UNIVERSITY OF SÃO PAULO
SCHOOL OF ARTS, SCIENCES AND HUMANITIES
GRADUATE PROGRAM IN COMPLEX SYSTEMS MODELING

Daniel Vartanian

Ecology of sleep and circadian phenotypes of the Brazilian population

São Paulo
2023

Daniel Vartanian

Ecology of sleep and circadian phenotypes of the Brazilian population

Preliminary version

Thesis presented to the School of Arts, Sciences and Humanities at University of São Paulo, as part of the requirements for the degree of Master of Science by the Graduate Program in Complex Systems Modeling.

Area of concentration: Fundamentals of complex systems.

Supervisor: Prof. Dr. Camilo Rodrigues Neto

São Paulo

2023

Cataloging in publication
Library
[Library name]

Vartanian, Daniel

Ecology of sleep and circadian phenotypes of the Brazilian population / Daniel Vartanian; supervisor, Camilo Rodrigues Neto. São Paulo, 2023

128p. : il

Thesis (Master of Science) – Graduate Program in Complex Systems Modeling, School of Arts, Sciences and Humanities, University of São Paulo, 2023.

Preliminary version

1. Chronotype. 2. Entrainment. 3. Sleep. 4. Chronobiology. 5. Complex systems. I. Rodrigues Neto, Camilo, super. II. Title.

ERRATA

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Master's Thesis, University of São Paulo].

This is the preliminary version of this thesis (version <1.0.0). Any required corrections will be listed here upon approval.

Thesis by Daniel Vartanian, under the title **Ecology of sleep and circadian phenotypes of the Brazilian population**, presented to the School of Arts, Sciences and Humanities at the University of São Paulo, as part of the requirements for the degree of Master of Science by the Graduate Program in Complex Systems Modeling, in the concentration area of Fundamentals of complex systems.

Approved on ____________________ , __________ .

Examination committee

Committee chair:

Prof. Dr. ____

Institution ____

Examiners:

Prof. Dr. ____

Institution ____

Evaluation ____

Prof. Dr. ____

Institution ____

Evaluation ____

Prof. Dr. ____

Institution ____

Evaluation ____

I dedicate this work to the skeptics, the radicals, the ignorant, the uncivilized, the subversives, the wild dogs, the irreducibles, the irreconcilables. To the true engines of change. To the destabilizers, who possess equal or greater importance than the stabilizers. To those who act on principle, even knowing that there is no ultimate reward or any meaning in life.

ACKNOWLEDGEMENTS

I would like to acknowledge and express my gratitude to the following persons and organizations:

Salete Perroni (Sal), my partner in life and in the fight for a better world.

My Mother, for her unconditional love.

My sister and my brother, for their love and companionship in life.

My friends in science, Alicia Rafaelly Vilefort Sales, Juliana Viana Mendes, and Maria Augusta Medeiros de Andrade.

My friend and Professor Humberto Miguel Garay Malpartida, for his support; for his principles; and for his integrity, which was demonstrated when the need arose.

Professor Camilo Rodrigues Neto, for introducing me to and teaching me about the science of complex systems since 2012; for supervising my dissertation; for the patience and the virtue in taking on and mediating the process of transitioning my master's supervision after the breakdown of relations with my former supervisor.

Professor Carlos Molina Mendes, for his speed, impartiality, patience, and virtuous approach in mediating the process of transitioning of my master's supervision.

My fellow friends: Alex Azevedo Martins; Amanda Moreira; Augusto Amado, Carina (Cacau) Prado; Ítalo Alves Bezerra do Nascimento; Júlia Mafra; Letícia Nery de Figueiredo; Marcelo Ricardo Fernandes Roschel; Reginaldo Noveli; Sílvia Capelanes; and Vanessa Simon Silva.

President Lula (Yes!), who saved Brazil from fascism and approved the long-overdue adjustments to graduate scholarships.

The local student movements, which truly support their category.

The Support Program for Student Permanence and Education (PAPFE) of USP, which enabled me to get this far.

The Coordination for the Improvement of Higher Education Personnel (CAPES), for funding this work and enabling my presence in graduate studies (Grant number: 88887.703720/2022-00).

Nullius in verba[1]

[1] The Royal Society. (n.d.). *History of the Royal Society*. https://royalsociety.org/about-us/history/

ABSTRACT

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Master's Thesis, University of São Paulo].

The text below is related to the **project** of this thesis. The final abstract can only be produced when the research is completed.

Theories related to sleep and circadian rhythms are already well-established in science. However, it is necessary to verify and test these same theories in more extensive samples to obtain a more accurate picture of the ecology of sleep and temporal phenotypes. This thesis undertakes this commitment, with the aim of mapping the expression of sleep-wake cycles and circadian phenotypes in the Brazilian adult population and investigating the hypothesis that latitude is associated with circadian rhythm regulation. The latitude hypothesis is based on the idea that regions located at latitudes near the poles have, on average, a lower annual incidence of sunlight compared to regions near the equator (latitude 0°). Therefore, it is deduced that regions near the equator have a stronger solar zeitgeber, which, according to chronobiology theories, could lead to a greater propensity for the synchronization of circadian rhythms in these populations, reducing the amplitude and diversity of circadian phenotypes. This would also give these populations a morning characteristic when compared to populations living far from the equator. To achieve the aforementioned objectives, this thesis project will rely on a data sample of sleep-wake cycle expression in the Brazilian population, composed of $120,265$ respondents covering all Brazilian states. This data was obtained in 2017 and is based on the Munich ChronoType Questionnaire (MCTQ), a widely validated questionnaire used to measure circadian phenotypes based on the sleep-wake cycle expression of individuals in their last four weeks. The results will contribute to the validation of chronobiology theories and will generate greater knowledge about the regulation of circadian rhythms and sleep-wake cycles in the Brazilian population.

Keywords: Chronobiology. Biological rhythms. Chronotype. Circadian phenotype. Sleep. Complex systems. Entrainment. Latitude. Ecology. MCTQ.

RESUMO

Vartanian, D. (2023). *Ecologia do sono e de fenótipos circadianos da população brasileira* [Dissertação de Mestrado, Universidade de São Paulo].

O texto abaixo está relacionado ao **projeto** desta dissertação. O resumo final só poderá ser produzido quando a pesquisa for finalizada.

Teorias relacionadas ao sono e aos ritmos circadianos já estão bem consolidadas na ciência. No entanto, é necessário verificar e testar essas mesmas teorias em amostras mais abrangentes para obter um retrato mais preciso da ecologia do sono e dos fenótipos temporais. Esta dissertação assume esse compromisso, tendo como objetivo mapear a expressão dos ciclos de sono-vigília e dos fenótipos circadianos da população adulta brasileira e investigar a hipótese de que a latitude está associada à regulação do ritmo circadiano. A hipótese da latitude se fundamenta na ideia de que regiões localizadas em latitudes próximas aos polos apresentam, em média, uma menor incidência de luz solar anual quando comparadas com regiões próximas da linha do equador (latitude 0°). Dessa forma, deduz-se que as regiões próximas ao equador apresentam um zeitgeber solar mais forte, o que, de acordo com as teorias da cronobiologia, pode gerar uma maior propensão à sincronização dos ritmos circadianos dessas populações, reduzindo a amplitude e a diversidade de fenótipos circadianos. Isso também daria a essas populações uma característica matutina quando comparadas com populações que vivem distantes da linha do equador. Para atingir os objetivos mencionados, o projeto irá contar com uma amostra de dados da expressão do ciclo sono-vigília da população brasileira composta por 120.265 respondentes que abrange todos os estados brasileiros. Essa amostra de dados foi obtida no ano de 2017 e se baseia no Munich ChronoType Questionnaire (MCTQ), um questionário amplamente validado e utilizado para mensurar fenótipos circadianos a partir da expressão do ciclo sono-vigília de indivíduos em suas últimas quatro semanas. Os resultados irão contribuir com a validação de teorias da cronobiologia e gerar conhecimento sobre a regulação do ritmo circadiano e dos ciclos de sono-vigília da população brasileira.

Palavras-chaves: Cronobiologia. Ritmos biológicos. Cronotipo. Fenótipo circadiano. Sono. Sistemas complexos. Entrainment. Latitude. Ecologia. MCTQ.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

$_{F}$

Subscript indicating a relation with work-free days

$_{W}$

Subscript indicating a relation with workdays

BT

Local time of going to bed

FD

Number of work-free days per week

GU

Local time of getting out of bed

HO

Horne & Ostberg’s morningness-eveningness questionnaire (same as *MEQ*)

LE

Light exposure

LE_{week}

Average weekly light exposure

MCTQ

Munich ChronoType Questionnaire

$MCTQ^{PT}$

Portuguese version of the MCTQ

MEQ

Morningness-Eveningness Questionnaire

MSF

Local time of mid-sleep on work-free days

$\mathbf{MSF_{sc}}$

Chronotype proxy. The midpoint between sleep onset and sleep end on work-free days. A sleep correction ($_{SC}$) is made when a possible sleep compensation related to a lack of sleep on workdays is identified.

MSW

Local time of mid-sleep on workdays

PRC

Phase response curve

SD

Sleep duration

$\mathbf{SD_{week}}$

Average weekly sleep duration

SE

Local time of sleep end

SI

“Sleep inertia”. Despite the name, this abbreviation represents the time that a person takes to get up after sleep end. It is used this way by the MCTQ authors.

SJL

Absolute social jetlag

$\mathbf{SJL_{rel}}$

Relative social jetlag

$\mathbf{SJL_{sc}}$

Jankowski’s sleep-corrected social jetlag

$\mathbf{SJL_{sc\text{-}rel}}$

Jankowski’s relative sleep-corrected social jetlag

$\mathbf{Sloss_{week}}$

Weekly sleep loss

SO

Local time of sleep onset

Slat

Sleep latency, i.e., time (duration) to fall asleep after deciding to sleep

SPrep

Local time of preparing to sleep

TBT

Total time in bed

WD

Number of workdays per week

LIST OF SYMBOLS

For an extensive list of chronobiology related symbols, please refer to Aschoff et al. (1965) and M. D. Marques and Oda (2012).

τ

Period of a rhythm in free flow. Only revealed under constant environmental conditions.

T

Zeitgeber period

ϕ

Phase

$\Delta\phi$

Phase shift

$+\Delta\phi$

Phase advance

$-\Delta\phi$

Phase delay

Ψ

Phase relation

CONTENTS

1 INTRODUCTION

You are currently viewing the preliminary print version of this master's thesis.

This document follows the collection of articles thesis format. This first chapter serves as an introduction to the thesis subject, providing its justification, aims, and a list of all projects and related activities produced during its development. The subsequent chapters consist of a series of articles connected to the thesis, with the exception of the last one, which encompasses a discussion and final remarks.

All analyses in this document are reproducible and were conducted using the R programming language along with the Quarto publishing system. It's worth noting that this type of thesis is best suited for online viewing. To access the digital version and see the latest research updates, please visit https://danielvartan.github.io/mastersthesis/.

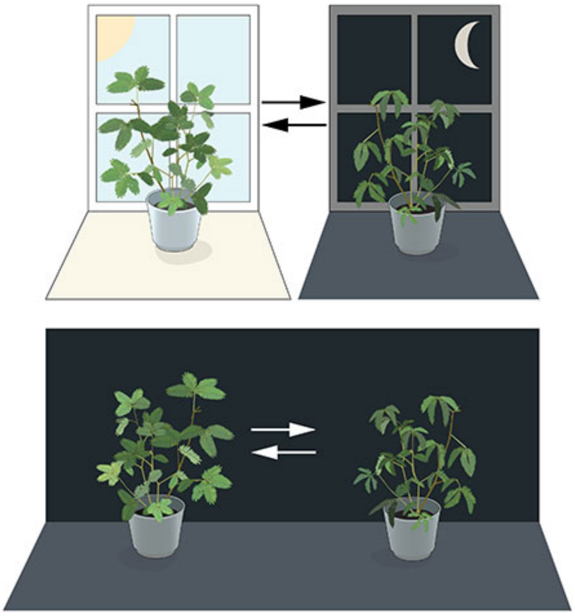
Given its preliminary nature, not all chapters are ready for reading. However, the author has chosen to display the entire state of the thesis rather than presenting only polished sections. This approach provides readers with a more comprehensive understanding of the work in progress. Chapters not suitable for reading will include a call block indicating their status.

1.1 A BRIEF INTRODUCTION TO CHRONOBIOLOGY

The dimension of time, manifest in the form of rhythms and cycles, like the alternating patterns of day and night as well as the annual transition of seasons, was consistently featured in the evolutionary journey of not only the human species but also all other life forms on our planet. These rhythms and cycles brought with them evolutionary pressures, resulting in the development of a temporal organization that allowed organisms to survive and reproduce in response to the conditions imposed by the environments they inhabited (Menna-Barreto, 2003; Pittendrigh, 1981). An example of this organization can be observed in the presence of different activity-rest patterns among living beings as they adapt to certain temporal niches, such as the diurnal behavior of humans and the nocturnal behavior of cats and some rodents (Foster & Kreitzman, 2005).

For years, scientists debated whether this organization was solely in response to environmental stimuli or if it was also present endogenously, internally, within organisms (Rotenberg et al., 2003). One of the early seminal studies describing a potential endogenous rhythmicity in living beings was conducted in 1729 by the French astronomer Jean Jacques d'Ortous de Mairan. De Mairan observed the movement of the sensitive plant (*mimosa pudica*) by isolating it from the light-dark cycle and found that the plant continued to move its leaves periodically (Figure 1) (Foster & Kreitzman, 2005; Rotenberg et al., 2003). The search for this internal timekeeper in living beings only began to solidify in the 20th century through the efforts of scientists like Jürgen Aschoff, Colin Pittendrigh, Franz Halberg, and Erwin Bünning, culminating in the establishment of the science known as chronobiology[1], with a significant milestone being the Cold Spring Harbor Symposium on Quantitative Biology: Biological Clocks in 1960 (*chrónos*, from Greek, meaning time; and *biology*, pertaining to the study of life) (Laboratory, n.d.; Rotenberg et al., 2003). However, the recognition of endogenous rhythmicity by the global scientific community truly came in 2017 when Jeffrey Hall, Michael Rosbash, and Michael Young were awarded the Nobel Prize in Physiology or Medicine for their discoveries of molecular mechanisms that regulate the circadian rhythm in fruit flies (*circā*, from Latin, meaning around, and *dĭes*, meaning day (Latinitium, n.d.) — a rhythm that expresses itself in approximately one day) (Nobel Prize Outreach AB, n.d.).

Figure 1 – Illustration of a circadian rhythm in the movement of the leaves of the sensitive plant (*mimosa pudica*) observed by Jacques d'Ortous de Mairan in 1729.



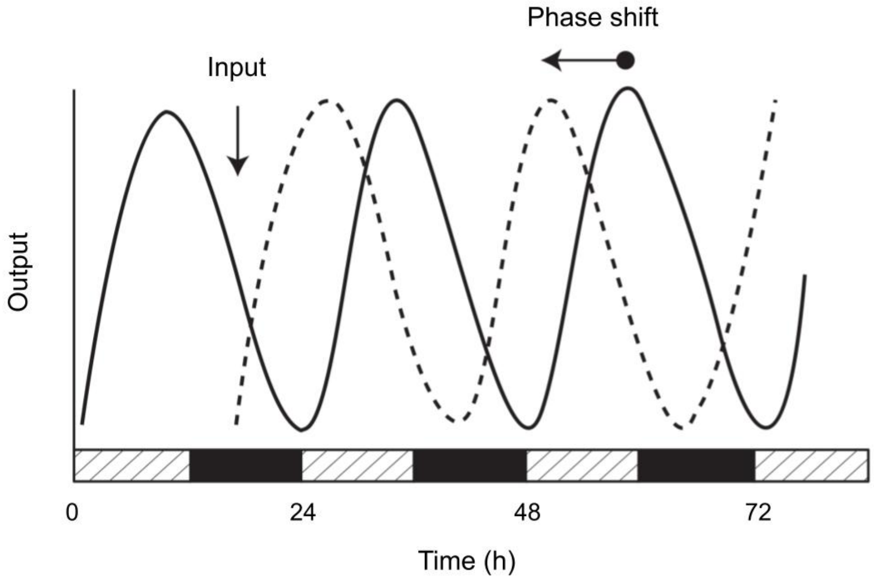
Source: Reproduction from Nobel Prize Outreach AB (n.d.).

[1] Some say the term *chronobiology* was coined by Franz Halberg during the Cold Spring Harbor Symposium on Quantitative Biology, vol. XXV (Menna-Barreto & Marques, 2023, p. 21).

Science has already demonstrated and described various biological rhythms and their impacts on organisms. These rhythms can occur at different levels, whether at a macro level, such as the menstrual cycle, or even at a micro level, such as rhythms expressed within cells (Roenneberg & Merrow, 2016). Like many other biological phenomena, these are complex systems present in all living beings, i.e., a emergence created by a large number of connected and interecticve agents that exhibit adaptive characteristics, all without the need of a central control (Boccara, 2010). It is understood today that the endogeneity of rhythms has provided organisms with an anticipatory capacity, allowing them to organize resources and activities before they are needed (N. Marques et al., 2003).

Despite the endogenous nature of these rhythms, they can still be regulated by the external environment. Signals (cues) from the environment that occur cyclically and have the ability to regulate biological rhythmic expression are called zeitgebers (from the German *zeit*, meaning time, and *geber*, meaning donor (Cambridge University Press, n.d.)). These zeitgebers act as synchronizers by entraining the phases of biological rhythms (Khalsa et al., 2003; Kuhlman et al., 2018) (see Figure 2). Among the known zeitgebers are, for example, meal timing and changes in environmental temperature (Aschoff, 1981; Roenneberg & Merrow, 2016). However, the most influential of them is the light-dark cycle. It is understood that the day/night cycle, resulting from the rotation of the Earth, has provided the vast majority of organisms with an oscillatory system with a periodic duration of approximately 24 hours (Kuhlman et al., 2018; Roenneberg, Kumar, & Merrow, 2007).

Figure 2 – Illustration of a circadian rhythm (output) whose phase is entrained in the presence of a zeitgeber (input). The rectangles represent the light-dark cycle.



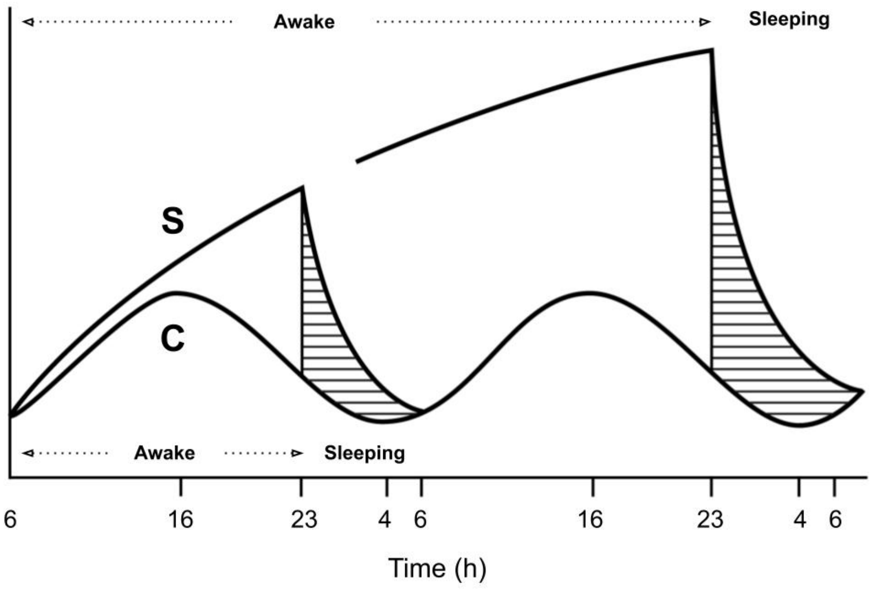
Source: Adapted from Kuhlman et al. (2018).

Naturally, the expression of this temporal organization varies from organism to organism, even among members of the same species, whether due to the different ways they are exposed to the environment or the differences in the expression of endogenous rhythmicity, which, in turn, results from gene expression (Roenneberg, Kuehnle, et al., 2007). The interaction between these two expressions, external and internal, of the environment and genotype, generates a signature, an observable characteristic, which is called a phenotype (Frommlet et al., 2016).

The various temporal characteristics of an organism can be linked to different oscillatory periods. Among these are circadian phenotypes, which refer to characteristics observed in rhythms with periods lasting about a day (Foster & Kreitzman, 2005). Another term used for these temporal phenotypes, as the name suggest, is *chronotype* (Ehret, 1974; Pittendrigh, 1993). This term is also often used to differentiate phenotypes on a spectrum ranging from morningness to eveningness (Horne & Ostberg, 1976; Roenneberg, Pilz, et al., 2019).

Sleep is a phenomenon that exhibits circadian expression. By observing the sleep characteristics of individuals, it is possible to assess the distribution of circadian phenotypes within the same population, thereby investigating their covariates and other relevant associations (Roenneberg, Wirz-Justice, & Merrow, 2003). This is because sleep regulation is understood as the result of the interaction between two processes: a homeostatic process (referred to as the S process), which is sleep-dependent and accumulates with sleep deprivation, and a circadian process (referred to as the C process), whose expression can be influenced by zeitgebers, such as the light-dark cycle (Borbély, 1982; Borbély et al., 2016) (Figure 3 illustrates these two process). Considering that the circadian rhythm (the C process) is present in sleep, its characteristics can be estimated if the S process can be controlled.

Figure 3 – Illustration of the interaction of the S process and the C process in sleep regulation. The figure depicts two scenarios: one without sleep deprivation and another with sleep deprivation. The y-axis represents the level of the process.



Source: Adapted from Borbély (1982).

Although many theories related to sleep and circadian rhythms are well-established in science, it is still necessary to verify and test them in larger samples to obtain a more accurate picture of the mechanisms related to the ecology of sleep and chronotypes. This project undertakes this commitment with the aim of investigating a hypothesis that is still relatively untested but widely accepted in chronobiology, which suggests that latitude is associated with the regulation of circadian rhythms (Hut et al., 2013; Leocadio-Miguel et al., 2014, 2017; Pittendrigh et al., 1991; Randler, 2008; Randler & Rahafar, 2017; Roenneberg, Wirz-Justice, & Merrow, 2003).

The latitude hypothesis is based on the idea that regions located at latitudes close to the poles, on average, experience less annual sunlight exposure compared to regions near the equator. Therefore, it is deduced that regions near latitude 0° have a stronger solar zeitgeber, which, according to chronobiology theories, should lead to a greater propensity for the synchronization of circadian rhythms in these populations with the light-dark cycle. This would reduce the amplitude and diversity of circadian phenotypes found due to a lower influence of individuals' characteristic endogenous periods. This would also give these populations a morningness characteristic when compared to populations living farther from the equator, where the opposite would occur – greater amplitude and diversity of circadian phenotypes and an eveningness characteristic compared to populations living near latitude 0° (Roenneberg, Wirz-Justice, & Merrow, 2003).

To achieve the mentioned objectives, this project will rely on a dataset of the sleep-wake cycle expression of the Brazilian population, consisting of $120,265$ respondents covering all states of the country. This dataset was collected in 2017 and is based on the Munich ChronoType Questionnaire (MCTQ), a widely validated scale used to measure chronotypes based on individuals' sleep-wake cycle expression in the last four weeks (Roenneberg, Wirz-Justice, & Merrow, 2003; Roenneberg et al., 2012).

1.2 THESIS JUSTIFICATION

Mapping the sleep-wake cycles and circadian phenotypes of Brazilians can contribute to the understanding of various phenomena related to sleep and chronobiology, such as the relationship between latitude and the regulation of circadian rhythms, the hypothesis tested by this thesis. However, in addition to contributing to the validation of theories and the advancement of scientific knowledge, the data, information, and knowledge generated by this project will also serve the public interest as a guide for public policies related to sleep and population health. Scientific literature is filled with studies pointing to negative associations with human health stemming from the disruption of biological rhythms. These range from fatigue (Tryon et al., 2004), deficits in cognitive performance (Dongen et al., 2003) , gastrointestinal problems (Fido & Ghali, 2008; Morito et al., 2014; Mortaş et al., 2020), mental disorders (Jones et al., 2005; Kalmbach et al., 2015; Roh et al., 2012) and even cancer (Lie et al., 2006; Papantoniou et al., 2015; Schernhammer et al., 2001).

This study will also produce the largest dataset of valid sleep-wake cycle expression among Brazilians ever recorded. For comparison, national epidemiological studies on sleep and circadian phenotypes such as those by Drager et al. (2022) and Leocadio-Miguel et al. (2017) worked with samples of $2,635$ and $12,884$ individuals, respectively. The sample of this project includes $120,265$ individuals in its raw state, covering all Brazilian states. Another advantage of the sample is its cross-sectional nature, as 98.173% of the data were collected during a single week (from October 15 to 21, 2017). This avoids potential distortions caused by seasonal effects.

1.3 THESIS AIMS

This project focuses on the ecology of sleep and circadian phenotypes (chronotypes) with the aim of providing answers to the following questions:

1. How are the sleep-wake cycles and circadian phenotypes of the adult Brazilian population characterized?
2. Is latitude associated with the regulation of circadian rhythms in humans?

The basic hypothesis to be tested is that populations residing near the equator (latitude 0°) have, on average, a shorter/more morning-oriented circadian phenotype compared to populations living near the Earth's poles (H1) (Hut et al., 2013; Leocadio-Miguel et al., 2014, 2017; Pittendrigh et al., 1991; Randler, 2008; Randler & Rahafar, 2017; Roenneberg, Wirz-Justice, & Merrow, 2003).

The primary objectives (PO) of the project are as follows:

A) Quantitatively describe the expression of sleep-wake cycles and circadian phenotypes of the Brazilian adult population at the end of the year 2017 (pre-pandemic).

B) Investigate and model the presence/absence of a significant association and effect between decimal degrees of latitude (independent variable (IV)) and circadian phenotypes (dependent variable (DV)) of the Brazilian population.

To achieve the primary objectives, the following secondary objectives (SO) have been outlined:

i) Conduct data cleaning, validation, and transformation processes on the obtained sample data.

ii) Collect secondary data on geolocation and solarimetric models and cross-reference them with the primary data.

iii) Develop algorithms for generating randomly sampled subsets adjusted to the proportions of the analyzed Brazilian regions, based on the latest Brazilian demographic census.

iv) Develop algorithms and models to help with the processing of MCTQ data and to simulate the complexity of the entrainment phenomena.

v) Evaluate and discuss the presence/absence of significant differences in the values of the corrected mid-sleep on free days (MSFsc) (DV) — a proxy for the expression of individuals' circadian phenotypes — based on decimal degrees of latitude (IV), while controlling for known covariates such as respondents' gender and age.

1.4 PROJECTS DEVELOPED

In addition to the main investigation, which is center on testing the latitude hypothesis, four additional projects/analyses were devised for this thesis. Each project was organized into a separate chapter, with the intention of crafting each chapter in a manner suitable for submission to a scientific journal. This organizational approach was influenced by the doctoral thesis of Reis (2020).

The first project involves a concise paper that delineates the resemblance observed among Portuguese translations of the MCTQ (Munich ChronoType Questionnaire) employed in scientific research. It's crucial to emphasize that, although the MCTQ functions as a self-report scale for assessing chronotypes, it primarily relies on objective temporal metrics (e.g., local bedtime, sleep latency duration) rather than more subjective factors such as perceived sleep quality. Essentially, it functions as a sleep diary. Nevertheless, these translations can exhibit noteworthy discrepancies. It's worth noting that the proper validation of MCTQ in Portuguese was only achieved in 2020 through the efforts of Reis (2020). The aim of this project is to assess the semantic similarity among these translations using a natural language model (NLM) known as Bidirectional

Encoder Representations from Transformers (BERT), developed by Google, and pre-trained on the Portuguese language (Devlin et al., 2018; Souza et al., 2020). By leveraging these semantic representation vectors, the translations will be evaluated based on cosine similarity.

The second project is an R package comprising a suite of tools designed for processing the MCTQ questionnaire. While it may appear to be a straightforward questionnaire, the MCTQ necessitates a considerable amount of date and time manipulation. This presents a challenge for many scientists, as handling date and time data can be particularly tricky, especially when dealing with extensive datasets. By creating a free, open-source and peer-reviewed R package, it becomes possible to standardize the analyses and enhance reproducibility for all research related to the MCTQ. This R package (Vartanian, 2023a)has already been developed and published on CRAN (The Comprehensive R Archive Network) and GitHub. It has been downloaded more than $6,000$ to this date, and underwent a peer review by the rOpenSci Initiative. Chapter 2 will serve as a manuscript for a publication regarding the package in the Journal of Statistical Software.

The third project is centered around the project's extensive MCTQ data sample, representing the largest dataset collected within a single country for this questionnaire thus far. This chapter serves as a crucial step in fulfilling one of the thesis primary objectives, which is to describe the sleep-wake cycle and circadian characteristics of the Brazilian population. Achieving this goal entails rigorous data cleaning and comprehensive data wrangling efforts. Furthermore, it functions as a means to facilitate the utilization of this valuable sample in future scientific research, while ensuring full compliance with ethical requirements.

The fourth project involves a rule-based model focusing on entrainment phenomena. Complex systems, such as biological rhythms, often exhibit the challenge of being described or represented concisely, as noted by David Krakauer (cited in Mitchell (2013)). Rule-based or agent-based models offer a means to simulate scenarios involving a multitude of agents and interactions. Models of this nature, underpinned by scientific theory-based rules, can provide valuable insights and enhance our comprehension of the various manifestations of entrainment phenomena within a population context. They offer an effective means to understand the implications of theory and test them against real-world data. An initial version of this package was developed as a Python package and is currently accessible on GitHub (see Vartanian, 2022b).

The fifth and final project is the test of the latitude hypothesis, which serves as the primary investigation. It's important to note that all the preceding projects converge into this one. The first project focuses on validating the MCTQ translation used for data collection. The second project involves the development of data processing tools. The third project is responsible for the necessary data manipulation to prepare it for analysis. The fourth project aims to offer valuable insights and guidance for the upcoming tasks.

All of these projects are developed using secure, open-source tools and adhere to the best international standards. They are designed to ensure 100% reproducibility and are accompanied by extensive documentation.

1.5 RELATED ACTIVITIES

During the development of this thesis, several activities and results have been accomplished. These activities are important to note, as they demonstrate the path taken to arrive at this final document.

1.5.1 **Courses**

The following graduate courses from the University of São Paulo (USP) were completed during the first year of the master's program.

- 2022/2: *SCX5000 - Mathematical and Computational Methods I* (10 credits) (Concept: **C**);
- 2022/2: *SCX5002 - Complex Systems I* (10 credits) (Concept: **A**);
- 2023/1: *SCX5001 - Mathematical and Computational Methods II* (10 credits) (Concept: **A**);
- 2023/1: *SCX5017 - Introduction to Data Science* (10 credits) (Concept: **A**);
- 2023/1: *EAH5001 - Pedagogic Preparation* (4 credits) (Concept: **A**).

Please note that the unfortunate **C** concept above happened in the same semester when the author broke relations with his former supervisor (*Mario Pedrazzoli*).

44 discipline credits were completed by this thesis publication date. An additional 12 special credits, related to an article publication (see Viana-Mendes et al. (2023)),

were requested and approved by the Graduate Program Coordination Commission (CCP) in accordance with program regulations. In total, 56 credits were earned. A minimum of 50 credits is required for the thesis defense.

1.5.2 **Teaching internship**

Scholarship students under the Coordination for the Improvement of Higher Education Personnel (CAPES) are required to participate in the Teaching Improvement Program (PAE). This internship is currently in progress and is scheduled to conclude in December 2023.

The internship responsibilities entail serving as an Assistant Professor for the undergraduate course *ACH0042 - Problem-Based Learning II* at USP. A comprehensive teaching plan (Vartanian, 2023b) was formulated during enrollment in the aforementioned graduate course *EAH5001*, and it is accessible through the following link.

Vartanian, D., Bernardes, M. E. M., & Rodrigues Neto, C. (2023). *Plano de ensino: ACH0042 - Resolução de Problemas II*. https://doi.org/10.13140/RG.2.2.33335.50086

1.5.3 **Publications**

The following article (Viana-Mendes et al., 2023) was published during the development of this thesis.

Viana-Mendes, J., Benedito-Silva, A. A., Andrade, M. A. M., **Vartanian, D.**, Gonçalves, B. da S. B., Cipolla-Neto, J., & Pedrazzoli, M. (2023). Actigraphic characterization of sleep and circadian phenotypes of PER3 gene VNTR genotypes. *Chronobiology International*. https://doi.org/10.1080/07420528.2023.2256858

1.5.4 **Translations**

As a member and package developer of the rOpenSci Initiative (based in Berkeley, CA), the author is actively contributing to the ongoing translation of the rOpenSci Developer Guide into Portuguese. The aim is to create a more inclusive environment for individuals in Brazil and other Portuguese-speaking countries when developing for the R programming language.

This endeavor is linked to the thesis, as the author's membership in rOpenSci began with the creation of the {mctq} R package (listed below).

1.5.5 **Conferences**

An abstract pertaining to the primary investigation was published and presented on a poster at the Sao Paulo School of Advanced Science on Ecology of Human Sleep and Biological Rhythms organized by the São Paulo Research Foundation (FAPESP). This international school hosted 100 participants, including students and young researchers, with a diverse representation of 50 individuals from various states within Brazil and an additional 50 from international backgrounds. The event took place from November 16, 2022, to November 26, 2022.

Vartanian, D., & Pedrazzoli, M. (2022). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Poster]. São Paulo Research Foundation; São Paulo School of Advanced Science on Ecology of Human Sleep and Biological Rhythms. https://doi.org/10.13140/RG.2.2.25343.07840

In the same semester (2022/2), the author also participated in USP's International Symposium on Scientific and Technological Initiation (SIICUSP) as both an examiner and a participant. As a participant, the author presented a research abstract related to the {actverse} R package for actigraphy data analysis, as detailed in Matias et al. (2022) and Vartanian (2022a). This project was conceived and developed by the author of this thesis and involved collaboration with two undergraduate students. Notably, this project achieved recognition, securing 2nd place in the category of *Earth and Exact Sciences*.

1.5.6 **Research compendia**

This thesis, along with all the accompanying research, is structured and organized within the research compendium provided below.

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Research compendium]. https://danielvartan.github.io/mastersthesis/

1.5.7 **Data plans**

This research has also produced and published the following open data model and data plan.

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Data Management Plan]. DMPHub. https://doi.org/10.48321/D1DW8P

1.5.8 **Softwares**

The following R packages, Quarto format (being used to write this thesis), and Python package were developed in relation with this thesis.

Vartanian, D. (2022). *{entrainment}: a rule-based model of the 24h light/dark cycle entrainment phenomenon* [Software, Python Package]. https://github.com/danielvartan/entrainment

Vartanian, D. (2023). *{mctq}: tools to process the Munich ChronoType Questionnaire (MCTQ)* [Software, R Package v0.3.2]. https://docs.ropensci.org/mctq/

Vartanian, D. (2023). *{lockr}: easily encrypt/decrypt files* [Software, R package v0.3.0]. https://github.com/danielvartan/lockr

Vartanian, D. (2023). *{lubritime}: an extension for the lubridate package* [Software, R package]. https://github.com/danielvartan/lubritime

Vartanian, D. (2023). *{abnt}: Quarto format for ABNT theses and dissertations* [Software, LaTeX/R format, v0.3.0]. https://github.com/danielvartan/abnt/

1.5.9 **Other projects**

The author is also currently working on the development of the project below.

Sales, A. R. V., Vartanian, D., Andrade, M. A. M., Pedrazzoli, M. (2023). *Associations between the duration and quality of sleep in third-trimester pregnant women and the duration of labor* [PhD project, University of Sao Paulo]. https://bit.ly/3S6O0MB

2 SIMILARITIES BETWEEN DIFFERENT VERSIONS OF THE MCTQ^{PT}

❗ Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

ℹ Target journal

1. Chronobiology International (IF 2022: 2.8/JCR | A1/2017-2020).
2. Journal of Biological Rhythms (IF 2022: 3.5/JCR | A2/2017-2020).

ℹ Note

The following study was performed by Daniel Vartanian (DV) and Camilo Rodrigues Neto (CR).

DV and **CR** contributed to the study's design. **DV** implemented the study, performed the statistical analysis, and authored the manuscript. All authors participated in discussions about the results and contributed to the final manuscript revision.

Future reference: Vartanian, D., & Rodrigues Neto, C. (2024). Similarities between different versions of the MCTQ^{PT}. *Chronobiology International*.

3 THE {MCTQ} R PACKAGE

! Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don’t recommend reading it.

ℹ Target journal

1. Journal of Statistical Software (IF 2022: 5.8/JCR | A1/2017-2020).
2. Journal of Open Source Software (B1/2017-2020).

ℹ Note

The following study was conducted by Daniel Vartanian (**DV**), Ana Amélia Benedito-Silva (**AA**), Mario Pedrazzoli (**MP**), and Camilo Rodrigues Neto (**CR**).

DV contributed to the conception, design, coding, and implementation of the software. **AA**, **MP**, and **CR** served as scientific advisors and reviewers. **DV** authored the manuscript. All authors discussed the results and revised the final manuscript.

Future reference: Vartanian, D., Benedito-Silva, A. A., Pedrazzoli, M., & Rodrigues Neto, C. (2024). {mctq}: tools to process the Munich ChronoType Questionnaire (MCTQ). *Journal of Statistical Software*.

4 ECOLOGY OF SLEEP AND CIRCADIAN PHENOTYPES OF THE BRAZILIAN POPULATION

❗ Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

ℹ Target journal

1. Chronobiology International (IF 2022: 2.8/JCR | A1/2017-2020).
2. Journal of Biological Rhythms (IF 2022: 3.5/JCR | A2/2017-2020).

ℹ Note

The following study was conducted by Daniel Vartanian (DV), Mario Pedrazzoli (MP), and Camilo Rodrigues Neto (CR).

DV conceived the study, contributed with the design, implementation, statistical analysis and authored the manuscript. **CR** contributed as a science adviser and reviewer. **DV** and **MP** were responsible for data collection. All authors actively participated in discussions regarding the results and contributed to the final manuscript.

Future reference: Vartanian, D., Pedrazzoli, M., & Rodrigues Neto, C. (2024). Ecology of sleep and circadian phenotypes of the Brazilian population. *Chronobiology International*.

5 RULE-BASED MODEL OF THE 24H LIGHT/DARK ENTRAINMENT PHENOMENON

❗ Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don't recommend reading it.

ℹ Target journal

1. Journal of Open Source Software (B1/2017-2020).

ℹ Note

The following study was conducted by Daniel Vartanian (DV) and Camilo Rodrigues Neto (CR).

DV was responsible for the design and software implementation. **CR** contributed as a science adviser and reviewer. **DV** wrote the manuscript. All authors discussed the results and revised the final manuscript.

Future reference: Vartanian, D, & Rodrigues Neto, C. (2024). {entrainment}: a rule-based model of the 24h light/dark cycle entrainment phenomenon. *Journal of Open Source*.

6 A BIOLOGICAL APPROACH FOR THE LATITUDINAL CLINE OF THE CHRONOTYPE

ℹ Note

You are reading the work-in-progress of this thesis.

This chapter should be readable but is currently undergoing final polishing.

⚠ Warning

The results shown here are **preliminary**, so please take them with a grain of salt.

The data has not yet been fully cleaned, balanced, and cross-referenced with the secondary databases. Think of these results as a low-resolution preview of the final results. The step-by-step analysis can be seen in the appendices section.

ℹ Target journal

1. Scientific Reports (IF 2022: 4.6/JCR | A1/2017-2020).

ℹ Note

The following study was performed by Daniel Vartanian (DV), Mario Pedrazzoli (MP) and Camilo Rodrigues Neto (CR).

DV contributed to the design and implementation of the study. **DV** and **MP** collected the data. **DV** and **CR** performed the statistical analysis. **DV** wrote the manuscript. All authors discussed the results and revised the final manuscript.

Future reference: Vartanian, D., Pedrazzoli, M., & Rodrigues Neto, C. (2024). A biological approach for the latitudinal cline of the chronotype. *Scientific Reports*.

Chronotypes are temporal phenotypes (Ehret, 1974; Pittendrigh, 1993). Observable traits, like weight and eye color. Our current understanding of these traits is that they are linked to our environment and are the result of evolution pressures for creating an inner temporal organization (Aschoff, 1989; Paranjpe & Sharma, 2005), a way that organisms found to anticipate events. Having such an important function in nature, these internal rhythms need to be closely aligned with environmental changes. The agents that shift these oscillations towards the environment are called zeitgebers and the shift phenomenon is called entrainment

(Roenneberg, Daan, & Merrow, 2003; Roenneberg et al., 2010). The main zeitgeber for humans is light exposure, particularly the light of the sun (Khalsa et al., 2003; Minors et al., 1991; Roenneberg, Kuehnle, et al., 2007). Considering the major role of light on entrainment, several studies hypothesized that the latitude shift of the sun could influence or even define the chronotypes of different populations (Horzum et al., 2015; Hut et al., 2013; Leocadio-Miguel et al., 2014, 2017; Pittendrigh et al., 1991; Randler & Rahafar, 2017). For example, populations that live close to the equator would be, on average, more entrained to the light-dark cycle and have morning-leaning characteristics. Here we test this hypothesis using a biological measure, the chronotype state, provided by the Munich ChronoType Questionnaire (Roenneberg, Wirz-Justice, & Merrow, 2003). We tested the latitude hypothesis on a sample with $76,744$ subjects living in different latitudes in Brazil. Our results show that, even with a wide, big, and aligned sample, the latitude is associated only with negligible effect sizes. The entrainment phenomenon appears to be much more complex than previously imagined, opening new questions and contradictions that need to be further investigated.

6.1 MAIN TEXT

6.1.1 **Introduction**

Humans can differ from one another in many ways. These observable traits, like hair color or height, are called phenotypes and are also presented in the way that our body functions.

A chronotype is a temporal phenotype (Ehret, 1974; Pittendrigh, 1993). This word is usually used to refer to endogenous circadian rhythms, i.e., rhythms which periods that are close to a day or 24 hours (*circa diem*). The current body of knowledge of Chronobiology, the science that studies biological rhythms, indicates that the evolution of these internal oscillators is linked to our oscillatory environment, like the day and night cycle, which, along with our evolution, created environmental pressures for the development of a temporal organization (Aschoff, 1989; Paranjpe & Sharma, 2005). A way in which an organism could predict events and better manage its needs, like storing food for the winter.

A temporal system wouldn't be of much use if it could not follow environmental changes. To those environmental signals that can regulate the biological rhythms are given the name zeitgeber (from the German Zeit, time, and Geber, giver). These zeitgebers produce inputs in our bodies that can shift and align those rhythms. This phenomenon is called entrainment (Roenneberg, Daan, & Merrow, 2003; Roenneberg et al., 2010).

The main zeitgeber known today is the light, particularly the sun's light (Khalsa et al., 2003; Minors et al., 1991; Roenneberg, Kuehnle, et al., 2007). Considering its influence in entraining the biological temporal system, several studies hypothesize that the latitudinal shift of the sun, related to the earth's axis, would produce, on average, different temporal traits in populations that live close to the equator line when compared to populations that live close to the planet's poles (Horzum et al., 2015; Hut et al., 2013; Leocadio-Miguel et al., 2014, 2017; Pittendrigh et al., 1991; Randler & Rahafar, 2017). That is because the latter ones would have greater oscillations in sun activity and an overall weak solar zeitgeber. This is the latitude hypothesis, that can also appear as an environmental hypothesis of circadian rhythm regulation.

Recently there have been attempts to test the latitude hypothesis in different settings, but, at least in humans, none of them have been successful in seeing a significant effect size related to the latitudinal cline. Some of these approaches worked with secondary data and with small samples. One of the most serious attempts of testing this hypothesis was made by Leocadio-Miguel et al. (2017) in 2017. They measured the chronotype of $12,884$ Brazillian subjects on a wide latitudinal spectrum using the Morningness–Eveningness Questionnaire (MEQ). Their results showed a negligible effect size. One possible reason for this is that the MEQ measures psychological traits and not biological states (Roenneberg, Winnebeck, & Klerman, 2019), i.e., the circadian oscillation itself, therefore, it's not the best way to answer the question (Leocadio-Miguel et al., 2014).

This article brings a novel attempt to test the latitude hypothesis, using, this time, a biological approach provided by the Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, & Merrow, 2003). Furthermore, the test was carried out on the biggest chronotype sample ever collected in a same country. A sample made of $76,744$ subjects, all living in the same timezone in Brazil, with only one week of difference between questionnaire responses.

6.1.2 Results

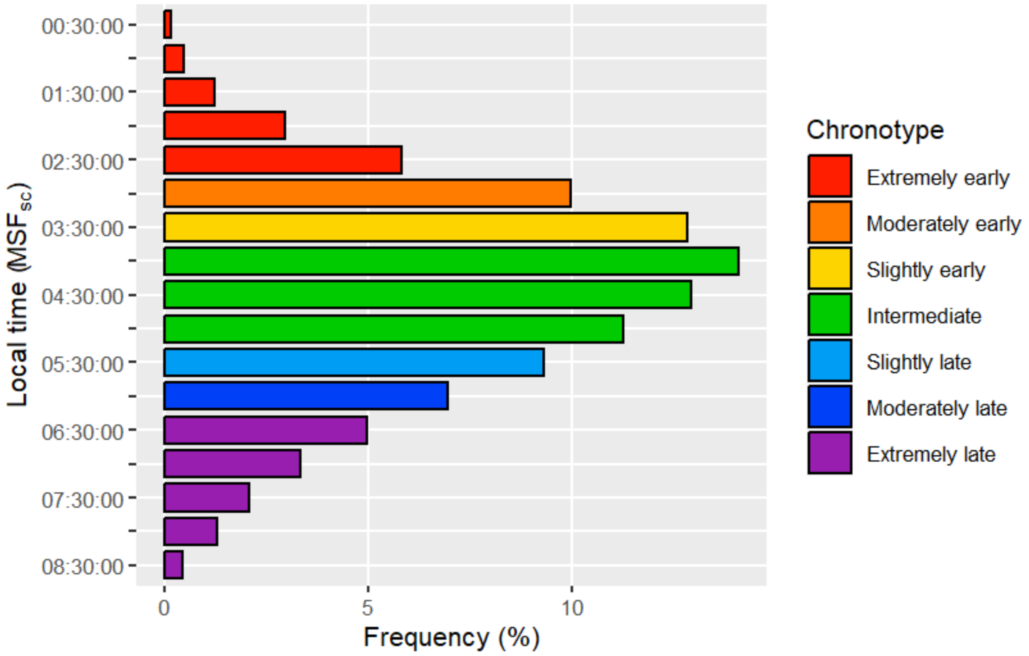
The local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, had an overall mean of 04:28:35. The distribution curve is shown in Figure 4.

That's the midsleep point of Brazilian subjects with an intermediate/average chronotype. One can imagine, following the 7-9h sleep recommendation for healthy adults of the American Academy of Sleep Medicine (AASM) (Watson et al., 2015), that this average person would, if he/she had no social obligations, typically wake up at about 08:28:35.

```r
source(here::here("R/utils.R"))

utc_minus_3_states <- c(
  "Amapá", "Pará", "Maranhão", "Tocantins", "Piauí", "Ceará",
  "Rio Grande do Norte", "Paraíba", "Pernambuco", "Alagoas", "Sergipe",
  "Bahia", "Distrito Federal", "Goiás", "Minas Gerais", "Espírito Santo",
  "Rio de Janeiro", "São Paulo", "Paraná", "Santa Catarina",
  "Rio Grande do Sul"
)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) |>
  dplyr::filter(state %in% utc_minus_3_states) |>
  dplyr::select(msf_sc, age, sex, state, latitude, longitude) |>
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

```r
source(here::here("R/plot_chronotype.R"))

data |>
  plot_chronotype(
    col = "msf_sc",
    x_lab = "Frequency (%)",
    y_lab = latex2exp::TeX("Local time ($MSF_{sc}$)"),
    col_width = 0.8,
    col_border = 0.6,
    text_size = env_vars$base_size,
    chronotype_cuts = FALSE,
    legend_position = "right"
  )
```

Figure 4 – Distribution of the local time of the midpoint between sleep onset and sleep end on work-free days (MSF$_{\text{sc}}$), MCTQ proxy for measuring the chronotype. The categorical cuts follow a quantile approach going from extremely early $(0| - 0.11)$ to the extremely late $(0.88 - 1)$.



Source: Created by the author.

The MSF$_{\text{sc}}$ curve had a skewness of 0.284 and a kurtosis of 2.773. However, the distribution was not normal accordingly to Kolmogorov-Smirnov test (D $= 0.03717$; p-value $= 2e - 16$) and D'Agostino Skewness test (Z3 $= 31.525$; p-value $= 2.2e - 16$).

A linear regression model was created with MSF$_{\text{sc}}$ as the response variable and with age and sex as predictors (R^2 $= 0.05373$; F$(2, 76741) = 2180$, p-value $= 2e - 16$), the two most known predictors for chronotype (Roenneberg, Kuehnle, et al. (2007)). A Box-Cox transformation of the response variable was needed to attend to the linear regression model assumptions ($\lambda = -1.1111$; MSF$_{\text{sc}}^{\lambda - 1} / \lambda$). All coefficients were significantly different than 0 (p-value $= 2e{-}16$) and, accordingly to D'Agostino Skewness test, the residuals were normal (Z3 $= -1.1906$; p-value $= 0.23383$). Residual homoscedasticity was verified by a Score Test for Heteroskedasticity ($\chi^2 = 0.00$; p-value $= 1$). No collinearity was found between the predictor variables (variance inflation factor: age $= 1.0012$; sex $= 1.0012$).

Another model was created on top of the first one, adding the latitude as a predictor variable (R^2 $= 0.060698$; F$(3, 76740) = 1650$, p-value $= 2e{-}16$). All coefficients were significantly different than 0 (p-value $= 2e{-}16$) and the residuals were normally distributed accordingly to the D'Agostino Skewness test, (Z3 $= 0.0742$; p-value $= 0.94085$). Residual homoscedasticity was verified by a Score Test for Heteroskedasticity ($\chi^2 = 0.00$; p-value $= 1$). No collinearity was found between the predictor variables (variance inflation factor: age $= 1.0065$; sex $= 1.0016$; \$latitude = 1.0056 \$). The longitude was not used as a predictor because it presented colinearity with the latitude variable.

An F test for nested models showed a significant reduction of the residual sum of squares (F$(1, 76740) = 568.94$, p-value $= 2e - 16$), meaning that the latitude seems to produce an effect on the chronotype. However, when estimating Cohen's f^2 effect size, the result was negligible (Cohen, 1992) $((0.06069 - 0.05373)/(1 - 0.06069) = 0.00740)$.

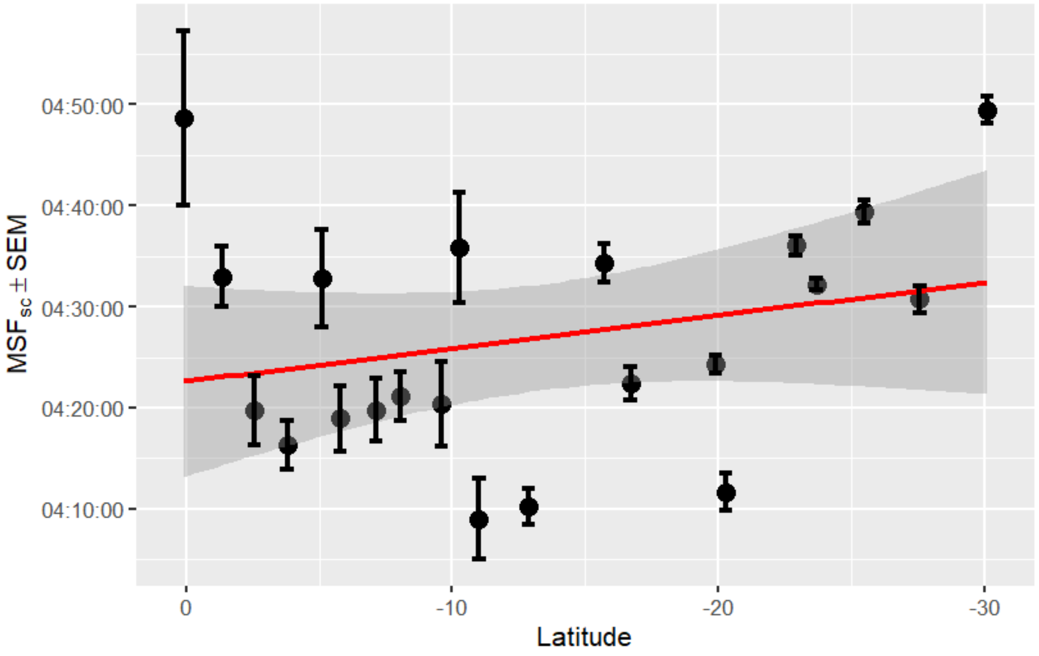
6.1.3 **Discussion**

The results show that even with a wide latitudinal spectrum and with a big and aligned sample of biological states the latitude effect does not reveal itself in a non-negligible size. Several studies indicate the existence of this effect on the chronotype (Hut et al., 2013; Leocadio-Miguel et al., 2017; Pittendrigh et al., 1991; Randler, 2008; Randler & Rahafar, 2017; Roenneberg, Wirz-Justice, & Merrow, 2003), but, at this time, at least in humans, no empirical evidence can support this claim. Our results are very similar to Leocadio-Miguel et al. (2017), which also found a negligible effect size (Cohen's $f^2 = 0.004143174$). The inconsistency of the latitude effect can be visualized in Figure 5.

```
source(here::here("R/plot_latitude_series.R"))

data ▷
```

```
  dplyr::filter(age <= 50) |>
  plot_latitude_series(
    col = "msf_sc",
    y_lab = latex2exp::TeX("$MSF_{sc} \\pm SEM$"),
    line_width = 2,
    point_size = 3,
    error_bar_width = 0.5,
    error_bar_linewidth = 1,
    error_bar = TRUE,
    text_size = env_vars$base_size
  )
```

Figure 5 – Distribution of mean aggregates of the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, in relation to latitude decimal degree intervals. Higher values of MSF_{sc} indicate a tendency toward a late chronotype. The red line represents a linear regression, and the shaded area indicates a pointwise 95% confidence interval.



Source: Created by the author.

Despite the lack of evidence, is not uncommon to hear talks insisting that this effect is real and already proven. We suspect that this behavior may be derived from a lack of understanding of statistical models and techniques. Although it may be logical and aligned with the overall theory for the evolution of biological temporal systems, it's our role as scientists to eliminate contractions, not pursue them.

As Karl Popper said, science begins and ends with questions (Popper, 1979). The absence of a strong entrainment with the solar zeitgeber shows that the entrainment phenomenon is more complex than we previously imagined. Other hypotheses for the human circadian entrainment, like the entrainment to self-selected light, proposed by Anna Skeldon and Derk-Jan Dijk (2021), need to be tested and may produce significant results.

It's important to notice that the results shown here are preliminary. The data still needs some cleaning and to be balanced with Brazil's latest population census. The latitude coordinates used in the analysis are related to the subject's state capital and, hence, have low resolution. Even with these results, it may be that a significant latitude effect can still appear at the end of the research.

Despite the several strengths that the dataset used in this study has, it is also important to notice its weaknesses and limitations. The fact that all the subjects were measured in the Spring season is one of them. Since the objective is to catch individuals in different seasonal patterns, the ideal moment to collect this kind of data is in the wintertime, when there is a greater insolation gradient between the equator and the poles. Another one is that this dataset can be influenced by the presence of a Daylight Saving Time (DST) event. This latter issue is explored in more detail in the methods section.

6.2 METHODS

6.2.1 **Ethics information**

Abiding by Brazilian law, all research involving human subjects must have the approval of a Research Ethics Committee (REC) affiliated with the Brazilian National Research Ethics Committee (CONEP). This approval request is ongoing (CAAE: 75588723.4.0000.5390).

6.2.2 **Measurement instrument**

Chronotypes were measured using the core version of the standard Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, & Merrow, 2003). MCTQ is a widely validated and widely used self-report questionnaire for measuring the sleep-wake cycle and chronotypes (Roenneberg, Winnebeck, & Klerman, 2019). It quantifies the chronotype as a state, a biological circadian phenotype, using as a proxy the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}). A sleep correction (SC) is made when a possible sleep compensation related to a lack of sleep on workdays is identified (Roenneberg, 2012).

Subjects were asked to complete an online questionnaire based on the MCTQ Portuguese translation created by Till Roenneberg & Martha Merrow for the EUCLOCK project (Roenneberg & Merrow, 2006) (statements mean cosine distance $= 0.921$). They were also asked to provide sociodemographic (e.g., age, sex), geographic (e.g., full residential address), anthropometric (e.g., weight, height), and work/study routine-related data. A deactivated version of the questionnaire can be seen at https://bit.ly/brchrono-form.

6.2.3 **Sample**

The sample is made up of $76,744$ Brazilian subjects. It was obtained in 2017 from October 15th to 21st by a broadcast of the online research questionnaire on a popular Sunday TV show with national reach (Globo, 2017). This amount of data collected in such a short time gave the sample a population cross-sectional characteristic.

A survey conducted in 2019 by the Brazilian Institute of Geography and Statistics (IBGE) (2021) found that 82.17% of Brazilian households had access to an internet connection. Therefore, this sample is likely to have a good representation of Brazil's population. Only residents of Brazilian states in the UTC-3 timezone, aged 18 years or older, were included in the final sample.

In order to verify if the sample size was adequate for the study of the phenomenon under investigation, a power analysis was conducted for nested multiple regression models using the G*Power software (Faul et al., 2007). The analysis used the parameters presented in Leocadio-Miguel et al. (2017) article for a multiple linear regression with 10 tested predictors and only 10 conceived predictors, considering a significance level of 0.05 (α) and a power of 0.95 ($1 - \beta$). The result showed that a sample of $5,895$ individuals would be necessary to test the hypothesis.

Daylight Saving Time (DST) began in Brazil at midnight on November 15th, 2017. Residents from the Midwest, Southeast, and South regions were instructed to set the clock forward by 1 hour. We believe that this event did not contaminate the data since it started on the same day of the data collection. It's important to notice that MCTQ asks subjects to relate their routine behavior, not how they behaved in the last few days. A possible effect of the DST on the sample is the production of an even later chronotype for populations near the planet's poles, amplifying a possible latitude effect. However, this was not shown on the hypothesis test.

Based on the 2022 census (Instituto Brasileiro de Geografia e Estatística, n.d.-a), Brazil had 52.263% of females and 47.737% of males with an age equal to or greater than 18 years old. The sample is skewed for female subjects, with 66.297% of females and 33.703% of male subjects.

The subject's mean age is 32.015 years (SD $= 9.252$; Max. $= 58.786$). Female subjects have a mean age of 31.787 years (SD $= 9.364$; Max. $= 58.786$) and male subjects 32.464 years (SD $= 9.012$; Max. $= 58.772$). For comparison, based on the 2022 census (Instituto Brasileiro de Geografia e Estatística, n.d.-c), Brazil's population with an age equal to or greater than 18 years old had a mean age of 44.277 years (SD $= 17.221$), with a mean age of 44.987 years (SD $= 17.511$) for female subjects and a mean age of 43.499 years (SD $= 16.864$) for male subjects.

Considering the five regions of Brazil, the sample is mostly skewed for the Southeast, the most populated region. According to Brazil's 2022 census (Instituto Brasileiro de Geografia e Estatística, 2022), the Southeast region is home to 41.784% of Brazil's population, followed by the Northeast (26.910%), South (14.741%), North (8.544%), and Midwest (8.021%) regions. 62.454% of the sample is located in the Southeast region, 11.797% in the Northeast, 17.861% in the South, 1.682% in the North, and 6.205% in the Midwest region. Note that a lack of subjects in the North and Midwest region is justified by the sample timezone inclusion criteria (UTC-3).

The sample latitudinal range was 30.211 decimal degrees (Min. $= -30.109$; Max. $= 0.10177$) with a longitudinal span of 16.378 decimal degrees (Min. $= -51.342$; Max. $= -34.964$). For comparison, Brazil has a latitudinal range of 39.024 decimal degrees (Min. $= -33.752$; Max. $= 5.2719$) and a longitudinal span of 39.198 decimal degrees (Min. $= -34.793$; Max. $= -73.991$).

The results shown in this article are just a preliminary view of the data analysis. The latitudes and longitudes of each subject are represented by the coordinates of his/her state's capital (a low resolution). The final results will have the latitude and longitude coordinates based on the subject's postal codes and will also use a balanced dataset following the latest Brazil census.

6.2.4 **Analysis**

The data wrangling and analysis followed the data science program proposed by Hadley Wickham and Garrett Grolemund (Wickham & Grolemund, 2016). All processes were made with the help of the R programming language (R. C. Team, 2023), RStudio IDE (P. Team, 2023), and several R packages. The tidyverse and rOpenSci package ecosystem and other R packages adherents of the tidy tools manifesto (Wickham & Bryan, 2023) were prioritized. The MCTQ data was analyzed using the `mctq` rOpenSci peer-reviewed package (Vartanian, 2023a). All processes were made in order to provide result reproducibility and to be in accordance with the FAIR principles (Wilkinson et al., 2016).

The study hypothesis was tested using nested models of multiple linear regressions. The main idea of nested models is to verify the effect of the inclusion of one or more predictors in the model variance explanation (i.e., the R^2) (Allen, 1997). This can be made by creating a restricted model and then comparing it with a full model. Hence, the hypothesis can be schematized as follows.

$$\begin{cases} \text{H}_0 : \text{R}^2_{\text{res}} >= \text{R}^2_{\text{full}} \\ \text{H}_a : \text{R}^2_{\text{res}} < \text{R}^2_{\text{full}} \end{cases}$$

In order to test a possible latitude association in predicting the chronotype, the full model was the restricted model with the addition of the latitude variable. The restricted model had the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}) as the response variable, MCTQ proxy for the chronotype, with sex and age as predictors.

A residual analysis was made to ensure the validity of the models before the hypothesis test. The hypothesis was tested using a 0.05 (α) significance level.

To favor the alternative hypothesis (H$_a$), not only the R^2 of the full model must be significantly larger than the R^2 of the restricted model, but the effect size must be at least considered small. To evaluate the effect size, Cohen's f^2 and his categorical parameters for size were used (Cohen, 1992). That means that, in order to favor (H$_a$), the effect size must be at least equal to or greater than 0.0219.

No blinding procedures were used during the analysis.

6.2.5 Data availability

The data that support the findings of this study are available from the corresponding author [DV]. Restrictions apply to the availability of these data, which were used under the approval of a Research Ethics Committee (REC) linked to the Brazilian National Research Ethics Committee (CONEP), hence it cannot be publicly shared. Data are, however, available from the author upon reasonable request and with CONEP approval.

6.2.6 Code availability

The research compendium of the project is available under the MIT license at https://github.com/danielvartan/mastersthesis. The code has all the steps from the raw data to the test results.

6.3 ACKNOWLEDGMENTS

Financial support was provided by the Coordination for the Improvement of Higher Education Personnel (CAPES) and by the University of Sao Paulo (USP) (Grant number: 88887.703720/2022-00).

6.4 ETHICS DECLARATIONS

6.4.1 **Competing interests**

The author declares that the study was carried out without any commercial or financial connections that could be seen as a possible competing interest.

6.5 ADDITIONAL INFORMATION

This manuscript shows only preliminary results and should not be considered a document ready for journal submission.

See the appendices section for supplementary information.

Correspondence can be sent to Daniel Vartanian (danvartan@gmail.com).

6.6 RIGHTS AND PERMISSIONS

7 DISCUSSION AND CONCLUSIONS

❗ Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don’t recommend reading it.

REFERENCES*

Allen, M. P. (1997). *Understanding regression analysis*. Plenum Press.

Andrade, T. (2023, January 27). *Tau*. Flyve.

Aschoff, J. (Ed.). (1981). *Biological rhythms* (Vol. 4). Plenum Press. https://doi.org/10.1007/978-1-4615-6552-9

Aschoff, J. (1989). Temporal orientation: Circadian clocks in animals and humans. *Animal Behaviour*, *37*, 881–896. https://doi.org/10.1016/0003-3472(89)90132-2

Aschoff, J., Klotter, K., & Wever, R. (1965). Circadian vocabulary: A recommended terminology with definitions. In *Circadian clocks*. North-Holland.

Bertalanffy, L. von. (1968). *General system theory: Foundations, development, applications*. George Braziller.

Boccara, N. (2010). *Modeling Complex Systems* (2nd ed.). Springer New York. https://doi.org/10.1007/978-1-4419-6562-2

Borbély, A. A. (1982). A two process model of sleep regulation. *Human Neurobiology*, *1*(3), 195–204. https://pubmed.ncbi.nlm.nih.gov/7185792/

Borbély, A. A., Daan, S., Wirz-Justice, A., & Deboer, T. (2016). The two-process model of sleep regulation: A reappraisal. *Journal of Sleep Research*, *25*(2), 131–143. https://doi.org/10.1111/jsr.12371

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brecht, B. (2000). *Poemas 1913-1956* (P. C. de Souza, Trans.; 5th ed.). Editora 34.

Cambridge University Press. (n.d.). *Cambridge dictionary*. Retrieved September 21, 2023, from https://dictionary.cambridge.org/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics* (4th ed.). Addison-Wesley.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved September 28, 2023, from https://arxiv.org/abs/1810.04805

Dongen, H. P. van, Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, *26*(2), 117–126. https://doi.org/10.1093/sleep/26.2.117

Drager, L. F., Pachito, D. V., Morihisa, R., Carvalho, P., Lobao, A., & Poyares, D. (2022). Sleep quality in the Brazilian general population: A cross-sectional study. *Sleep Epidemiology*, *2*, 100020. https://doi.org/10.1016/j.sleepe.2022.100020

Ehret, C. F. (1974). The sense of time: Evidence for its molecular basis in the eukaryotic gene-action system. In *Advances in Biological and Medical Physics* (pp. 47–77, Vol. 15). Elsevier. https://doi.org/10.1016/B978-0-12-005215-8.50009-7

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

* In accordance with the American Psychological Association (APA) Style, 7th edition.

Fido, A., & Ghali, A. (2008). Detrimental effects of variable work shifts on quality of sleep, general health and work performance. *Medical Principles and Practice*, *17*(6), 453–457. https://doi.org/10.1159/000151566

Foster, R. G., & Kreitzman, L. (2005). *Rhythms of life: The biological clocks that control the daily lives of every living thing*. Profile Books.

Frommlet, F., Bogdan, M., & Ramsey, D. (2016). *Phenotype and genotype: The search for influential genes* (Vol. 18). Springer London. https://doi.org/10.1007/978-1-4471-5310-8

Globo. (2017, October 15). *Metade da população se sente mal no horário de verão, revela pesquisa*. Fantástico. https://globoplay.globo.com/v/6219513/

Hair, J. F. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Horne, J. A., & Ostberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*(2), 97–110.

Horzum, M. B., Randler, C., Masal, E., Beşoluk, Ş., Önder, İ., & Vollmer, C. (2015). Morningness–eveningness and the environment hypothesis – a cross-cultural comparison of Turkish and German adolescents. *Chronobiology International*, *32*(6), 814–821. https://doi.org/10.3109/07420528.2015.1041598

Hut, R. A., Paolucci, S., Dor, R., Kyriacou, C. P., & Daan, S. (2013). Latitudinal clines: An evolutionary view on biological rhythms. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1765), 20130433. https://doi.org/10.1098/rspb.2013.0433

Instituto Brasileiro de Geografia e Estatística. (n.d.-a). *Tabela 1378 - população residente, por situação do domicílio, sexo e idade, segundo a condição no domicílio e compartilhamento da responsabilidade pelo domicílio* (Tabela). Tabela. Sistema IBGE de Recuperação Automática (SIDRA). https://sidra.ibge.gov.br/tabela/1378

Instituto Brasileiro de Geografia e Estatística. (n.d.-b). *Tabela 4714: População Residente, Área territorial e Densidade demográfica* (Tabela). Tabela. Sistema IBGE de Recuperação Automática (SIDRA). Retrieved November 11, 2023, from https://sidra.ibge.gov.br/Tabela/4714

Instituto Brasileiro de Geografia e Estatística. (n.d.-c). *Tabela 9514: População residente, por sexo, idade e forma de declaração da idade* (Tabela). Tabela. Sistema IBGE de Recuperação Automática (SIDRA). Retrieved November 11, 2023, from https://sidra.ibge.gov.br/tabela/9514

Instituto Brasileiro de Geografia e Estatística. (2021). Pesquisa nacional por amostra de domicílios contínua: Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal 2019. *Instituto Brasileiro de Geografia e Estatística*. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101794_informativo.pdf

Instituto Brasileiro de Geografia e Estatística. (2022). *Censo demográfico 2022 : População por idade e sexo : Resultados do universo*. IBGE. Retrieved November 11, 2023, from https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=73102

Jones, S. H., Hare, D. J., & Evershed, K. (2005). Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder. *Bipolar Disorders*, *7*(2), 176–186. https://doi.org/10.1111/j.1399-5618.2005.00187.x

Kalmbach, D. A., Pillai, V., Cheng, P., Arnedt, J. T., & Drake, C. L. (2015). Shift work disorder, depression, and anxiety in the transition to rotating shifts: The role of sleep reactivity. *Sleep Medicine*, *16*(12), 1532–1538. https://doi.org/10.1016/j.sleep.2015.09.007

Khalsa, S. B. S., Jewett, M. E., Cajochen, C., & Czeisler, C. A. (2003). A phase response curve to single bright light pulses in human subjects. *The Journal of Physiology*, *549*(3), 945–952. https://doi.org/10.1113/jphysiol.2003.040477

Kuhlman, S. J., Craig, L. M., & Duffy, J. F. (2018). Introduction to chronobiology. *Cold Spring Harbor Perspectives in Biology*, *10*(9), a033613. https://doi.org/10.1101/cshperspect.a033613

Laboratory, C. S. H. (n.d.). *1960: Biological clocks, vol. XXV*. Retrieved July 17, 2023, from https://symposium.cshlp.org/site/misc/topic25.xhtml

Latinitium. (n.d.). *Latin dictionaries*. Latinitium. Retrieved September 21, 2023, from https://latinitium.com/latin-dictionaries/

Leocadio-Miguel, M. A., Louzada, F. M., Duarte, L. L., Areas, R. P., Alam, M., Freire, M. V., Fontenele-Araujo, J., Menna-Barreto, L., & Pedrazzoli, M. (2017). Latitudinal cline of chronotype. *Scientific Reports*, *7*(1), 5437. https://doi.org/10.1038/s41598-017-05797-w

Leocadio-Miguel, M. A., Oliveira, V. C. D., Pereira, D., & Pedrazzoli, M. (2014). Detecting chronotype differences associated to latitude: A comparison between Horne–Östberg and Munich Chronotype questionnaires. *Annals of Human Biology*, *41*(2), 107–110. https://doi.org/10.3109/03014460.2013.832795

Levins, R. (1998). Dialectics and systems theory. *Science & Society*, *62*(3), 375–399. Retrieved September 21, 2023, from https://www.jstor.org/stable/40403729

Lie, J.-A. S., Roessink, J., & Kjærheim, K. (2006). Breast cancer and night work among norwegian nurses. *Cancer Causes & Control*, *17*(1), 39–44. https://doi.org/10.1007/s10552-005-3639-2

Marques, M. D., & Oda, G. (2012). Glossário. *Revista da Biologia*, *9*(3). Retrieved September 21, 2023, from https://www.revistas.usp.br/revbiologia/article/view/114816

Marques, N., Golombek, D. A., & Moreno, C. (2003). Adaptação temporal. In *Cronobiologia: Princípios e aplicações* (3. ed. rev. e ampl., pp. 55–98). Editora da Universidade de São Paulo.

Matias, V. A., Serrano, C., Vartanian, D., Pedrazzoli, M., & Benedito-Silva, A. A. (2022, September 3). *Ecology of sleep and circadian phenotypes of the Brazilian population [Poster]*. São Paulo. https://doi.org/10.13140/RG.2.2.25343.07840

Menna-Barreto, L. (2003). O tempo na biologia. In *Cronobiologia: Princípios e aplicações* (3. ed. rev. e ampl., pp. 25–30). Editora da Universidade de São Paulo.

Menna-Barreto, L., & Marques, N. (Eds.). (2023, August 16). *História e perspectivas da cronobiologia no brasil e na américa latina*. Editora da Universidade de São Paulo.

Minors, D. S., Waterhouse, J. M., & Wirz-Justice, A. (1991). A human phase-response curve to light. *Neuroscience Letters*, *133*(1), 36–40. https://doi.org/10.1016/0304-3940(91)90051-T

Mitchell, M. (2013). *Introduction to complexity*. Retrieved September 21, 2023, from https://www.complexityexplorer.org/courses/1-https://www.complexityexplorer.org/courses/1

Morito, Y., Aimi, M., Ishimura, N., Shimura, S., Mikami, H., Okimoto, E., Sato, S., Ishihara, S., Kushiyama, Y., Katsube, T., Adachi, K., & Kinoshita, Y. (2014). Association between sleep disturbances and abdominal symptoms. *Internal Medicine*, *53*(19), 2179–2183. https://doi.org/10.2169/internalmedicine.53.2591

Mortaş, H., Bilici, S., & Karakan, T. (2020). The circadian disruption of night work alters gut microbiota consistent with elevated risk for future metabolic and gastrointestinal pathology. *Chronobiology International*, *37*(7), 1067–1081. https://doi.org/10.1080/07420528.2020.1778717

Nobel Prize Outreach AB. (n.d.). *Press release*. The Nobel Prize. Retrieved September 28, 2023, from https://www.nobelprize.org/prizes/medicine/2017/press-release/

Papantoniou, K., Castaño□Vinyals, G., Espinosa, A., Aragonés, N., Pérez□Gómez, B., Burgos, J., Gómez□Acebo, I., Llorca, J., Peiró, R., Jimenez□Moleón, J. J., Arredondo, F., Tardón, A., Pollan, M., & Kogevinas, M. (2015). Night shift work, chronotype and prostate cancer risk in the MCC□S pain case□control study. *International Journal of Cancer*, *137*(5), 1147–1157. https://doi.org/10.1002/ijc.29400

Paranjpe, D. A., & Sharma, V. K. (2005). Evolution of temporal order in living organisms. *Journal of Circadian Rhythms*, *3*. https://doi.org/10.1186/1740-3391-3-7

Pittendrigh, C. S. (1960). Circadian rhythms and the circadian organization of living systems. *Cold Spring Harbor Symposia on Quantitative Biology*, *25*, 159–184. https://doi.org/10.1101/SQB.1960.025.01.015

Pittendrigh, C. S. (1981). Circadian systems: General perspective. In *Biological rhythms* (pp. 57–80, Vol. 4). Plenum Press. https://doi.org/10.1007/978-1-4615-6552-9

Pittendrigh, C. S. (1993). Temporal organization: Reflections of a darwinian clock-watcher. *Annual Review of Physiology*, *55*(1), 17–54. https://doi.org/10.1146/annurev.ph.55.030193.000313

Pittendrigh, C. S., Kyner, W. T., & Takamura, T. (1991). The amplitude of circadian oscillations: Temperature dependence, latitudinal clines, and the photoperiodic time measurement. *Journal of Biological Rhythms*, *6*(4), 299–313. https://doi.org/10.1177/074873049100600402

Popper, K. R. (1979). *Objective knowledge: An evolutionary approach*. Oxford University Press.

Randler, C. (2008). Morningness□eveningness comparison in adolescents from different countries around the world. *Chronobiology International*, *25*(6), 1017–1028. https://doi.org/10.1080/07420520802551519

Randler, C., & Rahafar, A. (2017). Latitude affects morningness-eveningness: Evidence for the environment hypothesis based on a systematic review. *Scientific Reports*, *7*(1), 39976. https://doi.org/10.1038/srep39976

Reis, C. (2020). *Sleep patterns in portugal* [Tese de doutorado]. Universidade de Lisboa. http://hdl.handle.net/10451/54147

Roenneberg, T. (2012). What is chronotype? *Sleep and Biological Rhythms*, *10*(2), 75–76. https://doi.org/10.1111/j.1479-8425.2012.00541.x

Roenneberg, T., Allebrandt, K. V., Merrow, M., & Vetter, C. (2012). Social jetlag and obesity. *Current Biology*, *22*(10), 939–943. https://doi.org/10.1016/j.cub.2012.03.038

Roenneberg, T., Daan, S., & Merrow, M. (2003). The art of entrainment. *Journal of Biological Rhythms*, *18*(3), 183–194. https://doi.org/10.1177/0748730403018003001

Roenneberg, T., Hut, R., Daan, S., & Merrow, M. (2010). Entrainment concepts revisited. *Journal of Biological Rhythms*, *25*(5), 329–339. https://doi.org/10.1177/0748730410379082

Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M., & Merrow, M. (2007). Epidemiology of the human circadian clock. *Sleep Medicine Reviews*, *11*(6), 429–438. https://doi.org/10.1016/j.smrv.2007.07.005

Roenneberg, T., Kumar, C. J., & Merrow, M. (2007). The human circadian clock entrains to sun time. *Current Biology*, *17*(2), R44–R45. https://doi.org/10.1016/j.cub.2006.12.011

Roenneberg, T., & Merrow, M. (2006). *EUCLOCK: Portuguese MCTQ*. Retrieved July 17, 2023, from http://web.archive.org/web/20141115175303/https://www.bioinfo.mpg.de/mctq/core_work_life/core/core.jsp?language=por_b

Roenneberg, T., & Merrow, M. (2016). The circadian clock and human health. *Current Biology*, *26*(10), R432–R443. https://doi.org/10.1016/j.cub.2016.04.011

Roenneberg, T., Pilz, L. K., Zerbini, G., & Winnebeck, E. C. (2019). Chronotype and social jetlag: A (self-) critical review. *Biology*, *8*(3), 54. https://doi.org/10.3390/biology8030054

Roenneberg, T., Winnebeck, E. C., & Klerman, E. B. (2019). Daylight saving time and artificial time zones – a battle between biological and social times. *Frontiers in Physiology*, *10*, 944. https://doi.org/10.3389/fphys.2019.00944

Roenneberg, T., Wirz-Justice, A., & Merrow, M. (2003). Life between clocks: Daily temporal patterns of human chronotypes. *Journal of Biological Rhythms*, *18*(1), 80–90. https://doi.org/10.1177/0748730402239679

Roh, J. H., Huang, Y., Bero, A. W., Kasten, T., Stewart, F. R., Bateman, R. J., & Holtzman, D. M. (2012). Disruption of the sleep-wake cycle and diurnal fluctuation of β-amyloid in mice with alzheimer's disease pathology. *Science Translational Medicine*, *4*(150). https://doi.org/10.1126/scitranslmed.3004291

Rotenberg, L., Marques, N., & Menna-Barreto, L. (2003). História e perspectivas da cronobiologia. In *Cronobiologia: Princípios e aplicações* (3. ed. rev. e ampl., pp. 31–53). Editora da Universidade de São Paulo.

Schernhammer, E. S., Laden, F., Speizer, F. E., Willett, W. C., Hunter, D. J., Kawachi, I., & Colditz, G. A. (2001). Rotating night shifts and risk of breast cancer in women participating in the nurses' health study. *JNCI Journal of the National Cancer Institute*, *93*(20), 1563–1568. https://doi.org/10.1093/jnci/93.20.1563

Skeldon, A. C., & Dijk, D.-J. (2021). Weekly and seasonal variation in the circadian melatonin rhythm in humans: Entrained to local clock time, social time, light exposure or sun time? *Journal of Pineal Research*, *71*(1), e12746. https://doi.org/10.1111/jpi.12746

Souza, F., Nogueira, R., & Lotufo, R. (2020, February 27). *Portuguese named entity recognition using BERT-CRF*. arXiv: 1909.10649 [cs]. https://doi.org/10.48550/arXiv.1909.10649

Team, P. (2023). *RStudio: Integrated development environment for R*. http://www.posit.co

Team, R. C. (2023). *R: A language and environment for statistical computing*. https://www.R-project.org

Tryon, W., Jason, L., Frankenberry, E., & Torresharding, S. (2004). Chronic fatigue syndrome impairs circadian rhythm of activity level. *Physiology & Behavior*, *82*(5), 849–853. https://doi.org/10.1016/S0031-9384(04)00303-8

Vartanian, D. (2022a). *{Actverse}: Tools for actigraphy data analysis* (Version 0.0.0.9000). https://docs.ropensci.org/mctq/

Vartanian, D. (2022b). *{Entrainment}: A rule-based model of the 24h light/dark cycle entrainment phenomenon* (Version 0.0.0.9000). https://github.com/danielvartan/entrainment

Vartanian, D. (2023a). *{Mctq}: Tools to process the Munich ChronoType Questionnaire (MCTQ)* (Version 0.3.2). https://docs.ropensci.org/mctq/

Vartanian, D. (2023b, July 12). *Plano de ensino: Ach0042 - resolução de problemas ii*. São Paulo. https://doi.org/10.13140/RG.2.2.33335.50086

Vartanian, D., & Pedrazzoli, M. (2017). *Questionário de cronotipo: Baseado no Munich ChronoType Questionnaire (MCTQ)*. https://web.archive.org/web/20171018043514/each.usp.br/gipso/mctq

Viana-Mendes, J., Benedito-Silva, A. A., Andrade, M. A. M., Vartanian, D., Gonçalves, B. d. S. B., Cipolla-Neto, J., & Pedrazzoli, M. (2023). Actigraphic characterization of sleep and circadian phenotypes of PER3 gene VNTR genotypes. *Chronobiology International*, *40*(9), 1244–1250. https://doi.org/10.1080/07420528.2023.2256858

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information*, *11*(9), 421. https://doi.org/10.3390/info11090421

Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F., & Tasali, E. (2015). Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Journal of Clinical Sleep Medicine*, *11*(6), 591–592. https://doi.org/10.5664/jcsm.4758

Wickham, H., & Bryan, J. (2023). *R packages* (2nd ed.). O'Reilly. https://r-pkgs.org/

Wickham, H., & Grolemund, G. (2016, December). *R for data science*. O'Reilly. https://r4ds.had.co.nz

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

GLOSSARY

For an extensive list of chronobiology related terms and definitions, please refer to Aschoff et al. (1965) and M. D. Marques and Oda (2012).

Chronotype

Any kind of temporal phenotype (Ehret, 1974; Pittendrigh, 1993). Usually, it refers to circadian phenotypes in a spectrum that goes from morningness to eveningness (Roenneberg, Wirz-Justice, & Merrow, 2003). It can also be seen as an organism's phase of entrainment (Roenneberg et al., 2012).

Circadian rhythm

A rhythm with a period close to a day/24h, an approximation to the period of the earth's rotation (Pittendrigh, 1960). From the Latin *circā*, around, and *dĭes*, day (Latinitium, n.d.). Example: the sleep-wake cycle.

Complex system

There are several definitions. Here are some that I found to be of use:

- "Systems that don't yield to compact forms of representation or description" (David Krakauer apud Mitchell (2013));
- "A system of many interacting parts where the system is more than just the sum of its parts" (Mark Newman apud Mitchell (2013));
- Systems with many connected agents that interact and exhibit self-organization and emergence behavior, all without the need for a central controller (adapted from Camilo Rodrigues Neto's definition, supervisor of this thesis);
- Dialectics at its finest (my working definition).

Entrainment

A shift and alignment of biological rhythms induced by a zeitgeber input (Kuhlman et al., 2018). For example: a shift/alignment of an organism's circadian rhythm when exposed to light.

Infradian rhythm

A rhythm with a period greater than a day/24h. From the Latin *infrā*, below (think in terms of period repetition), and *dĭes*, day (Latinitium, n.d.). Example: the menstrual cycle.

Period

Cycle duration of an oscillation. In a more technical way, the duration between two identical and consecutive phases in an oscillation (Kuhlman et al., 2018).

System theory

Two definitions can be of use:

- Science or discipline that investigates models, principles, and laws that are valid to systems in general (Bertalanffy, 1968);
- "The attempt of a reductionist scientific tradition to come to terms with complexity, nonlinearity, and change through sophisticated mathematical and computational techniques, *a groping toward a more dialectical understanding* that is held back by its philosophical biases and the institutional and economic contexts of its development" (Levins, 1998).

Ultradian rhythm

A rhythm with a period below a day/24h. From the Latin *ultrā*, beyond (think in terms of period repetition), and *dĭes*, day (Latinitium, n.d.). Example: the cardiac cycle.

Zeitgeber

Any periodic environmental signal/cue that can influence or regulate biological rhythms. From the German *zeit*, time, and *geber*, donor (Cambridge University Press, n.d.). Two main well known zeitgebers are light exposure and environment temperature (Pittendrigh, 1960).

APPENDIX A – CHAPTER 2 SUPPLEMENTAL MATERIAL

Note

You are reading the work-in-progress of this thesis.

This chapter should be readable but is currently undergoing final polishing.

A.1 BASE TEXTS

See Vartanian and Pedrazzoli (2017) to visualize the data questionnaire.

See Roenneberg and Merrow (2006) to visualize the EUCLOCK Portuguese questionnaire.

See Reis (2020) to learn more about the MCTQ^{PT} questionnaire. It's important to note that the MCTQ^{PT} was not included in the validation article. To obtain full access to the questionnaire statements, you should contact the main author of the article.

Two control texts were used, one from Andrade (2023) and another from Brecht (2000).

```
data_text <- c(
  "Você vai para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "Depois de ir para a cama, você decide dormir às ___ horas.",
  "Você precisa de ___ para dormir.",
  "Você acorda às ___ horas.",
  "Você se levanta ___ depois de despertar.",
  "Você vai para a cama às ___ horas.",
  "",
  "Depois de ir para a cama, você decide dormir às ___ horas.",
  "Você precisa de ___ para dormir.",
  "Você acorda às ___ horas.",
  "Você se levanta ___ depois de despertar."
)

euclock_text <- c(
```

```r
  "vou para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "às ___ horas, decido dormir.",
  "Eu necessito ___ minutos para adormecer.",
  "acordo às ___ horas,",
  "passados ___ minutos, me levanto.",
  "vou para a cama às ___ horas.",
  "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
  "às ___ horas, decido dormir.",
  "Eu necessito ___ minutos para adormecer.",
  "acordo às ___ horas,",
  "passados ___ minutos, me acordo."
)

mctq_pt_text <- c(
  "Vou para a cama às ___ horas.",
  "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
  "Às ___ horas estou pronto para adormecer.",
  "Necessito de ___ minutos para adormecer.",
  "Acordo às ___ horas.",
  "Após ___ minutos, levanto-me.",
  "Vou para a cama às ___ horas.",
  "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
  "Às ___ horas estou pronto para adormecer.",
  "Necessito de ___ minutos para adormecer.",
  "Acordo às ___ horas.",
  "Após ___ minutos, levanto-me."
)

# See: Andrade, T. (2023). Acronomia. In T. Andrade, Tau (Chapter 1). Flyve.
control_text_1 <- c(
  "Eles eliminaram o tempo, definitivamente.",
  "Removeram todos os relógios, de parede, de pulso, de bolso...",
  "Talvez esses objetos fossem realmente obsoletos àquela altura",
  "mas sim, foi deliberado: era um projeto mundial.",
  "Mas a situação é bem pior do que parece a princípio.",
  "Não foi apenas qualquer possibilidade de aferição do tempo",
  "exterminaram a própria capacidade de produzi-lo.",
  "Primeiro marcaram o 'Grande dia da entrega'.",
  "Um comboio de carros de lixo passou pelas ruas",
  "recolhendo todos os tipos de relógio",
  "e cronômetro que estavam de posse das pessoas.",
  "De mecanismos empoeirados e engrenagens enferrujadas a dispositivos modernos"
)

# See: Brecht, B. (2000). Quem se defende. In B. Brecht, Poemas 1913-1956
#      (5th ed., p. 73; Paulo César de Souza, Trans.). Editora 34.
control_text_2 <- c(
  "Quem se defende porque lhe tiram o ar",
  "Ao lhe apertar a garganta, ",
  "para este há um parágrafo",
  "Que diz: ele agiu em legítima defesa. ",
  "Mas",
  "O mesmo parágrafo silencia",
  "Quando vocês se defendem porque lhes tiram o pão.",
  "E no entanto morre quem não come, ",
  "e quem não come o suficiente",
  "Morre lentamente. ",
  "Durante os anos todos em que morre",
  "Não lhe é permitido se defender."
)
```

```
data_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = data_text,
```

```r
    meta = list(id = "data")
  )

euclock_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = euclock_text,
    meta = list(id = "euclock")
  )

mctq_pt_text_textreuse <-
  textreuse::TextReuseTextDocument(
    text = mctq_pt_text,
    meta = list(id = "mctq_pt")
  )

control_text_1_textreuse <-
  textreuse::TextReuseTextDocument(
    text = control_text_1,
    meta = list(id = "control_1")
  )

control_text_2_textreuse <-
  textreuse::TextReuseTextDocument(
    text = control_text_2,
    meta = list(id = "control_2")
  )
```

```r
# See
# <https://huggingface.co/neuralmind/bert-base-portuguese-cased>
# to learn more.

rutils:::assert_internet()

```

```r
text_embed <- function(text) {
  checkmate::assert_character(text)

  text |>
    text::textEmbed(
      model = "neuralmind/bert-base-portuguese-cased",
      layers = - 2,
      dim_name = TRUE,
      aggregation_from_layers_to_tokens = "concatenate",
      aggregation_from_tokens_to_texts = "mean",
      aggregation_from_tokens_to_word_types = NULL,
      keep_token_embeddings = TRUE,
      tokens_select = NULL,
      tokens_deselect = NULL,
      decontextualize = FALSE,
      model_max_length = NULL,
      max_token_to_sentence = 4,
      tokenizer_parallelism = FALSE,
      device = "gpu",
      logging_level = "error"
    )
}

data_text_textembed <- text_embed(data_text)
euclock_text_textembed <- text_embed(euclock_text)
mctq_pt_text_textembed <- text_embed(mctq_pt_text)
control_text_1_textembed <- text_embed(control_text_1)
control_text_2_textembed <- text_embed(control_text_2)
```

A.2 TEXT SIMILARITY

See Wang and Dong (2020) to learn more.

For a quick explanation, see https://youtu.be/e9U0QAFbfLI.

```r
text_distance <- function(x, y) {
  checkmate::assert_list(x, len = 2)
  checkmate::assert_list(y, len = 2)

  methods <- c(
    "binary", "cosine", "canberra", "euclidean", "manhattan", "maximum",
    "minkowski", "pearson"
  )

  for (i in methods) {
    cli::cli_alert_info(paste0(
      "Method: {.strong {stringr::str_to_title(i)}}"
      ))

    test <-
      text::textSimilarity(
        x$texts$texts,
        y$texts$texts,
        method = i,
        center = TRUE,
        scale = FALSE
      )

    cli::cli_bullets(c(">" = "Line by line"))
    print(test)

    cli::cli_bullets(c(">" = "Overall mean"))
    print(mean(test))

    cli::cat_line()
  }
}
```

```
text_representation <- function(x, y) {
  checkmate::assert_class(x, "TextReuseTextDocument")
  checkmate::assert_class(y, "TextReuseTextDocument")

  cli::cli_alert_info(paste0("Method: {.strong Jaccard similarity}"))
  print(textreuse::jaccard_similarity(x, y))
  cli::cat_line()

  cli::cli_alert_info(paste0("Method: {.strong Jaccard bag similarity}"))
  print(textreuse::jaccard_bag_similarity(x, y))
  cli::cat_line()
}
```

A.2.1 How similar is the *data questionnaire* when compared to the *EUCLOCK questionnaire*?

A.2.1.1 Text distance

```
text_distance(data_text_textembed, euclock_text_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
#> > Overall mean
#> [1] 1
#> i Method: Cosine
#> > Line by line
#>  [1] 0.9911730 1.0000000 0.9639984 0.9662432 0.9604119 0.9557896 0.9911730
#>  [8] 0.1559853 0.9639984 0.9662432 0.9604119 0.9497428
#> > Overall mean
#> [1] 0.9020976
#> i Method: Canberra
#> > Line by line
#>  [1] -218.3367    1.0000 -335.1895 -318.9419 -335.9000 -373.7989 -218.3367
#>  [8] -642.5522 -335.1895 -318.9419 -335.9000 -381.4067
#> > Overall mean
#> [1] -317.7912
#> i Method: Euclidean
#> > Line by line
#>  [1]  -1.504474   1.000000  -4.058168  -3.976768  -4.144779  -4.705653
#>  [7]  -1.504474 -19.453535  -4.058168  -3.976768  -4.144779  -5.071087
#> > Overall mean
#> [1] -4.633221
#> i Method: Manhattan
#> > Line by line
#>  [1]  -53.58509    1.00000 -108.96057 -105.53439 -111.46668 -124.16400
#>  [7]  -53.58509 -198.94244 -108.96057 -105.53439 -111.46668 -131.16054
#> > Overall mean
#> [1] -101.03
#> i Method: Maximum
#> > Line by line
#>  [1]   0.52944993   1.00000000   0.37664773   0.07405534   0.11978415
#>  [6]   0.31058226   0.52944993 -14.91353795   0.37664773   0.07405534
#> [11]   0.11978415   0.16830817
#> > Overall mean
#> [1] -0.9362311
#> i Method: Minkowski
#> > Line by line
#>  [1]  -1.504474   1.000000  -4.058168  -3.976768  -4.144779  -4.705653
#>  [7]  -1.504474 -19.453535  -4.058168  -3.976768  -4.144779  -5.071087
#> > Overall mean
#> [1] -4.633221
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9911730 1.0000000 0.9639984 0.9662432 0.9604119 0.9557896 0.9911730
#>  [8] 0.1559853 0.9639984 0.9662432 0.9604119 0.9497428
#> > Overall mean
#> [1] 0.9020976
```

A.2.1.2 Text representation

> **Note**: The maximum value for the Jaccard bag similarity is 0.5.

```
text_representation(euclock_text_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0.2173913
#> i Method: Jaccard bag similarity
#> [1] 0.1446541
```

A.2.2 **How similar is the *data questionnaire* when compared to the *MCTQ^{PT} questionnaire*?**

A.2.2.1 Text distance

```
text_distance(data_text_textembed, mctq_pt_text_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
#> > Overall mean
#> [1] 1
#> i Method: Cosine
#> > Line by line
#>  [1] 0.9901437 0.9898982 0.9687513 0.9575260 0.9882873 0.9598702 0.9901437
#>  [8] 0.1601500 0.9687513 0.9575260 0.9882873 0.9598702
#> > Overall mean
#> [1] 0.9066005
#> i Method: Canberra
#> > Line by line
#>  [1] -227.9938 -247.6044 -335.9297 -349.9493 -225.4809 -353.5263 -227.9938
#>  [8] -631.6228 -335.9297 -349.9493 -225.4809 -353.5263
#> > Overall mean
#> [1] -322.0823
#> i Method: Euclidean
#> > Line by line
#>  [1]  -1.662187  -1.729814  -3.807963  -4.603019  -1.810249  -4.458696
#>  [7]  -1.662187 -19.380367  -3.807963  -4.603019  -1.810249  -4.458696
#> > Overall mean
#> [1] -4.482867
#> i Method: Manhattan
#> > Line by line
#>  [1]  -57.28537  -58.97985 -102.26048 -119.55311  -59.76706 -117.55850
#>  [7]  -57.28537 -193.79964 -102.26048 -119.55311  -59.76706 -117.55850
#> > Overall mean
#> [1] -97.13571
#> i Method: Maximum
#> > Line by line
#>  [1]   0.60554241   0.60396075   0.41795957   0.01856209   0.39220500
#>  [6]   0.28932291   0.60554241 -14.95654231   0.41795957   0.01856209
#> [11]   0.39220500   0.28932291
#> > Overall mean
#> [1] -0.9087831
#> i Method: Minkowski
#> > Line by line
#>  [1]  -1.662187  -1.729814  -3.807963  -4.603019  -1.810249  -4.458696
#>  [7]  -1.662187 -19.380367  -3.807963  -4.603019  -1.810249  -4.458696
#> > Overall mean
#> [1] -4.482867
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9901437 0.9898982 0.9687513 0.9575260 0.9882873 0.9598702 0.9901437
#>  [8] 0.1601500 0.9687513 0.9575260 0.9882873 0.9598702
#> > Overall mean
#> [1] 0.9066005
```

A.2.2.2 Text representation

> **Note**: The maximum value for the Jaccard bag similarity is 0.5.

```
text_representation(mctq_pt_text_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0.1052632
#> i Method: Jaccard bag similarity
#> [1] 0.09815951
```

A.2.3 How similar is the *data questionnaire* when compared to the *Control Text 1*?

A.2.3.1 Text distance

```
text_distance(data_text_textembed, control_text_1_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
#> > Overall mean
#> [1] 1
#> i Method: Cosine
#> > Line by line
#>  [1] 0.9050224 0.8904954 0.8996243 0.8864538 0.8587490 0.8921112 0.8880434
#>  [8] 0.1887362 0.8882691 0.8727013 0.8732868 0.8433085
#> > Overall mean
#> [1] 0.8239001
#> i Method: Canberra
#> > Line by line
#>  [1] -469.3314 -485.1504 -483.8301 -506.1996 -504.6484 -490.3147 -478.4438
#>  [8] -632.5119 -492.0333 -490.9512 -496.6430 -517.2764
#> > Overall mean
#> [1] -503.9445
#> i Method: Euclidean
#> > Line by line
#>  [1]  -7.299638  -7.859283  -7.547236  -8.207869  -8.889088  -8.131165
#>  [7]  -8.090413 -18.891223  -7.834767  -8.892308  -8.472040  -9.696238
#> > Overall mean
#> [1] -9.150939
#> i Method: Manhattan
#> > Line by line
#>  [1] -174.9785 -191.0842 -182.5955 -197.5195 -212.8954 -196.0854 -196.2141
#>  [8] -181.0755 -192.3436 -212.1597 -198.4904 -233.2716
#> > Overall mean
#> [1] -197.3928
#> i Method: Maximum
#> > Line by line
#>  [1]  -0.1064962  -0.1808311  -0.2100441  -0.3340142  -0.3991810  -0.4304202
#>  [7]  -0.1888392 -15.1616240  -0.4010140  -0.9773587  -0.4897276  -0.6558742
#> > Overall mean
#> [1] -1.627952
#> i Method: Minkowski
#> > Line by line
#>  [1]  -7.299638  -7.859283  -7.547236  -8.207869  -8.889088  -8.131165
#>  [7]  -8.090413 -18.891223  -7.834767  -8.892308  -8.472040  -9.696238
#> > Overall mean
#> [1] -9.150939
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9050224 0.8904954 0.8996243 0.8864538 0.8587490 0.8921112 0.8880434
#>  [8] 0.1887362 0.8882691 0.8727013 0.8732868 0.8433085
#> > Overall mean
#> [1] 0.8239001
```

A.2.3.2 Text representation

```
text_representation(control_text_1_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0
#> i Method: Jaccard bag similarity
#> [1] 0
```

A.2.4 How similar is the *data questionnaire* when compared to the *Control Text 2*?

A.2.4.1 Text distance

```
text_distance(data_text_textembed, control_text_2_textembed)
#> i Method: Binary
#> > Line by line
#>  [1] 1 1 1 1 1 1 1 1 1 1 1 1
#> > Overall mean
#> [1] 1
#> i Method: Cosine
#> > Line by line
#>  [1] 0.9060095 0.9037570 0.8737437 0.8856842 0.7641104 0.9035066 0.9119419
#>  [8] 0.2057337 0.8851267 0.9036579 0.8678450 0.9198303
#> > Overall mean
#> [1] 0.8275789
#> i Method: Canberra
#> > Line by line
#>  [1] -464.0989 -465.0249 -497.9481 -493.0060 -520.5817 -473.6470 -454.5072
#>  [8] -627.1888 -485.4858 -469.6126 -488.5316 -461.6104
#> > Overall mean
#> [1] -491.7702
#> i Method: Euclidean
#> > Line by line
#>  [1]  -7.178356  -7.285667  -9.025508  -8.001721 -10.985822  -7.356390
#>  [7]  -6.953186 -18.747793  -7.828154  -7.079656  -8.591087  -6.716923
#> > Overall mean
#> [1] -8.812522
#> i Method: Manhattan
#> > Line by line
#>  [1] -173.8983 -180.4570 -213.3055 -192.0396 -188.6964 -182.8793 -171.7383
#>  [8] -174.9841 -188.5230 -170.6919 -210.6607 -166.2871
#> > Overall mean
#> [1] -184.5134
#> i Method: Maximum
#> > Line by line
#>  [1]  -0.39784696  -0.04475208  -0.72665224  -0.28434217  -5.73594077
#>  [6]  -0.20936910   0.05190219 -15.30314612  -0.55310996  -0.52536512
#> [11]  -0.10743206  -0.14595067
#> > Overall mean
#> [1] -1.9985
#> i Method: Minkowski
#> > Line by line
#>  [1]  -7.178356  -7.285667  -9.025508  -8.001721 -10.985822  -7.356390
#>  [7]  -6.953186 -18.747793  -7.828154  -7.079656  -8.591087  -6.716923
#> > Overall mean
#> [1] -8.812522
#> i Method: Pearson
#> > Line by line
#>  [1] 0.9060095 0.9037570 0.8737437 0.8856842 0.7641104 0.9035066 0.9119419
#>  [8] 0.2057337 0.8851267 0.9036579 0.8678450 0.9198303
#> > Overall mean
#> [1] 0.8275789
```

A.2.4.2 Text representation

```
text_representation(control_text_2_textreuse, data_text_textreuse)
#> i Method: Jaccard similarity
#> [1] 0
#> i Method: Jaccard bag similarity
#> [1] 0
```

APPENDIX B – CHAPTER 3 SUPPLEMENTAL MATERIAL

> ❗ Important
>
> You are reading the work-in-progress of this thesis.
>
> This chapter is currently a dumping ground for ideas, and I don’t recommend reading it.

APPENDIX C – CHAPTER 4 SUPPLEMENTAL MATERIAL

> **Note**
>
> You are reading the work-in-progress of this thesis.
>
> This chapter should be readable but is currently undergoing final polishing.

C.1 DATA WRANGLING

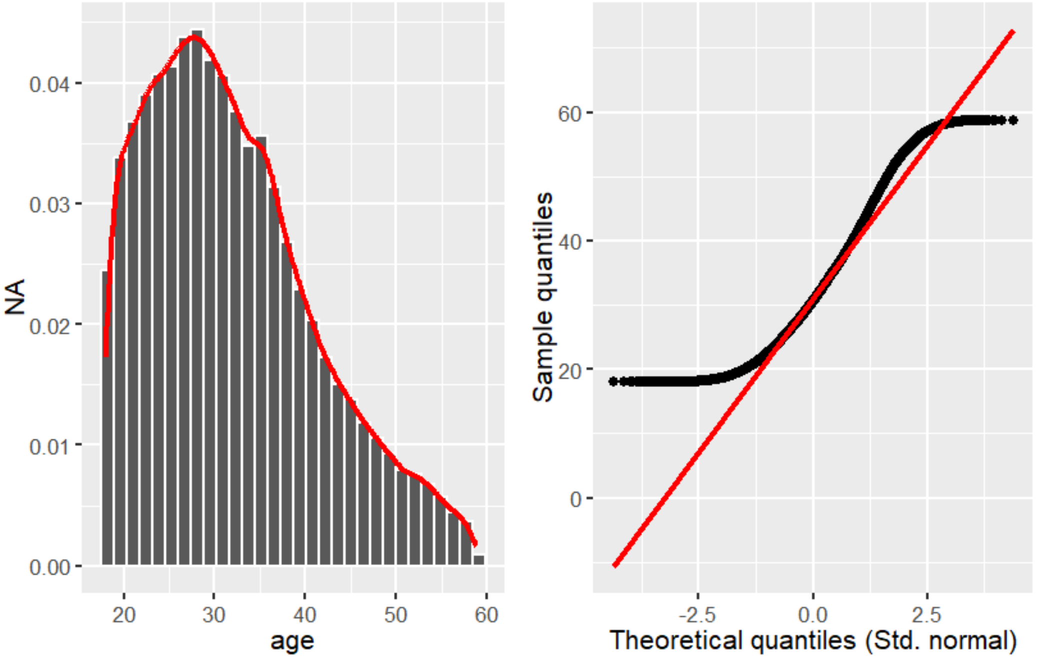
The data wrangling processes were performed using the `targets` R package. The full pipeline can be seen in the `_targets.R` file at the root of the research compendium.

```r
library(targets)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) ▷
  dplyr::select(
    age, sex, state, region, latitude, longitude, height, weight, work, study,
    msf_sc, sjl, le_week,
    ) ▷
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

C.2 DISTRIBUTION OF MAIN VARIABLES

```r
source(here::here("R/test_normality.R"))
source(here::here("R/utils.R"))

col <- "age"

stats <- data ▷
  magrittr::extract2(col) ▷
  test_normality(
    name = col,
    threshold = hms::parse_hms("12:00:00"),
    remove_outliers = FALSE,
    iqr_mult = 1.5,
    log_transform = FALSE,
    density_line = TRUE,
    text_size = env_vars$base_size,
    print = TRUE
  ) ▷
  rutils:::shush()
#> # A tibble: 14 x 2
#>   name    value
#>   <chr>   <chr>
#> 1 n       79198
#> 2 n_rm_na 79198
#> 3 n_na    0
#> 4 mean    31.9838074965417
#> 5 var     85.2414919292643
#> 6 sd      9.23263190695179
#> # i 8 more rows
```

Figure 6 – Frequencies of age among the sample subjects.



Source: Created by the author.

```r
stats$stats ▷ list_as_tibble()
```

Table 1 – Statistics related to the age variable.

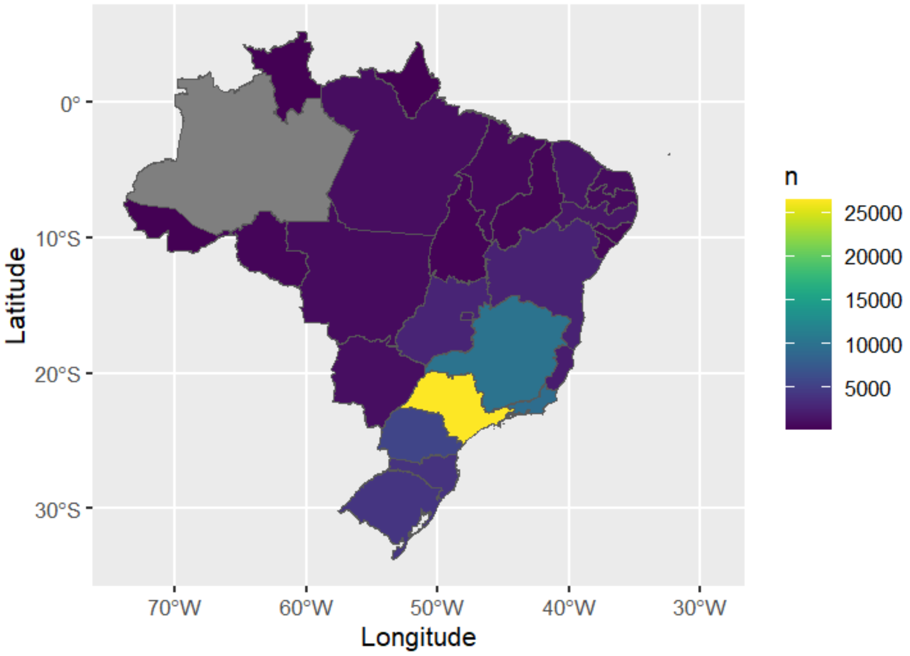
name	value
n	79198
n_rm_na	79198
n_na	0
mean	31.9838074965417
var	85.2414919292643
sd	9.23263190695179
min	18
q_1	24.7222222222222
median	30.5388888888889
q_3	37.61875
max	58.7861111111111
iqr	12.8965277777778
skewness	0.665751526654394
kurtosis	2.82381488030798

Source: Created by the author.

C.3 GEOGRAPHIC DISTRIBUTION

```r
source(here::here("R/plot_brazil_uf_map.R"))

rutils:::assert_internet()

brazil_uf_map <-
  data ▷
  plot_brazil_uf_map(option = "viridis", text_size = env_vars$base_size)
```

Figure 7 – Geographic distribution of the sample subjects.

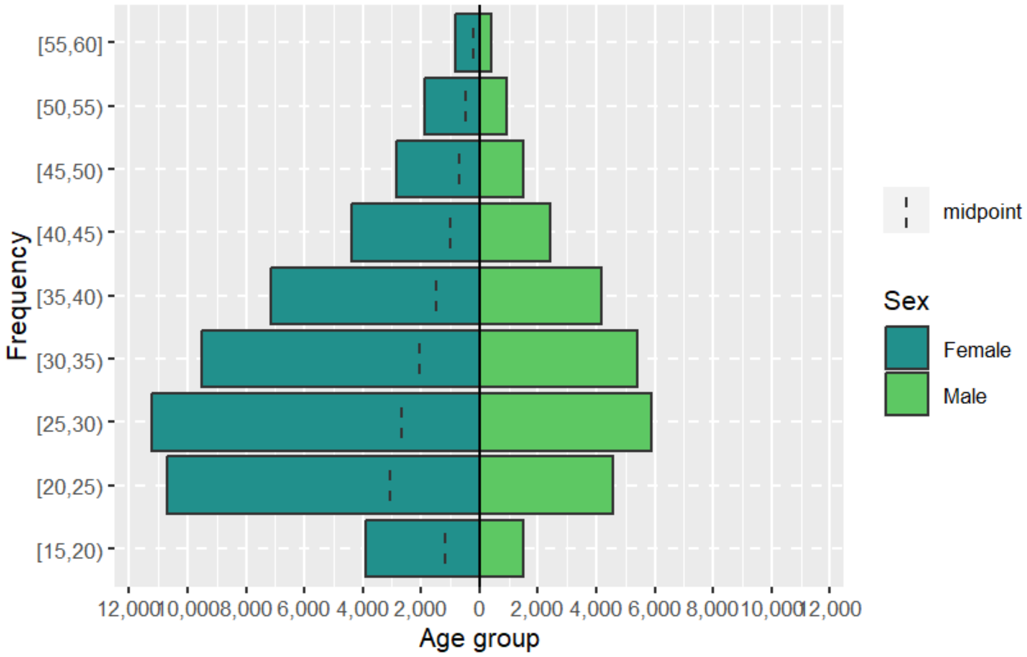


Source: Created by the author.

C.4 AGE PYRAMID

```r
source(here::here("R/plot_age_pyramid.R"))

age_pyramid <-
  data ▷
  plot_age_pyramid(
    interval = 10,
    na_rm = TRUE,
    text_size = env_vars$base_size
  )
```

Figure 8 – Age pyramid of the sample subjects.

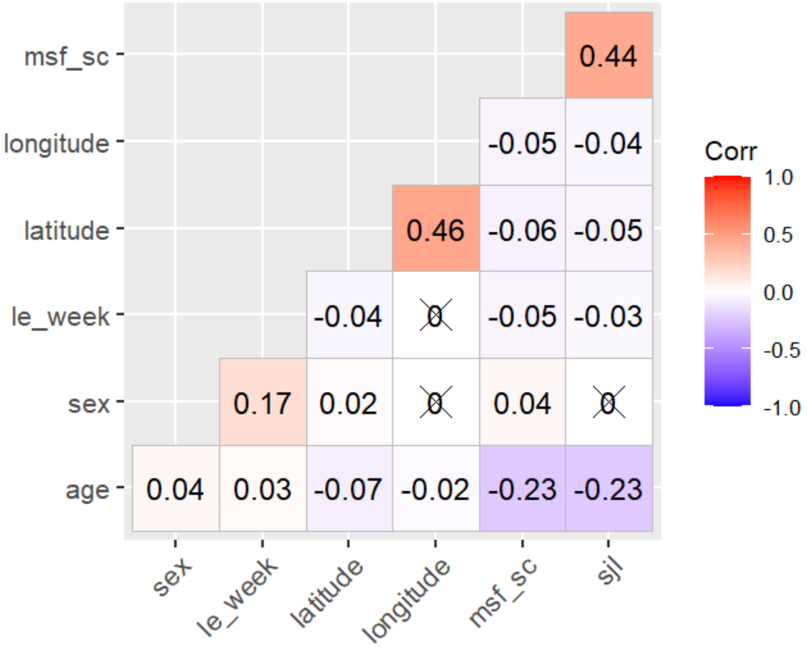


Source: Created by the author.

C.5 CORRELATION MATRIX

```
source(here::here("R/plot_ggcorrplot.R"))

cols <- c("sex", "age", "latitude", "longitude", "msf_sc", "sjl", "le_week")

ggcorrplot <-
  data ▷
  plot_ggcorrplot(
    cols = cols,
    na_rm = TRUE,
    text_size = env_vars$base_size,
    hc_order = TRUE
    )
```

Figure 9 – Correlation matrix of the main variables.



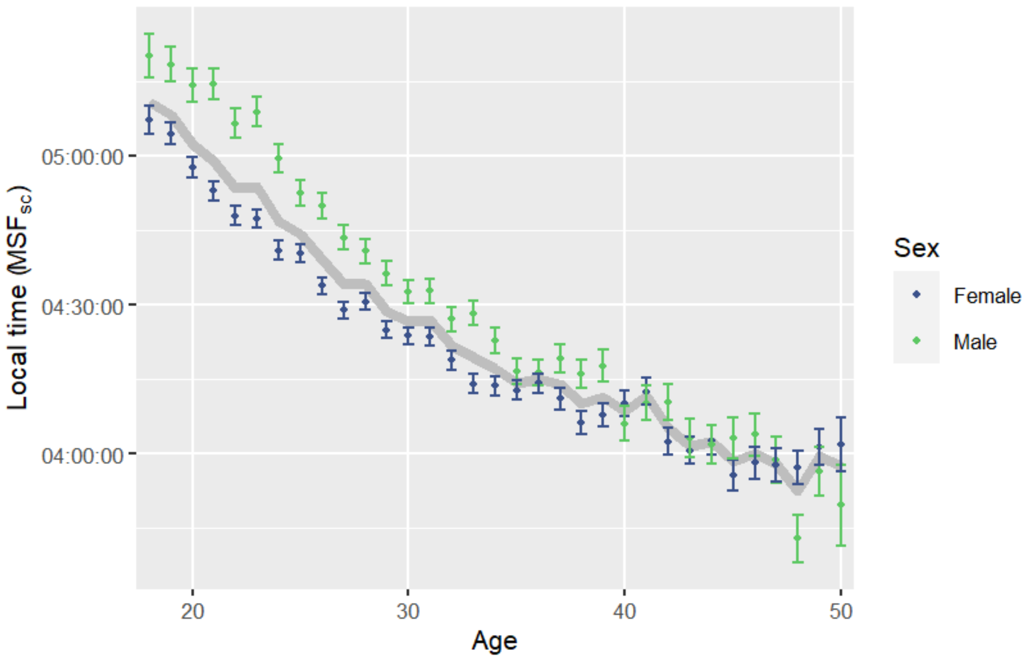
Source: Created by the author.

C.6 AGE AND SEX SERIES

C.6.1 **Age/sex *versus* chronotype**

```
source(here::here("R/plot_age_series.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("Local time ($MSF_{sc}$)")

data ▷
  dplyr::filter(age ⩽ 50) ▷
  plot_age_series(
    col = col,
    y_lab = y_lab,
    line_width = 2,
    boundary = 0.5,
    point_size = 1,
    error_bar_width = 0.5,
    error_bar_linewidth = 0.5,
    error_bar = TRUE,
    text_size = env_vars$base_size
    )
```

Figure 10 – Relation between age and chronotype, divided by sex. Chronotype is represented by the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype. The gray line represents both sex. Vertical lines represent the standard error of the mean (SEM).



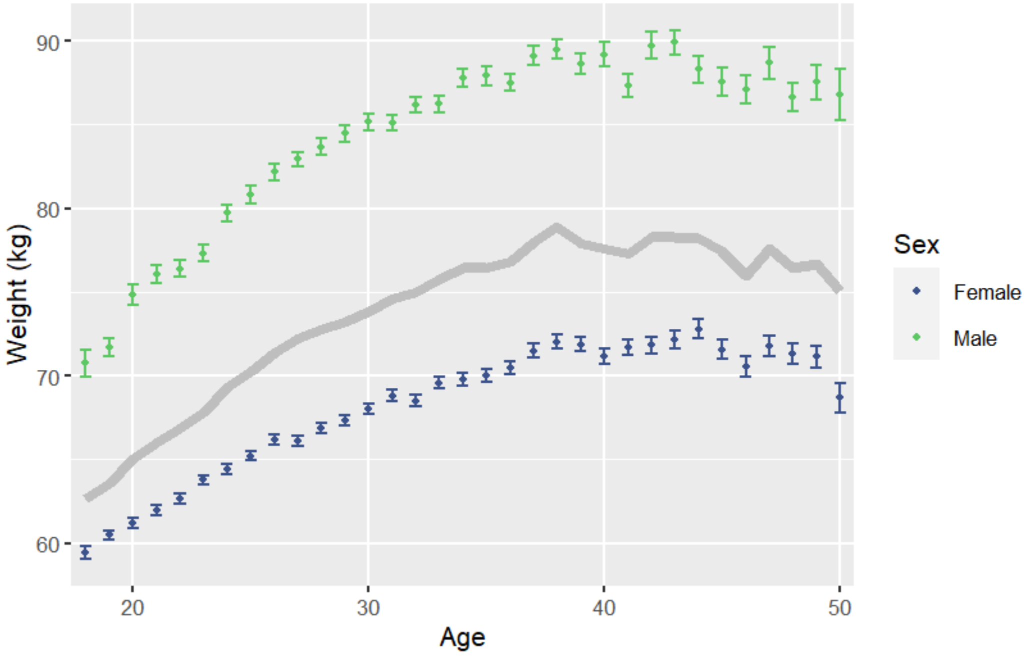
Source: Created by the author. Based on the data visualization found in Roenneberg, Kuehnle, et al. (2007).

C.6.2 Age/sex *versus* weight

```
source(here::here("R/plot_age_series.R"))

col <- "weight"
y_lab <- "Weight (kg)"

data |>
```

```r
dplyr::filter(age <= 50) |>
plot_age_series(
  col = col,
  y_lab = y_lab,
  line_width = 2,
  boundary = 0.5,
  point_size = 1,
  error_bar_width = 0.5,
  error_bar_linewidth = 0.5,
  error_bar = TRUE,
  text_size = env_vars$base_size
  )
```

Figure 11 – Relation between age and weight (kg), divided by sex. The gray line represents both sex. Vertical lines represent the standard error of the mean (SEM).

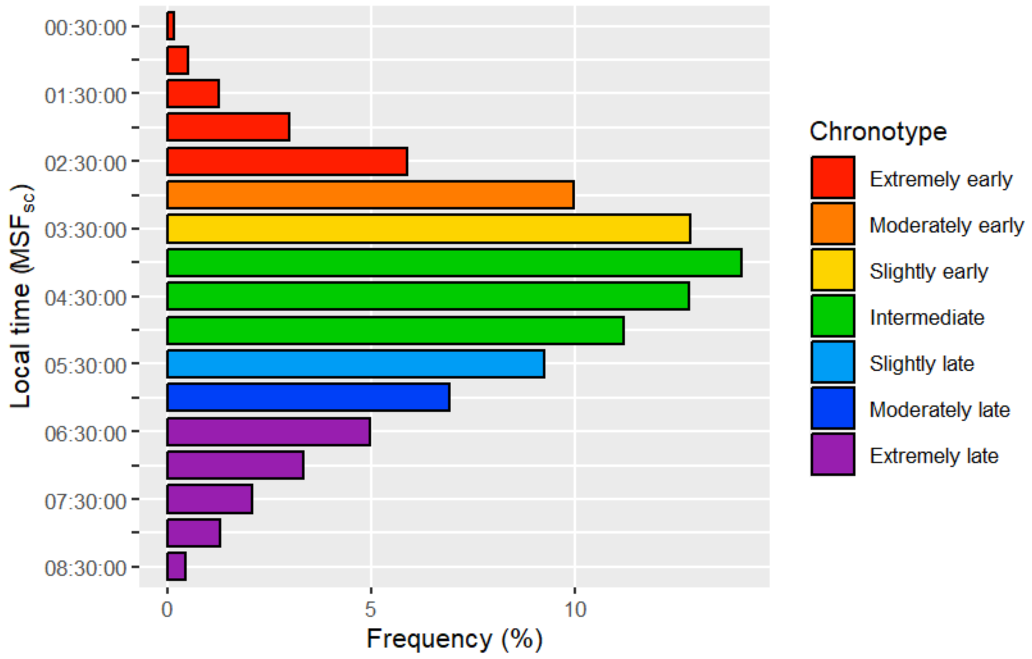


Source: Created by the author. Based on the data visualization found in Roenneberg, Kuehnle, et al. (2007).

C.7 CHRONOTYPE DISTRIBUTION

```r
source(here::here("R/plot_chronotype.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("Local time ($MSF_{sc}$)")

data |>
  plot_chronotype(
    col = col,
    x_lab = "Frequency (%)",
    y_lab = y_lab,
    col_width = 0.8,
    col_border = 0.6,
    text_size = env_vars$base_size,
    legend_position = "right",
    chronotype_cuts = FALSE
  )
```

Figure 12 – Distribution of the local time of the midpoint between sleep onset and sleep end on work-free days (MSF$_{sc}$), MCTQ proxy for measuring the chronotype. The categorical cuts follow a quantile approach going from extremely early $(0| - 0.11)$ to the extremely late $(0.88 - 1)$.

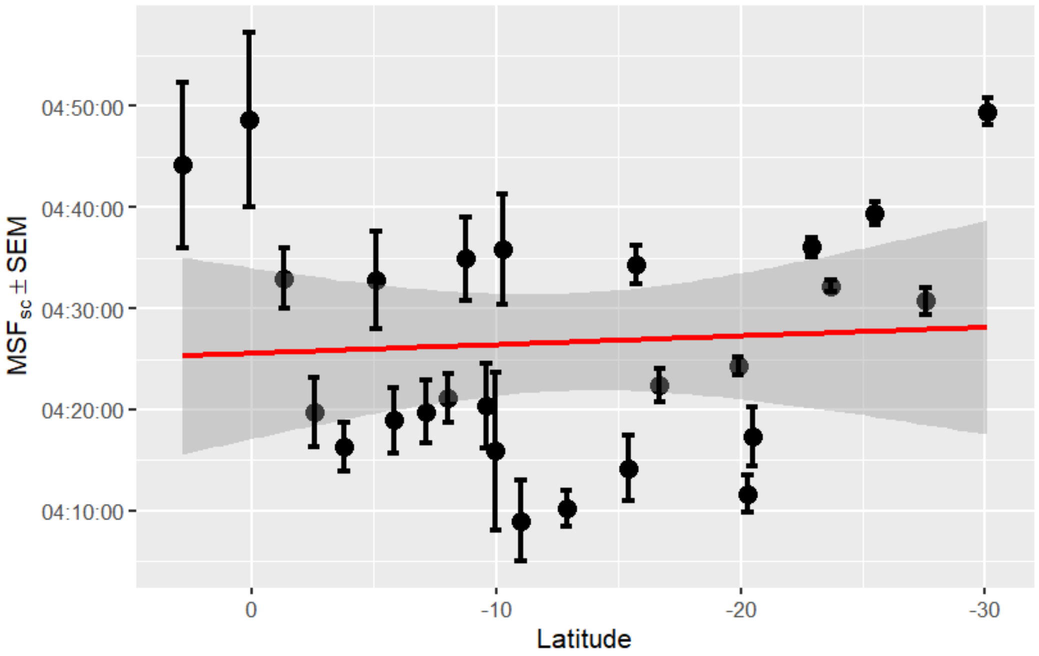


Source: Created by the author. Based on the data visualization found in Roenneberg, Pilz, et al. (2019).

C.8 LATITUDE SERIES

```
source(here::here("R/plot_latitude_series.R"))

col <- "msf_sc"
y_lab <- latex2exp::TeX("$MSF_{sc} \\pm SEM$")

data ▷
  dplyr::filter(age ⩽ 50) ▷
  plot_latitude_series(
    col = col,
    y_lab = y_lab,
    line_width = 2,
    point_size = 3,
    error_bar_width = 0.5,
    error_bar_linewidth = 1,
    error_bar = TRUE,
    text_size = env_vars$base_size
  )
```

Figure 13 – Distribution of mean aggregates of the local time of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, in relation to latitude decimal degree intervals. Higher values of MSF_{sc} indicate a tendency toward a late chronotype. The red line represents a linear regression, and the shaded area indicates a pointwise 95% confidence interval.



Source: Created by the author. Based on the data visualization found in Leocadio-Miguel et al. (2017).

C.9 STATISTICS

C.9.1 Numerical variables

```
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

col <- "msf_sc"

data |>
```

```r
  magrittr::extract2(col) |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

Table 2 – Statistics related to the MSF_{sc} variable.

name	value
n	79198
n_rm_na	79198
n_na	0
mean	04:28:17.770957
var	08:05:53.49992
sd	01:26:51.406096
min	00:22:30
q_1	03:26:25.714286
median	04:20:42.857143
q_3	05:25:42.857143
max	08:31:04.285714
iqr	01:59:17.142857
skewness	0.284586184927996
kurtosis	2.77321491178072

Source: Created by the author.

C.9.2 **Sex**

```r
# See <https://sidra.ibge.gov.br> to learn more.

library(magrittr)

rutils:::assert_internet()

# Brazil's 2022 census data
census_data <-
  sidrar::get_sidra(x = 9514) %>% # Don't change the pipe
  dplyr::filter(
    Sexo %in% c("Homens", "Mulheres", "Total"),
    stringr::str_detect(
      Idade,
      "^(1[8-9]|[2-9][0-9]) (ano|anos)$|^100 anos ou mais$"
    ),
    .[[17]] == "Total"
    ) ▷
  dplyr::transmute(
    sex = dplyr::case_when(
      Sexo == "Homens" ~ "Male",
      Sexo == "Mulheres" ~ "Female",
      Sexo == "Total" ~ "Total"
    ),
    value = Valor
  ) ▷
  dplyr::group_by(sex) ▷
  dplyr::summarise(n = sum(value)) ▷
  dplyr::ungroup()

census_data <-
  dplyr::bind_rows(
    census_data ▷
      dplyr::filter(sex ≠ "Total") ▷
      dplyr::mutate(
        n_rel = n / sum(n[sex ≠ "Total"]),
        n_per = round(n_rel * 100, 3)
      ),
    census_data ▷
      dplyr::filter(sex == "Total") ▷
      dplyr::mutate(n_rel = 1, n_per = 100)
  ) ▷
  dplyr::as_tibble() ▷
  dplyr::arrange(sex)

count <- data ▷
  dplyr::select(sex) ▷
  dplyr::group_by(sex) ▷
  dplyr::summarise(n = dplyr::n()) ▷
  dplyr::ungroup() ▷
  dplyr::mutate(
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
  ) ▷
  dplyr::arrange(dplyr::desc(n_rel)) ▷
  dplyr::bind_rows(
    dplyr::tibble(
      sex = "Total",
      n = nrow(tidyr::drop_na(data, sex)),
      n_rel = 1,
      n_per = 100
    )
  )

count <-
  dplyr::left_join(
    count, census_data,
    by = "sex",
    suffix = c("_sample", "_census")
  ) ▷
  dplyr::mutate(
    n_rel_diff = n_rel_sample - n_rel_census,
    n_per_diff = n_per_sample - n_per_census
  ) ▷
```

```r
  dplyr::relocate(
    sex, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
    n_per_sample, n_per_census, n_per_diff
  )

count |> dplyr::select(sex, n_per_sample, n_per_census, n_per_diff)
```

Table 3 – Frequencies of sex among subjects compared with data from Brazil’s 2022 census.

sex	n_per_sample	n_per_census	n_per_diff
Female	66.243	52.263	13.98
Male	33.757	47.737	-13.98
Total	100.000	100.000	0.00

Source: Created by the author, based on data from Brazil’s 2022 census (Instituto Brasileiro de Geografia e Estatística (n.d.-c)).

```r
sum(count$n_per_diff)
#> [1] -7.105427e-15
```

C.9.3 Age and sex

```r
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

value <- "Male"

data |>
  dplyr::filter(sex == value) |>
  magrittr::extract2("age") |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

Table 4 – Statistics related to male subject’s age.

name	value
n	26735
n_rm_na	26735
n_na	0
mean	32.4343759740665
var	80.9906211885464
sd	8.99947893983571
min	18
q_1	25.5388888888889
median	31.2583333333333
q_3	37.9319444444444
max	58.7722222222222
iqr	12.3930555555556
skewness	0.617696405622681
kurtosis	2.84390555184727

Source: Created by the author.

```r
# See <https://sidra.ibge.gov.br> to learn more.

library(magrittr)

rutils:::assert_internet()

# Brazil's 2022 census data
census_data <-
  sidrar::get_sidra(x = 9514) %>% # Don't change the pipe
  dplyr::filter(
    Sexo %in% c("Homens", "Mulheres", "Total"),
    stringr::str_detect(
      Idade,
      "^(1[8-9]|[2-9][0-9]) (ano|anos)$|^100 anos ou mais$"
    ),
    .[[17]] == "Total"
    ) |>
  dplyr::transmute(
    sex = dplyr::case_when(
      Sexo == "Homens" ~ "Male",
      Sexo == "Mulheres" ~ "Female",
      Sexo == "Total" ~ "Total"
    ),
    age = as.numeric(stringr::str_extract(Idade, "\\d+")),
    value = Valor
  ) |>
  dplyr::group_by(sex) |>
  dplyr::summarise(
    mean = stats::weighted.mean(age, value),
    sd = sqrt(Hmisc::wtd.var(age, value))
  ) |>
  dplyr::ungroup() |>
  dplyr::mutate(
    min = c(18, 18, 18),
    max = c(100, 100, 100)
  ) |>
  dplyr::relocate(sex, mean, sd, min, max) |>
  dplyr::as_tibble()

count <- data |>
  dplyr::select(sex, age) |>
  dplyr::group_by(sex) |>
  dplyr::mutate(sex = as.character(sex)) |>
  dplyr::summarise(
    mean = mean(age, na.rm = TRUE),
    sd = stats::sd(age, na.rm = TRUE),
    min = min(age, na.rm = TRUE),
    max = max(age, na.rm = TRUE)
    ) ▷
  dplyr::ungroup() ▷
  dplyr::bind_rows(
    dplyr::tibble(
      sex = "Total",
      mean = mean(data$age, na.rm = TRUE),
      sd = stats::sd(data$age, na.rm = TRUE),
      min = min(data$age, na.rm = TRUE),
      max = max(data$age, na.rm = TRUE)
    )
  )

count <-
  dplyr::left_join(
    count,
    census_data,
    by = "sex",
    suffix = c("_sample", "_census")
  ) ▷
  dplyr::mutate(mean_diff = mean_sample - mean_census) ▷
  dplyr::relocate(
    sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census,
    min_sample, min_census, max_sample, max_census
  )

count ▷
  dplyr::select(
    sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census
    )
```

Table 5 – Mean and standard deviation (sd) of subjects' age by sex compared with data from Brazil's 2022 census.

sex	mean_sample	mean_census	mean_diff	sd_sample	sd_census
Female	31.75420	44.98722	-13.23302	9.340939	17.51132
Male	32.43438	43.49903	-11.06465	8.999479	16.86385
Total	31.98381	44.27680	-12.29299	9.232632	17.22133

Source: Created by the author, based on data from Brazil's 2022 census (Instituto Brasileiro de Geografia e Estatística (n.d.-c)).

```
sum(count$mean_diff)
#> [1] -36.59066
```

C.9.4 **Longitudinal range**

C.9.4.1 Sample

```
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

stats <-
  data ▷
  magrittr::extract2("longitude") ▷
  stats_sum(print = FALSE)

abs(stats$max - stats$min)
#> [1] 33.023
stats ▷ list_as_tibble()
```

Table 6 – Statistics related to subject’s residential longitude.

name	value
n	79198
n_rm_na	79198
n_na	0
mean	-45.9455401815147
var	18.9406905927715
sd	4.35209037047388
min	-67.9869962
q_1	-48.4296364
median	-46.9249578
q_3	-43.7756411
max	-34.9639996
iqr	4.6539953
skewness	0.0156480710174436
kurtosis	5.78918700160139

Source: Created by the author.

C.9.4.2 Brazil

```
change_sign <- function(x) x * (-1)

## Ponta do Seixas, PB (7° 09′ 18″ S, 34° 47′ 34″ O)
min <-
  measurements::conv_unit("34 47 34", from = "deg_min_sec", to = "dec_deg") ▷
  as.numeric() ▷
  change_sign()

## Nascente do rio Moa, AC (7° 32′ 09″ S, 73° 59′ 26″ O)
max <-
  measurements::conv_unit("73 59 26", from = "deg_min_sec", to = "dec_deg") ▷
  as.numeric() ▷
  change_sign()

```

```r
min
#> [1] -34.79278
max
#> [1] -73.99056
abs(max - min)
#> [1] 39.19778
```

C.9.5 Latitudinal range

C.9.5.1 Sample

```r
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

stats <-
  data |>
  magrittr::extract2("latitude") |>
  stats_sum(print = FALSE)

abs(stats$max - stats$min)
#> [1] 32.91596
stats |> list_as_tibble()
```

Table 7 – Statistics related to subject’s residential latitude.

name	value
n	79198
n_rm_na	79198
n_na	0
mean	-20.8338507528991
var	40.2956396934244
sd	6.34788466289554
min	-30.1087672
q_1	-23.6820636
median	-23.6820636
q_3	-19.9026404
max	2.8071961
iqr	3.7794232
skewness	1.40629570823769
kurtosis	4.67433697579443

Source: Created by the author.

C.9.5.2 Brazil

```
change_sign <- function(x) x * (-1)

## Arroio Chuí, RS (33° 45′ 07″ S, 53° 23′ 50″ O)
min <-
  measurements::conv_unit("33 45 07", from = "deg_min_sec", to = "dec_deg") ▷
  as.numeric() ▷
  change_sign()

## Nascente do rio Ailã, RR (5° 16′ 19″ N, 60° 12′ 45″ O)
max <-
  measurements::conv_unit("5 16 19", from = "deg_min_sec", to = "dec_deg") ▷
  as.numeric()

min
```

```
#> [1] -33.75194
max
#> [1] 5.271944
abs(max - min)
#> [1] 39.02389
```

C.9.6 Region

```
# See <https://sidra.ibge.gov.br> to learn more.

rutils:::assert_internet()

# Brazil's 2022 census data
census_data <-
  sidrar::get_sidra(x = 4714, variable = 93, geo = "Region") ▷
  dplyr::select(dplyr::all_of(c("Valor", "Grande Região"))) ▷
  dplyr::transmute(
    col = `Grande Região`,
    n = Valor,
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
    ) ▷
  dplyr::mutate(
    col = dplyr::case_when(
      col == "Norte" ~ "North",
      col == "Nordeste" ~ "Northeast",
      col == "Centro-Oeste" ~ "Midwest",
      col == "Sudeste" ~ "Southeast",
      col == "Sul" ~ "South"
    )
  ) ▷
  dplyr::as_tibble() ▷
  dplyr::arrange(dplyr::desc(n_rel))
```

```r

count <- data |>
  magrittr::extract2("region") |>
  stats_sum(print = FALSE) |>
  magrittr::extract2("count") |>
  dplyr::mutate(
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
    ) |>
  dplyr::arrange(dplyr::desc(n_rel))

count <-
  dplyr::left_join(
    count, census_data, by = "col", suffix = c("_sample", "_census")
  ) |>
  dplyr::mutate(
    n_rel_diff = n_rel_sample - n_rel_census,
    n_per_diff = n_per_sample - n_per_census
  ) |>
  dplyr::relocate(
    col, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
    n_per_sample, n_per_census, n_per_diff
  )

count |> dplyr::select(col, n_per_sample, n_per_census, n_per_diff)
```

Table 8 – Frequencies of residential regions among subjects compared with data from Brazil’s 2022 census.

col	n_per_sample	n_per_census	n_per_diff
Southeast	60.565	41.777	18.788
South	17.122	14.742	2.380
Northeast	11.538	26.914	-15.376
Midwest	8.287	8.021	0.266
North	2.489	8.546	-6.057

Source: Created by the author, based on data from Brazil’s 2022 census (Instituto Brasileiro de Geografia e Estatística (n.d.-b)).

```
sum(count$n_per_diff)
#> [1] 0.001
```

C.9.7 State

```
source(here::here("R/stats_sum.R"))

data |>
  magrittr::extract2("state") |>
  stats_sum(print = FALSE) |>
  magrittr::extract2("count") |>
  dplyr::mutate(
    n_rel = n / sum(n),
    n_per = round(n_rel * 100, 3)
    ) |>
  dplyr::arrange(dplyr::desc(n_rel))
```

Table 9 – Frequencies of residential states among subjects compared with data from Brazil's 2022 census.

col	n	n_rel	n_per
São Paulo	26379	0.3330766	33.308
Minas Gerais	10115	0.1277179	12.772
Rio de Janeiro	9381	0.1184500	11.845
Paraná	5517	0.0696609	6.966
Rio Grande do Sul	4097	0.0517311	5.173
Santa Catarina	3946	0.0498245	4.982
Goiás	2674	0.0337635	3.376
Bahia	2522	0.0318442	3.184
Espírito Santo	2091	0.0264022	2.640
Distrito Federal	2087	0.0263517	2.635
Pernambuco	1550	0.0195712	1.957
Ceará	1398	0.0176520	1.765
Mato Grosso do Sul	1014	0.0128034	1.280
Pará	938	0.0118437	1.184
Rio Grande do Norte	789	0.0099624	0.996
Mato Grosso	788	0.0099497	0.995
Paraíba	773	0.0097603	0.976
Maranhão	652	0.0082325	0.823
Sergipe	533	0.0067300	0.673
Alagoas	526	0.0066416	0.664
Rondônia	401	0.0050633	0.506
Piauí	395	0.0049875	0.499
Tocantins	268	0.0033839	0.338
Acre	132	0.0016667	0.167
Roraima	119	0.0015026	0.150
Amapá	113	0.0014268	0.143

Source: Created by the author, based on data from Brazil's 2022 census (Instituto Brasileiro de Geografia e Estatística (n.d.-b)).

APPENDIX D – CHAPTER 5 SUPPLEMENTAL MATERIAL

! Important

You are reading the work-in-progress of this thesis.

This chapter is currently a dumping ground for ideas, and I don’t recommend reading it.

APPENDIX E – CHAPTER 6 SUPPLEMENTAL MATERIAL

ℹ Note

You are reading the work-in-progress of this thesis.

This chapter should be readable but is currently undergoing final polishing.

E.1 HYPOTHESIS

Statement

> Populations residing near the equator (latitude 0°) exhibit, on average, a shorter/morning circadian phenotype when compared to populations residing near the poles of the planet (Horzum et al., 2015; Hut et al., 2013; Leocadio-Miguel et al., 2014, 2017; Pittendrigh et al., 1991; Randler & Rahafar, 2017).

The study hypothesis was tested using nested models of multiple linear regressions. The main idea of nested models is to verify the effect of the inclusion of one or more predictors in the model variance explanation (i.e., the R^2) (Allen, 1997). This can be made by creating a restricted model and then comparing it with a full model. Hence, the hypothesis can be schematized as follows.

$$\begin{cases} \text{H}_0 : \text{R}^2_{\text{res}} >= \text{R}^2_{\text{full}} \\ \text{H}_a : \text{R}^2_{\text{res}} < \text{R}^2_{\text{full}} \end{cases}$$

The general equation for the F-test (Allen, 1997, p. 113) :

$$\text{F} = \frac{\text{R}^2_F - \text{R}^2_R / (k_F - k_R)}{(1 - \text{R}^2_F) / (\text{N} - k_F - 1)}$$

Where:

- R^2_F = Coefficient of determination for the **full** model
- R^2_R = Coefficient of determination for the **restricted** model
- k_F = Number of independent variables in the full model
- k_R = Number of independent variables in the restricted model
- N = Number of observations in the sample

$$\text{F} = \frac{\text{Additional Var. Explained}/\text{Additional d.f. Expended}}{\text{Var. unexplained}/\text{d.f. Remaining}}$$

It's important to note that, in addition to the F-test, it's assumed that for R^2_{res} to differ significantly from R^2_{full}, there must be a non-negligible effect size between them. This effect size can be calculated using Cohen's f^2 (Cohen, 1988, 1992):

$$f^2 = \frac{\text{R}^2_F - \text{R}^2_R}{1 - \text{R}^2_F}$$

$$f^2 = \frac{\text{Additional Var. Explained}}{\text{Var. unexplained}}$$

E.2 A BRIEF LOOK ON GENERAL LINEAR MODELS

See DeGroot and Schervish (2012, pp. 699-707, pp. 736-754) and Hair (2019, pp. 259-370) to learn more.

"[…] A problem of this type is called a problem of multiple linear regression because we are considering the regression of Y on k variables $X_1, \dots, X_k$, rather than on just a single variable X, and we are assuming also that this regression is a linear function of the parameters $\beta_0, \dots, \beta_k$. In a problem of multiple linear regressions, we obtain n vectors of observations ($x_{i1}. \dots, x_{ik}, Y_i$), for $i = 1, \dots, n$. Here x_{ij} is the observed value of the variable X_j for the ith observation. The EY is given by the relation

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

(DeGroot & Schervish, 2012, p. 738)

E.2.1 **Definitions**

Residuals/Fitted Values

For $i = 1, \dots, n$, the observed values of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are called *fitted values*. For $i = 1, \dots, n$, the observed values of $e_i = y_i - \hat{y}_i$ are called *residuals* (DeGroot & Schervish, 2012, p. 717).

"[…] regression problems in which the observations $Y_i, \ldots, Y_n$ […] we shall assume that each observation Y_i has a normal distribution, that the observations $Y_1, \ldots, Y_n$ are independent, and that the observations $Y_1, \ldots, Y_n$ have the same variance σ^2. Instead of a single predictor being associated with each Y_i, we assume that a p-dimensional vector $z_i = (z_{i0}, \ldots, z_{ip-1})$ is associated with each Y_i" (DeGroot & Schervish, 2012, p. 736).

General Linear Model The statistical model in which the observations $Y_1, \ldots, Y_n$ satisfy the following assumptions (DeGroot & Schervish, 2012, p. 738).

E.2.2 Assumptions

Assumption 1

Predictor is known. Either the vectors $z_1, \ldots, z_n$ are known ahead of time, or they are the observed values of random vectors $Z_1, \ldots, Z_n$ on whose values we condition before computing the joint distribution of ($Y_1, \ldots, Y_n$) (DeGroot & Schervish, 2012, p. 736).

Age and sex are known predictors for the chronotype (Roenneberg, Kuehnle, et al., 2007).

Assumption 2

Normality. For $i = 1, \ldots, n$, the conditional distribution of Y_i given the vectors $z_1, \ldots, z_n$ is a normal distribution (DeGroot & Schervish, 2012, p. 737).

(Normality of the error term distribution (Hair, 2019, p. 287)).

As it will be seen in the next topics, without any transformation, the chronotype variable does not have a normal distribution. However, this can be satisfied with a Box-Cox transformation (see Box and Cox (1964)).

A residual diagnostics will test the assumption of normality of the error term distribution.

Assumption 3

Linear mean. There is a vector of parameters $\beta = (\beta_0, \ldots, \beta_{p-1})$ such that the conditional mean of Y_i given the values $z_1, \ldots, z_n$ has the form

$$z_{i0}\beta_0 + z_{i1}\beta_1 + \cdots + z_{ip-1}\beta_{p-1}$$

for $i = 1, \dots, n$ (DeGroot & Schervish, 2012, p. 737).

(Linearity of the phenomenon measured (Hair, 2019, p. 287)).

The hypothesis assumes a linear relation.

Assumption 4

> **Common variance**. There is as parameter σ^2 such the conditional variance of Y_i given the values $z_1, \dots, z_n$ is σ^2 for $i = 1, \dots n$.

(Constant variance of the error terms (Hair, 2019, p. 287))

The presence of unequal variances (heteroscedasticity) will be tested with a residual diagnostics.

Assumption 5

> **Independence**. The random variables $Y_1, \dots, Y_n$ are independent given the observed $z_1, \dots, z_n$ (DeGroot & Schervish, 2012, p. 737).

(Independence of the error terms (Hair, 2019, p. 287)).

This will also be tested with a residual diagnostics.

E.3 DATA PREPARATION

Outlier treatment (for now): 1.5x Interquartile range (IQR) for age and chronotype (MSF$_{sc}$).

```r
is_outlier <- function(x, method = "iqr", iqr_mult = 1.5, sd_mult = 3) {
  checkmate::assert_numeric(x)
  checkmate::assert_choice(method, c("iqr", "sd"))
  checkmate::assert_number(iqr_mult)
  checkmate::assert_number(sd_mult)

  if (method == "iqr") {
    iqr <- stats::IQR(x, na.rm = TRUE)
    min <- stats::quantile(x, 0.25, na.rm = TRUE) - (iqr_mult * iqr)
    max <- stats::quantile(x, 0.75, na.rm = TRUE) + (iqr_mult * iqr)
  } else if (method == "sd") {
    min <- mean(x, na.rm = TRUE) - (sd_mult * stats::sd(x, na.rm = TRUE))
    max <- mean(x, na.rm = TRUE) + (sd_mult * stats::sd(x, na.rm = TRUE))
  }

  dplyr::if_else(x >= min & x <= max, FALSE, TRUE, missing = FALSE)
}
```

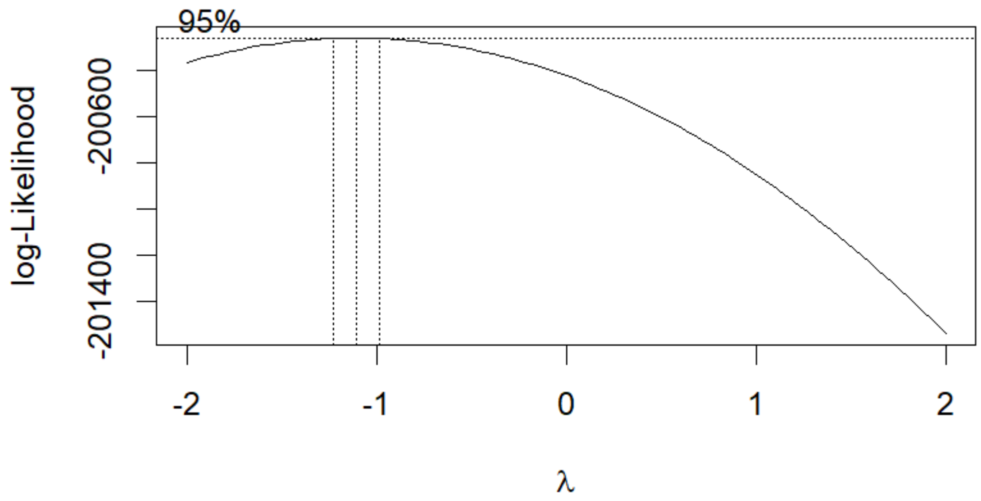
```
source(here::here("R/utils.R"))

utc_minus_3_states <- c(
  "Amapá", "Pará", "Maranhão", "Tocantins", "Piauí", "Ceará",
  "Rio Grande do Norte", "Paraíba", "Pernambuco", "Alagoas", "Sergipe",
  "Bahia", "Distrito Federal", "Goiás", "Minas Gerais", "Espírito Santo",
  "Rio de Janeiro", "São Paulo", "Paraná", "Santa Catarina",
  "Rio Grande do Sul"
)

data <-
  targets::tar_read("geocoded_data", store = here::here("_targets")) |>
  dplyr::filter(state %in% utc_minus_3_states) |>
  dplyr::select(msf_sc, age, sex, state, latitude, longitude) |>
  dplyr::mutate(msf_sc = transform_time(msf_sc)) |>
  tidyr::drop_na(msf_sc, age, sex, latitude)
```

E.4 RESTRICTED MODEL

E.4.1 Model building

```
box_cox <- MASS::boxcox(msf_sc ~ age + sex, data = data)
```

Table 10 – Profile of log-likelihoods for the parameter (λ) of the Box-Cox power transformation for the restricted model.



Source: Created by the author. See Box and Cox (1964) to learn more.

```r
lambda <- box_cox$x[which.max(box_cox$y)]

lambda
#> [1] -1.1111
```

```r
res_model <- stats::lm(
  ((msf_sc^lambda - 1) / lambda) ~ age + sex, data = data
)
```

```r
broom::tidy(res_model)
```

Table 11 – Summarized information about the components of the restricted model.

term	estimate	std.error	statistic	p.value
(Intercept)	0.9	0	513579298.250	0
age	0.0	0	-65.128	0
sexMale	0.0	0	13.020	0

Source: Created by the author.

```
broom::glance(res_model) ▷ tidyr::pivot_longer(cols = dplyr::everything())
```

Table 12 – Summarized statistics about the restricted model.

name	value
r.squared	0.05373
adj.r.squared	0.05371
sigma	0.00000
statistic	2178.87560
p.value	0.00000
df	2.00000
logLik	1106194.89709
AIC	-2212381.79419
BIC	-2212344.80126
deviance	0.00000
df.residual	76741.00000
nobs	76744.00000

Source: Created by the author.

```
res_model ▷ summary()
#>
#> Call:
#> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex,
#>     data = data)
#>
#> Residuals:
```

```
#>           Min            1Q        Median            3Q           Max
#> -0.0000004859 -0.0000000911 -0.0000000031  0.0000000916  0.0000004204
#>
#> Coefficients:
#>                      Estimate       Std. Error     t value Pr(>|t|)
#> (Intercept)  0.8999976603602  0.0000000017524 513579298.2   <2e-16 ***
#> age         -0.0000000033812  0.0000000000519       -65.1   <2e-16 ***
#> sexMale      0.0000000132309  0.0000000010162        13.0   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.000000133 on 76741 degrees of freedom
#> Multiple R-squared:  0.0537, Adjusted R-squared:  0.0537
#> F-statistic: 2.18e+03 on 2 and 76741 DF,  p-value: <2e-16
```

E.4.2 **Residual diagnostics**

E.4.2.1 Normality

```
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

res_model ▷
  stats::residuals() ▷
  stats_sum(print = FALSE) ▷
  list_as_tibble()
```

Table 13 – Statistics about the restricted model residuals.

name	value
n	76744
n_rm_na	76744
n_na	0
mean	6.60699976667332e-23
var	0.0000000000000176852866826985
sd	0.000000132986039427823
min	-0.000000485865195534305
q_1	-0.0000000911138016567908
median	-0.00000000313530324787135
q_3	0.000000091553820345483
max	0.000000420368932360539
iqr	0.000000182667622002274
skewness	-0.0105262146639209
kurtosis	2.82813923301771

Source: Created by the author.

```
source(here::here("R/normality_sum.R"))

res_model ▷
  stats::residuals() ▷
  normality_sum()
```

Table 14 – Normality tests about the restricted model residuals.

test	p_value
Anderson-Darling	0.00000
Bonett-Seier	0.00000
Cramer-von Mises	0.00000
D’Agostino Omnibus Test	NA
D’Agostino Skewness Test	0.23383
D’Agostino Kurtosis Test	NA
Jarque–Bera	0.00000
Lilliefors (K-S)	0.00000
Pearson chi-square	0.00000
Shapiro-Francia	NA
Shapiro-Wilk	NA

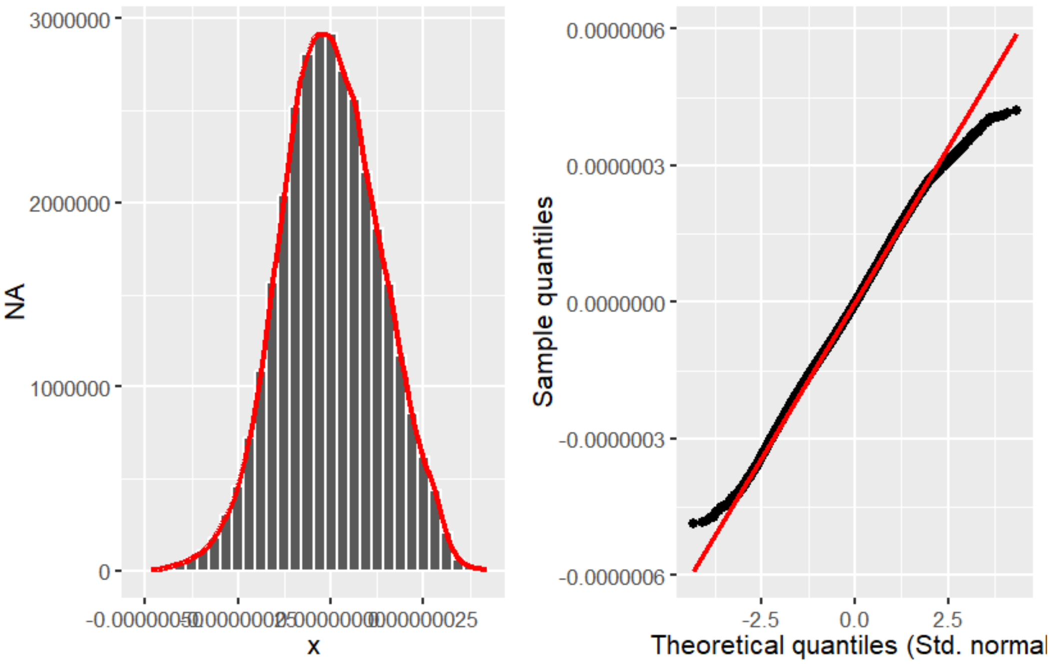
Source: Created by the author.

Correlation between observed residuals and expected residuals under normality.

```
res_model ▷ olsrr::ols_test_correlation()
#> [1] 0.99929
```

```
source(here::here("R/test_normality.R"))

# res_model ▷ olsrr::ols_plot_resid_qq()

qq_plot <- res_model ▷
  stats::residuals() ▷
  plot_qq(print = FALSE)

hist_plot <- res_model ▷
  stats::residuals() ▷
  plot_hist(print = FALSE)

cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)
```

Figure 14 – Histogram of the restricted model residuals with a kernel density estimate, along with a quantile-quantile (Q-Q) plot between the residuals and the theoretical quantiles of the normal distribution.

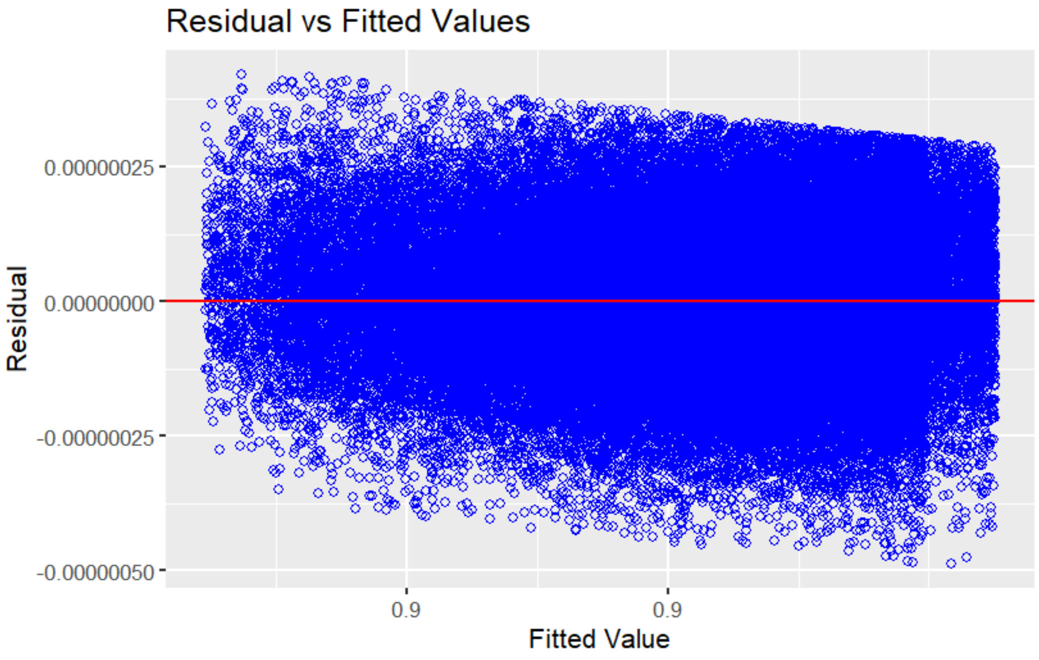


Source: Created by the author.

E.4.2.2 Common variance

```r
res_model ▷ olsrr::ols_plot_resid_fit()
```

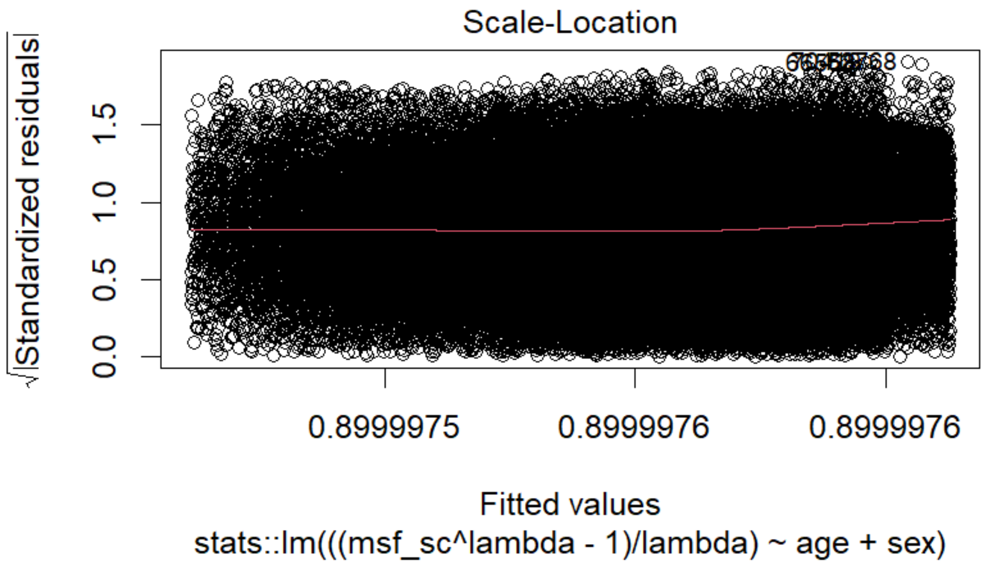
Figure 15 – Relation between the fitted values of the restricted model and its residuals.



Source: Created by the author.

```r
res_model |> plot(3)
```

Figure 16 – Relation between the fitted values of the restricted model and its standardized residuals.



Source: Created by the author.

```
res_model |> olsrr::ols_test_breusch_pagan()
#>
#>  Breusch Pagan Test for Heteroskedasticity
#>  -----------------------------------------
#>  Ho: the variance is constant
#>  Ha: the variance is not constant
#>
#>                          Data
#>  --------------------------------------------------------
#>  Response : ((msf_sc^lambda - 1)/lambda)
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>         Test Summary
#>  -----------------------------
#>  DF            =    1
#>  Chi2          =    70149.3586
#>  Prob > Chi2   =    0.0000
```

```
res_model |> olsrr::ols_test_score()
#>
#>  Score Test for Heteroskedasticity
#>  ---------------------------------
#>  Ho: Variance is homogenous
#>  Ha: Variance is not homogenous
#>
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>       Test Summary
#>  ------------------------
#>  DF            =    1
#>  Chi2          =    0.000
#>  Prob > Chi2   =    1.000
```

E.4.2.3 Independence

Variance inflation factor (VIF)

> "Indicator of the effect that the other independent variables have on the standard error of a regression coefficient. The variance inflation factor is directly related to the tolerance value ($\text{VIF}_i = 1/\text{TO}L$). Large VIF values also indicate a high degree of collinearity or multicollinearity among the independent variables" (Hair, 2019, p. 265).

```
res_model ▷ olsrr::ols_coll_diag()
#> Tolerance and Variance Inflation Factor
#> ---------------------------------------
#>   Variables Tolerance    VIF
#> 1       age    0.9988 1.0012
#> 2   sexMale    0.9988 1.0012
#>
#>
#> Eigenvalue and Condition Index
#> ------------------------------
#>   Eigenvalue Condition Index intercept      age   sexMale
#> 1   2.422418          1.0000  0.011753 0.011936 0.0669897
#> 2   0.538450          2.1211  0.015824 0.018848 0.9280439
#> 3   0.039132          7.8679  0.972423 0.969216 0.0049664
```

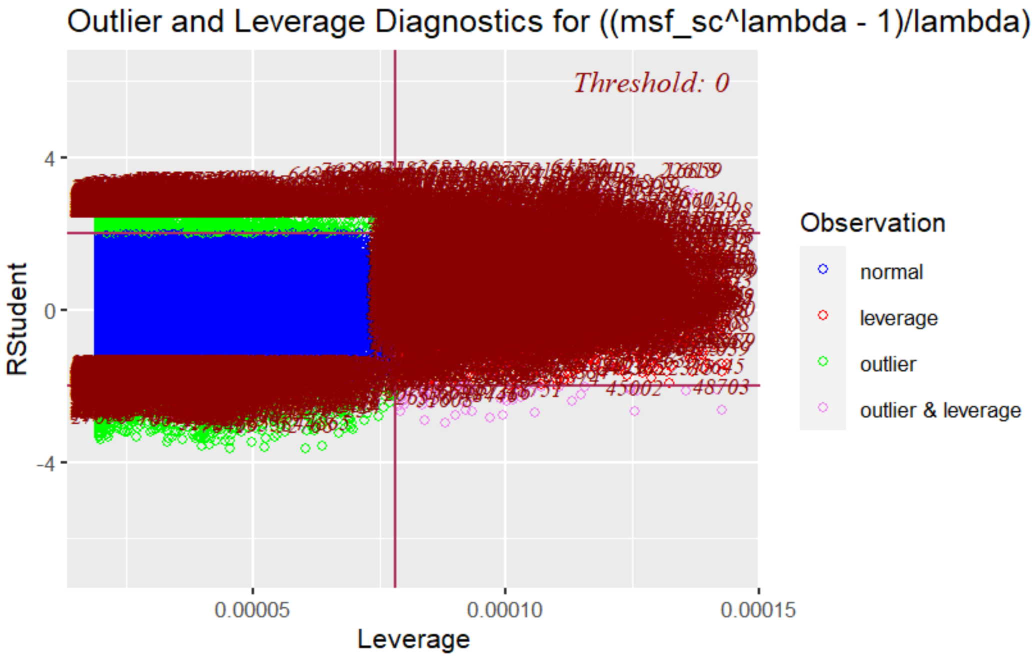
E.4.2.4 Measures of influence

Leverage points

> "Type of *influential observation* defined by one aspect of influence termed *leverage*. These observations are substantially different on one or more independent variables, so that they affect the estimation of one or more *regression coefficients*" (Hair, 2019, p. 262).

```
res_model ▷ olsrr::ols_plot_resid_lev()
```

Figure 17 – Relation between the restricted model studentized residuals and their leverage/influence.



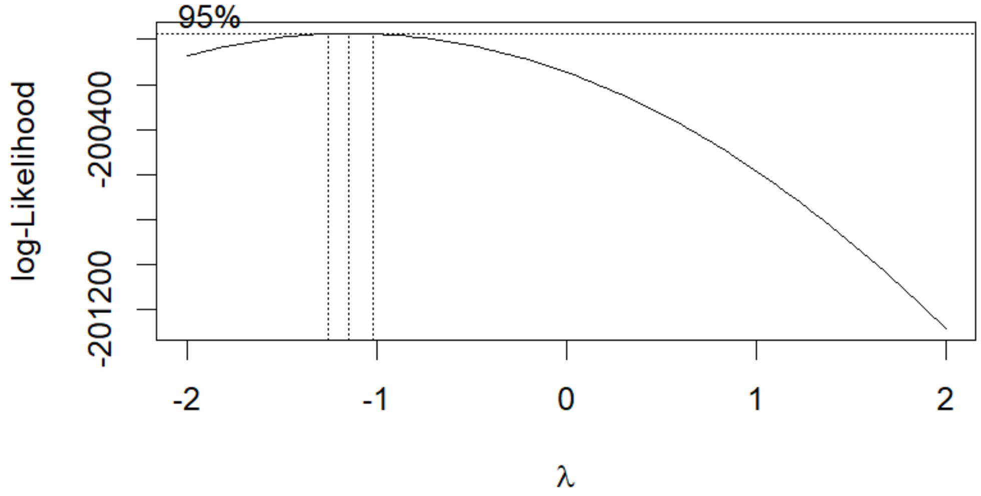
Source: Created by the author.

E.5 FULL MODEL

E.5.1 **Model building**

```r
box_cox <- MASS::boxcox(
  msf_sc ~ age + sex + latitude, data = data
  )
```

Table 15 – Profile of log-likelihoods for the parameter (λ) of the Box-Cox power transformation for the full model.



Source: Created by the author. See Box and Cox (1964) to learn more.

```r
box_cox$x[which.max(box_cox$y)] # lambda
#> [1] -1.1515
```

```r
lambda # The same lambda of the restricted model
#> [1] -1.1111
```

```r
full_model <- stats::lm(
  ((msf_sc^lambda - 1) / lambda) ~ age + sex + latitude,
  data = data
  )
```

```r
broom::tidy(full_model)
```

Table 16 – Summarized information about the components of the full model.

term	estimate	std.error	statistic	p.value
(Intercept)	0.9	0	391908052.847	0
age	0.0	0	-66.928	0
sexMale	0.0	0	13.558	0
latitude	0.0	0	-23.852	0

Source: Created by the author.

```r
broom::glance(full_model) |>
  tidyr::pivot_longer(cols = dplyr::everything())
```

Table 17 – Summarized statistics about the full model.

name	value
r.squared	0.06070
adj.r.squared	0.06066
sigma	0.00000
statistic	1652.97928
p.value	0.00000
df	3.00000
logLik	1106478.33068
AIC	-2212946.66136
BIC	-2212900.42021
deviance	0.00000
df.residual	76740.00000
nobs	76744.00000

Source: Created by the author.

```r
full_model |> summary()
#>
#> Call:
#> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex +
#>     latitude, data = data)
```

```
#>
#> Residuals:
#>          Min            1Q        Median            3Q           Max
#> -0.0000004874 -0.0000000911 -0.0000000034  0.0000000912  0.0000004328
#>
#> Coefficients:
#>                    Estimate       Std. Error     t value Pr(>|t|)
#> (Intercept)  0.8999976247783  0.0000000022965 391908052.9   <2e-16 ***
#> age         -0.0000000034710  0.0000000000519       -66.9   <2e-16 ***
#> sexMale      0.0000000137296  0.0000000010127        13.6   <2e-16 ***
#> latitude    -0.0000000018222  0.0000000000764       -23.9   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.000000132 on 76740 degrees of freedom
#> Multiple R-squared:  0.0607, Adjusted R-squared:  0.0607
#> F-statistic: 1.65e+03 on 3 and 76740 DF,  p-value: <2e-16
```

E.5.2 Residual diagnostics

E.5.2.1 Normality

```r
source(here::here("R/stats_sum.R"))
source(here::here("R/utils.R"))

full_model |>
  stats::residuals() |>
  stats_sum(print = FALSE) |>
  list_as_tibble()
```

Table 18 – Statistics about the full model residuals.

name	value
n	76744
n_rm_na	76744
n_na	0
mean	4.85272564733669e-24
var	0.0000000000000175551361304561
sd	0.000000132495796652O3
min	-0.000000487410752460545
q_1	-0.0000000910649425186321
median	-0.000000003374344652286
q_3	0.0000000911899588839585
max	0.000000432826012898983
iqr	0.000000182254901402591
skewness	0.000655994107765645
kurtosis	2.82688323293117

Source: Created by the author.

```r
source(here::here("R/normality_sum.R"))

full_model |>
  stats::residuals() |>
  normality_sum()
```

Table 19 – Normality tests about the full model residuals.

test	p_value
Anderson-Darling	0.00000
Bonett-Seier	0.00000
Cramer-von Mises	0.00000
D’Agostino Omnibus Test	NA
D’Agostino Skewness Test	0.94085
D’Agostino Kurtosis Test	NA
Jarque–Bera	0.00000
Lilliefors (K-S)	0.00000
Pearson chi-square	0.00000
Shapiro-Francia	NA
Shapiro-Wilk	NA

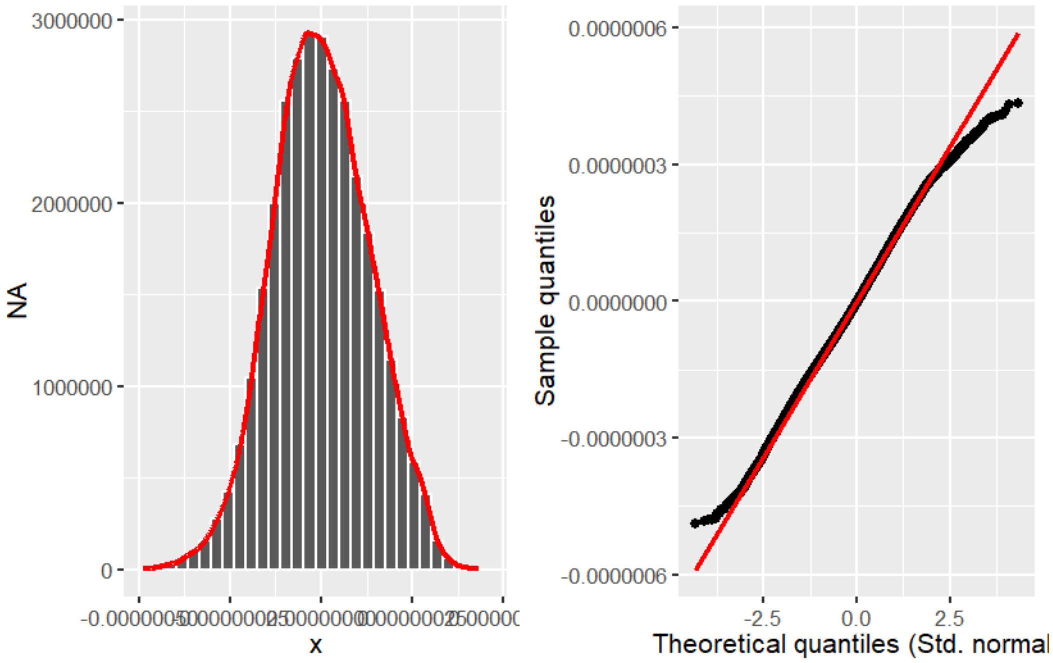
Source: Created by the author.

Correlation between observed residuals and expected residuals under normality.

```r
full_model ▷ olsrr::ols_test_correlation()
#> [1] 0.99929
```

```r
source(here::here("R/test_normality.R"))

hist_plot <- full_model ▷
  stats::residuals() ▷
  plot_hist(print = FALSE)

qq_plot <- full_model ▷
  stats::residuals() ▷
  plot_qq(print = FALSE)

cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)
```

Figure 18 – Histogram of the full model residuals with a kernel density estimate, along with a quantile-quantile (Q-Q) plot between the residuals and the theoretical quantiles of the normal distribution.

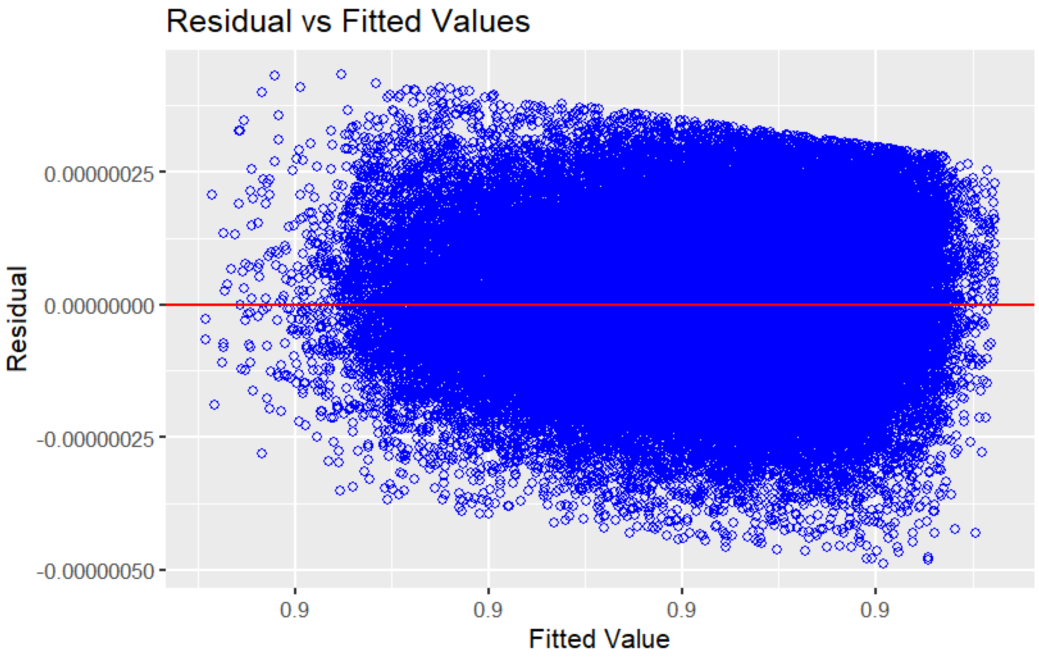


Source: Created by the author.

E.5.2.2 Common variance

```
full_model ▷ olsrr::ols_plot_resid_fit()
```

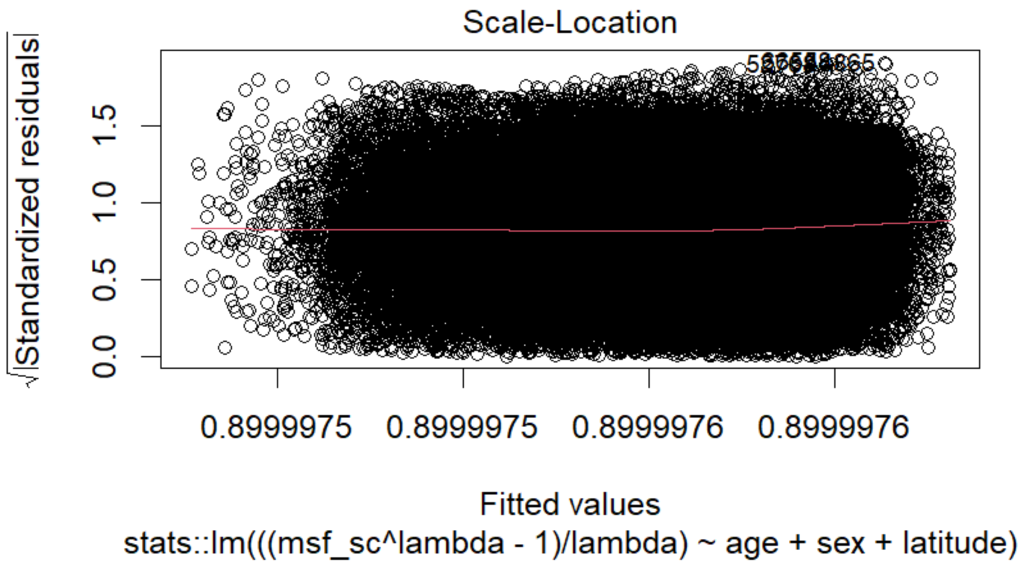
Figure 19 – Relation between the fitted values of the full model and its residuals.



Source: Created by the author.

```
full_model ▷ plot(3)
```

Figure 20 – Relation between the fitted values of the full model and its standardized residuals.



Source: Created by the author.

```
full_model |> olsrr::ols_test_breusch_pagan()
#>
#>  Breusch Pagan Test for Heteroskedasticity
#>  -----------------------------------------
#>  Ho: the variance is constant
#>  Ha: the variance is not constant
#>
#>                           Data
#>  --------------------------------------------------------
#>  Response : ((msf_sc^lambda - 1)/lambda)
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>         Test Summary
#>  -----------------------------
#>  DF            =    1
#>  Chi2          =    70101.1634
#>  Prob > Chi2   =    0.0000
```

```
full_model |> olsrr::ols_test_score()
#>
#>  Score Test for Heteroskedasticity
#>  ---------------------------------
#>  Ho: Variance is homogenous
#>  Ha: Variance is not homogenous
#>
#>  Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
#>
#>       Test Summary
#>  ------------------------
#>  DF            =    1
#>  Chi2          =    0.000
#>  Prob > Chi2   =    1.000
```

E.5.2.3 Independence

Variance inflation factor (VIF)

> “Indicator of the effect that the other independent variables have on the standard error of a regression coefficient. The variance inflation factor is directly related to the tolerance value (VIF$_i = 1/\text{TO}L$). Large VIF values also indicate a high degree of collinearity or multicollinearity among the independent variables” (Hair, 2019, p. 265).

```
full_model |> olsrr::ols_coll_diag()
#> Tolerance and Variance Inflation Factor
#> ---------------------------------------
#>   Variables Tolerance    VIF
#> 1       age   0.99354 1.0065
#> 2   sexMale   0.99838 1.0016
#> 3  latitude   0.99441 1.0056
#>
#>
#> Eigenvalue and Condition Index
#> ------------------------------
#>   Eigenvalue Condition Index  intercept       age   sexMale  latitude
#> 1   3.312504          1.0000 0.00377395 0.0064918 0.0304493 0.0068553
#> 2   0.584652          2.3803 0.00328127 0.0064143 0.9588857 0.0083393
#> 3   0.073700          6.7042 0.00040414 0.5063551 0.0023826 0.5659326
#> 4   0.029145         10.6609 0.99254063 0.4807389 0.0082824 0.4188728
```

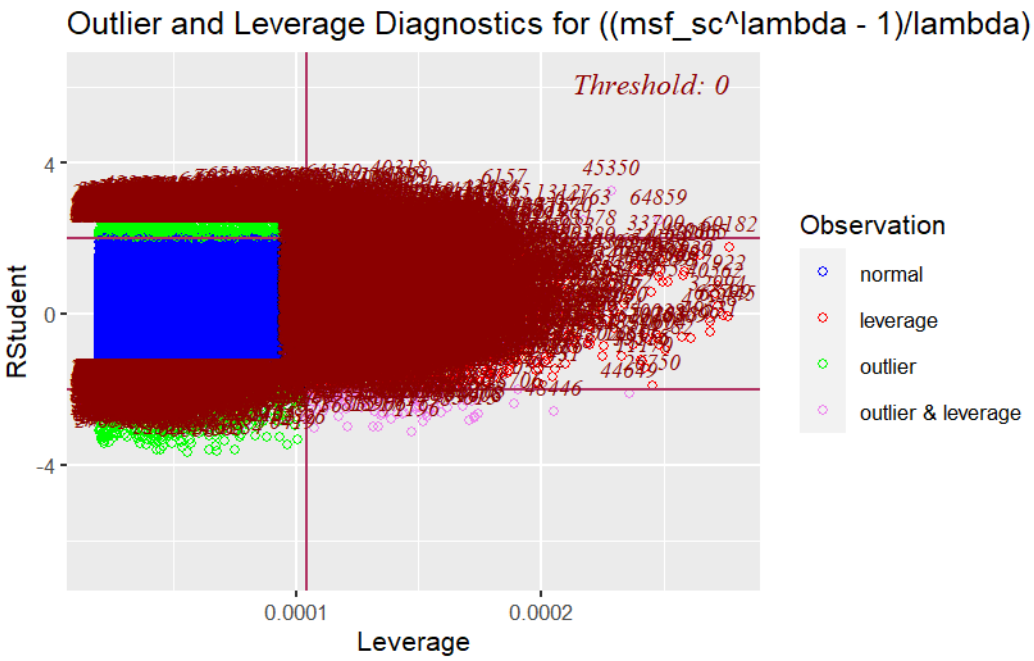
E.5.2.4 Measures of influence

Leverage points

> “Type of *influential observation* defined by one aspect of influence termed *leverage*. These observations are substantially different on one or more independent variables, so that they affect the estimation of one or more *regression coefficients*” (Hair, 2019, p. 262).

```r
full_model ▷ olsrr::ols_plot_resid_lev()
```

Figure 21 – Relation between the full model studentized residuals and their leverage/influence.



Source: Created by the author.

E.6 HYPOTHESIS TEST

$$\begin{cases} \mathrm{H}_0 : \mathrm{R}^2_{\text{res}} >= \mathrm{R}^2_{\text{full}} \\ \mathrm{H}_a : \mathrm{R}^2_{\text{res}} < \mathrm{R}^2_{\text{full}} \end{cases}$$

$$\mathrm{F} = \frac{\mathrm{R}^2_F - \mathrm{R}^2_R/(k_F - k_R)}{(1 - \mathrm{R}^2_F)/(\mathrm{N} - k_F - 1)}$$

$$\mathrm{F} = \frac{\text{Additional Var. Explained}/\text{Additional d.f. Expended}}{\text{Var. unexplained}/\text{d.f. Remaining}}$$

```r
source(here::here("R/utils-stats.R"))

dplyr::tibble(
  name = c("r_squared_res", "r_squared_full", "diff"),
  value = c(
  r_squared(res_model), r_squared(full_model),
  r_squared(full_model) - r_squared(res_model)
  )
)
```

Table 20 – Comparison between the coefficients of determination (R^2) of the restricted and full model.

name	value
r_squared_res	0.05373
r_squared_full	0.06070
diff	0.00696

Source: Created by the author.

```
print(stats::anova(res_model, full_model))
#> Analysis of Variance Table
#>
#> Model 1: ((msf_sc^lambda - 1)/lambda) ~ age + sex
#> Model 2: ((msf_sc^lambda - 1)/lambda) ~ age + sex + latitude
#>   Res.Df           RSS Df        Sum of Sq   F Pr(>F)
#> 1  76741 0.00000000136
#> 2  76740 0.00000000135  1 0.00000000000999 569 <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
source(here::here("R/utils-stats.R"))

n <- nrow(data)
k_res <- length(stats::coefficients(res_model)) - 1
k_full <- length(stats::coefficients(full_model)) - 1

((r_squared(full_model) - r_squared(res_model)) / (k_full - k_res)) /
```

```r
  ((1 - r_squared(full_model)) / (n  - k_full - 1))
#> [1] 568.94
```

$$f^2 = \frac{\text{R}_F^2 - \text{R}_R^2}{1 - \text{R}_F^2}$$

$$f^2 = \frac{\text{Additional Var. Explained}}{\text{Var. unexplained}}$$

```r
source(here::here("R/cohens_f_squared.R"))
source(here::here("R/utils-stats.R"))

cohens_f_squared_summary(
  adj_r_squared(res_model),
  adj_r_squared(full_model)
  )
```

Table 21 – Effect size between the restricted and full model based on Cohen's f^2.

name	value
f_squared	0.00740068896515648
effect_size	Negligible

Source: Created by the author. See Cohen (1988); Cohen (1992) to learn more.