Programa Anual de Investigación 2022

Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Integrantes:
Marentes Jimenez Priscila Arleen | priscila.marentes@inegi.org.mx
Cuellar Rio Manuel | manuel.cuellar@inegi.org.mx
Clemente Aréchiga Luis Martin | luis.clemente@inegi.org.mx
Ruiz Ortega Juan | juan.ruiz@inegi.org.mx
Villaseñor Garcia Elio Atenógenes | elio.villasenor@inegi.org.mx
Cabrera Zamora Irving Gibran | irving.cabrera@inegi.org.mx
Figueroa Martínez Alejandra | alejandra.figueroa@inegi.org.mx
Díaz Edgar Oswaldo | oswaldo.diaz@inegi.org.mx

Dirección General de Integración, Análisis e Investigación
Noviembre

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Contenido

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Resumen

Para el Instituto Nacional de Estadística y Geografía, INEGI, es de suma importancia preservar la privacidad del informante dado que es el que proporciona los datos para la producción de información estadística y geográfica. Para lograr este propósito se han realizado acciones para robustecer las medidas para reducir los riesgos en la identificación de los informantes, así como la identificación de mejores prácticas para fortalecer la confidencialidad estadística en la información que se genera en el instituto.

El presente reporte de investigación muestra los resultados del marco experimental para las pruebas de re-identificación en los microdatos generados por el INEGI, de los programas de información: Encuesta Nacional de Ocupación y Empleo (ENOE) y Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2020 nueva serie, además de utilizar el los datos abiertos del Censo de Población y Vivienda 2020, y el Marco Geoestadístico para referir geográficamente los resultados obtenidos, en el uso de mejores prácticas para evaluar métricas de riesgos en variables estratégicas, en un escenario controlado permitiendo aplicar métodos de re-identificación y generar datos sintéticos.

El objetivo es analizar el riesgo de una re-identificación de los informantes en los microdatos que genera el Instituto e integrar un prototipo de datos sintéticos que permita generar mecanismos para evitar la identificación de los Informantes en el sistema de la publicación de microdatos.

Contar con un mecanismo para la generación de datos sintéticos, le permitirá al Instituto fortalecer sus capacidades para proteger la confidencialidad de los datos que reporta, con un mínimo efecto restrictivo a su oferta de información.

Para el desarrollo del proyecto se hace uso de métricas para evaluar el riesgo de re-identificación, se realizan pruebas de re-identificación en los microdatos y se realizan algoritmos para la generación de datos sintéticos.

Palabras clave: re-identificación, datos sintéticos, confidencialidad y privacidad de datos

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Problemática

Actualmente el INEGI no cuenta con un mecanismo para la generación de datos sintéticos que le permita fortalecer sus capacidades para proteger la confidencialidad de los datos que reporta, con un mínimo efecto restrictivo a su oferta de información.

Además, es importante indagar si en los microdatos que publica el Instituto, existe alguna probabilidad de que se materialice el riesgo de re-identificación de los informantes, para tomar medidas preventivas y correctivas en su caso. De esta manera se contribuye a preservar la seguridad de la información y la confidencialidad estadística.

En el contexto de las Oficinas Nacionales de Estadística, se tiene acceso a grandes volúmenes de datos. Se debe encontrar el equilibrio entre el acceso a la información, la interoperabilidad de los datos y la preservación de la seguridad y confidencialidad estadística de la información sin afectar la calidad de ésta. La sociedad camina cada vez más hacia la transparencia, hacia la creación y el tratamiento de datos. Según Gil (2016) en este escenario, establecer un umbral de seguridad alto es una exigencia básica.

Como se establece en los principios sobre la confidencialidad estadística, los beneficios públicos de cualquier proyecto de integración de datos deberían ser suficientes para compensar cualquier preocupación por la privacidad o la confidencialidad estadística sobre el uso de los datos y los riesgos para la integridad del sistema. [1]

La integración de datos debe ocurrir en un entorno seguro y de una manera que no presente riesgos para la integridad del sistema estadístico oficial y cualquier identificador directo asociado con los datos que se integrarán debe eliminarse lo antes posible una vez finalizado el proceso de integración.

Las oficinas nacionales de estadística deben identificar y considerar adecuadamente todos los beneficios, las preocupaciones sobre la privacidad y los riesgos como parte de su proceso de producción.

El análisis de la privacidad y confidencialidad estadística requieren de una gestión cuidadosa de los riesgos de identificación indirecta, por ello es importante indagar si en los microdatos que publica el Instituto, existe alguna probabilidad de que se materialice el riesgo de re-identificación de los informantes, para tomar medidas preventivas y correctivas en su caso. De esta manera se contribuye a preservar la seguridad de la información y la confidencialidad estadística.

Desde la perspectiva de la comunidad investigadora, el apoyo a la investigación basada en microdatos debería ser un componente importante de cualquier sistema estadístico oficial. El acceso a los microdatos permite a los analistas calcular los efectos marginales en lugar de solo los promedios, permite la replicación de investigaciones importantes, facilita mejoras en la calidad de los datos y aumenta la gama de productos derivados de las recopilaciones estadísticas. [2]

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

La comunidad investigadora también ve la importancia de investigar métodos mejorados de protección de la confidencialidad que aumenten la utilidad de los datos subyacentes, por lo tanto, contar con un mecanismo para la generación de datos sintéticos, le permitirá al INEGI fortalecer sus capacidades para proteger la confidencialidad de los datos que reporta, con un mínimo efecto restrictivo a su oferta de información.

Importancia y utilidad del tema

El proyecto favorece el equilibrio entre privacidad, la calidad y la desagregación de la información en un contexto de ecosistema de datos sujetos a variantes dado el desarrollo tecnológico. Desarrollar el proyecto sobre privacidad de la información estadística y geográfica con datos sintéticos para microdatos permitirá al INEGI:

- Reducir el riesgo de identificación directa e indirecta de los informantes del sistema previo a la difusión y buscando afectar lo menos posible la utilidad de la Información para el conocimiento del tema.
- Disponer de una alternativa para compartir información en forma de microdatos manteniendo la privacidad de los informantes.
- Establecer estrategias para el aprovechamiento de información de terceros, dado que es una manera de compartir datos en forma de los registros que se utilizan para la recopilación de los datos y difundirlos sin que estos datos correspondan a levantamientos reales pero que si preservan propiedades estadísticas de los datos originales.
- Proponer nuevas estrategias para la publicación de los resultados de operativos que incorporen riesgos en materia de confidencialidad.

Cada vez son más los usuarios de la información con características acorde con el desarrollo de la tecnología. Disponen del conocimiento y las herramientas para explotar la masa de datos disponibles en la web, así como los datos estructurados en diferentes grados de desagregación. En este sentido la demanda de microdatos por parte de académicos e investigadores es más recurrente en donde está de por medio el dato confidencial. Lo anterior obliga que los métodos tradicionales de producción de información tiendan a renovarse.

La generación de datos sintéticos con base en información real representa una alternativa para cubrir la demanda de estos usuarios especializados sin poner en riesgo la privacidad del informante. Esto significa retos para proponer un cambio de paradigma en este sentido, el no considerarlo muestra un rezago en una de las funciones esenciales de la institución que hace referencia a proporcionar información a la sociedad a nivel de desagregación que cumpla con las necesidades de los usuarios.

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Hipótesis

Mediante la aplicación de técnicas de generación de datos sintéticos es posible compartir conjuntos de microdatos de manera segura sin comprometer la privacidad de los informantes.

Objetivos

El objetivo general es analizar el riesgo de que se materialice una re-identificación de los informantes en los microdatos que genera el Instituto e integrar un prototipo de datos sintéticos que permita generar mecanismos para evitar la identificación de los Informantes en el sistema de la publicación de microdatos. Los objetivos específicos son:

- Analizar la probabilidad que se presente el riesgo de re-identificación de los Informantes en los microdatos del INEGI.
- Reducir riesgos en la identificación de los informantes.
- Identificar mejores prácticas para fortalecer la confidencialidad estadística en los microdatos generados en el Instituto.
- Integrar un prototipo de datos sintéticos que permita generar mecanismos para evitar la identificación de los Informantes.

Entidades beneficiadas

El proyecto favorece el equilibrio entre el acceso a la información, la interoperabilidad de los datos y la preservación de la seguridad y confidencialidad estadística de la información sin afectar la calidad de ésta.

Desarrollar el proyecto sobre privacidad de la información estadística y geográfica con datos sintéticos para microdatos permitirá al INEGI reducir el riesgo de re-identificación directa e indirecta de los informantes previo a la difusión y buscando afectar lo menos posible la utilidad y calidad de la Información para el conocimiento del tema. De esta manera, se podrá disponer de una alternativa para compartir información en forma de microdatos manteniendo la privacidad de los informantes.

Además, se podrán establecer estrategias para el aprovechamiento de información de terceros, dado que es una manera de compartir datos en forma de los registros que se utilizan para la recopilación de los datos y difundirlos sin que estos datos correspondan a levantamientos reales pero que si preservan propiedades estadísticas de los datos originales. Por lo tanto, el instituto podrá proponer nuevas estrategias para la publicación de los resultados de operativos que incorporen riesgos en materia de confidencialidad.

Cada vez son más los usuarios de la información con características acorde con el desarrollo de la tecnología. Disponen del conocimiento y las herramientas para explotar la masa de datos disponibles en

la web, así como los datos estructurados en diferentes grados de desagregación. En este sentido la demanda de microdatos por parte de académicos e investigadores es más recurrente en donde está de por medio el dato confidencial. Lo anterior obliga que los métodos tradicionales de producción de información tiendan a renovarse.

La generación de datos sintéticos con base en información real representa una alternativa para cubrir la demanda de estos usuarios especializados sin poner en riesgo la privacidad del informante. Esto significa que existen retos significativos para proponer un cambio de paradigma, con la finalidad de reducir riesgos en las funciones esenciales de la institución sobre proporcionar información a la sociedad con un nivel de desagregación que cumpla con las necesidades de los usuarios.

Por lo tanto, los principales beneficiados son los informantes del sistema y los usuarios especializados que buscan información en el sitio Institucional por internet.

Dentro de INEGI, las direcciones de áreas involucradas y beneficiadas por este proyecto son:

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Estado del Arte

Los datos sintéticos han existido de una forma u otra durante décadas. Está en juegos de ordenador como simuladores de vuelo y simulaciones científicas de todo, desde átomos hasta galaxias. Donald B. Rubin, un profesor de estadística de Harvard estaba ayudando a las ramas del gobierno de Estados Unidos a resolver problemas como un recuento insuficiente, especialmente de personas pobres en un censo, cuando se le dio una idea. Lo describió en un artículo de 1993 a menudo citado como el nacimiento de datos sintéticos.

«*El término datos sintéticos en ese documento refiriéndose a múltiples conjuntos de datos simulados*», Donald B. Rubin. [Emeritus Professor of Statistics at Harvard University].

En el informe de 156 páginas de Sergey I. Nikolenko del Instituto Steklov de Matemáticas en San Petersburgo, Rusia, cita 719 artículos sobre datos sintéticos. Nikolenko concluye que «los datos sintéticos son esenciales para un mayor desarrollo del deep learning.

Dadas las preocupaciones y las políticas gubernamentales sobre la privacidad, eliminar información personal de un conjunto de datos es una práctica cada vez más común. Esto se llama anonimización de datos para el texto, un tipo de datos estructurados utilizados en industrias como las finanzas y la área de la salud pueden ser generados datos aumentados y/o anonimizados a los cuales no son considerados datos sintéticos.

Los métodos para datos sintéticos con un enfoque particular en la privacidad, permite implementar una herramienta para la generación con calidad, y un equilibrio en la similitud al dato real, considerando percepción e inferencia para reducir falsas expectativas en el uso en entornos de analítica y ciencia de datos contribuyendo a la democratización de los datos sensibles. Los datos sintéticos proporcionan herramientas prometedoras para mejorar la imparcialidad, el sesgo y la solidez de los sistemas de aprendizaje automático, a lo cual el uso estratégico depende del modelo de información con base a los objetivos para generar datos sintéticos, actualmente en el Grupo de Alto Nivel de la Comisión Económica de las Naciones Unidas para Europa, exponen casos de uso de oficinas de estadística internacionales para generar conocimiento colaborativo en estas técnica mencionada, a lo cual este reporte cita las mejores prácticas en el marco experimental con los programas de información estadística y geográfica generada por el Instituto.

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

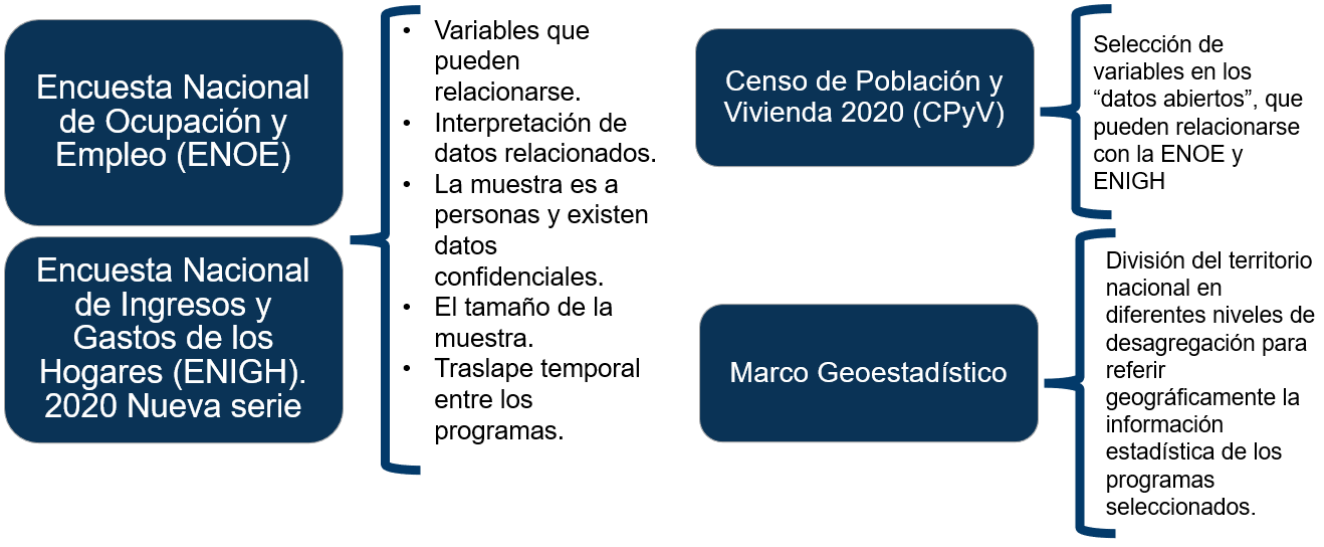
Proceso General de Solución

En contexto, el hackeo ético hace referencia al modo en que una persona, conocida como hacker, utiliza todos sus conocimientos e implementa herramientas sobre informática y ciberseguridad, para detectar vulnerabilidades en el entorno de tecnologías de información, a lo cual el propósito del proyecto referido en este documento permite implementar ciencia de datos en un marco experimental en generar pruebas para lograr identificar de manera directa o indirecta datos confidenciales de la información que genera el Instituto.

Un dato confidencial es un dato que permite identificar, directa o indirectamente, a los informantes que proporciona información individual. Incluye los datos clasificados con ese carácter en la legislación, así como los secretos de carácter bancario, fiduciario, industrial, comercial, fiscal, bursátil, postal o de cualquier otro tipo cuya titularidad corresponda a las personas. [3]

Los microdatos son valores de las variables asociadas a cada una de las unidades de observación, para un programa de Información es un conjunto de actividades, que se pueden repetir, que describen el propósito y contexto de un conjunto de Procesos que se llevarán a cabo cada periodo de tiempo para producir Información.

Los siguientes programas de información fueron seleccionados considerando lo siguiente:



Los datos utilizados de la Encuesta Nacional de Ocupación y Empleo (ENOE), son:

- Población objetivo Población de 12 y más años.
- Método de recolección Entrevista cara a cara y telefónica.
- Unidades de observación los hogares y residentes habituales de las viviendas particulares.

- Tamaño de la muestra 23 555 viviendas efectivas (16 948 visitadas cara a cara; 6 607 por teléfono).
- Alcance de resultados Nacional.
- Periodo de levantamiento 6 de julio al 2 de agosto de 2020

Los datos utilizados de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2020 nueva serie, son:

- Características de la vivienda
- Residentes e identificación de hogares en la vivienda
- Características sociodemográficas de los residentes de la vivienda
- Equipamiento y servicios del hogar
- Condición de actividad y características ocupacionales de los integrantes del hogar de 12 y más años.
- Ingreso corriente total (monetario y no monetario) de los hogares.
- Percepciones financieras y de capital de los hogares y sus integrantes.
- Gasto corriente monetario de los hogares.
- Erogaciones financieras y de capital de los hogares.
- Dimensiones de las carencias

Riesgos para considerar en el contexto

La anonimización es una técnica que se aplica a los datos personales para obtener una desidentificación irreversible (proceso que se utiliza para evitar que se revele la identidad de una persona).

Algunos tipos de anonimización son:

a) Supresión de atributo.

Descripción: se refiere a la eliminación de una parte completa de los datos (también referido como "columna" en bases de datos y hojas de cálculo) en un conjunto de datos.

Uso recomendado: cuando no se requiere un atributo en el conjunto de datos anonimizado, o cuando el atributo no se puede anonimizar adecuadamente con otra técnica. Esta técnica debe aplicarse al comienzo del proceso de anonimización, ya que es una manera fácil de disminuir la identificabilidad en este punto.

Implementación: eliminar los atributos o columnas que contienen los datos; es importante que la eliminación sea real y no solo se oculten las columnas.

b) Enmascaramiento de caracteres.

Descripción: se refiere al cambio de los caracteres de un valor de datos, por ejemplo, mediante el uso de un símbolo constante ("*" o "x"). El enmascaramiento suele ser parcial, es decir, se aplica solo en algunos caracteres del atributo.

Uso recomendado: cuando el valor de los datos es una cadena de caracteres y ocultar una parte es suficiente para proporcionar el grado de anonimato requerido.

Implementación: dependiendo de la naturaleza del atributo, reemplazar los caracteres apropiados con un símbolo elegido. Dependiendo del tipo de atributo, puede decidir reemplazar un número fijo de caracteres (por ejemplo, para números de tarjetas de crédito) o un número variable de caracteres (por ejemplo, para la dirección de correo electrónico).

c) Seudoanonimización.

Descripción: se refiere a la sustitución de datos de identificación con valores inventados. La seudoanonimización también se conoce como codificación.

Uso recomendado: cuando los valores de datos deben distinguirse de forma exclusiva y donde no se debe mantener ningún carácter ni ninguna otra información implícita del atributo original.
Implementación: reemplazar los valores de atributo respectivos con valores inventados. Una forma de hacer esto es generar previamente una lista de valores inventados y seleccionar al azar de esta lista para reemplazar cada uno de los valores originales. Los valores inventados deben ser únicos y no deben tener relación con los valores originales (de modo que uno no pueda derivar los valores originales de los seudónimos).

d) Generalización.

Descripción: se refiere a la reducción deliberada en la precisión de los datos. Por ejemplo, convertir la edad de una persona en un rango de edad, o una ubicación precisa en una ubicación menos precisa. Esta técnica también se conoce como recodificación.

Uso recomendado: para valores que pueden generalizarse y seguir siendo útiles para el propósito previsto.

Implementación: diseñar categorías y reglas para traducir datos. Suprimir cualquier registro que aún se destaque después de la traducción (es decir, la generalización).

e) Intercambio.

Descripción: se refiere a reorganizar los datos de modo que los valores de los atributos individuales todavía estén representados en el conjunto de datos, pero en general, no se corresponden con los registros originales. Esta técnica también se conoce como barajar y permutar.

Uso recomendado: cuando el análisis posterior solo necesita mirar datos agregados, o el análisis está en el nivel intra-atributo; en otras palabras, no es necesario analizar las relaciones entre los atributos a nivel de registro.

Implementación: identificar qué atributos intercambiar. Luego, para cada uno, intercambiar o reasignar los valores de los atributos a cualquier registro en el conjunto de datos.

f) Perturbación de datos.

Descripción: se refiere a modificar los valores del conjunto de datos original para ser ligeramente diferentes.

Uso recomendado: para cuasi identificadores (generalmente números y fechas) que pueden identificarse potencialmente cuando se combinan con otras fuentes de datos, y son aceptables ligeros cambios en el valor.

Implementación: depende de la técnica exacta de perturbación de datos utilizada. Estos incluyen redondear y agregar ruido aleatorio.

g) Agregación de datos.

Descripción: se refiere a convertir un conjunto de datos de una lista de registros a valores resumidos.

Uso recomendado: cuando no se requieren registros individuales y los datos agregados son suficientes para el propósito. Esta técnica es adecuada sí y sólo sí se agregan datos del mismo tipo.

Implementación: las formas típicas incluyen el uso de totales o promedios. También podría ser útil discutir con el receptor de datos acerca de la utilidad esperada y encontrar un compromiso adecuado.

h) Supresión de valores individuales o supresión de casillas.

Descripción: consiste en eliminar el valor de una celda dentro de un tabulado cuando a partir de dicho valor es posible realizar la identificación directa o indirecta.

Uso recomendado: en los tabulados donde se presenta información que incluye variables que pueden conducir a una identificación.

Implementación: identificar las celdas a partir de cuyos valores es posible realizar la identificación directa o indirecta y sustituir el contenido por un carácter especial [4]

Además, los riesgos de la anonimización son los siguientes:

- Singularización: La posibilidad de extraer de un conjunto de datos algunos registros (o todos los registros) que identifican a una persona.
- Vinculabilidad: La capacidad de vincular como mínimo dos registros de un único interesado o de un grupo de interesados, ya sea en un modelo de datos o mezcla con otros modelos de datos distintos. Al identificar o determinar que dos registros están asignados al mismo grupo de personas, pero no puede singularizar a las personas en este grupo, no hay singularización, pero si vinculabilidad.
- Inferencia: La posibilidad de deducir con una probabilidad significativa el valor de un atributo a partir de los valores de un conjunto de otros atributos.

Es importante no cometer los siguientes errores en el proceso de anonimización:

- Pensar que los datos seudonimizados son datos anonimizados. No constituyen información anonimizada, ya que permiten singularizar a los interesados y vincularlos entre conjuntos de datos diferentes. La probabilidad de que el seudoanonimato admita la identificabilidad es muy alta.
- Los datos correctamente anonimizados no los exime del ámbito de aplicación sobre protección de datos.
- No actuar con especial precaución al manejar información anonimizada, especialmente cuando esta se utiliza (con frecuencia en combinación con otros datos) para tomar decisiones que causan efectos (aunque sea indirectamente) en las personas.

En la producción de Información Estadística y Geográfica, las Unidades del Estado deberán implementar medidas para asegurar que la difusión de la Información se realiza de manera que los Informantes del Sistema y, en general, las personas físicas o morales objeto de la Información no sean identificados de manera directa o indirecta, por lo que previo a la entrega de resultados deben evaluar el riesgo de Identificación de acuerdo con lo siguiente:

a) Se considera que el nivel de riesgo es alto cuando la Identificación es inmediata. Con sólo acceder a la Información es posible reconocer a la persona física o moral a la que corresponden los datos; o cuando la Identificación se deduce de la combinación de diferentes variables contenidas en el mismo producto de difusión de la Información;
b) Se considera que el nivel de riesgo es medio cuando la Identificación se logra realizando la suma o resta de algunas clases o agrupamientos de este tabulado y combinando el resultado con otros productos de Información Estadística o Geográfica;
c) Se considera que el nivel de riesgo es bajo cuando la Identificación se logra combinando la Información con distintos repositorios públicos y privados de datos utilizando técnicas de análisis, software y equipo de cómputo, y
d) Se considera que el nivel de riesgo es nulo cuando no es posible realizar la Identificación por cualquier medio o por la combinación de cualquier variable de Información.[5]

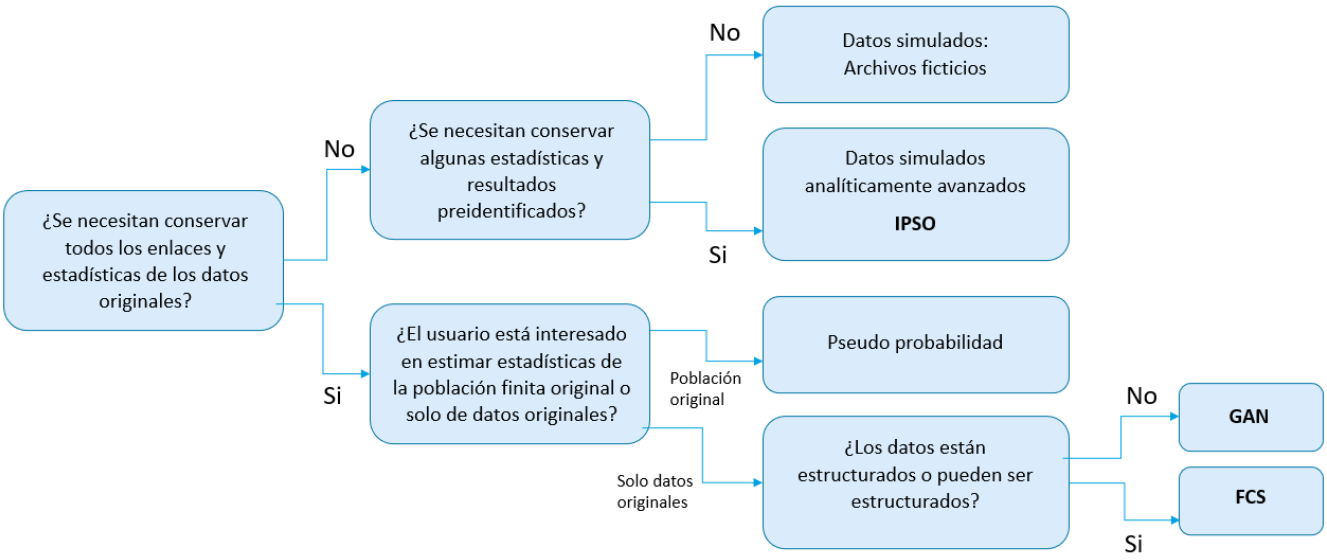
Estrategia implementada para pruebas de re-identificación

Con base en las mejores prácticas internacionales los métodos utilizados para generar métricas en la evaluación del riesgo de re-identificación, son las siguientes:

- k-anonimato (k-anonymity): Un conjunto de datos publicados tiene la propiedad de k-anonimato (o es k-anónimo) si la información de todas y cada una de las personas contenidas en ese conjunto es idéntica al menos con otras k-1 personas que también aparecen en dicho conjunto.
- l-diversidad (l-Diversity): L-diversidad es una extensión del k-anonimato. Requiere que cada bloque de k-anonimato contenga al menos l valores "bien representados" para el atributo sensible t-cercanía.
- t-cercanía (t-Closeness): refina aún más el concepto de l-diversidad. Una clase tiene t-cercanía si la distancia entre la distribución de un atributo sensible en la clase y la distribución en toda la tabla no es mayor a un valor t. Una tabla tiene t-cercanía si todas sus clases tienen t-cercanía.

Estrategia implementada para generar datos sintéticos

Con base en mejores prácticas internacionales Grupo de Alto Nivel de la Comisión Económica de las Naciones Unidas para Europa, la estrategia esta descrita en el siguiente diagrama de flujo.



Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Método para generar datos sintéticos Fully Conditional Specification (FCS)

Conceptualmente las características de los datos originales se obtienen de la distribución conjunta de todas las variables de las tablas originales, estas distribuciones no son conocidas a priori, por lo que se estiman usando modelación. El modelaje de distribuciones conjuntas en un solo paso es complejo, por lo que FCS descompone la distribución conjunta en una serie de distribuciones condicionales univariadas.

$$\boldsymbol{f}\boldsymbol{X1}, \boldsymbol{X2},\ldots ,\boldsymbol{Xp} = \boldsymbol{f}\boldsymbol{X1}\times\boldsymbol{f}\boldsymbol{X2}|\boldsymbol{X1}\times\cdots\times \boldsymbol{f}\boldsymbol{Xp}|\boldsymbol{X1},\boldsymbol{X2},\ldots, \boldsymbol{Xp}$$

- Modelar la distribución univariada $\boldsymbol{f}\boldsymbol{X1}$ con base a los datos originales
- Generar valores a partir del modelo no condicional para obtener valores sintéticos de $\boldsymbol{X1}$
- Modelar la distribución condicional $\boldsymbol{f}\boldsymbol{X2}|\boldsymbol{X1}$ basada en los datos originales
- Generate valores a partir del modelo $\boldsymbol{X1}$, $\boldsymbol{syn}$ como entrada para obtener valores sintéticos $\boldsymbol{X2}$
- Repetir el punto 3 y 4 hasta la última variable $\boldsymbol{Xp}$

Consideraciones

Este método es fácil de entender y explicar. Debido a que el objetivo es la distribución del conjunto de datos, este método pretende preservar (en teoría) todas las relaciones entre todas las variables. No es necesario que las relaciones de interés se conozcan antes del proceso de creación. Además, debido a que se deriva de la imputación, este enfoque naturalmente tiene un gran parecido operacionalmente hablando con la edición de datos.

Para datos sesgados (como datos comerciales o económicos), la presencia de valores atípicos sigue siendo un desafío en términos de divulgación o control de divulgación percibido. Con muchas variables, el proceso puede llevar mucho tiempo.

Publicación de microdatos sintéticos al público y análisis de prueba	Educación	Pruebas de Tecnología
Recomendado	Puede ser usado. Si los análisis realizados y las conclusiones estadísticas están predeterminadas, podría consumir demasiado tiempo en comparación con otros métodos.	Se puede usar, pero podría ser demasiado avanzado en comparación con la necesidad analítica real

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Método para generar datos sintéticos Information Preserving Statistical Obfuscation (IPSO)

En este método los datos originales se consideran estar conformados por dos subconjuntos de variables: la matriz X es información no confidencial y la matriz Y es información confidencial. Este método asume la normalidad multivariada en la distribución de un modelo de regresión lineal don la matriz X es la componente independiente y **Y** es la dependiente. Se ajusta el modelo $\boldsymbol{Y} = \beta \boldsymbol{X} + \varepsilon$ en los datos originales para obtener y se calcula. Se agrega ruido normalmente distribuido a para crear los valores sintéticos Y′.

$Y = X \ \beta + \Sigma$

IPSO añade un pasos extra para forzar la igualdad o.

$\boldsymbol{\beta original} = \boldsymbol{\beta\ synthetic}$ and $\boldsymbol{\Sigma\ original} = \boldsymbol{\Sigma\ synthetic}$

Consideraciones

Al igual que el Modelo de Especificación condicional completo es fácil de entender y explicar. En este método es posible conservar exactamente algunos parámetros preidentificados y estadísticos. Por lo tanto, cualquier análisis que se base en la normalidad (multivariante) producirá exactamente los mismos resultados en los datos originales y sintéticos. IPSO se puede implementar como parte de otro proceso para generar conjuntos de datos sintéticos.

Estos métodos híbridos pueden usarse para aliviar los supuestos de la distribución normal.
La distribución normal para todas las variables es un supuesto que rara vez es cierto.

Publicación de microdatos sintéticos al público y análisis de prueba	Educación	Pruebas de Tecnología
Recomendado si todos los análisis están relacionados con regresiones lineales; de lo contrario, no es recomendado.	Recomendado si todos los análisis están relacionados con regresiones lineales; de lo contrario, no es recomendado.	Se puede usar, pero podría ser demasiado avanzado en comparación con la necesidad analítica real.

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Método para generar datos sintéticos Generative Adversarial Networks (GAN)

- Con mejoras en la tecnología y la capacidad computacional, la implementación de Machine learning se ha vuelto más fácil y accesible.
- Los enfoques de Machine learning se han empleado cada vez más para generar conjuntos de datos sintéticos.
- Más específicamente, el uso de modelos de Deep learning se ha vuelto atractivo debido a su capacidad para extraer de grandes conjuntos de datos un modelo predictivo muy poderoso.
- La red adversarial generativa (GAN) (Goodfellow, et al., 2014) es un modelo generativo destacado utilizado para la generación de datos sintéticos.

La idea detrás de las GAN es crear dos modelos de redes neuronales. Uno es llamado el generador el cual recibe valores aleatorios y los transforma en un registro. El otro modelo es llamado el discriminador, recibe un registro y trata de discernir si es real o sintético. Estos dos modelos son entrenados de manera iterativa con el objetivo de vencer a su contraparte hasta que se logra un equilibrio entre ambos. El generador es entrenado para poder crear registros que parezcan reales, mientras el discriminador es entrenado para encontrar las sutiles diferencias entre los datos reales y generados.

Consideraciones

GAN se puede utilizar para generar conjuntos de datos continuos, discretos y también de texto, al tiempo que garantiza que se conservan la distribución y los patrones subyacentes de los datos originales. Puede generar conjuntos de datos totalmente sintéticos. Apunta a preservar todas las relaciones entre variables. Puede manejar datos no estructurados.

GAN puede verse tan complejo de entender, explicar o implementar cuando hay un conocimiento mínimo de las redes neuronales. A menudo hay una crítica asociada a las redes neuronales por falta de transparencia o por ser una caja negra. El método consume mucho tiempo y requiere una gran demanda de recursos computacionales.

Publicación de microdatos sintéticos al público y análisis de prueba	Educación	Pruebas de Tecnología
Recomendado especialmente en presencia de texto o datos no estructurados.	Puede ser usado. Si los análisis realizados y las conclusiones estadísticas están predeterminadas, podría consumir demasiado tiempo en comparación con otros métodos.	Se puede usar, pero podría ser demasiado avanzado en comparación con la necesidad analítica real.

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Datos simulados: desde archivos ficticios hasta archivos sintéticos más avanzados analíticamente

Consideraciones

Los procesos de simulación son fáciles de entender y pueden crear datos completamente seguros cuando no se utiliza información relacionada con los datos originales. Puede generar archivos totalmente sintéticos. Para tipos de simulaciones más avanzados, se puede conservar algún valor analítico.

Por lo general, no permite satisfacer necesidades analíticas complejas.

Publicación de microdatos sintéticos al público y análisis de prueba	Educación	Pruebas de Tecnología
No recomendado.	Se puede usar si el entrenamiento no requiere valor analítico en los datos.	Recomendado

Datos simulados analíticamente avanzados

Publicación de microdatos sintéticos al público y análisis de prueba	Educación	Pruebas de Tecnología
Recomendado si los análisis realizados están relacionados con los resultados preidentificados que debían conservarse en el proceso de síntesis. De lo contrario, no es recomendado.	Recomendado si los análisis realizados están relacionados con los resultados preidentificados que debían conservarse en el proceso de síntesis. De lo contrario, no recomendado.	Se puede usar, pero podría ser demasiado avanzado en comparación con la necesidad analítica real.

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Marco experimental

Con base al Modelo de Gestión de Proyectos MGP de la Dirección del Laboratorio de Ciencia de Datos y Métodos Modernos de Producción de Información, son descritas las siguientes fases.

Fase de fuentes, extracción y carga de datos

La selección de las fuentes de información en consenso con las áreas generadoras de las Unidades Administrativas involucradas fue que en el portal del sitio del INEGI por internet https://www.inegi.org.mx/microdatos/, de los programas de información referidos seleccionar los microdatos para generar una colección de datos requerida la cual por medio de técnicas de "scraping" fueron ingestados los microdatos en el lago de datos, que sirven de insumo para las siguientes fases.

Obtención de la tabla de datos

Obtenemos la tabla sociodemográfica de la ENOE I trimestre 2020. Para este ejercicio la obtenemos del lago de datos que tiene la versión correspondiente a la versión publica en el sitio del INEGI en la sección de microdatos.

Para acceder al lago de datos es necesario autenticarnos con usuario y contraseña.

In [3]:
```
opciones_almacenamiento={
        "key": input("Usuario: "),
        "secret": getpass(prompt = 'Contraseña: '),
        "client_kwargs": {"endpoint_url": "http://lcidn4.inegi.gob.mx:9100"}
    }
```

Out [3]:
```
Usuario:  irving.cabrera
Contraseña:  ·····
```

In [4]:
```
ruta_archivo = "s3://infoenoe-public/datosabiertos/2020/conjunto_de_datos_enoe_2020_1t_csv/conjunto_de_datos_sdem_enoe_2020_1t/conjunto_de_
```

In [5]:
```
tabla_sdem = pd.read_csv(
    ruta_archivo,
    storage_options = opciones_almacenamiento,
    encoding = "latin-1",
    low_memory=False,
    dtype= "str",
    na_values = [" ",""]
)
```

In [6]:
```
tabla_sdem
```

Out [6]:

	r_def	loc	mun	est	est_d	ageb	t_loc	cd_a	ent	con	...	ma48me1sm	p14apoyos	scian	t_tra	emp_ppal	tue_ppal	trans_ppal	mh_fil2	mh_col	sec_in
0	0	NaN	2	10	122	0	1	1	9	40001	...	0	2	5	1	2	2	0	3	6	4
1	0	NaN	2	10	122	0	1	1	9	40001	...	0	2	6	1	2	2	0	3	2	2
2	0	NaN	2	10	122	0	1	1	9	40001	...	1	2	5	1	1	2	0	3	1	4
3	0	NaN	2	10	122	0	1	1	9	40001	...	0	2	0	1	0	0	0	0	0	0
4	0	NaN	2	10	122	0	1	1	9	40001	...	0	2	0	1	0	0	0	0	0	0

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Fase recuperación de Información

Implementación de una estrategia para seleccionar el conjunto de microdatos, de la colección obtenida de la anterior fase descrita, observando las estructuras, variables, tipos de datos no relacionales, con base al diccionario proporcionado por las Unidades Administrativas involucradas en el proyecto, teniendo como resultado un “conjunto de datos”.

Filtrado de datos

Reducimos la tabla a solo las columnas uy filas de interés. Las cuáles serán

- columnas
  - columnas indicadas como de identificación
  - columnas indicadas como sensibles
- filas
  - filas que correspondan a residentes que no sean ausentes definitivos

In [7]:

```
variables_interes = [
    "ent", #Entidad
    "cd_a", # Ciudad autorrepresentada
    "mun", # Municipio
    "sex", # Sexo
    "eda", # Edad
    "cs_p13_1", # nivel escolar
    "pos_ocu", # Clasificación por posición en la ocupación
    "ingocup" # Ingreso mensual
]
```

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Fase procesamiento de Datos

Implementación y aplicación de técnicas estadísticas e informáticas para analizar los datos recolectados en las fases anteriores, permitiendo la generación de tablas virtuales relacionando las variables de los programas de información seleccionados en conjunto con los involucrados en el proyecto.

In [8]:

```
sdem_interes = tabla_sdem.loc[tabla_sdem["c_res"] != "2" ,variables_interes].copy()
sdem_interes.loc[sdem_interes["cd_a"].astype(int) > 46,"cd_a"] = np.NaN
sdem_interes.loc[sdem_interes["eda"] == "99","eda"] = np.NaN
sdem_interes.loc[sdem_interes["cs_p13_1"] == "99","cs_p13_1"] = np.NaN
sdem_interes.loc[sdem_interes["pos_ocu"] == "0","pos_ocu"] = np.NaN
sdem_interes.loc[sdem_interes["ingocup"] == "0","ingocup"] = np.NaN

sdem_interes.info(show_counts=True)
```

Out [8]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 409283 entries, 0 to 417481
Data columns (total 8 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   ent       409283 non-null  object
 1   cd_a      260801 non-null  object
 2   mun       405870 non-null  object
 3   sex       409283 non-null  object
 4   eda       409268 non-null  object
 5   cs_p13_1  381063 non-null  object
 6   pos_ocu   184064 non-null  object
 7   ingocup   134386 non-null  object
dtypes: object(8)
memory usage: 28.1+ MB
```

Fase modelo de aprendizaje

La estrategia para esta fase es con base al marco experimental realizar pruebas con el “conjunto de datos” más los métodos seleccionados considerando la siguiente tabla:

Pruebas de re-identificación

Método \ Programa	Programa de información seleccionado
k-anonimato (k-anonymity) = M1	Iteración de pruebas con M1
l-diversidad (l-diversity) = M2	Iteración de pruebas con M2
t-cercanía (t-closeness) = M3	Iteración de pruebas con M3
Resultado	Conjunto de datos resultados para Data Set

K-anonimity

Un conjunto de datos publicados tiene la propiedad de K-anonimity (k-anonimato) si la información de todas y cada una de las personas contenidas en ese conjunto es idéntica al menos con otras k-1 personas que también aparecen en dicho conjunto.

Por lo que agruparemos estos datos respecto a ciertos identificadores, en este caso utilizaremos los datos de

- Entidad
- Ciudad autorrepresentada
- Municipio
- Sexo
- Edad

In [9]:
```
variables_identificacion =[
    "ent", #Entidad
    "cd_a", # Ciudad autorrepresentada
    "mun", # Municipio
    "sex", # Sexo
    "eda", # Edad
]
```

In [10]:
```
# Esta función genera los conteos respecto a los valores únicos de las variables de identificación
# genera las agrupaciones primero respecto al a primera variable de identificación, después la primera y segunda
# hasta que finalmente genera el conteo de todas las variables de identificación

def obtener_k_anonimity(tabla,variables_identificacion):
    k_anonimity = (tabla[variables_identificacion]
        .value_counts(ascending=True)
        .rename("k_anonimity")
        .to_frame()
        .dropna()
    )
    return k_anonimity
```

Obtenemos los conteos y buscamos el valor mínimo ya que este es valor de la K-anonimity

In [11]:
```
for i in range(0,len(variables_identificacion)):
    v_i = variables_identificacion[:(i+1)]
    k_a = obtener_k_anonimity(sdem_interes,v_i)
    nivel = k_a["k_anonimity"].min()
    print(f"k-anonimity repecto a las variables ({', '.join(v_i)}): {nivel}")
    print(k_a.head())
```

Out [11]:
```
k-anonimity repecto a las variables (ent): 8052
     k_anonimity
ent
3           8052
9           9310
23          9474
6           9842
13         10157
k-anonimity repecto a las variables (ent, cd_a): 314
          k_anonimity
ent cd_a
30  10            314
10  6            3096
3   40           4969
30  12           5167
28  10           5305
k-anonimity repecto a las variables (ent, cd_a, mun): 6
```

l-diversity

Es una extensión del K-anonimity. Requiere que cada bloque de K-anonimity contenga al menos *l* valores "bien representados" para el atributo sensible.

Es decir, generamos el conteo no solo respecto a variables de identificación sino también respecto a variables sensibles.

Se recomienda usar variables sensibles del tipo categóricas. Ya que datos continuos pueden producir demasiados grupos

In [12]:
```
# tomamos un nivel de  K-anonimity mayor a cero para observar el efecto de agregar variables sensibles
variables_identificacion =[
    "ent", #Entidad
    "cd_a", # Ciudad autorrepresentada
]
variables_sensibles = [
    "cs_p13_1", # nivel escolar
    "pos_ocu" # Clasificación por posición en la ocupación
]
```

In [13]:
```
# Función para calcular los conteos te variables de identificación y variables sensibles
# Para cada variable sensible calcula los conteos de las variables de identificación más las variables sensibles
def obtener_l_diversity(tabla,variables_identificacion,variable_sensible):
    l_diversity = (tabla[variables_identificacion + [variable_sensible]]
        .value_counts(ascending=True)
        .rename("l_diversity")
        .to_frame()
    )
    return l_diversity
```

Obtenemos los conteos y buscamos el valor mínimo ya que este es valor de la *l*-diversity

In [14]:
```
for v_s in variables_sensibles:
    l_d = obtener_l_diversity(sdem_interes,variables_identificacion,v_s)
    nivel = l_d["l_diversity"].min()
    print(f"l-diversity repecto a las variables {', '.join(variables_identificacion + [v_s])}: {nivel}")
    print(l_d.head())
```

Out [14]:
```
l-diversity repecto a las variables ent, cd_a, cs_p13_1: 1
                  l_diversity
ent cd_a cs_p13_1
10  6    9                  1
23  41   9                  1
30  30   9                  2
12  13   9                  3
4   42   9                  4
l-diversity repecto a las variables ent, cd_a, pos_ocu: 5
                 l_diversity
ent cd_a pos_ocu
30  10   2                 5
         4                 6
8   9    4                13
2   44   4                16
8   20   4                24
```

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

t-closness

t-closness mide que tanto se acercan la distribución de valores sensibles en cada clase y la correspondiente distribución en la tabla original.

Cuando la distribución de valores sensibles es muy cercana a la distribución original, los elementos individuales son más difíciles de identificar y el valor de *t*-closness será cercano a 0.

In [15]:
```
# tomamos un nivel bajo de variables de identificación como base
variables_identificacion =[
    "ent", #Entidad
    "cd_a", # Ciudad autorrepresentada
]
variables_sensibles = [
    "cs_p13_1", # nivel escolar
    "pos_ocu" # Clasificación por posición en la ocupación
]
```

In [16]:
```
# Función de frecuencias totales respecto a un valor sensible. Obtiene la frecuencia total y genera una tabla con estos valores para cada c
def get_frec_total(tabla, var_id, var_sen):
    # Obtenemos todas los elementos distintos de variables de identificación
    iden = (
        tabla[var_id] # limitamos a las variables de identificación
        .groupby(var_id).sum() # no se suma solo obtenemos los elementos distintos
        .assign(unos=lambda x: 1) # generamos una columna de unos para hacer una union todos con todos
        .reset_index() # quitamos las variables del índex y las ponemos como columnas para unirlas
    )

    # calculamos la distribución total
    dis_tot = (
        tabla
        .value_counts(subset=var_sen,normalize=True) # calculamos las frecuencias relativas
        .to_frame(name="freq_tot") # renombramos la columna de frecuencias
        .assign(unos=lambda x: 1) # generamos una columna de unos para hacer una unión todos con todos
        .reset_index() # quitamos las variables del índex y las ponemos como columnas para unirlas
    )
```

Para cada elemento de los datos sensibles calculamos la *t*-closness

```
umbral = 0.05
for s in variables_sensibles:
    t_c = get_t_closeness(sdem_interes, variables_identificacion, s).copy()
    mask = t_c["t-closenes"] > umbral
    print(f"Grupos con t-closeness mayor a {umbral} para las variables {', '.join(variables_identificacion)}, {s} : {(t_c > umbral).sum()[0
    print(t_c[mask])
```

```
Grupos con t-closeness mayor a 0.05 para las variables ent, cd_a, cs_p13_1 : 41
          t-closenes
ent cd_a
32  32      1.042426
25  24      0.924242
16  15      0.912086
26  25      0.889281
8   9       0.885871
13  43      0.871840
18  27      0.871779
27  18      0.858904
3   40      0.850851
22  36      0.836921
24  7       0.795913
9   1       0.787975
```

Nota: El acceso al ambiente generado para el marco experimental esta en la siguiente referencia https://git.inegi.org.mx/laboratorio-de-ciencia-de-datos/data-privacy-by-design/-/tree/re_identificacion/

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Pruebas para generar datos sintéticos

Método\ Programa	Programa de información seleccionado
Fully Conditional Specification (FCS)	Iteración de pruebas con FCS
Information Preserving Statistical Obfuscation (IPSO)	Iteración de pruebas con IPSO
Generative Adversarial Networks (GAN)	Iteración de pruebas con GAN

Variables de acceso al lago

En la siguiente celda se establecen variables para acceder al lago de datos. Es necesario tener accesos adecuados para obtener los datos.

In [3]:
```
Sys.setenv("AWS_ACCESS_KEY_ID" = readline(prompt = "Usuario: "),
           "AWS_SECRET_ACCESS_KEY" = getPass(prompt = "Contraseña: "),
          "AWS_S3_ENDPOINT"= "lcidn4.inegi.gob.mx:9100"
          )
ruta_archivo <- "s3://infoenoe-private/ENOE_SDEMT120_CAMPOS_SELECCIONADOS.csv"
```

Out [3]:
```
Usuario:  irving.cabrera
Contraseña:  ·····
```

Acceso a los datos del lago

Para este ejercicio se obtuvieron datos de la Encuesta de Ocupación y Empleo. Estos difieren de los datos disponibles públicamente ya que contienen información real de localidad y manzana que son omitidas por cuestiones de privacidad.

Debido a la naturaleza sensible de estos datos, el acceso a esta información es controlado solo a usuarios autorizados. Así también se omitirá desplegar información particular de los datos originales.

In [4]:
```
tabla_sdem  <- as_tibble(s3read_using(read_csv, col_types =cols(.default ="c"), object=ruta_archivo, opts = list(use_https = F)))
```

Variables de interés

Para este ejercicio no requerimos hacer uso de FAC el cual indica el factor de expansión. Por lo que las variables de interés serán:

- **LOC:** Localidad

Debido a la naturaleza sensible de estos datos, el acceso a esta información es controlado solo a usuarios autorizados. Así también se omitirá desplegar información particular de los datos originales.

In [4]:
```
tabla_sdem  <- as_tibble(s3read_using(read_csv, col_types =cols(.default ="c"), object=ruta_archivo, opts = list(use_https = F)))
```

Variables de interés

Para este ejercicio no requerimos hacer uso de FAC el cual indica el factor de expansión. Por lo que las variables de interés serán:

- **LOC:** Localidad
- **MUN:** Municipio
- **T_LOC:** Tamaño de localidad
- **MAN:** Manzana
- **CD_A:** Ciudad auto representada
- **ENT:** Entidad
- **AGEB:** Área GeoEstadística Básica
- **SEX:** Sexo
- **EDA:** Edad
- **NAC_DIA:** Dia de nacimiento
- **NAC_MES:** Mes de nacimiento
- **NAC_ANIO:** Año de nacimiento
- **CS_P13_1:** Nivel escolar
- **POS_OCU:** Posición en la ocupación
- **INGOCUP:** Ingreso del personal ocupado

Como se puede observar, las variables son en general variables de identificación o variables sensibles. Por lo que la aplicación de datos sintéticos resulta ideal.

In [5]:
```
variables_interes <- c('LOC','MUN','T_LOC','MAN','CD_A','ENT','AGEB','SEX','EDA','NAC_DIA','NAC_MES','NAC_ANIO','CS_P13_1','POS_OCU
variables_interes
```

Secuencia de visita

La manera que FCS funciona es que va modelando una variable a la vez. Por lo que es importante establecer el orden en que las columnas son sintetizadas.

Comenzando con las variables "independientes" y después las variables que dependen de las primeras y así sucesivamente.

In [10]:
```
my_visitsequence <- names( c(
    my_visitsequence["ENT"]
    ,my_visitsequence["CD_A"]
    ,my_visitsequence["MUN"]
    ,my_visitsequence["LOC"]
    ,my_visitsequence["T_LOC"]
    ,my_visitsequence["AGEB"]
    ,my_visitsequence["MAN"]
    ,my_visitsequence["SEX"]
    ,my_visitsequence["EDA"]
    ,my_visitsequence["NAC_DIA"]
    ,my_visitsequence["NAC_MES"]
    ,my_visitsequence["NAC_ANIO"]
    ,my_visitsequence["CS_P13_1"]
    ,my_visitsequence["POS_OCU"]
    ,my_visitsequence["INGOCUP"]
)
)
my_visitsequence
```

Out [10]: 'ENT' · 'CD_A' · 'MUN' · 'LOC' · 'T_LOC' · 'AGEB' · 'MAN' · 'SEX' · 'EDA' · 'NAC_DIA' · 'NAC_MES' · 'NAC_ANIO' · 'CS_P13_1' · 'POS_OCU' · 'INGOCUP'

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

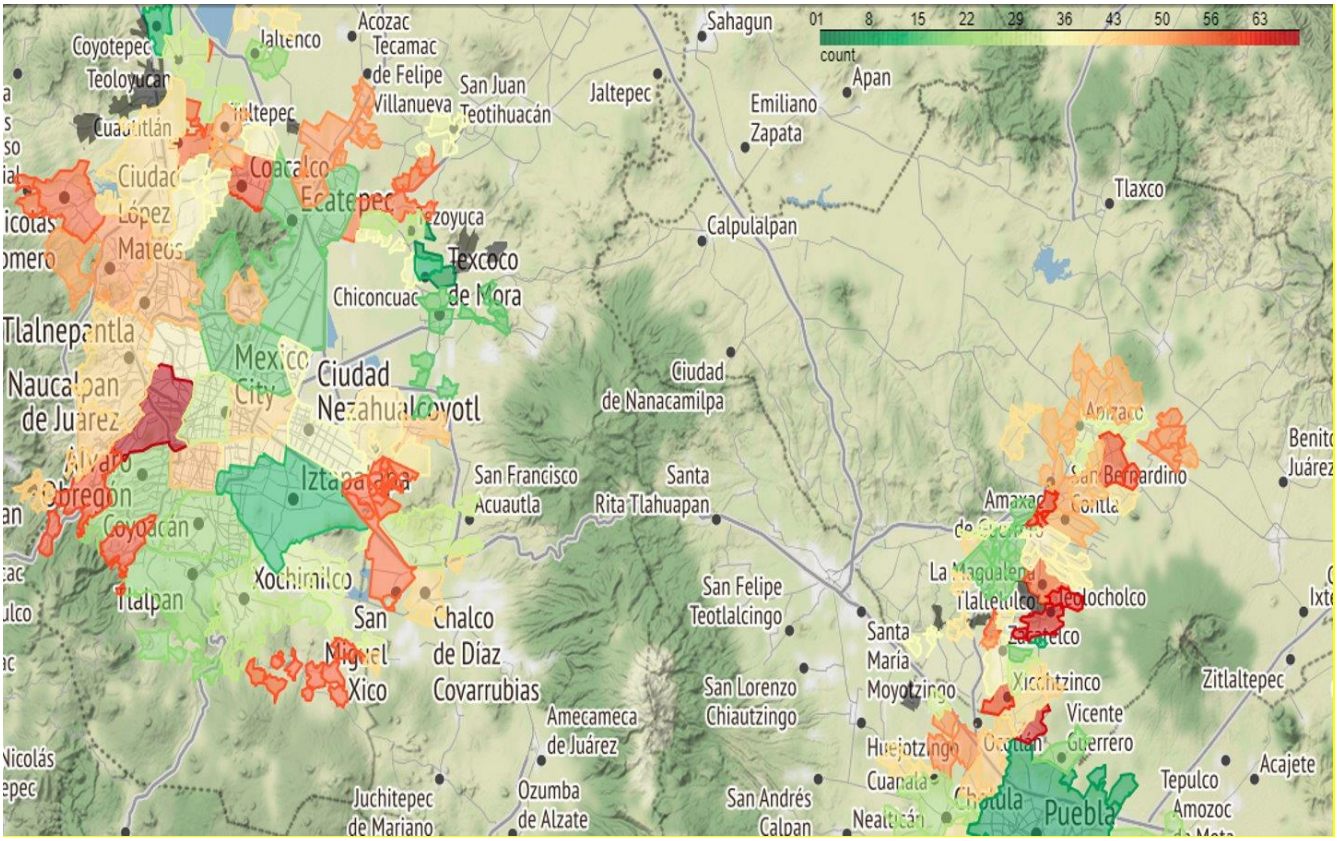
Fase presentación de resultados para evaluación y validación

Para presentar el comportamiento conforme al marco experimental del proyecto fue realizada la siguiente estrategia:

Pruebas de re-identificación

De los (conjunto de datos resultados) obtenidos por cada programa, es generada una visualización considerando las variables del Censo de Población y Vivienda 2020 y el Marco Geoestadístico.

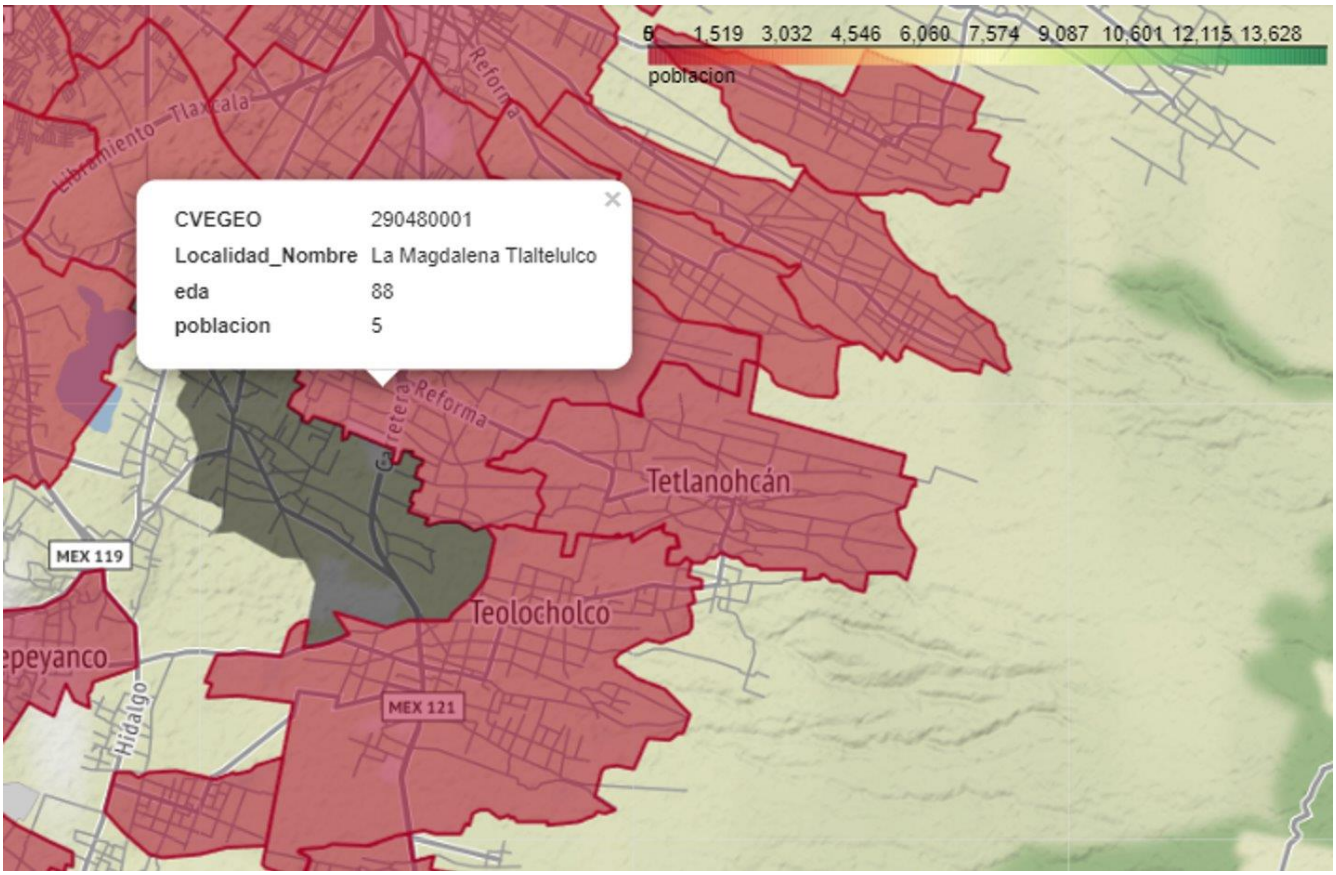
Re-identificación ENOE con K-anonimity 1 a nivel localidad



Apartir de los resultados se seleccionan los K-anonimity igual a 1, los cuales representa que solo existe una persona encuestada de determinada edad, la escala de colores de verde a rojo contabiliza los grupos de edades en donde se encontró una persona ubicando geográficamente las localidades que tienen una persona de X edad hasta 70 personas con X edades diferentes

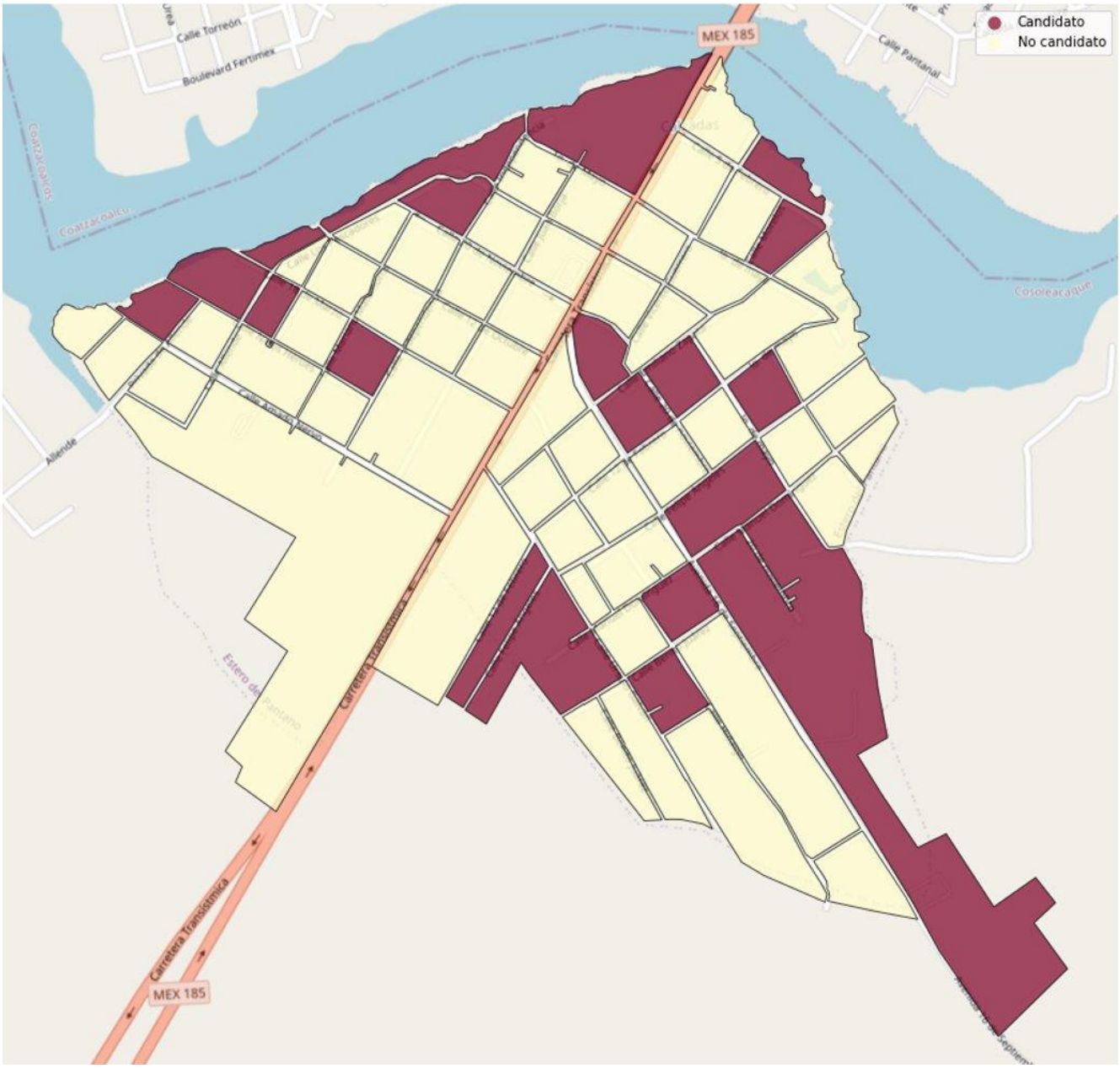
Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Re-identificación ENOE con K-anonimity 1 unión CPV2020 nivel localidad



Los registros de K-anonimity igual a 1 contienen la información de entidad, municipio y ciudad auto representada, esta última se compone de una o más localidades, con esta información fue posible unirla con la información del CPV2020 para así ubicar el total de población que existe en esa localidad por grupos de edades, en el ejemplo se ubicó que en la localidad de la Magdalena en Tlaxcala la encuesta se levantó a una persona del sexo femenino de 88 años en una población donde existen solo 5 mujeres mayores de 85 años

Re-identificación ENOE con K-anonimity 1 unión CPV2020 nivel manzana



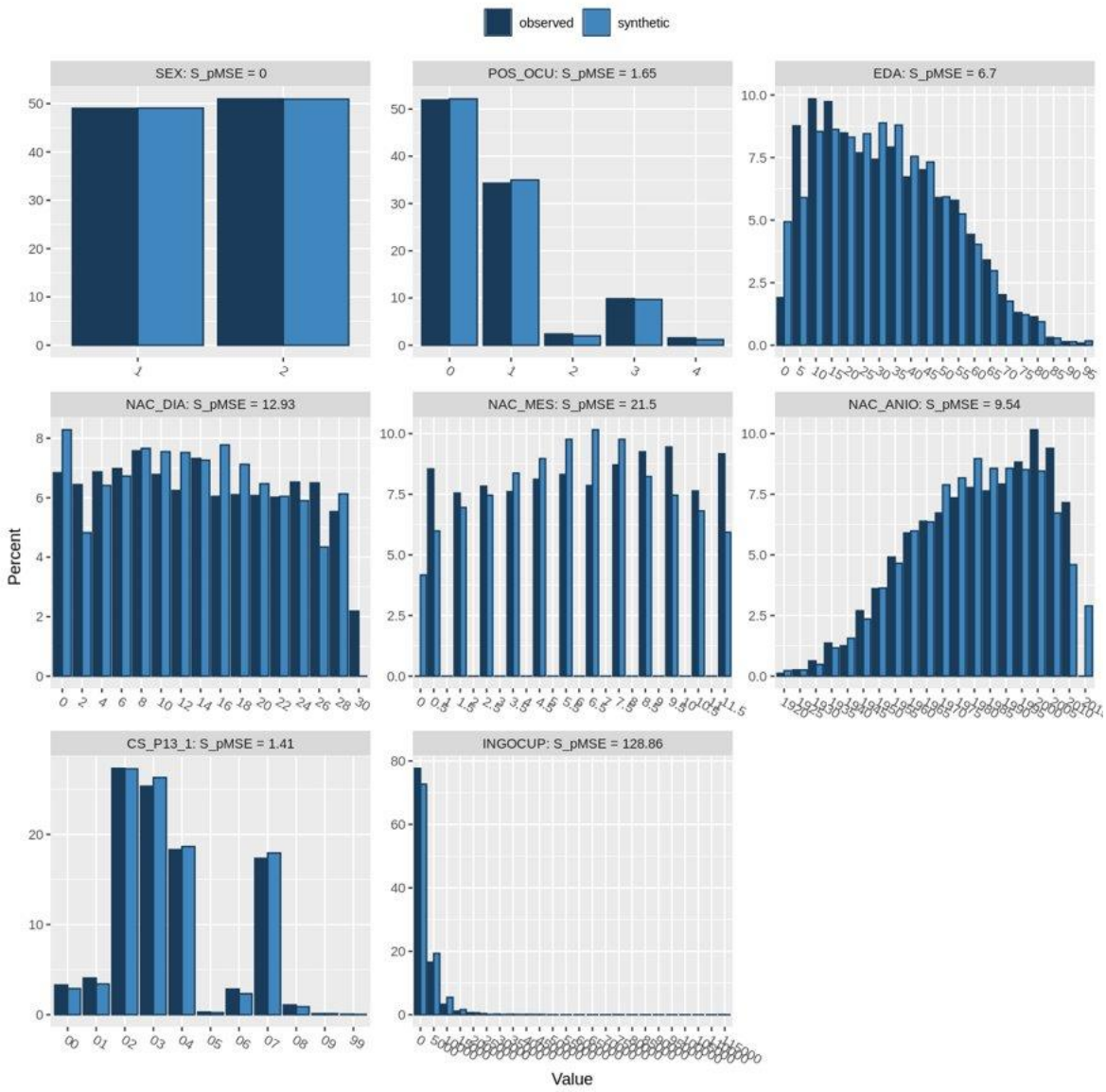
Con los registros de K-anonimity igual a 1 se regresa a la encuesta del ENOE para identificar las edades y sexo del resto de los integrantes del hogar, para hacer una re-identificación con en CPV2020 a nivel manzana se filtran por cada hogar las manzanas que corresponde a la entidad, municipio y ciudad auto representada (una o más localidades) además se filtran los grupos de edades y sexo que componen el hogar, en donde su valor debe ser diferente de 0

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Pruebas para generar datos sintéticos

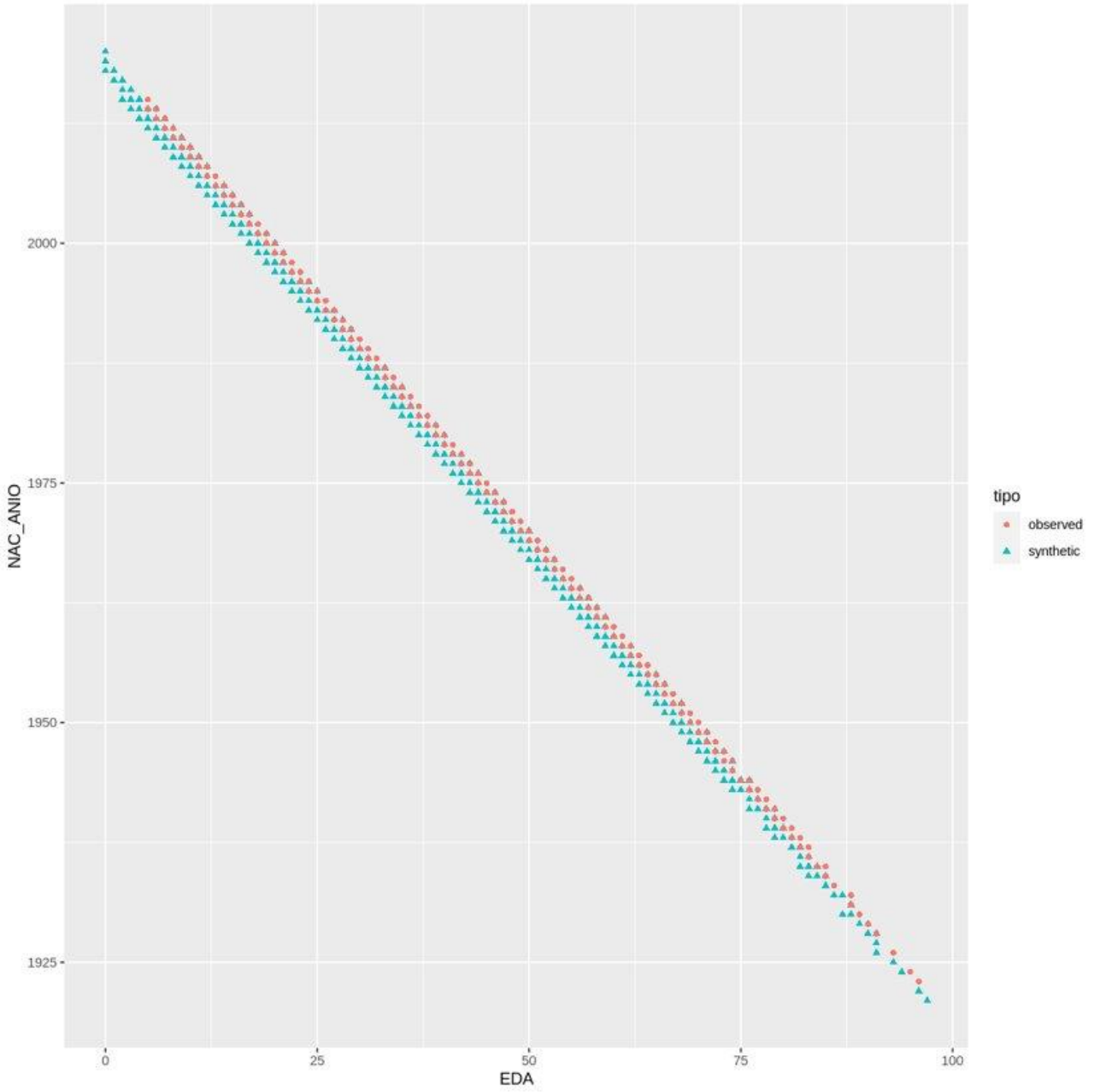
De los resultados obtenidos, fue generado un notebook basado en el IDLE Jupyter LAB, que muestra los métodos y los conjuntos de datos involucrados en el entrenamiento mostrando el equilibrio estadístico del dato real al dato sintético.

Distribución de la población en datos observados y sintéticos para cada variable



Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

Valor del año de nacimiento respecto a la Edad para datos observados y sintéticos

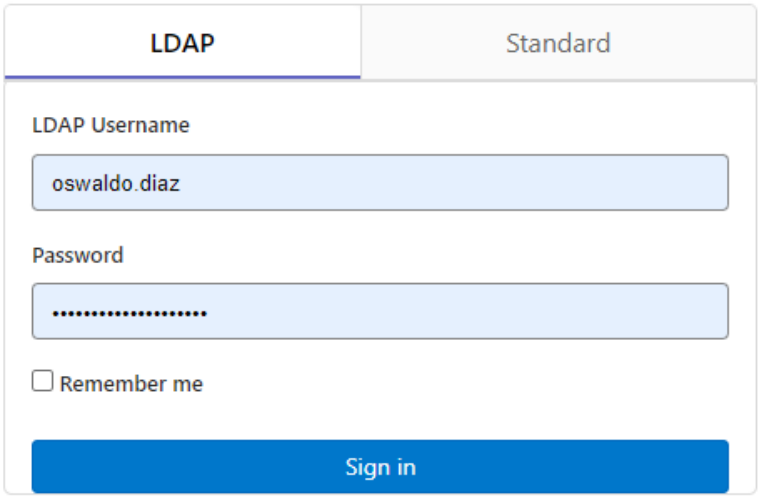


Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

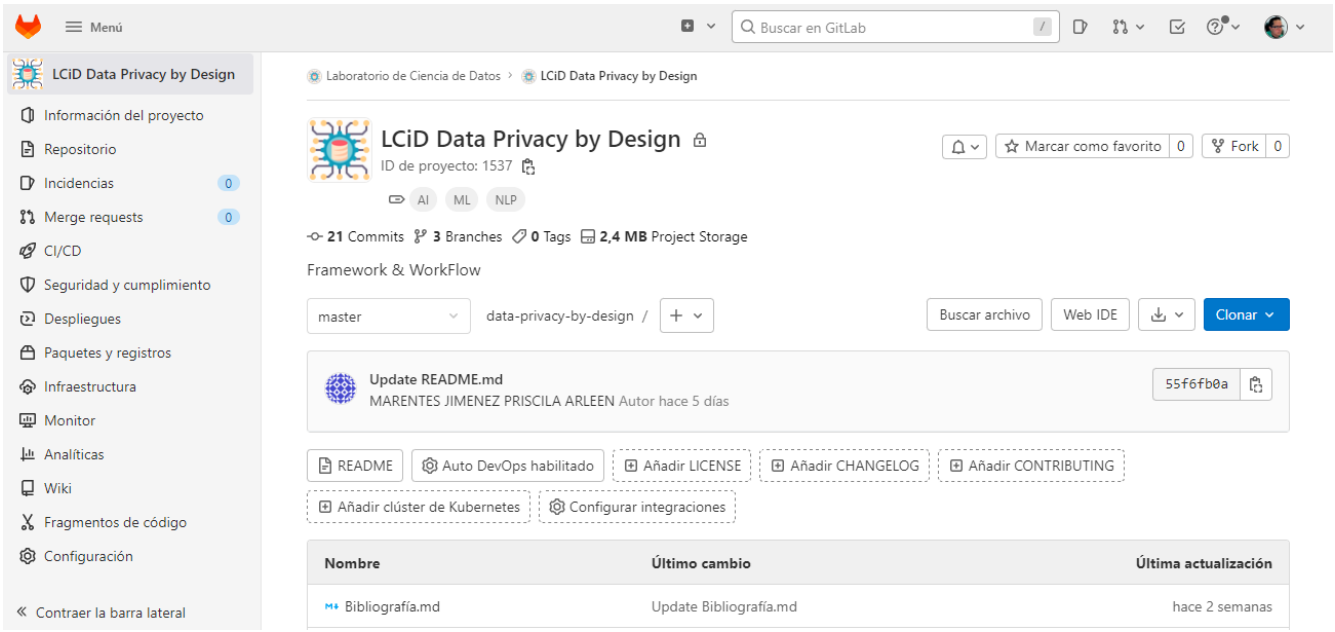
Fase entrega de productos de datos

Para esta fase el proceso de implementación en la arquitectura con base a tecnologías de información de las Unidades Administrativas involucradas en el proyecto fue generado un repositorio donde son descritos los componentes requeridos para habilitar el escenario para el marco experimental en la re-identificación y generación de datos sintéticos, considerando la siguiente referencia
https://git.inegi.org.mx/laboratorio-de-ciencia-de-datos/data-privacy-by-design

El acceso es con las credenciales institucionales

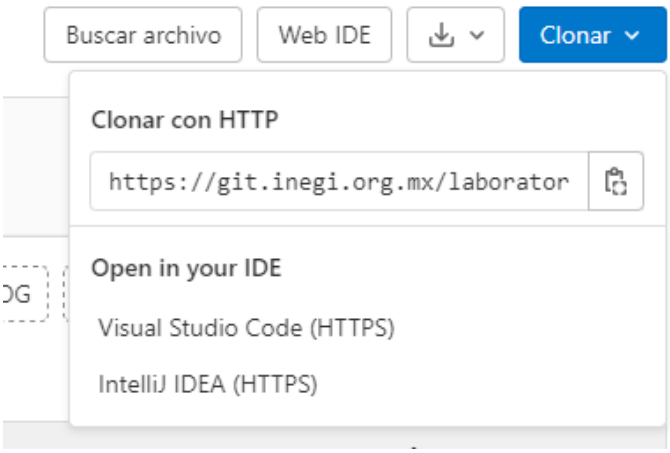


Muestra la siguiente pantalla



En el apartado llamado “master”, dando un clic muestra el apartado para los componentes utilizados en la re-identificación y datos sintéticos, como lo muestra la siguiente pantalla

Una vez seleccionado lo requerido en la parte superior derecha hay un apartado para poder clonar el proyecto en la infraestructura de tecnologías de información habilitada



Proceso técnico para clonar el proyecto

- Instalación del software GIT https://git-scm.com/doc
- Instalación del software para contenerización https://docs.docker.com/get-docker/
- Posterior a la instalación y configuración del software requerido en una terminal para line a de comandos ejecutar

Git clone https://git.inegi.org.mx/laboratorio-de-ciencia-de-datos/data- p rivacy-by-design.git
Nota: en el grupo colaborativo a este proyecto podemos apoyar para la instalación, configuración e implementación de los componentes anteriores mencionados.

Niveles de madurez alcanzados

Los Niveles de Madurez Tecnológica, están basados en el sistema de medición / métrica del proyecto sistemático que respalda la investigación hasta la operación, la aplicación, del producto o servicio, descritos por la por la NASA (Mankins, 1995), considerando los siguientes:

- Nivel 1: Investigación básica, trabajo experimental o teórico realizado principalmente para adquirir nuevos conocimientos de los fundamentos subyacentes de los fenómenos y hechos observables, sin ninguna aplicación o uso particular en vista. La investigación básica puede orientarse o dirigirse hacia algunos campos amplios de interés general, con el objetivo explícito de una gama de aplicaciones futuras.
- Nivel 2: Investigación aplicada, investigación original realizada para adquirir nuevos conocimientos. Sin embargo, se dirige principalmente hacia un objetivo u objetivo práctico y específico. La investigación aplicada se lleva a cabo para determinar los posibles usos de los resultados de la investigación básica o para determinar nuevos métodos o formas de lograr objetivos específicos y predeterminados.
- Nivel 3: Prueba de concepto para sistema, proceso, producto, servicio o herramienta; esto puede considerarse una fase temprana del desarrollo experimental; Se pueden incluir estudios de viabilidad.
- Nivel 4: Evaluación exitosa del sistema, subsistema, proceso, producto, servicio o herramienta en un laboratorio u otro entorno experimental; esto puede considerarse una fase intermedia de desarrollo.
- Nivel 5: Evaluación exitosa del sistema, proceso del subsistema, producto, servicio o herramienta en el entorno relevante a través de pruebas y creación de prototipos; esto puede considerarse la etapa final de desarrollo antes de que comience la demostración.
- Nivel 6: Demostración de un prototipo de sistema, subsistema, proceso, producto, servicio o herramienta en un entorno relevante o de prueba (potencial demostrado).

- Nivel 7: Prototipo de sistema, proceso, producto, servicio o herramienta demostrado en un entorno operativo u otro entorno relevante (funcionalidad demostrada en un entorno del mundo casi real; componentes del subsistema totalmente integrados en el sistema).
- Nivel 8: Sistema, proceso, producto, servicio o herramienta finalizados probados, y se ha demostrado que funciona o funciona como se espera en el entorno del usuario; formación de usuarios y documentación completada; aprobación otorgada por el operador o usuario.
- Nivel 9: Sistema, proceso, producto, servicio o herramienta implementado y utilizado de forma rutinaria.

Con base a los niveles anteriores descritos el proyecto llamado "Privacidad de la información estadística y geográfica con datos sintéticos para microdatos", su nivel de madurez evaluado por los involucrados es 6.

Marco Normativo

Se deben determinar cuáles son las variables que de manera directa o indirecta pueden aportar en la identificación de una persona. Para tal fin es necesario clasificar cada una de las variables, teniendo en cuenta las siguientes categorías:

- Variables de identificación geográfica.
- Variables de identificación directa (datos personales)
- Variables de carácter sensible o confidencial (datos personales sensibles)

Sustento:
Ley federal de transparencia y acceso a la información pública.
Artículo 113

- Se considera información confidencial:
  - I. La que contiene datos personales concernientes a una persona física identificada o identificable;
  - II. Los secretos bancario, fiduciario, industrial, comercial, fiscal, bursátil y postal, cuya titularidad corresponda a particulares, sujetos de derecho internacional o a sujetos obligados cuando no involucren el ejercicio de recursos públicos, y
  - III. Aquella que presenten los particulares a los sujetos obligados, siempre que tengan el derecho a ello, de conformidad con lo dispuesto por las leyes o los tratados internacionales.

Ley general de protección de datos personales en posesión de sujetos obligados
Artículo 3 fracciones IX y X

- Datos personales: Cualquier información concerniente a una persona física identificada o identificable. Se considera que una persona es identificable cuando su identidad pueda determinarse directa o indirectamente a través de cualquier información.

- Datos personales sensibles: Aquellos que se refieran a la esfera más íntima de su titular, o cuya utilización indebida pueda dar origen a discriminación o conlleve un riesgo grave para éste. De manera enunciativa más no limitativa, se consideran sensibles los datos personales que puedan revelar aspectos como origen racial o étnico, estado de salud presente o futuro, información genética, creencias religiosas, filosóficas y morales, opiniones políticas y preferencia sexual.

Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica

Artículo 3 fracciones VI, VII y VIII

- VI. Dato confidencial: Dato que permite identificar, directa o indirectamente, a los Informantes del Sistema y que revela información individual. Incluye los datos clasificados con ese carácter en la legislación, así como los secretos de carácter bancario, fiduciario, industrial, comercial, fiscal, bursátil, postal o de cualquier otro tipo cuya titularidad corresponda a los informantes del Sistema;
- VII. Datos geográficos: Conjunto de atributos representados a través de números y caracteres que describen o identifican fenómenos espaciotemporales relativos al territorio;
- VIII. Datos georreferenciados: Conjunto de números o caracteres que tiene asignadas coordenadas geográficas de acuerdo con un sistema de coordenadas determinado;

Catálogo de datos personales: Criterios y resoluciones para su tratamiento
(Elaborado por la Unidad de Transparencia de la SEMARNAT)

Variable	Justificación de por qué es un dato personal
Sexo	Que el INAI en sus Resoluciones 1588/16 y RRA 0098/17 determinó que el sexo es considerado un dato personal, pues con él se distinguen las características biológicas y fisiológicas de una persona y que la harían identificada o identificable, por ejemplo, sus órganos reproductivos, cromosomas, hormonas, etcétera; de esta manera se considera que este dato incide directamente en su ámbito privado y, por ende, en su intimidad, conforme a lo dispuesto en el artículo 113, fracción I, de la Ley federal de Transparencia y Acceso a la Información Pública.
Edad y fecha de nacimiento	Que el INAI en la Resolución RRA 0098/17 señaló que tanto la fecha de nacimiento como la edad son datos personales, toda vez que los mismos consisten en información concerniente a una persona física identificada o identificable. Ambos datos están estrechamente relacionados, toda vez que, al dar a conocer la fecha de nacimiento, se revela la edad de una persona. Se trata de datos personales confidenciales, en virtud de que al darlos a conocer se afectaría la intimidad de la persona titular de los mismos. Por lo anterior, el INAI considera procedente su clasificación, en términos del artículo 113, fracción I, de la Ley Federal de Transparencia y Acceso a la Información Pública.

Estado civil	Que en la Resolución RRA 0098/17 el INAI señaló que el estado civil constituye un atributo de la personalidad que se refiere a la posición que ocupa una persona en relación con la familia; en razón de lo anterior, por su propia naturaleza es considerado como un dato personal, en virtud de que incide en la esfera privada de los particulares y, por ello, es clasificado con fundamento en el artículo 113, fracción I, de la Ley Federal de Transparencia y Acceso a la Información Pública.
Profesión u ocupación	Que en su Resolución RRA 1024/16, el INAI determinó que la profesión u ocupación de una persona física identificada constituye un dato personal que, incluso, podría reflejar el grado de estudios, preparación académica, preferencias o ideología. Por lo que se actualiza su clasificación como información confidencial de conformidad con el artículo 113, fracción I de la Ley de la materia.
Ingresos por concepto de renta (patrimonio)	Por lo que respecta a los ingresos por concepto de renta (patrimonio), este Comité de Transparencia analizó que se trata de información concerniente a una persona física a través de la cual puede ser identificada o identificable, por lo que actualiza el supuesto previsto en los artículos 116, primer párrafo de la LGTAIP, artículo 113, fracción I de la LFTAIP, aunado a que requieren el consentimiento de los particulares para permitir el acceso al mismo de conformidad con lo dispuesto en los artículos 120 primer párrafo de la LGTAIP, primer párrafo del artículo 117 de la LFTAIP.
Información relativa al estado de salud	Descripción del estado de salud, condición o riesgos, registros, anotaciones, en su caso, constancias y certificaciones correspondientes de la atención médica del paciente, por ende, datos personales que han de protegerse con fundamento en los artículos 113, fr. I, y segundo transitorio LFTAIP, 3, fr. II, 18, fr. II, y 21 LFTAIPG, 37 y 40 RLFTAIPG. [Ver Expediente clínico]

Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica

Artículo 11.- En la producción de Información Estadística y Geográfica, las Unidades del Estado deberán implementar medidas para:

V. Asegurar que la difusión de la Información se realiza de manera que los Informantes del Sistema y, en general, las personas físicas o morales objeto de la Información no sean identificados de manera directa o indirecta, por lo que previo a la entrega de resultados deben evaluar el riesgo de Identificación de acuerdo con lo siguiente:

a) Se considera que el nivel de riesgo es alto cuando la Identificación es inmediata. Con sólo acceder a la Información es posible reconocer a la persona física o moral a la que corresponden los datos; o cuando la Identificación se deduce de la combinación de diferentes variables contenidas en el mismo producto de difusión de la Información;

b) Se considera que el nivel de riesgo es medio cuando la Identificación se logra realizando la suma o resta de algunas clases o agrupamientos de este tabulado y combinando el resultado con otros productos de Información Estadística o Geográfica;
c) Se considera que el nivel de riesgo es bajo cuando la Identificación se logra combinando la Información con distintos repositorios públicos y privados de datos utilizando técnicas de análisis, software y equipo de cómputo, y

d) Se considera que el nivel de riesgo es nulo cuando no es posible realizar la Identificación por cualquier medio o por la combinación de cualquier variable de Información.

Las Unidades del Estado, cuando contraten a un tercero para la producción de la Información, deben verificar que las medidas referidas en el presente artículo queden establecidas en los contratos respectivos, así como vigilar su cumplimiento.

Conclusiones

- Integrar un prototipo de datos sintéticos que permita generar mecanismos para utilizar información sin la identificación de los Informantes.

- Evaluar los niveles de madurez en la privacidad de datos sensibles para la toma de decisiones

- Referente a las pruebas de re-identificación, los resultados permiten identificar que la materialización del riesgo es bajo en los programas de información considerados. Lo anterior conforme al artículo 11 fracción V, inciso, c, de la Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica.

- En relación con los datos sintéticos, es posible considerarlos como una alternativa adicional para la publicación de microdatos, la cual también evita la re-identificación de los Informantes.

Referencias

[1] UNECE. (2009). Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes. Ginebra: UNECE.
[2] UNECE. (2007). Managing Statistical Confidentiality & Microdata Access. Principles and guidelines of good practice. Ginebra: UNECE.
[4] INEGI. (diciembre 18, 2021). Guía para la Gestión de la Confidencialidad en la Información Estadística y Geográfica. agosto 1, 2022, de SNIEG Sitio web: https://www.snieg.mx/Documentos/Normatividad/Vigente/guia-gestion-confiden_infestgeog.pdf
[5] INEGI. (octubre 29, 2021). Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica. agosto 2, 2022, de SNIEG Sitio web: https://www.snieg.mx/Documentos/Normatividad/Vigente/politica_inf_estad_geog.pdf
[6] Grupo de trabajo sobre protección de datos del artículo 29. (octubre, 2018). *Dictamen 05/2014 sobre técnicas de anonimización. Adoptado el 10 de abril de 2014*. 1 agosto, 2022, de Gahazas Sitio web: https://gahazas.files.wordpress.com/2018/10/wp216_es_-tc3a9cnicas-de-anonimizacic3b3n.pdf
[7] INEGI. (enero 25, 2021). *Presentación de resultados*. agosto 3, 2022, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/programas/ccpv/2020/doc/Censo2020_Principales_resultados_ejecutiva_EUM.pdf
[8] INEGI. (septiembre, 2020). *Encuesta Nacional de Ocupación y Empleo (Nueva Edición) (ENOE)*. agosto 4, 2022, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/programas/enoe/15ymas/doc/enoe_n_presentacion_ejecutiva_0720.pdf
[9] INEGI. (julio 28, 2021). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020.* agosto 3, 2022, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/programas/enigh/nc/2020/doc/enigh2020_ns_presentacion_resultados.pdf
[10] INEGI. (octubre 29, 2021). *Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica*. agosto 2, 2022, de SNIEG Sitio web: https://www.snieg.mx/Documentos/Normatividad/Vigente/politica_inf_estad_geog.pdf
[11] INEGI. (diciembre 18, 2021). *Guía para la Gestión de la Confidencialidad en la Información Estadística y Geográfica*. agosto 1, 2022, de SNIEG Sitio web: https://www.snieg.mx/Documentos/Normatividad/Vigente/guia-gestion-confiden_infestgeog.pdf
[12] Lahera G. (febrero 16, 2022). *Los Datos Sintéticos y sus Beneficios para las Empresas Data-Driven*. agosto 5, 2022, de Medium Sitio web: https://gmnl.medium.com/los-datos-sint%C3%A9ticos-y-sus-beneficios-para-las-empresas-data-driven-bd68552a71b8
[13] Lan L, You L, Zhang Z, Fan Z, Zhao W, Zeng N, Chen Y & Zhou X. (12 mayo, 2020). *Adversarial generative networks and their applications in biomedical informatics*. agosto 1, 2022, de Frontiers in Public Health Sitio web: https://www-frontiersin-org.translate.goog/articles/10.3389/fpubh.2020.00164/full?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=op,sc
[14] Montagud, N. (noviembre 6, 2020). *Redes neuronales profundas: qué son y cómo funcionan*. agosto 2, 2022, de Psicología y mente Sitio web: https://psicologiaymente.com/cultura/redes-neuronales-profundas

[15] Oswaldo Díaz, Elio Villaseñor. (2021). Preservación de la privacidad mediante el uso de datos sintéticos. Laboratorio de Ciencia de Datos. 2021, De INEGI.
[16] Rivera, M. (septiembre 13, 2019). *DCGAN: Redes Generadoras Antagónicas Convolucionales Profundas.* agosto 4, 2022, de CIMAT Sitio web: http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje_profundo/dcgan/dcgan.html
[17] Urcuqui López, C., Peña, M., Osorio Quintero, J., & Navarro, A. (2018). Ciberseguridad: un enfoque desde la ciencia de datos. https://doi.org/10.18046/EUI/ee.4.2018
[18] Vega, Belén & Rubio-Escudero, Cristina & Riquelme, José & Nepomuceno-Chamorro, Isabel. (2020). Creation of Synthetic Data with Conditional Generative Adversarial Networks. 10.1007/978-3-030-20055-8_22.
[19] Das, Hari Prasanna & Spanos, Costas. (2022). Conditional Synthetic Data Generation for Personal Thermal Comfort Models.
[20] Yale, Andrew & Dash, Saloni & Dutta, Ritik & Guyon, Isabelle & Pavao, Adrien & Bennett, Kristin. (2020). Generation and Evaluation of Privacy Preserving Synthetic Health Data. Neurocomputing. 416. 10.1016/j.neucom.2019.12.136.
[21] Garcia Torres, Douglas. (2018). Generation of Synthetic Data with Generative Adversarial Networks.
[22] Park, Noseong & Mohammadi, Mahmoud & Gorde, Kshitij & Jajodia, Sushil. (2018). Data Synthesis based on Generative Adversarial Networks.
[23] Aziira, A & Setiawan, N & Soesanti, I. (2020). Generation of Synthetic Continuous Numerical Data Using Generative Adversarial Networks. Journal of Physics: Conference Series. 1577. 012027. 10.1088/1742-6596/1577/1/012027.
[24] Architecture, Forensic. (2022). Experiments in Synthetic Data. 10.1002/9781119815075.ch50.
[25] Horvath, Blanka. (2022). Synthetic Data for Deep Learning. Quantitative Finance. 22. 423-425. 10.1080/14697688.2022.2048062.
[26] Nikolenko, S.I. (2021). Synthetic Data for Basic Computer Vision Problems. In: Synthetic Data for Deep Learning. Springer Optimization and Its Applications, vol 174. Springer, Cham. https://doi.org/10.1007/978-3-030-75178-4_6
[27] M. Hittmeir, A. Ekelhart and R. Mayer, "Utility and Privacy Assessments of Synthetic Data for Regression Tasks," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5763-5772, doi: 10.1109/BigData47090.2019.9005476.
[28] Zhang, Tianwei. (2018). Privacy-preserving Machine Learning through Data Obfuscation.
[29] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648
[30] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). -Diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007), 52 pages. DOI = 10.1145/1217299.1217302 http://doi.acm.org/10.1145/1217299.1217302
[31] Li, Ninghui; Li, Tiancheng; Venkatasubramanian, Suresh (2007). "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity". t-Closeness: Privacy beyond k-anonymity and l-diversity (PDF). pp. 106–115. doi:10.1109/ICDE.2007.367856
[32] UNECE. (2007). Managing Statistical Confidentiality & Microdata Access. Principles and guidelines of good practice. Ginebra: UNECE.

[33] UNECE. (2009). Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes. Ginebra: UNECE.
[34] Ley federal de transparencia y acceso a la información pública. Artículo 113. 20 de mayo de 2021.
[35] Ley general de protección de datos personales en posesión de sujetos obligados. Artículo 3. 26 de enero de 2017.
[36] INEGI. (octubre 29, 2021). Política para la Gestión de la Confidencialidad en la Información Estadística y Geográfica. agosto 2, 2022, de SNIEG Sitio web: https://www.snieg.mx/Documentos/Normatividad/Vigente/politica_inf_estad_geog.pdf

Proyecto: Privacidad de la información estadística y geográfica con datos sintéticos para microdatos

A N E X O S

Programa de trabajo

Actividad	Fecha de entrega
• Definir programas de información para el caso de uso.	14 de marzo
• Analizar medidas para evaluar el riesgo de privacidad y confidencialidad estadística.	8 de abril
• Definir aspectos conceptuales y técnicos de microdatos de los programas de información seleccionados	27 de junio
• Definir variables de los programas de información seleccionados, que incrementen el riesgo de re-identificación.	12 de julio
• Analizar medidas para evaluar el riesgo de seguridad y privacidad en los programas de información seleccionados	5 de agosto
• Reportar avances en el Comité de Seguridad y Confidencialidad Estadística de la información.	20 de septiembre
• Pruebas de re-identificación con las variables de los programas de información.	17 de octubre
• Generar reporte preliminar de investigación.	17 de octubre
• Generar datos sintéticos de los programas de información seleccionados.	14 de noviembre

• Generar reporte final de investigación:
  • Resultados de pruebas de re-identificación.
  • Propuesta de datos sintéticos.
	30 de noviembre

Arquitectura para realizar el proyecto