Perinatal and neonatal mortality analysis in Colombia

Problem definition and project overview

As of 2011, The World Health Organization (WHO for its acronym) estimates state that about 4 million neonates die every year, and nearly 41% of all under-five child deaths are among newborn infants, babies in their first 28 days of life or the neonatal period [1]. Accordingly, perinatal and neonatal mortality is now a part of the 2030 agenda for Sustainable Development of the United Nations and is also one of the topics of interest to the Colombian Ministry of Health [2].

Our aim is to carry out an analysis of different databases related to perinatal and neonatal death, as well as demographic, economic, social and geographic data for Colombia, that can help us understand which features may influence and contribute to the said death rate in different locations in the country by looking for possible correlations that may constitute the first step to the search for solutions to this global health issue.

Impact of a possible solution to the problem.

Characterization of the social, economic, demographic and geological variables that can be correlated and may have an impact on the newborn and fetal death rates is of capital importance for the localization of the most affected regions, as well as for the identification of causal relationships between said factors and, upon inclusion of larger databases, identification and prediction of new factors, possibly distribution of genetic disorders and others. Such an analysis may eventually lead to the implementation of more specialized and well-designed social programs as well as health campaigns, among other humanitarian initiatives.

Dataset description.

An entire dataset of the Colombian population was sourced from the National Administrative Department of Statistics site (DANE for its acronym in Spanish) with geographical, geospatial and demographic features as well as perinatal and neonatal information [3].The unified information gathered has more than 1’855.962 data points about births, as well as fetal and neonatal deaths, described by about 60 or more columns. This could potentially be complemented by the resource distributions through the country and region characteristics, such as water quality, electricity, health facilities and education access, as well as poverty indicators and some others, sourced from the Humanitarian Data Exchange (HDX).

Since we have several possible databases with lots of features that are not directly related to newborn infants and death rates it is challenging to dig into such an amount of information and find relationships that eventually can lead us to the most insightful ones. Moreover, It will

be really time-consuming to clean up, summarize and visualize all of that amount of data in order to show it in an organized and user-friendly way.
This dataset includes in its majority a set of categorical variables, which are counted as ordered natural numbers starting from 1 (1,2,3,...).
From this information we can highlight the following fields:

Common fields on all of the datasets (Fetal death, newborn death, alive newborn):

- **Location:** Department, City/Municipality of death or born. Categorical variables with Colombian standardized encoding called Divipola (for its acronym in Spanish of División Político-Administrativa)
- **Habitual residence**: Country, Department, City. Categorical values. Divipola encoding.
- **Gender**: Categorical variable (Male, Female, Unknown).
- **Born type**: Categorical variable (Natural, Cesarean, Instrumental, Ignored, No data)
- **Pregnancy type**: Categorical variable (Normal, Twin, Triple, Multiple, Ignored)**Pregnancy time:** Categorical variable (<22 weeks, <27 weeks, …)
- **Mother's age:** Categorical variable. Encoded in ranges of 4. (10-14,15-19,...)
- **Amount of children alive:** Discrete variable
- **Amount of children not alive:** Discrete variable

Fetal/newborn death

- **Death cause identification method:** Categorical value. (Necropsy, Medical history, Lab. tests, familiar interviews, no data)
- **Death cause:** Categorical variable. Encoded with the International Statistical Classification of Diseases and Related Health Problems, a medical classification list by the World Health Organization (CIE-10)
- **Diagnostic:** Direct, indirect, Precedent medical history, others. Categorical variable. Encoded with CIE-10
- **Probable death:** Categorical variable. (Natural, Violent, In studies)

Alive newborn

- **Weight**: Categorical variable. Encoded on ranges of 0.5 pounds (<1.000, <1.499,...)
- **Height:** Categorical variable. Encoded on ranges of 10 centimeters (<20 , <29, …)

Methods

One of the bases of this project is the data visualization which allows for the identification, comparison, summarization and explanation of the perinatal and neonatal death rates in Colombia, with the objective to find trends and behavior that can support the programs of prevention and reduction of said rate in this country.
Additionally, this project provides the analytical and visual representation tools to continue the control over this topic as new data is sourced by DANE, and support in a good way the future of the country on this matter.

Below are some of the graphics that are planned to be carried out throughout the project. These charts are designed to help recognize patterns that can contribute to the analysis of this problem:

- Maps that relate geolocation with outcomes of the pregnancies (fetal, non-fetal and neonatal data) at the departmental level in Colombia. Through this visualization, the areas and populations that may be most affected by this event will be identified.
- Charts that allow the exploration of possible correlations between the pregnancy outcome and the descriptive variables that have been classified as variables of interest.
- Interactive graphics that allow visualizing the behavior of pregnancy results in Colombia over the years.

Models

In order to find possible causes for the perinatal and neonatal death rates in Colombia, we will use the following set of techniques:

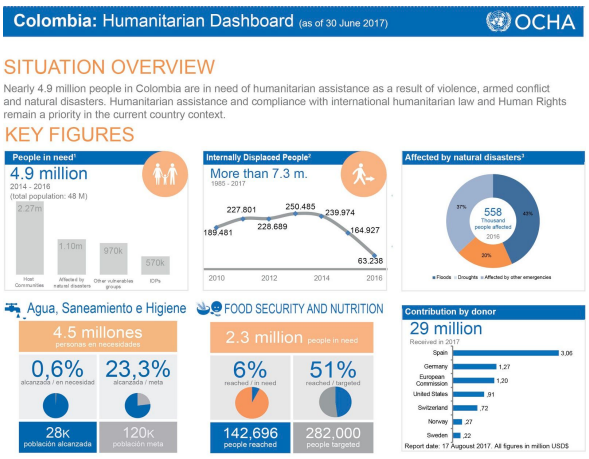
- Descriptive analysis: Descriptive analysis of the data will be used throughout the project with a view to identifying statistical values to achieve the proposed objectives.
- Georeferenced maps: Through this visualization tool and the departmental descriptive statistics, the aim is to identify and analyze different possible areas where the perinatal and neonatal death rate is high.
- Correlations: Based on the previous methods, we intend to explore possible correlations between variables that may influentiate the outcome of the pregnancy.
- Clusters analysis: With the help of clustering algorithms, it is intended to find groups of characteristics, with a view to predicting the outcome of the pregnancy.
- Decision trees: Based on its branched structure and its conditional control sentences, it will be possible to analyze the different options and consequences, with a view to predictive trends in pregnancy outcomes that depend on temporal, geographic and demographic data.
- Random forest: In case the results of the decision trees do not meet the expectations in the prediction of the pregnancy result based on different categorical variables over time, it is possible to iteratively perform said algorithm with different parameters within the same data, seeking to reduce possible noisy models, so as to generate more robust predictions.

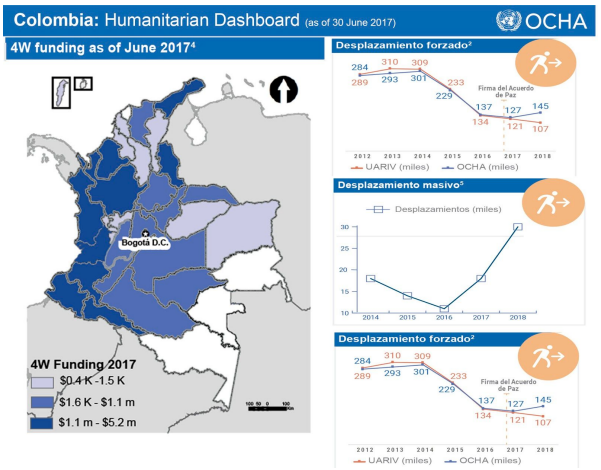
Interface

The final front-end product will feature ideally three landing pages: an overview page, an analytics page and a recommendations page. The overview page will have some summarized statistics and maybe some context about Colombia's current situation in relation to this topic. An Analytics page with a heat map showing perinatal and neonatal death rates by department (Colombia's geographical subdivisions), or even to some more geographical detail (it should be at least by department). If time permits, this heat map should allow to see the different rates interactively over the years. The second landing page will also have some visualizations of the historical data, and graphs showing possible correlations between the

pregnancy outcome and the most relevant descriptive variables. The recommendations page could have a different version depending on the outcome of our analysis, ranging from some conclusions and general or variable-specific recommendations (for example geographical recommendations), to something as sophisticated as the user seeing how the outcome (death rate in this case) is predicted to change as some relevant variables are changed.

Some examples of the sought graphics are given below (taken from [4,5]).





Milestones

Version 1: Create a dashboard containing plots and maps relating the outcomes of the pregnancies (fetal, non-fetal and neonatal data) with spatial localization at the levels of municipalities and departments for one specific year. This will be accompanied by a set of visualizations and descriptive statistics that will allow us to explore correlations between pregnancy-descriptive variables (such as mother's age, gestation time, etc) and the outcome of the pregnancy itself (born alive, fetal death, etc).

Version 2: A principal component analysis (PCA) will be performed on the dataset in order to identify the most influential variables, followed by clustering and a decision tree analysis. Subsequently, test whether these clusters are somehow correlated with geographical distribution of the cases, that is, are the clustered cases geographically close to each other?

Version 3: Enrich the original dataset so it now contains socio-demographic variables such as access to public services, poverty indicators, and others. Build interactive graphics that allow the visualization of the behavior of the pregnancy outcomes in Colombia throughout the years.

Version 4: Build predictive and decision-making models with the newly enriched dataset. The aim of these will be to predict trends in pregnancy outcomes dependent on temporal as well as geographical and demographical data. Predictions will be drawn for the upcoming years and interactive graphics will be crafted showing these, as well as the most influential variables determining the predicted behavior. A similar analysis will be performed for other countries (similar to Colombia) in order to compare what is the situation of our country, and

how it has evolved over the years with respect to the others. This last part aims for a more efficient and directed analysis of public policies in this matter for the short-to-middle term.

Timeline

Date	Deliverable	Details
Week 3	The idea should be finalized and start on scoping	The idea is finalized and data being sourced. The scope document building task was distributed as: **Pedro:** Dataset + Concerns **Andres:** Milestones + Timeline **Jenny:** Methods + Models **Angelica:** Interface **Melissa:** Methods + Models **Santiago:** General revisitings.
Week 4	project scoping complete, datasets sourced, started basic EDA.	The full final document is checked and edited by the whole team. EDA is being done by each integrant individually on the first dataset.
Week 5	datasets sourced (including those to be used for version 3). Start thinking on features from version 2.	Version 1 backend completed to >70%. Frontend is being built alongside. At this point, we will be able to decide whether we go ahead and start working on features of version 2.
Week 6	basic EDA completed. Frontend technology has been selected. Ideas for version 2 started to be implemented.	Backend for version 1 fully functional. Each integrant brings advanced and already-implemented ideas for the frontend. Decisions on the frontend are taken at this point (colors, design, interactivity, etc). Features of version 2 are being implemented already.
Week 7	Backend completed for version 1. Frontend near 70% completed. In a new branch: Version 2 backend ~40%. Start thinking about modifications to frontend.	EDA is completed for original dataset. All version 1 features have been implemented. Frontend is now in an advanced state. Version 2 features are being implemented on top of this, and modifications to frontend are being implemented to include these features.
Week 8	Final documents building started. Temptative final product completed. Start building version 3 features.	This temptative final project contains features from versions 1 and 2. In case time is not enough for version 3, this product should make a good project.
Week 9	Completion of the product (main document, full application)	The team will try to include features from version 3 into the final product, although it

		is not expected to be completed.
Week 10	New features included, application tested and documents completed.	Features will no longer be included from now on. ~30% of version 3 features are expected to be in the final product.
Week 11	Product adjustments.	Application and document details are being fixed.
Week 12	Final presentation is now being prepared.	The whole team is now working on the final presentation.

Concerns

One of the main concerns is the domain knowledge since none of the team members has a background in this field, which makes it necessary to dig into it, look for documentation or papers that explain this phenomenon and try to understand all the terminology used in this area. As a team, we have also identified a lack of information or scientific research related to this field in Colombia, this is not a well or wide-known topic in this country so we decided to gather foreign information and try to compare with the situation of our country. Another concerning point is that most of the information that we could use to enrich the main databases are very wide so cleaning, analyzing and summarizing tasks could be extremely demanding and time-consuming, thus we will eventually need to narrow a little bit the search of databases in order to avoid overwhelming among the team.

Exploratory Data Analysis (EDA)

Data Cleaning. In order to concatenate the 3 datasets, some column names were changed and values standardized across datasets. It was observed that the dataset contained death reports with information from people of all ages, so it was reduced to the data that effectively contribute to the objective analysis, that is, data associated with over-one-year-old people was discarded. Afterwards, missing information rate was computed and it was possible to identify columns with huge amount of missing or unspecified data. Figure 1 shows a table sorting the columns with the highest missing-value rate, the ones with a score of 70% or more were chosen to be analysed individually, in order to determine which ones can safely be discarded.

	Porcentaje faltantes
OCUPACION	99.978045
C_MCM1	99.828022
C_ANT32	99.758498
MAN_MUER	99.502360
CODOCUR	99.487724
C_ANT12	99.396246
C_ANT22	99.348677
C_DIR12	99.235245
C_PAT2	98.455853
C_ANT3	91.982875
C_PAT1	89.974020
FECHA_NACM	80.965275
C_ANT2	80.723773
GRU_ED1	73.182334
ULTCURPAD	71.528413
C_ANT1	69.830583

Figure 1. Missing-values rate on columns

-

A simple exploratory analysis was carried out for each variable, aimed to allow a heuristic determination of the information content. In addition, some columns were imputed based on other variables, namely, the missing values of the field related to the number of alive successful pregnancy deliveries was imputed based on the total of pregnancies and the number of born-dead children, and in the other way around to calculate the missing values of born-dead field. Furthermore, some categorical values were relabeled, since they were given as non-self-explanatory integer values, or unified to build new features; e. g. department and municipality were used to create the zip codes and individual day, month and year to create the complete date, and some other transformations.

EDA. Initially, some easily understood variables were plotted as discriminated distributions (births, fetal deaths and newborn deaths). The plots shown in figure 2 were obtained as a first approximation to the EDA by considering only a subset of the data for the sake of efficiency. Since all of those variables do not give enough information by itself, they were discriminated by our variable of interest (pregnancy outcome, figure 3); It shows already some important features of the dataset in a more detailed way; particularly, it can be seen that the weight distribution for newborns is a normal-like, and most of the fetal deaths are concentrated below 1.0 kg, which is an expected conclusion since fetal deaths don't make it to the end of the normal nine-month pregnancy period and therefore they do not reach the total of their development. Although we have considerably less data for the non-fetal deaths, we can see that the weight distribution for this group differs from that for the newborns, in that its mean is displaced to the left and has a peak below 1 kg. Weight can be thus proposed to be a good factor indicating newborn viability.

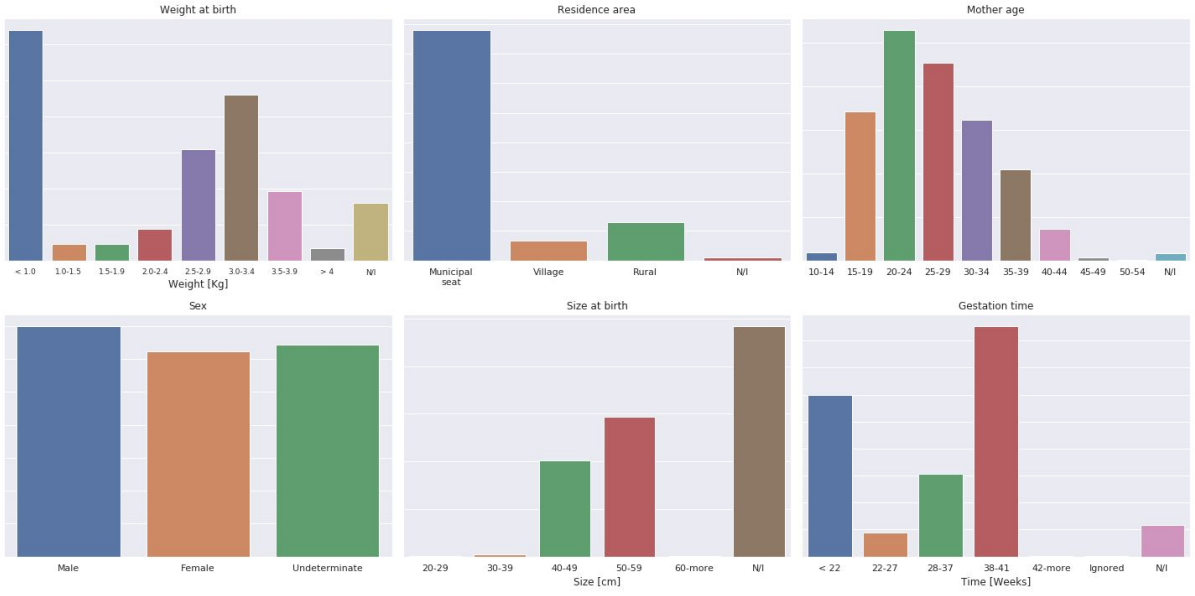


Figure 2. Distribution of some of the variables of interest.

-

Although distributions for age of the mother are similar across categories, it can be concluded that pregnancy outcome is also sensitive to this variable since the probability of a fetal death increments with mother's age, as can be seen from the respective plot, where the distribution for fetal deaths is more spread out. Some general features are that the peak is at (25-30 ?), and it is an interesting (and worrying?) situation the number of mothers in the age range 10-14, and even 15-19 years old.
Some general characteristics of mother's age distribution:

- The number of pregnancies in the age range of 10-14 is higher than wanted, as this kind of pregnancies are treated as a public health issue.
- The peak of normal distributions is found in the ages of 20-24 years old
- In the subset data used for this graph, there are more fetal deaths than live births in the age range 35-49, it would be interesting to see if this observation holds true for the complete data set. Nonetheless, this suggests a greater risk as mothers age, which matches public health information available.

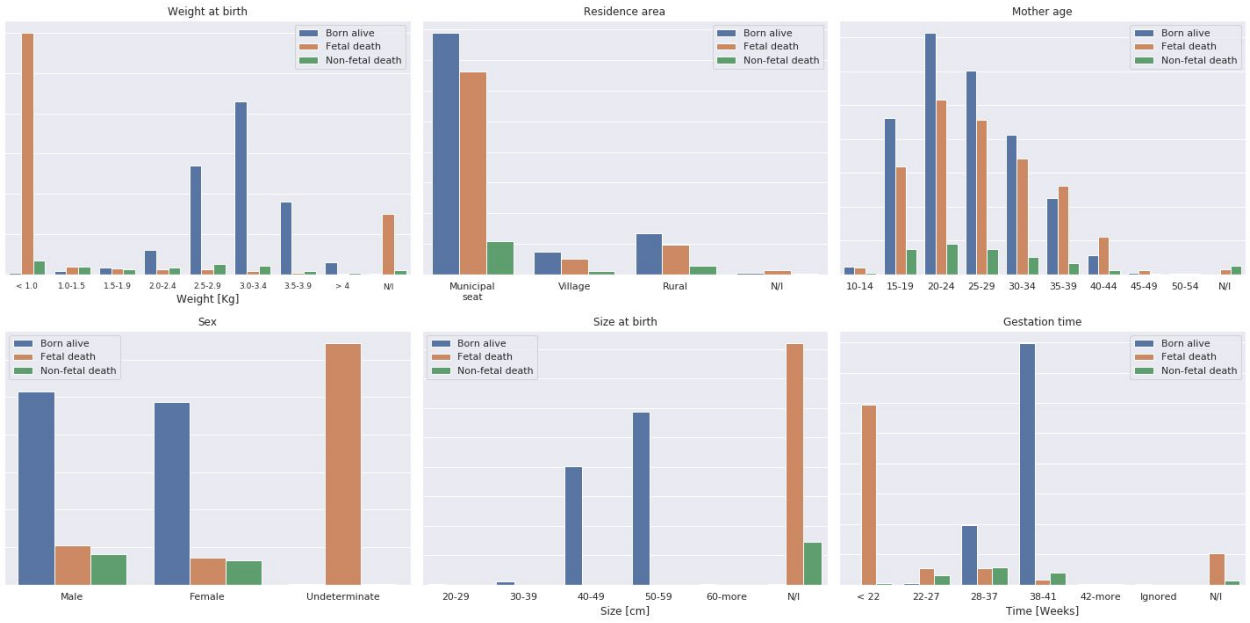


Figure 3. Distribution of some of the variables of interest, discriminated by final pregnancy outcome.

The sex and pregnancy time plots can be understood together as follows: most of the fetal deaths occur before 22 weeks, which may affect the sex determination procedure since at this stage the fetus is not yet fully developed; that way can the sex-undetermined peak in the sex plot (which is due to fetal deaths only) be understood. Additionally there does not seem to be a relationship between gender and the probability that the infant will be born alive or suffer a fetal or non-fetal death.
Finally, there does not seem to be substantial differences among distributions in residence area, that is, the fact that some mother lives in a city or in a rural location does not seem to correlate substantially with the outcome of the pregnancy.
From the graph of gestation time, it can be seen that the infant is more likely to suffer a non-fetal death if the infant is born in the 28-37 week period; the effect of this is however apparently weaker as compared to others, such as weight or mother's age, so it needs to be further explored to confirm this claim.
From the baby length graph, we can see that we only have available data for non-dead newborns.
Available variables were explored in order to formulate some questions to the dataset, which will be analysed in future analysis. These are the questions we have come up with so far:

- Is there any difference between newborns whose mothers reside in a different place where the kids are born?
- Is the distribution of ages of the fathers different from that of the mothers? Discriminate between newborn, deaths, etc. See if there's any correlation.
- Look at age aggregations for non-fetal deaths. How is the distribution of ages of

newborn deaths?
- Look up cultural-racial identification. Are distributions any different?
- Influence of multiplicity of pregnancy on outcome.
- Compare mother's and father's educative levels. Are these distributions any different? Look for combinations of these two, maybe data can already show inequity?
- Influence of number of pregnancies, successful pregnancies and deaths for a given mother on outcome. Check distribution, geographical distribution, correlation with race, culture, age, among others.

Next steps:

Once the variables and the cleaning of the data set for a single year have been established, during week 6 of the course, the team will work on obtaining a complete database with as much as 10 years data, or as close to it as possible, that meets the requirements that have been determined in this week's work such as maximum total null data and criteria such as relevancy of the information. The above in order to study the behavior of the mortality rate of this population over time in Colombia. Additionally, we will begin to look for the factors that explain or are related to the question posed for this project, the team will therefore be applying techniques such as: descriptive statistical time series analysis, visualization, correlation between variables and clusters analysis.

References.

[1] Newborn death and illness, World Health Organization, 2011. https://www.who.int/pmnch/media/press_materials/fs/fs_newborndealth_illness/en/
[2] Resolution adopted by the General Assembly on 25 September 2015. United Nations, 2015, p. 16 https://undocs.org/A/RES/70/1
[3] COLOMBIA - Estadísticas Vitales. Posted: 24 Jan, 2020. Source; Dane. http://microdatos.dane.gov.co/index.php/catalog/652/get_microdata
[4] Colombia: Humanitarian Dashboard (as of 30 June 2017) - Colombia, Source: OCHA, Posted: 29 Aug 2017, Originally published: 14 Aug 2017.
[5] Colombia: Dashboard Humanitario 2019 (marzo de 2019) - Colombia, Source: OCHA, Posted: 27 May 2019, Originally published 20 May 2019.