
	Sex: Female	-0.01	0.07	-0.08	0.94
	Intervention : Age	-0.26	0.07	-3.51	0.00
	Age : Sex	0.20	0.10	2.10	0.04
z2	(Intercept)	1.00	0.05	19.13	0.00
	Intervention: Yes	-0.07	0.04	-1.60	0.11
	Age: Over 65	0.21	0.07	2.79	0.01
	Sex: Female	0.16	0.05	3.09	0.00
	Intervention : Age	-0.02	0.06	-0.36	0.72
	Age : Sex	-0.13	0.07	-1.71	0.09
z3	(Intercept)	0.25	0.02	12.79	0.00
	Intervention: Yes	-0.02	0.02	-1.26	0.21
	Age: Over 65	0.04	0.03	1.43	0.15
	Sex: Female	-0.06	0.02	-3.21	0.00
	Intervention : Age	0.05	0.02	2.35	0.02
	Age : Sex	-0.02	0.03	-0.88	0.38

the interactions appear significant, namely those between the intervention and age, and age and sex. None of the interactions with time are especially significant, which justifies leaving them out of further analyses.

Leaving in only the signicant terms of the saturated model, as well as age and sex, the model we are left with is

$$\mathbf{z} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{int} + \boldsymbol{\beta}_2 x_{sex} + \boldsymbol{\beta}_3 x_{age} + \boldsymbol{\beta}_4 x_{int} x_{age} + \boldsymbol{\beta}_5 x_{sex} x_{age} + \boldsymbol{\epsilon}.$$

The results can be seen in Table 9. Intervention and its interaction with age significantly affect the first coordinate, but in opposite directions. While those with a tracker are more inclined toward exercise, being older than 65 cancels this out completely. The interaction between age and sex also positively affects the first coordinate, meaning older women are more likely to exercise.

Age by itself is only significant with regard to the second coordinate, with older people being more inclined towards lighter activity. Sex on the other hand is siginificant to both the secondary coordinates, indicating that women prefer lighter activity and sleep more compared to men.

The normality of the data can be tested with various standard distributional tests. We can begin with some visual inspections, using the same balance coordinates as earlier. Figures 5 and 6 show histograms and normal Q-Q plots for the

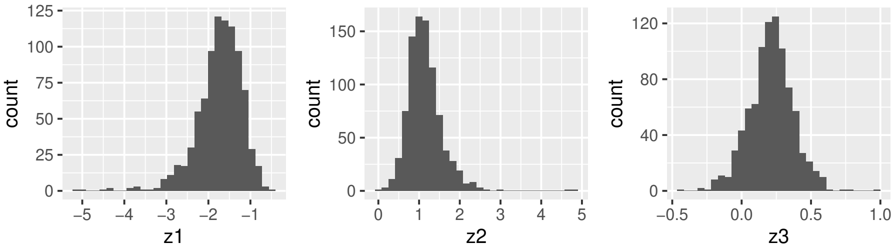

Figure 5: Histograms for balance coordinates. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

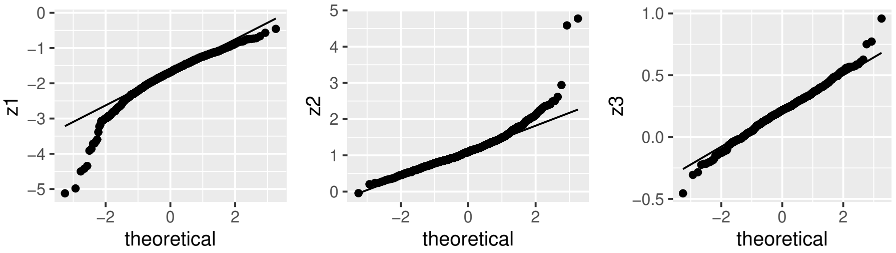

Figure 6: Normal Q-Q-plots for balance coordinates. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

balances. The histograms seem to indicate normal distributions, although the first two coordinates seem to have elongated tails. This is also reflected in the quantile plots, which show the first and second coordinates dipping quite far away from the theoretical line. This could indicate that the data are not normally distributed.

We can also perform a battery of formal Anderson-Darling normality tests for ilr-transformed coordinates. After transforming, univariate normality tests are performed for each individual coordinate, and multivariate tests are performed for combinations of the coordinates. A robust method for performing these tests is provided by the R-package robCompositions (Templ, Hron, and Filzmoser 2011). Other ways of testing normality are discussed in Section 3.3.

Based on the tests, it seems that the normality of the data can't be assumed. Each test has an extremely low p-value ($p < 0.01$) , which means that the null hypothesis of normally distributed data must be rejected. The results agree with different tests or different coordinate transformations, such as the Shapiro-Wilk test performed on balances.

While failing the normality tests may indicate that approaches such as linear models are not appropriate, we can still justify them with asymptotic arguments. As stated in Theorem 1, asymptotic normality holds for the averages of random compositions at sufficiently large sample sizes, implying that the inferences based on the previous models can be seen to be at least approximately valid.

Table 10: Point estimates for the balance coordinates of different groups. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

		Control			Intervention		
Sex	Age	z1	z2	z3	z1	z2	z3
Male	Under 65	-1.82	1.00	0.25	-1.60	1.00	0.25
	Over 65	-1.92	1.20	0.29	-1.95	1.18	0.34
Female	Under 65	-1.83	1.16	0.18	-1.60	1.16	0.18
	Over 65	-1.72	1.23	0.20	-1.75	1.21	0.25

Since the final model uses only categorical variables, we can combine the estimated coefficients to form estimates for each combination of categories. These are presented in Table 10. Level 0.95 confidence intervals for the estimates are presented in Table 11. The same general conclusions as before can be drawn. The table especially makes it easy to see that the intervention positively affects the ratio of physical activity and sedentariness, but only among those who are under 65. In the control group, younger men are more inclined to physical activity than older men, but the reverse is true for women. In the intervention group, both sexes are more inclined to physical activity when under 65. Men are more inclined to heavier activity than women, and as both get older, the proportion of light activity compared to heavier activity increases.

As noted in Definition 13, it is simple to transform a set of ilr coordinates back into compositional form. Since we have estimated the sets of coordinates for the different groups, we can also transform these into estimated average compositions. These are found in Table 12. The compositions have again been closed to 1440 minutes, or 24 hours. The compositions yield largely the same conclusions as obtained from Table 10, but in a quantitative form. For example, for participants under 65 the activity tracker seems to be connected to a roughly 40 minute increase in light activity over the course of the day.

5.2 Within-composition regression

In addition to the background variables, we can study how the different activity types affect each other by using within-composition regression. For instance, we may choose to study how the time spent while sedentary is affected by the time spent on other activities. For this, we can form three models, one for each activity type to be used as an explanatory variable. Pivot coordinates are formed for each model so that sedentariness is the pivot variable and the covariate being studied is the second variable. Each model has the form

$$z_1 = \beta_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon,$$

where z_1 is the first pivot coordinate, z_2 is the second coordinate, which contains the covariate being studied, and z_3 is the third coordinate, containing the rest of

Table 11: Level 0.95 confidence intervals for the point estimates of balance coordinates for different groups. z1: Physical activity against sedentariness and sleep. z2: Light physical activity against moderate and vigorous activity. z3: Sedentary behaviour against sleep.

		z1		z2		z3	
Sex	Age	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Control							
Male	Under 65	-1.96	-1.69	0.89	1.10	0.21	0.28
	Over 65	-2.23	-1.60	0.95	1.45	0.19	0.38
Female	Under 65	-2.09	-1.57	0.95	1.36	0.11	0.26
	Over 65	-2.36	-1.08	0.74	1.73	0.02	0.38
Intervention							
Male	Under 65	-1.84	-1.36	0.89	1.10	0.21	0.28
	Over 65	-2.53	-1.37	0.82	1.54	0.20	0.47
Female	Under 65	-1.98	-1.23	0.95	1.36	0.11	0.26
	Over 65	-2.65	-0.86	0.60	1.82	0.02	0.47

Table 12: Estimated average compositions for movement behaviours during 24 hours, obtained from balance coordinates via an inverse transform. The closure constant is set to 1440 (minutes). Sed.: Sedentariness. LPA: Light Physical Activity. MVPA: Moderate to Vigorous Physical Activity.

		Control				Intervention			
Sex	Age	Sed.	LPA	MVPA	Sleep	Sed.	LPA	MVPA	Sleep
Male	Under 65	703.52	192.69	47.15	496.65	675.06	231.69	56.69	476.56
	Over 65	719.57	202.66	37.04	480.73	746.38	192.81	36.31	464.51
Female	Under 65	669.47	213.28	41.62	515.63	640.76	255.80	49.92	493.52
	Over 65	656.84	244.89	42.77	495.51	683.68	233.80	42.06	480.46

Table 13: Fitted models for within-composition regression. Each model has been formed by creating a pivot transform with sedentariness as the pivot variable. The models were then manipulated so that the second coordinate contains all relative information about the explanatory part being studied, and the coordinates were used as explanatory variables for the model.

Explanatory part	Term	Estimate	Std. Error	t-statistic	p-value
	(Intercept)	0.10	0.02	4.81	0.0
Light physical activity	z2	-0.44	0.02	-18.45	0.0
	z3	-0.72	0.01	-54.39	0.0
	(Intercept)	0.10	0.02	4.81	0.0
Moderate-to-Vigorous physical activity	z2	-0.40	0.01	-31.44	0.0
	z3	-0.74	0.02	-30.85	0.0
	(Intercept)	0.10	0.02	4.81	0.0
Sleep	z2	0.84	0.02	43.23	0.0
	z3	-0.02	0.02	-1.04	0.3

the relative information.

The summaries of the fitted models are shown in Table 13. As explained in Section 3.4.2, only the z_2 parameter should be considered meaningful. It can be seen that sleep has a positive coefficient, whereas the two physical activity classes have negative ones. This means that as the proportion of sleep among the three explanatory parts increases, the proportion of sedentariness compared to the other three types also increases. On the other hand, increasing the amount of physical activity leads to the proportion of sedentariness decreasing. Sleep seems to have the strongest effect, and MVPA the weakest. The results seem to indicate that replacing sleep with physical activity leads to a decreasing amount of sedentariness.

6 Discussion

Based on just the activity data, it seems that using an activity bracelet does indeed have a positive effect on physical activity. This is not particularly surprising, given that this is what the bracelets were designed for. Other factors such as sex and age also had some effect, which may indicate starting points for further research. Since the participants of the study were relatively aged, it might be worthwile extending the study to younger people.

Compositional models seem to fit analysis of movement behaviours quite well. It might also be worth considering what other areas of public health could benefit from compositional analysis. Aspects of daily life such as food consumption could easily be adapted into compositional form, and it is easy to imagine that analysing them could lead to interesting results. On the other hand, as technology for measuring things such as brain patterns or body functions continue to advance, compositional analysis might become relevant in analysing the data produced.

The models used to analyse the data were basic linear models. Due to the questionably normality of the data, it might be worthwile to analyse them with more advanced models that could additionally take into account possible unusual features. For example, individuals may differ in how they retain the increased activity level after receiving a bracelet. Furthermore, in the models used the time points were treated as independent. In reality, each individual's measurements were probably heavily dependent on their first measurement. Taking this type of dependency into account requires using techniques such as random effect models. Since random compositions are largely based around the regular normal distribution, there are no major theoretical obstacles to applying longitudinal analysis techniques to compositional data, and mixed-effect models have been used successfully in the past (Wang, Wang, and Wang 2019).

While this thesis covers most of the basic theory necessary for analysing compositional data, many interesting applications were left out due to space constraints or for falling outside the scope of the analysis. These include topics such as compositional cluster analysis, compositional correlation analysis, compositional processes and so on. The field is also constantly undergoing new developments, so it is worthwile to keep up-to-date on the latest research.

Appendix A R packages used in the thesis

Package	**Version**	**Reference**
haven	2.3.1	Wickham and Miller 2020
dplyr	1.0.2	Wickham et al. 2020
robCompositions	2.3.0	Templ, Hron, and Filzmoser 2011
compositions	2.0-0	van den Boogaart, Tolosana-Delgado, and Bren 2020
ggplot2	3.3.3	Wickham 2016
ggtern	3.3.0	Hamilton and Ferry 2018
knitr	1.30	Xie 2020
kableExtra	1.3.1	Zhu 2020

Table A1: A list of R packages used for handling, analysing and visualising the data.

References

Aitchison, John. 1986. *The Statistical Analysis Of Compositional Data.* Chapman & Hall.

Buccianti, Antonella, Glòria Mateu-Figueras, and Vera Pawlowsky-Glahn. 2006. "Compositional data analysis in the geosciences: from theory to practice." Geological Society of London.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. 2003. "Isometric logratio transformations for compositional data analysis." *Mathematical Geology* 35:279–300. doi:`https://doi.org/10.1023/A:1023818214614`.

Filzmoser, Peter, Kapel Hron, and Matthias Templ. 2018. *Applied Compositional Data Analysis : With Worked Examples in R.* Springer. doi:https://doi.org/10.1007/978-3-319-96422-5.

Hamilton, Nicholas E., and Michael Ferry. 2018. "ggtern: Ternary diagrams using ggplot2." *Journal of Statistical Software, Code Snippets* 87 (3): 1–17. doi:10.18637/jss.v087.c03.

Janssen, Ian, Anna E. Clarke, Valerie Carson, Jean-Philippe Chaput, Lora M. Giangregorio, Michelle E. Kho, Veronica J. Poitras, et al. 2020. "A systematic review of compositional data analysis studies examining associations between sleep, sedentary behaviour, and physical activity with health outcomes in adults." *Applied Physiology, Nutrition, and Metabolism* 45 (10): 248–257. doi:https://doi.org/10.1139/apnm-2020-0160.

Leskinen, Tuija, Kristin Suorsa, Miika Tuominen, Anna Pulakka, Jaana Pentti, Eliisa Löyttyniemi, Ilkka Heinonen, Jussi Vahtera, and Sari Stenholm. 2021. "The effect of consumer-based activity tracker intervention on physical activity among recent retirees—an RCT study." *Medicine & Science in Sports & Exercise.*

Migueles, Jairo H., Alex V. Rowlands, Florian Huber, Séverine Sabia, and Vincent T. van Hees. 2019. "GGIR: A research community–driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data." *Journal for the Measurement of Physical Behaviour* 2:188–196. doi:https://doi.org/10.1123/jmpb.2018-0063.

Pawlowsky-Glahn, Vera, and Antonella Buccianti. 2011. *Compositional Data Analysis : Theory And Applications.* Wiley.

Pawlowsky-Glahn, Vera, Juan J. Egozcue, and Raimon Tolosana-Delgado. 2015. *Modelling And Analysis Of Compositional Data.* Wiley.

Pedišić, Željko. 2014. "Measurement issues and poor adjustments for physical activity and sleep undermine sedentary behaviour research—the focus should shift to the balance between sleep, sedentary behaviour, standing and activity." *Kinesiology* 46:135–146. https://hrcak.srce.hr/123743.

Templ, Matthias, Karel Hron, and Peter Filzmoser. 2011. *robCompositions: An R-package For Robust Statistical Analysis Of Compositional Data,* 341–355. John Wiley / Sons. ISBN: 978-0-470-71135-4.

van den Boogaart, K. Gerald, Raimon Tolosana-Delgado, and Matevz Bren. 2020. *compositions: Compositional Data Analysis.* R package version 2.0-0. https://CRAN.R-project.org/package=compositions.

Wang, Zhichao, Huiwen Wang, and Shanshan Wang. 2019. "Linear mixed-effects model for multivariate longitudinal compositional data." *Neurocomputing* 335:48–58. ISSN: 0925-2312. doi:https://doi.org/10.1016/j.neucom.2019.01.043. https://www.sciencedirect.com/science/article/pii/S0925231219300694.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics For Data Analysis.* Springer-Verlag New York. ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *dplyr: A Grammar Of Data Manipulation.* R package version 1.0.2. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Evan Miller. 2020. *haven: Import And Export 'SPSS', 'Stata' And 'SAS' Files.* R package version 2.3.1. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2020. *knitr: A General-Purpose Package For Dynamic Report Generation In R.* R package version 1.30. https://yihui.org/knitr/.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table With 'kable' And Pipe Syntax.* R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra.