ampliasaúde
observatório do período perinatal

Exploration Trails:
Data visualization methodology

The **Exploration Trails** are an interactive and customized tool that offers exploratory visualization, download, and sharing of maternal and neonatal health data using pollution data for 5,565 Brazilian municipalities over seven years (2012-2019).

SUMMARY

REALIZATION

UFRJ LABVIS FIOCRUZ

FINANCING

BILL & MELINDA GATES foundation CNPq FAPERJ

SUPPORT

PCDaS

THE OBJECTIVE OF THE AMPLIA SAÚDE (EXTENDED HEALTH) PROJECT

Maternal, neonatal, and children under-five mortality has been reduced in Brazil over the last few years, mainly by means of public health policies specific to the mother-infant axis within the Sustainable Development Goal (SDG) 3. A healthy pregnancy, birth care, sufficient breastfeeding, and timely vaccination favor children's growth and development. The fetus developing environment is essential for the child's onward life. It is mainly affected by social determinants and the health system (Figure 1), that is, demographic factors and the maternal socioeconomic conditions that make them more or less vulnerable, besides the access to specialized care, prenatal care, and vaccination. Environmental conditions consider only those in the immediate surroundings, such as access to drinking water, sewage treatment, and garbage collection. However, extreme weather conditions, environmental disasters, and pollution can directly affect human health as well as the availability and access to health services. Climate and environmental determinants are increasingly frequent, and their impact must be measured in maternal-neonatal health. Based on the National Health System (SUS) principle of comprehensiveness, the Amplia Saúde (Extended Health) project aimed to expand the analysis model of maternal and neonatal morbimortality indicators by integrating air pollution-related indicators.

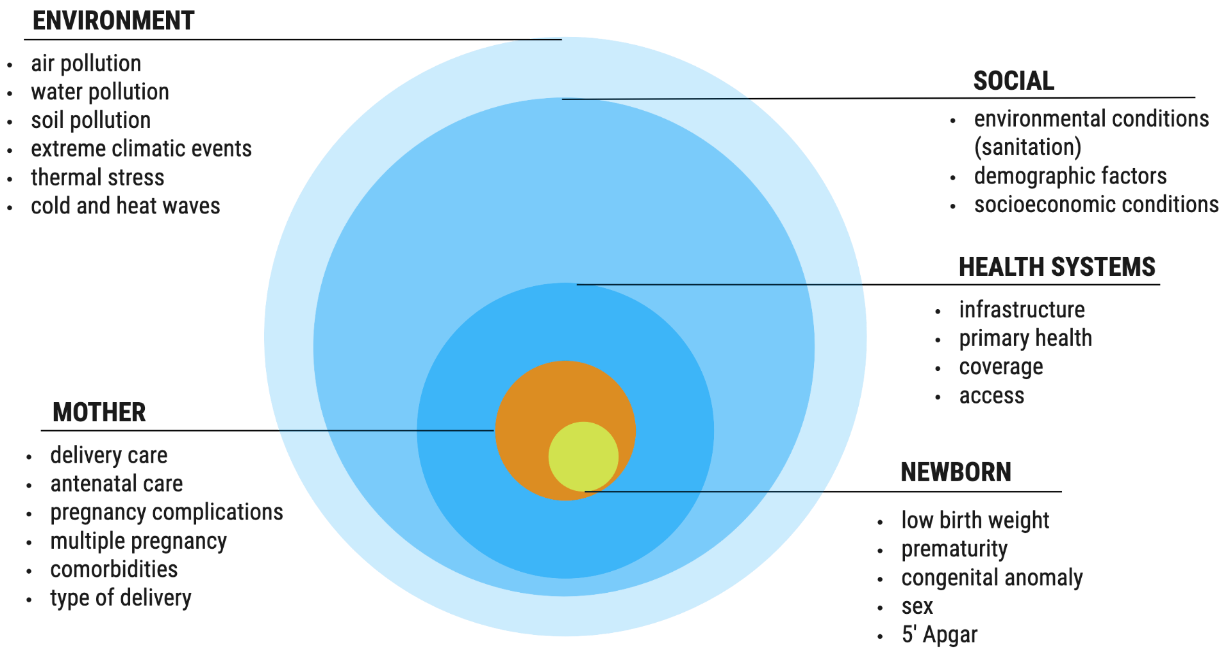

Figure 1. The conceptual framework for analyzing neonatal morbimortality

REALIZATION
UFRJ
LAB VIS
FIOCRUZ
FINANCING
BILL & MELINDA GATES foundation
CNPq
FAPERJ
SUPPORT
PCDaS

THE TOOL'S SPECIFIC OBJECTIVE

- Offer an **innovative visual analysis approach** of maternal and infant health data from the perinatal period, considering environmental factors (air pollution)
- Create an **interactive tool that allows exploring** this relationship

DATA SOURCES

The visualization tool uses data made openly available by the following government agencies:

- DATASUS, Department of Informatics of the Unified Health System of Brazil - Ministry of Health
- The “Queimadas” (Fires) Program of the National Institute for Space Research (INPE)
- The Brazilian Institute of Geography and Statistics (IBGE)

Infant and fetal death surveillance has been mandatory in public and private health services of the Brazilian Unified Health System (SUS) since 2010 by decree of the Ministry of Health. Thus, we specifically adopted data from DATASUS Mortality Information System (SIM) and Live Births Information System (SINASC) from 2012 onwards. Besides mortality data, we also employed the Fetal Death Declaration (DOFET-SIM) databases, which were extracted, processed, and enriched by the PCDaS (Platform of Data Science applied to Health, the Institute of Scientific and Technological Communication and Information in Health from the Oswaldo Cruz Foundation-ICICT), using ETL (Extract, Transform, Load) methods.

We used INPE’s Environmental Information System Integrated to Health (SISAM), developed through a partnership between INPE, PAHO/WHO (the Pan American Health Organization/World Health Organization), and FUNDEP (Foundation for Research Development). We used data on 2.5 micron air pollutants, (PM2.5) fine particulate matter, from 2012 to 2019, the last year available in the database. While other data, such as ozone concentration, carbon monoxide, and wind speed and direction, were available, they lacked the necessary consistency and were thus abandoned after the first tests. SISAM's data were downloaded by an automatic data collector (web scraping crawler). The database offers up to four daily PM2.5 measurements, although variations were identified in different municipalities. Thus, we opted for a weekly aggregation of pollution data

We employed the IBGE's 2010 Human Development Index (HDI), the last census carried out so far (2022), and the geolocation of the municipalities.

Only aggregated secondary data were used throughout the *Amplia Saúde* project, with no possibility of individual identification, including no common field that could link the events in the SINASC and SIM databases.

RELEVANT INFORMATION ON DATA RELATIONSHIP

The data tool developed within the *Amplia Saúde* project comprises a set of exploratory visualizations that combine a vast perinatal data volume with socioeconomic, geographic, and air pollution incidence attributes. The exploratory visualizations can identify data patterns, offer insights, and formulate or test hypotheses.

Since they gather and combine heterogeneous and large-scale data (Big Data) retrieved from unlinked databases, these visualizations are not indicated to seek direct cause-and-effect relationships between air pollution and health outcomes for newborns, requiring further studies that substantiate the identified correlations.

THE PROJECT'S ASSUMPTIONS

As it addresses large volumes of data collected over several years in different formats from diverse locations, we had to conduct a preliminary survey of the available databases and verify their quality, besides possible aggregation and interaction, creation of indicators, or use of existing indicators. In this initial stage, a diagram was created on the Miro platform (Figure 2), where data from sources related to maternal-neonatal health and environmental pollution were listed. Subsequently, some of the selected databases were not employed, as was the case with the National Registry of Health Facilities (CNES) and the Hospital and Outpatient Information System (SIH).

The Tool provides exploratory, on-demand, and detailed information. It is a layer aimed at those seeking to perform specific queries, formulate or test hypotheses, and make data-driven decisions. Therefore, we had to create a customized architecture for the tool (Figure 3). We adopted an architecture where the data is stored in and retrieved from a cloud server to generate interactive visualizations. However, the visualizations themselves are built by code running in the user's browser. An additional layer was introduced in the system architecture (labeled "middleware" in Figure 3). It allows compressing the data sent over the network from the server to the client and faster access to data to work around performance issues typical of large data visualization systems.

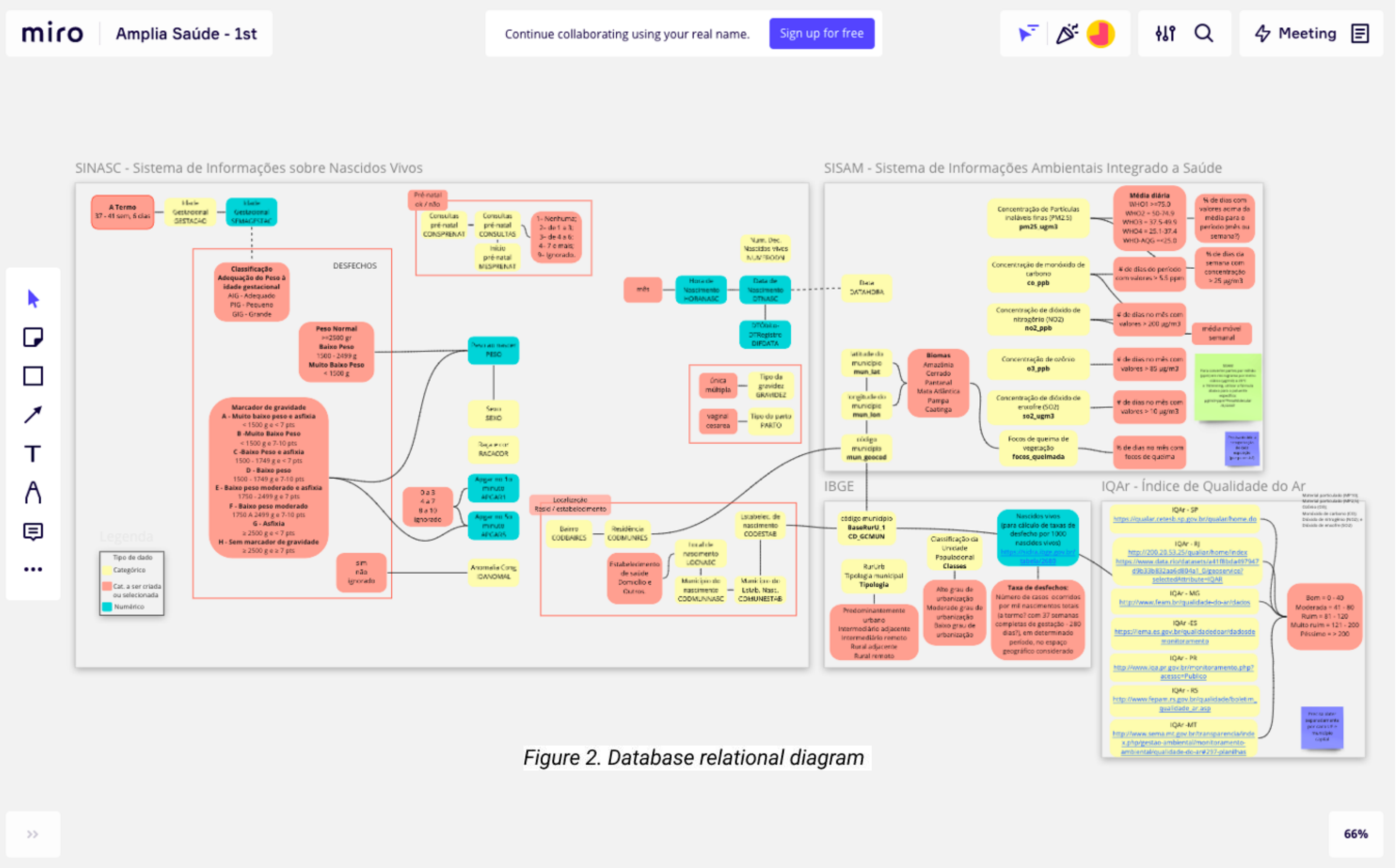
miro
Ampla Saúde - 1st
Continue collaborating using your real name.
Sign up for free
Meeting
SINASC - Sistema de Informações sobre Nascidos Vivos
A Termo 37 - 41 sem, 6 dias
Idade Gestacional GESTACAO
Idade Gestacional SEMAGESTAC
DESFECHOS
Classificação Adequação do Peso à idade gestacional AIG - Adequado PIG - Pequeno GIG - Grande
Peso Normal >=2500 gr Baixo Peso 1500 - 2499 g Muito Baixo Peso < 1500 g
Marcador de gravidade A - Muito baixo peso e asfixia < 1500 g e < 7 pts B -Muito Baixo Peso < 1500 g e 7-10 pts C -Baixo Peso e asfixia 1500 - 1749 g e < 7 pts D - Baixo peso 1500 - 1749 g e 7-10 pts E - Baixo peso moderado e asfixia 1750 - 2499 g e 7 pts F - Baixo peso moderado 1750 A 2499 g e 7-10 pts G - Asfixia ≥ 2500 g e < 7 pts H - Sem marcador de gravidade ≥ 2500 g e ≥ 7 pts
sim não ignorado
Pré-natal ok / não
Consultas pré-natal CONSPRENAT
Consultas pré-natal CONSULTAS
Início pré-natal MESPRENAT
1- Nenhuma; 2- de 1 a 3; 3- de 4 a 6; 4- 7 e mais; 9- Ignorado.
Peso ao nascer PESO
Sexo SEXO
Raça e cor RACACOR
0 a 3 4 a 7 8 a 10 Ignorado
Apgar no 1o minuto APGAR1
Apgar no 5o minuto APGAR5
Anomalia Cong. IDANOMAL
mês
Hora de Nascimento HORANASC
Data de Nascimento DTNASC
DTÓbito-DTRegistro DIFDATA
Num. Dec. Nascidos vivos NUMERODN
única múltipla
Tipo da gravidez GRAVIDEZ
vaginal cesarea
Tipo do parto PARTO
Localização Resid / estabelecimento
Bairro CODBAIRES
Residência CODMUNRES
Estabelecimento de saúde Domicílio e Outros
Local de nascimento LOCNASC
Município do nascimento CODMUNNASC
Estabelec. de nascimento CODESTAB
Município do Estab. Nasc. CODMUNESTAB
Legenda
Tipo de dado
Categórico
Cat. a ser criada ou selecionada
Numérico
SISAM - Sistema de Informações Ambientais Integrado a Saúde
Data DATAHORA
latitude do município mun_lat
longitude do município mun_lon
código município mun_geocod
Biomas Amazônia Cerrado Pantanal Mata Atlântica Pampa Caatinga
Concentração de Partículas inaláveis finas (PM2.5) pm25_ugm3
Concentração de monóxido de carbono co_ppb
Concentração de dióxido de nitrogênio (NO2) no2_ppb
Concentração de ozônio o3_ppb
Concentração de dióxido de enxofre (SO2) so2_ugm3
Focos de queima de vegetação focos_queimada
Média diária WHO1 >=75.0 WHO2 = 50-74.9 WHO3 = 37.5-49.9 WHO4 = 25.1-37.4 WHO-AQG =<25.0
% de dias com valores acima da média para o período (mês ou semana?)
# de dias do período com valores > 5.5 ppm
% de dias da semana com concentração > 25 µg/m3
# de dias no mês com valores > 200 µg/m3
média móvel semanal
# de dias no mês com valores > 85 µg/m3
# de dias no mês com valores > 10 µg/m3
% de dias no mês com focos de queima
IBGE
código município BaseRurU_1 CD_GCMUN
RurUrb Tipologia municipal Tipologia
Predominantemente urbano Intermediário adjacente Intermediário remoto Rural adjacente Rural remoto
Classificação da Unidade Populacional Classes
Alto grau de urbanização Moderado grau de urbanização Baixo grau de urbanização
Nascidos vivos (para cálculo de taxas de desfecho por 1000 nascidos vivos) https://sidra.ibge.gov.br/tabela/2680
Taxa de desfechos: Número de casos ocorridos por mil nascimentos totais (a termo? com 37 semanas completas de gestação - 280 dias?), em determinado período, no espaço geográfico considerado
IQAr - Índice de Qualidade do Ar
IQAr - SP https://qualar.cetesb.sp.gov.br/qualar/home.do
IQAr - RJ http://200.20.53.25/qualiar/home/index https://www.data.rio/datasets/a41ff8bda497947d9b33b832a6d804a1_0/geoservice?selectedAttribute=IQAR
IQAr - MG http://www.feam.br/qualidade-do-ar/dados
IQAr -ES https://iema.es.gov.br/qualidadedoar/dadosdemonitoramento
IQAr - PR http://www.iap.pr.gov.br/monitoramento.php?acesso=Publico
IQAr - RS http://www.fepam.rs.gov.br/qualidade/boletim_qualidade_ar.asp
IQAr -MT http://www.sema.mt.gov.br/transparencia/index.php/gestao-ambiental/monitoramento-ambiental/qualidade-do-ar#297-planilhas
Bom = 0 - 40 Moderada = 41 - 80 Ruim = 81 - 120 Muito ruim = 121 - 200 Péssimo = > 200
66%

Figure 2. Database relational diagram

REALIZATION
UFRJ
LAB VIS
FIOCRUZ

FINANCING
BILL & MELINDA GATES foundation
CNPq
FAPERJ

SUPPORT
PCDaS

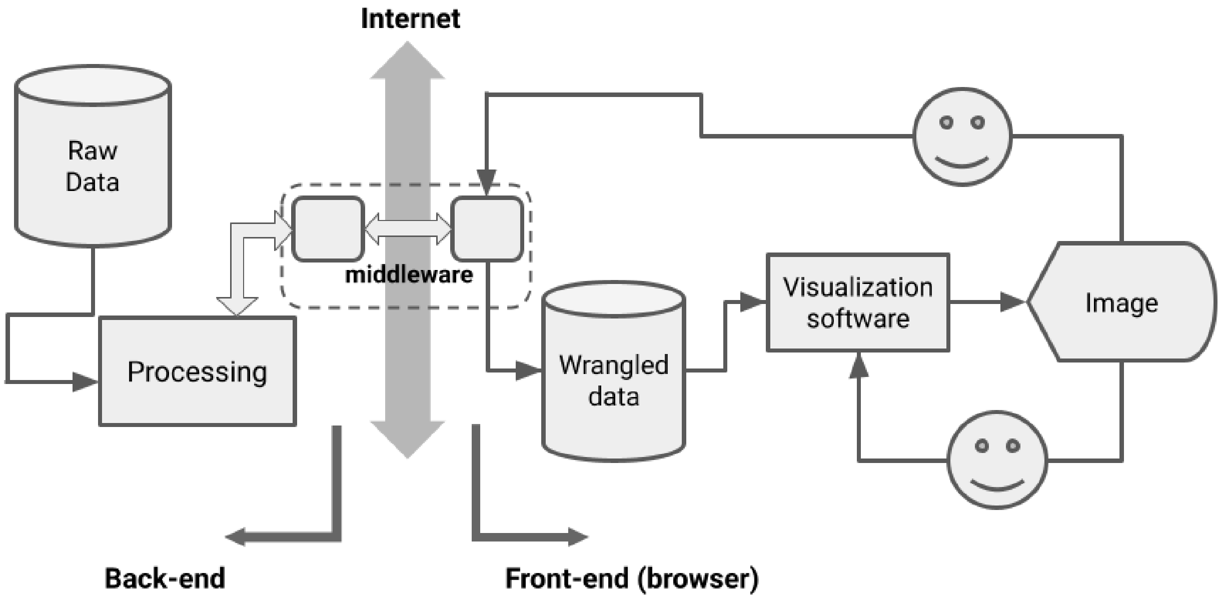

Figure 3. Data/tool system architecture

We had to make several design-related decisions in producing visualizations due to the large data volume that can be displayed in multiple ways according to choices made by the user. For instance, we limited to four the number of municipalities that can be plotted simultaneously in the Trails visualization. Similarly, when comparing municipalities with extreme pollution values, the air pollution graphs of municipalities with low pollution levels may show no variation, because the axis scale was fixed to cover the wide range of values.

Five design assumptions were determined due to the several challenges observed throughout data structuring and while creating the visualizations, which will be detailed below.

DATA AGGREGATION AND TEMPORAL RELATIONSHIP APPROXIMATION

Maternal-neonatal health data and pollution data have different temporal granularities. While the former are usually organized and analyzed by month or year, the latter are measured several times daily. We had to create a single aggregation mode to allow the simultaneous visual analysis of the two datasets (health and air pollution). Thus, all data were aggregated by epidemiological week and year, considering their use in the visualization groups of the exploratory trails and the map/scatter plot.

SELECTION OF SINGLETON BIRTHS

We decided to exclude those born from multiple pregnancies from the project database to allow a basis for comparing pregnancy periods and because fetal morbimortality increases in pregnancies with two or more fetuses in-uterus.

POLLUTION AND HUMAN HEALTH RATES

We adopted the benchmark values determined by CONAMA (National Council for the Environment) on air quality in daily measurements of particulate matter 25 µ (PM2.5) to visually indicate the pollution levels unsuitable for human health, which in turn were based on the WHO recommendations of daily (15µg/m3) and annual (5µg/m3) thresholds without prejudice to health (*World Health Organization. Air Quality Guidelines - Update 2021. Copenhagen, Denmark: WHO Regional Office for Europe*).

REORGANIZING DATA BY CONCEPTION RATE

The main challenge to visualizing and analyzing the possible effects of pollution on pregnancies and maternal-neonatal health lies in the secondary data-imposed limitation. In other words, we cannot monitor pregnancies and births individually. Women are pregnant at different gestational periods at any one time, and pregnancies are of varying lengths. Thus, the same air pollution peak period can occur in women in early pregnancy while others are in the middle or at the end of their pregnancy. Furthermore, the period of most susceptibility for the fetus is known to occur between the 22nd and 38th gestation weeks. However, we cannot observe the effect over different gestation periods when analyzing the exposure to pollution over a set of pregnancies. From these considerations, we decided to adopt an estimated conception date, which was calculated by subtracting the length of gestation in weeks from the date of birth. In the visualization, we used the term “window of susceptibility” to demarcate this period between the 22nd and 38th weeks.

HIGHLIGHTED DATA

The SIM and SINASC databases have more than one hundred attributes when considering the original attributes and those added later at enrichment. However, birth weight is a critical variable since this information is crucial for calculating several indicators, including gestational age birthweight adequacy. After establishing the relevance of the weight attribute, only cases with no missing birthweight were included. Gestational age in weeks, necessary for the calculations, was included in the SINASC database in 2009 but found to be consistent only from 2012 onwards. This date determined the onset of the period covered by the visualizations.

DEFINITION AND INDICATORS

The tool comprises two sets of visualizations: **Map** and **Trails**.

MAP

The Map visualization offers several alternatives for comparing municipalities, filtering, and allows the selection of up to four municipalities, considering the in-depth analysis to be conducted later with the Trails visualization. On the map itself, we can observe the size of the population of each municipality and the mean annual value of a primary variable chosen among six possible ones, facilitating the selection based on geographic positioning. This page also contains a scatter plot to analyze up to three variables from all available municipalities per the desired filtering. A bar chart showing the variation of the primary variable over the years for up to four municipalities is also available.

KEY INDICATORS:

- ➔ Pollution
- ➔ Mean annual pollution
- ➔ Low birthweight
- ➔ Neonatal Mortality
- ➔ Perinatal Mortality
- ➔ Infant Mortality

TRAILS

In the Trails, we can investigate the municipalities selected on the Map and optionally modify this choice. The Trails comprise three investigative sets: Live Births, Birth Characteristics, and Perinatal and Infant Mortality. All have a line graph with the data of air pollutants in parallel with other graphs related explicitly to the neonatal health data addressed. Neonatal health line graphs will vary by attributes selected in the main menu or the comparison menu.

LIVE BIRTHS

KEY INDICATORS:

- ➔ Pollutant
- ➔ Birth weight
- ➔ (Birthweight) Adequacy to gestational age
- ➔ Total live births

ATTRIBUTES:

- ➔ **Birth weight**: Normal, Low weight, Extremely low weight, Unknown
- ➔ **Gestational age**: Post-term, Late term, Full term, Early term, Moderate preterm, Very preterm, Extremely preterm, Unknown
- ➔ **Gestational age adequacy**: LGA, AGA, SGA, Unknown
- ➔ **Sex**: Female, Male, Unspecified
- ➔ **Congenital anomaly**: No, Yes
- ➔ **Mother’s ethnicity/skin color**: White, Black, Brown, Other, Unknown
- ➔ **Mother’s Schooling**: Higher Education, High School, Elementary School, Incomplete Elementary School, Illiterate, Unknown
- ➔ **Mother’s Marital Status**: Single, Married or Common-Law Marriage, Widowed, Separated or Divorced, Unknown

REALIZATION
UFRJ LABVIS FIOCRUZ
FINANCING
BILL & MELINDA GATES foundation CNPq FAPERJ
SUPPORT
PCDaS

- Mother's age: 10-14, 15-19, 20-29, 30-39, 40-49, 50+, Unknown
- Prenatal visits: 7+, 4-6, 1-3, 0, Unknown
- Delivery type: Vaginal, Cesarean, Unspecified
- Induced labor: Yes, No
- Prenatal control Adequacy: Not performed, Inadequate, Intermediate, Adequate, More than adequate, Unknown

BIRTHS CHARACTERISTICS

KEY INDICATORS:

- Pollutant
- Birth weight
- Robson groups
- Total live births

ATTRIBUTES:

- Birth weight: Normal, Low weight, Extremely low weight, Unknown
- Gestational age: Post-term, Late term, Full term, Early term, Moderate preterm, Very preterm, Extremely preterm, Unknown
- Adequacy (of birthweight) to Gestational age: LGA (large for gestational age), AGA (adequate for gestational age), SGA (small for gestational age), Unknown
- Sex: Female, Male, Unspecified
- Congenital anomaly: No, Yes
- Mother's ethnicity/skin color: White, Black, Brown, Other, Unknown
- Mother's Schooling: Higher Education, High School, Elementary School, Incomplete Elementary School, Illiterate, Unknown
- Mother's Marital Status: Single, Married or Common-Law Marriage, Widowed, Separated or Divorced, Unknown
- Mother's age: 10-14, 15-19, 20-29, 30-39, 40-49, 50+, Unknown
- Apgar 5th Minute: Good vitality, Moderate asphyxia, Severe asphyxia, Unknow
- Delivery type: Vaginal, Cesarean, Unspecified
- Induced labor: Yes, No
- Prenatal Adequacy: Not performed, Inadequate, Intermediate, Adequate, More than adequate, Unknown

PERINATAL AND INFANT MORTALITY

KEY INDICATORS:

- Pollutant
- Fetal deaths
- Deaths versus delivery
- Total deaths

ATTRIBUTES:

- Birth weight: Normal, Low weight, Extremely low weight, Unknown
- Sex: Female, Male, Unspecified
- Mother's ethnicity/skin color: White, Black, Brown, Other, Unknown
- Mother's Schooling: Higher Education, High School, Elementary School, Incomplete Elementary School, Illiterate, Unknown
- Mother's age: 10-14, 15-19, 20-29, 30-39, 40-49, 50+, Unknow
- Delivery type: Vaginal, Cesarean, Unspecified

Reference for citation:

KOSMINSKY, D.; ESPERANÇA, C.; ILLARRAMENDI, X.; GIANNELLA, J. R.; MAURO, R.; GOMES JUNIOR, S. C.; BARCELLOS, L. **Exploration Trails:** Data visualization methodology. 2023. Available on: <http://ampliasaude.org/en/metodologia/>