

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

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

Dissertation 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 Modeling Complex Systems (PPG-SCX).

Area of Concentration: Fundamentals of complex systems.

Supervisor: Prof. Dr. Camilo Rodrigues Neto

São Paulo

2023

ERRATA

This is the development version of the thesis (version <1.0.0). Any necessary corrections will be listed here after its approval.

Qualifying exam text by Daniel Vartanian, titled **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 Modeling Complex Systems (PPG-SCX), 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 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 guiding my dissertation; for the patience and the virtue in taking on and mediating the process of transitioning my guidance in my master's degree after the breakdown of relations with my former supervisor.

Professor [Carlos Molina Mendes](#), for the speed, impartiality, patience, and virtue in mediating the process of transitioning my guidance in my master's degree.

My fellow friends: Alex Azevedo Martins; Amanda Moreira; Augusto Amado, Carina (Cacau) Prado; Cauê Teles; Ítalo Alves Bezerra do Nascimento; Júlia Mafra; Leonardo Kazuhiko Kawazoe; Letícia Nery de Figueiredo; and Reginaldo Noveli.

[President Lula](#) (Yes!), who saved Brazil from fascism and approved the long-overdue adjustments to postgraduate 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 postgraduate studies.

Nullius in verba
(The Royal Society, n.d.)

ABSTRACT

Vartanian, D. (2023). *Ecology of sleep and circadian phenotypes of the Brazilian population* [Master's thesis]. School of Arts, Sciences and Humanities, University of São Paulo, São Paulo. <https://danielvartan.github.io/mastersthesis/>

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. chronotype. circadian phenotypes. sleep. sleep-wake cycle. entrainment. latitude. ecology. mctq.

RESUMO

Vartanian, D. (2023). *Ecologia do sono e de fenótipos circadianos da população brasileira* [Dissertação de Mestrado]. Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo. <https://danielvartan.github.io/mastersthesis/>

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. cronotipo. fenótipos circadianos. sono. ciclo sono-vigília. entrainment. latitude. ecologia. mctq.

LIST OF FIGURES

Figure 1 – Illustration of a circadian rhythm in the movement of the leaves of the sensitive plant (<i>mimosa pudica</i>) observed by Jacques d’Ortous de Mairan in 1729. Reproduction from Nobel Prize Outreach AB (n.d.).	22
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. Adaptation based on Kuhlman, Craig, and Duffy (2018).	23
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 process. Adaptation based on A. A. Borbély (1982).	24
Figure 4 – Distribution 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 ($0.88 - 1$).	42
Figure 5 – Distribution of mean aggregates of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, with latitude decimal degree intervals. Higher values of MSF_{sc} indicate a late chronotype tendency.	45

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

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

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

SJL_{rel}

Relative social jetlag

SJL_{sc}

Jankowski’s sleep-corrected social jetlag

SJL_{sc-rel}

Jankowski’s relative sleep-corrected social jetlag

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 Marques & 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

LIST OF TERMS AND DEFINITIONS

For an extensive list of chronobiology related terms and definitions, please refer to Aschoff et al. (1965) and Marques & 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 (Horne & Ostberg, 1976; Roenneberg et al., 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 *dies*, 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 *infra*, below (think in terms of period repetition), and *dies*, 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 *ultra*, beyond (think in terms of period repetition), and *dies*, 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).

Table of contents

	Table of contents	17
1	INTRODUCTION	20
1.1	Foreword	20
1.2	A brief introduction to chronobiology	20
1.3	Thesis justification	25
1.4	Thesis aims	26
1.5	Projects developed	27
1.6	Related activities	29
1.6.1	Classes	29
1.6.2	Teaching internship	30
1.6.3	Publications	30
1.6.4	Conferences	31
1.6.5	Research compendia	31
1.6.6	Data plans	32
1.6.7	Softwares	32
1.6.8	Other projects	32
2	SIMILARITIES BETWEEN DIFFERENT VERSIONS OF THE MCTQ^{PT}	34
3	THE MCTQ R PACKAGE	35
4	ECOLOGY OF SLEEP AND CIRCADIAN PHENOTYPES OF THE BRAZILIAN POPULATION	36
5	RULE-BASED MODEL OF THE 24H LIGHT/DARK ENTRAINMENT PHENOMENON	37
6	A BIOLOGICAL APPROACH FOR THE LATITUDINAL CLINE OF THE CHRONOTYPE	38
6.1	Main text	39
6.1.1	Introduction	39
6.1.2	Results	41

6.1.3	Discussion	43
6.2	Methods	46
6.2.1	Ethics information	46
6.2.2	Measurement instrument	46
6.2.3	Sample	47
6.2.4	Analysis	49
6.2.5	Data availability	50
6.2.6	Code availability	50
6.3	Acknowledgments	50
6.4	Ethics declarations	50
6.4.1	Competing interests	50
6.5	Additional information	50
6.6	Rights and permissions	51
7	DISCUSSION AND CONCLUSIONS	52
	REFERENCES	53

Appendices 62

A	APPENDICE: CHAPTER 2 SUPPLEMENTAL MATERIAL	62
A.1	Load and embed texts	62
A.2	Text similarity	65
A.2.1	How similar is the <i>data questionnaire</i> when compared to the <i>EUCLOCK questionnaire</i> ?	67
A.2.2	How similar is the <i>data questionnaire</i> when compared to the <i>MCTQ^{PT} questionnaire</i> ?	69
B	APPENDICE: CHAPTER 3 SUPPLEMENTAL MATERIAL	72
C	APPENDICE: CHAPTER 4 SUPPLEMENTAL MATERIAL	73
C.1	Data wrangling	73
C.2	Distribution of main variables	74
C.3	Geographic distribution	76

C.4	Age pyramid	77
C.5	Correlation matrix	78
C.6	Age series	79
C.7	Chronotype	81
C.8	Latitude series	82
C.9	Statistics	83
C.9.1	Numerical variables	83
C.9.2	Sex	84
C.9.3	Sex and Age	87
C.9.4	Longitudinal range	91
C.9.5	Latitudinal range	93
C.9.6	Region	95
C.9.7	State	97
D	APPENDICE: CHAPTER 5 SUPPLEMENTAL MATERIAL	99
E	APPENDICE: CHAPTER 6 SUPPLEMENTAL MATERIAL	100
E.1	Hypothesis	100
E.2	Assumptions	101
E.3	Data preparation	102
E.4	Restricted model	103
E.4.1	Residual diagnostics	105
E.4.2	Heteroskedasticity	109
E.4.3	Collinearity diagnostics	110
E.4.4	Measures of influence	111
E.5	Full model	112
E.5.1	Residual diagnostics	115
E.5.2	Heteroskedasticity	118
E.5.3	Collinearity diagnostics	120
E.5.4	Measures of influence	121
E.6	Nested regression models test	121
E.7	Group test	123

1 INTRODUCTION

Note

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

1.1 Foreword

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

The analyses presented in this document are entirely reproducible. They were conducted using the [R programming language](#) in conjunction with the [Quarto](#) publishing system (an evolution of [R Markdown](#)). It's worth noting that this type of thesis, given its nature, is better suited for viewing in a **digital format**. To access the digital version of the thesis, as well as the latest research updates, please visit <https://danielvartan.github.io/mastersthesis/>.

As one might expect from a preliminary version, not all chapters are ready for reading. However, the author has chosen to display the entire state of the thesis rather than showing only one or two polished sections. This approach provides readers with a more comprehensive understanding of the work in progress. Chapters that are not suitable for reading will include a call block indicating their status.

1.2 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 (C. S. Pittendrigh 1981; Menna-Barreto 2003). 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 be-

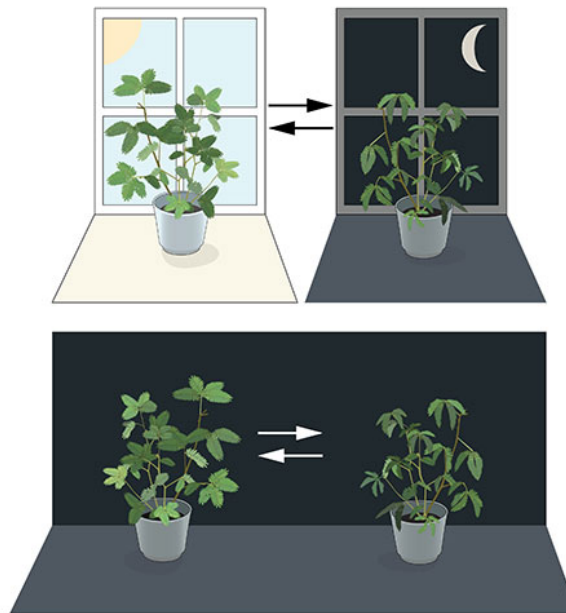
havior of humans and the nocturnal behavior of cats and some rodents (Foster and Kreitzman 2005).

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, Marques, and Menna-Barreto 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 and Kreitzman 2005; Rotenberg, Marques, and Menna-Barreto 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, and Erwin Bünning, culminating in the establishment of the science known as chronobiology, 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) (Rotenberg, Marques, and Menna-Barreto 2003; Cold Spring Harbor Laboratory n.d.) . 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](<https://www.nobelprize.org/prizes/medicine/2017/press-release/>)(<https://www.nobelprize.org/prizes/medicine/2017/press-release/>) for their discoveries of molecular mechanisms that regulate the circadian rhythm in fruit flies (*circā*, from Latin, meaning around, and *dīes*, meaning day (Latinium n.d.) — a rhythm that expresses itself in approximately one day) (Nobel Prize Outreach AB n.d.).

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 and 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 interective 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 ca-

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.

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



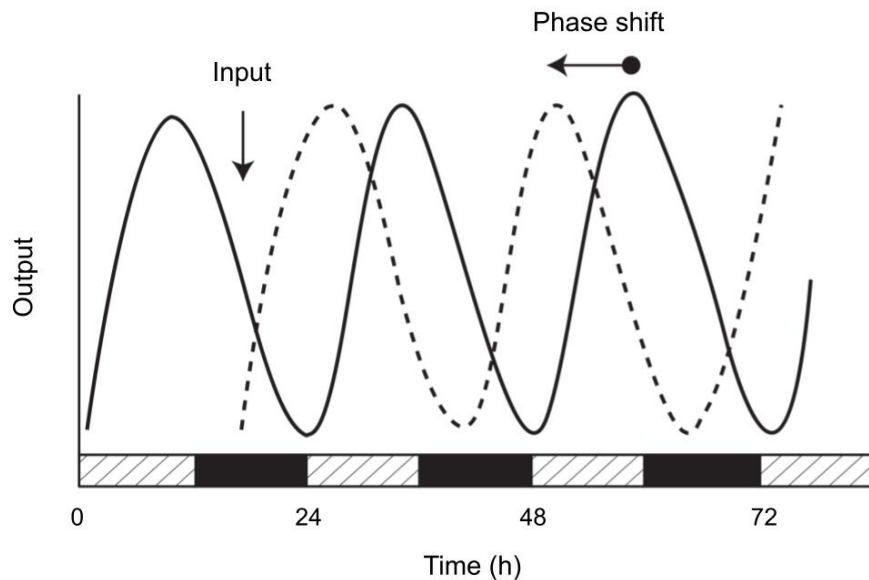
capacity, allowing them to organize resources and activities before they are needed (N. Marques, Golombek, and Moreno 2003).

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, Craig, and Duffy 2018) (see Figure 2). Among the known *zeitgebers* are, for example, meal timing and changes in environmental temperature (Aschoff 1981; Roenneberg and 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, Craig, and Duffy 2018; Roenneberg, Kumar, and Merrow 2007).

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

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.

Adaptation based on Kuhlman, Craig, and Duffy (2018).



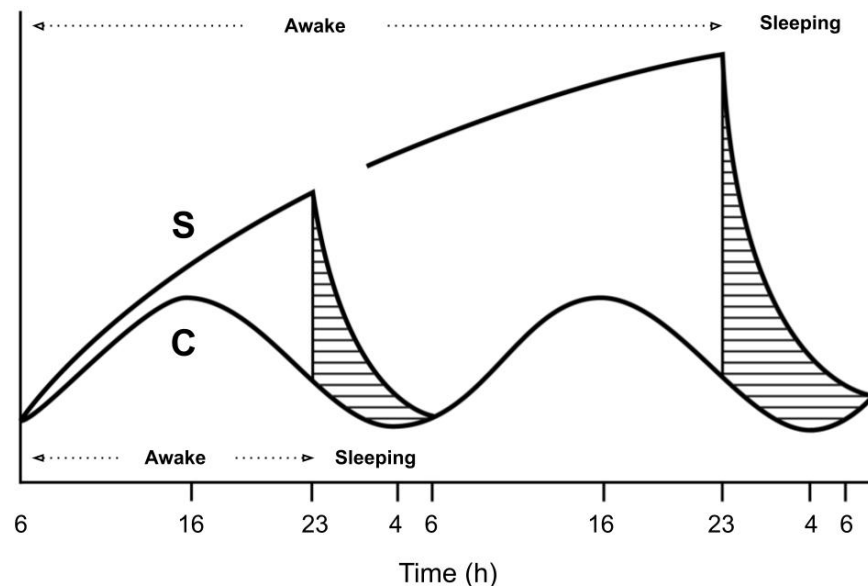
they are exposed to the environment or the differences in the expression of endogenous rhythmicity, which, in turn, results from gene expression (Roenneberg et al. 2007). The interaction between these two expressions, external and internal, of the environment and genotype, generates a signature, an observable characteristic, which is called a phenotype (Frommlet, Bogdan, and Ramsey 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 and Kreitzman 2005). Another term used for these temporal phenotypes, as the name suggest, is *chronotype* (Ehret 1974; C. S. Pittendrigh 1993). This term is also often used to differentiate phenotypes on a spectrum ranging from morningness to eveningness (Horne and Ostberg 1976; Roenneberg 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, and Mellow 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 (A. A. Borbély 1982; Alexander A. 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 process. Adaptation based on A. A. 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; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler 2008; Randler and Rahafar 2017; Roenneberg, Wirz-Justice, and 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, and Mellow 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, and Mellow 2003; Roenneberg et al. 2012).

1.3 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 and Ghali 2008; Morito et al. 2014; Mortaş, Bilici, and Karakan 2020), mental disorders (Jones,

Hare, and Evershed 2005; Kalmbach et al. 2015; Roh et al. 2012) and even cancer (Lie, Roessink, and Kjærheim 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, and a preliminary data analysis estimated that 89.25% of them are valid. 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.

1.4 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; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler 2008; Randler and Rahafar 2017; Roenneberg, Wirz-Justice, and Mellow 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.

1.5 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, Nogueira, and Lotufo 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 (2023d)) has already been developed and published on [CRAN](#) (The Comprehensive R Archive Network) and [GitHub](#). It have been downloaded more than 6000 to this date, and underwent 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.

1.6 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.

1.6.1 Classes

The following post-graduate classes at 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*).

In total, 44 credits were completed until the thesis publication date. Another 12 special credits related to an article publication (see Viana-Mendes et al. (2023)) were requested, according to [program regulations](#), to the Post-Graduate Program Coordination Commission (CCP) and should be evaluated by the end of October 2023. A minimum of 50 credits is needed for the thesis defense.

1.6.2 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 2023f) 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>

1.6.3 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>

1.6.4 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.

1.6.5 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/>

1.6.6 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>

1.6.7 Softwares

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

Vartanian, D. (2022). *{entrainment}: a rule-based model of the 24h light/dark cycle entrainment phenomenon* [Software, Python library]. <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). *{tesesusp}: a Quarto format for USP theses and dissertation* [Software, LaTeX/R format, v0.1.0]. <https://github.com/danielvartan/tesesusp/>

1.6.8 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 [Master's project]. School of Arts, Sciences and Humanities, University of Sao Paulo, São 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

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 use 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

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

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

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 CHRONO-TYPE

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.

Target

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

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; C. S. 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 and Sharma 2005). A way that organisms have 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, and 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, Waterhouse, and Wirz-Justice 1991; Roenneberg 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. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and 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, and Merrow 2003). We tested the latitude hypothesis on a sample with 73,825 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; C. S. 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 and Sharma 2005).

A way in which an organism could predict events and better manage its needs, like storing food for the winter.

But 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, and Merrow 2003; Roenneberg et al. 2010).

The main *zeitgeber* known today is the light, particularly the sun's light (Khalsa et al. 2003; Minors, Waterhouse, and Wirz-Justice 1991; Roenneberg 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. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and 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 Brazilian 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, and 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, and Merrow 2003). Furthermore, the test was carried out

on the biggest chronotype sample ever collected in a same country. A sample made of 73,825 subjects, all living in the same timezone in Brazil, with only one week of difference between questionnaire responses.

6.1.2 Results

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 04:26:51. 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 08:26:51.

```

1 library(dplyr)
2 library(here)
3 library(targets)
4 library(tidyr)
5
6 data <-
7   targets::tar_read("geocoded_data", store = here::here("_targets")) |>
8   dplyr::select(
9     age, sex, state, region, latitude, longitude, height, weight, work, study,
10    msf_sc, sjl, le_week,
11    ) |>
12   tidyr::drop_na(msf_sc, age, sex, latitude)

```

```

1 library(here)
2 library(latex2exp)
3
4 source(here::here("R/plot_chronotype.R"))
5

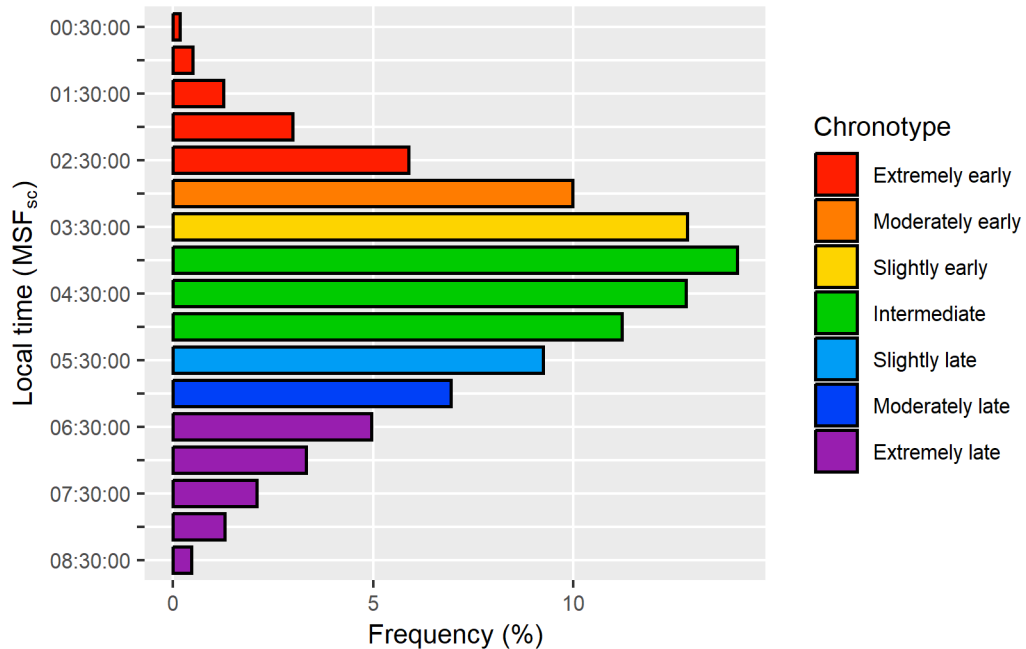
```

```

6 data |>
7   plot_chronotype(
8     col = "msf_sc",
9     x_lab = "Frequency (%)",
10    y_lab = latex2exp::TeX("Local time ($MSF_{sc}$)"),
11    col_width = 0.8,
12    col_border = 0.6,
13    text_size = 10,
14    chronotype_cuts = FALSE,
15    legend_position = "right"
16  )

```

Figure 4 – Distribution 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 ($0.88 - 1$).



The MSF_{sc} curve had a skewness of 0.291 and a kurtosis of 2.766. However, the distribution was not normal according to Kolmogorov-Smirnov test ($D = 0.0373$; $p\text{-value} < 2e - 16$) and D'Agostino Skewness test ($Z3 = 31.544$; $p\text{-value} < 2.2e - 16$).

A linear regression model was created with MSF_{sc} as the response variable and with age and sex as predictors ($R^2 = 0.054433$; $F(2, 73822) = 2130$, p-value $< 2e - 16$). A Box-Cox transformation of the response variable was needed to attend to the linear regression model assumptions ($\lambda = -1.1919$; $MSF_{sc}^{\lambda-1}/\lambda$). All coefficients were significantly different than 0 (p-value $< 2e - 16$) and, accordingly to D'Agostino Skewness test, the residuals were normal ($Z3 = -1.2704$; p-value < 0.2039). 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 = 1.0014; sex = 1.0014).

Another model was created on top of the first one, adding the latitude as a predictor variable ($R^2 = 0.06204$; $F(3, 73821) = 1630$, 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, ($Z3 = 0.0703$; p-value < 0.944). 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 = 1.0067; sex = 1.0018; 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 ($F(1, 73821) = 600$, 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) ($((0.06204 - 0.054433)/(1 - 0.06204) = 0.0081102)$).

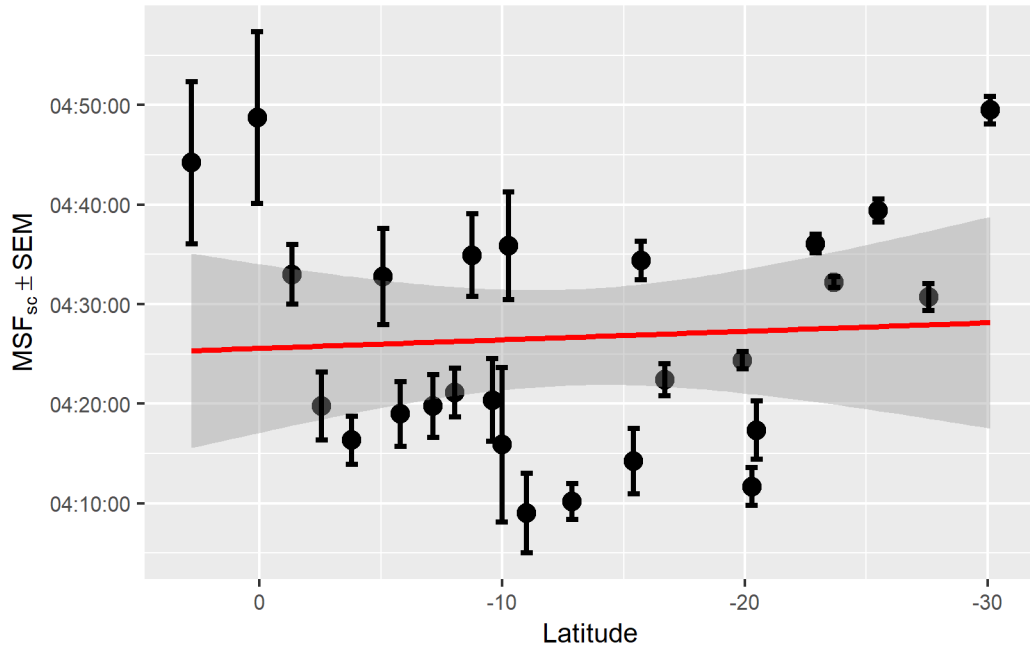
6.1.3 Discussion

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; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler 2008; Randler and Rahafar 2017; Roenneberg, Wirz-Justice, and 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.

```
1 library(dplyr)
2 library(here)
3 library(latex2exp)
4
5 source(here::here("R/plot_latitude_series.R"))
6
7 data |>
8   dplyr::filter(age <= 50) |>
9   plot_latitude_series(
10     col = "msf_sc",
11     y_lab = latex2exp::TeX("$MSF_{sc} \\pm SEM$"),
12     line_width = 2,
13     point_size = 3,
14     error_bar_width = 0.5,
15     error_bar_linewidth = 1,
16     error_bar = TRUE,
17     text_size = 10
18   )
```

Figure 5 – Distribution of mean aggregates of the midpoint between sleep onset and sleep end on work-free days (MSF_{sc}), MCTQ proxy for measuring the chronotype, with latitude decimal degree intervals. Higher values of MSF_{sc} indicate a late chronotype tendency.



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 (Skeldon and 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.

6.2.2 Measurement instrument

Chronotypes were measured using the core version of the standard Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, and Mellow 2003). MCTQ is a widely validated and widely used self-report questionnaire for measuring the sleep-wake cycle and chronotypes (Roenneberg, Winnebeck, and 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 Mellow for the EUCLOCK project (Roenneberg and Mellow 2006) (statements mean cosine distance = 0.921). They were also asked to provide sociodemographic (e.g., age, gender), 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 73,825 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 made in 2019 by the Brazilian Institute of Geography and Statistics (IBGE)²⁷ saw that 82.17% of Brazilians' homes had access to an internet connection. Hence, this sample can have a good diversity of Brazil's population. Only Brazilian residents in states with UTC-3 timezone and with an age equal to or greater than 18 years old 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 2010 census (Instituto Brasileiro de Geografia e Estatística, n.d.), Brazil had 51.793% of females and 48.207% of males with an age equal to or greater

than 18 years old. The sample is skewed for female subjects, with 66.341% of females and 33.659% of male subjects.

The subject's mean age is 32.017 years (SD = 9.242; Max. = 58.786). Female subjects have a mean age of 31.770 years (SD = 9.340; Max. = 58.786) and male subjects 32.504 years (SD = 9.026; Max. = 58.772). For comparison, based on the 2010 census²⁹, Brazil's population with an age equal to or greater than 18 years old had a mean age of 41.032 years (SD = 9.242), with a mean age of 41.645 years (SD = 16.907) for female subjects and a mean age of 40.373 years (SD = 16.200) for male subjects. Data related to age and sex from the last Brazil census (2022) were not available before this article was submitted.

Considering the 5 major regions of Brazil, the sample is mostly skewed for the Southeast, the most populated region. According to Brazil's 2022 census, 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 and Grolemund 2016). All processes were made with the help of the R programming language (R Core Team 2023), RStudio IDE (Posit Team 2023), and several R packages. The tidyverse and rOpenSci package ecosystem and other R packages adherents of the tidy tools manifesto (Wickham and Bryan 2023) were prioritized. The MCTQ data was analyzed using the `mctq` rOpenSci peer-reviewed package (Vartanian 2023d). 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.

$$\begin{cases} H_0 : R_{\text{res}}^2 \geq R_{\text{full}}^2 \\ H_a : R_{\text{res}}^2 < R_{\text{full}}^2 \end{cases}$$

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 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

This article is released under the [Creative Commons Attribution 4.0 International License](#), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as be given appropriate credit to the original author and the source, provide a link to the Creative Commons license, and indicate if changes were made.

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 ¹

- Allen, Michael Patrick. 1997. *Understanding Regression Analysis*. New York: Plenum Press.
- Aschoff, Jürgen, ed. 1981. *Biological Rhythms*. Vol. 4. Handbook of Behavioral Neurobiology. New York, NY: Plenum Press. <https://doi.org/10.1007/978-1-4615-6552-9>.
- . 1989. “Temporal Orientation: Circadian Clocks in Animals and Humans.” *Animal Behaviour* 37 (June): 881–96. [https://doi.org/10.1016/0003-3472\(89\)90132-2](https://doi.org/10.1016/0003-3472(89)90132-2).
- Aschoff, Jürgen, K. Klotter, and R. Wever. 1965. “Circadian Vocabulary: A Recommended Terminology with Definitions.” In *Circadian Clocks*. North-Holland.
- Bertalanffy, Ludwig von. 1968. *General System Theory: Foundations, Development, Applications*. New York, NY: George Braziller.
- Boccara, Nino. 2010. *Modeling Complex Systems*. 2nd ed. Graduate Texts in Physics. New York, NY: 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, Alexander A., Serge Daan, Anna Wirz-Justice, and Tom Deboer. 2016. “The Two-Process Model of Sleep Regulation: A Reappraisal.” *Journal of Sleep Research* 25 (2): 131–43. <https://doi.org/10.1111/jsr.12371>.
- Bussab, Wilton de Oliveira. 1988. *Análise de variância e de regressão: uma introdução*. 2nd ed. Métodos quantitativos. São Paulo: Atlas.
- Cambridge University Press. n.d. “Cambridge Dictionary.” Accessed September 21, 2023. <https://dictionary.cambridge.org/>.
- Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112 (1): 155–59. <https://doi.org/10.1037/0033-2909.112.1.155>.
- Cold Spring Harbor Laboratory. n.d. “1960: Biological Clocks, Vol. XXV.” Accessed July 17, 2023. <https://symposium.cshlp.org/site/misc/topic25.xhtml>.
- Dalpiaz, David. n.d. *Applied Statistics with R*. <https://book.stat420.org/>.
- DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics*. 4th ed. Boston: Addison-Wesley.

¹ According to the APA style - American Psychological Association.

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." <http://arxiv.org/abs/1810.04805>.
- Dongen, Hans P.A. van, Greg Maislin, Janet M. Mullington, and David F. Dinges. 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): 117–26. <https://doi.org/10.1093/sleep/26.2.117>.
- Drager, Luciano F., Daniela Vianna Pachito, Rogerio Morihisa, Pedro Carvalho, Abner Lobao, and Dalva Poyares. 2022. "Sleep Quality in the Brazilian General Population: A Cross-Sectional Study." *Sleep Epidemiology* 2 (December): 100020. <https://doi.org/10.1016/j.sleepe.2022.100020>.
- Dudek, Bruce. 2020. *Linear Models with R: Emphasis on 2-IV Models: Basics of Multiple Regression*. <https://bcdudek.net/regression1/>.
- Ehret, Charles F. 1974. "The Sense of Time: Evidence for Its Molecular Basis in the Eukaryotic Gene-Action System." In *Advances in Biological and Medical Physics*, 15:47–77. Elsevier. <https://doi.org/10.1016/B978-0-12-005215-8.50009-7>.
- Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. <https://doi.org/10.3758/BF03193146>.
- Fido, Abdullahi, and Adel Ghali. 2008. "Detrimental Effects of Variable Work Shifts on Quality of Sleep, General Health and Work Performance." *Medical Principles and Practice* 17 (6): 453–57. <https://doi.org/10.1159/000151566>.
- Foster, Russell G., and Leon Kreitzman. 2005. *Rhythms of Life: The Biological Clocks That Control the Daily Lives of Every Living Thing*. London: Profile Books.
- Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Thousand Oaks, CA: Sage.
- Frey, Bruce B., ed. 2022. *The SAGE Encyclopedia of Research Design*. 2nd ed. Thousand Oaks, CA: SAGE Publications. <https://doi.org/10.4135/9781071812082>.
- Frommlet, Florian, Małgorzata Bogdan, and David Ramsey. 2016. *Phenotype and Genotype: The Search for Influential Genes*. Vol. 18. Computational Biology. London: Springer London. <https://doi.org/10.1007/978-1-4471-5310-8>.

- Globo. 2017. "Metade Da População Se Sente Mal No Horário de Verão, Revela Pesquisa." *Fantástico*. October 15, 2017. <https://globoplay.globo.com/v/6219513/>.
- Hair, Joseph F. 2019. *Multivariate Data Analysis*. 8th ed. Andover, Hampshire: Cengage.
- Horne, J. A., and O. Ostberg. 1976. "A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms." *International Journal of Chronobiology* 4 (2): 97–110.
- Horzum, Mehmet Barış, Christoph Randler, Ercan Masal, Şenol Beşoluk, İsmail Önder, and Christian Vollmer. 2015. "Morningness–Eveningness and the Environment Hypothesis – a Cross-Cultural Comparison of Turkish and German Adolescents." *Chronobiology International* 32 (6): 814–21. <https://doi.org/10.3109/07420528.2015.1041598>.
- Hut, Roelof A., Silvia Paolucci, Roi Dor, Charalambos P. Kyriacou, and Serge Daan. 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. "Tabela 1378 - População Residente, Por Situação Do Domicílio, Sexo e Idade, Segundo a Condição No Domicílio e Compartilhamento Da Responsabilidade Pelo Domicílio." Tabela. Sistema IBGE de Recuperação Automática (SIDRA). <https://sidra.ibge.gov.br/tabela/1378>.
- Johnson, Richard, and Dean Wichern. 2013. *Applied Multivariate Statistical Analysis: Pearson New International Edition*. 6th ed. Harlow, UK: Pearson.
- Jones, Steven Huntley, Dougal Julian Hare, and Kate Evershed. 2005. "Actigraphic Assessment of Circadian Activity and Sleep Patterns in Bipolar Disorder." *Bipolar Disorders* 7 (2): 176–86. <https://doi.org/10.1111/j.1399-5618.2005.00187.x>.
- Kalmbach, David A., Vivek Pillai, Philip Cheng, J. Todd Arnedt, and Christopher L. Drake. 2015. "Shift Work Disorder, Depression, and Anxiety in the Transition to Rotating Shifts: The Role of Sleep Reactivity." *Sleep Medicine* 16 (12): 1532–38. <https://doi.org/10.1016/j.sleep.2015.09.007>.
- Khalsa, Sat Bir S., Megan E. Jewett, Christian Cajochen, and Charles A. Czeisler. 2003. "A Phase Response Curve to Single Bright Light Pulses in Human Subjects." *The Journal of Physiology* 549 (3): 945–52. <https://doi.org/10.1113/jphysiol.2003.040477>.

- Kuhlman, Sandra J., L. Michon Craig, and Jeanne F. Duffy. 2018. "Introduction to Chronobiology." *Cold Spring Harbor Perspectives in Biology* 10 (9): a033613. <https://doi.org/10.1101/cshperspect.a033613>.
- Kuhn, Max, and Julia Silge. 2022. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse*. Sebastopol, CA: O'Reilly Media. <https://www.tmwr.org/>.
- Latinitium. n.d. "Latin Dictionaries." Latinitium. Accessed September 21, 2023. <https://latinitium.com/latin-dictionaries/>.
- Leocadio-Miguel, Mario André, Fernando Mazzili Louzada, Leandro Lourenção Duarte, Roberta Peixoto Areas, Marilene Alam, Marcelo Ventura Freire, John Fontenele-Araujo, Luiz Menna-Barreto, and Mario Pedrazzoli. 2017. "Latitudinal Cline of Chronotype." *Scientific Reports* 7 (1): 5437. <https://doi.org/10.1038/s41598-017-05797-w>.
- Leocadio-Miguel, Mario André, Valéria Clarisse De Oliveira, Danyella Pereira, and Mario Pedrazzoli. 2014. "Detecting Chronotype Differences Associated to Latitude: A Comparison Between Horne--Östberg and Munich Chronotype Questionnaires." *Annals of Human Biology* 41 (2): 107–10. <https://doi.org/10.3109/03014460.2013.832795>.
- Levins, Richard. 1998. "Dialectics and Systems Theory." *Science & Society* 62 (3): 375–99. <https://www.jstor.org/stable/40403729>.
- Lie, Jenny-Anne S., Jolanta Roessink, and Kristina Kjærheim. 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, Mirian David, and Gisele Oda. 2012. "Glossário." *Revista da Biologia* 9 (3). <https://www.revistas.usp.br/revbiologia/article/view/114816>.
- Marques, Nelson, Diego A. Golombek, and Claudia Moreno. 2003. "Adaptação temporal." In *Cronobiologia: princípios e aplicações*, 3. ed. rev. e ampl., 55–98. Acadêmica 12. São Paulo: Editora da Universidade de São Paulo.
- Matias, Vinicius Alves, Cassio Serrano, Daniel Vartanian, Mario Pedrazzoli, and Ana Amélia Benedito-Silva. 2022. "Ecology of Sleep and Circadian Phenotypes of the Brazilian Population [Poster]." São Paulo, September 3. <https://doi.org/10.13140/RG.2.2.25343.07840>.

- Menna-Barreto, Luiz. 2003. "O tempo na biologia." In *Cronobiologia: princípios e aplicações*, 3. ed. rev. e ampl., 25–30. Acadêmica 12. São Paulo: Editora da Universidade de São Paulo.
- Minors, David S., James M. Waterhouse, and Anna Wirz-Justice. 1991. "A Human Phase-Response Curve to Light." *Neuroscience Letters* 133 (1): 36–40. [https://doi.org/10.1016/0304-3940\(91\)90051-T](https://doi.org/10.1016/0304-3940(91)90051-T).
- Mitchell, Melanie. 2013. "Introduction to Complexity." Online course. 2013. <https://www.complexityexplorer.org/courses/1>–<https://www.complexityexplorer.org/courses/1>.
- Morito, Yoshiya, Masahito Aimi, Norihisa Ishimura, Shino Shimura, Hironobu Mikami, Eiko Okimoto, Shuuichi Sato, et al. 2014. "Association Between Sleep Disturbances and Abdominal Symptoms." *Internal Medicine* 53 (19): 2179–83. <https://doi.org/10.2169/internalmedicine.53.2591>.
- Mortaş, Hande, Saniye Bilici, and Tarkan Karakan. 2020. "The Circadian Disruption of Night Work Alters Gut Microbiota Consistent with Elevated Risk for Future Metabolic and Gastrointestinal Pathology." *Chronobiology International* 37 (7): 1067–81. <https://doi.org/10.1080/07420528.2020.1778717>.
- Nobel Prize Outreach AB. n.d. "Press Release." The Nobel Prize. Accessed September 28, 2023. <https://www.nobelprize.org/prizes/medicine/2017/press-release/>.
- Papantoniou, Kyriaki, Gemma Castaño-Vinyals, Ana Espinosa, Nuria Aragonés, Beatriz Pérez-Gómez, Javier Burgos, Inés Gómez-Acebo, et al. 2015. "Night Shift Work, Chronotype and Prostate Cancer Risk in the MCC-S Pain Case-control Study." *International Journal of Cancer* 137 (5): 1147–57. <https://doi.org/10.1002/ijc.29400>.
- Paranjpe, Dhanashree A, and Vijay Kumar Sharma. 2005. "Evolution of Temporal Order in Living Organisms." *Journal of Circadian Rhythms* 3 (May). <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: 159–84. <https://doi.org/10.1101/SQB.1960.025.01.015>.
- . 1981. "Circadian Systems: General Perspective." In *Biological Rhythms*, 4:57–80. Handbook of Behavioral Neurobiology. New York, NY: Plenum Press. <https://doi.org/10.1007/978-1-4615-6552-9>.

- . 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, Colin S., Walter T. Kyner, and Tsuguhiko Takamura. 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, Karl R. 1979. *Objective Knowledge: An Evolutionary Approach*. Rev. ed. Oxford University Press.
- Posit Team. 2023. “RStudio: Integrated Development Environment for R.” C. Posit Software. <http://www.posit.co>.
- R Core Team. 2023. “R: A Language and Environment for Statistical Computing.” R. R Foundation for Statistical Computing. <https://www.R-project.org>.
- Randler, Christoph. 2008. “Morningness–eveningness Comparison in Adolescents from Different Countries Around the World.” *Chronobiology International* 25 (6): 1017–28. <https://doi.org/10.1080/07420520802551519>.
- Randler, Christoph, and Arash Rahafar. 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, Cátia. 2020. “Sleep patterns in portugal.” Tese de doutorado, Lisboa: Universidade de Lisboa. Repositório da Universidade de Lisboa. <http://hdl.handle.net/10451/54147>.
- Roenneberg, Till. 2012. “What Is Chronotype?” *Sleep and Biological Rhythms* 10 (2): 75–76. <https://doi.org/10.1111/j.1479-8425.2012.00541.x>.
- Roenneberg, Till, Karla V. Allebrandt, Martha Merrow, and Céline Vetter. 2012. “Social Jetlag and Obesity.” *Current Biology* 22 (10): 939–43. <https://doi.org/10.1016/j.cub.2012.03.038>.
- Roenneberg, Till, Serge Daan, and Martha Merrow. 2003. “The Art of Entrainment.” *Journal of Biological Rhythms* 18 (3): 183–94. <https://doi.org/10.1177/0748730403018003001>.
- Roenneberg, Till, Roelof Hut, Serge Daan, and Martha Merrow. 2010. “Entrainment Concepts Revisited.” *Journal of Biological Rhythms* 25 (5): 329–39. <https://doi.org/10.1177/0748730410379082>.

- Roenneberg, Till, Tim Kuehnle, Myriam Juda, Thomas Kantermann, Karla Allebrandt, Marijke Gordijn, and Martha Merrow. 2007. "Epidemiology of the Human Circadian Clock." *Sleep Medicine Reviews* 11 (6): 429–38. <https://doi.org/10.1016/j.smr.2007.07.005>.
- Roenneberg, Till, C. Jairaj Kumar, and Martha Merrow. 2007. "The Human Circadian Clock Entrain to Sun Time." *Current Biology* 17 (2): R44–45. <https://doi.org/10.1016/j.cub.2006.12.011>.
- Roenneberg, Till, and Martha Merrow. 2006. "EUCLOCK: Portuguese MCTQ." 2006. http://web.archive.org/web/20141115175303/https://www.bioinfo.mpg.de/mctq/core_work_life/core/core.jsp?language=por_b.
- . 2016. "The Circadian Clock and Human Health." *Current Biology* 26 (10): R432–43. <https://doi.org/10.1016/j.cub.2016.04.011>.
- Roenneberg, Till, Luísa K. Pilz, Giulia Zerbini, and Eva C. Winnebeck. 2019. "Chronotype and Social Jetlag: A (Self-) Critical Review." *Biology* 8 (3): 54. <https://doi.org/10.3390/biology8030054>.
- Roenneberg, Till, Eva C. Winnebeck, and Elizabeth B. Klerman. 2019. "Daylight Saving Time and Artificial Time Zones – a Battle Between Biological and Social Times." *Frontiers in Physiology* 10 (August): 944. <https://doi.org/10.3389/fphys.2019.00944>.
- Roenneberg, Till, Anna Wirz-Justice, and Martha Merrow. 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, Jee Hoon, Yafei Huang, Adam W. Bero, Tom Kasten, Floy R. Stewart, Randall J. Bateman, and David M. Holtzman. 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., Nelson Marques, and Luiz Menna-Barreto. 2003. "História e perspectivas da cronobiologia." In *Cronobiologia: princípios e aplicações*, 3. ed. rev. e ampl., 31–53. Acadêmica 12. São Paulo: Editora da Universidade de São Paulo.
- Schernhammer, E. S., F. Laden, F. E. Speizer, W. C. Willett, D. J. Hunter, I. Kawachi, and G. A. Colditz. 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): 1563–68. <https://doi.org/10.1093/jnci/93.20.1563>.

- Skeldon, Anne C., and Derk-Jan Dijk. 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, Fábio, Rodrigo Nogueira, and Roberto Lotufo. 2020. "Portuguese Named Entity Recognition Using BERT-CRF." arXiv. <https://doi.org/10.48550/arXiv.1909.10649>.
- The Royal Society. n.d. "History of the Royal Society." Accessed September 9, 2023. <https://royalsociety.org/about-us/history/>.
- Tryon, W, L Jason, E Frankenberry, and S Torresharding. 2004. "Chronic Fatigue Syndrome Impairs Circadian Rhythm of Activity Level." *Physiology & Behavior* 82 (5): 849–53. [https://doi.org/10.1016/S0031-9384\(04\)00303-8](https://doi.org/10.1016/S0031-9384(04)00303-8).
- Vartanian, Daniel. 2022a. "{Actverse}: Tools for Actigraphy Data Analysis." R. <https://docs.ropensci.org/mctq/>.
- . 2022b. "{Entrainment}: A Rule-Based Model of the 24h Light/Dark Cycle Entrainment Phenomenon." Python. <https://github.com/danielvartan/entrainment>.
- . 2022c. "Ecology of Sleep and Circadian Phenotypes of the Brazilian Population [Poster]." São Paulo, November 19. <https://doi.org/10.13140/RG.2.2.25343.07840>.
- . 2023a. "Ecology of Sleep and Circadian Phenotypes of the Brazilian Population [Research Compendium]." Research compedium. R/TeX/Quarto. <https://github.com/danielvartan/tesesusp>.
- . 2023b. "{Lockr}: Easily Encrypt/Decrypt Files." R. <https://github.com/danielvartan/lockr>.
- . 2023c. "{Lubritime}: An Extension for the Lubridate Package." R. <https://github.com/danielvartan/lubritime>.
- . 2023d. "{Mctq}: Tools to Process the Munich ChronoType Questionnaire (MCTQ)." R. <https://docs.ropensci.org/mctq/>.
- . 2023e. "{Tesesusp}: A Quarto Format for USP Theses and Dissertation." TeX. <https://github.com/danielvartan/tesesusp>.
- . 2023f. "Plano de ensino: ACH0042 - Resolução de Problemas II." São Paulo. <https://doi.org/10.13140/RG.2.2.33335.50086>.
- . 2023g. "Ecologia do sono e de fenótipos circadianos da população brasileira [Data Management Plan]." DMPHub. <https://doi.org/10.48321/D1DW8P>.

- Vartanian, Daniel, and Mario Pedrazzoli. 2017. "Questionário de Cronotipo: Baseado No Munich ChronoType Questionnaire (MCTQ)." 2017. <https://web.archive.org/web/20171018043514/each.usp.br/gipso/mctq>.
- Viana-Mendes, Juliana, Ana Amélia Benedito-Silva, Maria Augusta Medeiros Andrade, Daniel Vartanian, Bruno da Silva Brandão Gonçalves, José Cipolla-Neto, and Mario Pedrazzoli. 2023. "Actigraphic Characterization of Sleep and Circadian Phenotypes of PER3 Gene VNTR Genotypes." *Chronobiology International*, September. <https://doi.org/10.1080/07420528.2023.2256858>.
- Wang, Jiapeng, and Yihong Dong. 2020. "Measurement of Text Similarity: A Survey." *Information* 11 (9): 421. <https://doi.org/10.3390/info11090421>.
- Watson, Nathaniel F., M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, et al. 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): 591–92. <https://doi.org/10.5664/jcsm.4758>.
- Wickham, Hadley, and Jennifer Bryan. 2023. *R Packages*. 2nd ed. O'Reilly. <https://r-pkgs.org/>.
- Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science*. O'Reilly. <https://r4ds.had.co.nz>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.

A APPENDICE: 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.

The analyses shown here are 100% reproducible. You can find all the code used in this thesis along with the computational notebooks at <https://github.com/danielvartan/mastersthesis/>.

A.1 Load and embed 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.

```

1 data_text <- c(
2   "Você vai para a cama às ___ horas.",
3   "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
4   "Depois de ir para a cama, você decide dormir às ___ horas.",
5   "Você precisa de ___ para dormir.",
6   "Você acorda às ___ horas.",
7   "Você se levanta ___ depois de despertar.",
8   "Você vai para a cama às ___ horas.",
9   "",
10  "Depois de ir para a cama, você decide dormir às ___ horas.",
11  "Você precisa de ___ para dormir.",
12  "Você acorda às ___ horas.",
13  "Você se levanta ___ depois de despertar."
14 )
15
16 euclock_text <- c(

```

```

17   "vou para a cama às ___ horas.",
18   "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
19   "às ___ horas, decido dormir.",
20   "Eu necessito ___ minutos para adormecer.",
21   "acordo às ___ horas,",
22   "passados ___ minutos, me levanto.",
23   "vou para a cama às ___ horas.",
24   "Algumas pessoas permanecem um tempo acordadas depois que vão se deitar.",
25   "às ___ horas, decido dormir.",
26   "Eu necessito ___ minutos para adormecer.",
27   "acordo às ___ horas,",
28   "passados ___ minutos, me acordo."
29 )
30
31 mctq_pt_text <- c(
32   "Vou para a cama às ___ horas.",
33   "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
34   "Às ___ horas estou pronto para adormecer.",
35   "Necessito de ___ minutos para adormecer.",
36   "Acordo às ___ horas.",
37   "Após ___ minutos, levanto-me.",
38   "Vou para a cama às ___ horas.",
39   "Algumas pessoas permanecem algum tempo acordadas depois de estarem na cama.",
40   "Às ___ horas estou pronto para adormecer.",
41   "Necessito de ___ minutos para adormecer.",
42   "Acordo às ___ horas.",
43   "Após ___ minutos, levanto-me."
44 )

1  # library(textreuse)
2
3  data_text_textreuse <-

```



```

4   textreuse::TextReuseTextDocument(
5     text = data_text,
6     meta = list(id = "data")
7   )
8
9   euclock_text_textreuse <-
10    textreuse::TextReuseTextDocument(
11      text = euclock_text,
12      meta = list(id = "euclock")
13    )
14
15   mctq_pt_text_textreuse <-
16    textreuse::TextReuseTextDocument(
17      text = mctq_pt_text,
18      meta = list(id = "mctq_pt")
19    )

```

```

1   # See
2   # <https://huggingface.co/neuralmind/bert-base-portuguese-cased>
3   # to learn more.
4
5   # library(checkmate)
6   # library(text)
7   # library(rutils)
8
9   rutils::assert_internet()
10
11   text_embed <- function(text) {
12     checkmate::assert_character(text)
13
14     text |>
15     text::textEmbed(

```

```

16     model = "neuralmind/bert-base-portuguese-cased",
17     layers = - 2,
18     dim_name = TRUE,
19     aggregation_from_layers_to_tokens = "concatenate",
20     aggregation_from_tokens_to_texts = "mean",
21     aggregation_from_tokens_to_word_types = NULL,
22     keep_token_embeddings = TRUE,
23     tokens_select = NULL,
24     tokens_deselect = NULL,
25     decontextualize = FALSE,
26     model_max_length = NULL,
27     max_token_to_sentence = 4,
28     tokenizer_parallelism = FALSE,
29     device = "gpu",
30     logging_level = "error"
31 )
32 }
33
34 data_text_textembed <- text_embed(data_text)
35 euclock_text_textembed <- text_embed(euclock_text)
36 mctq_pt_text_textembed <- text_embed(mctq_pt_text)

```

A.2 Text similarity

See Wang and Dong (2020) to learn more.

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

```

1 # See `?text::textSimilarity` to learn more.
2
3 # library(checkmate)
4 # library(cli)
5 # library(text)

```

```

6
7 text_distance <- function(x, y) {
8   checkmate::assert_list(x, len = 2)
9   checkmate::assert_list(y, len = 2)
10
11   methods <- c(
12     "binary", "cosine", "canberra", "euclidean", "manhattan", "maximum",
13     "minkowski", "pearson"
14   )
15
16   for (i in methods) {
17     cli::cli_alert_info(paste0(
18       "Method: {.strong {stringr::str_to_title(i)}}")
19     ))
20
21     test <-
22       text::textSimilarity(
23         x$texts$texts,
24         y$texts$texts,
25         method = i,
26         center = TRUE,
27         scale = FALSE
28       )
29
30     cli::cli_bullets(c(">" = "Line by line"))
31     print(test)
32
33     cli::cli_bullets(c(">" = "Overall mean"))
34     print(mean(test))
35
36     cli::cat_line()

```

```

37   }
38 }

1  # See `?textreuse::jaccard_similarity` to learn more.
2
3  # library(checkmate)
4  # library(cli)
5  # library(textreuse)
6
7  text_representation <- function(x, y) {
8    checkmate::assert_class(x, "TextReuseTextDocument")
9    checkmate::assert_class(y, "TextReuseTextDocument")
10
11    cli::cli_alert_info(paste0("Method: {.strong Jaccard similarity}"))
12    print(textreuse::jaccard_similarity(x, y))
13    cli::cat_line()
14
15    cli::cli_alert_info(paste0("Method: {.strong Jaccard bag similarity}"))
16    print(textreuse::jaccard_bag_similarity(x, y))
17    cli::cat_line()
18 }

```

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

Text distance | Embedded (Semantic) test

```

1  # See `?text::textSimilarity` to learn more.
2
3  text_distance(data_text_textembed, euclock_text_textembed)
4
5  #> i Method: Binary
6  #> > Line by line
7
8  #> [1] 1 1 1 1 1 1 1 1 1 1 1 1

```

```

7  #> > Overall mean
8  #> [1] 1
9  #> i Method: Cosine
10 #> > Line by line
11 #> [1] 0.991 1.000 0.964 0.966 0.960 0.956 0.991 0.156 0.964 0.966 0.960 0.950
12 #> > Overall mean
13 #> [1] 0.902
14 #> i Method: Canberra
15 #> > Line by line
16 #> [1] -218 1 -335 -319 -336 -374 -218 -643 -335 -319 -336 -381
17 #> > Overall mean
18 #> [1] -318
19 #> i Method: Euclidean
20 #> > Line by line
21 #> [1] -1.50 1.00 -4.06 -3.98 -4.14 -4.71 -1.50 -19.45 -4.06 -3.98
22 #> [11] -4.14 -5.07
23 #> > Overall mean
24 #> [1] -4.63
25 #> i Method: Manhattan
26 #> > Line by line
27 #> [1] -53.6 1.0 -109.0 -105.5 -111.5 -124.2 -53.6 -198.9 -109.0 -105.5
28 #> [11] -111.5 -131.2
29 #> > Overall mean
30 #> [1] -101
31 #> i Method: Maximum
32 #> > Line by line
33 #> [1] 0.5294 1.0000 0.3766 0.0741 0.1198 0.3106 0.5294 -14.9135
34 #> [9] 0.3766 0.0741 0.1198 0.1683
35 #> > Overall mean
36 #> [1] -0.936
37 #> i Method: Minkowski
38 #> > Line by line

```

```

39 #> [1] -1.50  1.00 -4.06 -3.98 -4.14 -4.71 -1.50 -19.45 -4.06 -3.98
40 #> [11] -4.14 -5.07
41 #> > Overall mean
42 #> [1] -4.63
43 #> i Method: Pearson
44 #> > Line by line
45 #> [1] 0.991 1.000 0.964 0.966 0.960 0.956 0.991 0.156 0.964 0.966 0.960 0.950
46 #> > Overall mean
47 #> [1] 0.902

```

Text representation

```

1 # The maximum value for the Jaccard bag similarity is 0.5.
2
3 text_representation(euclock_text_textreuse, data_text_textreuse)
4 #> i Method: Jaccard similarity
5 #> [1] 0.217
6 #> i Method: Jaccard bag similarity
7 #> [1] 0.145

```

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

Text distance | Embedded (Semantic) test

```

1 # See `?text::textSimilarity` to learn more.
2
3 text_distance(data_text_textembed, mctq_pt_text_textembed)
4 #> i Method: Binary
5 #> > Line by line
6 #> [1] 1 1 1 1 1 1 1 1 1 1 1 1
7 #> > Overall mean

```

```

8  #> [1] 1
9  #> i Method: Cosine
10 #> > Line by line
11 #> [1] 0.990 0.990 0.969 0.958 0.988 0.960 0.990 0.160 0.969 0.958 0.988 0.960
12 #> > Overall mean
13 #> [1] 0.907
14 #> i Method: Canberra
15 #> > Line by line
16 #> [1] -228 -248 -336 -350 -225 -354 -228 -632 -336 -350 -225 -354
17 #> > Overall mean
18 #> [1] -322
19 #> i Method: Euclidean
20 #> > Line by line
21 #> [1] -1.66 -1.73 -3.81 -4.60 -1.81 -4.46 -1.66 -19.38 -3.81 -4.60
22 #> [11] -1.81 -4.46
23 #> > Overall mean
24 #> [1] -4.48
25 #> i Method: Manhattan
26 #> > Line by line
27 #> [1] -57.3 -59.0 -102.3 -119.6 -59.8 -117.6 -57.3 -193.8 -102.3 -119.6
28 #> [11] -59.8 -117.6
29 #> > Overall mean
30 #> [1] -97.1
31 #> i Method: Maximum
32 #> > Line by line
33 #> [1] 0.6055 0.6040 0.4180 0.0186 0.3922 0.2893 0.6055 -14.9565
34 #> [9] 0.4180 0.0186 0.3922 0.2893
35 #> > Overall mean
36 #> [1] -0.909
37 #> i Method: Minkowski
38 #> > Line by line
39 #> [1] -1.66 -1.73 -3.81 -4.60 -1.81 -4.46 -1.66 -19.38 -3.81 -4.60

```

```

40 #> [11] -1.81 -4.46
41 #> > Overall mean
42 #> [1] -4.48
43 #> i Method: Pearson
44 #> > Line by line
45 #> [1] 0.990 0.990 0.969 0.958 0.988 0.960 0.990 0.160 0.969 0.958 0.988 0.960
46 #> > Overall mean
47 #> [1] 0.907

```

Text representation

```

1 # The maximum value for the Jaccard bag similarity is 0.5.
2
3 text_representation(mctq_pt_text_textreuse, data_text_textreuse)
4 #> i Method: Jaccard similarity
5 #> [1] 0.105
6 #> i Method: Jaccard bag similarity
7 #> [1] 0.0982

```


B APPENDICE: 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.

C APPENDICE: 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.

The analyses here are 100% reproducible. You can find all the code used in this thesis along with the computational notebooks at <https://github.com/danielvartan/masterthesis/>.

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](#).

```
1 # library(dplyr)
2 # library(here)
3 library(targets)
4 # library(tidyr)
5
6 data <-
7   targets::tar_read("geocoded_data", store = here::here("_targets")) |>
8   dplyr::select(
9     age, sex, state, region, latitude, longitude, height, weight, work, study,
10     msf_sc, sjl, le_week,
11   ) |>
12   tidyr::drop_na(msf_sc, age, sex, latitude)
```

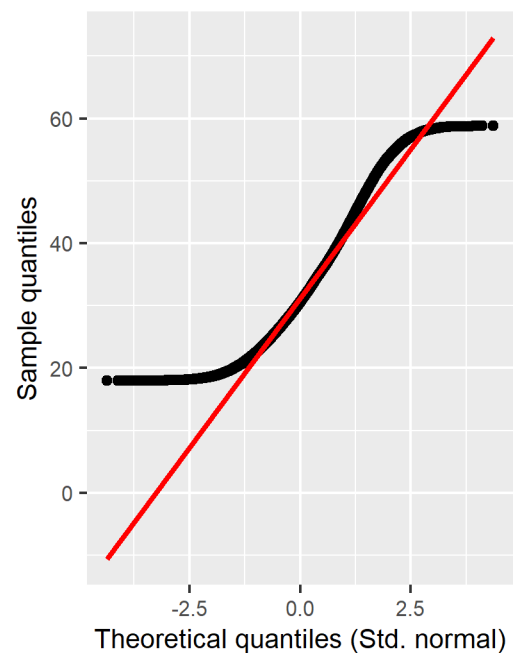
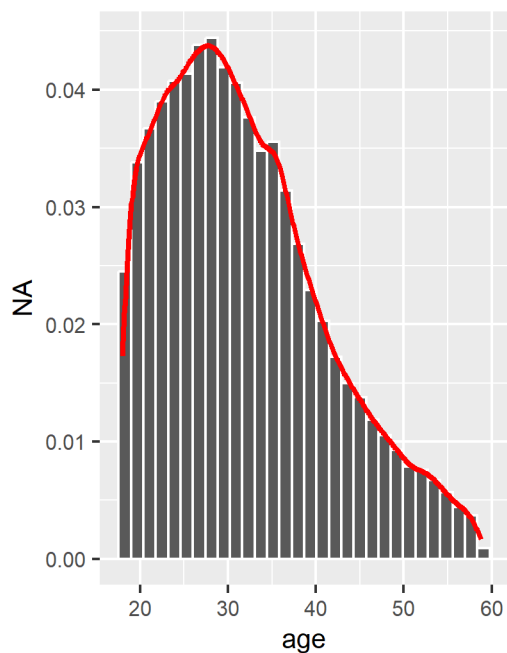
C.2 Distribution of main variables

```
1 # label: fig-ap-ch-4-age-var-distribution
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7 # library(hms)
8 # library(magrittr)
9
10 source(here::here("R/test_normality.R"))
11 source(here::here("R/utils.R"))
12
13 col <- "age"
14
15 stats <- data |>
16   magrittr::extract2(col) |>
17   test_normality(
18     name = col,
19     threshold = hms::parse_hms("12:00:00"),
20     remove_outliers = FALSE,
21     iqr_mult = 1.5,
22     log_transform = FALSE,
23     density_line = TRUE,
24     text_size = text_size,
25     print = TRUE
26   )
27 #> Registered S3 method overwritten by 'quantmod':
28 #>   method                from
29 #>   as.zoo.data.frame zoo
30 #> # A tibble: 14 x 2
```

```

31 #>   name      value
32 #>   <chr>    <chr>
33 #> 1 n        79198
34 #> 2 n_rm_na  79198
35 #> 3 n_na      0
36 #> 4 mean      31.9838074965417
37 #> 5 var       85.2414919292643
38 #> 6 sd        9.23263190695179
39 #> # i 8 more rows

```



```

1
2 stats$stats |> list_as_tibble()

```

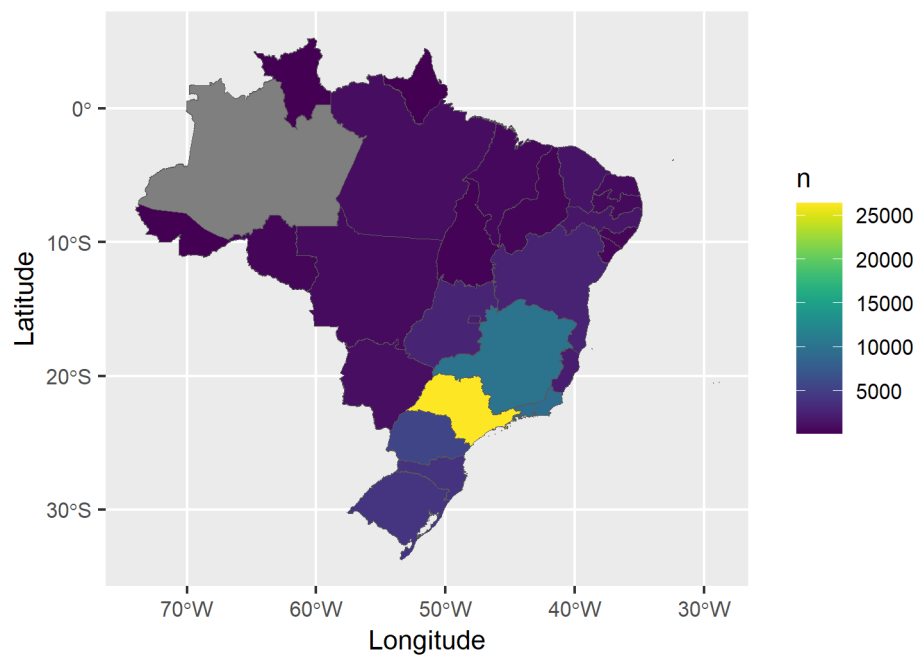
name	value
n	79198
n_rm_na	79198
n_na	0
mean	31.9838074965417
var	85.2414919292643
sd	9.23263190695179
min	18
q_1	24.7222222222222
median	30.5388888888889
q_3	37.61875
max	58.7861111111111
iqr	12.8965277777778
skewness	0.665751526654394
kurtosis	2.82381488030798

C.3 Geographic distribution

```

1 # label: fig-ap-ch-4-brazil-uf-map
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7 # library(rutils)
8
9 source(here::here("R/plot_brazil_uf_map.R"))
10
11 rutils:::assert_internet()
12
13 brazil_uf_map <-
14   data |>
15   plot_brazil_uf_map(option = "viridis", text_size = 10)

```

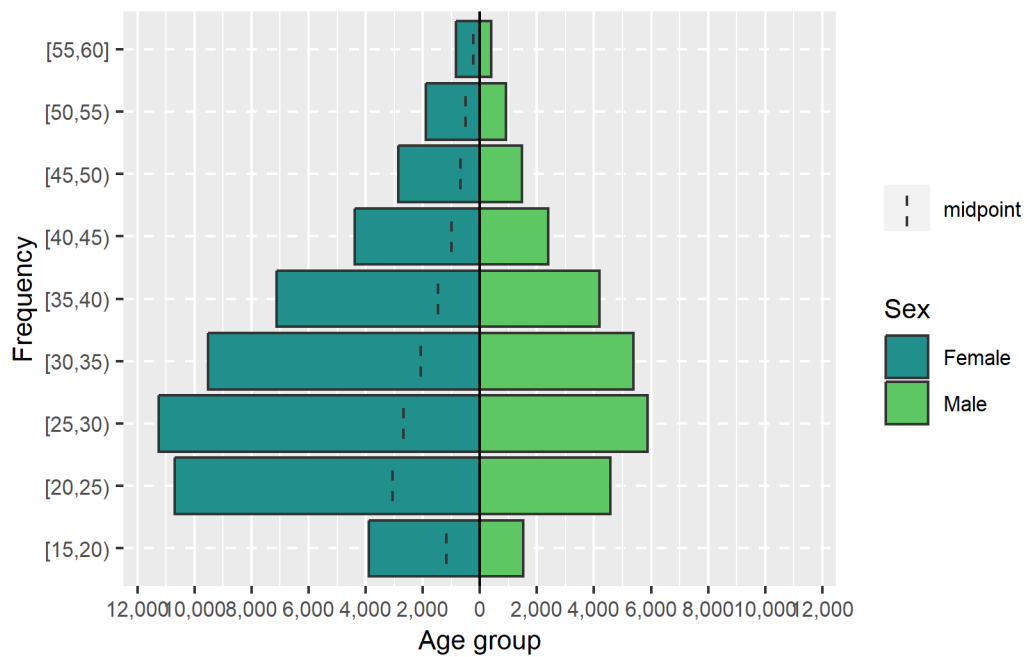


C.4 Age pyramid

```

1 # label: fig-ap-ch-4-age-pyramid
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7
8 source(here::here("R/plot_age_pyramid.R"))
9
10 age_pyramid <-
11   data |>
12   plot_age_pyramid(
13     interval = 10,
14     na_rm = TRUE,
15     text_size = text_size
16   )

```



C.5 Correlation matrix

```

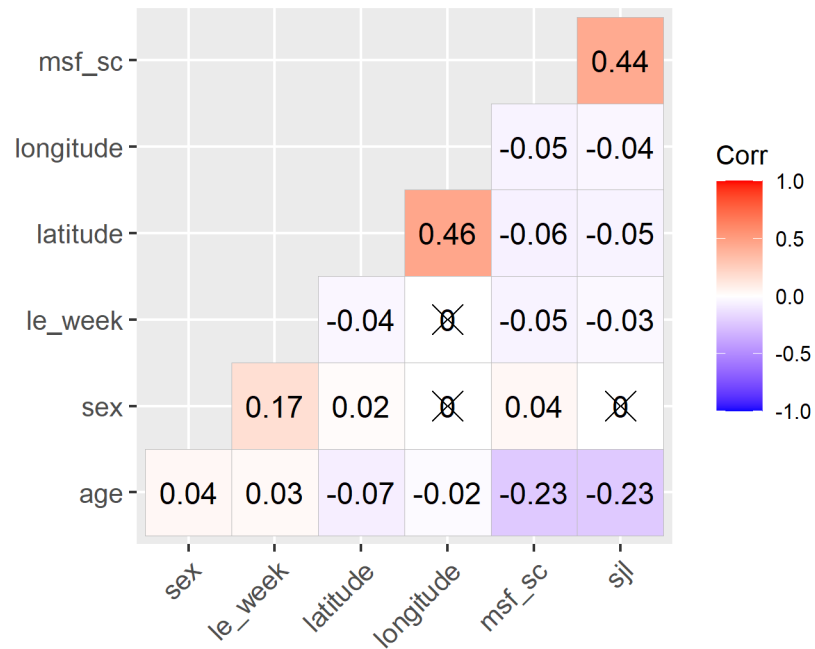
1 # label: fig-ap-ch-4-correlation-matrix
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7
8 source(here::here("R/plot_ggcorrplot.R"))
9
10 cols <- c("sex", "age", "latitude", "longitude", "msf_sc", "sjl", "le_week")
11
12 ggcorrplot <-
13   data |>
14   plot_ggcorrplot(
15     cols = cols,
16     na_rm = TRUE,
17     text_size = text_size,

```

```

18   hc_order = TRUE
19   )

```



C.6 Age series

```

1  # label: fig-ap-ch-4-age-series
2  #| message: false
3  #| warning: false
4  #| code-fold: false
5
6  # library(here)
7  # library(latex2exp)
8
9  source(here::here("R/plot_age_series.R"))
10 #>
11 #> Attaching package: 'magrittr'
12 #> The following object is masked from 'package:rlang':
13 #>
14 #>   set_names

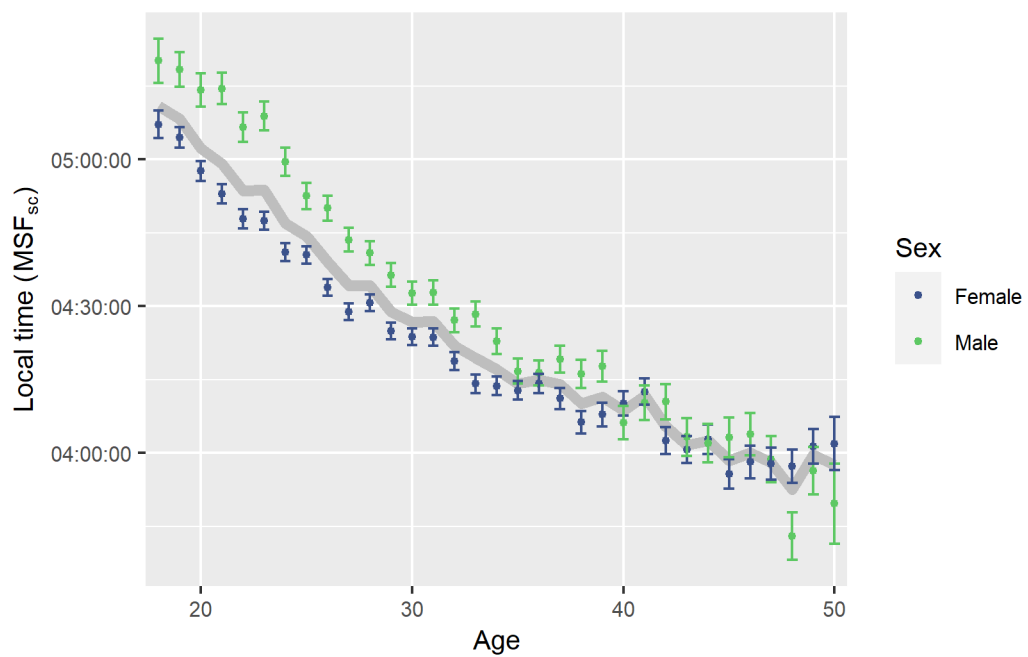
```



```

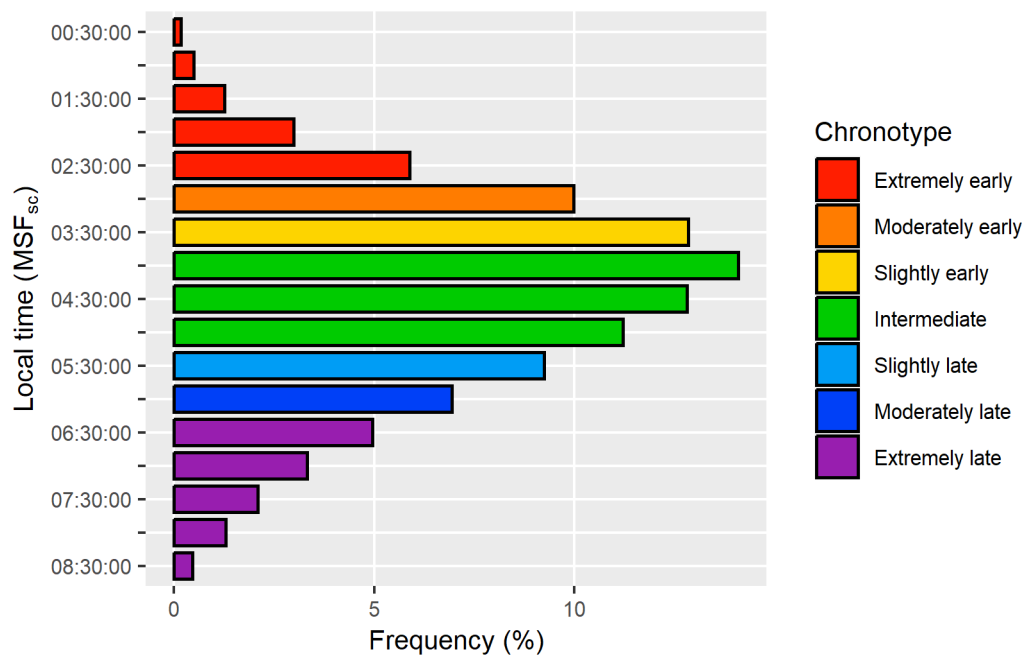
15
16 col <- "msf_sc"
17 y_lab <- latex2exp::TeX("Local time ( $MSF_{sc}$ )")
18
19 data |>
20   dplyr::filter(age <= 50) |>
21   plot_age_series(
22     col = col,
23     y_lab = y_lab,
24     line_width = 2,
25     boundary = 0.5,
26     point_size = 1,
27     error_bar_width = 0.5,
28     error_bar_linewidth = 0.5,
29     error_bar = TRUE,
30     text_size = text_size
31   )

```



C.7 Chronotype

```
1 # label: fig-ap-ch-4-chronotype
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7 # library(latex2exp)
8
9 source(here::here("R/plot_chronotype.R"))
10
11 col <- "msf_sc"
12 y_lab <- latex2exp::TeX("Local time ( $MSF_{sc}$ )")
13
14 data |>
15   plot_chronotype(
16     col = col,
17     x_lab = "Frequency (%)",
18     y_lab = y_lab,
19     col_width = 0.8,
20     col_border = 0.6,
21     text_size = text_size,
22     legend_position = "right",
23     chronotype_cuts = FALSE
24   )
```



C.8 Latitude series

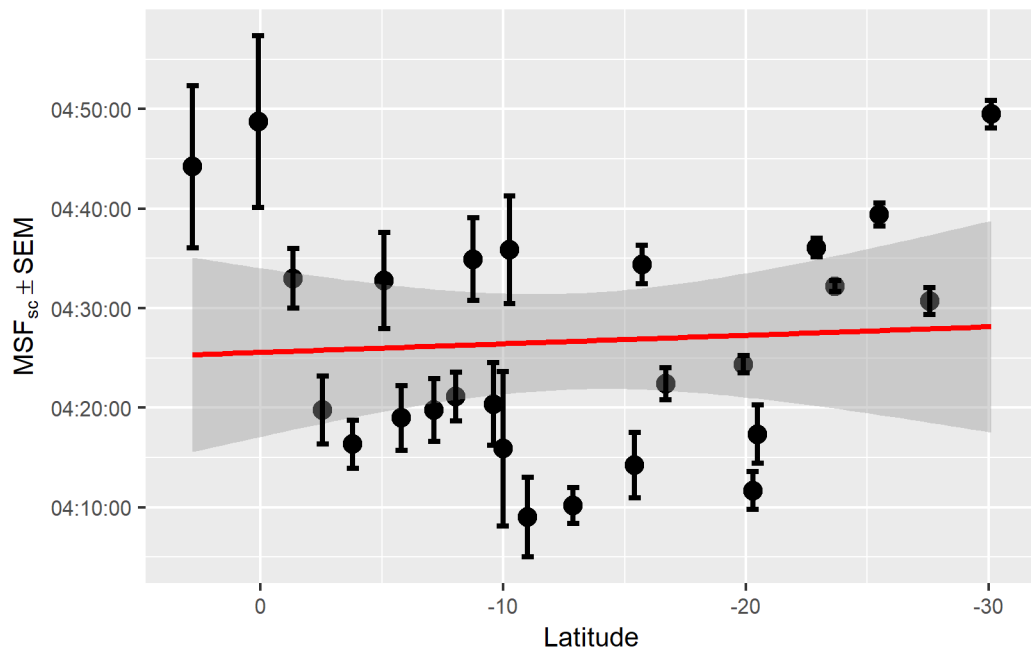
```

1 # label: fig-ap-ch-4-latitude-series
2 #| message: false
3 #| warning: false
4 #| code-fold: false
5
6 # library(here)
7 # library(latex2exp)
8
9 source(here::here("R/plot_latitude_series.R"))
10
11 col <- "msf_sc"
12 y_lab <- latex2exp::TeX("$MSF_{sc} \\pm SEM$")
13
14 data |>
15   dplyr::filter(age <= 50) |>
16   plot_latitude_series(
17     col = col,
```

```

18     y_lab = y_lab,
19     line_width = 2,
20     point_size = 3,
21     error_bar_width = 0.5,
22     error_bar_linewidth = 1,
23     error_bar = TRUE,
24     text_size = text_size
25 )

```



C.9 Statistics

C.9.1 Numerical variables

```

1 # library(here)
2 # library(magrittr)
3
4 source(here::here("R/stats_sum.R"))
5 source(here::here("R/utils.R"))
6
7 col <- "age"

```

```

8
9 data |>
10   magrittr::extract2(col) |>
11   stats_sum(print = FALSE) |>
12   list_as_tibble()

```

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

C.9.2 Sex

```

1 # See <https://sidra.ibge.gov.br> to learn more.
2
3 # library(dplyr)
4 library(magrittr)
5 # library(sidrar)
6 # library(stringr)
7 # library(rutils)
8
9 rutils:::assert_internet()
10
11 # Brazil's 2010 census data
12 census_data <-

```

```

13  sidrar::get_sidra(x = 1378) %>% # Don't change the pipe
14  dplyr::filter(
15    Sexo %in% c("Homens", "Mulheres", "Total"),
16    stringr::str_detect(Idade, "^(1[8-9]|[2-9][0-9]+) (ano|anos)$"),
17    .[[13]] == "Total",
18    .[[19]] == "Total"
19  ) |>
20  dplyr::transmute(
21    sex = dplyr::case_when(
22      Sexo == "Homens" ~ "Male",
23      Sexo == "Mulheres" ~ "Female",
24      Sexo == "Total" ~ "Total"
25    ),
26    value = Valor
27  ) |>
28  dplyr::group_by(sex) |>
29  dplyr::summarise(n = sum(value)) |>
30  dplyr::ungroup()
31
32  census_data <-
33  dplyr::bind_rows(
34    census_data |>
35      dplyr::filter(sex != "Total") |>
36      dplyr::mutate(
37        n_rel = n / sum(n[sex != "Total"]),
38        n_per = round(n_rel * 100, 3)
39      ),
40    census_data |>
41      dplyr::filter(sex == "Total") |>
42      dplyr::mutate(n_rel = 1, n_per = 100)
43  ) |>
44  dplyr::as_tibble() |>

```

```

45   dplyr::arrange(sex)
46
47   count <- data |>
48     dplyr::select(sex) |>
49     dplyr::group_by(sex) |>
50     dplyr::summarise(n = dplyr::n()) |>
51     dplyr::ungroup() |>
52     dplyr::mutate(
53       n_rel = n / sum(n),
54       n_per = round(n_rel * 100, 3)
55     ) |>
56     dplyr::arrange(dplyr::desc(n_rel)) |>
57     dplyr::bind_rows(
58       dplyr::tibble(
59         sex = "Total",
60         n = nrow(tidyr::drop_na(data, sex)),
61         n_rel = 1,
62         n_per = 100
63       )
64     )
65
66   count <-
67     dplyr::left_join(
68       count, census_data,
69       by = "sex",
70       suffix = c("_sample", "_census")
71     ) |>
72     dplyr::mutate(
73       n_rel_diff = n_rel_sample - n_rel_census,
74       n_per_diff = n_per_sample - n_per_census
75     ) |>
76     dplyr::relocate(

```

```

77     sex, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
78     n_per_sample, n_per_census, n_per_diff
79   )
80
81   ## Difference from the sample to the census
82   count

```

sex	n_sample	n_census	n_rel_sample	n_rel_census	n_rel_diff	n_per_sample
Female	52463	69631672	0.662	0.518	0.144	66.2
Male	26735	64809723	0.338	0.482	-0.144	33.8
Total	79198	134441395	1.000	1.000	0.000	100.0

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

sex	n_per_sample	n_per_census	n_per_diff
Female	66.2	51.8	14.4
Male	33.8	48.2	-14.4
Total	100.0	100.0	0.0

```

1 sum(count$n_per_diff)
2 #> [1] -7.11e-15

```

C.9.3 Sex and Age

```

1 # library(here)
2 # library(dplyr)
3 # library(magrittr)
4
5 source(here::here("R/stats_sum.R"))
6 source(here::here("R/utils.R"))
7
8 value <- "Male"
9
10 data |>

```



```

11   dplyr::filter(sex == value) |>
12   magrittr::extract2("age") |>
13   stats_sum(print = FALSE) |>
14   list_as_tibble()

```

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

```

1   # See <https://sidra.ibge.gov.br> to learn more.
2
3   # library(dplyr)
4   library(magrittr)
5   # library(sidrar)
6   # library(stats)
7   # library(stringr)
8   # library(rutils)
9
10  rutils:::assert_internet()
11
12  # Brazil's 2010 census data
13  census_data <-
14    sidrar::get_sidra(x = 1378) %>% # Don't change the pipe
15    dplyr::filter(
16      Sexo %in% c("Homens", "Mulheres", "Total"),

```

```

17   stringr::str_detect(Idade, "^(1[8-9]|[2-9][0-9]+) (ano|anos)$"),
18   .[[13]] == "Total",
19   .[[19]] == "Total"
20 ) |>
21 dplyr::transmute(
22   sex = dplyr::case_when(
23     Sexo == "Homens" ~ "Male",
24     Sexo == "Mulheres" ~ "Female",
25     Sexo == "Total" ~ "Total"
26   ),
27   age = as.numeric(stringr::str_extract(Idade, "\\d+")),
28   value = Valor
29 ) |>
30 dplyr::group_by(sex) |>
31 dplyr::summarise(
32   mean = stats::weighted.mean(age, value),
33   sd = sqrt(Hmisc::wtd.var(age, value))
34 ) |>
35 dplyr::ungroup() |>
36 dplyr::mutate(
37   min = c(18, 18, 18),
38   max = c(100, 100, 100)
39 ) |>
40 dplyr::relocate(sex, mean, sd, min, max) |>
41 dplyr::as_tibble()
42
43 count <- data |>
44   dplyr::select(sex, age) |>
45   dplyr::group_by(sex) |>
46   dplyr::mutate(sex = as.character(sex)) |>
47   dplyr::summarise(
48     mean = mean(age, na.rm = TRUE),

```

```

49     sd = stats::sd(age, na.rm = TRUE),
50     min = min(age, na.rm = TRUE),
51     max = max(age, na.rm = TRUE)
52   ) |>
53   dplyr::ungroup() |>
54   dplyr::bind_rows(
55     dplyr::tibble(
56       sex = "Total",
57       mean = mean(data$age, na.rm = TRUE),
58       sd = stats::sd(data$age, na.rm = TRUE),
59       min = min(data$age, na.rm = TRUE),
60       max = max(data$age, na.rm = TRUE)
61     )
62   )
63
64   count <-
65     dplyr::left_join(
66       count,
67       census_data,
68       by = "sex",
69       suffix = c("_sample", "_census")
70     ) |>
71     dplyr::mutate(mean_diff = mean_sample - mean_census) |>
72     dplyr::relocate(
73       sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census,
74       min_sample, min_census, max_sample, max_census
75     )
76
77   count

```

sex	mean_sample	mean_census	mean_diff	sd_sample	sd_census	min_sample
Female	31.8	41.6	-9.89	9.34	16.9	18
Male	32.4	40.4	-7.94	9.00	16.2	18
Total	32.0	41.0	-9.05	9.23	16.6	18

```

1
2 count |>
3   dplyr::select(
4     sex, mean_sample, mean_census, mean_diff, sd_sample, sd_census
5   )

```

sex	mean_sample	mean_census	mean_diff	sd_sample	sd_census
Female	31.8	41.6	-9.89	9.34	16.9
Male	32.4	40.4	-7.94	9.00	16.2
Total	32.0	41.0	-9.05	9.23	16.6

```

1
2 sum(count$mean_diff)
3 #> [1] -26.9

```

C.9.4 Longitudinal range

Sample

```

1 # library(here)
2 # library(dplyr)
3 # library(magrittr)
4
5 source(here::here("R/stats_sum.R"))
6 source(here::here("R/utls.R"))
7
8 stats <-
9   data |>
10   magrittr::extract2("longitude") |>
11   stats_sum(print = FALSE)

```

```

12
13 abs(stats$max - stats$min)
14 #> [1] 33
15 stats |> list_as_tibble()

```

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

Brazil

```

1 # library(measurements)
2
3 change_sign <- function(x) x * (-1)
4
5 ## Ponta do Seixas, PB (7° 09' 18" S, 34° 47' 34" O)
6 min <-
7   measurements::conv_unit("34 47 34", from = "deg_min_sec", to = "dec_deg") |>
8   as.numeric() |>
9   change_sign()
10
11 ## Nascente do rio Moa, AC (7° 32' 09" S, 73° 59' 26" O)
12 max <-
13   measurements::conv_unit("73 59 26", from = "deg_min_sec", to = "dec_deg") |>

```

```

14   as.numeric() |>
15   change_sign()
16
17   min
18   #> [1] -34.8
19   max
20   #> [1] -74
21   abs(max - min)
22   #> [1] 39.2

```

C.9.5 Latitudinal range

Sample

```

1   # library(here)
2   # library(dplyr)
3   # library(magrittr)
4
5   source(here::here("R/stats_sum.R"))
6   source(here::here("R/utls.R"))
7
8   stats <-
9     data |>
10    magrittr::extract2("latitude") |>
11    stats_sum(print = FALSE)
12
13   abs(stats$max - stats$min)
14   #> [1] 32.9
15   stats |> list_as_tibble()

```

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

Brazil

```

1  # library(measurements)
2
3  change_sign <- function(x) x * (-1)
4
5  ## Arroio Chuí, RS (33° 45' 07" S, 53° 23' 50" O)
6  min <-
7    measurements::conv_unit("33 45 07", from = "deg_min_sec", to = "dec_deg") |>
8    as.numeric() |>
9    change_sign()
10
11  ## Nascente do rio Ailã, RR (5° 16' 19" N, 60° 12' 45" O)
12  max <-
13    measurements::conv_unit("5 16 19", from = "deg_min_sec", to = "dec_deg") |>
14    as.numeric()
15
16  min
17  #> [1] -33.8
18  max
19  #> [1] 5.27

```

```

20 abs(max - min)
21 #> [1] 39

```

C.9.6 Region

```

1  # See <https://sidra.ibge.gov.br> to learn more.
2
3  # library(dplyr)
4  # library(sidrar)
5  # library(stats)
6  # library(stringr)
7  # library(rutils)
8
9  rutils:::assert_internet()
10
11 # Brazil's 2022 census data
12 census_data <-
13   sidrar::get_sidra(x = 4714, variable = 93, geo = "Region") |>
14   dplyr::select(dplyr::all_of(c("Valor", "Grande Região"))) |>
15   dplyr::transmute(
16     col = `Grande Região`,
17     n = Valor,
18     n_rel = n / sum(n),
19     n_per = round(n_rel * 100, 3)
20   ) |>
21   dplyr::mutate(
22     col = dplyr::case_when(
23       col == "Norte" ~ "North",
24       col == "Nordeste" ~ "Northeast",
25       col == "Centro-Oeste" ~ "Midwest",
26       col == "Sudeste" ~ "Southeast",
27       col == "Sul" ~ "South"

```



```

28   )
29   ) |>
30   dplyr::as_tibble() |>
31   dplyr::arrange(dplyr::desc(n_rel))
32
33 count <- data |>
34   magrittr::extract2("region") |>
35   stats_sum(print = FALSE) |>
36   magrittr::extract2("count") |>
37   dplyr::mutate(
38     n_rel = n / sum(n),
39     n_per = round(n_rel * 100, 3)
40   ) |>
41   dplyr::arrange(dplyr::desc(n_rel))
42
43 count <-
44   dplyr::left_join(
45     count, census_data, by = "col", suffix = c("_sample", "_census")
46   ) |>
47   dplyr::mutate(
48     n_rel_diff = n_rel_sample - n_rel_census,
49     n_per_diff = n_per_sample - n_per_census
50   ) |>
51   dplyr::relocate(
52     col, n_sample, n_census, n_rel_sample, n_rel_census, n_rel_diff,
53     n_per_sample, n_per_census, n_per_diff
54   )
55
56 ## Difference from the sample to the census
57 count

```

col	n_sample	n_census	n_rel_sample	n_rel_census	n_rel_diff	n_per_sample
Southeast	47966	84847187	0.606	0.418	0.188	60.56
South	13560	29933315	0.171	0.147	0.024	17.12
Northeast	9138	54644582	0.115	0.269	-0.154	11.54
Midwest	6563	16287809	0.083	0.080	0.003	8.29
North	1971	17349619	0.025	0.085	-0.061	2.49

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

col	n_per_sample	n_per_census	n_per_diff
Southeast	60.56	41.78	18.781
South	17.12	14.74	2.381
Northeast	11.54	26.91	-15.372
Midwest	8.29	8.02	0.266
North	2.49	8.54	-6.055

```
1 sum(count$n_per_diff)
2 #> [1] 0.001
```

C.9.7 State

```
1 # library(here)
2 # library(dplyr)
3 # library(magrittr)
4
5 source(here::here("R/stats_sum.R"))
6
7 data |>
8   magrittr::extract2("state") |>
9   stats_sum(print = FALSE) |>
10  magrittr::extract2("count") |>
11  dplyr::mutate(
12    n_rel = n / sum(n),
13    n_per = round(n_rel * 100, 3)
14  ) |>
15  dplyr::arrange(dplyr::desc(n_rel))
```

col	n	n_rel	n_per
São Paulo	26379	0.333	33.308
Minas Gerais	10115	0.128	12.772
Rio de Janeiro	9381	0.118	11.845
Paraná	5517	0.070	6.966
Rio Grande do Sul	4097	0.052	5.173
Santa Catarina	3946	0.050	4.982
Goiás	2674	0.034	3.376
Bahia	2522	0.032	3.184
Espírito Santo	2091	0.026	2.640
Distrito Federal	2087	0.026	2.635
Pernambuco	1550	0.020	1.957
Ceará	1398	0.018	1.765
Mato Grosso do Sul	1014	0.013	1.280
Pará	938	0.012	1.184
Rio Grande do Norte	789	0.010	0.996
Mato Grosso	788	0.010	0.995
Paraíba	773	0.010	0.976
Maranhão	652	0.008	0.823
Sergipe	533	0.007	0.673
Alagoas	526	0.007	0.664
Rondônia	401	0.005	0.506
Piauí	395	0.005	0.499
Tocantins	268	0.003	0.338
Acre	132	0.002	0.167
Roraima	119	0.002	0.150
Amapá	113	0.001	0.143

D APPENDICE: 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.

E APPENDICE: 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.

The analyses here are 100% reproducible. You can find all the code used in this thesis along with the computational notebooks at <https://github.com/danielvartan/masterthesis/>.

E.1 Hypothesis

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. 2017, 2014; Colin S. Pittendrigh, Kyner, and Takamura 1991; Randler and 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.

$$\begin{cases} H_0 : R_{\text{res}}^2 \geq R_{\text{full}}^2 \\ H_a : R_{\text{res}}^2 < R_{\text{full}}^2 \end{cases}$$

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

$$F = \frac{R_F^2 - R_R^2 / (k_F - k_R)}{(1 - R_F^2) / (N - k_F - 1)}$$

Where:

- R_F^2 = Coefficient of determination for the **full** model
- R_R^2 = 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

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

E.2 Assumptions

See DeGroot and Schervish (2012, 736–38) to learn more.

Warning

The predictor is known. Either the vectors z_1, \dots, z_n are known ahead of time, or they are the observed values of random vectors Z_1, \dots, Z_n on whose values we condition before computing the joint distribution of (Y_1, \dots, Y_n) .

Warning

Normality. For $i = 1, \dots, n$, the conditional distribution of Y_i given the vectors z_1, \dots, z_n is a normal distribution (**normality assumption**).

Warning

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

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

for $i = 1, \dots, n$ (**zero error mean assumption**).

Warning

Common variance. The observations Y_1, \dots, Y_n have the same variance σ^2 (**homoscedasticity assumption**).

Warning

Independence. The random variables Y_1, \dots, Y_n are independent given the observed z_1, \dots, z_n (**independent errors assumption**).

E.3 Data preparation

```

1  # library(dplyr)
2  # library(here)
3  library(targets)
4  # library(tidyr)
5
6  source(here::here("R/utils.R"))
7
8  utc_minus_3_states <- c(
9    "Amapá", "Pará", "Maranhão", "Tocantins", "Piauí", "Ceará",
10   "Rio Grande do Norte", "Paraíba", "Pernambuco", "Alagoas", "Sergipe",
11   "Bahia", "Distrito Federal", "Goiás", "Minas Gerais", "Espírito Santo",
12   "Rio de Janeiro", "São Paulo", "Paraná", "Santa Catarina",
13   "Rio Grande do Sul"
14 )
15
16 data <-
17   targets::tar_read("geocoded_data", store = here::here("_targets")) |>
18   dplyr::filter(state %in% utc_minus_3_states) |>
19   dplyr::select(msf_sc, age, sex, state, latitude, longitude) |>
20   dplyr::mutate(msf_sc = transform_time(msf_sc)) |>
21   tidyr::drop_na(msf_sc, age, sex, latitude)

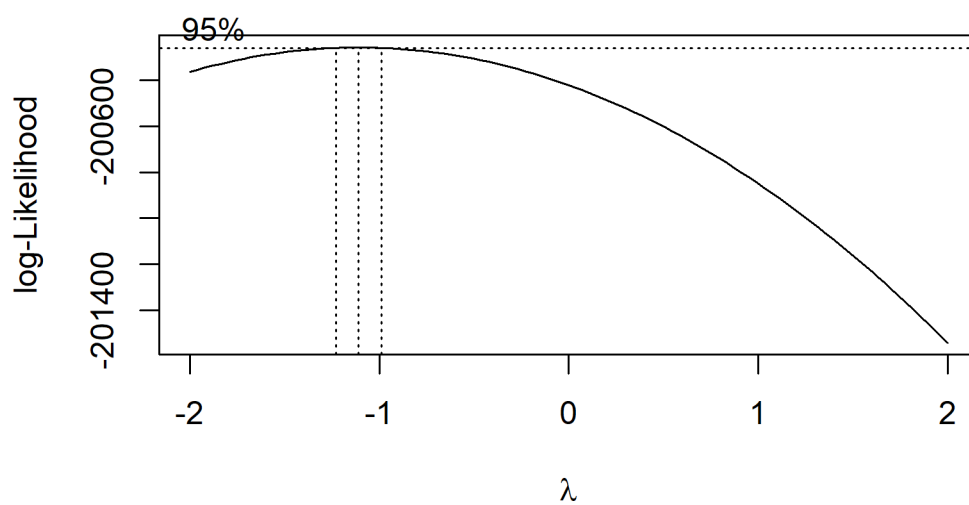
```

E.4 Restricted model

```

1 # library(MASS)
2
3 box_cox <- MASS::boxcox(msf_sc ~ age + sex, data = data)

```



```

1 lambda <- box_cox$x[which.max(box_cox$y)]
2
3 lambda
4 #> [1] -1.1111

```

```

1 # library(stats)
2
3 res_model <- stats::lm(
4   ((msf_sc^lambda - 1) / lambda) ~ age + sex, data = data
5 )

```

```

1 # library(broom)
2
3 broom::tidy(res_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

```

1 # library(broom)
2 # library(dplyr)
3 # library(tidyr)
4
5 broom::glance(res_model) |> tidyr::pivot_longer(cols = dplyr::everything())

```

name	value
r.squared	0.054
adj.r.squared	0.054
sigma	0.000
statistic	2178.876
p.value	0.000
df	2.000
logLik	1106194.897
AIC	-2212381.794
BIC	-2212344.801
deviance	0.000
df.residual	76741.000
nobs	76744.000

```

1 # library(olsrr)
2
3 # res_model |> olsrr::ols_regress()
4 res_model |> summary()
5 #>
6 #> Call:
7 #> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex,
8 #> data = data)
9 #>
10 #> Residuals:
11 #> Min 1Q Median 3Q Max
12 #> -0.0000004859 -0.0000000911 -0.0000000031 0.0000000916 0.0000004204

```

```

13 #>
14 #> Coefficients:
15 #>               Estimate      Std. Error    t value Pr(>|t|)
16 #> (Intercept)  0.8999976603602  0.00000000017524  513579298.2   <2e-16 ***
17 #> age          -0.00000000033812  0.00000000000519    -65.1   <2e-16 ***
18 #> sexMale      0.00000000132309  0.00000000010162     13.0   <2e-16 ***
19 #> ---
20 #> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21 #>
22 #> Residual standard error: 0.000000133 on 76741 degrees of freedom
23 #> Multiple R-squared:  0.0537, Adjusted R-squared:  0.0537
24 #> F-statistic: 2.18e+03 on 2 and 76741 DF,  p-value: <2e-16

```

E.4.1 Residual diagnostics

Normality and zero mean error assumption.

```

1 # library(here)
2 # library(stats)
3
4 source(here::here("R/stats_sum.R"))
5 source(here::here("R/utils.R"))
6
7 res_model |>
8   stats::residuals() |>
9   stats_sum(print = FALSE) |>
10  list_as_tibble()

```

name	value
n	76744
n_rm_na	76744
n_na	0
mean	6.60699976667332e-23
var	0.00000000000000176852866826985
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

```

1 # See `?moments::agostino.test` & `?fBasics::dagoTest()` to learn more.
2
3 # library(fBasics)
4 # library(moments)
5
6 source(here::here("R/normality_sum.R"))
7
8 res_model |>
9   stats::residuals() |>
10  normality_sum()

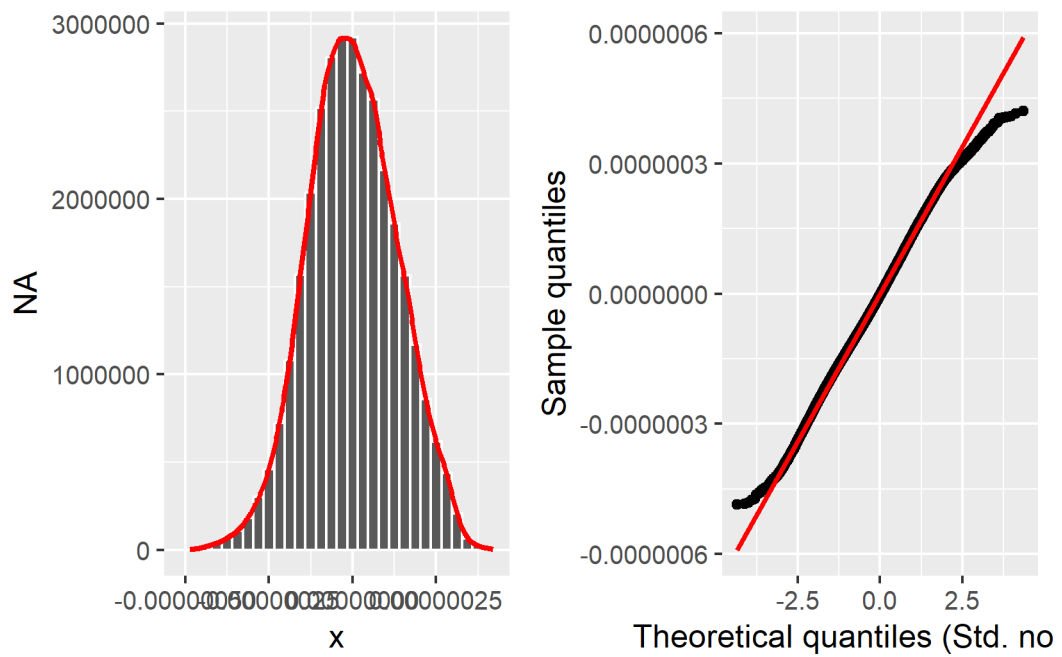
```

test	p_value
Anderson-Darling	0.000
Bonett-Seier	0.000
Cramer-von Mises	0.000
D'Agostino Omnibus Test	NA
D'Agostino Skewness Test	0.234
D'Agostino Kurtosis Test	NA
Jarque-Bera	0.000
Lilliefors (K-S)	0.000
Pearson chi-square	0.000
Shapiro-Francia	NA
Shapiro-Wilk	NA

Correlation between observed residuals and expected residuals under normality.

```
1 # library(olsrr)
2
3 res_model |> olsrr::ols_test_correlation()
4 #> [1] 0.99929
```

```
1 # library(cowplot)
2 # library(olsrr)
3 # library(stats)
4
5 source(here::here("R/test_normality.R"))
6
7 # res_model |> olsrr::ols_plot_resid_qq()
8
9 qq_plot <- res_model |>
10   stats::residuals() |>
11   plot_qq(print = FALSE)
12
13 hist_plot <- res_model |>
14   stats::residuals() |>
15   plot_hist(print = FALSE)
16
17 cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)
```

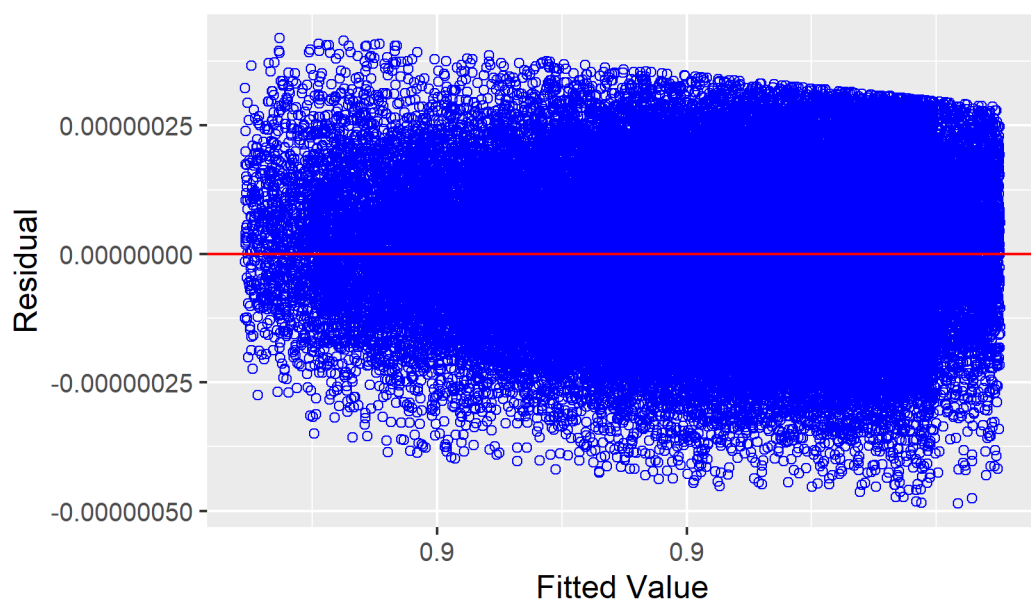


```

1 # library(olsrr)
2
3 # Linear mean assumption
4
5 res_model |> olsrr::ols_plot_resid_fit()

```

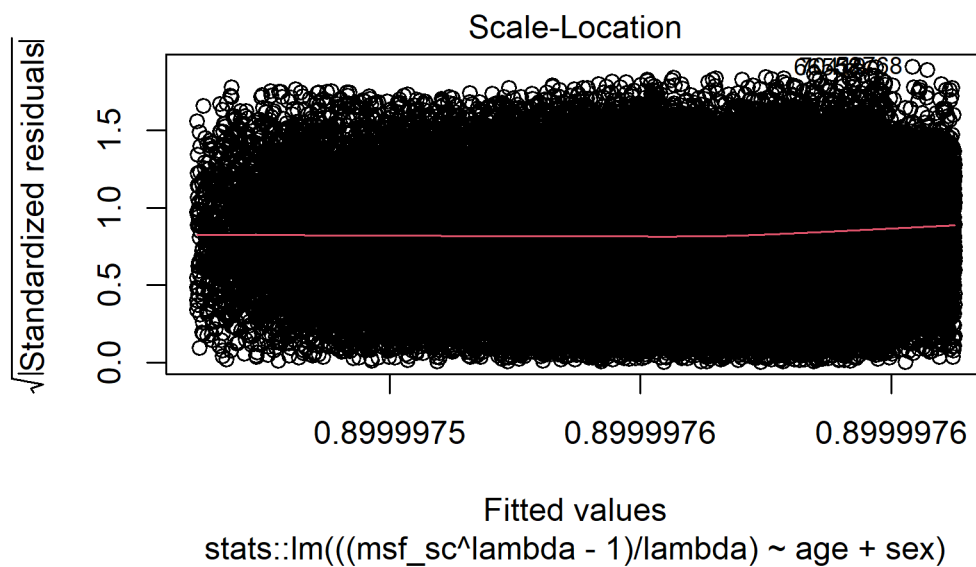
Residual vs Fitted Values



```

1 res_model |> plot(3)

```



E.4.2 Heteroskedasticity

Homoscedasticity assumption.

```

1 # library(olsrr)
2
3 # "It test whether variance of errors from a regression is dependent on the values of a indep
4
5 res_model |> olsrr::ols_test_breusch_pagan()
6 #>
7 #> Breusch Pagan Test for Heteroskedasticity
8 #> -----
9 #> Ho: the variance is constant
10 #> Ha: the variance is not constant
11 #>
12 #> Data
13 #> -----
14 #> Response : ((msf_sc^lambda - 1)/lambda)
15 #> Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
16 #>

```

```

17 #>          Test Summary
18 #> -----
19 #> DF          =      1
20 #> Chi2         =    70149.3586
21 #> Prob > Chi2  =     0.0000

```

```

1 # library(olsrr)
2
3 res_model |> olsrr::ols_test_score()
4 #>
5 #> Score Test for Heteroskedasticity
6 #> -----
7 #> Ho: Variance is homogenous
8 #> Ha: Variance is not homogenous
9 #>
10 #> Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
11 #>
12 #>          Test Summary
13 #> -----
14 #> DF          =      1
15 #> Chi2         =     0.0000
16 #> Prob > Chi2  =     1.0000

```

E.4.3 Collinearity diagnostics

Independence assumption.

```

1 # library(olsrr)
2
3 res_model |> olsrr::ols_coll_diag()
4 #> Tolerance and Variance Inflation Factor
5 #> -----
6 #> Variables Tolerance    VIF

```

```

7 #> 1      age      0.9988 1.0012
8 #> 2    sexMale      0.9988 1.0012
9 #>
10 #>
11 #> Eigenvalue and Condition Index
12 #> -----
13 #> Eigenvalue Condition Index intercept      age      sexMale
14 #> 1      2.422418              1.00000 0.011753 0.011936 0.0669897
15 #> 2      0.538450              2.12111 0.015824 0.018848 0.9280439
16 #> 3      0.039132              7.8679 0.972423 0.969216 0.0049664

```

The variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables (e.g., VIF equal to 1 = variables are not correlated).

```

1 # library(car)
2
3 res_model |> car::vif()
4 #>      age      sex
5 #> 1.0012 1.0012

```

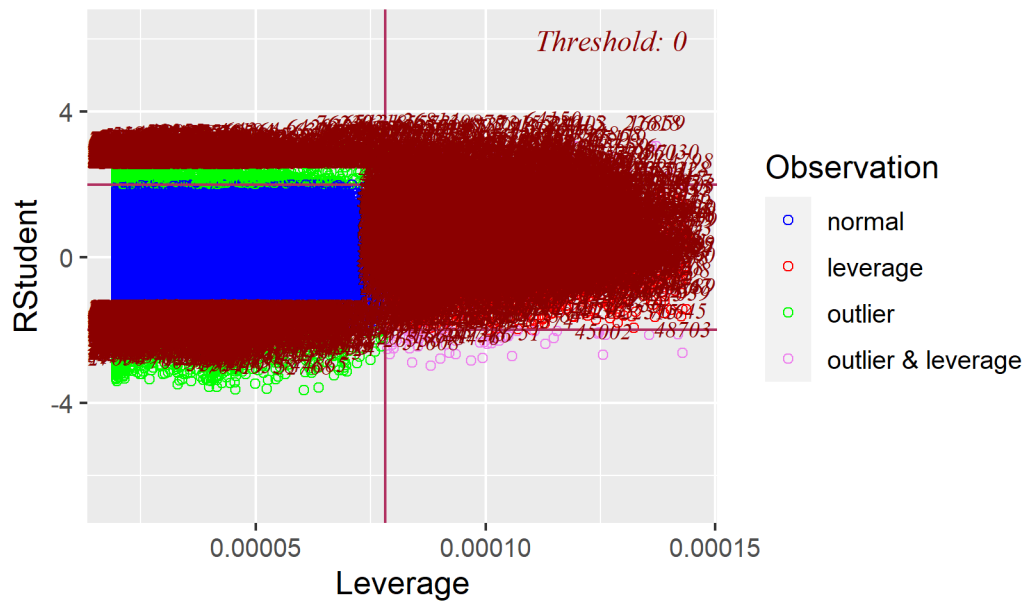
E.4.4 Measures of influence

```

1 # library(olsrr)
2
3 res_model |> olsrr::ols_plot_resid_lev()

```


Outlier and Leverage Diagnostics for ((msf_sc^lambda - 1

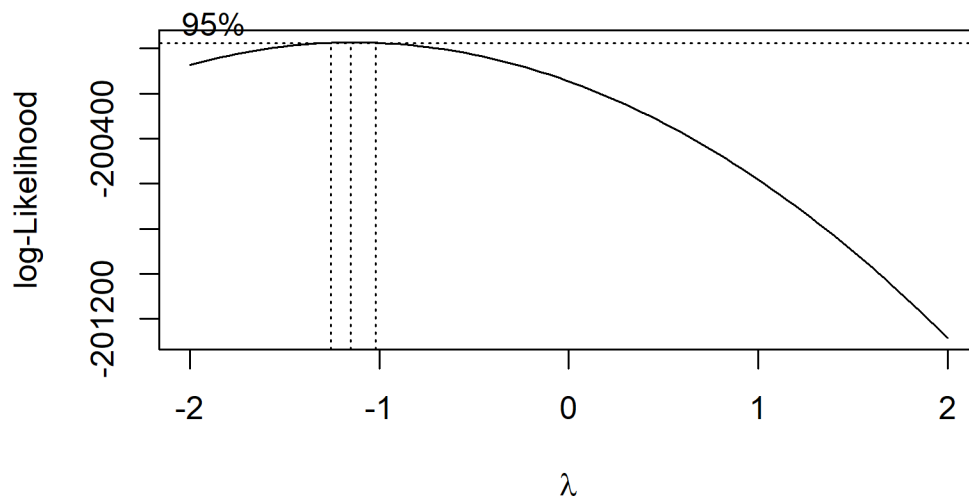


E.5 Full model

```

1 # library(MASS)
2
3 box_cox <- MASS::boxcox(
4   msf_sc ~ age + sex + latitude, data = data
5 )

```



```

1
2 box_cox$x[which.max(box_cox$y)] # lambda
3 #> [1] -1.1515

```

```

1 lambda # The same lambda of the restricted model
2 #> [1] -1.1111

```

```

1 # library(stats)
2
3 full_model <- stats::lm(
4   ((msf_sc^lambda - 1) / lambda) ~ age + sex + latitude,
5   data = data
6 )

```

```

1 # library(broom)
2
3 # ?broom::tidy.lm
4 broom::tidy(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

```

1 # library(broom)
2 # library(dplyr)
3 # library(tidyr)
4
5 # ?broom::glance.lm
6 broom::glance(full_model) |>
7   tidyr::pivot_longer(cols = dplyr::everything())

```

name	value
r.squared	0.061
adj.r.squared	0.061
sigma	0.000
statistic	1652.979
p.value	0.000
df	3.000
logLik	1106478.331
AIC	-2212946.661
BIC	-2212900.420
deviance	0.000
df.residual	76740.000
nobs	76744.000

```

1 # full_model |> olsrr::ols_regress()
2 full_model |> summary()
3 #>
4 #> Call:
5 #> stats::lm(formula = ((msf_sc^lambda - 1)/lambda) ~ age + sex +
6 #>     latitude, data = data)
7 #>
8 #> Residuals:
9 #>           Min             1Q           Median             3Q            Max
10 #> -0.00000004874 -0.0000000911 -0.0000000034  0.0000000912  0.0000004328
11 #>
12 #> Coefficients:
13 #>             Estimate      Std. Error    t value Pr(>|t|)
14 #> (Intercept)  0.8999976247783  0.0000000022965 391908052.9  <2e-16 ***
15 #> age          -0.0000000034710  0.0000000000519   -66.9    <2e-16 ***
16 #> sexMale       0.0000000137296  0.0000000010127    13.6    <2e-16 ***
17 #> latitude     -0.0000000018222  0.0000000000764   -23.9    <2e-16 ***
18 #> ---
19 #> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
20 #>
21 #> Residual standard error: 0.000000132 on 76740 degrees of freedom

```

```

22 #> Multiple R-squared:  0.0607, Adjusted R-squared:  0.0607
23 #> F-statistic: 1.65e+03 on 3 and 76740 DF,  p-value: <2e-16

```

E.5.1 Residual diagnostics

Normality and zero mean error assumption.

```

1 # library(here)
2 # library(stats)
3
4 source(here::here("R/stats_sum.R"))
5 source(here::here("R/utils.R"))
6
7 full_model |>
8   stats::residuals() |>
9   stats_sum(print = FALSE) |>
10  list_as_tibble()

```

name	value
n	76744
n_rm_na	76744
n_na	0
mean	4.85272564733669e-24
var	0.0000000000000175551361304561
sd	0.00000013249579665203
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

```

1 # library(here)
2 # library(stats)
3

```

```

4 source(here::here("R/normality_sum.R"))
5
6 full_model |>
7   stats::residuals() |>
8   normality_sum()

```

test	p_value
Anderson-Darling	0.000
Bonett-Seier	0.000
Cramer-von Mises	0.000
D'Agostino Omnibus Test	NA
D'Agostino Skewness Test	0.941
D'Agostino Kurtosis Test	NA
Jarque-Bera	0.000
Lilliefors (K-S)	0.000
Pearson chi-square	0.000
Shapiro-Francia	NA
Shapiro-Wilk	NA

Correlation between observed residuals and expected residuals under normality.

```

1 # library(olsrr)
2
3 full_model |> olsrr::ols_test_correlation()
4 #> [1] 0.99929

```

```

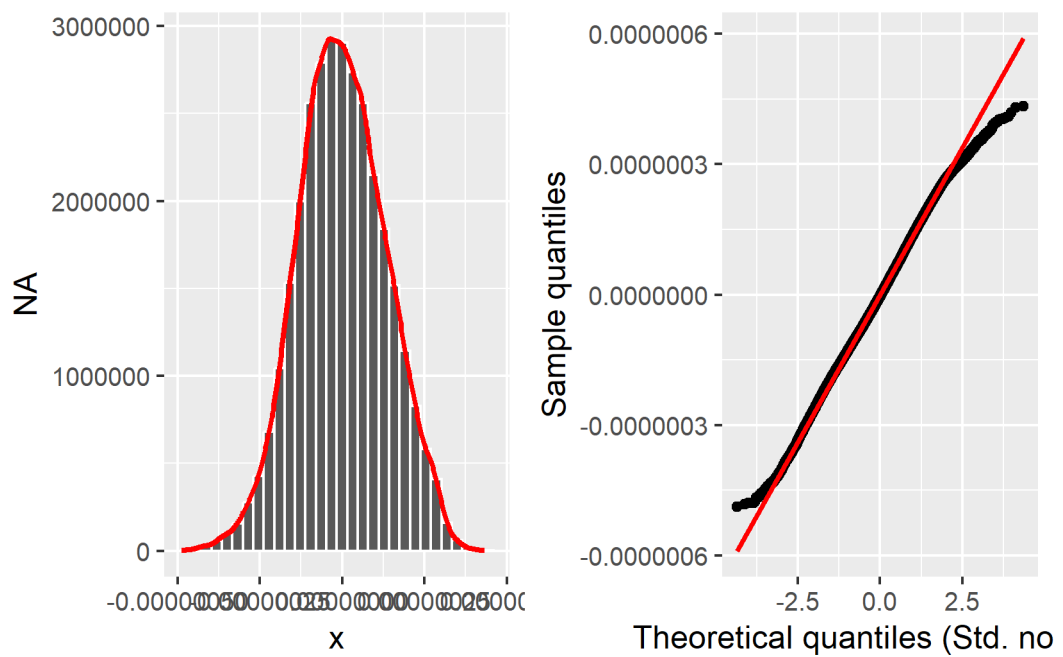
1 # library(here)
2 # library(cowplot)
3 # library(olsrr)
4 # library(stats)
5
6 source(here::here("R/test_normality.R"))
7
8 # full_model |> olsrr::ols_plot_resid_qq()
9
10 hist_plot <- full_model |>
11   stats::residuals() |>

```

```

12   plot_hist(print = FALSE)
13
14   qq_plot <- full_model |>
15     stats::residuals() |>
16     plot_qq(print = FALSE)
17
18   cowplot::plot_grid(hist_plot, qq_plot, ncol = 2, nrow = 1)

```

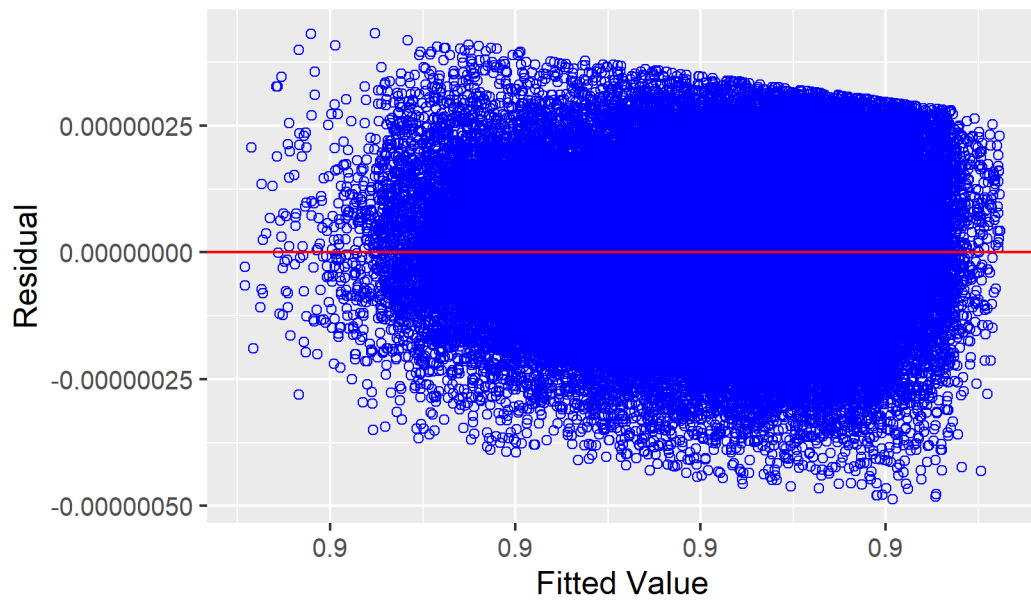


```

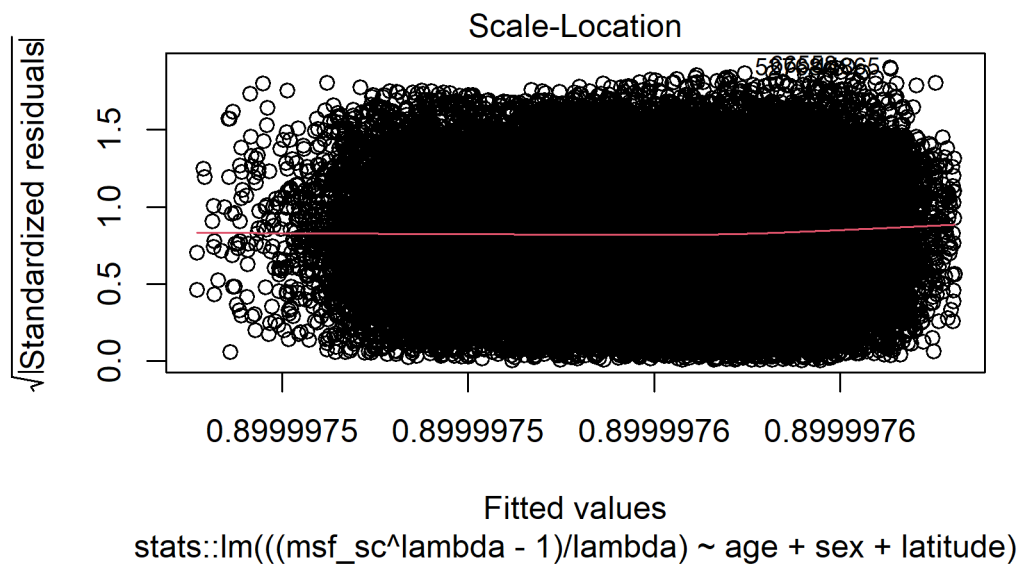
1   # library(olsrr)
2
3   full_model |> olsrr::ols_plot_resid_fit()

```

Residual vs Fitted Values



```
1 full_model |> plot(3)
```



E.5.2 Heteroskedasticity

Homoscedasticity assumption.

```

1 # library(olsrr)
2
3 full_model |> olsrr::ols_test_breusch_pagan()
4 #>
5 #> Breusch Pagan Test for Heteroskedasticity
6 #> -----
7 #> Ho: the variance is constant
8 #> Ha: the variance is not constant
9 #>
10 #>                                Data
11 #> -----
12 #> Response : ((msf_sc^lambda - 1)/lambda)
13 #> Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
14 #>
15 #>          Test Summary
16 #> -----
17 #> DF          =      1
18 #> Chi2         =    70101.1634
19 #> Prob > Chi2  =     0.0000

```

```

1 # library(olsrr)
2
3 full_model |> olsrr::ols_test_score()
4 #>
5 #> Score Test for Heteroskedasticity
6 #> -----
7 #> Ho: Variance is homogenous
8 #> Ha: Variance is not homogenous
9 #>
10 #> Variables: fitted values of ((msf_sc^lambda - 1)/lambda)
11 #>
12 #>          Test Summary

```



```

13 #> -----
14 #> DF          =      1
15 #> Chi2         =      0.000
16 #> Prob > Chi2  =      1.000

```

E.5.3 Collinearity diagnostics

Independence assumption.

```

1 # library(olsrr)
2
3 full_model |> olsrr::ols_coll_diag()
4 #> Tolerance and Variance Inflation Factor
5 #> -----
6 #> Variables Tolerance   VIF
7 #> 1      age  0.99354 1.0065
8 #> 2  sexMale  0.99838 1.0016
9 #> 3 latitude  0.99441 1.0056
10 #>
11 #>
12 #> Eigenvalue and Condition Index
13 #> -----
14 #> Eigenvalue Condition Index  intercept      age  sexMale  latitude
15 #> 1  3.312504                1.0000 0.00377395 0.0064918 0.0304493 0.0068553
16 #> 2  0.584652                2.3803 0.00328127 0.0064143 0.9588857 0.0083393
17 #> 3  0.073700                6.7042 0.00040414 0.5063551 0.0023826 0.5659326
18 #> 4  0.029145                10.6609 0.99254063 0.4807389 0.0082824 0.4188728

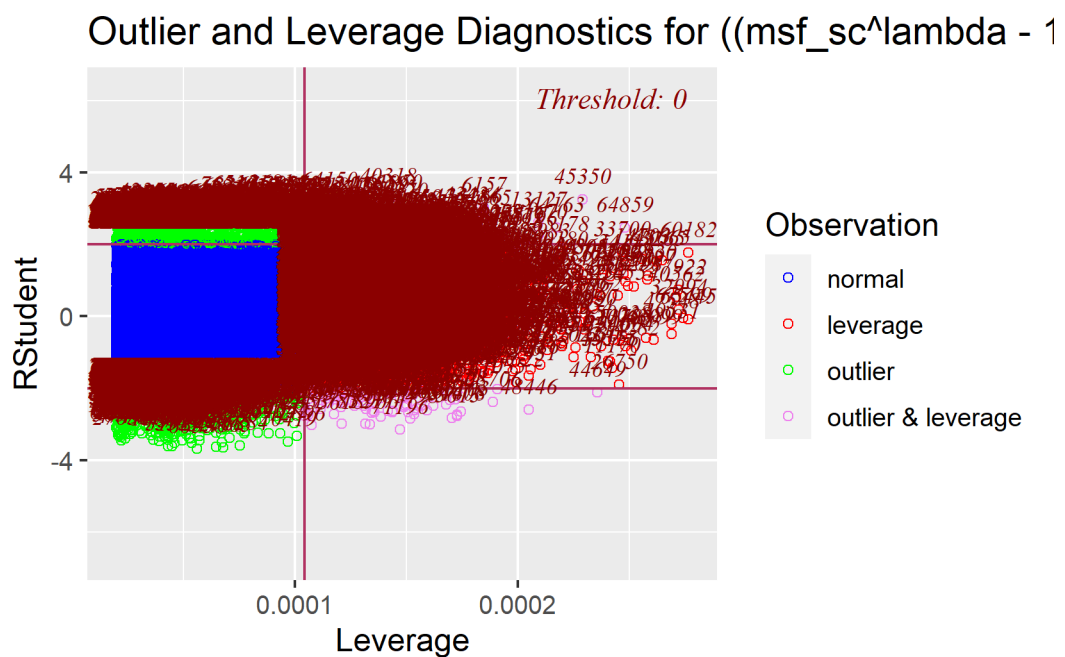
```

The variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables (e.g., VIF equal to 1 = variables are not correlated).

```
1 # library(car)
2
3 full_model |> car::vif()
4 #>      age      sex latitude
5 #>  1.0065  1.0016  1.0056
```

E.5.4 Measures of influence

```
1 # library(olsrr)
2
3 full_model |> olsrr::ols_plot_resid_lev()
```



E.6 Nested regression models test

$$\begin{cases} H_0 : R_{\text{res}}^2 \geq R_{\text{full}}^2 \\ H_a : R_{\text{res}}^2 < R_{\text{full}}^2 \end{cases}$$

$$F = \frac{R_F^2 - R_R^2 / (k_F - k_R)}{(1 - R_F^2) / (N - k_F - 1)}$$

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

```

1 # library(dplyr)
2 # library(here)
3
4 source(here::here("R/utls-stats.R"))
5
6 dplyr::tibble(
7   name = c("r_squared_res", "r_squared_full", "diff"),
8   value = c(
9     r_squared(res_model), r_squared(full_model),
10    r_squared(full_model) - r_squared(res_model)
11  )
12 )

```

name	value
r_squared_res	0.054
r_squared_full	0.061
diff	0.007

```

1 # library(stats)
2
3 stats::anova(res_model, full_model)

```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
76741	0	NA	NA	NA	NA
76740	0	1	0	568.94	0

```

1 # library(stats)
2 # library(here)
3
4 source(here::here("R/utls-stats.R"))
5

```

```

6  n <- nrow(data)
7  k_res <- length(stats::coefficients(res_model)) - 1
8  k_full <- length(stats::coefficients(full_model)) - 1
9
10 ((r_squared(full_model) - r_squared(res_model)) / (k_full - k_res)) / ((1 - r_squared(full_m
11 #> [1] 568.94

```

$$f^2 = \frac{R_F^2 - R_R^2}{1 - R_F^2}$$

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

```

1  # library(here)
2
3  source(here::here("R/cohens_f_squared.R"))
4  source(here::here("R/utlis-stats.R"))
5
6  cohens_f_squared_summary(
7    adj_r_squared(res_model),
8    adj_r_squared(full_model)
9  )

```

name	value
f_squared	0.00740068896515648
effect_size	Negligible

E.7 Group test

$$\begin{cases} H_0 : MSF_{sc}^{0^{\circ}} \geq MSF_{sc}^{30^{\circ}} \\ H_a : MSF_{sc}^{0^{\circ}} < MSF_{sc}^{30^{\circ}} \end{cases}$$

```

1 # library(dplyr)
2 # library(here)
3 # library(magrittr)
4
5 source(here::here("R/stats_sum.R"))
6 source(here::here("R/utils.R"))
7
8 group_1 <- "Amapá" # Boa vista (0° 2' 18.84" N, 51° 3' 59.1" W)
9
10 msf_sc_group_1 <-
11   data |>
12     dplyr::filter(state == group_1) |>
13     magrittr::extract2("msf_sc")
14
15 stats_sum_group_1 <-
16   data |>
17     dplyr::filter(state == group_1) |>
18     magrittr::extract2("msf_sc") |>
19     stats_sum(print = FALSE)
20
21 stats_sum_group_1 |> list_as_tibble()

```

name	value
n	113
n_rm_na	113
n_na	0
mean	103564.475347661
var	28720417.6933099
sd	5359.14337308771
min	92100
q_1	99000
median	104164.285714286
q_3	107142.857142857
max	114450
iqr	8142.85714285714
skewness	-0.0419134073448402
kurtosis	2.19330502090315

```

1 # library(dplyr)
2 # library(here)
3 # library(magrittr)
4
5 source(here::here("R/stats_sum.R"))
6 source(here::here("R/utls.R"))
7
8 group_2 <- "Rio Grande do Sul" # Porto Alegre (30° 01' 58" S, 51° 13' 48" O)
9
10 msf_sc_group_2 <- data |>
11   dplyr::filter(state == group_2) |>
12   magrittr::extract2("msf_sc")
13
14 stats_sum_group_2 <-
15   data |>
16   dplyr::filter(state == group_2) |>
17   magrittr::extract2("msf_sc") |>
18   stats_sum(print = FALSE)
19
20 stats_sum_group_2 |> list_as_tibble()

```

name	value
n	4097
n_rm_na	4097
n_na	0
mean	103651.485407441
var	26397872.5737977
sd	5137.88600241361
min	88092.8571428571
q_1	99857.1428571429
median	103328.571428571
q_3	107057.142857143
max	117064.285714286
iqr	7200
skewness	0.263604552812467
kurtosis	2.69168212740063

```

1 # library(stats)
2
3 stats::t.test(msf_sc_group_1, msf_sc_group_2, alternative = "less")
4 #>
5 #> Welch Two Sample t-test
6 #>
7 #> data: msf_sc_group_1 and msf_sc_group_2
8 #> t = -0.17, df = 118, p-value = 0.43
9 #> alternative hypothesis: true difference in means is less than 0
10 #> 95 percent confidence interval:
11 #>      -Inf 759.34
12 #> sample estimates:
13 #> mean of x mean of y
14 #>      103564      103651

```

```

1 # library(dplyr)
2
3 dplyr::tibble(
4   name = c("mean_group_1", "mean_group_2", "diff"),
5   value = c(
6     stats_sum_group_1$mean, stats_sum_group_2$mean,
7     stats_sum_group_2$mean - stats_sum_group_1$mean
8   )
9 )

```

name	value
mean_group_1	103564.48
mean_group_2	103651.49
diff	87.01

See Frey (2022, 224–27) to learn more.

$$d = \frac{\mu_1 - \mu_2}{\sigma_e}$$

```
1 # library(effsize)
2
3 effsize::cohen.d(msf_sc_group_1, msf_sc_group_2)
4 #>
5 #> Cohen's d
6 #>
7 #> d estimate: -0.016915 (negligible)
8 #> 95 percent confidence interval:
9 #>      lower      upper
10 #> -0.20387  0.17004
```