New Methodologies in the World of Official Statistics

Instructions: Click on the link to access each author's presentation.

Chair: Alejandro Ruiz

Participants:

Gonzalo Peraza:* Generating Synthetic People and Households from Incomplete Aggregated Counts Using Mathematical Programming

Elio Villaseñor: Enhancing Data Integration and Utilization: An Open Source Platform for Consolidating Statistical Office Outputs

Jairo Alberto Fuquene Patino: A Bayesian Approach to estimate the completeness of death registration

Humberto Vaquera Huerta:* Use of modern statistical methods in evaluating the impact of public policy programs

Flor Martínez:* A distribution regression aproach to estimate the rural-urban well-being in Mexico.

* Work presentation not available or non-existent

IAOS-ISI 2024
MEXICO CONFERENCE

An Open Source Platform for Consolidating Statistical Office Outputs

17 May 2024

Elio A. Villaseñor G.[1], Irving G. Cabrera Z.[1], Oswaldo Díaz[1], Alejandra Figueroa M.[1], Ricardo A. Olvera N.[1], Ignacio Agloni J.[2], Klaus Lehmann M.[2], Juan E. Concha O.[2], Catalina Quijada E.[2], Felipe S. Jimenez A.[2]

[1]*Instituto Nacional de Estadística y Geografía (INEGI), México*

[2]*Instituto Nacional de Estadísticas (INE), Chile*

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

Contents

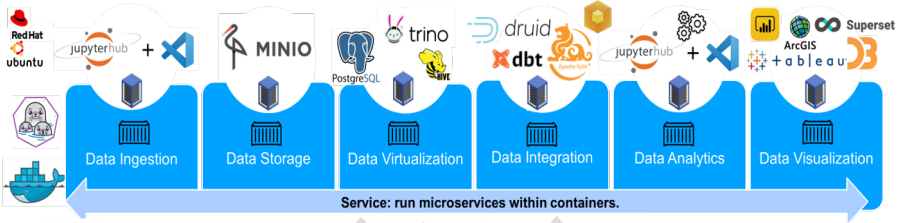
IOS-ISI 2024 MEXICO CONFERENCE

Data Lake Storage - NAS

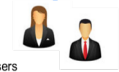
WS Cluster Sandbox

IP-NIC-Media Access Control Address

On-premise IT transversal infrastructure

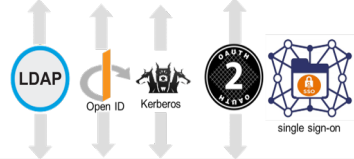


Data Science Lab



user | password | token

Users
- Data engineers.
- Data scientist.
- Data Developer integrator.
- Researchers.



Cybersecurity

ci/cd pipeline

Regulatory Framework

Chile-Mexico Cooperation Found Project

Project Name: "Development of an open-source technology platform focused on integrating various information sources to enhance the utilization of statistical office products."

Objective of the Project

Generate a collaborative environment that allows interaction with processes, procedures, roles, for the development of an open-source technological platform aimed at integrating various sources of information, which allow storing, transforming, processing, visualizing, large volumes of digital information, with privacy levels

Key Developments

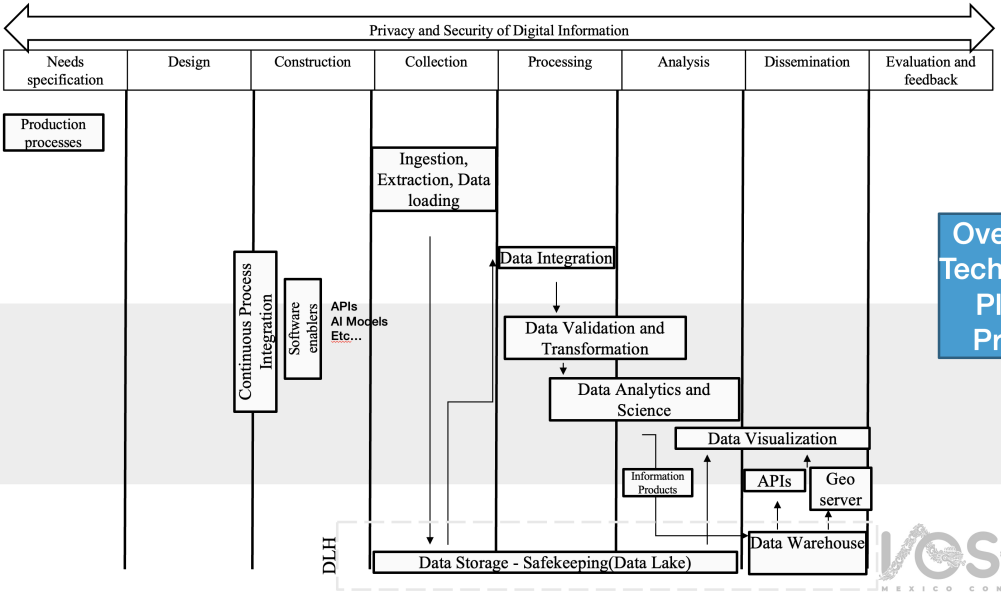
- Bi-weekly working meetings with INE and INEGI since June 2023.
- Review of documents on data lakes, virtualization, and statistical data utilization platforms globally.
- Deployment of LCiD's technology stack in INE Chile's on-premise IT infrastructure..
- Successful paradata analysis tests from surveys (ENUSC, ENUT).

Current Challenges at INE

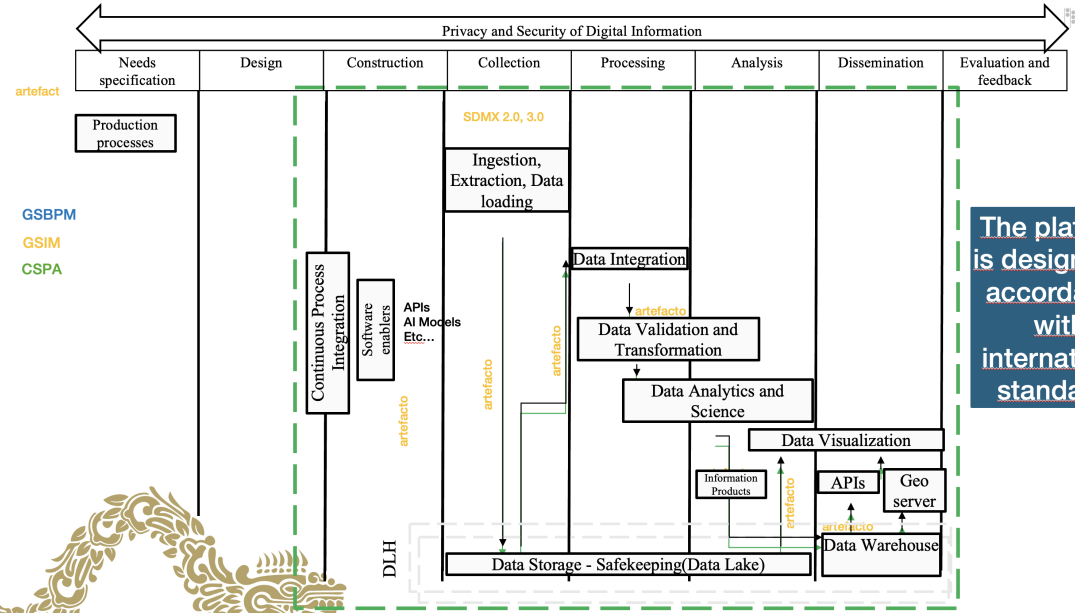
- Utilizing unstructured data sources like satellite images and managing large structured datasets like census data.
- Specific challenges include land cover mapping, updating cartographic frameworks, and processing census data.

Platform Overview

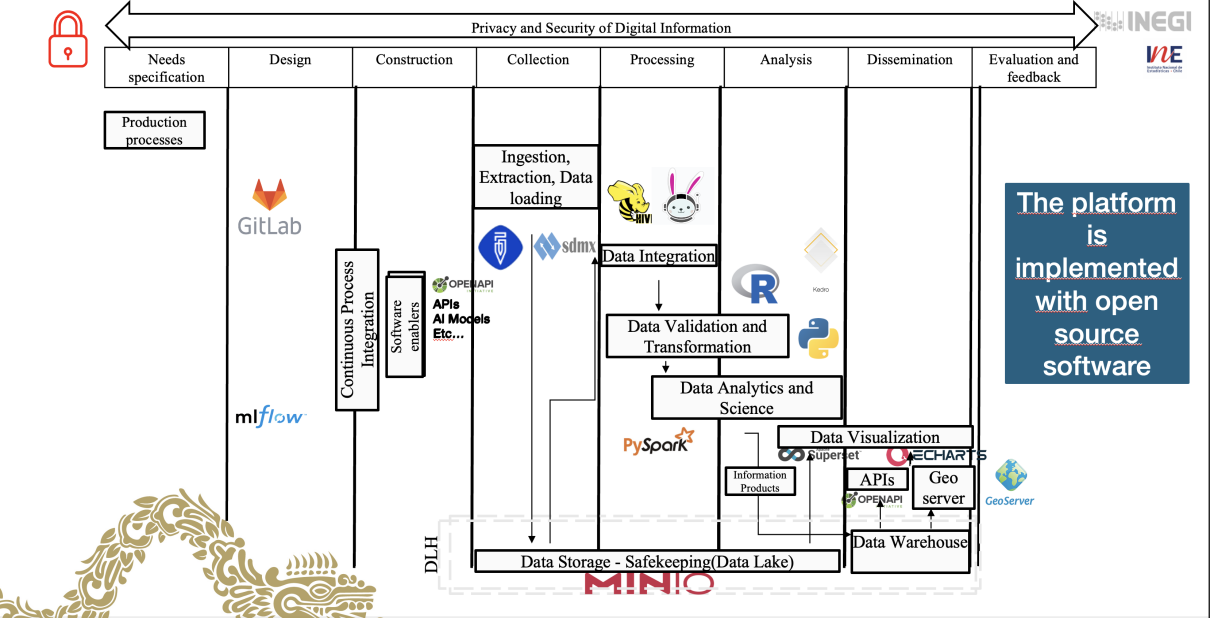
- **Design Phases:** Needs detection, design, construction, data collection, processing, analysis, dissemination, feedback.
- **Functionalities:** Data ingestion and extraction, validation, data integration, transformation, analytics, visualization, storage (Data Lake), continuous process integration, and Data Warehouse support.



Overview of Technological Platform Proposal



The platform is designed in accordance with international standards


Privacy and Security of Digital Information
Needs specification
Design
Construction
Collection
Processing
Analysis
Dissemination
Evaluation and feedback
Production processes
GitLab
mlflow
Continuous Process Integration
Software enablers
APIs
AI Models
Etc...
Ingestion, Extraction, Data loading
sdmx
Data Integration
Data Validation and Transformation
Data Analytics and Science
PySpark
Data Visualization
Information Products
APIs
Geo server
GeoServer
Data Warehouse
DLH
Data Storage - Safekeeping(Data Lake)
MINIO
The platform is implemented with open source software

Opportunities

- Utilizing INEGI's internally implemented platform for support and guidance.
- Handling large data volumes and integrating non-traditional data sources through advanced data lakes.
- Enhancing analytical capabilities with Big Data and AI for deeper, predictive, and prescriptive analytics.

Challenges

- Ensuring data governance to facilitate the use of advanced data science tools.
- Managing technological innovation to harness new technologies effectively.
- Addressing change management needs related to technological advancements.
- Developing human capital to keep up with modern statistical practices.

Plans for 2024-2025

- Implementation daily paradata monitoring from surveys (INE).
- Workflow of strategic programs, REP and RUE, implemented within the technology platform (INE).
- Survey Solutions Integration with Data Lake for Warehousing (INE).
- Deployment of a collaborative development environment (INE-INEGI).
- Launch platform and promote a practice community of practice (INE-INEGI).

IAOS-ISI 2024
MEXICO CONFERENCE

Thank you

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

IOS-ISI 2024
MEXICO CONFERENCE

A Bayesian approach to estimate the completeness of death registration

Jairo Fúquene-Patiño
Department of Statistics
University of California, Davis

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

A Bayesian approach to estimate the completeness of death registration

Jairo Fuquene, Department of Statistics, UC Davis (Joint work with **Tim Adair,** The Nossal Institute for Global Health, Melbourne School of Population and Global Health, The University of Melbourne, Technical Advisory Group on COVID-19 Mortality Assessment. World Health Organization).

IOS-ISI 2024 MEXICO CONFERENCE

Table of Contents

Introduction

- **Measures of civil registration and vital statistics (CRVS) systems play a critical role for improving the access of people to health care and education.**

- **Equitable Completeness of Death Registration (CDR) around world allow policymaking in statistical and ministry of health institutions.**

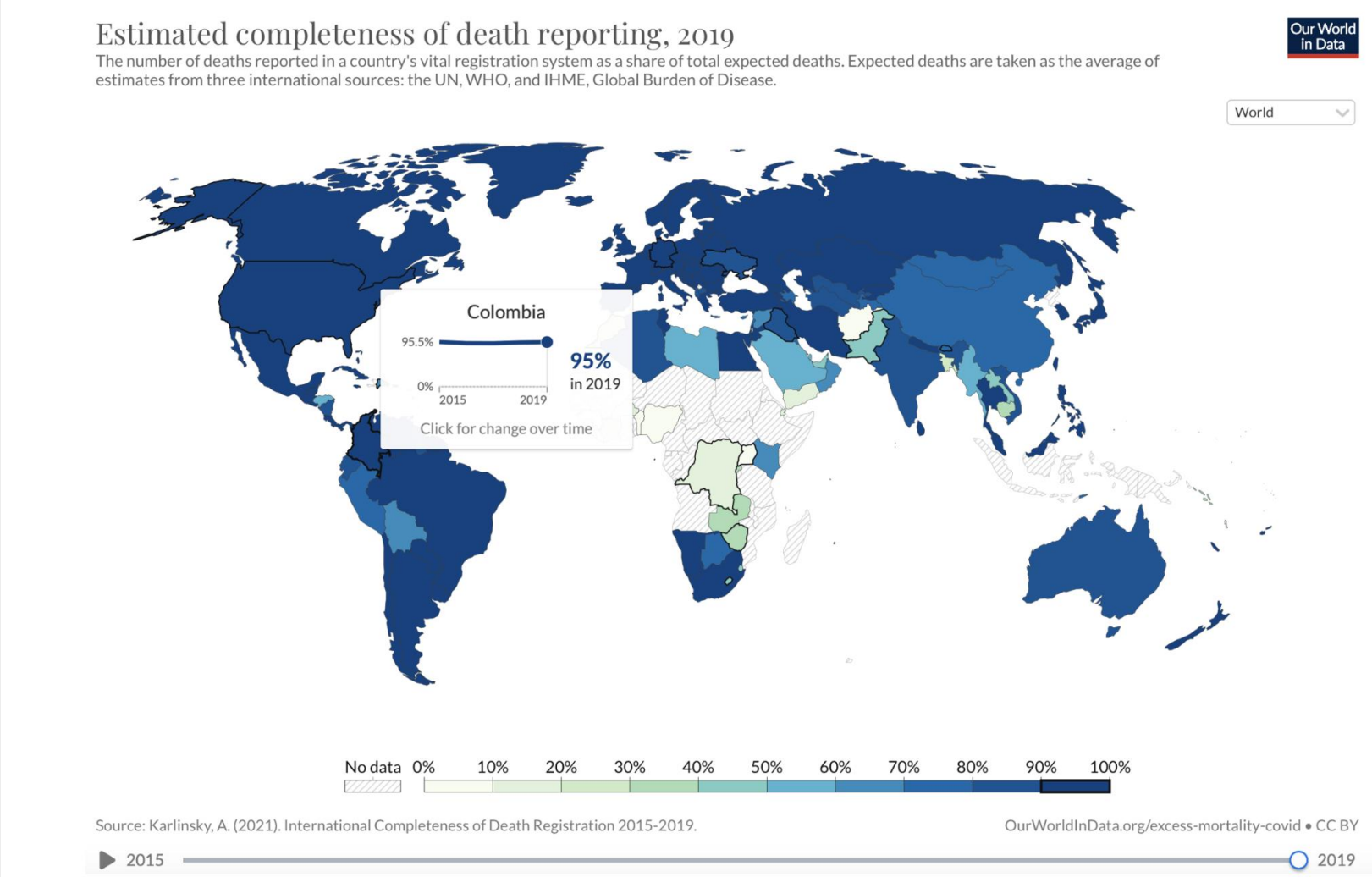
- **The models are applied to demographic models which considers demographic covariates to predict completeness and which has been implemented in several low and high income countries.**

Introduction

- **The use of our approach can allow institutions to improve the model parameter estimates and prediction of completeness of death registration. Our new models are based on a dataset updated to 2019, which uses Global Burden of Disease death estimates based on the GBD 2019 and now comprises 120 countries and 2,748 country-years from 1970-2019.**

- **To illustrate the effectiveness of our proposal at national and subnational levels, we consider the completeness of death registration in the departments of Colombia in 2017 and use the reported completeness from the Colombian Population Census in 2018 as our comparator dataset.**

CDR international sources: the UN, WHO, and IHME, Global Burden of Disease



CDR important for the National Statistical Office

- **The primary source of mortality data for national governments should be a complete civil registration and vital statistics (CRVS) [Mikkelsen et al., 2015]. However, only an estimated 59% of all deaths that occur globally are registered, with almost all unregistered deaths occurring in low- and middle-income countries [Dicker et al., 2018].**

- **Reliable measurement of the completeness of death registration – that is, the percentage of deaths that are registered – is important firstly to measure the extent of under-registration of deaths and to inform interventions that aim to attain complete registration. The completeness of registration can also be used to adjust data from a death registration system and, together with model life tables, produce estimates of key mortality indicators.**

Various methods

- **Death distribution methods (DDMs).**
- **Capture-recapture methods.**
- **Population estimated by the Global Burden of Disease (GBD) Study or United Nations (UN) World Population Prospects.**

The Bayesian models

- **The models use only relatively limited data (which are readily available at subnational levels), provide completeness estimates for the most recent year of death registration data, and are not reliant on assumptions about population dynamics such as closed migration that adversely affect subnational estimates.**

- **The frequentist models have been applied in several settings, including to estimate excess mortality from COVID-19 in Peru, to calculate subnational completeness for Indian states and 2,844 Chinese counties, to monitor completeness of community death reporting systems in Bangladesh, and to estimate cause-specific mortality rates to track attainment of Sustainable Development Goals in Myanmar. Also, in other latin American Countries, Colombia, Ecuador and Brazil at subnational levels.**

The model

$$y_{ij} = \theta_{ij} + \epsilon_{ij}, \qquad j = 1, \ldots, n_i,$$
$$\theta_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + u_i, \qquad i = 1, \ldots, m,$$

The models

- **The models use only relatively limited data (which are readily available at subnational levels), To estimate the completeness of death registration for both sexes (males and females) we have**

- **Model 1: $x_{ij} = (1, RegCDR_{ij}, RegCDR^2_{ij}, \%65^2_{ij}, \ln(5q0)_{ij}, C5q0_{ij}, year_{ij})$,**
- **Model 2: $x_{ij} = (1, RegCDR_{ij}, RegCDR^2_{ij}, \%65^2_{ij}, \ln(5q0)_{ij}, year_{ij})$,**

RegCDR is the registered crude death rate (registered deaths divided by population multiplied by 1000), RegCDR2 is the registered crude death rate squared, %65 is the fraction of the population aged 65 years and over, ln(5q0) is the natural log of the estimate of the true under-five mortality rate, C5q0 is the is the completeness of registered under-five deaths (estimated as the under-five mortality rate from registration data divided by the estimate of the true under-five mortality rate), and year is calendar year of death. The under-five mortality rate and the population aged 65 years and over are important drivers of the registered crude death rates in populations. We also consider models for males and females independently with the same predictors but sharing the same information for C5q0 which are likely to be similar for males and females.

Deviance measures and evaluation of the goodness of fit

- **To evaluate the performance of the models and GL priors we consider two deviance measures, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) given by,**

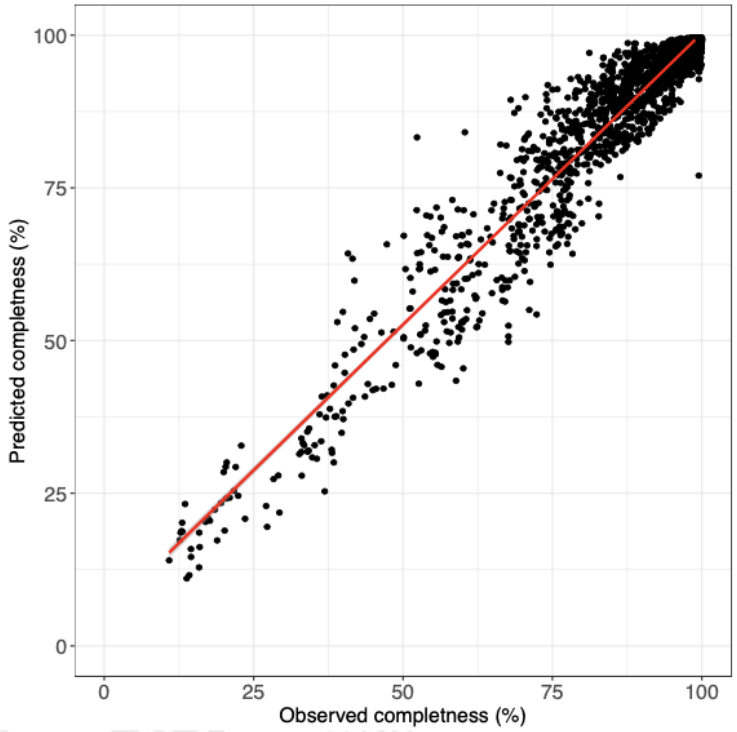
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mid \hat{\delta}_{ij} - c_{ij} \mid, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\hat{\delta}_{ij} - c_{ij})^2}.$$

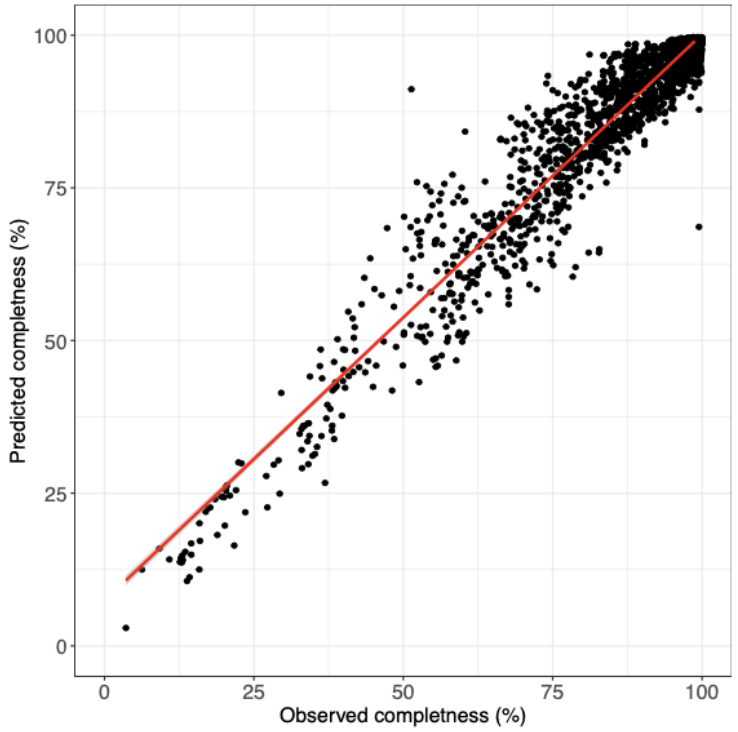
- **We also evaluate the goodness of fit of the models using an estimate of the R-square as follows,**

$$\text{R-square} = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(c_{ij} - \bar{c}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(c_{ij} - \hat{\delta}_{ij,-u})},$$

Bayesian models Females

Model 1 (Females)	Model 2 (Females)
R-square= 0.848	R-square=0.811
RMSE= 2.638	RMSE= 2.646
MAE= 4.239	MAE=4.376

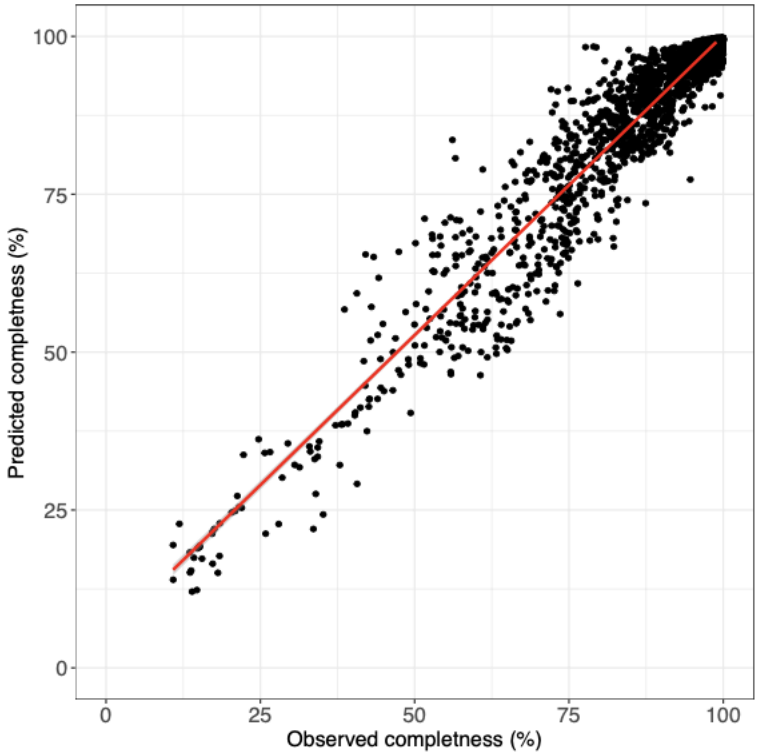


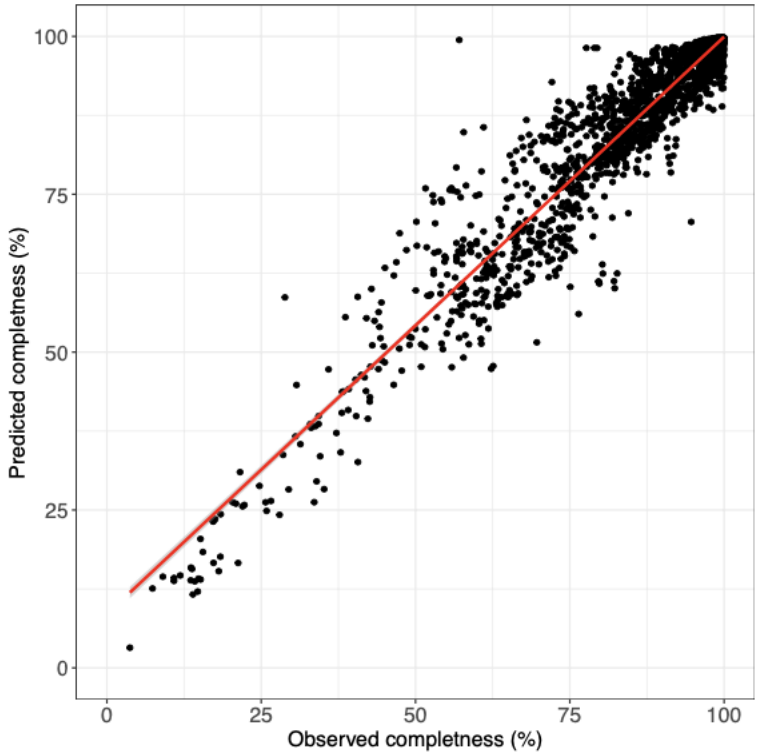


IOS-ISI 2024
MEXICO CONFERENCE

Bayesian models Males

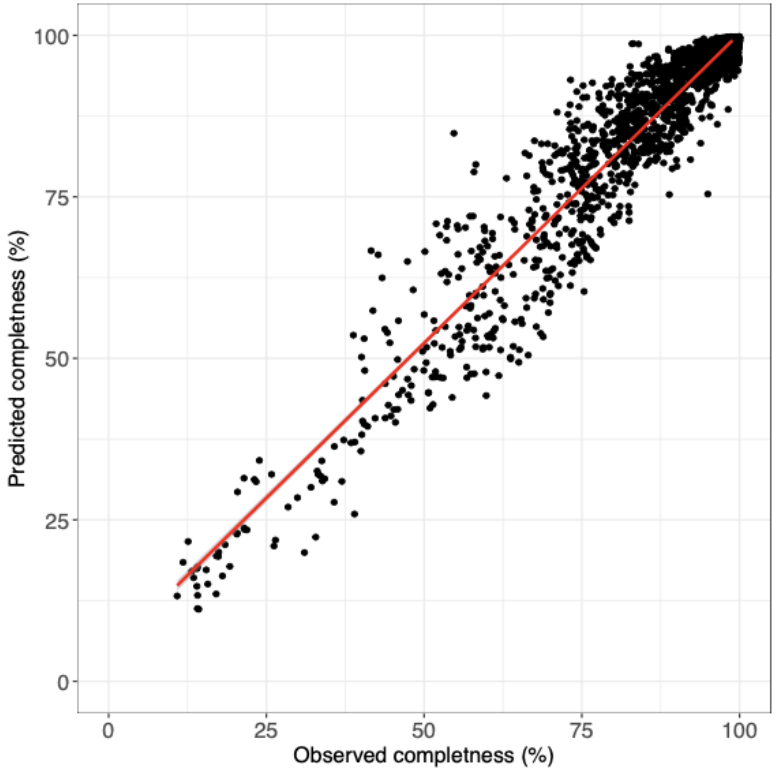
Model 1 (Males)	Model 2 (Males)
R-square = 0.823	R-square = 0.780
RMSE = 2.705	RMSE = 2.784
MAE = 4.358	MAE = 4.685

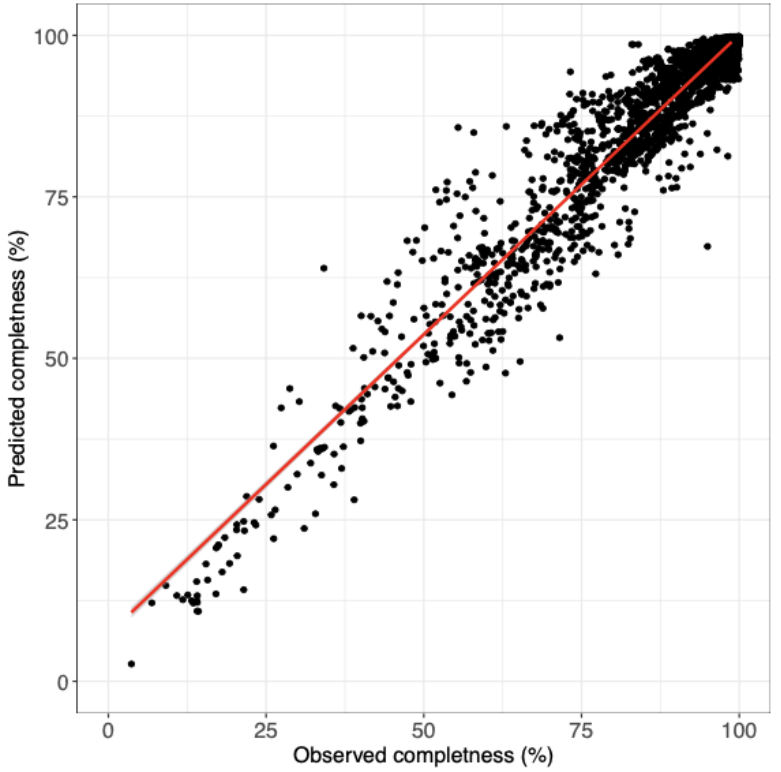




Bayesian models Both Sexes

Model 1 (Both)	Model 2 (Both)
R-square = 0.836	R-square =0.789
RMSE = 2.617	RMSE = 2.615
MAE = 4.280	MAE = 4.455





IOS-ISI 2024 MEXICO CONFERENCE

Deviance measures and evaluation of the goodness of fit different levels of completeness

	RMSE (model 1) × 100	MAE (model 1) × 100	RMSE (model 2) × 100	MAE (model 2) × 100
Both Sex				
90$<$100	1.87	1.18	2.08	1.25
80$<$90	5.36	4.49	5.10	3.99
60$<$80	7.92	6.63	7.33	5.66
30$<$60	9.14	6.90	10.10	7.47
$<$30	4.67	3.78	4.84	3.33
Females				
90$<$100	1.89	1.24	2.06	1.32
80$<$90	5.42	4.50	5.26	4.15
60$<$80	7.80	6.39	7.26	5.60
30$<$60	8.53	6.59	9.70	7.36
$<$30	5.15	4.34	4.34	3.58
Males				
90$<$100	1.90	1.20	2.12	1.29
80$<$90	5.24	4.38	5.05	3.89
60$<$80	7.94	6.61	7.72	5.96
30$<$60	8.69	6.29	10.27	7.85
$<$30	5.39	4.59	6.24	3.95

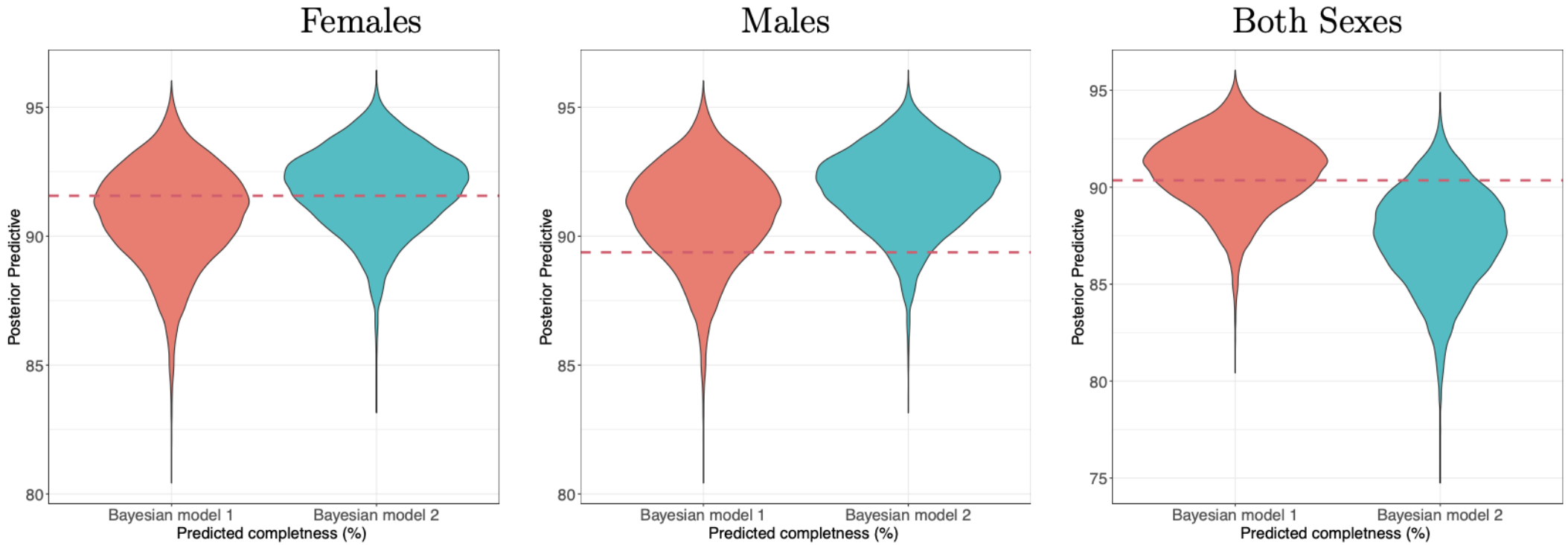
National level implementation

To evaluate the performance and demonstrate the utility of the models we use them to estimate the completeness of death registration for both sexes, males and females for the Country of Colombia and the corresponding regional division by 33 departments in 2017, k = 1, ..., 33. The results are compared to data from the Colombian population census 2018.

Deviance measures and evaluation of the goodness (national level)

		Model 1	Model 2
Both sexes	100× MAE	5.99	5.36
	100× MSE	0.94	0.56
	$\#\mid \hat{\delta}_k - c_k \mid < 10\%$	29	29
Males	100× MAE	6.17	5.50
	100× MSE	0.85	0.57
	$\#\mid \hat{\delta}_k - c_k \mid < 10\%$	29	30
Females	100× MAE	5.89	5.20
	100× MSE	0.98	0.53
	$\#\mid \hat{\delta}_k - c_k \mid < 10\%$	29	28

Marginal posterior predictive distributions of the completeness at the national level under Models 1 and 2. The red lines illustrate the observed values obtained from Census 2018.

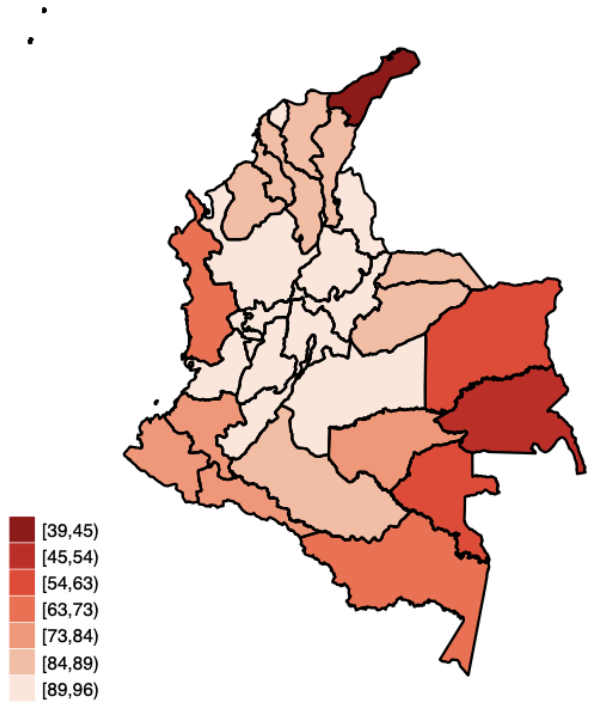


IOS-ISI 2024
MEXICO CONFERENCE

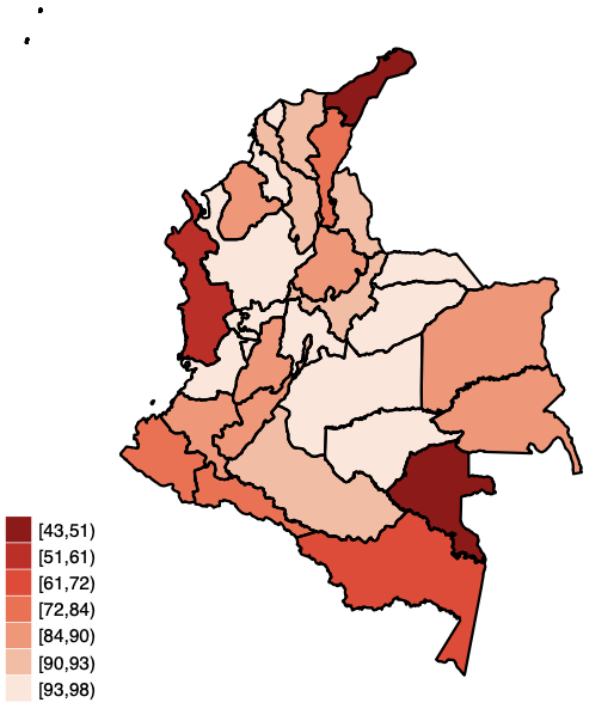
Subnational level implementation

Departments with some of the higher mortality rates and historically affected by problems in the civil registration system present important lower values of completeness according to Census.

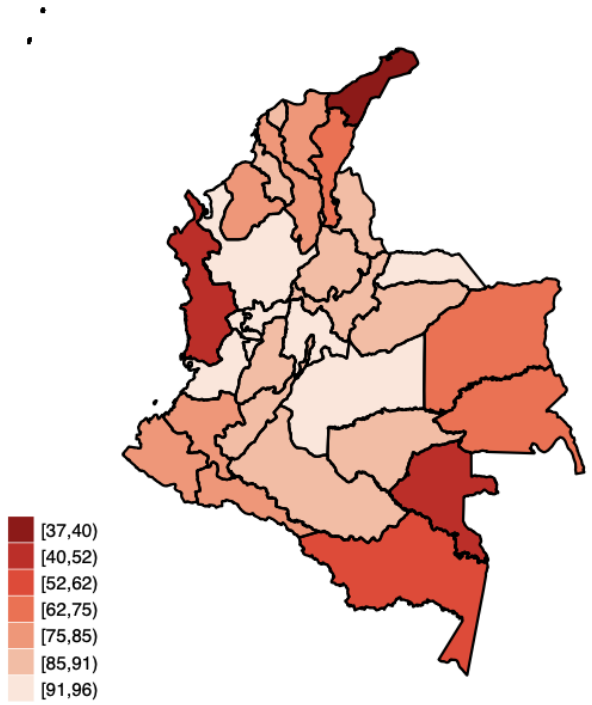
Census 2018 (Both-sexes)



Model 1 (Both Sexes)



Model 2 (Both Sexes)

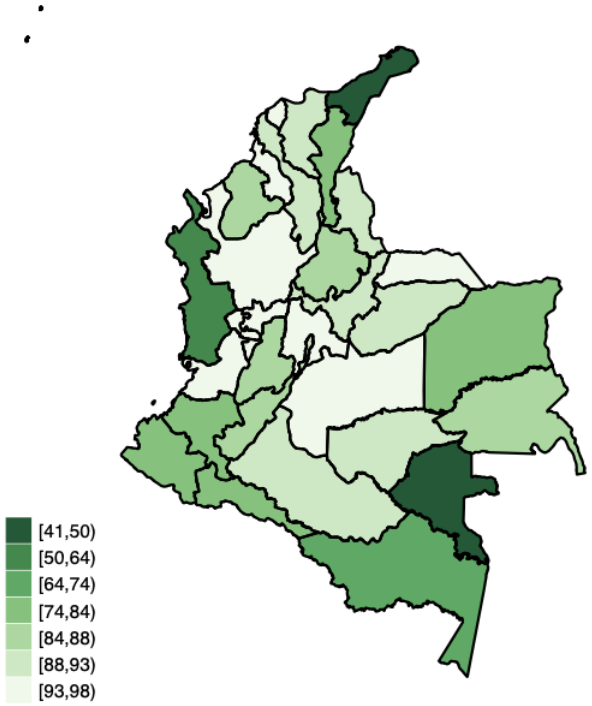


IOS-ISI 2024 MEXICO CONFERENCE

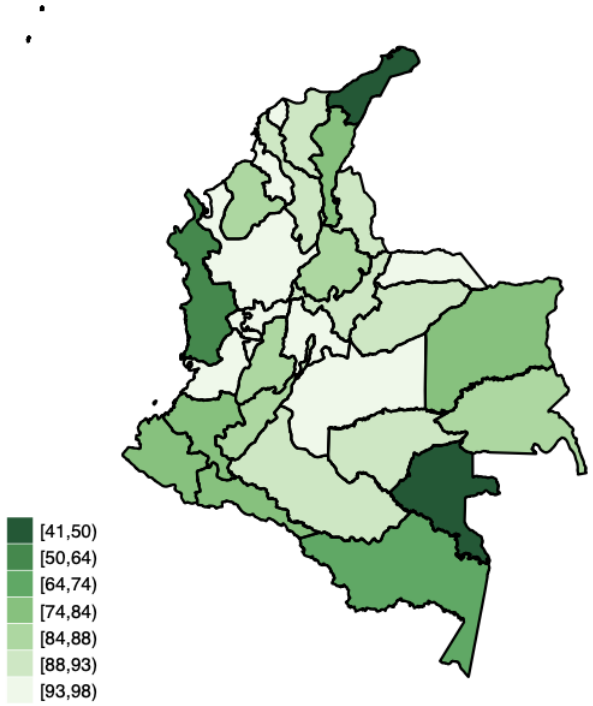
Subnational level implementation

Departments with some of the higher mortality rates and historically affected by problems in the civil registration system present important lower values of completeness according to Census.

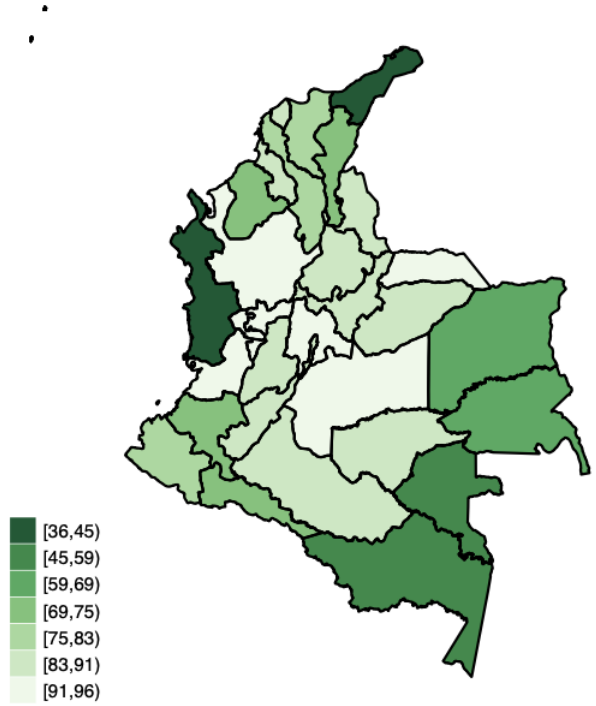
Census 2018 (Female)



Model 1 (Female)



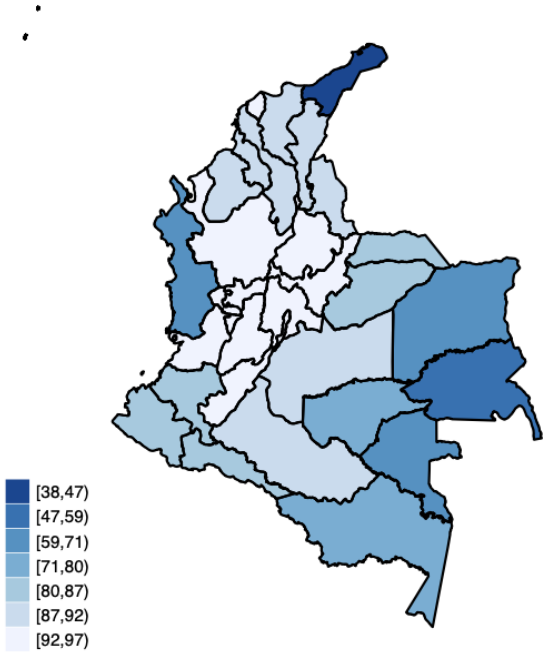
Model 2 (Female)



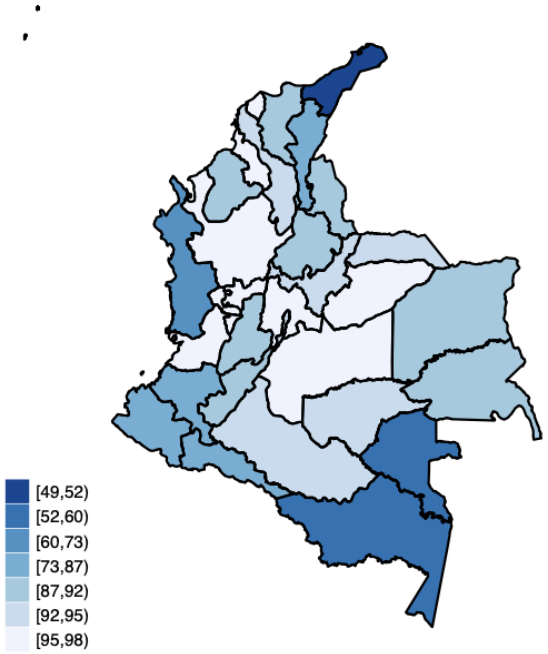
Subnational level implementation

Departments with some of the higher mortality rates and historically affected by problems in the civil registration system present important lower values of completeness according to Census.

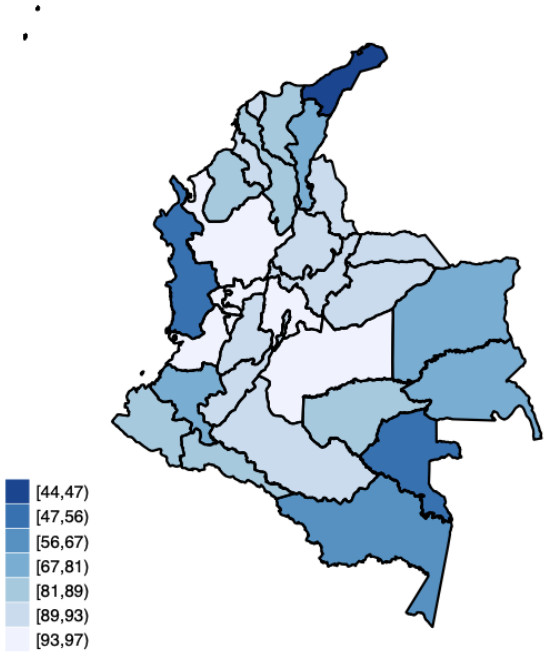
Census 2018 (Male)



Model 1 (Male)



Model 2 (Male)



Conclusions

- **Our proposal to estimate the completeness of death registration under a Bayesian framework is inspired by demographic models which include information typically available from multiple sources, e.g., surveys, censuses and administrative records.**

- **The use of a Bayesian framework to extend the existing frequentist models further strengthens the models effectiveness at estimating completeness of death registration. These models are immensely useful to enable estimation of completeness of death registration in a timely manner using available data, as demonstrated by their wide use in several different settings. In particular, they overcome the limitations of existing methods which rely on inaccurate assumptions of population dynamics and often provide estimates that lack timeliness.**

IOS-ISI 2024
MEXICO CONFERENCE

Thank you (Gracias!)

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS