

Singularidades en Curvas Poblacionales para el AMBA 1991-2022

Fernando Meseri

24 de junio de 2024



Instituto Tecnológico Buenos Aires

Argentina-2024

1. Introducción

La información estadística que brindan las proyecciones de población representa un insumo vital en la implementación de políticas estatales. Dichas proyecciones es común encontrarlas a nivel País o Provincia, pero resulta particularmente importante poder contar con dichas proyecciones con un mayor nivel de desagregación (Municipios). En el presente trabajo se analiza en particular los municipios de la zona AMBA, Área Metropolitana de Buenos Aires (Argentina), para el período 1991-2022. El AMBA técnicamente incluye 40 municipios: Almirante Brown, Avellaneda, Berazategui, Berisso, Brandsen, Campana, Cañuelas, Ensenada, Escobar, Esteban Echeverría, Exaltación, Ezeiza, Florencio Varela, Gral. Las Heras, Gral. Rodríguez, Gral. San Martín, Hurlingham, Ituzaingó, José C. Paz, La Matanza, La Plata, Lanús, Lomas de Zamora, Luján, Malvinas Argentinas, Marcos Paz, Merlo, Moreno, Morón, Quilmes, Pilar, Presidente Perón, San Fernando, San Isidro, San Miguel, San Vicente, Tigre, Tres de Febrero, Vicente López, Zárate. Ocupa un territorio de aproximadamente 3.833km^2 y concentra 35 % de la población nacional, siendo el área geográfica más poblada del país y configurándose históricamente, como el núcleo central del sistema urbano argentino. El Instituto Nacional de estadísticas y Censos (INDEC) en general trabaja diferenciadamente 24(veinticuatro) de estos municipios del AMBA , siendo los que se analizarán en este trabajo. A saber: Almirante Brown, Avellaneda, Berazategui, Esteban Echeverría Ezeiza, Florencio Varela, General San Martín, Hurlingham, Ituzaingó, José C. Paz, La Matanza, Lanús, Lomas de Zamora, Malvinas Argentinas Merlo, Moreno, Morón, Quilmes, San Fernando, San Isidro, San Miguel, Tigre, Tres de Febrero, Vicente López.

Justamente el INDEC ha estimado la población dichos municipios para el período 2010-2025(INDEC, 2015)[12]. Si se analiza los errores de estas proyecciones respecto al valor arrojado para el censo 2022(INDEC, 2022)[10] , surge que **La Matanza** presenta una desviación importante respecto al error promedio encontrado para otros departamentos. En el presente trabajo partir del análisis de datos censales y variables indirectas se analizarán las curvas poblacionales del AMBA, se buscará estimar la población para el año 2022, para luego relevar el error respecto al censo 2022 según las distintas metodologías aplicadas. Se analizará puntualmente también La Matanza.

2. Revisión bibliográfica

2.1. Marco Conceptual

La información estadística que brindan las proyecciones de población en general es utilizada en 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(INDEC, 2015)[12].

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 nivel de desagregación espacial municipal (Departamentos).

2.2. 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)[13].

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 et al., 2001)[1]. Una forma de realizar estas predicciones ha sido mediante métodos matemáticos de extrapolación en base a la información censal previa(Álvarez et al., 2001)[1].

El INDEC provee proyecciones de población por departamento para el período 2010-2025(INDEC, 2013)[13], particularmente para todos los municipios del AMBA.Dichas estimaciones se apoyan en el crecimiento intercensal que comprende la evolución en conjunto de las tres variables básicas del análisis demográfico: la fecundidad, la mortalidad y la migración del período 2001–2010.Se destaca que el crecimiento de la población en Argentina observado en este período a nivel departamental pone en evidencia las diferencias geográficas que existen en la dinámica poblacional, con un comportamiento heterogéneo.

2.3. 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 et al., 2001)[1].

Ciertamente la Serie de Análisis Demográfico de INDEC utiliza este método para la proyecciones Nacionales y Provinciales(INDEC, 2015)[12]. Asimismo, para población de países desarrollados, también se ha utilizado el modelo de regresión logística en este tipo de predicciones(Gupta, Bhattacharya y Chattyopadhyay, 2012)[6]. 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)[7].

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 et al., 2001)[1].

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 y Giri, 2018)[4]. También se destaca el uso de árboles regresión y clasificación para el agrupamiento o clustering en distintas clases, tomando como input información censal.

3. 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.-

4. 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(INDEC, 2013)[13].

Las proyecciones del INDEC(INDEC, 2015)[12] 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.

5. 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)[10]. El trabajo se limita a demostrar la singularidad o no en el dato poblacional de la Matanza, sin intentar explicar las causas del fenómeno demográfico que pudiese estar detrás de esta singularidad.

6. 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 factores socio-demográficos en grandes muestras muestren patronales similares, o valores semejantes.

7. 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.
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. Realizar proyecciones para el censo 2022 mediante data mining (decision tree regression algoritm, others), utilizando los 3 censos anteriores (1991-2001-2010)
5. Comparar las curvas censales y su ajuste con las proyecciones realizadas para 2022.-
6. Determinar el error de las distintas proyecciones para 2022 para todos los municipios, comparar con la situación de la Matanza.
7. 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.

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. La **Tasa Bruta de Natalidad(TBN)** es el cociente entre el número de nacimientos ocurridos durante un período determinado, generalmente un año calendario, y la población media del período.
9. La **Tasa Bruta de Mortalidad (TBM)** es el cociente entre el número de defunciones ocurridas durante un período determinado, generalmente un año calendario, y la población media del período.
10. La **Tasa de Crecimiento Vegetativo(TCV)** es la diferencia entre la Tasa Bruta de Natalidad y la Tasa Bruta de Mortalidad de un período determinado, generalmente un año calendario
11. La **Tasa Global de Fecundidad(TGF)** es el número de hijos que en promedio tendría una mujer de una cohorte hipotética de mujeres que durante su vida fértil tuvieran sus hijos de acuerdo a las tasas de fecundidad por edad del período en estudio y no estuvieran expuestas al riesgo de mortalidad desde el nacimiento hasta el término de su período fértil.
12. La **Tasa de Mortalidad Infantil(TMI)** expresa el cociente entre el número de muertes de menores de un año acaecidas en la población de un área geográfica durante un período determinado, generalmente un año calendario, y los nacidos vivos en esa área durante el mismo período.
13. La **Matricula en ciclo primario común(Mat1ria)** representa la cantidad de alumnos que ingresan en primer año del ciclo primario.

8. Metodologías

8.1. Técnica

En esta sección se detallarán las metodologías utilizadas para realizar predicciones del valor poblacional de todos los departamentos del AMBA para el Año 2022, basados en los 3 censos anteriores y, dependiendo del método, variables sintomáticas típicas del fenómeno demográfico. Es importante destacar que ciertas metodologías tradicionales no son susceptibles de ser utilizadas debido al nivel de agregación analizado (Departamento).

8.1.1. Variables Censales

De los censos poblacionales se obtuvieron las siguientes atributos o variables.

1. Código Depto. Código del Departamento (Unidad Administrativa) según INDEC.
2. Población.
3. Varones.
4. Mujeres.
5. Viviendas Particulares Totales.
6. Viviendas Colectivas Totales.
7. Índice de Masculinidad.

8.1.2. Variables Sintomáticas

En este caso para algunas metodologías se utilizan variables sintomáticas como input en las estimaciones de población a realizar. A saber:

- 1. Tasa Bruta de Natalidad(TBN)**
- 2. Tasa Bruta de Mortalidad (TBM)**
- 3. Tasa de Crecimiento Vegetativo(TCV)**
- 4. Tasa Global de Fecundidad(TGF)**
- 5. Tasa de Mortalidad Infantil(TMI)**
- 6. Matricula en ciclo primario común(Mat1ria)**

A partir del relevamiento de datos públicos gubernamentales, de fuentes como el Programa de Análisis Demográfico(INDEC, 2022)[11] e información del Ministerio de Educación Nacional (Educación, 2022)[5] se confeccionó un compendio de estas variables para los años de interés. En el siguiente cuadro 1 puede observarse el valor de las mismas para el periodo analizado.

| Jurisdicción | Año | TMI | TGF | TBN | TBM | TCV | Mat1ria |
|---------------------|------------|------------|------------|------------|------------|------------|----------------|
| BUENOS AIRES | 1980 | 28.4 | 3.0 | 22.1 | 8.1 | 14.0 | |
| BUENOS AIRES | 1991 | 24.2 | 2.6 | 18.4 | 7.9 | 10.5 | 1752994.0 |
| BUENOS AIRES | 2001 | 15.0 | 2.3 | 16.9 | 8.2 | 8.7 | 1658221.0 |
| BUENOS AIRES | 2010 | 12.0 | 2.5 | 18.9 | 8.4 | 10.5 | 1667278.0 |
| BUENOS AIRES | 2022 | 7.9 | 1.89 | 9.7 | 8.8 | 0.899 | 1767473.0 |

Cuadro 1: Variables Sintomáticas para cálculo de estimaciones.Provincia de Buenos Aires

8.2. Metodologías Tradicionales

8.2.1. Regresión Lineal

En una primera aproximación se utilizó la regresión lineal. Esta técnica estadística se utiliza para entender la relación entre dos variables. Una de estas variables es conocida como la variable independiente (X), y la otra es la variable dependiente (Y). La idea principal es encontrar una línea recta que mejor se ajuste a los datos. Esta línea recta se llama "línea de regresión". La regresión lineal busca minimizar la distancia vertical entre cada punto de datos y la línea de regresión. Esto significa que la línea de regresión pasa lo más cerca posible de todos los puntos de datos. En este caso se utilizó como variable independiente el **Año** del censo y como variable dependiente obviamente la **Población**.

8.2.2. Estimaciones INDEC

Se considerará también las proyecciones realizadas por el INDEC que indican la "Población estimada al 1 de julio de cada año calendario por sexo, según partido. Provincia de Buenos Aires. Años 2010-2025"(INDEC, 2015)[12]. Estas proyecciones se apoyan en el crecimiento intercensal que comprende la evolución en conjunto de las tres variables básicas del análisis demográfico: la fecundidad, la mortalidad y la migración del período 2001–2010. Los resultados de las estimaciones departamentales de población son coherentes y consistentes con las proyecciones de población nacionales Serie de análisis demográfico N 35 (INDEC, 2013a) y provinciales vigentes (INDEC, 2013b)(INDEC, 2015)[12].

8.3. Metodologías Data Minings

8.3.1. CART, Random Forest y LightGBM

Si bien en este caso los datos son dispersos ("sparse"), se aplicarán árboles de decisión simples, random forest y otros algoritmos de clasificación tal cual lo expuesto en el artículo "Exploring New Models for Population Prediction in Detecting Demographic Phase Change for Sparse Census Data"(Chawda, Rane y Giri, 2018)[4].

Los árboles de decisión son una técnica de aprendizaje automático y análisis de datos que se utiliza tanto para problemas de clasificación como de regresión. Su objetivo es tomar decisiones o hacer predicciones basadas en una serie de preguntas o condiciones. Debido al sesgo habitual en los árboles individuales se analiza también utilizando Random Forest. Es una técnica de aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y la robustez de las predicciones. Es un método de ensamble, lo que significa que construye varios modelos y combina sus resultados para obtener una predicción final.

Por último se utiliza LightGBM es un framework de aprendizaje automático que utiliza el algoritmo de boosting para construir modelos predictivos. Boosting es una técnica de ensamble que crea un fuerte modelo predictivo a partir de una combinación de modelos más débiles, generalmente árboles de decisión. Es una implementación del algoritmo de gradient boosting diseñado para ser eficiente y escalable, tanto en términos de tiempo de entrenamiento como de uso de memoria.

8.4. Herramientas

Para el guardado y manipulación de los datos se va a trabajar con una base de datos POSTGRES (Local). Se desarrolló un proceso de ETL en Python y SQL mediante primera instancia crear todas las tablas necesarias, para luego producir la ingesta y transformar los diferentes inputs que van a componer el modelo de datos.

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

9. Obtención de datos fuente y Análisis exploratorio de datos

En primera instancia se busca recopilar los datos de los censos poblacionales correspondientes desde 1991 elaborados por el INDEC (Instituto Nacional de Estadísticas y Censos, Argentina). Se obtuvieron desde los repositorios web del INDEC los datos correspondientes a los censos de 1991, 2001, 2010 y 2022(UN-CEPAL, 2023)[2](CEPAL, 2023)[3](INDEC, 2022)[10]. En general el INDEC 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, presencia de espacios en blanco no uniformes, así como presentan descripciones y comentarios justamente en las primeras líneas del archivo. En general no respetan formatos uniformes según la dimensión de análisis (población por Sexo, población por edad, etc) de los datos dentro del mismo período censal. Asimismo los formatos de presentación cambian de censo a censo. Por otra parte, el formato de presentación de los archivos ".csv" así como las variables relevadas se han visto modificadas desde el censo 1991 hasta el 2022. Por ejemplo el censo 1991 no relevó la cantidad de habitantes en viviendas colectivas y particulares.

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. Se removieron espacios en blanco y comentarios en la parte superior, eliminaron columnas a la izquierda. Luego fue necesario unificar y adaptar todos los archivos a la misma cantidad y nomenclatura de columnas. Siempre buscando lograr un archivo ".csv" con la estructura de "Headers" seguido de los datos, sin espacios en blanco.

Es importante aclarar que existe una base consolidada -(CEPAL, 2023)[3]-(CEPAL, 2023)[3] con toda la información censal para los censos 2001 y 2010, la misma lamentablemente no está disponible para los censos de 1991 y 2022. Trabaja con REDATAM 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.

Durante la fase de investigación de fuentes se analizó y manipuló este software sin resultados alentadores. Esta base sólo puede accederse mediante este software específico, no disponiendo de back ups de la base de datos accesible y en general las salidas de las consultas manuales por interfase de usuario son también archivos ".csv" que requieren post-proceso.

Finalmente las variables censales definidas, comunes al período 1991-2022, se indican en el cuadro 2.

9.1. Dimensión Departamento- Slowly Changing Dimension

Al analizar los distintos censos desde 1991 hasta 2022 se observa claramente que los departamentos del AMBA han modificado su división político-administrativa en este período. Es decir, se han creado partidos nuevos, así como en otros casos partidos han cedido superficie, o bien se han fusionado otros. Esto general que la dimensión 'departamento'

Cuadro 2: Descripción de Variables

| Variable | Description |
|--------------|--|
| 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. |

cambia a particularmente desde el censo 1991 al censo 2001.

Se analiza el enfoque **Slowly Changing Dimensions** para bases de datos. En general se trata de impactar cambios en las dimensiones del modelo. Este concepto fue introducido por Kimball(Kimball, 1996)[14], el mismo se refiere a casos donde las dimensiones sufren pequeñas modificaciones que afectan el modelo lógico del almacén de datos.

El caso de Slowly Changing Dimensions **TIPO 1** corresponde a cambios esporádicos, no permanentes, que implican una modificación en algún atributo de la dimensión. En definitiva es una corrección de este atributo, se ejecuta mediante UPDATE de las filas involucradas sin conservar el valor anterior del atributo. La instancia de Slowly Changing Dimension **TIPO 2** esta basada en la teoría de datos temporales, donde para la versión que es afectada se produce un versionado. Es decir, se inserta una nueva tupla cada vez que se produce un cambio. Se extiende la dimensión generalmente con columnas "FROM" y "TO" que indican fechas de validez para los atributos correspondientes a ese valor de la dimensión. Obviamente esto afecta las consultas y todos los procesos de ETL involucrados en el almacén de datos. Por último se habla de Slowly Changing Dimension **TIPO 3** cuando si bien se desea mantener el cambio y relevan el mismo sin sobrescribir la tupla, no se mantiene el historial de la dimensión. En este caso sólo se almacena el estado actual y pasado de la misma. Simplemente se agrega una columna por cada atributo sujeto a cambios.

Es por esto que utiliza en este caso para la dimensión Departamento el enfoque **SCD 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 cuanto a la granularidad utilizada. Esto resulta particularmente importante para analizar las superficies de los departamentos.

En lo que respecta a las variables poblacionales se han reclasificado los datos de 1991 para corresponder con la división político-administrativa de año 2001. 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" (INDEC, 2023)[9].

En el cuadro 3 puede verse la dimensión departamento con los campos agregados y los comentarios correspondientes en cada caso.

9.2. Proyecciones INDEC 2010-2025

Se incorpora también al conjunto de datos que se analizarán proyecciones realizadas por el INDEC que indican la "Población estimada al 1 de julio de cada año calendario por sexo, según partido. Provincia de Buenos Aires. Años 2010-2025" (INDEC, 2015)[12]. Dichas estimaciones se apoyan en el crecimiento intercensal que comprende la evolución en conjunto de las tres variables básicas del análisis demográfico: la fecundidad, la mortalidad y la migración del período 2001–2010. Las mismas se utilizarán para comparar los resultados de los censos, acompañadas de proyecciones realizadas por otros métodos.

La descripción de este conjunto puede verse en el cuadro 4, mientras el cuadro 5 muestra la estructura típica del dataset. En forma gráfica se dividen los departamentos según la población inicial para el año 2010. Se gráfica por un lado aquellos departamentos con una población menor a 300.000 habitantes, figura 1. Luego aquellos con una

Cuadro 3: Dimensión Departamento

| Codigo | Departamento | PartidoFrom | Codigo From | Sup1991 | Sup2001 | Comentarios |
|--------|---------------------|--------------------|-------------|---------|---------|---|
| 6028 | Almirante Brown | | | 122.0 | 122.0 | |
| 6035 | Avellaneda | | | 55.0 | 55.0 | |
| 6091 | Berazategui | | | 188.0 | 188.0 | |
| 6260 | Esteban Echeverría | | | 377.0 | 120.0 | (1) Superficie modificada, cede tierras a Cañuelas y San Vicente y para la creación de Ezeiza y Presidente Perón. Leyes provinciales 11.550 del 20/10/1994 y 11.480 del 25/11/1993. |
| 6270 | Ezeiza | Esteban Echeverría | 6260 | 0.0 | 223.0 | (2) Se crea con tierras del partido de Esteban Echeverría. Ley provincial 11.550 del 20/10/1994. |
| 6274 | Florencio Varela | | | 206.0 | 190.0 | (3) Sup.modificada, cede tierras a Presidente Perón. Ley 11.480 del 25/11/1993. |
| 6371 | General San Martín | | | 56.0 | 56.0 | |
| 6408 | Hurlingham | Morón | 6568 | 0.0 | 36.0 | (4) Se crea con tierras del partido de Morón. Ley provincial 11.610 del 28/12/1994. |
| 6410 | Ituzaingó | Morón | 6568 | 0.0 | 39.0 | (5) Se crea con tierras del partido de Morón. Ley provincial 11.610 del 28/12/1994. |
| 6412 | José C. Paz | General Sarmiento | 6005 | 0.0 | 50.0 | (6) Se crea con tierras del partido de General Sarmiento. Ley provincial 11.551 del 20/10/1994. |
| 6427 | La Matanza | | | 323.0 | 323.0 | |
| 6434 | Lanús | | | 45.0 | 45.0 | |
| 6490 | Lomas de Zamora | | | 89.0 | 89.0 | |
| 6515 | Malvinas Argentinas | General Sarmiento | 6005 | 0.0 | 63.0 | (7) Se crea con tierras del partido de General Sarmiento e incorpora un sector del partido de Pilar. Ley provincial 11.551 del 20/10/1994. |
| 6539 | Merlo | | | 170.0 | 170.0 | |
| 6560 | Moreno | | | 180.0 | 180.0 | |
| 6568 | Morón | | | 131.0 | 56.0 | (8) Superficie modificada, cede tierras para la creación de los partidos de Hurlingham e Ituzaingó. Ley 11.610 del 28/12/1994. |
| 6658 | Quilmes | | | 125.0 | 125.0 | |
| 6749 | San Fernando | | | 924.0 | 924.0 | |
| 6756 | San Isidro | | | 48.0 | 48.0 | |
| 6760 | San Miguel | General Sarmiento | 6005 | 0.0 | 80.0 | (9) Se crea con tierras del partido de General Sarmiento. Ley provincial 11.551 del 20/10/1994. |
| 6805 | Tigre | | | 360.0 | 360.0 | |
| 6840 | Tres De Febrero | | | 46.0 | 46.0 | |
| 6861 | Vicente López | | | 39.0 | 39.0 | |
| 6005 | General Sarmiento | | | 196.0 | 0.0 | Desaparece para el año 2001 |

población de entre 300.000 y 700.000, figura 2. Por último La Matanza cuya población supera ampliamente los 700.000 habitantes, figura 3

| Column | Non-Null Count | Dtype |
|--------------|----------------|--------|
| id | 384 non-null | int64 |
| CodigoDpto | 384 non-null | object |
| ano | 384 non-null | int64 |
| Departamento | 384 non-null | object |
| Poblacion | 384 non-null | int64 |
| Varones | 384 non-null | int64 |
| Mujeres | 384 non-null | int64 |

Cuadro 4: Summary del dataset public.proyecciones.

| id | CodigoDpto | ano | Departamento | Poblacion | Varones | Mujeres |
|-----------|-------------------|------------|---------------------|------------------|----------------|----------------|
| 1 | 6028 | 2010 | Almirante Brown | 557025 | 273352 | 283673 |
| 25 | 6028 | 2011 | Almirante Brown | 561349 | 275570 | 285779 |
| 49 | 6028 | 2012 | Almirante Brown | 565509 | 277794 | 287715 |
| 73 | 6028 | 2013 | Almirante Brown | 569911 | 279980 | 289931 |
| 97 | 6028 | 2014 | Almirante Brown | 574263 | 282143 | 292120 |
| 121 | 6028 | 2015 | Almirante Brown | 578513 | 284281 | 294232 |
| 145 | 6028 | 2016 | Almirante Brown | 582541 | 286295 | 296246 |
| 169 | 6028 | 2017 | Almirante Brown | 586564 | 288288 | 298276 |
| 193 | 6028 | 2018 | Almirante Brown | 590418 | 290157 | 300261 |
| 217 | 6028 | 2019 | Almirante Brown | 594270 | 292051 | 302219 |
| | | | | | | |

Cuadro 5: Proyecciones del valor población por departamento Año 2010 a 2025 .Fuente INDEC

9.3. 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, 2024)[8], de donde se obtiene la capa SIG para todos los departamentos de la República Argentina. Luego se cruza con el listado de departamentos del AMBA, generando un dataset con las características geográficas sumadas a los datos censales recopilados desde 1991 a 2022. La capa geográfica incorpora las variables que se detallan en el cuadro 6.

9.4. Población

Se detallan a continuación los datos recopilados según nivel de agregación. En primera instancia se generó un dataset con todos los departamentos del AMBA, sus características geográficas, así como los valores de las variables censales correspondientes a los censos 1991-2001-2010-2022. En el cuadro 7 se puede observar el esquema del dataset.

El resumen del Dataset "AMBA Censo" se observa en el cuadro 8. Se puede notar que existen valores nulos para el valor de población, que se corresponden con los casos de municipios de desaparecen en la reorganización administrativa desde 1991 a 2001. A saber: **Esteban Echeverría, Ezeiza, Florencio Varela, Hurlingham, Ituzaingó, José C. Paz, Malvinas Argentinas, Morón, San Miguel y General Sarmiento**.

Se destaca también que para el censo 1991 no se detalla el total de viviendas particulares y colectivas.

9.4.1. Población. Vista geográfica

En este caso se ofrece una visualización geográfica del dataset. Utilizando el software QGIS conectado directamente a la base de datos AMBA (POSTGRES) se pudo disponer la información de cada Departamento con su geolocalización. Se detalla en la figura 4 la población total para cada departamento (Censos 1991 -2022). Se observa claramente que el distrito más poblado es La Matanza.

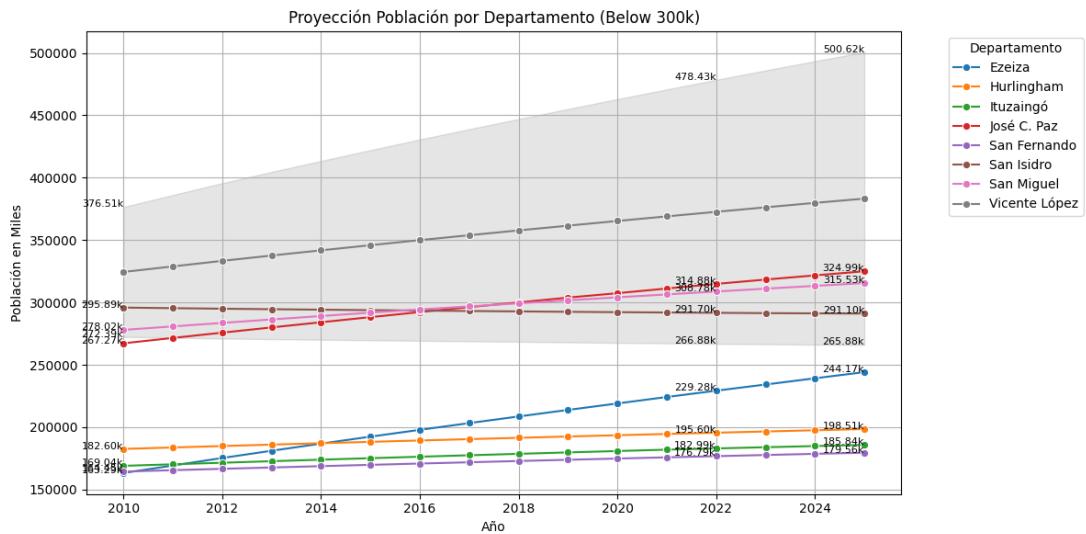


Figura 1: Población por departamento 2010-2025.INDEC

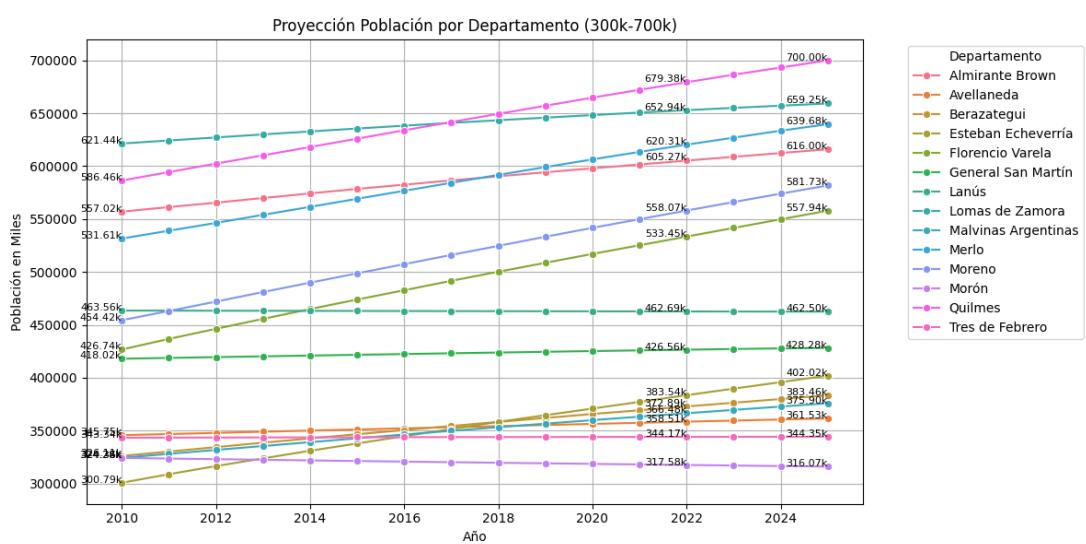


Figura 2: Población por departamento 2010-2025.INDEC

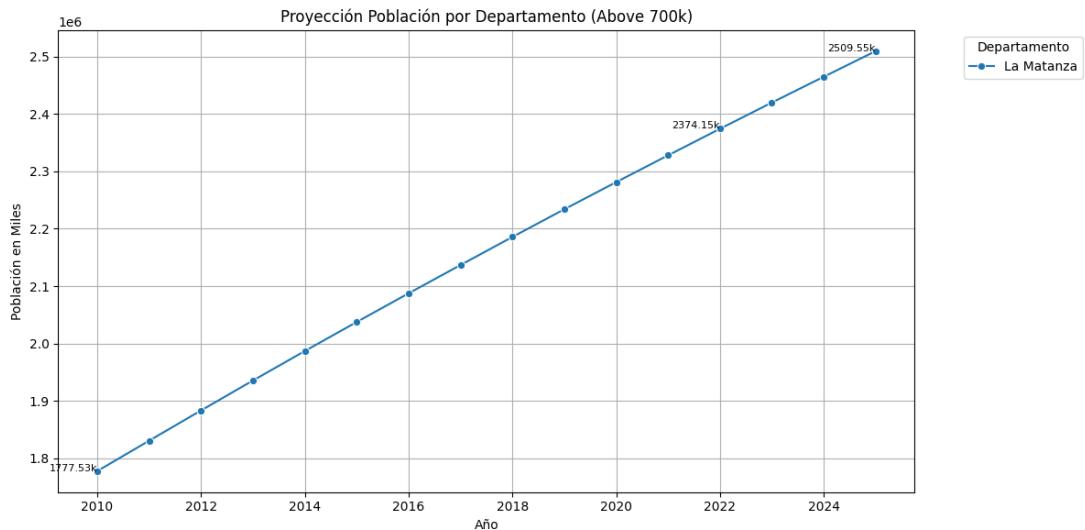


Figura 3: Población por departamento 2010-2025. INDEC

Cuadro 6: Descripción de Variables Geográficas

| Variable | Nombre | Descripción |
|----------|---------------------|---|
| geom | geometry | Polígono WKT con los límites del departamento |
| fna | Nombre geográfico | Nombre completo que se utiliza para designar un objeto en un mapa o carta. Está formado por el término genérico y el término específico. Ejemplo: río Mendoza. |
| gna | Término genérico | Parte del nombre geográfico que indica el tipo de objeto que identifica. Ejemplo: río, monte, glaciar, establecimiento. |
| nam | Término específico | Parte de un nombre geográfico que acompaña al término genérico y que identifica e individualiza un objeto geográfico determinado. Ejemplo: Paraná en río Paraná; Upsala en glaciar Upsala; Las Marías en establecimiento Las Marías; Esperanza en el caso de bahía Esperanza. |
| in1 | Código INDEC | Código único de vías de circulación asignado por el Instituto Nacional de Estadística y Censos de la República Argentina. |
| fdc | Fuente de captura | Identificación del nombre y tipo de fuente utilizada para capturar la información. Puede incluir fecha y otros datos adicionales. |
| sag | Autoridad de fuente | Nombre de la autoridad responsable de la información utilizada. |

| nam | cod_dept | anio | pob | var | muj | vivpart | vivtotal | sup | ind_masc | dens_pob |
|-----------------|-----------------|-------------|------------|------------|------------|----------------|-----------------|------------|-----------------|-----------------|
| Almirante Brown | 06028 | 1991 | 450698.0 | 222042.0 | 228656.0 | nan | nan | 157.87 | 97.1 | 2854.87 |
| Almirante Brown | 06028 | 2001 | 515556.0 | 252454.0 | 263102.0 | 143543.0 | 88.0 | 157.87 | 96.0 | 3265.70 |
| Almirante Brown | 06028 | 2010 | 552902.0 | 270247.0 | 282655.0 | 156218.0 | 78.0 | 157.87 | 95.6 | 3502.26 |
| Almirante Brown | 06028 | 2022 | 585852.0 | 281842.0 | 301779.0 | 184403.0 | 60.0 | 157.87 | 93.4 | 3710.98 |
| Avellaneda | 06035 | 1991 | 344991.0 | 164243.0 | 180748.0 | nan | nan | 68.54 | 90.9 | 5033.43 |
| Avellaneda | 06035 | 2001 | 328980.0 | 155450.0 | 173530.0 | 117200.0 | 59.0 | 68.54 | 89.6 | 4799.82 |
| Avellaneda | 06035 | 2010 | 342677.0 | 162264.0 | 180413.0 | 121307.0 | 68.0 | 68.54 | 89.9 | 4999.66 |
| Avellaneda | 06035 | 2022 | 370939.0 | 174572.0 | 194911.0 | 144988.0 | 64.0 | 68.54 | 89.6 | 5412.01 |
| | | | | | | | | | | |

Cuadro 7: Dataset Censos AMBA

Cuadro 8: Summary de columnas de Censos AMBA

| Column | Non-Null Count | Dtype |
|---------------|-----------------------|--------------|
| nam | 96 | object |
| cod_dept | 96 | object |
| anio | 96 | object |
| pob | 90 | float64 |
| var | 90 | float64 |
| muj | 90 | float64 |
| vivpart | 72 | float64 |
| vivtotal | 72 | float64 |
| sup | 96 | object |
| ind_masc | 90 | object |
| dens_pob | 90 | object |

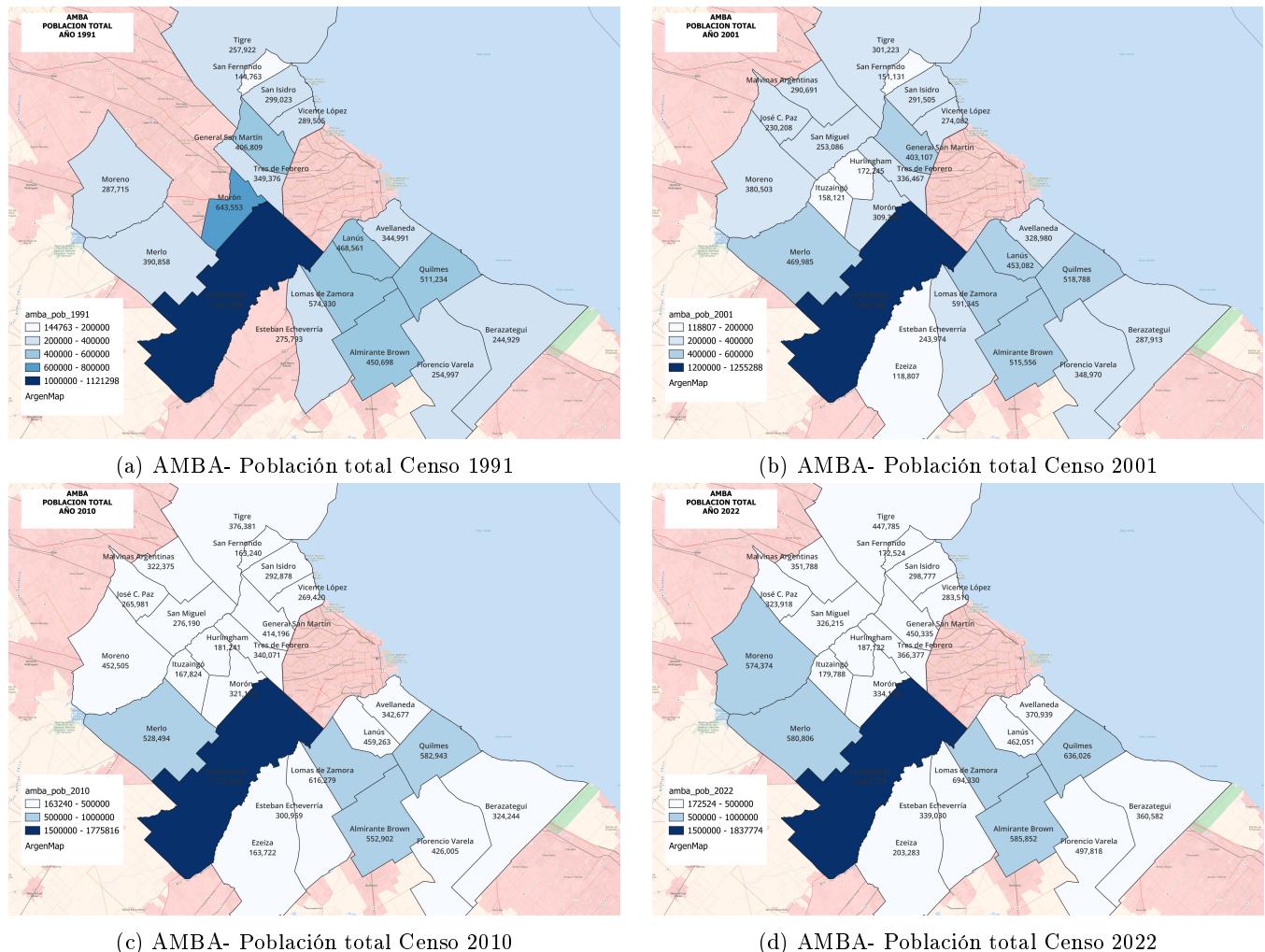


Figura 4: Población Total Censos 1991 2022

9.4.2. Población. Ratios de crecimiento

Se determinaron los ratios de crecimiento entre censos consecutivos para cada departamento. Para este análisis se trata de manera difrenciada aquellos municipios que han sufrido modificaciones administrativas desde 1991 a 2001: **Esteban Echeverría, Ezeiza, Florencio Varela, Hurlingham, Ituzaingó José C. Paz, Malvinas Argentinas, Morón, San Miguel y General Sarmiento.** Donde al desmembrarse en varios partidos, resignando superficie, algunas tasas o ratios de crecimiento se muestran negativos. Asimismo partidos que han visto incrementada su superficie presentan tasas de crecimiento no representativas del fenómeno demográfico.

Por otra parte partidos que desaparecen o se crean en el año 2001 no se consideran el valor del ratio para el intervalo 1991-2001.

En el cuadro 9 se puede observar los valores obtenidos. Para el año en cuestión la tasa surge de dividir el valor poblacional del censo de este año sobre el valor poblacional del censo anterior, expresado en porcentaje.

Es decir:

$$\text{Growth Ratio} = \left(\frac{\text{Población}_{(\text{Censo } X)}}{\text{Población}_{(\text{Censo } X - 1)}} - 1 \right) \times 100 \quad (1)$$

| nam | anio | pob | growth_ratio |
|--------------------|------|----------|--------------|
| Almirante Brown | 2001 | 515556.0 | 14.39 |
| Almirante Brown | 2010 | 552902.0 | 7.24 |
| Almirante Brown | 2022 | 585852.0 | 5.96 |
| Avellaneda | 2001 | 328980.0 | -4.64 |
| Avellaneda | 2010 | 342677.0 | 4.16 |
| Avellaneda | 2022 | 370939.0 | 8.25 |
| Berazategui | 2001 | 287913.0 | 17.55 |
| Berazategui | 2010 | 324244.0 | 12.62 |
| Berazategui | 2022 | 360582.0 | 11.21 |
| Esteban Echeverría | 2001 | 243974.0 | -11.54 |
| Esteban Echeverría | 2010 | 300959.0 | 23.36 |
| Esteban Echeverría | 2022 | 339030.0 | 12.65 |
| | | | |

Cuadro 9: Extracto. Tasa de crecimiento Intercensal

A partir de análisis estadístico simple sobre estas tasas se observa una gran disparidad en los valores, siendo que se trata de un sector geográfico de similares características socio-demográficas. Cuadro 10

| Estadístico | Valor |
|--------------------|-------|
| Count | 66 |
| Mean | 9.3 |
| Standard Deviation | 13.1 |
| Minimum | -51.9 |
| 25 % Percentile | 3.0 |
| 50 % Percentile | 8.1 |
| 75 % Percentile | 16.5 |
| Maximum | 41.5 |

Cuadro 10: Summary Statistics for Growth Ratio

A partir de la elaboración de un gráfico boxplot se determinaron los **Outliers** para estas tasas de crecimiento.

En este caso en la figura 5 se observan 4 outliers, tanto en valores positivos como negativos. Como se comentó anteriormente no se analizarán los casos con modificaciones administrativas desde 1991 a 2001. Cuadro 3. Se desestima el caso de Morón para el período 1991-2001 donde la tasa es negativa de 51.9 %, debido a la cesión de tierras para la creación de distintos partidos.

Se grafica a continuación (Figura 6) la curva población de los 3 outliers destacados :Florencio Varela ,La Matanza y Ezeiza.

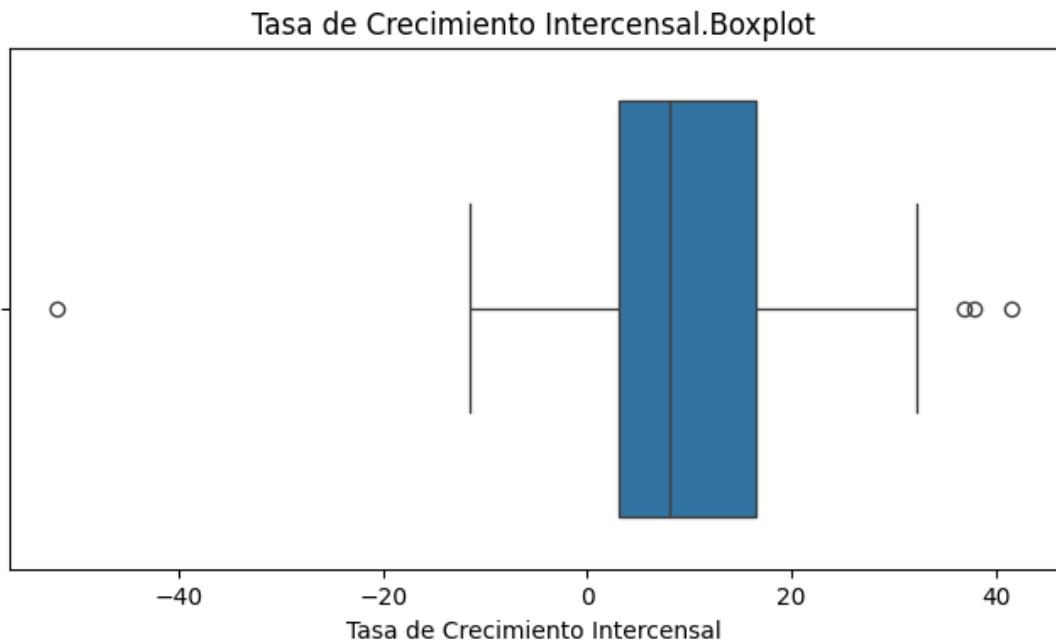


Figura 5: Box plot Tasas de crecimiento intercensal 1991-2022

| id | nam | codDept | anio | pob | growthRatio |
|-----------|------------------|----------------|-------------|------------|--------------------|
| 54 | Morón | 06568 | 2001 | 309380.0 | -51.92 |
| 78 | Florencio Varela | 06274 | 2001 | 348970.0 | 36.85 |
| 13 | La Matanza | 06427 | 2010 | 1775816.0 | 41.47 |
| 29 | Ezeiza | 06270 | 2010 | 163722.0 | 37.80 |

Cuadro 11: Outliers para Tasas de crecimiento intercensal

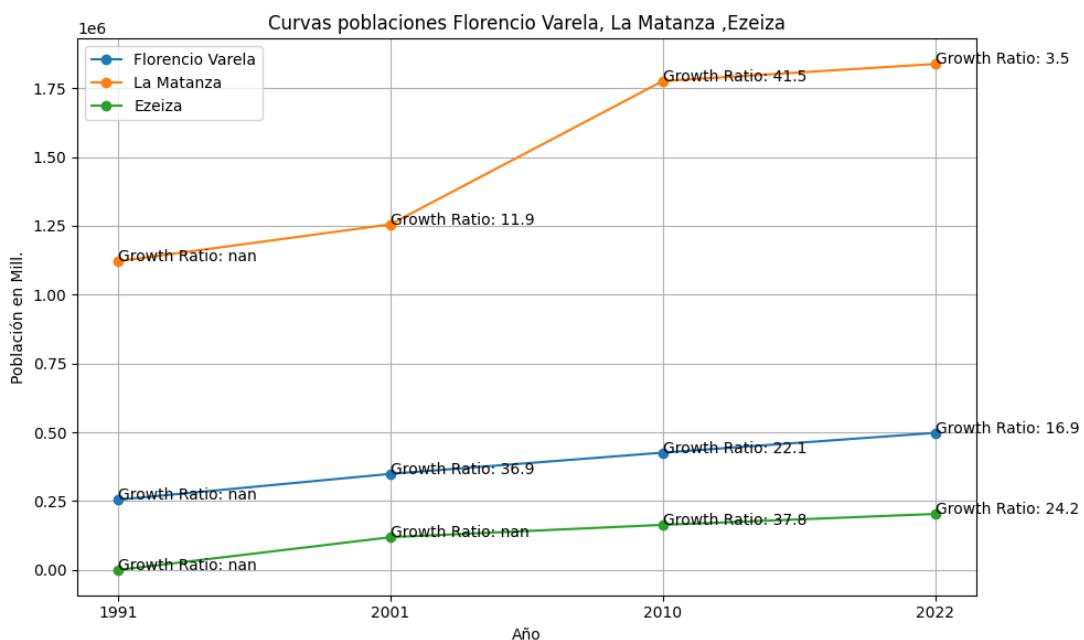


Figura 6: Curvas de Población por departamento 1991-2022

9.4.3. Tasas de crecimiento intercensal.Coefficiente de Variación

Con el objetivo de profundizar el análisis de las tasas de crecimiento se procedió a calcular luego la desviación estandar y el coeficiente de variación por Departamento.

En primera instancia se agrupa por departamento para determinar la media y la desviación estándar de cada subset de datos. Se define el coeficiente de variación (CV) como:

$$CV_{(Departamento)} = \left(\frac{\text{Std.Dev}_{(Departamento)}}{\text{Mean}_{(Departamento)}} \right) \times 100 \quad (2)$$

El coeficiente de variación es una medida de la dispersión alrededor de la media de la población. En este caso sería curvas poblacionales con gran dispersión en su tasas de crecimiento para los censos del periodo analizado año 1991 -2022.-

En la figura 7 se observa el resultado .Se destaca en particular los valores extremos tanto negativos como positivos.A saber : Vincente López, Morón, Tres de Febrero, Avellaneda ,La Matanza , Esteban Echeverría, General San Martín. Cabe recordar en este instancia que los partidos de Esteban Echeverría, Florencio Varela y Morón sufrieron modificaciones administrativa en el período 1991-2001. En la figura 8 así como en la figura 9 puede observarse la curvas poblacionales de los departamentos destacados.

En estos casos se prestará particular atención a las predicciones hechas por el INDEC así como las metodologías que se implementen para la estimación de la curva poblacional, es esperable un mayor error de predicción en estos departamentos cuyas curvas prensitan singularidades respecto a los municipios inmediatamente aledaños.

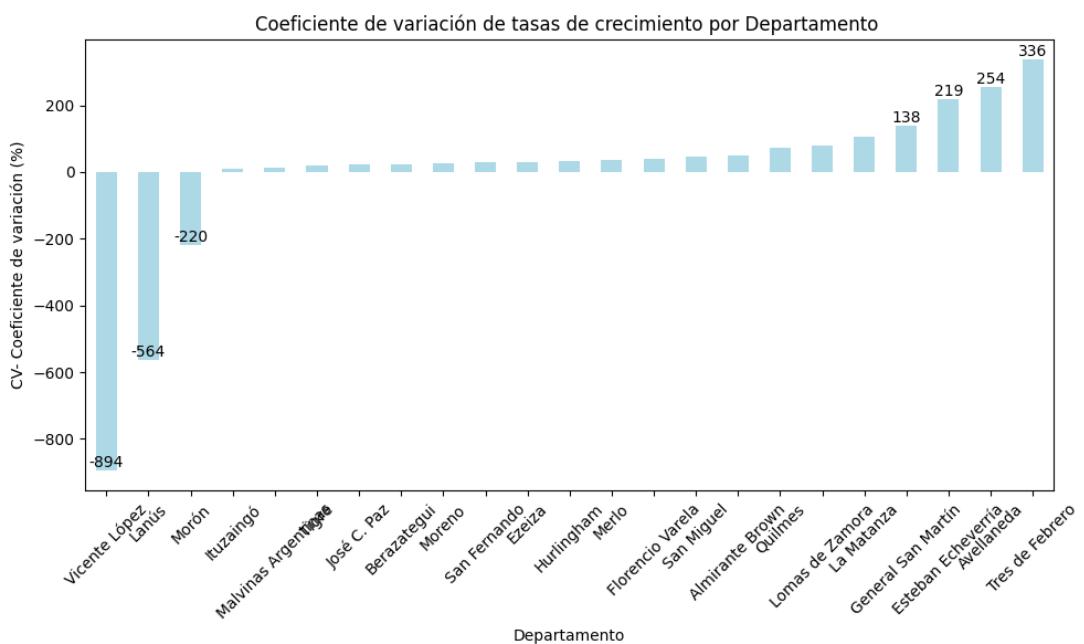


Figura 7: CV.Coefficiente de variación por departamento -1991-2022

9.5. Densidad de Población

En el dataset se incorpora la densidad de población en habitantes/km².Si analizamos la distribución de los valores de densidad para cada censo, puede observarse distribuciones homogeneas simétricas,con un incremento en al mediana de cada población censal desde 2001 a 2022. El análisis univariado de esta variable puede observarse en la figura 10 y la figura 11.

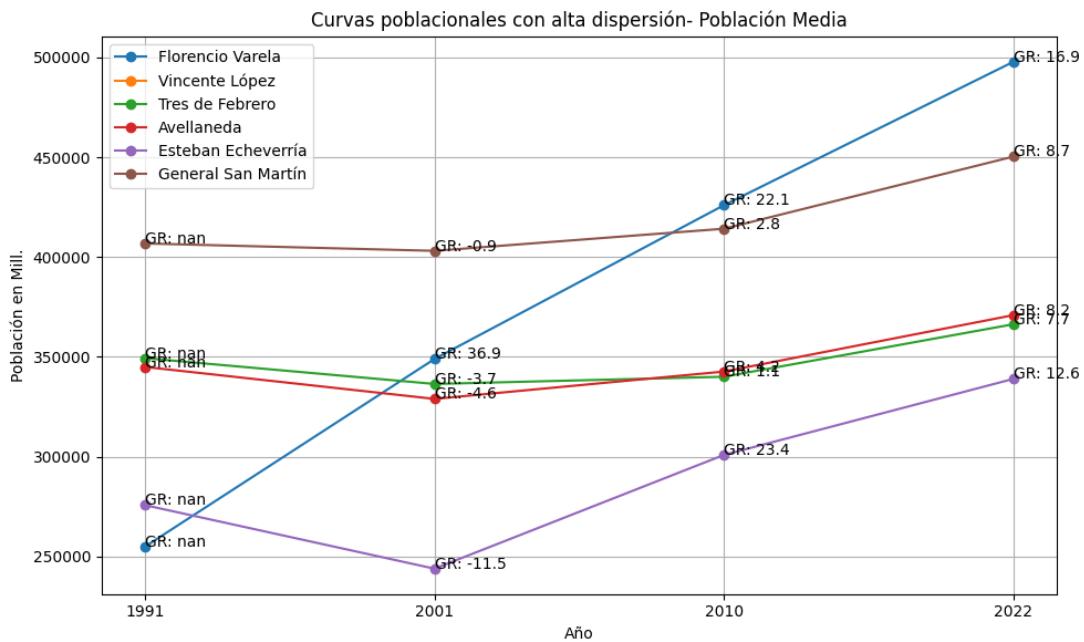


Figura 8: Curvas poblacionales de alta dispersión - Población Media

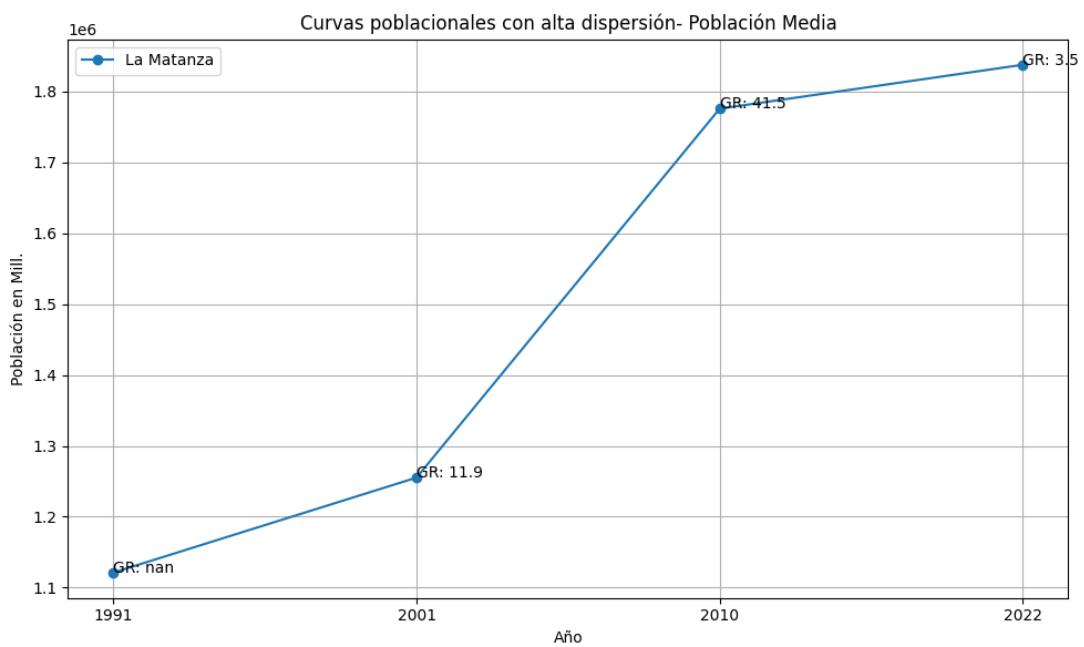


Figura 9: Curvas poblacionales de alta dispersión - Población Alta

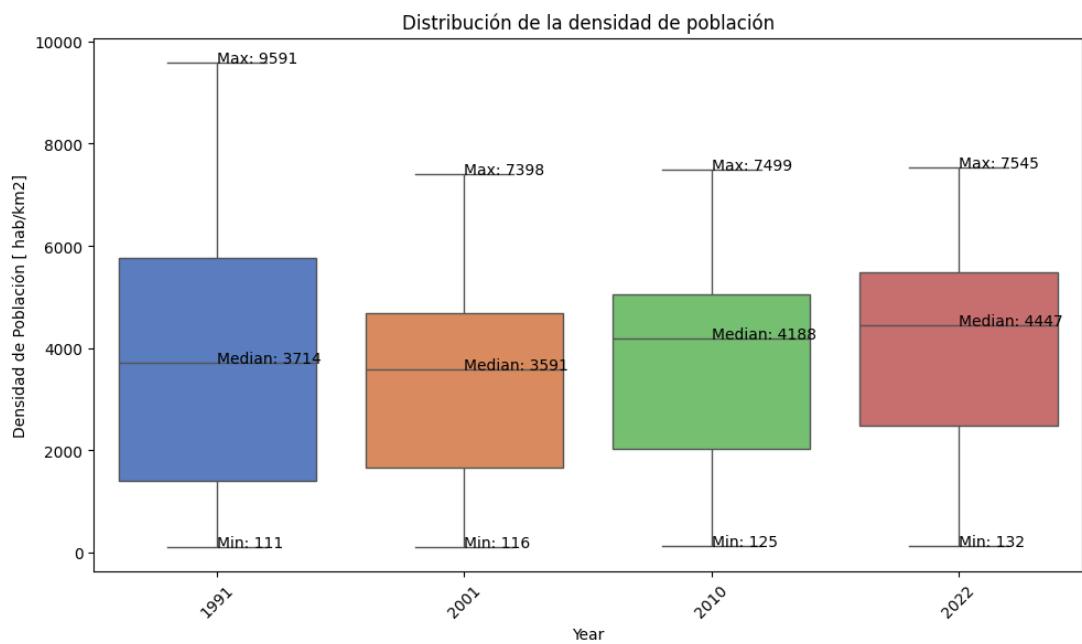


Figura 10: Densidad de Población. Análisis Univariado

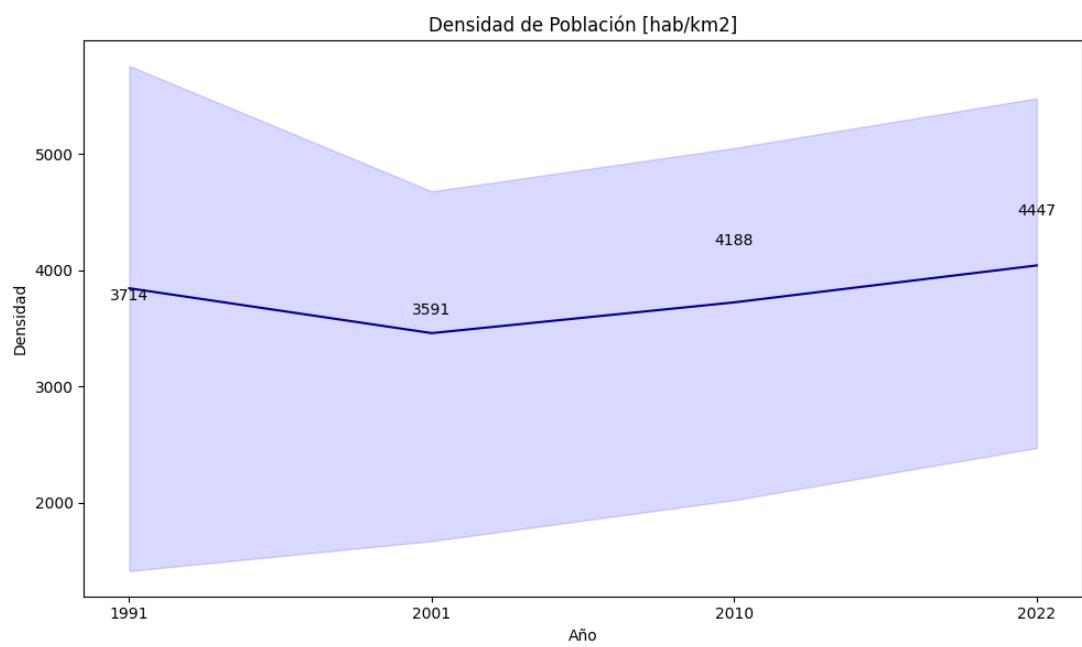
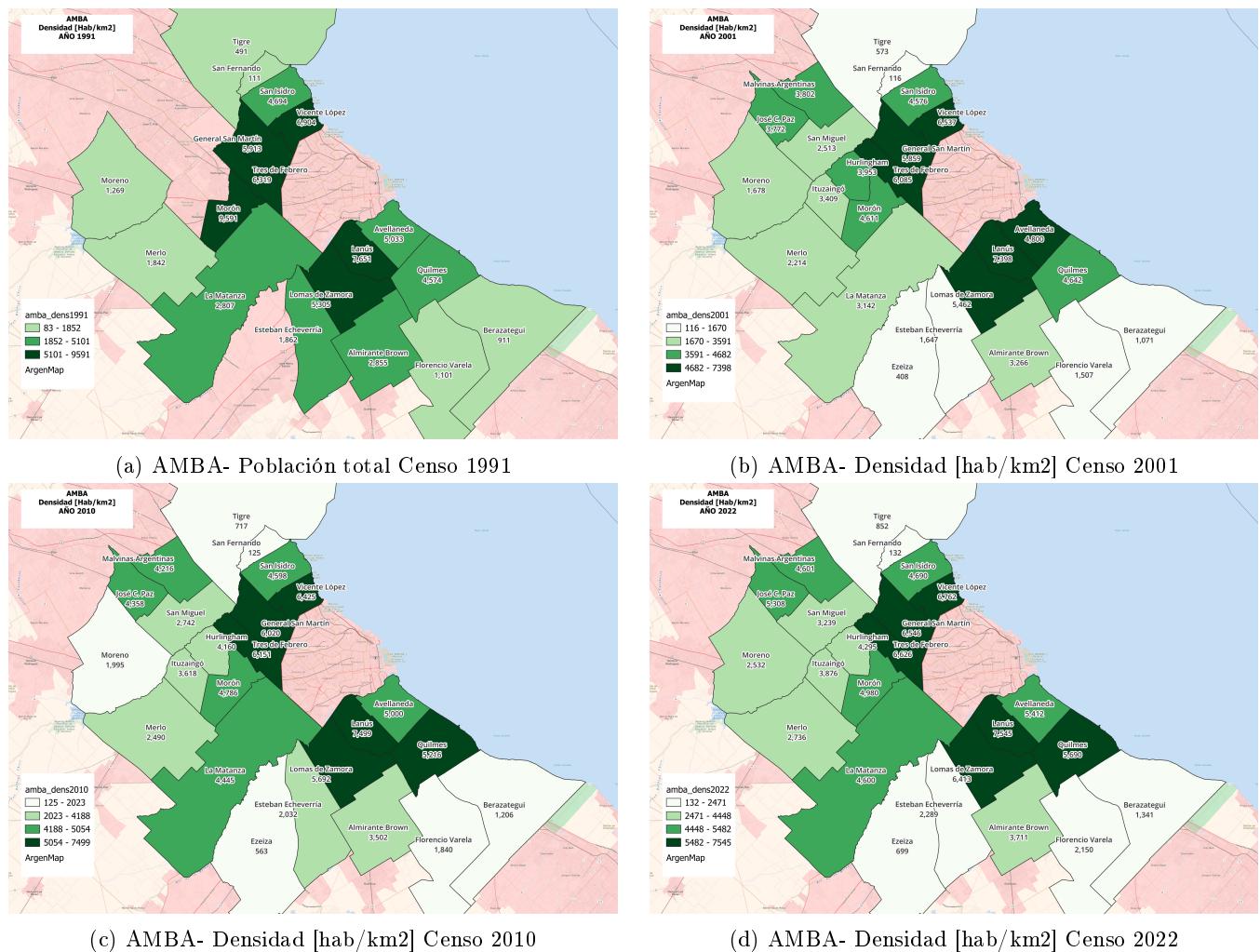


Figura 11: Densidad de Población. Análisis Univariado

9.5.1. Densidad de Población. Vista geográfica

Se presenta también la distribución geográfica de la densidad por departamento para los censos analizados. Si bien el distrito más poblado es La Matanza, debido su extensión no es el departamento con mayor densidad poblacional en [hab/km²] figura 12, los distritos más densamente poblados son Lanús y Vicente López.

Figura 12: AMBA- Densidad [hab/km²] Censos 1991 2022

9.6. Tipo de Vivienda

A partir del censo 2001 se incorpora como atributo descriptivo en cada censo la composición o tipo de la vivienda, ya sea de tipo particular o colectiva. Se detalla en cada caso la cantidad de viviendas particulares('vivpart') en determinado departamento, así como la cantidad de viviendas colectivas presentes. Se presenta entonces la posibilidad de realizar un análisis univariado del porcentaje de viviendas particulares respecto al total de viviendas para un determinado departamento y año censal.

Se agrega al dataset una nueva columna definida como:

$$\text{VivPart \%} = \frac{\text{VivPart}}{\text{VivPart} + \text{VivColTot}} \times 100 \quad (3)$$

Surge de inmediato que mayoría de la población vive en viviendas particulares (mayor a 99.9%). Al analizar el comportamiento de este indicador se observa un leve incremento sostenido en el tiempo desde 1991 hasta 2022. Es decir que se observan cada vez más peso de las viviendas particulares. El análisis univariado de este indicador puede observarse en la figura 13 así como en la figura 14

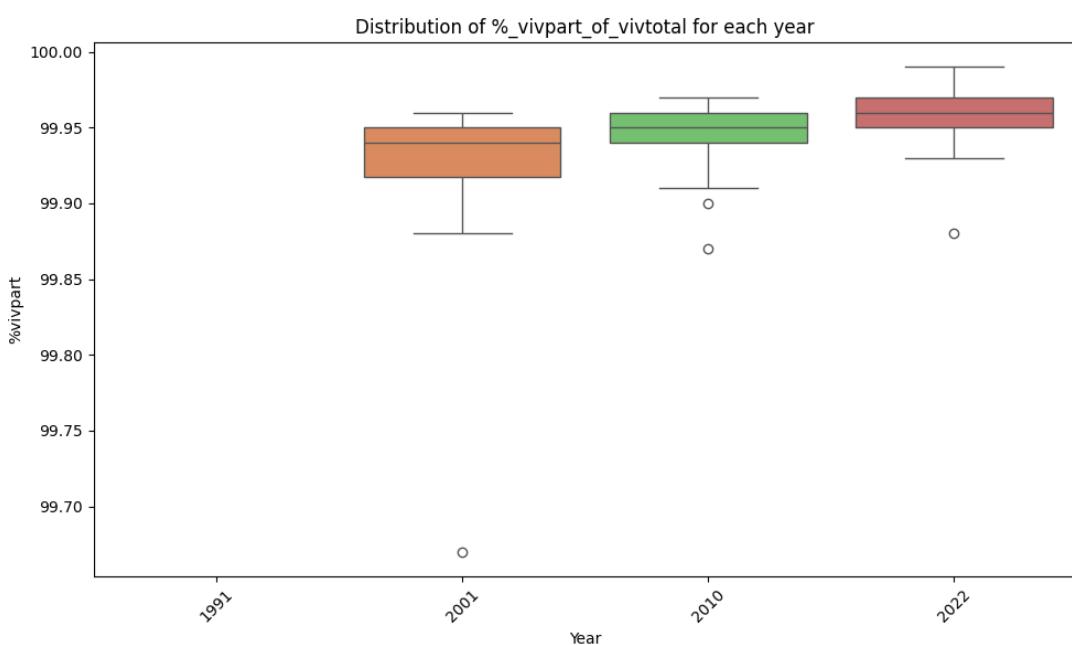


Figura 13: Composición de viviendas. Análisis Univariado

9.7. Índice de Masculinidad

Un indicador habitual de las muestras poblacionales es el índice de masculinidad. Resultado de dividir el número de hombres entre el número de mujeres de una unidad geográfica o administrativa, generalmente multiplicado por 100 expresado como el número de hombres por cada 100 mujeres.

Es decir:

$$\text{IndMasc} = \frac{\text{Varones}}{\text{Mujeres}} \times 100 \quad (4)$$

Al analizar el comportamiento del índice a lo largo del tiempo se observa un descenso sostenido del mismo desde 1991 hasta 2022. Particularmente el máximo de la muestra presenta un descenso de 5 puntos porcentuales para el año 2022, así como 2 puntos porcentuales la mediana.

Esto puede observarse en la figura 15 así como en la figura 16.

10. Compartiva de los Modelos de Predicción

En base a los censos 1991,2001 y 2010 ,sumado a las variables sintomáticas antes descriptas, se pretende ajustar y entrenar los modelos para luego predecir la población total de cada departamento del AMBA para el Año 2022. Al

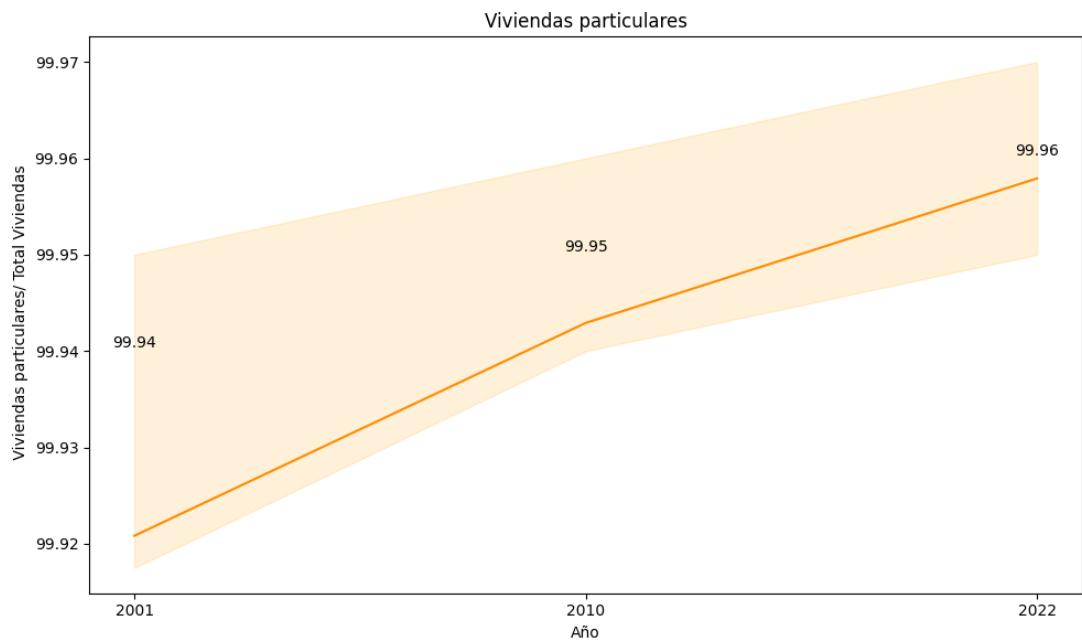


Figura 14: Composición de viviendas. Análisis Univariado

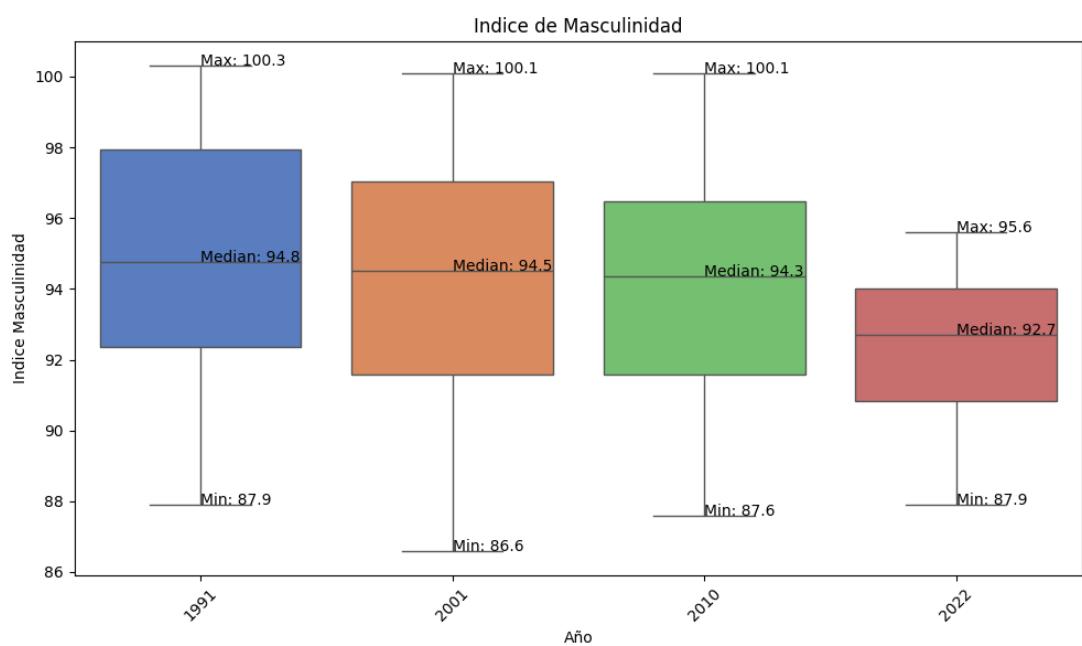


Figura 15: Índice de Masculinidad. Análisis Univariado

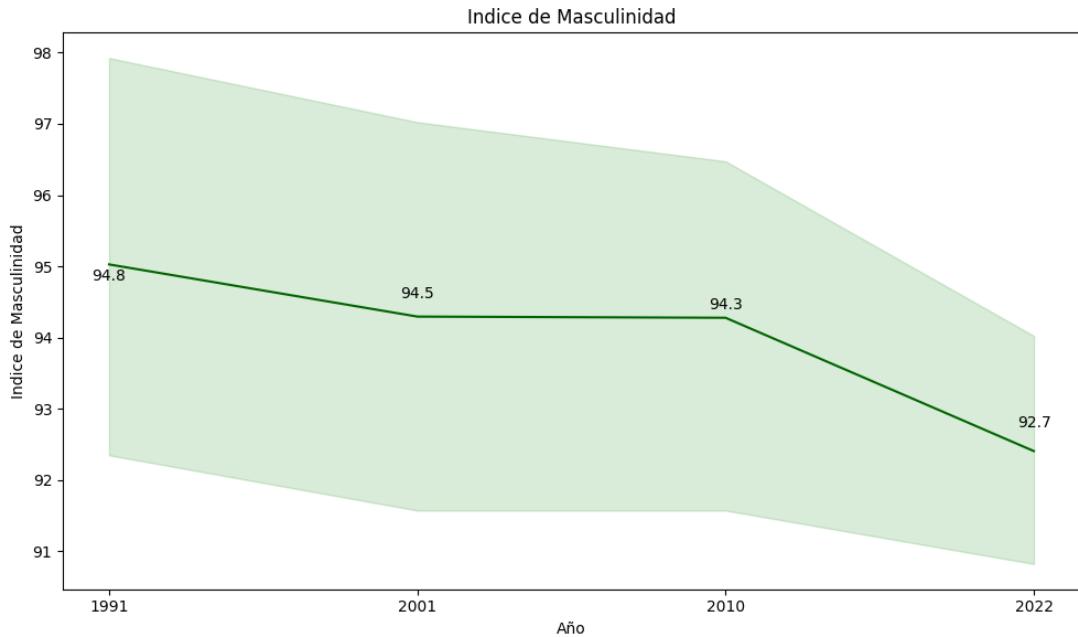


Figura 16: Índice de Masculinidad. Análisis Univariado

comparar con los resultados publicados por INDEC para el censo 2022 (INDEC, 2022)[10] se determinó la precisión de cada metodología aplicada.

Muchas de estas herramientas tiene problemas con los valores nulos o faltantes. Es por esto que los departamentos que no existían en 1991, demandaron un tratamiento especial.

10.1. Variables Censales

Respecto a las variables censales podemos decir que presentan una correlación lineal directa muy importante y no aportan variabilidad. Por este motivo no son atributos significativos para los modelos propuestos. Para su análisis se utilizó la matriz de correlación, la misma puede verse en la figura 17

10.2. Errores típicos

Para la determinar la precisión y realizar la comparativa entre los modelos aplicados se recurrió a los siguientes errores típicos.

El Error Cuadrático Medio (Mean Squared Error, MSE) es una medida de la calidad de un estimador. Se calcula promediando el cuadrado de los errores (diferencias entre los valores predichos y los valores reales).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Se calculó también la Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE) es la raíz cuadrada del MSE. Proporciona una medida de la magnitud promedio del error en las mismas unidades que los valores predichos.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Asimismo se determinó el Error Porcentual Absoluto Medio (Mean Absolute Percentage Error, MAPE) es una medida de precisión que expresa el error como un porcentaje. Se calcula promediando el valor absoluto de los errores

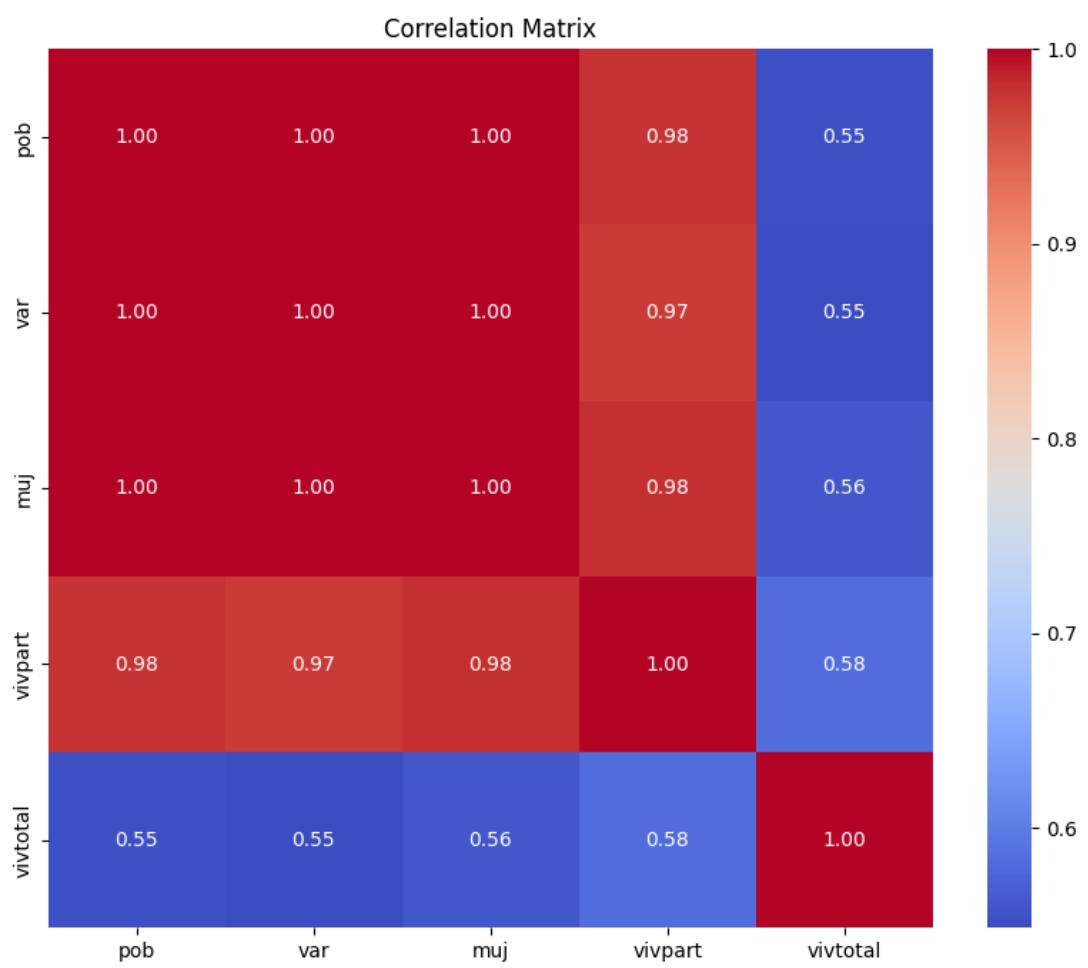


Figura 17: Matriz de correlación .Variables Censales

porcentuales.

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

10.3. Subset de Datos para Entrenamiento

Como se mencionó anteriormente se trabajó con los los censos de los años 1991 , 2001 y 2010. A este conjunto se le agregó el valor de las variables sintomáticas para la Jurisdicción provincial- Provincia de Buenos Aires para dichos años. Se toma el enfoque planteado por(2001)(Álvarez et al., 2001)[1], entendiendo que el comportamiento de estas variables puede ayudar a explicar fenómenos a nivel departamental.

Un esquema del dataset utilizado como input de los modelos puede verse en el cuadro 14.

10.4. Predicciones Población año 2022

En base al subset de datos predefinido, se entrenaron los distintos modelos para luego predecir el valor de población total en cada departamento para el año 2022. Las predicciones realizadas corresponde a Regresión Lineal, árbol de decisión (CART), Random Forest , LightGMB y también se analizan las estimaciones realizadas por el INDEC en el año 2010(INDEC, 2015)[12]. Para cada estimación se obtuvieron los errores típicos. Particularmente para la comparativa de los los modelos se trabaja con el Error Porcentual Absoluto Medio (Mean Absolute Percentage Error, MAPE).

Los resultados obtenidos implican que las metodologías tradicionales aproximan mejor este tipo de datos dispersos "sparse" poblacionales.Tanto la regresión Lineal , como las proyecciones realizadas por el INDEC presentan mejor precisión- menor error- y una menor desviación estándar. Este comportamiento se puede observar en la figura 18, así como el detalle de los estadísticos correspondientes se muestra en el cuadro 12. Las algoritmos de data mining presenta dificultades debido a las características particulares de la información censal ,así como el hecho de estar trabajando sólo con la información de 3(tres) Censos Nacionales.Esto sumado la granularidad analizada en este caso, nivel departamental,que hace difícil enriquecer el dataset con variables sintomáticas a este nivel y se debe recurrir a información agregada a nivel Provincial. Estos modelos presentan valores de error notablemente mayores y una amplia dispersión de resultados.

| Method | Q1 | Median | Q3 | IQR | Lower Bound | Upper Bound |
|-------------------|-----|--------|------|------|-------------|-------------|
| Linear Regression | 3.2 | 4.5 | 6.3 | 3.1 | -1.4 | 11.0 |
| Regression Tree | 6.2 | 10.5 | 17.6 | 11.5 | -11.1 | 34.9 |
| Random Forest | 9.1 | 14.1 | 18.2 | 9.1 | -4.5 | 31.8 |
| LightGBM | 9.4 | 15.5 | 24.7 | 15.3 | -13.5 | 47.6 |
| INDEC | 3.3 | 5.1 | 6.6 | 3.3 | -1.6 | 11.6 |

Cuadro 12: Estadísticos MAPE por Modelo

Los outliers para estos modelos se resumen en el cuadro 13. Si se analizan los modelos con mejor precisión, en Regresión Lineal tanto Esteban Echeverría(MAPE:11.2 %) como la Matanza(MAPE:41.6 %) son outliers, donde la predicción estuvo muy alejada del valor poblacional del censo 2022. Para el caso de las proyecciones según INDEC, aparecen también Esteban Echeverría (MAPE:13.1 %), Ezeiza (MAPE:12.8 %) y la Matanza (MAPE:29.2 %) como outliers. El caso de Esteban Echeverría tiene que ver con la cesión de territorio y cambios administrativos desde el año 1991 al 2001. Esto implicó que su población descendiera entre 1991 y 2001 un -11,5 % mientras que entre 2001 y 2010 creció un 23.3 %. La Matanza presenta entonces singularidades y se comporta distinto de los departamentos aledaños.

| Departamento | MAPE _{LR} | MAPE _{RT} | MAPE _{RF} | MAPE _{LGB} | MAPE _{Pred₁INDEC} |
|--------------------|--------------------|--------------------|--------------------|---------------------|---------------------------------------|
| Esteban Echeverría | 11.2 | 28.0 | 17.8 | 19.3 | 13.1 |
| Ezeiza | 10.0 | 41.6 | | | 12.8 |
| La Matanza | 34.4 | 3.4 | 21.7 | 24.6 | 29.2 |
| Moreno | 4.5 | 21.2 | 32.2 | 35.0 | 2.8 |

Cuadro 13: Outliers por modelo. MAPE .Proyecciones Población 2022

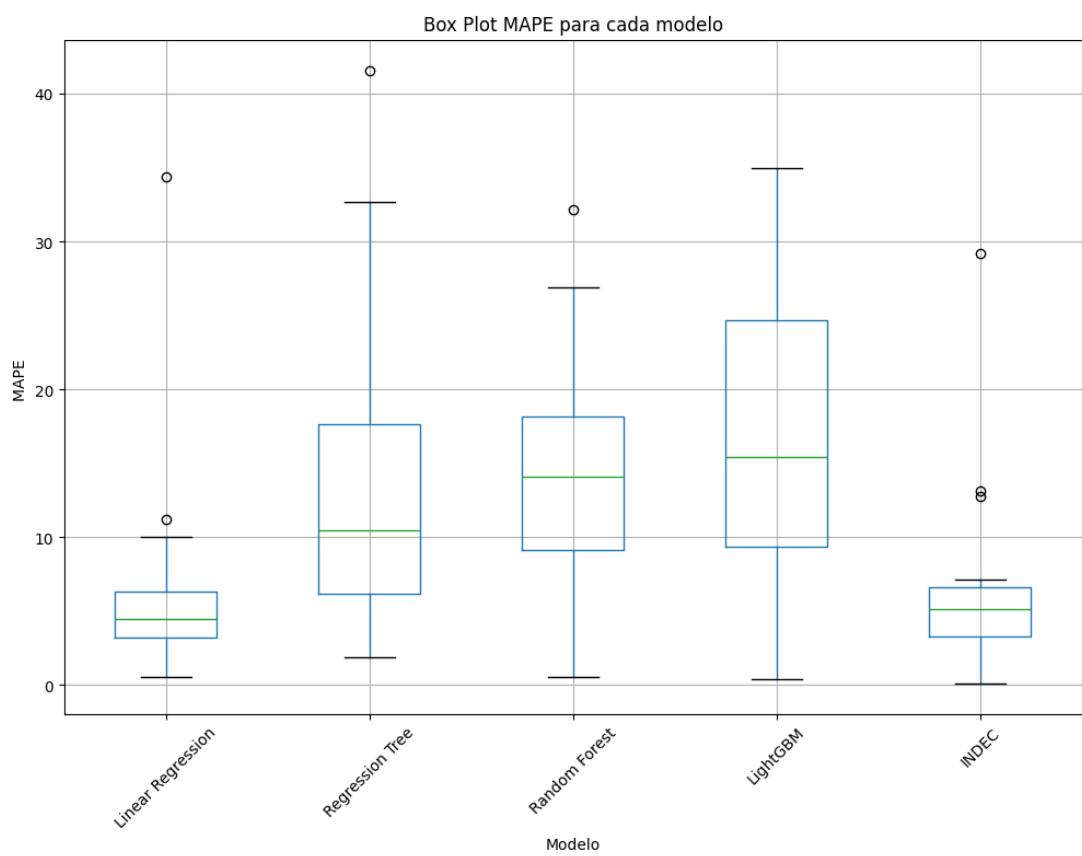


Figura 18: Box plot del Error Porcentual Absoluto Medio para cada modelo, sobre la predicción población año 2022

| Departamento | cod _{depto} | ano | pob | var | muj | vivpart | vivtotal | sup | ind _{masc} | dens _{pob} | TMI | TGF | TBN | TBM | TCV | Mat1ria |
|--------------------|----------------------|------|----------|----------|----------|----------|----------|--------|---------------------|---------------------|------|-----|------|-----|------|-----------|
| Almirante Brown | 6028 | 1991 | 450698.0 | 222042.0 | 228656.0 | nan | nan | 157.87 | 97.1 | 2854.87 | 24.2 | 2.6 | 18.4 | 7.9 | 10.5 | 1752994.0 |
| Almirante Brown | 6028 | 2001 | 515556.0 | 252454.0 | 263102.0 | 143543.0 | 88.0 | 157.87 | 96.0 | 3265.7 | 15.0 | 2.3 | 16.9 | 8.2 | 8.7 | 1658221.0 |
| Almirante Brown | 6028 | 2010 | 552902.0 | 270247.0 | 282655.0 | 156218.0 | 78.0 | 157.87 | 95.6 | 3502.26 | 12.0 | 2.5 | 18.9 | 8.4 | 10.5 | 1667278.0 |
| Avellaneda | 6035 | 1991 | 344991.0 | 164243.0 | 180748.0 | nan | nan | 68.54 | 90.9 | 5033.43 | 24.2 | 2.6 | 18.4 | 7.9 | 10.5 | 1752994.0 |
| Avellaneda | 6035 | 2001 | 328980.0 | 155450.0 | 173530.0 | 117200.0 | 59.0 | 68.54 | 89.6 | 4799.82 | 15.0 | 2.3 | 16.9 | 8.2 | 8.7 | 1658221.0 |
| Avellaneda | 6035 | 2010 | 342677.0 | 162264.0 | 180413.0 | 121307.0 | 68.0 | 68.54 | 89.9 | 4999.66 | 12.0 | 2.5 | 18.9 | 8.4 | 10.5 | 1667278.0 |
| Berazategui | 6091 | 1991 | 244929.0 | 120870.0 | 124059.0 | nan | nan | 268.91 | 97.4 | 910.82 | 24.2 | 2.6 | 18.4 | 7.9 | 10.5 | 1752994.0 |
| Berazategui | 6091 | 2001 | 287913.0 | 141163.0 | 146750.0 | 81511.0 | 38.0 | 268.91 | 96.2 | 1070.67 | 15.0 | 2.3 | 16.9 | 8.2 | 8.7 | 1658221.0 |
| Berazategui | 6091 | 2010 | 324244.0 | 158608.0 | 165636.0 | 96029.0 | 37.0 | 268.91 | 95.8 | 1205.77 | 12.0 | 2.5 | 18.9 | 8.4 | 10.5 | 1667278.0 |
| Esteban Echeverría | 6260 | 1991 | 275793.0 | 136784.0 | 139009.0 | nan | nan | 148.12 | 98.4 | 1861.96 | 24.2 | 2.6 | 18.4 | 7.9 | 10.5 | 1752994.0 |
| Esteban Echeverría | 6260 | 2001 | 243974.0 | 120110.0 | 123864.0 | 70535.0 | 26.0 | 148.12 | 97.0 | 1647.14 | 15.0 | 2.3 | 16.9 | 8.2 | 8.7 | 1658221.0 |
| Esteban Echeverría | 6260 | 2010 | 300959.0 | 147980.0 | 152979.0 | 88164.0 | 26.0 | 148.12 | 96.7 | 2031.86 | 12.0 | 2.5 | 18.9 | 8.4 | 10.5 | 1667278.0 |

Cuadro 14: Subset de datos input. Primeras 15 filas. Censos 1991, 2001, 2010 enriquecidos con las variables sintomáticas

11. Análisis de Singularidades. Curva Poblacional La Matanza

Al analizar la curva poblacional de La Matanza, figura 19, se observan cambios significativos en el ratio de crecimiento intercensal, a lo largo de los distintos censos. Si se observa el año 2001, el ratio intercensal es de 11.9% que se ubica un 75 % por encima del crecimiento(6.8%) promedio para los municipios del AMBA. En el año 2010 se observa un ratio de 41.5 % que representa un 250 % por encima del crecimiento promedio(11.8%). Esto representa un salto importante, siendo el municipio con mayor crecimiento en este periodo. Mientras que para el año 2022 se observa un ratio muy menor del orden de 3.5 %, un 70 % por debajo del crecimiento promedio (10.6 %).

Esto implica que La Matanza es uno de los departamentos con mayor desviación estandar y coeficiente de variación en ratios de crecimiento, tal como se describió en apartados anteriores. El crecimiento promedio de los departamentos del AMBA para el periodo 1991 a 2010 es de 9,9 % , mientras que La Matanza en el mismo periodo creció un 26,7 % siendo uno de los municipios con mayor crecimiento poblacional. Obviamente el promedio cae al incorporar el año 2022, dando lugar a que La Matanza en toda la serie se acerque al crecimiento promedio de otros municipios. Seguramente estas singularidades hayan dificultado la predicción del valor poblacional para el departamento con todas las metodologías aplicadas, incluyendo aquellas que pudieron resultar más efectivas en la mayoría de los casos.

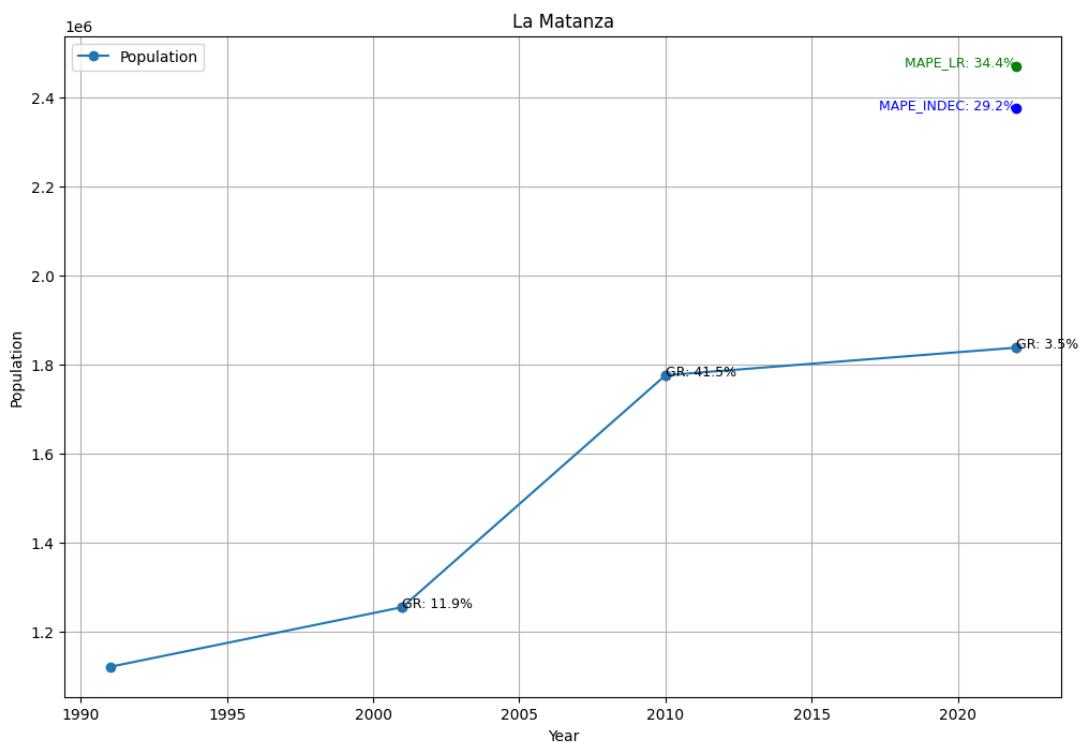


Figura 19: Curva Poblacional La Matanza.

12. Conclusión

En resumen, puede decirse que para la predicción de valores de población a nivel departamental se destacan las metodologías tradicionales, como regresión lineal o bien metodologías que se apoyan en el crecimiento intercensal que comprende la evolución en conjunto de las tres variables básicas del análisis demográfico: la fecundidad, la mortalidad y la migración.

Sin embargo debe considerarse que la elaboración de proyecciones de población de áreas menores resulta compleja debido a la imposibilidad de aplicar un método estrictamente demográfico, tal como el método de los componentes, que requiere la estimación y proyección independiente de cada una de las variables del crecimiento de la población. A este nivel de desagregación, los hechos vitales presentan fluctuaciones anuales más acentuadas cuanto menor es el número de población y consecuentemente de nacimientos y defunciones, que pueden afectar las estimaciones de la fecundidad y la mortalidad.

Por otra parte estos datos suelen difundirse al público en un nivel de Provincia y no existen datos por Departamento. Asimismo, se hace casi imposible la determinación de la migración interna, que suele ser un elemento muy importante del crecimiento de dichas áreas. Esto se debe a la dificultad de obtener estimaciones de saldos migratorios consistentes a nivel departamental y a la complejidad para su proyección futura, por tratarse de un factor estrechamente asociado a las condiciones económicas y sociales del momento.

Por otro lado puede decirse que las metodologías de data mining no han producido en este caso resultados precisos. Presentan dificultades debido a las características particulares de la información censal, el hecho de estar trabajando sólo con la información de 3(tres) Censos Nacionales. Esto sumado al hecho de la granularidad analizada en este caso, nivel departamental, lo que hace difícil enriquecer el dataset con variables sintomáticas a este nivel y se debe recurrir a información agregada a nivel Provincial. Estos modelos presentan valores de error notablemente mayores y una amplia dispersión de resultados.

Ciertamente si bien en la generalidad de los casos es posible estimar un valor población futuro con un grado de confianza aceptable, algunos departamentos presentan comportamiento singulares en sus curvas poblacionales que dificultan su predicción. Las razones de estas singularidades escapan al alcance del presente trabajo, bien pueden deberse a causales del complejo fenómeno demográfico así como también posibles errores en el relevamiento censal.

13. Bibliografía

Referencias

- [1] Gustavo. Álvarez et al. *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, 2001, pág. 36. ISBN: 9213217749.
- [2] UN-CEPAL. *CENSO 2001 Procesamiento del Censo Nacional de Población, Hogares y Viviendas 2001*. 2023. URL: <https://redatam.indec.gob.ar/argbin/RpWebEngine.exe/PortalAction?BASE=CPV2001ARG> (visitado 30-10-2023).
- [3] UN- CEPAL. *CENSO 2010 Procesamiento del Censo Nacional de Población, Hogares y Viviendas 2010 - Cuestionario Ampliado*. 2023. URL: <https://redatam.indec.gob.ar/binarg/RpWebEngine.exe/Portal?BASE=CPV2010A&lang=ESP> (visitado 30-10-2023).
- [4] Manan Chawda, Rutuja Rane y Srinanth Giri. «Demographic Progress Analysis of Census Data Using Data Mining». En: 2018, págs. 1894-1897. ISBN: 978-1-5386-1974-2.
- [5] Ministerio de Educación. <https://data.educacion.gob.ar/reporte-matricula.php>. 2022.
- [6] Arindam Gupta, Sabyasachi Bhattacharya y Asis Kumar Chattyopadhyay. «Exploring New Models for Population Prediction in Detecting Demographic Phase Change for Sparse Census Data». En: (2012).
- [7] Nazrul Hoque. «Evaluation of small area population estimates produced by Housing Unit, Ratio-correlation, and Component Method II compared to 2000 Census counts». En: *Canadian Studies in Population* 39, No. 1–2 (2012), págs. 91-108.
- [8] IGM. *Dataset GEO IGM–Insititu Geográfico Nacional, Argentina Capas SIG - Departamento*. 2024. URL: <https://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG> (visitado 30-03-2024).
- [9] INDEC. *CENSO 2001 Censo Nacional de Población, Hogares y Viviendas del año 2001*. 2023. URL: https://www.indec.gob.ar/micro_sitios/webcenso/index.asp (visitado 30-11-2023).
- [10] INDEC. *CENSO 2022 Resultados Previsionales*. 2022. URL: https://www.censo.gob.ar/index.php/datos_provisionales (visitado 30-10-2023).
- [11] INDEC. *DEMOGRAFICO Programa de Análisis Demográfico de la Dirección de Estadísticas Poblacionales*. 2022. URL: <https://www.indec.gob.ar/indec/web/Institucional - Indec - IndicadoresDemograficos> (visitado 10-02-2024).
- [12] INDEC. *Estimaciones de población por sexo, departamento y año calendario 2010-2025*. 1 ed. 2015.
- [13] INDEC. *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. 2013. ISBN: 978-950-896-433-5.
- [14] Ralph Kimball. «The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data Warehouses-Bom». En: (1996).

14. ANEXO

14.1. 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 o sin tilde, con espacios o abreviaturas en el campo.

De esta forma en la ingesta (ETL) se buscaba el código de departamento correspondiente, que funciona como clave única. El formato del diccionario puede verse en el siguiente cuadro15.

| CodigoDpto | Departamento |
|------------|------------------------|
| 6005 | General Sarmiento |
| 6005 | General Sarmiento (4) |
| 6260 | Esteban Echeverría (1) |
| 6260 | Esteban Echeverria |
| 6260 | Esteban Echeverría |
| 6270 | Ezeiza |
| 6270 | Ezeiza (2) |
| 6274 | Florencio Varela (3) |
| 6274 | Florencio Varela |
| 6274 | Florencio Varela (3) |

Cuadro 15: Diccionario. Primeras 10 filas

14.2. Diagrama de entidad relación

A modo descriptivo se indica el diagrama de entidad relación de la base de datos confeccionada para este trabajo. Figura 20.

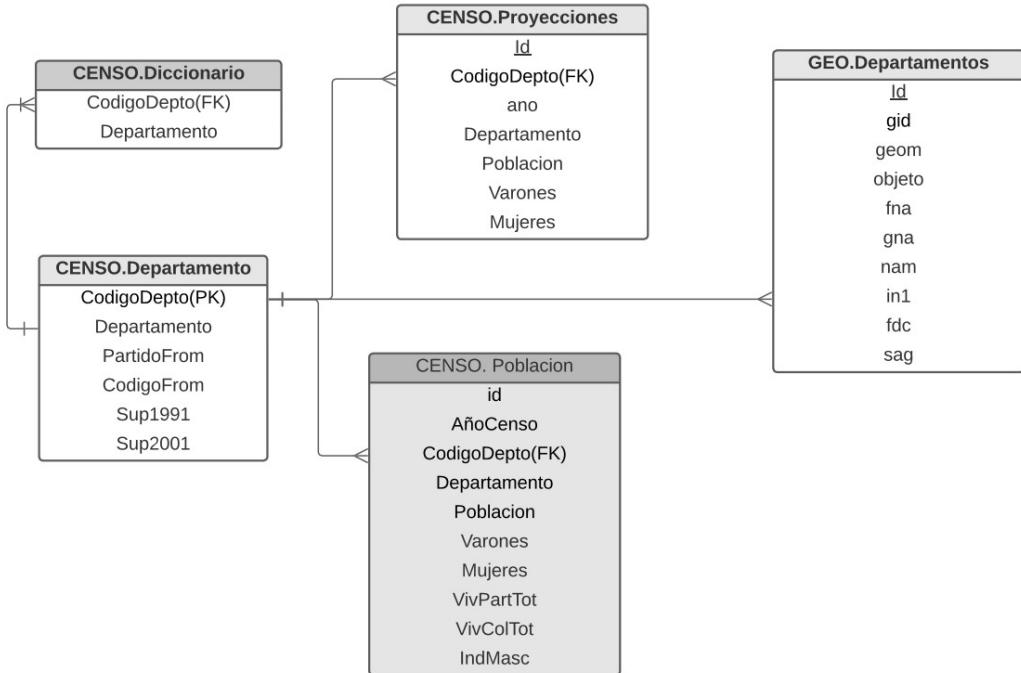


Figura 20: Diagrama de Entidad Relación

14.2.1. Script Creación Tablas.sql

```
1      -- ## POBLACION
2      CREATE TABLE IF NOT EXISTS public.poblacion (
3          Id SERIAL PRIMARY KEY,
4          "AnoCenso" VARCHAR(50),
5          "CodigoDpto" VARCHAR(50),
6          "Departamento" VARCHAR(150),
7          "Poblacion" INT,
8          "Varones" INT,
9          "Mujeres" INT,
10         "VivPartTot" INT,
11         "VivColectTot" INT,
12         "IndMasc" FLOAT,
13         "Superficie" INT,
14         "DensPob" FLOAT
15     );
16
17     TRUNCATE TABLE public.poblacion;
18
19     -- ## DIM Departamento
20     CREATE TABLE IF NOT EXISTS public.DimDepto (
21         Id SERIAL PRIMARY KEY,
22         "CodigoDpto" VARCHAR(50),
23         "Departamento" VARCHAR(150),
24         "PartidoFrom" VARCHAR(150),
25         "CodigoFrom" VARCHAR(50),
26         "Sup1991" INT,
27         "Sup2001" INT,
28         "IsAMBA" BOOLEAN,
29         "Comentarios" VARCHAR(550)
30     );
31
32     TRUNCATE TABLE public.DimDepto;
33
34     -- ## DICCIONARIO DATOS PARTIDO CODIGO
35     CREATE TABLE IF NOT EXISTS public.diccionario (
36         "CodigoDpto" VARCHAR(50),
37         "Departamento" VARCHAR(150)
38     );
39
40     TRUNCATE TABLE public.diccionario;
41
42     -- ## PROYECCIONES
43     CREATE TABLE IF NOT EXISTS public.proyecciones (
44         Id SERIAL PRIMARY KEY,
45         "CodigoDpto" VARCHAR(50),
46         "ano" INT,
47         "Departamento" VARCHAR(150),
48         "Poblacion" INT,
49         "Varones" INT,
50         "Mujeres" INT
51     );
52
53     TRUNCATE TABLE public.proyecciones;
```

```

55
56
57      ----- GEOMETRY TABLES ---
58      -- 1) departamento created from QGIS
59      -- 2) amba+censos
60
61
62      ---Now weed need to alter to add    COLUMNS para data de los censos
63
64      CREATE TABLE IF NOT EXISTS geo.amba_pob AS
65          SELECT *
66          FROM geo."vAMBOgeom"
67          ALTER TABLE geo.amba_pob
68              ADD COLUMN pob1991 INT,
69              ADD COLUMN pob2001 INT,
70              ADD COLUMN pob2010 INT,
71              ADD COLUMN pob2022 INT;
72      ######
73          ALTER TABLE geo.amba_pob
74              ADD COLUMN dens1991 INT,
75              ADD COLUMN dens2001 INT,
76              ADD COLUMN dens2010 INT,
77              ADD COLUMN dens2022 INT;
78
79          UPDATE geo.amba_pob AS am
80              SET pob1991 = c.pob
81              FROM public.v_censos_amba c
82              WHERE am.cod_depto = c.cod_depto AND anio='1991';
83      --- Update Poblacion de los censos
84          UPDATE geo.amba_pob AS am
85              SET pob2001 = c.pob
86              FROM public.v_censos_amba c
87              WHERE am.cod_depto = c.cod_depto AND anio='2001';
88
89          UPDATE geo.amba_pob AS am
90              SET pob2010 = c.pob
91              FROM public.v_censos_amba c
92              WHERE am.cod_depto = c.cod_depto AND anio='2010';
93          UPDATE geo.amba_pob AS am
94              SET pob2022 = c.pob
95              FROM public.v_censos_amba c
96              WHERE am.cod_depto = c.cod_depto AND anio='2022';
97
98      --- Update DENSIDAD de los censos
99          UPDATE geo.amba_pob AS am
100             SET dens1991 = c.dens_pob
101             FROM public.v_censos_amba c
102             WHERE am.cod_depto = c.cod_depto AND anio='1991';
103         UPDATE geo.amba_pob AS am
104             SET dens2001 = c.dens_pob
105             FROM public.v_censos_amba c
106             WHERE am.cod_depto = c.cod_depto AND anio='2001';
107
108         UPDATE geo.amba_pob AS am
109             SET dens2010 = c.dens_pob
110             FROM public.v_censos_amba c

```

```

111 WHERE am.cod_depto = c.cod_depto AND anio='2010';
112 UPDATE geo.amba_pob AS am
113 SET dens2022 = c.dens_pob
114 FROM public.v_censos_amba c
115 WHERE am.cod_depto = c.cod_depto AND anio='2022';

```

Listing 1: CreateTables.sql

14.2.2. Script PopulateTables.sql

```

1   -- Step 1: Insert data from CSV file without the primary key column
2   --- POBLACION START ---
3   COPY public.poblacion ("AnoCenso", "CodigoDpto", "Departamento", "Poblacion"
4   ", "Varones", "Mujeres", "VivPartTot", "VivColectTot", "IndMasc",
5   "Superficie","DensPob")
6   FROM 'C:/Temp/1991_A~1.CSV'
7   WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
8   ENCODING 'UTF8');

9
10
11
12
13
14   COPY public.poblacion ("AnoCenso", "CodigoDpto", "Departamento", "Poblacion"
15   ", "Varones", "Mujeres", "VivPartTot", "VivColectTot", "IndMasc",
16   "Superficie","DensPob")
17   FROM 'C:/Temp/1991_Resto.CSV'
18   WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
19   ENCODING 'UTF8');

20
21
22
23
24   COPY public.poblacion ("AnoCenso", "CodigoDpto", "Departamento", "Poblacion"
25   ", "Varones", "Mujeres", "VivPartTot", "VivColectTot", "IndMasc",
26   "Superficie","DensPob")
27   FROM 'C:/Temp/2001.CSV'
28   WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
29   ENCODING 'UTF8');

30
31
32
33
34   COPY public.poblacion ("AnoCenso", "CodigoDpto", "Departamento", "Poblacion"
35   ", "Varones", "Mujeres", "VivPartTot", "VivColectTot", "IndMasc",
36   "Superficie","DensPob")
37   FROM 'C:/Temp/2010.CSV'
38   WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
39   ENCODING 'UTF8');

40
41
42
43
44   COPY public.poblacion ("AnoCenso", "CodigoDpto", "Departamento", "Poblacion"
45   ", "Varones", "Mujeres", "VivPartTot", "VivColectTot", "IndMasc",
46   "Superficie","DensPob")
47   FROM 'C:/Temp/2022.CSV'
48   WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
49   ENCODING 'UTF8');

50
51
52
53
54   -- POBLACION END-----
55   ---- DICCIONARIO ----
56   COPY public.diccionario ("CodigoDpto", "Departamento")
57   FROM 'C:/Temp/DiccionarioPartidosCodigo.CSV'

```

```

32      WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
33          ENCODING 'UTF8');

34
35      ---      DIM Depto ---
36      COPY public.dimdepto (
37          "CodigoDpto",
38          "Departamento",
39          "PartidoFrom",
40          "CodigoFrom",
41          "Sup1991",
42          "Sup2001",
43          "IsAMBA",
44          "Comentarios")
45      FROM 'C:/Temp/DIM Departamento.CSV'
46      WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
47          ENCODING 'UTF8');

48
49      ---      Proyeccion 2025 ---
50      COPY public.proyecciones (
51          "CodigoDpto",
52          "ano",
53          "Departamento",
54          "Poblacion",
55          "Varones",
56          "Mujeres")
57      FROM 'C:/Temp/proy_1025.CSV'
58      WITH (FORMAT csv, HEADER true, DELIMITER ';;', QUOTE ''', ESCAPE '''',
          ENCODING 'UTF8');

```

Listing 2: PopulateTables.sql