INSTITUTO TECNOLÓGICO BUENOS AIRES – ITBA

|  |
| --- |
| itba_logo-2000x664 |
|  |
| |  |  |  | | --- | --- | --- | | **Fernando MESERI** | **4/14/23** |  | |

ESPECIALIZACIÓN EN CIENCIA DE DATOS

**Singularidades en las Curvas Poblacionales en el AMBA 1991-2022**

AUTOR: **Meseri, Fernando (Leg. N° 503.801)**

Profesores: Alicia Laura Mon ; Dr. Horacio Kuna.-

TALLER DE TRABAJO FINAL INTEGRADOR

BUENOS AIRES,

PRIMER CUATRIMESTRE 2023

1. Introducción

La información estadística que brindan las proyecciones de población es un insumo principal en la implementación de políticas estatales. Las proyecciones poblaciones es común encontrarlas a nivel País o Provincia, pero resulta particularmente importante poder contar con dichas proyecciones con un mayor nivel de desagregación, municipio/departamento.

En este caso se analiza en particular los municipios del AMBA, Argentina para el período 1991-2022. Las proyecciones del Instituto Nacional de estadísticas y Censos (INDEC) han estimado la población dichos municipios. El valor arrojado para el CENSO 2022 en el caso de La Matanza presenta una desviación importante respecto al error promedio encontrado entre las proyecciones y el valor relevando en el CENSO 2022. A partir del análisis de datos censales y variables indirectas se analiza las curvas poblacionales, el error en las proyecciones INDEC respecto a lo arrojado en el CENSO 2022 así como el caso puntual de La Matanza.

1. Revisión bibliográfica

Marco Conceptual

La información estadística que brindan las proyecciones de población constituye una herramienta fundamental para la planificación de políticas públicas de corto, mediano y largo plazo. Permite estimar demanda potencial de bienes y servicios en distintas áreas como Salud, Educación, entre otras - Instituto Nacional de Estadísticas y Censos (INDEC,2013) [1]. El estado puede de esta forma determinar los recursos presupuestarios necesarios para satisfacer estas demandas. En la provincia de Buenos Aires ciertos aspectos del presupuesto son asignados en base a la población de cada municipio. Es necesario entonces contar con la información en un alto nivel de desagregación espacial(municipios).

Marco Teórico

La elaboración de proyecciones de población es una tarea compleja que debe ser realizada a través de un análisis exhaustivo que permita considerar los censos anteriores como también registros vitales y estimaciones de migración. (INDEC,2013) [1]. En general se ha utilizado en método de las componentes para elaborar dichas proyecciones. Mas esta metodología no ha podido ser replicada al nivel de las jurisdicciones más elementales, departamentos, por cuanto la información no es suficientemente confiable y la inestabilidad de la migración interna no admite formulación de hipótesis a mediano plazo. (Álvarez,2001) [2]. Una forma de realizar estas predicciones ha sido mediante métodos matemáticos de extrapolación en base a la información censal previa. (Álvarez,2001) [2].

El INDEC provee proyecciones de población por departamento para el período 2010-2025(INDEC, 2015) [3], particularmente para todos los municipios del AMBA. Se destaca que el crecimiento de la población en Argentina observado entre 2001 y 2010 a nivel departamental pone en evidencia las diferencias geográficas que existen en la dinámica poblaciones, con un comportamiento heterogéneo.

Estado del arte

Históricamente se observa conceso en la utilización del Método de las Componentes para la determinación de proyecciones poblacionales a nivel País o Provincia. El mismo contempla el crecimiento poblacional intercensal y proyecta cada una de las variables determinantes de forma independiente -fecundidad, mortalidad y migración (Álvarez,2001) [2]. Ciertamente la Serie de Análisis Demográfico de INDEC utiliza este método para la proyecciones Nacionales y Provinciales (INDEC, 2015) [3].

Asimismo, para población de países desarrollados, también se ha utilizado el modelo de regresión logística en este tipo de predicciones (Gupta, Bhattacharya,Chattyopadhyay ,2012) [4]. Pero este modelo tiene ciertas limitaciones cuando se aplica a data censal dispersa en el tiempo, especialmente para países en desarrollo. Generalmente en estos casos las tasas de crecimiento relativo presentan tendencias inusuales, distintas a la tendencia decreciente de la regresión logística. Gupta et al., (2012) proponen modelos simplificados y variantes de Tsoularis and Wallace Model (TWM) que han proporcionado mejores resultados.

A mayor nivel de desagregación se trabaja con métodos alternativos, como puede ser extrapolación matemática, Ratio- Correlation Method, Housing Unit Method, entre otros (Hoque, 2012)[5]. Por otra parte el centro Latinoamericano y Caribeño de Demografía (CELADE) ha promovido la utilización de otras técnicas para mejorar las estimaciones poblaciones derivadas de la extrapolación matemática. Se utiliza la metodología de variables sintomáticas, que permite establecer correlaciones a las tendencias poblacionales con información de variables indirectamente asociadas al fenómeno de crecimiento poblacional, a saber: nacimientos y defunciones, matrícula escolar, permisos de construcción, otros (Álvarez,2001)[2].

En lo que respecta a técnicas propias de ciencias de datos para análisis de información censal, se desatacan los siguientes usos: la utilización de data mining para búsqueda de patrones en la información censal, predicciones y forecasting utilizando modelos ARIMA e inducción con árboles de decisión (Chawda, Rane, Giri, 2018) [6]. También se destaca el uso de árboles regresión y clasificación para el agrupamiento o clustering en distintas clases, tomando como input información censal.

1. Definición del problema

El municipio de La Matanza una singularidad en su curva de crecimiento poblacional, tanto número de habitantes como tasas intercensales, respecto a los municipios aledaños del AMBA para el período 1991-2022.-

1. Justificación del estudio

La información estadística que brindan las proyecciones de población constituye una herramienta fundamental para la planificación de políticas públicas de corto, mediano y largo plazo. Permite estimar demanda potencial de bienes y servicios en distintas áreas como Salud, Educación, entre otras - Instituto Nacional de Estadísticas y Censos (INDEC,2013) [1].

Las proyecciones del INDEC en la serie demográfica N °38(INDEC, 2015) [3] han estimado la población para los municipios del AMBA. El valor arrojado para el censo 2022 en el caso de La Matanza presenta una desviación importante respecto al error promedio encontrado para el resto de los municipios. Es por esto que se pretende analizar la singularidad en la curva poblacional de La Matanza.

1. Alcances del trabajo y limitaciones

El alcance del trabajo es básicamente el análisis y estudio de datos censales. Comprende principalmente la utilización de datos censales del AMBA para el período 1991-2022. Analizar las curvas poblacionales, su comparación y el error en las proyecciones INDEC respecto a lo arrojado en el CENSO 2022 (INDEC ,2022) [7]. El trabajo se limita a demostrar la singularidad o no en el dato poblacional de la Matanza, sin intentar explicar las causas del hecho, particularmente en lo que se refiere al fenómeno demográfico que pudiese estar detrás de esta singularidad.

1. Hipótesis

Es posible demostrar la curva de crecimiento poblacional de la Matanza presenta una singularidad respecto a los municipios aledaños (AMBA) en situaciones socio-demográficas similares para el período 1991-2022.

Preguntas

1. ¿Se puede estimar una tasa de crecimiento promedio de la población urbana/suburbana en base a los 4 Censos anteriores?
2. ¿Se puede individualizar tasas de crecimiento distintas por municipio??
3. ¿Considerando las curvas de crecimiento reales de los municipios y las proyecciones realizadas, dado un intervalo de confianza se puede realmente afirmar que la Matanza presenta una singularidad?
4. ¿Es esperable que municipios aledaños crezcan poblacionalmente en el mismo orden de magnitud?
5. La apertura de los datos por edad, sexo y nivel de habitantes en el hogar muestra cierto comportamiento esperable, algún patrón. Se asemeja a los valores encontrado para La Matanza para Censo 2022.-

Respuestas

1. Sí, es posible determinar la tasa de crecimiento en función a los censos anteriores, utilizando estadística descriptiva.
2. Sí, es posible, ya que el censo incluye la desagregación por departamento/municipio. Luego estos se pueden agrupar en sectores de interés, ej. AMBA.
3. Si, en principio. Las causas de este fenómeno no son de interés del presente trabajo. El fenómeno demográfico es complejo, multicausal. Aun así, se pueden establecer valores de referencia acordes a la región, donde las características socio-demográficas son similar (conurbano).
4. Sí, municipios urbanos suburbanos con iguales características socio-demográficas es esperable que tengan tasas de crecimiento similares.
5. Es esperable que estos factures socio-demográficos en grandes muestras muestren patronales similares, o valores semejantes.
6. Objetivos

El objetivo del presente trabajo consiste en demostrar que la tasa de crecimiento y curva poblacional de la Matanza desde 1991 hasta 2022 presenta una singularidad respecto a aquella desarrollada por los otros municipios de conurbano bonaerense. Tomando como base los datos obtenidos en los 4 últimos censos nacionales 1991-2001-2010 y 2022.

OBJETIVOS ESPECÍFICOS

1. Procesar y unificar la información censal desde 1991 a 2022 para distintos niveles de granularidad, ya que en algunos casos no se presentan todas las variables. Se requiere un pre procesamiento y limpieza de datos importante.
2. Exploratory Data Analysis. Construir las curvas poblaciones y comparaciones geográficas de las mismas. Analizar Data mining patterns.
3. Realizar proyecciones para población censo 2022 con metodologías tradicionales modernas utilizando las variables en forma individual o combinadas, tomando como base los 3 censos anteriores (1991-2001-2010).
4. Utilizar data mining para rellenar los tiempos intercensales de las variables de interés.
5. Realizar proyecciones para el censo 2022 mediante data mining (ARIMA, decision tree regression algoritm, others), utilizando los 3 censos anteriores (1991-2001-2010)
6. Comparar las curvas censales y su ajuste con las proyecciones realizadas para 2022.-
7. Determinar el error de las distintas proyecciones para 2022 para todos los municipios, comparar con la situación de la Matanza.
8. Inferir si existe alguna metodología que ajuste mejor las proyecciones para el caso del AMBA.

Variables

Se trata de una hipótesis multivariable donde las variables podrían tener una relación causa-efecto sobre el fenómeno analizado.

- Tipos de Variables

1. Población Total. Variable Dependiente. Cuantitativa. Cantidad de personas que habitan un determinado sector del territorio, en distintos niveles de agregación. INDEC
2. Tasa de Natalidad. Independiente. Cuantitativa. Refiere a la relación que existe entre el número de nacimientos ocurridos en un cierto periodo y la cantidad total de población existente en el área geográfica.
3. Tasa de mortalidad. Independiente. Cuantitativa. Es la proporción de personas que fallecen respecto al total de la población en un período de tiempo, usualmente expresada en tanto por mil (‰) por año.
4. Sexo. Contextual. Categórica. Sexo de las personas, por nivel de desagregación. INDEC
5. Edad. Contextual. Cuantitativa. Edad de las personas, por nivel de desagregación. INDEC
6. Nivel educativo. Contextual. Categórica Ordinal. Se refiere al nivel de educación máximo alcanzado por un personal hasta el momento. Primario, secundario, terciario, Universitario. INDEC
7. Cantidad de habitantes en el hogar. Contextual. Cuantitativa. Se determina como la cantidad de personas conviviendo de forma permanente bajo el mismo techo. INDEC
8. Asistencia escolar. Contextual. Categórica Ordinal. Refiere a la asistencia o no a establecimientos educativos, desagregados por edad.
9. Metodologías
   1. Técnicas

Para dominio del problema, particularmente debido al nivel de desagregación analizado-departamentos-, se utilizarán las siguientes metodologías en el trabajo.

Metodologías Tradicionales:

Metodología de variables sintomáticas (distribución por prorrateo; dist. Proporcional; método de tasas variables, correlación de razones, correlación de diferencias, correlación de tasas) (Álvarez,2001) [2].

Metodologías Data Mining: CART, ARIMA, decision tree regression and classification algoritms (Chawda, Rane, Giri, 2018) [6].

* 1. Herramientas

Como herramientas se contemplan los lenguajes Python y SQL. Los datos se trabajarán en base de datos POSTGRES, para visualización geográfica se utilizará QGIS y OpenJump. En cuento al scripting se gestionará mediante Jupyter Notebooks.

1. Referencias bibliográficas

[1] Instituto Nacional de Estadística y Censos - INDEC. (2013). *Proyecciones provinciales de población por sexo y grupo de edad 2010-2014. -* Instituto Nacional de Estadística y Censos - INDEC, 2013. E-Book. (Vol. 1a ed.).

[2]Álvarez, G., United Nations. Economic Commission for Latin America and the Caribbean., & CELADE (Organization). División de Población. (2001). *Estimación de población en áreas menores mediante variables sintomáticas : una aplicación para los departamentos de la República Argentina (1991 y 1996).* Naciones Unidas, CEPAL/ECLAC.

[3] Instituto Nacional de Estadística y Censos - INDEC. (2015). *Estimaciones de población por sexo, departamento y año calendario 2010-2025.* (1a. ed.).

[4] Gupta, Arindam , Gupta, Arindam , Gupta, Arindam (2012).*Exploring New Models for Population Prediction in Detecting Demographic Phase Change for Sparse Census Data*

[5] Hoque, Nazrul(2012). *Evaluation of small area population estimates produced by Housing Unit, Ratio-correlation, and Component Method II compared to 2000 Census counts.* Canadian Studies in Population 39, No. 1–2, pg. 91-108.

[6] Chawda, Manan, Rane, Rutuja, Giri, Skrinanth(2018) 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT) .*Demographic Progress Analysis of Census Data Using Data Mining ,* pg 1894-1897

[7] INDEC CENSO 2022- [Archivo de datos] Recuperado de https://www.censo.gob.ar/index.php/datos\_provisionales/

|  |  |  |  |
| --- | --- | --- | --- |
| [8] INDEC. Censo Nacional de Población y Vivienda 1991 y Censo Nacional | | | |
| de Población, Hogares y Viviendas 2001. |  |  |  |

-----------------------------------------

[9] https://redatam.org/es

<https://redatam.indec.gob.ar/argbin/RpWebEngine.exe/PortalAction?BASE=CPV2010A>

<https://redatam.indec.gob.ar/argbin/RpWebEngine.exe/PortalAction?BASE=CPV2001ARG>

**[10] IGM – Capas SIG - Departamento** [Archivo de datos] Recuperado de /**https://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG**

LinkS

DESARROLLO

1. Obtención de datos desde las fuentes. Consolidación Dataset

En primera instancia se busca recopilar los datos de los CENSOS poblacionales correspondientes desde 1991. Es decir Censos 1991, 2001, 2010 y 2022. En general el INDEC (colocar Fuentes ,refencias link a paginas??) dispone la información en formatos CSV o bien archivos Microsoft Excel de difícil procesado para su ingesta por parte de un proceso de ETL. En primera instancia los archivos no cuentan con estructura de encabezados en la primera línea, presentan descripciones y comentarios en las primeras líneas y no respetan formatos uniformes según la dimensión de análisis de los datos dentro del mismo periodo censal. Por otra parte, el formato de presentación y las variables relevadas se han visto modificadas desde el censo 1991 hasta el 2022. Una vez analizada las variables comunes a todos estos censos, fue necesario un pre-proceso manual de los archivos para lograr uniformidad y consistencia de forma de que sean tomados por el proceso de ETL .

En cuanto a las variables censales comunes al período analizado se toman las indicadas en la Tabla 1.

|  |  |
| --- | --- |
| **TABLA 1 . Variables CENSOS INDEC** | |
|  |  |
| **Variable** | **Descripción** |
| **Censo** | Año del censo indicado como entero de 4 dígitos. |
| **CodigoDepto** | Código de 4 dígitos con el INDEC indica los municipios / departamentos censales. |
| **Población** | Cantidad de personas que habitan un determinado sector del territorio, en distintos niveles de agregación. |
| **Varones** | Cantidad de personas de sexo Masculino . |
| **Mujeres** | Cantidad de personas de sexo Femenino. |
| **VivPartTot** | Cantidad de viviendas particulares totales para ese nivel de agregación. |
| **VivColectTot** | Cantidad de viviendas Colectivas totales para ese nivel de agregación. |
| **IndMasc** | Índice de masculinidad. |
| **Superficie** | Superficie total del partido a la fecha del censo. |

Es importante aclarar que existe una base consolidada -REDATAM [9]- con toda la información censal para los censos 2001 y 2010, la misma no está disponible para los censos de 1991 y 2022. REDATAM [9] es un software para procesamiento estadístico especializado en microdatos de censos de población y vivienda, encuestas y estadísticas vitales, desarrollado por el CELADE - División de Población de la CEPAL, de las Naciones Unidas. A su vez esta base sólo puede accederse mediante este software específico, no disponiendo de back ups y en general las salidas de las consultas son también archivos CSV no normalizados que requieren post-proceso.

## Dimensión Departamento- Slowly Changing Dimension

Al analizar los distintos censos desde 1991 hasta 2022 los departamentos del AMBA han modificado su división político – administrativa. Es decir, se han creado varios partidos nuevos, así como fusionado otros. Esto general que la dimensión departamento cambia a través de los distintos censos. Es por esto que se adopta un enfoque tipo Slowly Changing dimensión tipo 3. Donde se impacta el cambio de departamento, conservando en el registro el valor anterior, de que departamento proviene y la superficie correspondiente asociada. Esto es necesario para poder luego procesar información consistente en cuento a la granularidad analizada. Mismo el INDEC consigna:

|  |  |
| --- | --- |
| *“Nota: con el fin de posibilitar la comparación entre los Censos 1991 y 2001, los datos que* | |
| *corresponden al año 1991 fueron reprocesados según la división político - administrativa* | |
| *vigente al año 2001.” [8] INDEC, 2001*  En la Tabla 2 puede verse la dimensión departamento con los campos agregados y los comentarios correspondientes en cada caso. |

## Diccionario de Departamentos

Debido a las inconsistencias y falta de normalización de los archivos CSV que provee el INDEC, fue necesaria la creación de un diccionario de departamento asociando las distintas acepciones del nombre de departamento con su correspondiente código para que el proceso de ETL pudiese utilizar los códigos de departamento como clave foránea común a todos los Censos Nacionales. Se encontraron casos donde el mismo departamento figura nombrado con tílde , sin tílde , con espacios o abreviaturas en el campo.

## Proyecciones INDEC 2010-2025

Se incorpora también al conjunto de datos que se analizarán proyecciones realizadas por el INDEC [3] que indican la “*Población estimada al 1 de julio de cada año calendario por sexo, según partido. Provincia de Buenos Aires. Años 2010-2025”.*

## Componente geográfico

Con el objetivo de enriquecer el dataset se incorpora una tabla con los polígonos geográficos de cada departamento. La misma se obtiene del Instituto Geográfico Nacional (IGM) [10], de donde se obtiene la capa SIG para todos los departamentos de la República Argentina.

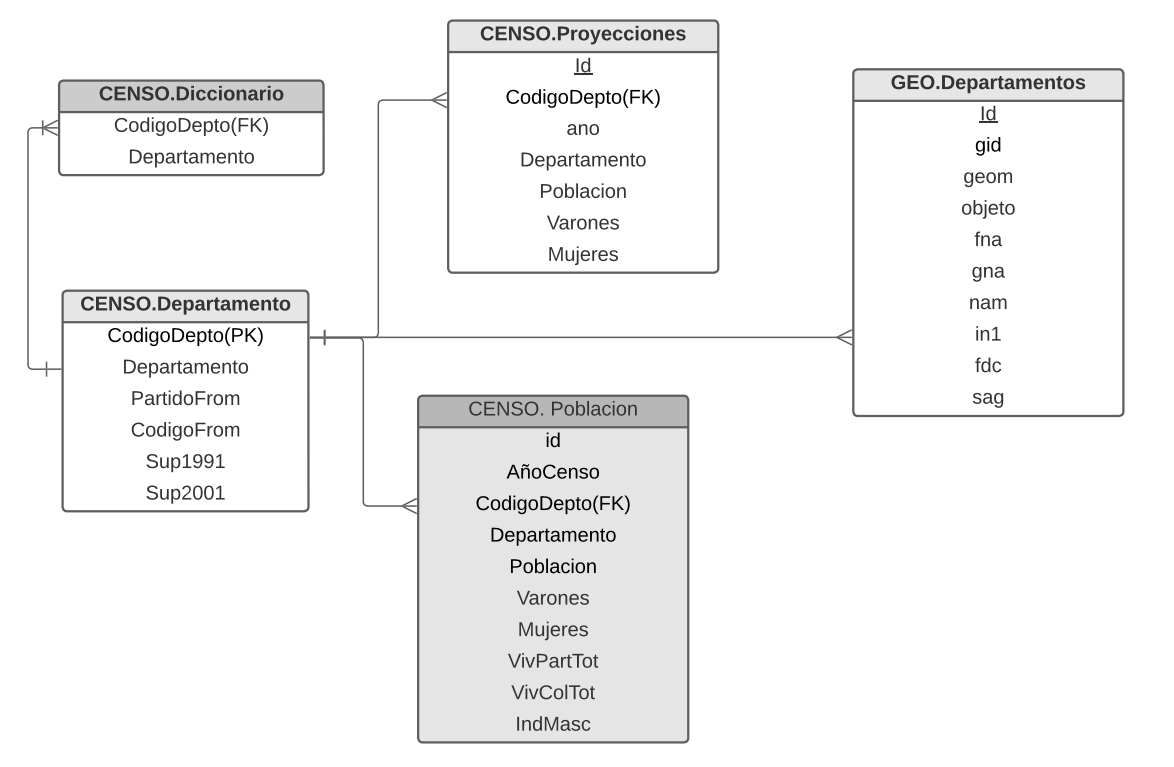
Tabla 3. Atributos geo.Departamento



## Herramientas

Para el guardado y manipulación de los datos se va a trabajar con un base de datos SQL ( SLQ Server- Local). Se desarrollo justamente un proceso de ETL en SQL Server Integration Services(SSIS) que en primera crear todas las tablas necesarias, para luego producir la ingesta , transformar en caso de que sea necesario, de los diferentes inputs que van a componer el modelo de datos.

## Diagrama entidad - relación



1. Descripción de funcionamiento de cada método a utilizar (Metodología de variables sintomáticas, Data Mining, etc.)
2. Implementación del ETL para extracción y transformación de datos
3. EDA- Análisis exploratorio y descriptivo de las variables
4. Comparación entre los modelos
5. Conclusiones