

Universidade Federal do Estado do Rio de Janeiro  
Departamento de Informática Aplicada  
Programa de Pós-Graduação em Informática

ANALYTICAL STUDY OF NEWBORN DATA IN BRAZIL – NEW CURVES

Bruno Cesar Barreiros Alves Pereira

Rio de Janeiro  
2025

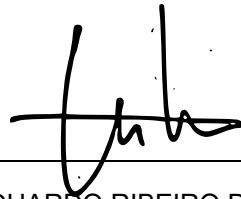
ANALYTICAL STUDY OF NEWBORN DATA IN BRAZIL – NEW CURVES  
Bruno Cesar Barreiros Alves Pereira

ANALYTICAL STUDY OF NEWBORN DATA IN BRAZIL – NEW CURVES

Bruno Cesar Barreiros Alves Pereira

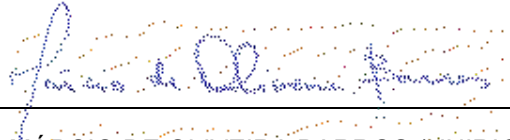
DISSERTAÇÃO DE MESTRADO APRESENTADA COMO REQUISITO OBRIGATÓRIO ESTIPULADO PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA (PPGI) DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada Por:



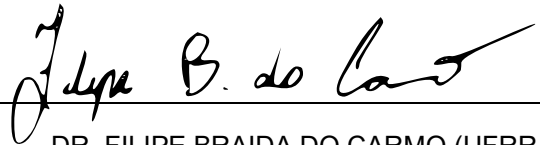
---

DR. CARLOS EDUARDO RIBEIRO DE MELLO (UNIRIO)



---

DR. MÁRCIO DE OLIVEIRA BARROS (UNIRIO)



---

DR. FILIPE BRAIDA DO CARMO (UFRRJ)

RIO DE JANEIRO, RJ – BRASIL

Catálogo informatizada pelo autor

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| P436 | <p>Pereira, Bruno Cesar Barreiros Alves<br/>Estudo Analítico de dados de nascidos vivos no Brasil -<br/>Novas Curvas / Bruno Cesar Barreiros Alves Pereira. -- Rio<br/>de Janeiro : UNIRIO, 2025.<br/>54f</p> <p>Orientador: Carlos Eduardo Ribeiro de Mello.<br/>Dissertação (Mestrado) - Universidade Federal do Estado<br/>do Rio de Janeiro, Programa de Pós-Graduação em Informática,<br/>2025.</p> <p>1. Análise de peso ao nascer . 2. Curvas percentis de<br/>peso ao nascer. I. de Mello, Carlos Eduardo Ribeiro,<br/>orient. II. Título.</p> |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Acknowledgements

To Grace Hopper, a visionary pioneer that is considered to be the mother of computing, software development, and design. She taught computers to *speak* human languages.

## **RESUMO**

Este estudo tem o objetivo de analisar o perfil mais recente dos recém-nascidos no Brasil, baseado em três importantes variáveis clínicas: peso ao nascer, idade gestacional e índice APGAR. Esta análise proporciona novos entendimentos e gera subsídios para políticas de saúde pública. 22,6 milhões de registros de nascimento ocorridos entre os anos de 2011 e 2019 foram extraídos das bases de dados do SINASC. Uma análise descritiva foi conduzida: ela identificou que os recém-nascidos estão ficando ligeiramente mais pesados no período, um perfil de "mãe média brasileira" foi identificado, uma tabela contendo as curvas de idade gestacional e percentis de peso foi apresentada como um dos resultados. Este estudo contribui ao oferecer novas perspectivas sobre as condições de nascimento que podem servir de base para futuras análises preditivas e orientar políticas de cuidado neonatal.

**Palavras-chave:** Peso ao nascer, Idade Gestacional, recém-nascido, crescimento fetal, tabela de crescimento.

## **ABSTRACT**

This study aims to analyze the profile of recent newborns from Brazil based on three fundamental clinical variables: birthweight, gestational age, and APGAR scores. This analysis offers new understandings and subsidies to public health policies. 22.6 million newborn registries from 2011 to 2019 were extracted from SINASC databases. A descriptive analysis is conducted; it identified that newborns in Brazil are getting slightly heavier over the period; this analysis also identified an average mother profile; a table containing gestational age and weight percentiles is also presented as a result. This study contributes by offering new panoramas regarding the conditions of birth, which can support future predictive analysis and orient neonatal care policies.

**Keywords:** Birth Weight, Gestational Age, Infant, Newborn, Fetal Growth, Growth Charts.

|                                                                         |           |
|-------------------------------------------------------------------------|-----------|
| <b>1 Context and motivation .....</b>                                   | <b>12</b> |
| <b>1.1 Research Goals.....</b>                                          | <b>14</b> |
| <b>2 Literature Review.....</b>                                         | <b>16</b> |
| <b>3 Materials .....</b>                                                | <b>20</b> |
| <b>3.1 The SINASC Database .....</b>                                    | <b>20</b> |
| <b>4 Methodology .....</b>                                              | <b>24</b> |
| <b>5 Statistical and Exploratory Analysis, and Discussions .....</b>    | <b>33</b> |
| <b>5.1 Statistical Analysis.....</b>                                    | <b>33</b> |
| <b>5.2 Exploratory Analysis and discussion .....</b>                    | <b>34</b> |
| <b>5.2 Updated Gestational Age to birthweight percentile table.....</b> | <b>48</b> |
| <b>6 Conclusion.....</b>                                                | <b>52</b> |
| <b>6.1 Final Considerations .....</b>                                   | <b>52</b> |
| <b>6.2 Risks and Limitations.....</b>                                   | <b>52</b> |
| <b>6.3 Suggestions for future research.....</b>                         | <b>53</b> |
| <b>7 References .....</b>                                               | <b>54</b> |

## List of Figures

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| Figure 1: Raw weight distribution from the 22 <sup>nd</sup> to 27 <sup>th</sup> gestational week. .... | 22 |
| Figure 2: Raw weight distribution by the Gestational Age estimation method. ....                       | 23 |
| Figure 3: Orange Data Miner interface and the data cleansing workflow .....                            | 24 |
| Figure 4: Diagram organizing every step of this analysis.....                                          | 25 |
| Figure 5: Shows the weight limit analysis process .....                                                | 30 |
| Figure 6: Raw weight distribution for the 40th gestational week .....                                  | 35 |
| Figure 7: Evolution of birth quantity by gestational week group in percentage.....                     | 38 |
| Figure 8 Evolution of the average mother's age .....                                                   | 39 |
| Figure 9: Evolution of the APGAR1 and APGAR5 averages .....                                            | 39 |
| Figure 10: Proportions of delivery type by mothers' race.....                                          | 41 |
| Figure 11: Average weight evolution grouped by Brazilian region.....                                   | 42 |
| Figure 12: Average mother's age grouped by Brazilian region.....                                       | 42 |
| Figure 13: Average weight, grouped by mothers' age group and skin color .....                          | 43 |
| Figure 14: Number of babies delivered by mothers' age group and skin color.....                        | 44 |
| Figure 15: Percentual of births by mother's race and education .....                                   | 45 |
| Figure 16: Average weight by region and mothers' age group over the period.....                        | 45 |
| Figure 17: Overview of babies born grouped by region and mothers' skin color.....                      | 46 |
| Figure 18: Evolution of birthweight by mother's profile.....                                           | 47 |

## List of Tables

|                                                                                                                              |    |
|------------------------------------------------------------------------------------------------------------------------------|----|
| Table 1: APGAR (1953) and APGAR et AL (1958) numbers of birth by type of delivery.....                                       | 17 |
| Table 2: Number of deliveries and neonatal deaths by APGAR1 from APGAR et al (1958).....                                     | 18 |
| Table 3: Grouping APGAR et al (1958) deliveries using APGAR (1953) classification and delivery type.....                     | 19 |
| Table 4: 19 fields from SINASC.....                                                                                          | 26 |
| Table 5: New fields that were created.....                                                                                   | 28 |
| Table 6: Weight filters applied by PEDREIRA et al. (2011).....                                                               | 29 |
| Table 7: New weight filter for gestational age.....                                                                          | 29 |
| Table 8: Evolution of registries after each step of data cleanse.....                                                        | 30 |
| Table 9: A database comprising 22.7 million newborn records has been established after the application of all filters .....  | 31 |
| Table 10: Central Values and Interval Widths used by Pedreira et al.(2011) to calculate their interpolated percentiles.....  | 31 |
| Table 11: A selected group of gestational ages and their respective raw weight percentiles before filtering corrections..... | 34 |
| Table 12: The same gestational ages, but after the filtering process .....                                                   | 35 |
| Table 13: Addition of a delta to Table 9 .....                                                                               | 36 |
| Table 14: Average APGAR1 score in Brazil by delivery type .....                                                              | 40 |
| Table 15: Number of deliveries by APGAR1 group and delivery type.....                                                        | 40 |
| Table 16: Weight chart using the same filters as PEDREIRA et al. (2011) .....                                                | 48 |
| Table 17: Updated birthweight to gestational age, singleton Brazilians from 2011 to 2019.....                                | 50 |
| Table 18: Comparison between PEDREIRA et al. (2011) and the current study .....                                              | 51 |

## List of Equations

|                                                              |    |
|--------------------------------------------------------------|----|
| Equation 1: Formula for a Z-test with two samples.....       | 33 |
| Equation 2: Relative percentual difference calculation ..... | 51 |

## 1 Context and motivation

Gestational age, birthweight, and APGAR scores are important indicators of a newborn's health condition. In common, they can only be measured at the delivery moment. Nevertheless, those indicators are closely related and well documented signals for the chances of neonatal deaths. Additionally, both low and high birthweight values can also be used as important estimators for the possibility of type 1 and 2 diabetes later in adulthood.

The APGAR score is named after Virginia Apgar; a north American physician specialized in obstetrics and anesthesiology considered a pioneer of neonatology studies. At first, there were several scattered evaluations being used such as breathing time, crying time, or state of depression. APGAR (1953) proposed a score to determine a conditions grade for the infant at birth.

This score is based on five basic health indicators: heart rate, respiratory effort, reflex irritability, muscle tone, and color. Each of those indicators would receive a rating from 0 to 2, then when all the individual ratings are summed up, a number ranging from 0 to 10 (highest possible score) is obtained. The first measurement of those five indicators occurs exactly 60 seconds (or one minute) after the complete birth of the baby, then a second measurement occurs at the 5<sup>th</sup> minute; they are named APGAR1 and APGAR5, respectively.

In the same article, APGAR (1953) categorizes the APGAR scores into 3 groups: 0, 1, and 2 scores are infants considered to be in 'poor condition'; scores of 3, 4, 5, 6, and 7 result in a 'fair condition'; finally, scores of 8, 9 and 10 are indicative of a 'good condition'. In that study, a newborn in the poor condition group had 14 percent of chance of death, while newborns in the fair and good conditions had respectively, 1.1 percent and 0.13 percent of a chance of deaths. "The score was especially useful in judging the need for resuscitative measures, such as respiratory assistance." (APGAR et al., 1958)

Neonatal death is defined by the World Health Organization (WHO) as "...deaths among live births during the first 28 completed days of life." (WORLD HEALTH ORGANIZATION, 2006). This period can be further sub-divided into early neonatal deaths (deaths on or before the 7<sup>th</sup> day after birth) and late neonatal deaths (after the 7<sup>th</sup> day until 28<sup>th</sup> days after birth) (UNICEF, 2014). PATHIRANA et al. (2016) points out that more than 40% of children's deaths under the age of 5 years occur during the first month of life. Despite the recent advancements in Medicine and neonatal care practices, high neonatal death rates persist in some regions and populations. This supposed contradiction indicates a necessity for deeper comprehension of the factors that contribute to this phenomenon.

Challenges associated with neonatal mortality have many facets, involving biological, cultural, socioeconomical factors and access to health services. PATHIRANA et al. (2016) mentions that the leading causes of neonatal death in 2013 were associated with: preterm birth complications, intrapartum related complications, neonatal sepsis, and other neonatal infections. As such, effectively identifying and understanding that those factors require a non-reductionist approach, beyond a traditional analysis by making use of advanced methodologies like Data Science.

HARDER et al (2009) performed a systematic review and meta-analysis aiming to quantitatively summarize evidence of prenatal risk factors that might have an influence on the risk of diabetes. They analyzed twelve studies, most of them based on European populations, and they analyzed two groups: birthweight value higher than 4000g, and birthweight lower than 2500g. The first and heavier birthweight group was associated with type 1 diabetes, each 1000g increase in birthweight resulted in 7% more chances of developing in diabetes later in life. Furthermore, the authors argue that the 2500g group had a non-significantly decreased risk of type 1 diabetes, but they added that a rapid weight-gain during the first year for those babies in the same lighter group can also precede type 1 diabetes.

LUBCHENCO et al. (1963) is one of the very first studies to propose reference data associating birthweight and gestational age. The authors constructed a chart describing the median weights for Caucasian infants born in one city in the state of Colorado, United States. After that, other more recent studies, in the nineties and early 2000's also assessed those parameters and proposed birthweight versus gestational age charts, however, most of them had in common the fact of being based on developed countries' populations.

"Most developed countries have created birthweight for Gestational Age reference charts that are based on cross sectional studies using registry data from their official population records" (TSAGKARI et al., 2020). Most developing countries lacked a national parameters table until very recently.

As a result, pediatricians in developing countries do not have access to extensive databases when comparing measured fetal or newborn weight with a reference table based on a distinct foreign population. As a result, a specific fetus or newborn measured by a non-representative reference weight table may lead to an incorrect medical decision, such as rushing into premature birth. Brazil was included among the nations in this situation, where until recently most clinical evaluations were being based on foreign data, as mentioned by PEDREIRA et al. (2011).

PEDREIRA et al. (2011) proposed a pioneer contribution with the first comprehensive charts with the ethnic-mixed population of Brazil. In their study, they analyzed and aggregated data of almost eight million Brazilians with birthdates ranging from 2003 to 2005. This study presented tables and graphics that provide a gender-specific reference of birthweight distribution, gestational age and the 3rd, 5th, 10th, 25th, 50th, 90th, 95th and 97th percentile curves for births between the 22 and 43 weeks.

## 1.1 Research Goals

One of the motivations for this dissertation is to update PEDREIRA et al. (2011)'s study with more recent data. This research employs the same data source derived from the Brazilian system: "Sistema de Informação sobre nascidos vivos" (SINASC, or Born Alive Information System, in a free translation). However, data has been collected from a much larger and recent interval of 9 years ranging between 2011 and 2019. The following years of 2020 and later were excluded from this study because of the COVID pandemic and its possible effects on the health systems.

The main goal of this study an exploratory analysis of information leading to pertinent discussions which results in the development of a reliable average profile of a typical Brazilian mother. Then, this analysis culminates in the construction a more recent reference of Brazilian newborn conditions in the form of a table demonstrating birthweight per gestational week. Such an updated reference table can be used by maternity hospitals, pediatricians, and the general health-care facilities to prepare and organize the eventual need for special attention of newborns considering their position on the percentile curve.

This updated table can also be used to assess the evolution of birthweights by the year and to understand if and when social programs like *Bolsa Família* (Family Allowance Benefit, in a free translation) - a social welfare and revenue transfer program created in 2003 that unified other smaller social aid programs that contributed to the reduction of poverty in Brazil; and *Mais Médicos* (More Doctors, in a free translation) - a program that lasted from 2013 until 2018 that was created to resolve the shortage of doctors in small municipalities and the fringes of large cities in Brazil – both programs were improving newborn health conditions based on their birthweight as an indicator.

The scope of this research concentrates on birthweights in the Brazilian national context focusing on applying advanced analytics techniques to comprehend the associated factors and to develop specific predictive models for the Brazilian reality. The research is performed considering demographic, socioeconomics and maternal health factors that affect neonatal health inside of Brazilian national frontiers.

This study proposes five different deliverable artifacts. The first is the detailed description of SINASC's dataset that was performed in the beginning of this chapter; the second is the reproduction of PEDREIRA et al. (2011)'s work generating an updated birthweight for gestational age chart; then a comparison between PEDREIRA et al. (2011)'s and this study's chart; in forth, a general analysis of the database aiming to identify if there are relevant differences in birthweight between demographical groups.

After that exploration and discussion of the evolution of births in a 10-year timespan. The following research questions can be asked:

RQ1: How does the birthweight develop during the investigated period?

a. Is there a population group that evolved differently?

The general research plan involves the following stages:

- 1) Bibliographic review: Literature analysis concerning birthweight in national and international contexts, including previous studies and relevant interventions.
- 2) Data collection: Identification and acquisition of nationwide databases that include demographic, medical history and socioeconomic conditions related to newborns.
- 3) Data preprocessing: Cleaning and organizing data while maintaining its integrity and consistency.
- 4) Descriptive analysis: Descriptive statistical analysis to understand variable distribution and correlation.
- 5) Development of an updated Gestational week percentile table
- 6) Interpretation and resulting discussions: Deepened result discussions, considering practical applications within national public health systems.
- 7) Final report: Compilation of findings in an in-depth report, including specific recommendations for enhanced health practices and national policies.
- 8) Scientific divulgation and presentation: Result presentation in national and international events, submission to national specialized periodicals.

## 2 Literature Review

PEDREIRA et al. (2011) collected data from SINASC live-birth certificates from all Brazilian states ranging from 2003 to 2005. That research applies data selection filters by considering the following factors: newborns who were only singletons, held no major congenital anomalies; and had gestational age between 22 and 42 weeks. Other records with missing data for major congenital anomalies, gestational age or birthweight were disregarded.

After the exclusions, almost 8 million newborn records were identified, subject to analysis, resulting in “more than double the sample size of the largest early studies” (PEDREIRA et al., 2011). Yet, it is still important to reiterate that all those newborns are Brazilian and that they purportedly did not filter the different races that form Brazil’s population, thus resulting in a birthweight-to-gestational age chart for the heterogeneous population of a developing country.

From 1996 until 2009, SINASC only offered the gestational age in discrete ranges: less than 22 weeks, 22 to 27 weeks, 28 to 31 weeks, 32 to 36 weeks, 37 to 41 weeks, and more than 41 weeks. PEDREIRA et al. (2011) had to apply a ‘shape-preserving piecewise cubic interpolation method’ to estimate the birthweight growth curves by week since they analyzed the 2003 to 2005 period.

From 2010 on, SINASC offers a numeric value for gestational weeks(age). This updated database also provides the last menstruation date, which can be used jointly with the birthdate to check the gestational age. Both the provided and calculated gestational age can be considered rather than interpolation curves for each week. This dissertation uses the provided gestational age in its analysis.

PEDREIRA et al. (2011) compares their weight to gestational week table for the Brazilian population with other four other international studies: showing that the Brazilian weight-per-week data was majorly lower than ALEXANDER et al. (1996)’s, regardless of percentile. When comparing with KRAMER et al. (2001) and BONELLIE et al. (2008) it is shown that Brazilian birthweights were lower in the 10<sup>th</sup> percentile, but higher on the 50<sup>th</sup> and 90<sup>th</sup> (except for the births later the 37 gestational weeks on). When comparing with ALSHIMMIRI et al. (2004), Brazilian birthweights were almost equivalent to the lower 10<sup>th</sup> percentile, but much lower on the 50<sup>th</sup> percentile then higher on the 90<sup>th</sup> percentile. This behavior might show a different flattening in the curves. In other words, an average Brazilian newborn weighs less than the average Kuwaiti newborn, and on the other hand, Brazilian higher-weight newborns are heavier than the heaviest newborns recorded in Kuwait.

It is interesting to note that the Brazilian chart produced lower percentiles for the majority of the gestational weeks when compared to ALEXANDER et al. (1996). Furthermore, the 50th percentile is generally higher for the Brazilian data at exceptionally low gestational ages (up to 26th weeks) and lower at term. This may be an indication of poorer survival rates at lower gestations in Brazil.” (PEDREIRA et

al., 2011) explains the different curves and speculates a probable reason for the higher birthweights at lower gestational weeks.

It is also important to remember that ALEXANDER et al. (1996) used data from the United States population, ALSHIMMIRI et al. (2004) considered the Kuwaiti population, BONELLIE et al. (2008) was based on Scottish singleton births, and KRAMER et al. (2001) used Canadian data. All the four aforementioned countries are developed countries, unlike Brazil.

No Brazilian updates to PEDREIRA et al. (2011) were found by the time of writing this study, but other four international articles cited them and generated birthweight versus gestational age charts for their respective countries. Both DAI et al. (2014) and YAO et al. (2018) developed new charts based on the contemporary Chinese population; TSAGKARI et al. (2020) produced a preprint study with an outcome of percentile charts for neonates in Greece; and SALEH GARGARI et al. (2017) proposed reference weight values for the Iranian population.

All the aforementioned studies applied similar filters, considering only singleton pregnancies and excluding cases with detected congenital anomalies.

APGAR (1953) applied the APGAR evaluation in 1021 infants born in the Sloane Hospital for Women, located in New York, in the United States, during her studies in the early fifties. The babies with best APGAR scores were the ones delivered spontaneously or low forceps delivery. This group also represented most deliveries.

APGAR et al. (1958) furthered their previous studies upon evaluating more than fifteen thousand infants delivered in the same North American hospital during the period between 1952 and 1956. Table 1 is a replication of a table from both studies, it shows the number of deliveries, their mean APGAR1 score (1953). This study added a column for a Percent of total deliveries. No information was provided for ‘Midforceps delivery’ neither for ‘Version and breech extraction’, in the 1958 study.

*Table 1: APGAR (1953) and APGAR et AL (1958) numbers of birth by type of delivery*

| <b>Delivery type</b>          | <b>Number of infants (1953)</b> | <b>Mean APGAR1 Score (1953)</b> | <b>Percent of total (1953)</b> | <b>Number of infants (1958)</b> | <b>Percent of total (1958)</b> |
|-------------------------------|---------------------------------|---------------------------------|--------------------------------|---------------------------------|--------------------------------|
| Low forceps or spontaneous    | 843                             | 8.4                             | 82.57%                         | 13248                           | 86.32%                         |
| Cesarean section              | 141                             | 6.8                             | 13.81%                         | 1582                            | 10.31%                         |
| Midforceps delivery           | 17                              | 6.9                             | 1.67%                          | n/a                             | n/a                            |
| Breech delivery               | 16                              | 6.7                             | 1.57%                          | 518                             | 3.38%                          |
| Version and Breech extraction | 4                               | 6.3                             | 0.39%                          | n/a                             | n/a                            |

It is possible to notice that spontaneous deliveries represent the vast majority of births, representing 86.32% of deliveries in the studies period; followed by cesarean section, but with only 10.31%.

APGAR (1953) declared that “It did not seem to be of value to separate the spontaneous deliveries from those in which low forceps were used”. So, the group ‘Low forceps or spontaneous’ is used as equivalent to vaginal delivery in comparisons further in this study.

Table 2 is a transcript from a figure from that APGAR et al. (1958). It shows the distribution of scores of newborns in all types of deliveries, the number of neonatal deaths, and the death rate for deliveries with the respective APGAR1 score.

*Table 2: Number of deliveries and neonatal deaths by APGAR1 from APGAR et al (1958)*

| APGAR1 | Spontaneous delivery |        |         | Cesarean section |        |         | Breech delivery |        |         |
|--------|----------------------|--------|---------|------------------|--------|---------|-----------------|--------|---------|
|        | Deliveries           | Deaths | Death % | Deliveries       | Deaths | Death % | Deliveries      | Deaths | Death % |
| 0      | 35                   | 9      | 25.71%  | 9                | 3      | 33.33%  | 8               | 3      | 37.50%  |
| 1      | 219                  | 35     | 15.98%  | 89               | 18     | 20.22%  | 58              | 24     | 41.38%  |
| 2      | 299                  | 19     | 6.35%   | 60               | 0      | 0.00%   | 45              | 14     | 31.11%  |
| 3      | 322                  | 8      | 2.48%   | 49               | 1      | 2.04%   | 24              | 3      | 12.50%  |
| 4      | 387                  | 10     | 2.58%   | 58               | 3      | 5.17%   | 30              | 2      | 6.67%   |
| 5      | 529                  | 8      | 1.51%   | 52               | 2      | 3.85%   | 48              | 3      | 6.25%   |
| 6      | 760                  | 8      | 1.05%   | 96               | 3      | 3.13%   | 59              | 2      | 3.39%   |
| 7      | 1406                 | 10     | 0.71%   | 195              | 2      | 1.03%   | 63              | 0      | 0%      |
| 8      | 3195                 | 12     | 0.38%   | 443              | 4      | 0.90%   | 98              | 2      | 2.4%    |
| 9      | 4019                 | 16     | 0.40%   | 397              | 3      | 0.76%   | 72              | 0      | 0%      |
| 10     | 2077                 | 2      | 0.10%   | 134              | 1      | 0.75%   | 13              | 0      | 0%      |

Based on this table, it is clear that there is a relation between low APGAR1 scores and a higher number of neonatal deaths, especially for breech deliveries.

Table 3 is forged by making use of APGAR (1953) classification of APGAR scores and the numbers from APGAR et al (1958) study.

*Table 3: Grouping APGAR et al (1958) deliveries using APGAR (1953) classification and delivery type.*

| <b>APGAR1 Group</b> | <b>Infant Number</b> | <b>% of total</b> | <b>Delivery type</b> | <b>Infant Number</b> | <b>% inside group</b> | <b>% of total</b> |
|---------------------|----------------------|-------------------|----------------------|----------------------|-----------------------|-------------------|
| Poor condition      | 822                  | 5.36%             | Spontaneous          | 553                  | 67.27%                | 3.60%             |
|                     |                      |                   | Cesarian             | 158                  | 19.22%                | 1.03%             |
|                     |                      |                   | Breech               | 111                  | 13.50%                | 0.72%             |
| Fair condition      | 4078                 | 26.57%            | Spontaneous          | 3404                 | 83.47%                | 22.18%            |
|                     |                      |                   | Cesarian             | 450                  | 11.03%                | 2.93%             |
|                     |                      |                   | Breech               | 224                  | 5.49%                 | 1.46%             |
| Good condition      | 10448                | 68.07%            | Spontaneous          | 9291                 | 88.93%                | 60.54%            |
|                     |                      |                   | Cesarian             | 974                  | 9.32%                 | 6.35%             |
|                     |                      |                   | Breech               | 183                  | 1.75%                 | 1.19%             |

From this table we see that spontaneous deliveries in good condition represent roughly 60% of deliveries; followed by spontaneous in fair condition with a bit over 22.28% of the total. The ‘poor condition’ group, which represents the higher risk of neonatal death, totals only 5.36% of all births. Among that group, 67% were spontaneous, and 19% were cesarian.

It is also possible to observe that the proportion of delivery types varies among the condition groups. While spontaneous deliveries are the preferred method among all score groups, the cesarean and breech percentual rises as the evaluation decreases. A possible cause-effect relation is that pre-natal evaluations determine that simpler or standard gestations tend to be directed to spontaneous deliveries while more complex situations, like the different feet-head orientations, might be advised to perform cesarian or breech deliveries.

### 3 Materials

This research study is developed using statistical, sociodemographic, and epidemiologic data derived from the Brazilian database, the SINASC. This is a public database made available by the Brazilian Health Ministry.

#### 3.1 The SINASC Database

Before the adoption of SINASC, the only information source regarding births came from the Civil Registry Statistic supplied by IBGE (*Instituto Brasileiro de Geografia e Estatística*, or Brazilian Geography and Statistics Institute in a free translation) (SZWARCWALD et al., 2002). But the data provided only contained standard notary office registries regarding legal proof for the birth event, but with no information more relevant to the health sector such as the childbirth conditions (SZWARCWALD, C. L. et al., 2019).

Brazilian Health Ministry implanted SINASC in 1990. It is a system that uses the *declaração de nascido vivo* (newborn declaration document, in a free translation) as a data source. This document is obligatory, issued by the health facility where the birth took place. Unlike Civil Registry documents, SINASC aims at characterizing the conditions of pregnancy, labor and birth (SZWARCWALD, C. L. et al., 2019).

The SINASC data structure has evolved since its implementation. The first version included information regarding the registry office; place of birth; mother's age, marital status, level of education, number of previous children, number of fetal losses/abortions, and city of residency; pregnancy duration in ranges (less than 22 weeks, 22 to 27 weeks, 28 to 31 weeks, 32 to 36 weeks, 37 to 41 weeks, and more than 41 weeks), multiplicity, type of deliver, and number of prenatal appointments; newborn's birth date, gender, birthweight, APGAR scores, skin color, weight, and a code for anomaly (if any).

In 1997 SINASC went through a small update in which the new registries now have two more fields: one for hour of birth and a flag informing if an anomaly was found. Since 2000 two more fields have been made available: the code of the health establishment where the birth took place, and a code for the State that informed the registry. Then, in 2006, two more fields: date of newborn declaration document, and a date for systematic batch receipt.

The year 2010 came with a revamped version of SINASC. The total number of fields more than doubled from the previous 27 to 56. The new fields regarded notarial bureaucracy: Type of Database Source, Registry Office Code, Registry Office registration number, date of registry, batch number, system version, batch date, and a calculated difference from birthdate and the batch date; new information about the mother: code of country of residence, nationality code for foreign mothers, code for

mother's city of birth, last school grade, birthdate, race/color, number of previous gestations, number of previous vaginal deliveries, and number of previous c-sections; new fields regarding the pregnancy: last menstruation date, gestational age in weeks, gestational age estimation method, number of pre-natal consultations, month of pre-natal initiation; delivery's presentation type (cephalic, pelvic or podalic) , a flag for induced labor, a flag for c-section before labor start, Robson group code, and two flags for death certificate; and at last: the father's birthdate.

In 2011, the system underwent a small update with a new race/color field for the newborn. Then, in 2012 there is another field for newborn's race/color, and an updated mother's education level that describes her highest completed status.

The penultimate update was in 2013 when the database received: a field a city code for the Registry Office; a Federative Unit (State) code for the newborn; a code for delivery assistance type; and yet another mother's more detailed education level.

Finally, the last update of 2014: an added registration date; the role of the professional that filled the document; a document from the professional; a new declaration date; a flag for first pregnancy; and the Kotelchuck pre-natal index.

Also in 2014, and only in that year, a discrepancy was found: the newborn gender was classified as M for male, and F for female. For all the other years before and after 2014 the gender was classified with the number 1 for male, and 2 for female.

Based on this history of changes, it can be understood that continuous efforts have been made to improve information and data quality through different Health Ministries and Governments. In special, we can notice a focus on new parturient and the gestation attributes that were added.

Mother's age field used to be manually fulfilled, but now it is the calculated difference between the mother's and newborn's birthdate; mother's education changed from total years of education to the highest completed level; mother's marital status was updated to reflect the new "*união estável*" (or stable union, in a free translation) that emerged with the new Brazilian Civil Code; mother's and newborn's race/color were adapted to capture eventual inequalities in access to health systems that could be determined due to differences in race/color; mother's fertility calculations were updated from the quantities for born alive or born dead to number of previous gestations, number of vaginal deliveries, number of c-sections, and number of fetal losses/abortions.

Another important update was the gestational age. The old system only collected the gestational ages in groups intervals, but since 2011 it has been a field calculated based on the difference between the mother's last menstruation date and the delivery date; and if the mother's last menstruation date is unknown, then a different estimation method can be applied and flagged. Finally, the number of pre-natal consultations were another piece of information that went from group intervals to an integer number.

Those updates allow for more detailed and accurate analysis regarding the mother's socioeconomic factors. For instance, it is now possible to compare the weights of

children delivered in the 38<sup>th</sup> week by a dark/mixed skinned, with stable union relationship and elementary education mother that lives in the north/northeastern regions against the remaining group of newborns.

Upon starting the data collection, an important difficulty to overcome is that several registries are clearly incorrect: some remarkably high birthweights were identified upon examining births of the 22<sup>nd</sup> week. While PEDREIRA et al. (2011) determines a 97<sup>th</sup> percentile of 926 grams for male and 903 grams for female, among the 7915 22<sup>nd</sup> weeks in our more recent data there are 1761 newborns weighing more than 1200 grams, which is clearly a data annotation error.

Figure 1 shows the raw weight distribution for the first week group, which starts on the 22<sup>nd</sup> week and ends on 27<sup>th</sup>. It is noticeable that those raw numbers seem to be in a bimodal distribution, which is an unexpected behavior that can only be explained by either weight and/or gestational age erroneous annotations/typing mistakes.

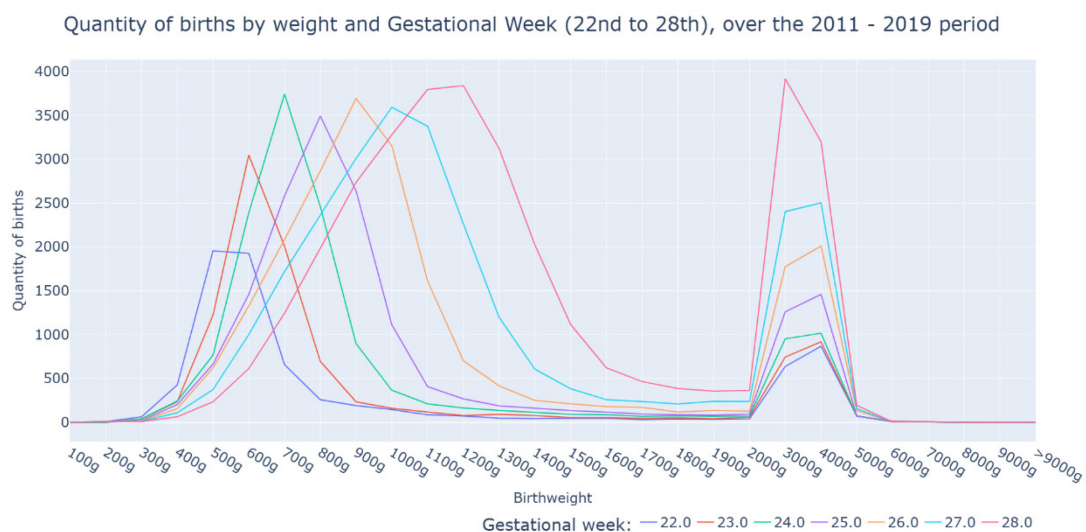


*Figure 1: Raw weight distribution from the 22<sup>nd</sup> to 27<sup>th</sup> gestational week.*

BONELLIE et al. (2008) excluded outliers weighing less than 250g. They used a method that assumed symmetric birthweight distribution. “The number of birthweights omitted as possible outliers was small and inspection of the omitted birthweights suggests that most of these could be explained by transcription errors.”

KRAMER et al. (2001) observed the same behavior on the Canadian databases, which they classified as on term births wrongly registered as low gestational weeks. They also assume that the birth weight for gestational age follows the Normal distribution. “We also assume that the observed gestational ages are a mixture of correct values and of term (40-week) gestations erroneously measured at the observed (recorded) gestational age. This latter assumption is consistent with the frequently observed curve “bump” at the upper percentiles for extremely preterm infants”.

In the same study, they applied an 'estimation-maximization algorithm' to estimate the likelihood of means, Standard Deviation, and the probability of an observed (and registered) gestational age of an individual newborn to be an on term (40 weeks). Then they resampled their observations and decided to either maintain or delete a register based on the estimated probability of the true gestational week be equal to 40.

Another kind of difficulty was identified when examining the incorrect birthweight values. There is a field named TPMETESTIM, which determines the method that was applied to estimate the gestational age. This field's dictionary consists of only three possible values: 1, *exame físico*; 2, *outro método*; and 9, *ignorado* (in a free translation, respectively: physical exam; other method; and not known).

But there are several registries where TPMETESTIM equals to 8, which is an undocumented value for this field. Upon further examining the field called DTULTMENST, that stands for last menstruation date and realizing that it is only filled when TPMETESTIM equals to 8, it can be inferred that the code 8 method must mean that the gestational age calculation method used was the difference between the birthdate and the last menstruation date.

Figure 2 shows the numbers of births in the 22<sup>nd</sup> week grouped by the Gestational Age estimation method. The four methods show the same modes, but 'Last menstruation date' and 'not known' seem to have more registries accumulated on the second and smaller local peak.

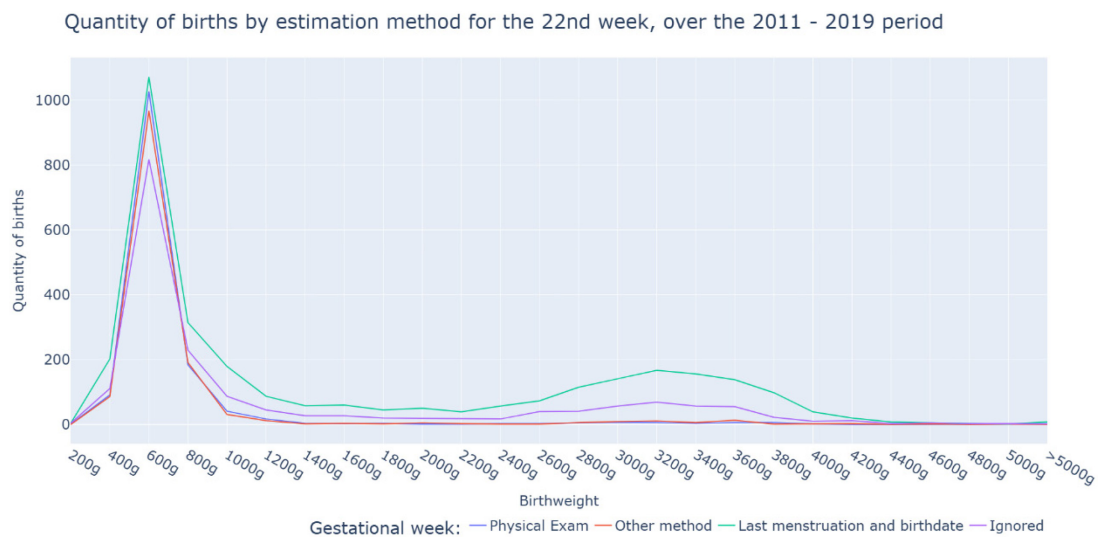


Figure 2: Raw weight distribution by the Gestational Age estimation method.

## 4 Methodology

The software used in this analysis is Orange Data Mining, an open-source data mining, data visualization and machine learning visual toolkit developed by a Slovenian group of researchers. They describe it as “a machine learning and data mining suite for data analysis through Python scripting and visual programming” (DEMŠAR et al., 2013). Out of the box Orange provides several widgets, which are visual components that facilitate data manipulation, visualization, classification, machine learning regressions algorithms, prediction methods, and unsupervised algorithms for clustering and data projection.

This project started in 1996 in the University of Ljubljana, the largest and oldest University of Slovenia. Its core was developed with C++ and the modules are developed in Python. Being an open-source project, it is also composed of an active community that develops more specific add-ons for different areas of science, such as, but not limited to: Bioinformatics, Education, Image Analytics, Network Analysis, and text mining.

Figure 3 is a capture of Orange Data Mining running on Windows 10 with an open ows file. On the left there is a toolbox with all the available widgets, and on the center, once can view the data cleansing workflow.

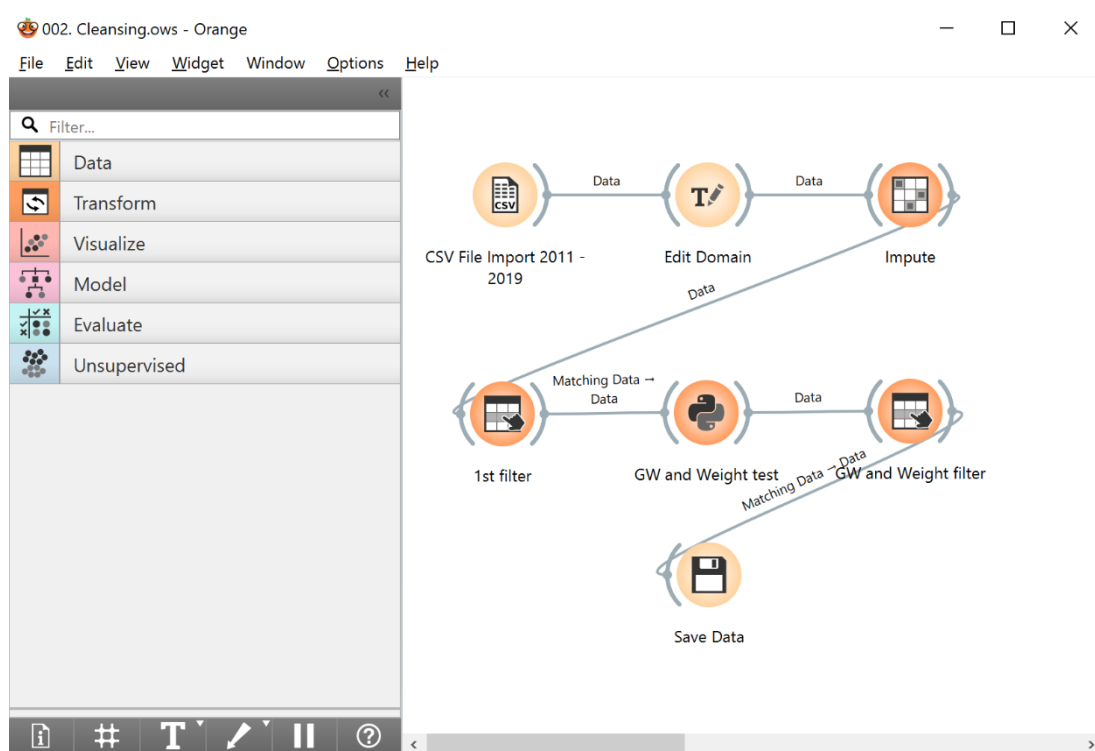


Figure 3: Orange Data Miner interface and the data cleansing workflow

This analysis also employs data visualization libraries provided by Plotly, a Canadian company founded in 2013. This open-source library offers tools to create many diverse types of graphs, and it is compatible with Python, JavaScript, R, and Java.

The methodology of this study is outlined in Figure 4. It resumes the seven substeps of data acquisition, preprocessing, and analysis: 9 yearly csv files were extracted from SINASC's website; only 19 relevant fields were selected to reduce performance bottlenecks; the data dictionary was used to translate the database codes into their meanings; a data cleansing process of imputing a standard value for missing data; registers with missing data, multiple pregnancies (twins or more), or newborn with anomalies as well as registries considered misclassified as pre-term were discarded; then the percentiles table and a descriptive analyses were conducted in parallel.

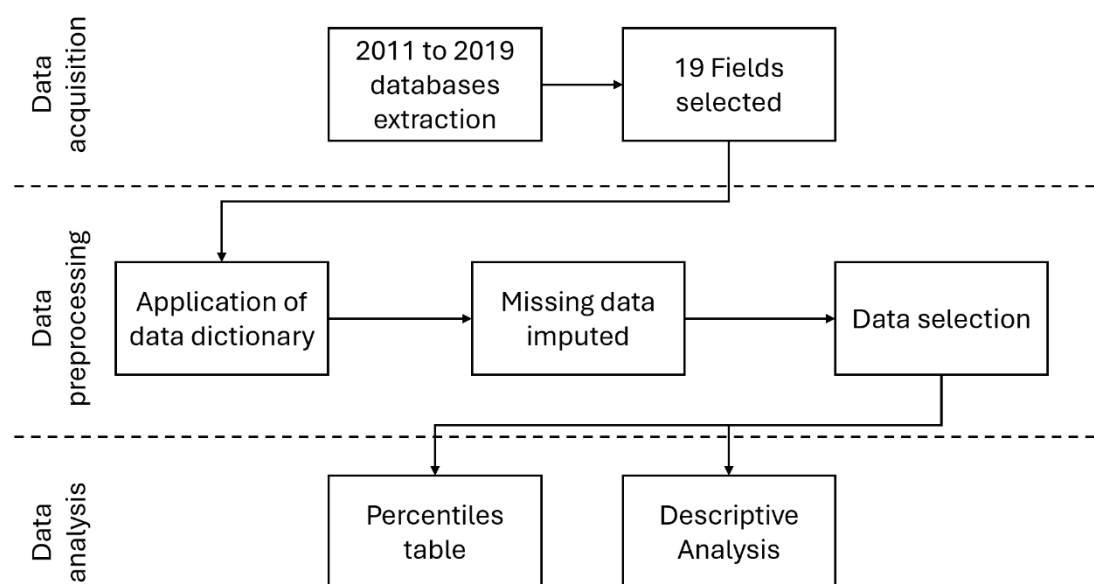


Figure 4: Diagram organizing every step of this analysis

The first step of this analysis was to retrieve the database files from SINASC's website. Nine comma separated values archives containing data for each birth of the years ranging from 2011 to 2019 were downloaded. This 9-year interval represents the latest three pre covid-19 pandemic triennia.

In that step, those nine files were imported to Orange and united in a whole csv file comprised of 26 million registries. It was necessary to select only the fields that are relevant to this study. This merged file was later saved as a unique csv to be accessed easily later.

The most recent SINASC file contains 69 data fields, but only the following 18 fields are used in this study. Table 4 contains the original field name, a translated description, and the possible values for categorical fields.

Table 4: 19 fields from SINASC

| Field name        | Description                                           | Possible values                                                                                                                                                                                                               |
|-------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>CONTADOR</i>   | Registry identification number                        |                                                                                                                                                                                                                               |
| CODMUNNASC        | IBGE birth municipality code                          |                                                                                                                                                                                                                               |
| <i>DTNASC</i>     | Birthdate in the format dd mm yyyy                    |                                                                                                                                                                                                                               |
| <i>DTULTMENST</i> | Last menstruation date in the format dd mm yyyy       |                                                                                                                                                                                                                               |
| <i>GESTACAO</i>   | Gestational week interval group                       | (1) less than 22 weeks;<br>(2) 22 to 27 weeks;<br>(3) 28 to 31 weeks;<br>(4) 32 to 36 weeks;<br>(5) 37 to 41 weeks;<br>(6) 42 weeks or more;<br>(9) ignored or no data.                                                       |
| <i>GRAVIDEZ</i>   | Multiple pregnancy flag                               | (1) Singleton pregnancy;<br>(2) Twin pregnancy;<br>(3) Triplet or more pregnancy;<br>(9) ignored or no data.                                                                                                                  |
| <i>PESO</i>       | Birthweight in grams                                  |                                                                                                                                                                                                                               |
| <i>IDANOMAL</i>   | Flag that tells if a major anomaly was identified     | (1) Yes;<br>(2) No;<br>(9) ignored or no data.                                                                                                                                                                                |
| <i>SEMAGESTAC</i> | Calculated Gestational week                           |                                                                                                                                                                                                                               |
| SEXO              | Newborn's Gender                                      | (0) ignored or no data;<br>(1) Male;<br>(2) Female.                                                                                                                                                                           |
| APGAR1            | Apgar Score on the first minute                       | From 0 to 10                                                                                                                                                                                                                  |
| APGAR5            | Apgar Score on the fifth minute                       | From 0 to 10                                                                                                                                                                                                                  |
| UFINFORM          | <i>Unidade Federativa</i> (Federal State) of registry |                                                                                                                                                                                                                               |
| RACACOR           | IBGE skin color/race classification for the newborn   | (1) <i>Branca</i> (White skinned);<br>(2) <i>Preta</i> (Dark skinned);<br>(3) <i>Amarela</i> (East Asian);<br>(4) <i>Parda</i> (Mixed race);<br>(5) <i>Indígena</i> (Brazilian Indigenous);<br>(9) <i>Ignorado</i> (Ignored). |
| RACACORMAE        | IBGE skin color/race classification for the mother    | (1) <i>Branca</i> (White skinned);<br>(2) <i>Preta</i> (Dark skinned);<br>(3) <i>Amarela</i> (East Asian);<br>(4) <i>Parda</i> (Mixed race);                                                                                  |

|           |                                   |                                                                                                                                                                                                                                                                                                                                                                |
|-----------|-----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|           |                                   | (5) <i>Indígena</i> (Brazilian Indigenous);<br>(9) <i>Ignorado</i> (Ignored).                                                                                                                                                                                                                                                                                  |
| ESMAE2010 | Mother's level of education       | (0) <i>Sem escolaridade</i> (No study);<br>(1) <i>Fundamental 1</i> (Elementary education);<br>(2) <i>Fundamental 2</i> (Intermediate education);<br>(3) <i>Médio</i> (Secondary education);<br>(4) <i>Superior Incompleto</i> (Incomplete College education);<br>(5) <i>Superior Completo</i> (Complete College education);<br>(9) <i>Ignorado</i> (Ignored). |
| ESTCIVMAR | Mother's Marital Status           | (1) <i>Solteira</i> (Single);<br>(2) <i>Casada</i> (Married);<br>(3) <i>Viúva</i> (Widowed).<br>(4) <i>Separada Judicialmente/Divorciada</i> (Divorced);<br>(5) <i>União estável</i> (Consensual Marriage);<br>(9) <i>Ignorado</i> (Ignored).                                                                                                                  |
| PARTO     | Type of delivery                  | (1) <i>Vaginal</i> (Vaginal);<br>(2) <i>Cesáreo</i> (Cesarean);<br>(9) <i>Ignorado</i> (Ignored).                                                                                                                                                                                                                                                              |
| TPMESTIM  | Gestational age estimation method | (1) <i>Exame físico</i> (Physical Exam);<br>(2) <i>Outro método</i> (Other method);<br>(8) <i>Undocumented</i> (Last menstruation date and birthdate);<br>(9) <i>Ignorado</i> (Ignored).                                                                                                                                                                       |

Please note that the fields named 'GESTACAO' and 'SEMAGESTAC' contains the same information: Gestational Week, but in different notations. The former is grouped in week intervals (e.g., 22 to 27 weeks); while the latter has been available since 2010 and it represents a calculated gestational week (e.g., 39<sup>th</sup> week).

An inconsistency was found in the file of the year 2014. The newborn gender for that year was classified as M (for male), and F (for female). All the other years, before and after 2014, were received as described in Table 4.

Some registers of AGPAR1 and APGAR5 were received as empty: 653,967, and 655,027, respectively. In total, 25,622,632 registries had both valid AGPAR1 and APGAR5 scores. Considering the volume of more than 26 million births, a 97.5% of completeness is acceptable. A more serious problem is that some registries were received with the invalid values of: '89', only one for APGAR1; and '99', 15,327 for

APGAR1; and 11,944 for APGAR5. A few of mother's age registries were also received as empty values: 202; and the invalid value of 99 years old: 97.

Table 5 shows new fields that were composed based on the original fields available in the SINASC database. The first digit from IBGE's municipality code represents the state region inside Brazil. Then, APGAR1 and APGAR5 scores were grouped based on APGAR (1953) experience. At last, the MOTHERS\_AGE\_GROUP is based on IBGE's population pyramid (AZEVEDO, [s.d.])

*Table 5: New fields that were created*

| Field name        | Description                                                 | Possible values                                                                                                                                                  |
|-------------------|-------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| REGION            | Brazilian region where the municipality of birth is located | (1) North;<br>(2) Northeast;<br>(3) Southeast;<br>(4) South;<br>(5) Central west.                                                                                |
| APGAR1_GROUP      | AGPAR1 grouped by condition                                 | (8, 9 or 10) Good condition;<br>(3, 4, 5, 6 or 7) Fair condition;<br>(0, 1, or 2) Poor condition.                                                                |
| APGAR5_GROUP      | APGAR5 grouped by condition                                 | (8, 9 or 10) Good condition;<br>(3, 4, 5, 6 or 7) Fair condition;<br>(0, 1, or 2) Poor condition.                                                                |
| MOTHERS_AGE_GROUP | Mother's age grouped                                        | 'Less than 10 years';<br>'10 to 14';<br>'15 to 19';<br>'20 to 24';<br>'25 to 29';<br>'30 to 34';<br>'35 to 39';<br>'40 to 44';<br>'45 to 49';<br>'50 and older'. |

The following step of data preprocessing consists of matching the codes from the databases to their respective meaning. Orange has a functionality called 'Edit Domain' that was used to transpose those original codes to their respective meanings (translated to English) provided by the data dictionary.

The next functionality employed ins named 'Impute', which is used to fill in incomplete data with a standard value. Numerical fields like weight, gestational week, municipality code, mother's age, and APGAR scores were filled with -1 (minus one);

then the categorical values of multiplicity, gender, anomaly flag, delivery type, newborns race, mother's marital, race, and education and were set to ignore.

The data preprocessing phase also encountered some obstacles. Some important fields had missing data, thus two stages of cleaning had to be executed. The first stage filtered out register with no weight, gender, or gestational week recorded, and also selected singletons with no anomalies. Those are the same filters that PEDREIRA et al. (2011) and the other similar studies applied.

PEDREIRA et al. (2011) also applied specific filters to SINASC's database to neutralize errors like the presented in Chapter 3, Figure 1. The second filter uses a maximum weight caps adapted from PEDREIRA et al. (2011), which is described in Table 6.

*Table 6: Weight filters applied by PEDREIRA et al. (2011)*

| <b>Gestational Age Group</b>                   | <b>Grams</b> |
|------------------------------------------------|--------------|
| Less than 22 weeks                             | 900          |
| From 22 <sup>nd</sup> to 27 <sup>th</sup> week | 1250         |
| From 28 <sup>th</sup> to 31 <sup>st</sup> week | 2000         |

The 1250g filter used by (PEDREIRA et al., 2011) seems adequate to their context, considering they only had the gestational groups available in the 2003-2005 databases. But upon observing registries within the gestational week granularity, it is noticeable that it is too strict with the upper weeks inside that group (noticeably 26<sup>th</sup> and 27<sup>th</sup> weeks) capping out a considerable part of the right end of the Normal distribution. The same behavior is observed within the 28<sup>th</sup> to 31<sup>st</sup> week group, so their filter also had to be adjusted.

Table 7 shows the new upper bounds for a registry to be considered valid. Consequently, newborns on any of the mentioned week groups presenting a weight higher than the respective upper bound are to be considered as data errors at this stage of study.

*Table 7: New weight filter for gestational age*

| <b>Gestational Age</b> | <b>Upper outlier filter</b> |
|------------------------|-----------------------------|
| 19                     | 1200                        |
| 20                     | 1200                        |
| 21                     | 1200                        |
| 22                     | 1300                        |
| 23                     | 1400                        |
| 24                     | 1500                        |
| 25                     | 1600                        |
| 26                     | 1800                        |
| 27                     | 2000                        |
| 28                     | 2100                        |
| 29                     | 2300                        |

|    |      |
|----|------|
| 30 | 2600 |
| 31 | 2800 |
| 32 | 3000 |
| 33 | 3300 |
| 34 | 3600 |
| 35 | 3800 |

A visual week-by-week analysis was conducted to determine the weight limits. Figure 5 shows an example of that analysis starting from the 22<sup>nd</sup> until the 28<sup>th</sup> stational Week.

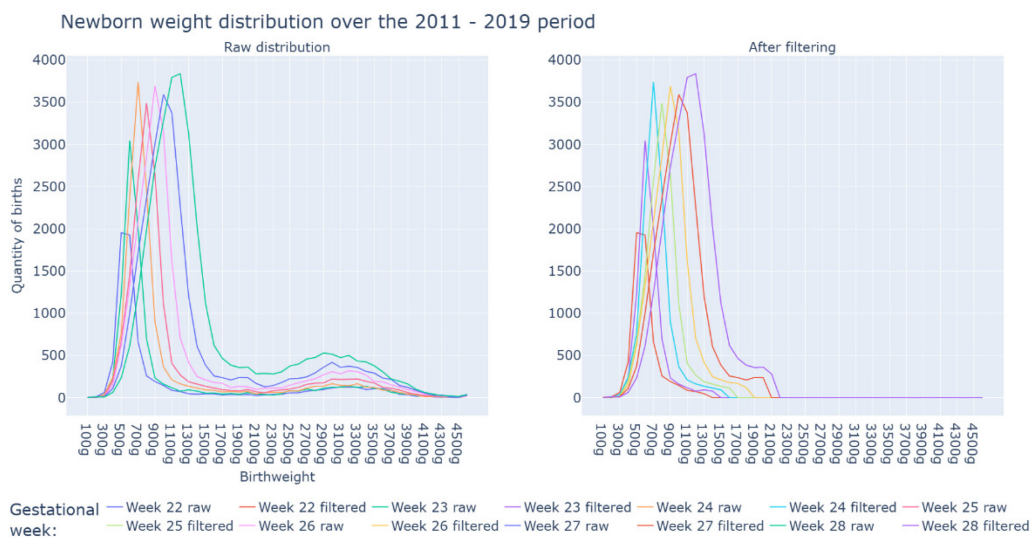


Figure 5: Shows the weight limit analysis process

Table 8 shows the number of registries left after each stage of data cleansing and the percentual compared to the raw numbers in the respective year. Special attention must be given to the year 2011 because it presented a count of more than a million registries missing the gestational week. Apart from that year, it is noticeable that the percentual of remaining registries increased consistently after every year, which shows that an effort in data validity and completeness must have been made in the period.

Table 8: Evolution of registries after each step of data cleanse

| Year | Raw register number | Register with no Gestational week | Register remaining after first filter | Register remaining after second filter |
|------|---------------------|-----------------------------------|---------------------------------------|----------------------------------------|
| 2011 | 2,913,160           | 1,339,644                         | 1,465,804 (50.32%)                    | 1,449,474 (49.76%)                     |
| 2012 | 2,905,789           | 239,756                           | 2,516,734 (86.61%)                    | 2,493,910 (85.83%)                     |
| 2013 | 2,904,027           | 132,353                           | 2,614,415 (90.03%)                    | 2,592,366 (89.27%)                     |
| 2014 | 2,979,259           | 86,972                            | 2,730,917 (91.66%)                    | 2,711,038 (91.00%)                     |
| 2015 | 3,017,668           | 78,839                            | 2,782,654 (92.21%)                    | 2,763,652 (91.58%)                     |
| 2016 | 2,857,800           | 56,968                            | 2,656,048 (92.94%)                    | 2,638,477 (92.33%)                     |
| 2017 | 2,923,535           | 51,842                            | 2,720,153 (93.04%)                    | 2,703,416 (92.47%)                     |

|      |           |        |                    |                    |
|------|-----------|--------|--------------------|--------------------|
| 2018 | 2,944,932 | 43,683 | 2,747,668 (93.30%) | 2,731,188 (92.74%) |
| 2019 | 2,849,146 | 37,092 | 2,671,249 (93.76%) | 2,656,077 (93.22%) |

The number of registries under or equal to the 18<sup>th</sup> week is irrelevant, so the Less than 22 weeks group started on the 19<sup>th</sup> week. Registers for the 46<sup>th</sup> and 48<sup>th</sup> gestational weeks were also ignored. Table 9 shows the number of registries by Gestational Age Groups for each of the examined trienniums:

*Table 9: A database comprising 22.7 million newborn records has been established after the application of all filters*

| Gestational age group                          | Number of registries per triennium |                         |                         |                         |
|------------------------------------------------|------------------------------------|-------------------------|-------------------------|-------------------------|
|                                                | 2003-2005                          | 2011-2013               | 2014-2016               | 2017-2019               |
| Total number of births (100%)                  | 7,991,111                          | 6,535,750               | 8,113,167               | 8,090,681               |
| Less than 22 weeks                             | 1,473<br>(0.0184%)                 | 812<br>(0.0124%)        | 1,030<br>(0.0127%)      | 1,154<br>(0.0143%)      |
| From 22 <sup>nd</sup> to 27 <sup>th</sup> week | 19,838<br>(0.2483%)                | 21,390<br>(0.3273%)     | 27,542<br>(0.3395%)     | 28,798<br>(0.3559%)     |
| From 28 <sup>th</sup> to 31 <sup>st</sup> week | 43,246<br>(0.5412%)                | 47,378<br>(0.7249%)     | 56,488<br>(0.6963%)     | 56,974<br>(0.7042%)     |
| From 32 <sup>nd</sup> to 36 <sup>th</sup> week | 380,433<br>(4.7607%)               | 613,675<br>(9.3895%)    | 693,385<br>(8.5464%)    | 675,609<br>(8.3505%)    |
| From 37 <sup>th</sup> to 41 <sup>st</sup> week | 7,460,161<br>(93.3557%)            | 5,590,467<br>(85.5367%) | 7,095,697<br>(87.4590%) | 7,125,515<br>(88.0706%) |
| 42 <sup>nd</sup> week or more                  | 85,960<br>(1.0757%)                | 262,028<br>(4.0091%)    | 239,025<br>(2.9461%)    | 202,631<br>(2.5045%)    |

After all these preparations, it is possible to advance to the fourth step. This stage consists of producing the gestational age/birthweight/percentile table with more recent data. PEDREIRA et al. (2011) had to apply a “shape-preserving piecewise cubic interpolation method” to calculate the percentiles for each gestational age parting from the gestational week interval, which was the smallest grain of information available in their dataset.

PEDREIRA et al. (2011)’s calculation process started with a vector that holds a central value and a width for each Gestational Age week group. Those values are shown in Table 10.

*Table 10: Central Values and Interval Widths used by Pedreira et al.(2011) to calculate their interpolated percentiles*

| Gestational age                                | Central Value | Interval band |
|------------------------------------------------|---------------|---------------|
| Less than 22 weeks                             | 21            | 1             |
| From 22 <sup>nd</sup> to 27 <sup>th</sup> week | 24.5          | 6             |
| From 28 <sup>th</sup> to 31 <sup>st</sup> week | 29.5          | 4             |
| From 32 <sup>nd</sup> to 36 <sup>th</sup> week | 34            | 5             |
| From 37 <sup>th</sup> to 41 <sup>st</sup> week | 39            | 5             |

|                               |    |   |
|-------------------------------|----|---|
| 42 <sup>nd</sup> week or more | 42 | 1 |
|-------------------------------|----|---|

PEDREIRA et al. (2011) has set an arbitrary number for the Central Value on the Less than 22 weeks and the 42<sup>nd</sup> week. The value for the remaining week groups is calculated as the mean of the initial and final values of each group.

After that, nine vectors are created: Each one comprising calculation of the percentiles (3<sup>rd</sup>, 5<sup>th</sup>, 10<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, 95<sup>th</sup> and 97<sup>th</sup>) for each of the 7-week groups.

Then the matlab function called `interp1` was called using the `pchip` (Piecewise Cubic Hermite Interpolating Polynomial) as interpolation method. This function transforms each of the nine percentile vectors that contained values for 7-week groups in nine percentile vectors with the interpolated values for every integer value between the minimal and maximal Central Value (respectively 21 and 42). In other words, it plots a curve using the percentiles for the 7-week groups, then estimates the values for each week between the 21<sup>st</sup> and the 42<sup>nd</sup>.

It is much simpler to calculate the percentiles per age because there is a finer calculated gestational week available in the dataset.

The last step can also be taken in parallel with the previous one. An exploratory analysis of the data is performed and a profile for the average Brazilian mother is produced.

## 5 Statistical and Exploratory Analysis, and Discussions

### 5.1 Statistical Analysis

At first, this study intended to apply statistical hypothesis tests to evaluate if there were significant changes in the whole population by comparing the whole 2003-2005 triennium to the most recent 2011-2019 years.

Hypothesis tests are important to evaluate populations based on sample data. They provide a structured approach to evaluate suppositions over a population based on empirical evidence.

Two mutually exclusive propositions are set in this kind of test. The null hypothesis (H0) is a declaration that there is no effect or correlation; the alternative hypothesis (H1) contradicts the null hypothesis and represents the point of view that the researcher wants to prove. The specific hypothesis test depends on the data available and the kind of question being asked.

Following, an estimation indicator is chosen: it could be an average, or standard deviation, among others. A significance level that determines the maximum acceptable false positive rate is also chosen. Then observations are computed. In the end, there is a decision to choose which of the contradicting hypothesis is true, and consequently, the other hypothesis is determined as false.

The Z-test is a statistical test used to determine if there is a significant difference between the average of a sample and the average of a population, or between two independent samples extracted from the same population. The STD (Standard Deviation) from the population ought to be known, and the sample size must be greater than 30. It is presumed that the population follows the Normal distribution. Equation 1 shows how those inputs form the Z number.

$$Z\_VALUE = \frac{SAMPLE1_{MEAN} - SAMPLE2_{MEAN}}{\sqrt{\frac{SAMPLE1_{STD}^2}{SAMPLE1_{SIZE}} + \frac{SAMPLE2_{STD}^2}{SAMPLE2_{SIZE}}}}$$

*Equation 1: Formula for a Z-test with two samples*

The Z-test for two samples utilizes the before-mentioned formula and returns a z-value. This value must be compared to the Standard Normal Table. In this table the critical z-value for a 5% of significancy in a bilateral distribution is 1.96.

For this test, the statistical Population is all newborns in Brazil, SAMPLE1 represents those born in the 2003-2005 period, and SAMPLE2 represents those born in the 2011-2019 period. Therefore, H0 affirms that the average weight for sample 1 equals the average weight for the sample 2. In other words, no significant change has occurred.

The alternative H1 affirms that those averages are different, resulting in a change in the population profile.

Our parameters for the Z-test are calculated as:  $SAMPLE1_{MEAN} = 3203.53$ ;  $SAMPLE2_{MEAN} = 3205.36$ ;  $SAMPLE1_{STD} = 534.98$ ;  $SAMPLE2_{STD} = 536.52$ ;  $SAMPLE1_{SIZE} = 7,991,111$ ;  $SAMPLE2_{SIZE} = 22,739,598$ . Those parameters when applied to the Equation 1 returns a z-value of 8.48.

Now, it is time to compare the z-value with the critical z-value: If the z-value is higher, then we accept the H1 hypothesis; otherwise H0 is considered true. Since 8.48 (z-value) is higher than 1.96 (critical z-value), culminating in the conclusion that H1 is true, which means that there was a significant statistical change in the average weight.

There are two issues with this primary evaluation: the first point is that the sample sizes of millions seem to be too big, leading to a denominator from the formula being a very small number, resulting in a big z-value after the division; the second is that declaring a difference in the global average of almost 3 grams may be interesting from a statistical point of view, but in the end of the day may not be as much relevant to the obstetrics and neonatology community.

## 5.2 Exploratory Analysis and discussion

An exploratory analysis can capture this evolution with much more attention to nuances, making this kind of analysis more interesting from a clinical point of view. Tail groups such as premature and postmature, which represent higher risk, have reduced their numbers over the studied period. There was also a significative change in on-term curves.

Table 11 shows the raw numbers as collected from SINASC database. It is noticeable that the 22<sup>nd</sup> and 25<sup>th</sup> weeks percentiles have been severely impacted by either weight and/or gestational age misannotations/typing mistakes.

*Table 11: A selected group of gestational ages and their respective raw weight percentiles before filtering corrections*

| Gestational Age | Birthweight Percentiles |        |                  |        |                  |        |                  |        |                  |        |
|-----------------|-------------------------|--------|------------------|--------|------------------|--------|------------------|--------|------------------|--------|
|                 | 5 <sup>th</sup>         |        | 25 <sup>th</sup> |        | 50 <sup>th</sup> |        | 75 <sup>th</sup> |        | 95 <sup>th</sup> |        |
|                 | Male                    | Female | Male             | Female | Male             | Female | Male             | Female | Male             | Female |
| 22              | 390                     | 365    | 485              | 460    | 560              | 535    | 890              | 980    | 3300             | 3204   |
| 25              | 520                     | 475    | 695              | 650    | 810              | 760    | 1000             | 960    | 3115             | 3055   |
| 30              | 935                     | 875    | 1295             | 1220   | 1560             | 1495   | 2250             | 2320   | 3365             | 3270   |
| 35              | 1880                    | 1800   | 2360             | 2270   | 2710             | 2620   | 3150             | 3050   | 3660             | 3550   |
| 40              | 2750                    | 2650   | 3157             | 3040   | 3430             | 3305   | 3720             | 3588   | 4095             | 3950   |
| 42              | 2688                    | 2595   | 3125             | 3010   | 3420             | 3300   | 3725             | 3595   | 4120             | 3975   |

It is noticeable from this table that those incorrect outliers would severely impact any analysis: they completely distort the 95<sup>th</sup> percentile for the 22<sup>nd</sup> gestational week.

Table 12 shows the same weeks in Table 11 but the percentile values have been corrected by the filtering process described in chapter 3, which is an adaptation of the process applied by PEDREIRA et al. (2011).

The filters applied in this study are: weight must be filled and inside the parameters set in Table 7; Gender filled and not ignored; Gestational Week Group must be not ignored; Gestational Week between 21<sup>st</sup> and 44<sup>th</sup>; Multiplicity is Singleton; and Anomaly is No.

Table 12: The same gestational ages, but after the filtering process

| Gestational Age | Birthweight Percentiles |        |                  |        |                  |        |                  |        |                  |        |
|-----------------|-------------------------|--------|------------------|--------|------------------|--------|------------------|--------|------------------|--------|
|                 | 5 <sup>th</sup>         |        | 25 <sup>th</sup> |        | 50 <sup>th</sup> |        | 75 <sup>th</sup> |        | 95 <sup>th</sup> |        |
|                 | Male                    | Female | Male             | Female | Male             | Female | Male             | Female | Male             | Female |
| 22              | 390                     | 365    | 475              | 450    | 532              | 510    | 612              | 595    | 940              | 944    |
| 25              | 505                     | 464    | 675              | 625    | 780              | 728    | 885              | 830    | 1190             | 1150   |
| 30              | 900                     | 840    | 1225             | 1145   | 1465             | 1390   | 1705             | 1650   | 2350             | 2400   |
| 35              | 1910                    | 1835   | 2395             | 2315   | 2735             | 2665   | 3130             | 3060   | 3600             | 3540   |
| 40              | 2758                    | 2660   | 3160             | 3045   | 3435             | 3305   | 3720             | 3590   | 4178             | 4025   |
| 42              | 2700                    | 2600   | 3130             | 3015   | 3420             | 3300   | 3730             | 3600   | 4205             | 4060   |

The weeks starting from 36<sup>th</sup> and later did not need to be adjusted because their distributions seemed to be more reliable, as can be seen in Figure 5.

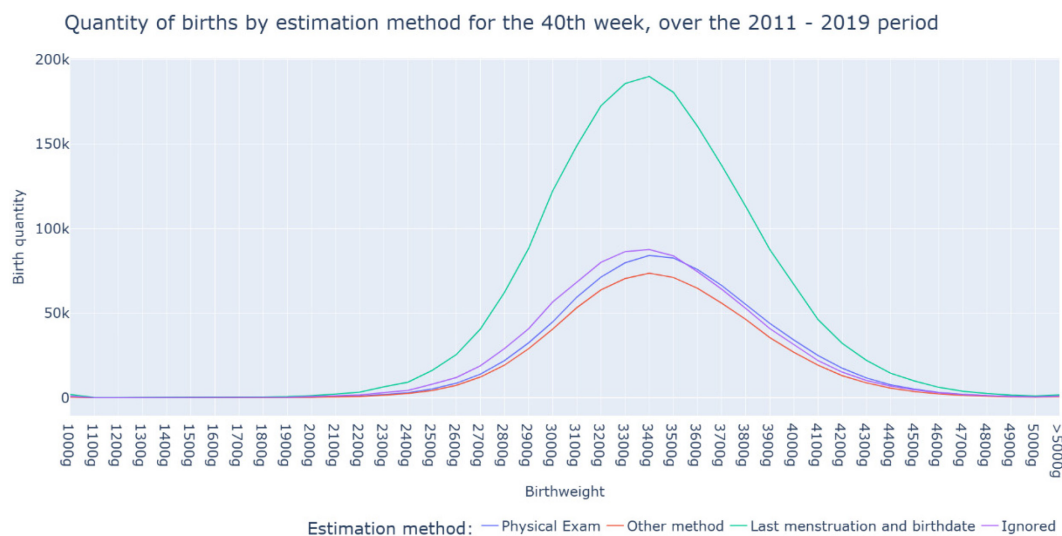


Figure 6: Raw weight distribution for the 40th gestational week

It is necessary to refer to Table 9 at this moment of the discussion. This table has been reorganized as a percentual increase or decrease of the number of births of the same gestational group when compared to the 2003 to 2005 period, and it is shown as Table 13.

*Table 13: Addition of a delta to Table 9*

| <b>Δ evolution compared to 2003-2005</b> | <b>2003-2005</b>     | <b>2011-2013</b>     |          | <b>2014-2016</b>     |          | <b>2017-2019</b>     |          |
|------------------------------------------|----------------------|----------------------|----------|----------------------|----------|----------------------|----------|
|                                          |                      | <b>Total</b>         | <b>Δ</b> | <b>Total</b>         | <b>Δ</b> | <b>Total</b>         | <b>Δ</b> |
| Total number of births                   | 7,991,111 (100%)     | 6,535,750 (100%)     | -18.21%  | 8,113,167 (100%)     | 1.53%    | 8,090,681 (100%)     | 1.25%    |
| Less than 22 weeks                       | 1,473 (0.0184%)      | 812 (0.0124%)        | -44.87%  | 1,030 (0.0127%)      | -30.07%  | 1,154 (0.0143%)      | -21.66%  |
| From 22nd to 27th week                   | 19,838 (0.2483%)     | 21,390 (0.3273%)     | 7.82%    | 27,542 (0.3395%)     | 38.83%   | 28,798 (0.3559%)     | 45.17%   |
| From 28th to 31st week                   | 43,246 (0.5412%)     | 47,378 (0.7249%)     | 9.55%    | 56,488 (0.6963%)     | 30.62%   | 56,974 (0.7042%)     | 31.74%   |
| From 32nd to 36th week                   | 380,433 (4.7607%)    | 613,675 (9.3895%)    | 61.31%   | 693,385 (8.5464%)    | 82.26%   | 675,609 (8.3505%)    | 77.59%   |
| From 37th to 41st week                   | 7,460,161 (93.3557%) | 5,590,467 (85.5367%) | -25.06%  | 7,095,697 (87.4590%) | -4.89%   | 7,125,515 (88.0706%) | -4.49%   |
| 42 weeks or more                         | 85,960 (1.0757%)     | 262,028 (4.0091%)    | 204.83%  | 239,025 (2.9461%)    | 178.07%  | 202,631 (2.5045%)    | 135.73%  |

Now it is possible to notice the numeric and percentage evolution of total births quantity among the four distinct three-year periods. The absolute number of live births slightly varied among the recent periods when compared to 2003 to 2005: the period of 2011 to 2013 is an exception because of a reduction of 18.21% - unfortunately more than a million of 2011 registers were excluded due to data quality issues -, but then the following periods of 2014 to 2016 and 2017 to 2019 resulted in a respective increase of 1.53% and 1.25% in the total number of births.

After analyzing the whole results, it is interesting to notice the variations found among each gestational age group over the same time periods. The first group, which consists of babies born in less than 22 weeks of gestation, has seen a percentual decrease when comparing the first period (2003-2005) to the three later triennia (2011-2013, 2014-2016 and 2017-2019): 44.87%, 30.07%, and 21.6%, respectively.

This is an exceptional group, the ones that have such brief pregnancies that they are evidently not prepared for delivery, hence the high mortality rates. Further studies must be held to understand if there is an increase in mortality rates for those very early weeks, or if those babies are taking longer to be delivered, hence being classified in a further gestational week group, or even if the filters applied to them at 1200g was excessively excluding.

The second and third groups had increased percentages when comparing the time periods as well. The 22<sup>nd</sup> week to 27<sup>th</sup> week group increase rates are: 7.82%, 38.83%, and 45.17%. Then the 28<sup>th</sup> to 31<sup>st</sup> week group resulted in 9.55%; 30.62%, and 31.74% when compared to the 2003 to 2005 triennium. Those groups also represent highly premature deliveries and seem to be experiencing increasingly better survival rates. But it is also important to note that those gestational weeks in this study underwent the selection filters described in Table 7, which are different when compared to what was applied by PEDREIRA et al. (2011) as specified in Table 6.

The fourth and sixth groups also presented significant percentage differences. From the 32<sup>nd</sup> to 36<sup>th</sup> week the difference was 61.31%, 82.26%, and 77.59% for the respective year groups; and then the 42 weeks or more saw an increase of a whopping 204.83%, 178.07% and 135.73% when compared to the first triennium.

Nonetheless, the 37<sup>th</sup> week to 41<sup>st</sup> week group, which is the expected pregnancy duration, and it is also called full term pregnancy, was the only one gestational age group that had their numbers reduced when comparing the three more recent periods to the original years from 2003 to 2005: a reduction of 25.06% in 2011 to 2013; 4.89% in 2014 to 2016; and 4.49% in 2017 to 2019. This percentual reduction is not necessarily a concern because the other gestational week groups are gradually increasing, it means that there is a reorganization inside of the whole 100% of those respective periods.

From a macro-view: Apart from the less than 22 weeks, all the other groups considered pre-term stretched their numbers. This means that preterm babies are having more chances of survival. On the other hand, the sixth group, which is composed of post-term babies, also increased their respective percentual over the years, so longer pregnancies are also observing more chances of survival.

Figure 6 presents an overview of the birth quantity grouped by the gestational age week group over the period being studied. Please note that the scale is logarithmic to make a percentual in different orders of magnitude more readable.

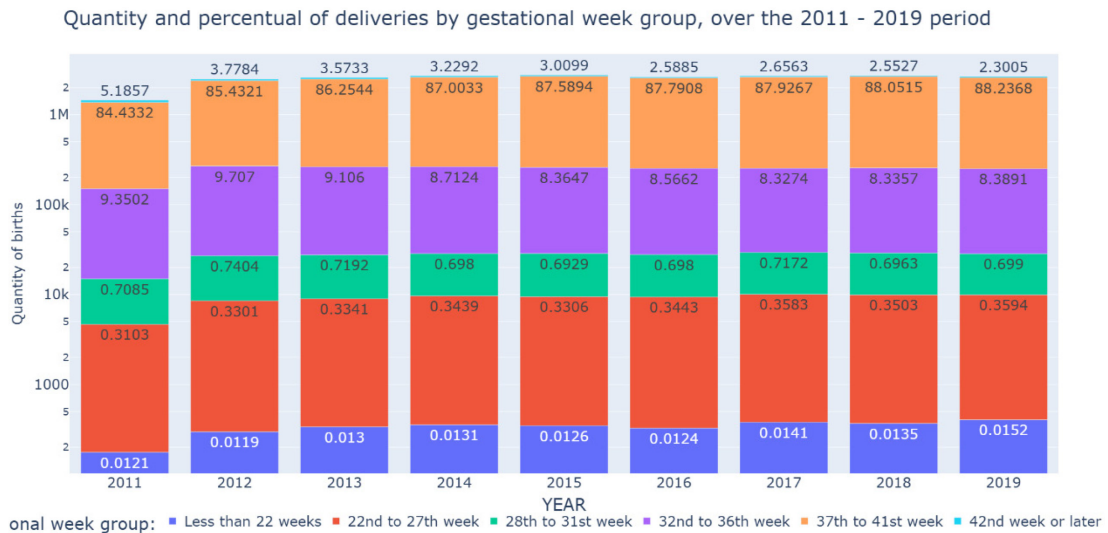


*Figure 7: Evolution of birth quantity by gestational week group in percentage*

It is noticeable with more detail how there is a gradual and consistent increase in the percentual of births that occurred in the full-term weeks, starting from 84.43% in 2011 and reaching as high as 88.24% in 2019. Moreover, the less than 22 weeks percentual has also increased from 0.0122% to 0.0153%, which can also indicate that more of the premature babies closer to the breaking point of neonatal death are surviving.

An important aspect in debate around the relation between the health of a newborn and its mother's age: it is known and acceptable that younger mothers have physiologies more suitable to give birth to healthy babies, but there is a tradeoff in the choosing to have a baby at a later age because the mother's financial situation could be better organized. Thus, providing better health conditions to the belated baby. Figure 7 shows the evolution on the average of the mother's age and one standard deviation interval across the studied period.

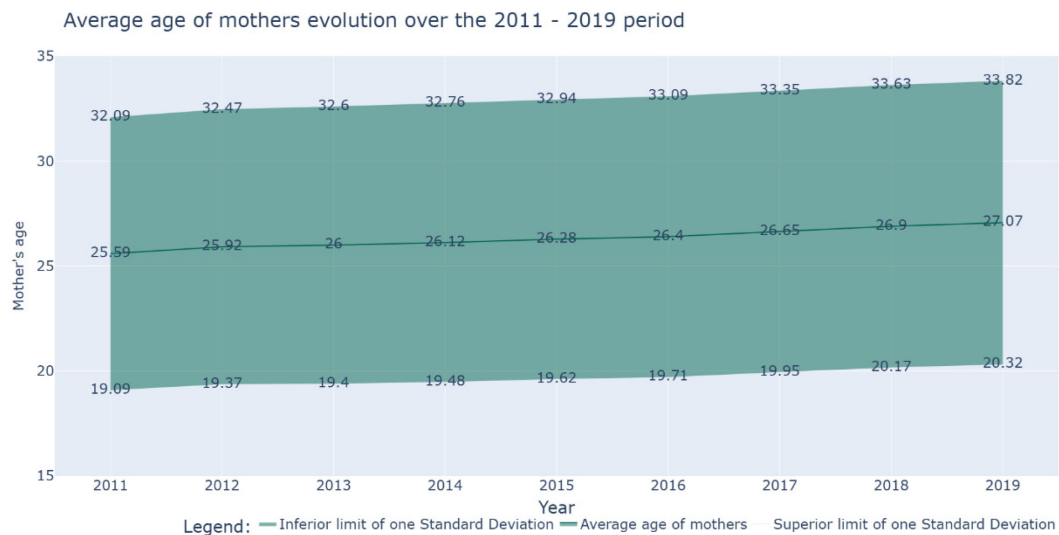


*Figure 8 Evolution of the average mother's age*

From Figure 7 we can see that not only is the average age of mothers increasing over the time period: from 25.6 years old in 2011 to 27 years old in 2019, almost one year and a half older in the period of just 9 years; but the one standard deviation window is also slightly increasing. This means that the age of the mother is getting somewhat more heterogeneous.

Besides the mother's age, the APGAR is an important indicator of newborn's health. Figure 8 analyses the evolution of APGAR1 and APGAR5 averages among the studied period.

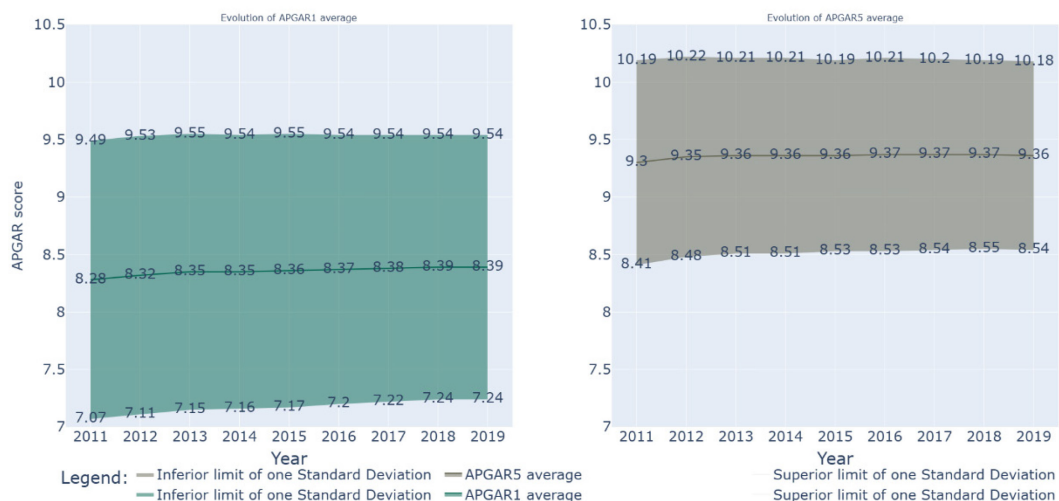


*Figure 9: Evolution of the APGAR1 and APGAR5 averages*

From this figure it is possible to see that there is a gradual increase in the average APGAR1 and APGAR5 scores. There is also an indication that both scores are becoming more homogeneous because their standard deviation window is becoming thinner: For example, in 2011 one Standard deviation for APGAR1 was calculated as

1.21, but in 2019 it was calculated as 1.15. Please note that even though the superior limit for one Standard deviation for the APGAR5 exceeds 10, the higher possible value for the APGAR score is 10.

Table 14 reveals the Brazilian average of APGAR1 and APGAR5 indicators, and the number and percentual of deliveries, grouped by delivery type. The infant number filters out the ones that presented invalid values for the APGAR1 indicator.

*Table 14: Average APGAR1 score in Brazil by delivery type*

| <b>Delivery type</b> | <b>Average Apgar1</b> | <b>Average Apgar5</b> | <b>Infant Number</b> | <b>%</b> |
|----------------------|-----------------------|-----------------------|----------------------|----------|
| Cesarean             | 8.38                  | 9.37                  | 12.432.082           | 56.04%   |
| Ignored              | 8.24                  | 9.28                  | 14.745               | 0.07%    |
| Vaginal              | 8.33                  | 9.34                  | 9.713.121            | 43.89%   |

This table shows that the preferred delivery method in Brazil is the cesarian, with roughly 56% of the total number of births; while the remaining 44% are vaginal deliveries. This proportion of delivery types is quite different from the numbers collected by APGAR et al. (1958) back in the fifties: 86% and 10% for Vaginal and Cesarean, respectively. The APGAR1 and APGAR5 averages do not seem to be affected by the choice of delivery, except on the very few times when the delivery type is ignored. APGAR5 average is higher than APGAR1 average, as expected.

Table 15 advances the investigation of the APGAR1 indicator. This table shows the number of deliveries by APGAR1 group and furthers the grouping by drilling down by Delivery type.

*Table 15: Number of deliveries by APGAR1 group and delivery type*

| <b>APGAR1 Group</b> | <b>Infant Number</b> | <b>% of total</b> | <b>Delivery type</b> | <b>Infant Number</b> | <b>% inside group</b> | <b>% of total</b> |
|---------------------|----------------------|-------------------|----------------------|----------------------|-----------------------|-------------------|
| Poor condition      | 149,851              | 0.67%             | Vaginal              | 84154                | 56.16%                | 0.38%             |
|                     |                      |                   | Cesarean             | 65572                | 43.76%                | 0.29%             |
|                     |                      |                   | Ignored              | 125                  | 0.08%                 | 0%                |
| Fair condition      | 2,522,713            | 11.34%            | Vaginal              | 1156392              | 45.84%                | 5.20%             |
|                     |                      |                   | Cesarean             | 1364220              | 54.08%                | 6.13%             |
|                     |                      |                   | Ignored              | 2101                 | 0.08%                 | 0.01%             |
| Good condition      | 19,576,289           | 87.98%            | Vaginal              | 8524038              | 43.54%                | 38.31%            |
|                     |                      |                   | Cesarean             | 11039622             | 56.39%                | 49.62%            |
|                     |                      |                   | Ignored              | 12629                | 0.06%                 | 0.06%             |

The substantial majority of babies are delivered among the good-conditioned group, which represents 88% of deliveries; the fair condition group accounts for 11.33% of deliveries; and the remaining 0.67% are considered to be in the poor condition group.

The Brazilian good condition group percentual is 20 percentual points higher than the ones from APGAR et al. (1958), and the poor condition group 5 percentual points lower than the ones from the same referred study. This is a significant indication that Brazilian babies are born healthier than the Americans from that extract.

The proportion of delivery type by APGAR1 evaluation presents another interesting difference: the percentual number of vaginal deliveries rises and the number of cesarean shrinks as the APGAR1 condition deteriorates in Brazil. Which is the contrary tendency of what is observed in APGAR et al. (1958).

Figure 9 investigates the relations between mothers' race and delivery type. From left to right: Brazillian Indigenous mothers tend deliver babies in north-american-like numbers, almost 80% of vaginal deliveries, but this can be explained by this group choosing more traditional methods and a difficulty of reaching medical facilities; both dark skinned and mixed race mothers deliveres are in a relative balance between cesarean and vaginal; then east asian and white skinned mothers deliver proportionaly more by cesarean, this could be an indicator that this socioeconomical group has more access to private medical facilities.

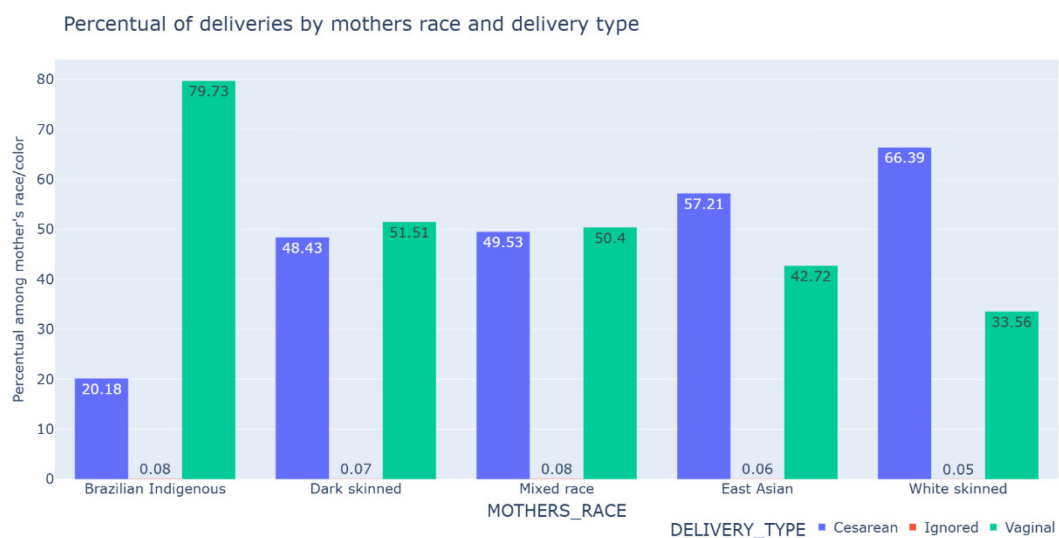


*Figure 10: Proportions of delivery type by mothers' race*

Now it is time to start dissecting other dimensions of the babies and their mothers to try to understand the trends and to build an average profile. Figure 10 starts this discussion by showing the average weight by Brazilian region over the years.

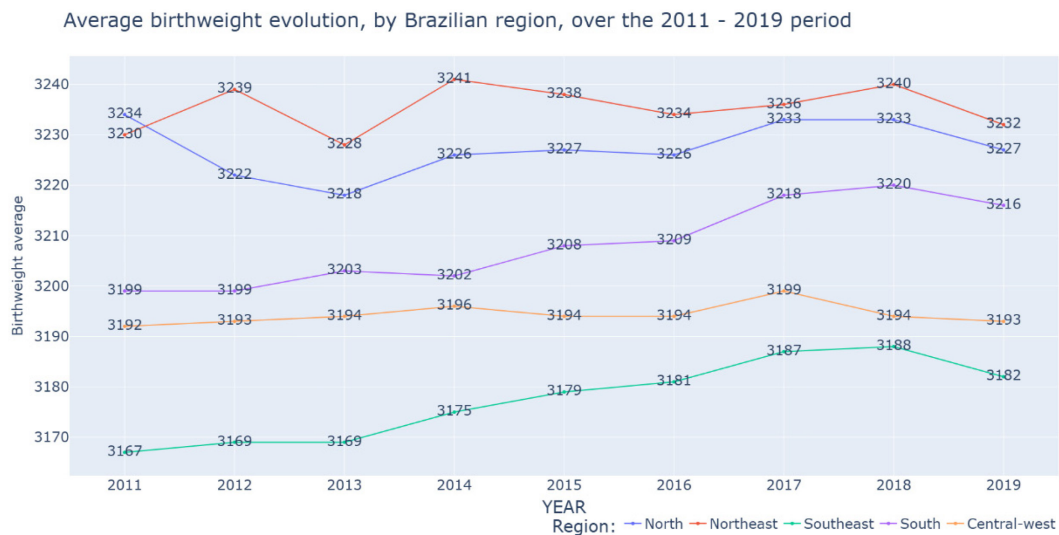


*Figure 11: Average weight evolution grouped by Brazilian region*

From Figure 10 it is possible to evaluate that babies in the Southern and East-southern regions are consistently getting bigger and heavier over the period; Babies from the North-eastern and Northern regions seem to have suffered a decay until 2013 then a period of growing heavier after 2014, while babies from the West-Central regions seems to be maintaining the same weight average pattern.

More interestingly, and maybe counter-intuitively, babies from the Northern and East-northern regions, which are the poorest regions from Brazil, are born heavier than all the babies delivered from the other three richer regions.

In order to investigate this seemingly contradiction, Figure 11 analyses the average age of the babies' mother, by region and over the time period of this study.

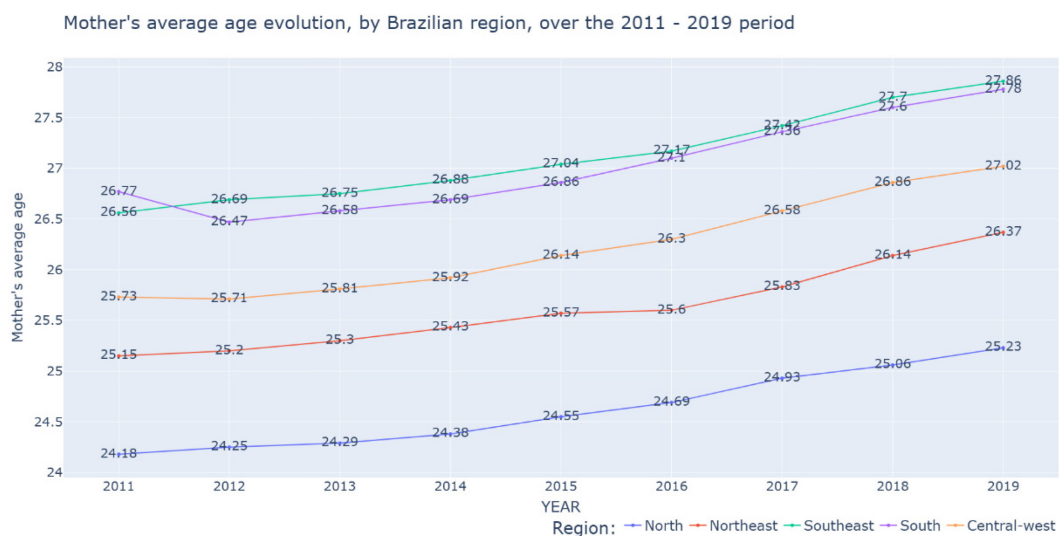


*Figure 12: Average mother's age grouped by Brazilian region*

From this figure we learn that mothers age is consistently increasing for all Brazilian regions; but more importantly: the North and East-North regions mothers are respectively, on average, around 2 and 1 years younger than the mothers from the South and East-South regions. This age gap is the first aspect that can help explain the average weight difference between regions.

IBGE organizes the population demographic pyramid in groups bins of 5 years. In this study, the mother's age is organized in the same way. The mothers were aggrouped in of the following ages: 'From 15 to 19', 'From 20 to 24', 'From 25 to 29', 'From 30 to 34', and 'From 35 to 39' years old. Mothers of 14 or younger and mothers over 40 years old were excluded from the next analysis because of their proportionally lower number of appearances. Figure 12 shows the evolution of the weights of babies born over the period of this study.

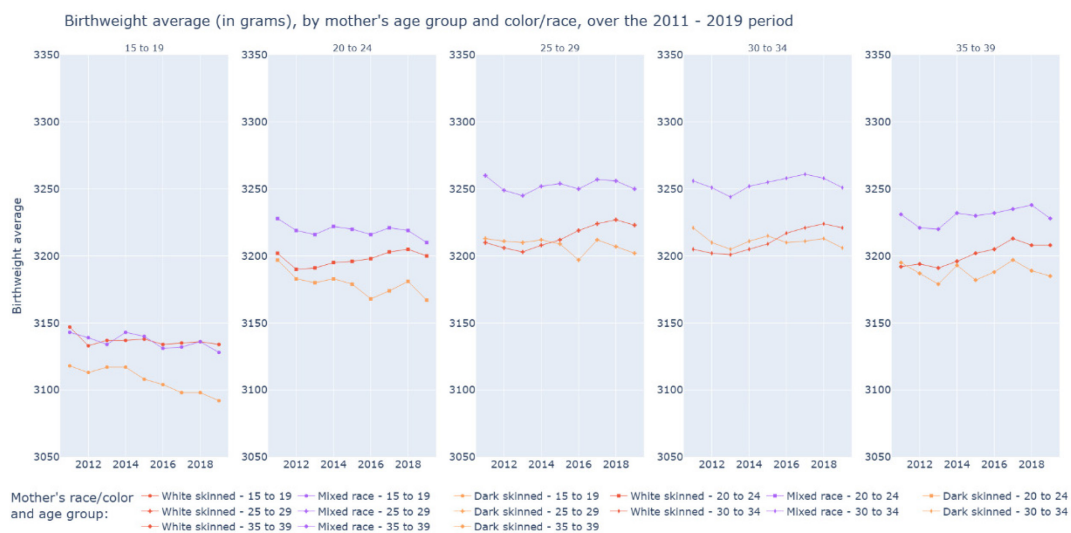


*Figure 13: Average weight, grouped by mothers' age group and skin color*

From Figure 12 it is possible to understand that: mothers of mixed race deliver heavier babies than white or dark skinned mothers in every age group and along the time period; all mothers from the '15 to 19' year old group seem to be delivering lighter babies; dark skinned mothers are delivering lighter babies along the time period; white skinned mothers are delivering heavier babies in the timespan; and mixed race mothers seems to be fluctuating or in a stagnant weight average over the period.

A different aspect is now analyzed by Figure 13. It shows the quantity of babies delivered over this study's time period, grouped by mothers' group age and skin color.

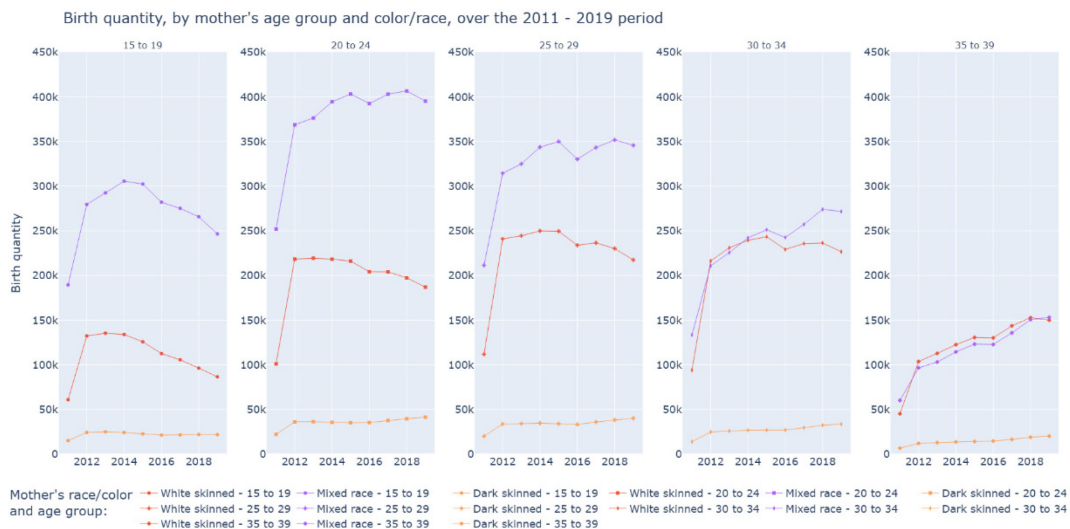


Figure 14: Number of babies delivered by mothers' age group and skin color

First, contextualizing with IBGE 2022 Brazilian census: 45.3% of Brazilians considered themselves as mixed race; then 43.5% considered themselves white skinned; and only 10.2% considered themselves dark-skinned. Important observations about this most recent census include: the expression 'considered themselves' is used because the census literally makes use of a self-declaration questionnaire, and more importantly: this was the first time since the census is being performed that the mixed race percentual is higher than the white skinned.

Figure 13 helps understand this movement: mixed race mothers delivered more babies than white mothers in all age groups, the only exception being the '35 to 39' years old. In other words, mixed race mothers are younger and in higher numbers. This can also explain the higher weight average.

The quantity of babies from dark and white skinned mothers has reduced in the '15 to 19' years old groups; and in the same age group for mixed race mothers after 2014.

The quantity of babies delivered from mixed-skinned mothers rises over time but lowers with the age group. On the other hand, white skinned mothers deliver less babies in the period, but the number grows with the passing of the years. This means that white mothers are taking longer to decide to have babies. Notice how white mothers will only surpass mixed-skinned mothers in the age group from '35 to 39' years old.

Figure 14 helps the understanding of this behavior. White skinned mothers have more years of education than all the other race groups: 45% of them have Complete Secondary education, 5% Incomplete College, and 26.6% of them reached Complete College Education. Mixed race mothers have a similar 47% of Complete Secondary education, but they fall behind with less than 3.5% with Incomplete College and less than 7% that have acquired College Diplomas.

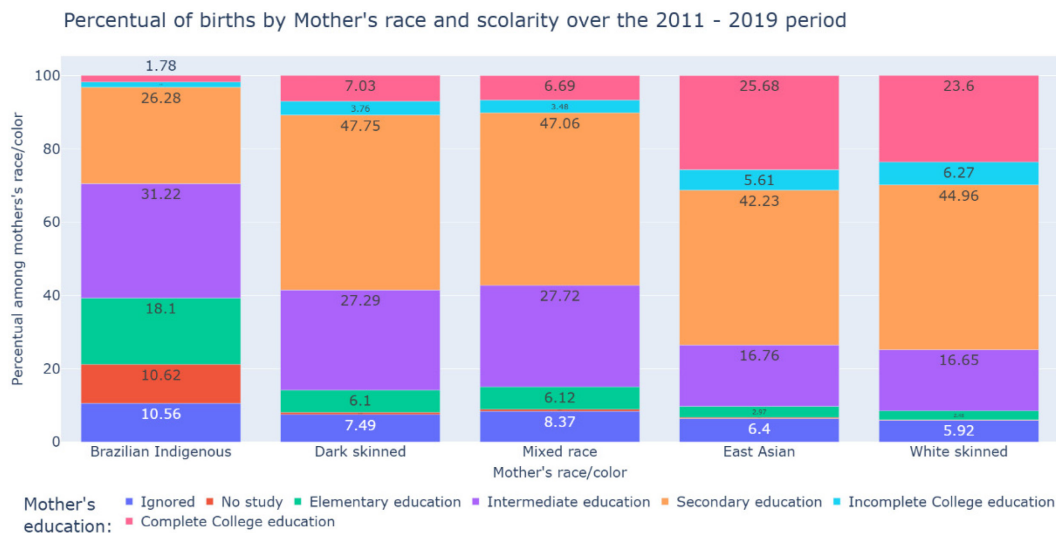


Figure 15: Percentual of births by mother's race and education

White skinned mothers seem to be prioritizing studies and work life over having babies when compared to mixed race mothers, this can explain the behavior seen in Figure 13 where the number of deliveries from white skinned mothers grows and the number of babies from mixed race mothers shrinks as they age.

From Figure 15 it is possible to observe that there is a positive temporal relationship between the newborn's weight and the mother's age in most age groups and regions.

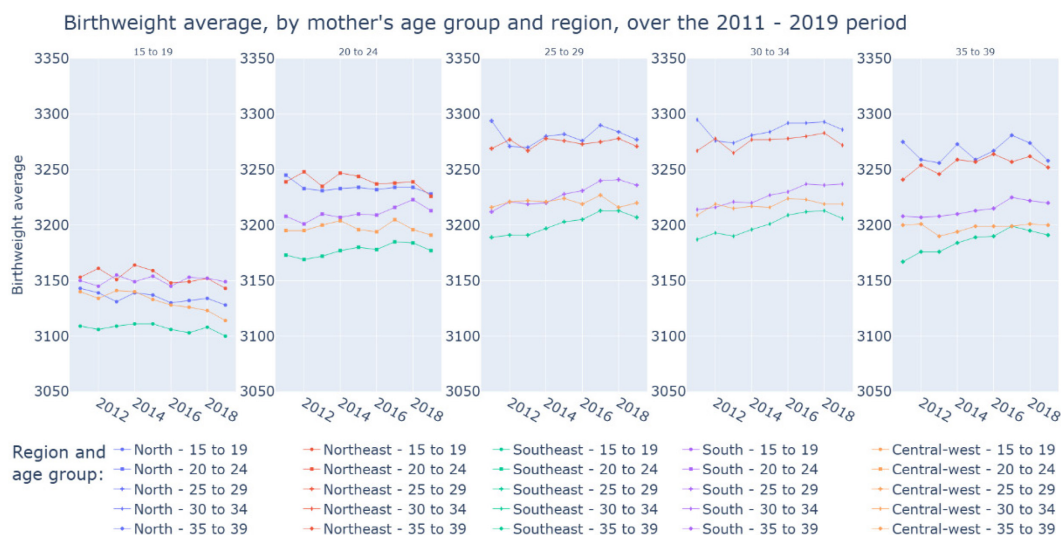


Figure 16: Average weight by region and mothers' age group over the period

The babies were, on average, heavier over the whole period. Exceptions include: mothers in the age group '15 to 19' years old, in every region, delivered babies weighing less; mothers in the age group '20 to 24' years old in North and Northeast region also delivered progressively lighter babies; and mothers in the age group '25 to 29' years old in the North and Northeast region remained in almost the same average over the period.

Figure 16 leads to a dividing path for the analysis and search for a profile that describes an *average* Brazilian mother. This figure shows two treemaps of the accumulated births in the whole period, but from two different points of view: on the left the births are group first by mother's race then Brazilian region; on the right the grouping was done first by region then by mother's race. A treemap is a visual representation of a tree, in this case: the *root* is Brazil, with the total of births that occurred between 2011 and 2019; then the intermediate group of boxes describes the mother's race (tree on the left) or the Brazilian region (tree on the right); and then the inner box describes the Brazilian region or mother's race, respectively for the left and right tree.

As each box represents a combination of mother's race and region; the box's size is proportional to the percentual of said combination among the whole, which is Brazilian total of births in the period. Inside each box, the first number is the count of births for the respective mother's race/region combination; the second number is the percentual of the inside box relative to the intermediate box; and the third number is the percentual relative to the whole Brazilian dataset.

Total birth quantity, by mother's color/race and brazilian region Treemaps

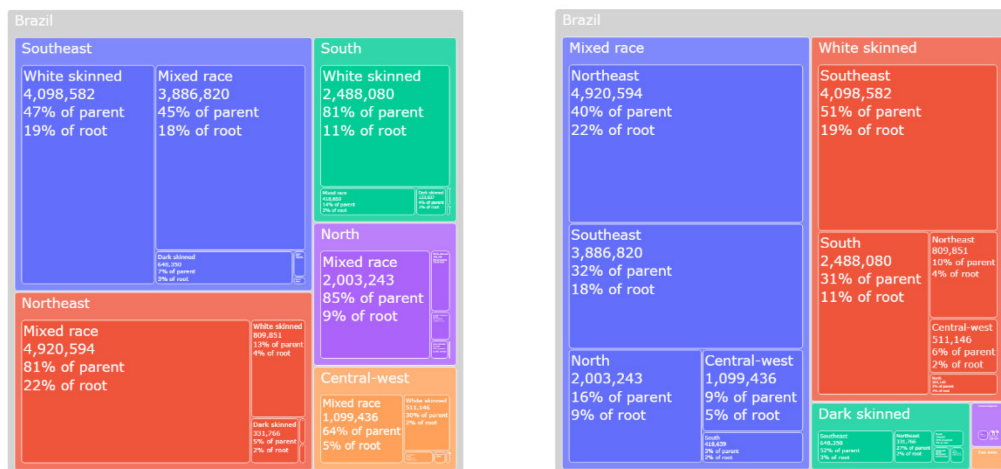


Figure 17: Overview of babies born grouped by region and mothers' skin color

Even though the numbers for mothers from the southeast region are higher than the numbers of mothers from each other region, the number of mixed-race mothers is higher than the mothers of other skin colors. The combination of Mixed-race mother with Northeast region summed up 22% of the Brazilian births, so these are the first dimensions to determine the profile of an average Brazilian mother.

The third attribute for the average Brazilian mother is the mother's age group, the aggregated group from 20 to 29 years old represents 49.4% of all births, and when considered with the mixed-race in the northeast region forms a group of 11% of all births occurred in the period.

From this profile: Mixed-race, Northeastern, from 20 to 29 years old, this study now analyses the weight distribution of children descendant from those mothers in comparison with children born from mothers that do not fit in that profile.

Figure 17 shows a comparison of the mean birthweight from prevenient from mothers categorized as the most common or “average Brazilian mother” compared to the rest of the population. That group presents a higher average weight, but this average seems to be decreasing over the period. On the other hand, the comparison group is consistently increasing their average, especially from 2013 on. This year of 2013 shows to be a lower peak for the average weight for both groups, further studies should be conducted to investigate this phenomenon.

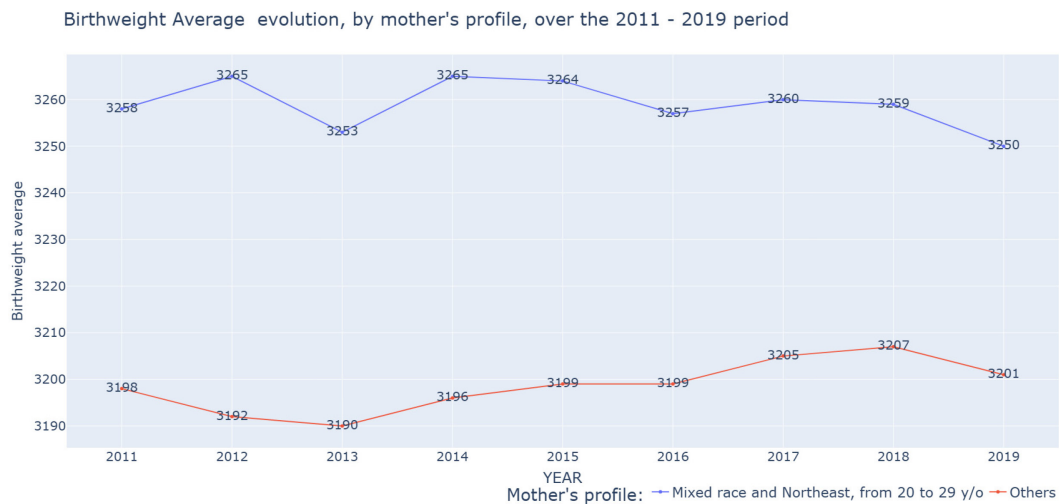


*Figure 18: Evolution of birthweight by mother's profile*

For the whole period, the results show that newborns delivered from the average mother presented an average weight of 3259g, a value higher in relation to the comparison group, which obtained an average of 3198g. This difference of approximately 60g suggests that the mother profile has a marginal influence on intrauterine growth.

It is observed when detailing each dimension separately that: children from mixed-race mothers weigh in average 3214g, in contrast with children from other races/colors, whose average weight is 3184g; in the regional dimension, newborns in the Northeast region weigh 3235g in average, compared to 3193g in the other regions; and by the mother's age criteria, those from 20 to 29 years delivered an average of 3221g, while the remaining age groups calculates an 3189g average.

When composing an “average Brazilian mother profile”: Mixed-race, from the Northeast and with age between 20 and 29 years old, this study aimed to portray a predominant configuration of gestations in Brazil during the period in question. When the birthweight of children from those mothers is compared to the remaining children, some differences can be observed that reflect biological and social factors.

Newborns from that group present an average weight close to the national average, yet slightly higher. This result can be associated with historical and structural conditions that characterize the Northeastern region, like socioeconomical and cultural

inequalities. At a first and shallow view, it may seem counter intuitive that the babies born in a region with lower Human Development Index are delivered heavier. But, 59.6% of northeastern population self-declared as mixed-race in IBGE's 2022 census. Moreover, most mothers in this group are aged between 20 to 29 years, a biologically favorable period for gestation, this factor may partially counterbalance the observed inequalities.

Therefore, this average mother profile constitutes a demographic portrait of the country with contradicting factors: while the mother's age may be considered closer to the biological ideal, other vulnerability factors are presented in regional and ethnical-racial inequalities. A comparative analysis of birthweight shows that while this group is close to the national average, it still carries structural marks that must be considered when formulating public health policies for infants and their mothers.

## 5.2 Updated Gestational Age to birthweight percentile table

The filters applied by PEDREIRA et al. (2011) puts a cap to the 27<sup>th</sup> and 31<sup>st</sup> weeks, that limit was not noticeable because the weekly curves were smoothed by mathematical functions applied to gestational age groups. But as this study uses the finer grain of the weekly data the jump from the highest weight percentiles from the 27<sup>th</sup> to 28<sup>th</sup> weeks and from the 31<sup>st</sup> to 32<sup>nd</sup> weeks it is noticeable. This reveals a necessity for weekly weight filters.

*Table 16: Weight chart using the same filters as PEDREIRA et al. (2011)*

| Gestational Age (Weeks) | Percentiles     |                 |                  |                  |                  |                  |                  |                  |                  |
|-------------------------|-----------------|-----------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                         | 3 <sup>rd</sup> | 5 <sup>th</sup> | 10 <sup>th</sup> | 25 <sup>th</sup> | 50 <sup>th</sup> | 75 <sup>th</sup> | 90 <sup>th</sup> | 95 <sup>th</sup> | 97 <sup>th</sup> |
| 26                      | 470             | 505             | 575              | 700              | 840              | 950              | 1050             | 1115             | 1155             |
| 27                      | 510             | 560             | 630              | 765              | 925              | 1050             | 1145             | 1190             | 1210             |
| 28                      | 590             | 640             | 730              | 895              | 1085             | 1260             | 1460             | 1655             | 1790             |
| 29                      | 665             | 725             | 825              | 1000             | 1210             | 1400             | 1595             | 1735             | 1828             |
| 30                      | 780             | 850             | 960              | 1150             | 1380             | 1580             | 1756             | 1850             | 1905             |
| 31                      | 874             | 950             | 1080             | 1290             | 1525             | 1725             | 1868             | 1930             | 1960             |
| 32                      | 1125            | 1230            | 1390             | 1700             | 2125             | 2975             | 3420             | 3655             | 3800             |
| 33                      | 1320            | 1430            | 1610             | 1925             | 2350             | 3010             | 3435             | 3655             | 3800             |

Table 17 presents the final deliverable of this study. It is an updated table of Gestational Age, birthweight percentiles. Differently than PEDREIRA et al. (2011), that had to apply interpolation formulas to calculate each of the percentiles to simulate each gestational age, this study could use the real gestational week numbers to calculate every percentile.

This table reveals some interesting behaviors: Male babies are mostly heavier than female babies; the only exceptions are the 90<sup>th</sup> percentiles and above for the 22<sup>nd</sup>, 23<sup>rd</sup> weeks; the 97<sup>th</sup> percentile for the 30<sup>st</sup> week; and the 95<sup>th</sup> and 97<sup>th</sup> percentile for the

31<sup>st</sup>, 32<sup>nd</sup>, and 34<sup>th</sup> weeks. As the number of samples for those weeks is low – especially the 22<sup>nd</sup> and 23<sup>rd</sup> weeks, there might be a sample bias happening.

The newborn weight in every percentile increases for every gestational week, which is an expected behavior as the fetus is continuously being nourished during their pregnancy. The exception happens after the 41<sup>st</sup> week, as the newborns average weight is lighter in the 42<sup>nd</sup> and then even lighter at the 43<sup>rd</sup> week. This could mean that the delivery of smaller babies might be postponed from the on-term period so they can grow inside the uterus for more weeks before birth.

*Table 17: Updated birthweight to gestational age, singleton Brazilians from 2011 to 2019*

| Gestational Age (Weeks) | Percentiles     |      |                 |      |                  |      |                  |      |                  |      |                  |      |                  |      |                  |      |                  |      |
|-------------------------|-----------------|------|-----------------|------|------------------|------|------------------|------|------------------|------|------------------|------|------------------|------|------------------|------|------------------|------|
|                         | 3 <sup>rd</sup> |      | 5 <sup>th</sup> |      | 10 <sup>th</sup> |      | 25 <sup>th</sup> |      | 50 <sup>th</sup> |      | 75 <sup>th</sup> |      | 90 <sup>th</sup> |      | 95 <sup>th</sup> |      | 97 <sup>th</sup> |      |
|                         | Male            | Fem  | Male            | Fem  | Male             | Fem  | Male             | Fem  | Male             | Fem  | Male             | Fem  | Male             | Fem  | Male             | Fem  | Male             | Fem  |
| 22                      | 360             | 340  | 390             | 365  | 425              | 400  | 475              | 450  | 532              | 510  | 612              | 595  | 787              | 800  | 940              | 944  | 1020             | 1060 |
| 23                      | 415             | 380  | 445             | 410  | 485              | 450  | 540              | 508  | 600              | 570  | 680              | 648  | 795              | 800  | 977              | 1000 | 1102             | 1115 |
| 24                      | 440             | 400  | 480             | 440  | 530              | 500  | 605              | 574  | 686              | 650  | 775              | 735  | 910              | 865  | 1100             | 1030 | 1233             | 1170 |
| 25                      | 465             | 420  | 505             | 464  | 570              | 520  | 675              | 625  | 780              | 728  | 885              | 830  | 1020             | 969  | 1190             | 1150 | 1336             | 1300 |
| 26                      | 495             | 460  | 535             | 499  | 610              | 560  | 740              | 680  | 880              | 825  | 1000             | 950  | 1160             | 1105 | 1350             | 1290 | 1500             | 1450 |
| 27                      | 540             | 505  | 595             | 550  | 675              | 620  | 830              | 760  | 990              | 930  | 1140             | 1090 | 1340             | 1284 | 1565             | 1525 | 1740             | 1710 |
| 28                      | 610             | 565  | 670             | 620  | 770              | 700  | 935              | 855  | 1120             | 1050 | 1300             | 1230 | 1500             | 1460 | 1740             | 1705 | 1875             | 1850 |
| 29                      | 695             | 650  | 760             | 700  | 865              | 800  | 1050             | 975  | 1260             | 1195 | 1465             | 1400 | 1720             | 1680 | 1965             | 1970 | 2100             | 2100 |
| 30                      | 825             | 780  | 900             | 840  | 1015             | 950  | 1225             | 1145 | 1465             | 1390 | 1705             | 1650 | 2086             | 2105 | 2350             | 2400 | 2460             | 2492 |
| 31                      | 950             | 890  | 1030            | 970  | 1170             | 1100 | 1410             | 1330 | 1680             | 1618 | 1995             | 1990 | 2480             | 2510 | 2635             | 2650 | 2700             | 2705 |
| 32                      | 1105            | 1050 | 1200            | 1146 | 1355             | 1295 | 1630             | 1560 | 1925             | 1878 | 2334             | 2375 | 2775             | 2800 | 2900             | 2902 | 2940             | 2950 |
| 33                      | 1322            | 1260 | 1430            | 1360 | 1605             | 1530 | 1900             | 1830 | 2235             | 2192 | 2740             | 2750 | 3085             | 3079 | 3196             | 3190 | 3240             | 3230 |
| 34                      | 1550            | 1480 | 1670            | 1600 | 1860             | 1790 | 2165             | 2095 | 2520             | 2470 | 3000             | 2960 | 3325             | 3280 | 3450             | 3420 | 3500             | 3486 |
| 35                      | 1780            | 1700 | 1910            | 1835 | 2100             | 2025 | 2395             | 2315 | 2735             | 2665 | 3130             | 3060 | 3450             | 3386 | 3600             | 3540 | 3670             | 3625 |
| 36                      | 2010            | 1930 | 2140            | 2060 | 2320             | 2240 | 2600             | 2510 | 2915             | 2815 | 3260             | 3150 | 3600             | 3495 | 3805             | 3700 | 3950             | 3835 |
| 37                      | 2250            | 2170 | 2372            | 2285 | 2545             | 2450 | 2805             | 2700 | 3090             | 2975 | 3390             | 3270 | 3686             | 3560 | 3880             | 3755 | 4005             | 3895 |
| 38                      | 2470            | 2385 | 2580            | 2490 | 2735             | 2635 | 2975             | 2865 | 3240             | 3120 | 3515             | 3390 | 3790             | 3655 | 3965             | 3830 | 4090             | 3950 |
| 39                      | 2585            | 2500 | 2695            | 2600 | 2842             | 2745 | 3080             | 2970 | 3340             | 3220 | 3615             | 3490 | 3890             | 3750 | 4060             | 3920 | 4180             | 4030 |
| 40                      | 2650            | 2555 | 2758            | 2660 | 2915             | 2805 | 3160             | 3045 | 3435             | 3305 | 3720             | 3590 | 4000             | 3855 | 4178             | 4025 | 4300             | 4150 |
| 41                      | 2645            | 2555 | 2760            | 2665 | 2930             | 2820 | 3190             | 3070 | 3470             | 3340 | 3770             | 3630 | 4055             | 3905 | 4240             | 4080 | 4365             | 4200 |
| 42                      | 2570            | 2494 | 2700            | 2600 | 2870             | 2770 | 3130             | 3015 | 3420             | 3300 | 3730             | 3600 | 4015             | 3880 | 4205             | 4060 | 4345             | 4195 |
| 43                      | 2550            | 2465 | 2680            | 2584 | 2850             | 2745 | 3110             | 3000 | 3400             | 3275 | 3700             | 3560 | 4000             | 3850 | 4185             | 4020 | 4300             | 4150 |
| 44                      | 2540            | 2450 | 2670            | 2580 | 2850             | 2750 | 3120             | 3000 | 3400             | 3280 | 3705             | 3565 | 4000             | 3850 | 4190             | 4020 | 4310             | 4150 |

Equation 2 is used to compare this study's results with the percentile published by PEDREIRA et al. (2011).

$$RPD = \frac{(PEREIRA2025_{perc} - PEDREIRA2011_{perc}) * 100}{PEREIRA2025_{perc}}$$

Equation 2: Relative percentual difference calculation

In this formula,  $PEREIRA2025_{perc}$  represents this study's percentile values, and  $PEDREIRA2011_{perc}$  are the percentiles published by PEDREIRA et al. (2011).

Table 18 shows a comparison of relative percentual differences for the 10<sup>th</sup>, 50<sup>th</sup> and 90<sup>th</sup> percentiles for male newborn between PEDREIRA et al. (2011), used as a reference, and the current study. A negative percentual in the percentual column means that the respective percentile is smaller when compared to the results of this study, and a positive percentual means that the respective percentile is larger in this study.

Table 18: Comparison between PEDREIRA et al. (2011) and the current study

| Gestational Week | 10 <sup>th</sup> Percentile |         |       | 50 <sup>th</sup> Percentile |         |        | 90 <sup>th</sup> Percentile |         |        |
|------------------|-----------------------------|---------|-------|-----------------------------|---------|--------|-----------------------------|---------|--------|
|                  | Pedreira                    | Pereira | %     | Pedreira                    | Pereira | %      | Pedreira                    | Pereira | %      |
| 22               | 388                         | 425     | 8.7%  | 576                         | 532     | -8.3%  | 827                         | 787     | -5.1%  |
| 23               | 451                         | 485     | 7.0%  | 651                         | 600     | -8.5%  | 925                         | 795     | -16.4% |
| 24               | 517                         | 530     | 2.5%  | 735                         | 686     | -7.1%  | 1,033                       | 910     | -13.5% |
| 25               | 582                         | 570     | -2.1% | 826                         | 780     | -5.9%  | 1,169                       | 1,020   | -14.6% |
| 26               | 642                         | 610     | -5.2% | 922                         | 880     | -4.8%  | 1,270                       | 1,160   | -9.5%  |
| 27               | 702                         | 675     | -4.0% | 1,026                       | 990     | -3.6%  | 1,601                       | 1,340   | -19.5% |
| 28               | 770                         | 770     | 0.0%  | 1,166                       | 1,120   | -4.1%  | 1,549                       | 1,500   | -3.3%  |
| 29               | 852                         | 865     | 1.5%  | 1,283                       | 1,260   | -1.8%  | 1,719                       | 1,720   | 0.1%   |
| 30               | 960                         | 1,015   | 5.4%  | 1,653                       | 1,465   | -12.8% | 1,933                       | 2,086   | 7.3%   |
| 31               | 1,124                       | 1,170   | 3.9%  | 1,698                       | 1,680   | -1.1%  | 2,249                       | 2,480   | 9.3%   |
| 32               | 1,327                       | 1,355   | 2.1%  | 1,982                       | 1,925   | -3.0%  | 2,615                       | 2,775   | 5.8%   |
| 33               | 1,543                       | 1,605   | 3.9%  | 2,265                       | 2,235   | -1.3%  | 2,962                       | 3,085   | 4.0%   |
| 34               | 1,750                       | 1,860   | 5.9%  | 2,500                       | 2,520   | 0.8%   | 3,220                       | 3,325   | 3.2%   |
| 35               | 1,966                       | 2,100   | 6.4%  | 2,698                       | 2,735   | 1.4%   | 3,397                       | 3,450   | 1.5%   |
| 36               | 2,205                       | 2,320   | 5.0%  | 2,888                       | 2,915   | 0.9%   | 3,548                       | 3,600   | 1.4%   |
| 37               | 2,436                       | 2,545   | 4.3%  | 3,060                       | 3,090   | 1.0%   | 3,679                       | 3,686   | 0.2%   |
| 38               | 2,628                       | 2,735   | 3.9%  | 3,202                       | 3,240   | 1.2%   | 3,795                       | 3,790   | -0.1%  |
| 39               | 2,750                       | 2,842   | 3.2%  | 3,300                       | 3,340   | 1.2%   | 3,900                       | 3,890   | -0.3%  |
| 40               | 2,923                       | 2,915   | -0.3% | 3,369                       | 3,435   | 1.9%   | 3,996                       | 4,000   | 0.1%   |
| 41               | 2,878                       | 2,930   | 1.8%  | 3,423                       | 3,470   | 1.4%   | 4,079                       | 4,055   | -0.6%  |
| 42               | 2,900                       | 2,870   | -1.0% | 3,450                       | 3,420   | -0.9%  | 4,150                       | 4,015   | -3.4%  |

The study published by PEDREIRA et al. (2011) represented a snapshot of Brazilian past conditions: babies born between 2003 and 2005. From that moment to today much has changed, but that table is still a good representation of our national population: specially for the 'on-term' gestational weeks the percentual difference of the percentiles is negligible.

## 6 Conclusion

The present study reviews the health status of newborns in Brazil from 2011 and 2019. Research Question 1 is answer is that birthweight showed a consistent pattern of increase during the investigated period; the Brazilian average mother generates newborns with an average weight consistent with the national average, but social and regional inequities still impact the numbers.

### 6.1 Final Considerations

The goal of this study is to analyze the most recent births occurred in Brazil. In general, their health indicators are improving: birthweight averages and APGAR scores are gradually increasing. There is also a change in the mother's profile: mixed race mothers are delivering babies earlier, heavier, and in more quantity than white skinned mothers. This behavior has the power to chance Brazilian racial group proportions and is observed by other studies.

In addition to that analysis, a new and updated birthweight gestational week percentile table is presented. This table is calculated using the new gestational week field that was made available in SINASC's database in 2010, which result in a finer and closer to reality numbers.

### 6.2 Risks and Limitations

Descriptive statistical analysis suffers from some limitations as described:

- **Inferences:** This kind of analysis just described what was observed, it does not indicate causes nor if the results can be applied to other populations;
- **Sample context dependency:** The results strongly depend on the sample characteristics (size, representativity, biases);
- **Sensibility to outliers and oversimplifications:** Measures of median and standard deviations might be distorted by extreme values and can hide variabilities or important patterns.
- **Lack of control over external variables:** Descriptive analyses do not isolate confusion factors nor control intervenient variables.

The validity of this study is subject to several potential risks as outlined below:

- **Compromised Internal validity** occurs when no hypothesis tests are applied in an exploratory analysis to evaluate if patterns are statistically significative or result of chance;
- **Selection bias** can occur if data is collected with a bias (i.e., by convenience);

- **Interpretation bias** is a risk when the researcher or the reader comes to casual conclusions when there is no basis for them;
- **Omission of relevant variables** when descriptive analysis does not reveal relationships between variables, crucial factors might be ignored.

### 6.3 Suggestions for future research

The defined period and context of this study is delimited by the starting year of 2011, one year after SINASC database was updated with important new fields, especially the Gestational Age; and the ending year of 2019, the year before COVID-19 pandemic started. To address this limitation and enhance the findings, future research could incorporate more recent years into the database.

This would allow research to compare if there was a change in the quantity and quality of newborns during the pandemic period. If there was a significant reduction in the number of newborns during the pandemic period, a future study could model the preceding years and extrapolate it using Machine Learning methods them to predict the number of new Brazilians that would be born if the pandemic has never happened.

Further investigation of the relation between social programs like *Bolsa Família* and *Mais Médicos* impacting the birthweight and APGAR scores in the granularity of city of birth. Analyze if the municipalities that received more attention from those programs have experienced a different evolution in birth health indicators when compared to their surrounding neighbors.

Utilize machine learning models to analyze relationships between birthweight and other various socioeconomic variables. Build classification machine learning models to categorize newborns based on their delivery evaluation and their mother's socioeconomic conditions.

## 7 References

ALEXANDER, G. R. et al. **A United States National Reference for Fetal Growth.** *Obstetrics & Gynecology*, [s.l.], v. 87, n° 2, p. 163, 1996. ISSN: 0029-7844.

ALSHIMMIRI, M. M. et al. **Birthweight percentiles by gestational age in Kuwait.** *Archives of Gynecology and Obstetrics*, [s.l.], v. 269, n° 2, p. 111–116, 2004. ISSN: 1432-0711, DOI: 10.1007/s00404-002-0434-0.

APGAR, V. **A proposal for a new method of evaluation of the newborn infant.** *Current Researches in Anesthesia & Analgesia*, [s.l.], v. 32, n° 4, p. 260–267, 1953. ISSN: 0099-8125.

APGAR, V et al. **Evaluation of the newborn infant; second report.** *Journal of the American Medical Association*, [s.l.], v. 168, n° 15, p. 1985–1988, 1958. ISSN: 0002-9955, DOI: 10.1001/jama.1958.03000150027007.

AZEVEDO, A. L. M. dos S. **IBGE - Educa | Jovens.** *IBGE Educa Jovens*. [s.d.]. Disponível em: <<https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18318-piramide-etaria.html>>. Acesso em: 23/jun./25.

BONELLIE, S. et al. **Centile charts for birthweight for gestational age for Scottish singleton births.** *BMC Pregnancy and Childbirth*, [s.l.], v. 8, n° 1, p. 5, 2008. ISSN: 1471-2393, DOI: 10.1186/1471-2393-8-5.

DAI, L. et al. **Birth Weight Reference Percentiles for Chinese.** *PLOS ONE*, [s.l.], v. 9, n° 8, p. e104779, 2014. ISSN: 1932-6203, DOI: 10.1371/journal.pone.0104779.

DEMŠAR, J. et al. **Orange: Data Mining Toolbox in Python.** *Journal of Machine Learning Research*, [s.l.], v. 14, n° 71, p. 2349–2353, 2013. ISSN: 1533-7928.

HARDER, T. et al. **Birth Weight, Early Weight Gain, and Subsequent Risk of Type 1 Diabetes: Systematic Review and Meta-Analysis.** *American Journal of Epidemiology*, [s.l.], v. 169, n° 12, p. 1428–1436, 2009. ISSN: 0002-9262, DOI: 10.1093/aje/kwp065.

KRAMER, M. S. et al. **A New and Improved Population-Based Canadian Reference for Birth Weight for Gestational Age.** *Pediatrics*, [s.l.], v. 108, n° 2, p. e35, 2001. ISSN: 0031-4005, DOI: 10.1542/peds.108.2.e35.

LUBCHENCO, L. O. et al. **INTRAUTERINE GROWTH AS ESTIMATED FROM LIVEBORN BIRTH-WEIGHT DATA AT 24 TO 42 WEEKS OF GESTATION.**

*Pediatrics*, [s.l.], v. 32, n° 5, p. 793–800, 1963. ISSN: 0031-4005, DOI: 10.1542/peds.32.5.793.

PATHIRANA, J. et al. **Neonatal death: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data.** *Vaccine*, [s.l.], v. 34, n° 49, p. 6027–6037, 2016. ISSN: 0264-410X, DOI: 10.1016/j.vaccine.2016.03.040.

PEDREIRA, C. E. et al. **Birth weight patterns by gestational age in Brazil.** *Anais da Academia Brasileira de Ciências*, [s.l.], v. 83, p. 619–625, 2011. ISSN: 0001-3765, 1678-2690, DOI: 10.1590/S0001-37652011005000008.

SALEH GARGARI, S. et al. **Normalized Weight of Iranian Newborns According to Gestational Age: A Nationwide Study.** *Archives of Iranian Medicine*, [s.l.], v. 20, n° 10, p. 626–632, 2017. ISSN: 1735-3947.

SZWARCWALD, C. L. et al. **Infant mortality estimation in Brazil: what do Ministry of Health data on deaths and live births say?** *Cadernos de Saúde Pública*, [s.l.], v. 18, p. 1725–1736, 2002. ISSN: 0102-311X, 1678-4464, DOI: 10.1590/S0102-311X2002000600027.

SZWARCWALD, C. L. et al. **Avaliação das informações do Sistema de Informações sobre Nascidos Vivos (SINASC), Brasil.** *Cadernos de Saúde Pública*, [s.l.], v. 35, p. e00214918, 2019. ISSN: 0102-311X, 1678-4464, DOI: 10.1590/0102-311X00214918.

UNICEF. **Levels and Trends in Child Mortality 2014.** *UNICEF DATA*. 2014. Disponível em: <<https://data.unicef.org/resources/levels-trends-child-mortality-report-2014/>>. Acesso em: 09/jan./24.

WORLD HEALTH ORGANIZATION. **Neonatal and perinatal mortality : country, regional and global estimates.** Geneva, 2006. ISSN: 9241563206.

YAO, F. et al. **New birthweight percentiles by sex and gestational age in Southern China and its comparison with the INTERGROWTH-21st Standard.** *Scientific Reports*, [s.l.], v. 8, n° 1, p. 7567, 2018. ISSN: 2045-2322, DOI: 10.1038/s41598-018-25744-7.