

# Reporte técnico

 Técnicas de  
Anonimización aplicadas  
al REBPE, 2022

Noviembre, 2024



<b>Organización</b>	Instituto Nacional de Estadística y Censos.
<b>Proyecto/Proceso</b>	Plan Estratégico Institucional - PEI
<b>Actividad</b>	Determinar la técnica más adecuada para anonimizar la base de datos del Registro Estadístico Base de Población de Ecuador (REBPE)
<b>Medio de verificación</b>	Informe de Técnicas de Anonimización aplicadas al REBPE

FIRMAS DE APROBACIÓN:		
ELABORADO POR:	REVISADO POR:	APROBADO POR:
<b>Sonia Herrera</b> Analista de Planificación y Metodologías de Registros Administrativos	<b>Marco Viteri</b> Responsable de Gestión de Planificación y Metodologías de Registros Administrativos	<b>Paúl Benavides</b> Director de Registros Administrativos



# Abstract

El reporte técnico presentado a continuación tiene como propósito determinar la técnica más adecuada para anonimizar la base de datos del Registro Estadístico Base de Población de Ecuador (REBPE). Se describen de forma general las técnicas utilizadas para anonimizar un dato, así como también, los algoritmos más comunes para verificar el grado de anonimización que debe ser aplicado en una base de datos. La clasificación de las variables dentro del conjunto de datos, permitió aplicar reglas para reducir la vulnerabilidad de re identificación. El enfoque metodológico aplicado es el algoritmo de k anonimato considerando los *quasi-identificadores* de los que dispone el REBPE. Adicionalmente, se ha implementado la librería del lenguaje de programación Python “pyCANON” para robustecer el análisis del anonimato a través de los algoritmos *l-diversity* y *t-closeness*.



## Contenido

Marco Teórico .....	1
Técnicas utilizadas para anonimizar un dato .....	1
1. Remover o suprimir atributos .....	1
2. Reemplazo de carácter .....	2
3. Permutación.- .....	2
4. Adición de ruido .....	2
5. Generalización.- .....	2
5.1 K- Anonimato .....	2
5.2 L- Diversity .....	3
Metodología .....	6
Enfoque .....	6
Objetivo General .....	6
Algoritmo de Anonimización del REBPE utilizando k-anonimato .....	6
Reglas para reducir la vulnerabilidad de re identificación .....	7
Reporte de anonimización: Propiedades de anonimato l-diversity y t-closeness .....	8
Referencias .....	9

### Índice de Tablas

<b>Tabla 1:</b> Antes de anonimización .....	1
<b>Tabla 2:</b> Después de anonimización .....	1
<b>Tabla 3:</b> Sin aplicar k anonimato .....	2
<b>Tabla 4:</b> Aplicando k anonimato .....	2
<b>Tabla 5:</b> Técnicas de anonimización vs. Riesgo de re-identificación .....	3
<b>Tabla 6:</b> Nivel de Riesgo y $f_k$ .....	4
<b>Tabla 7:</b> Frecuencia k anonimato de Riesgo alto .....	7

### Índice de Ilustraciones

<b>Ilustración 1:</b> Resultados de anonimato utilizando pyCANON .....	8
------------------------------------------------------------------------	---



## Marco Teórico

La necesidad de publicar datos abiertos en la medida de mantener informada a la sociedad consolida el desarrollo de herramientas que garanticen la privacidad en la publicación de la información (Díaz y López, 2022).

Estudios ejecutados en el Buró de Censos de Estados Unidos, Díaz y López (2022, pág. 1) revelaron que no basta con remover variables de identificación de la base de datos para evitar que un individuo sea identificado, para esto se clasificaron en tres conceptos clave los atributos de la base de datos: i) identificadores, variables de identificación directa, por ejemplo: nombre, número de identificación. ii) quasi-identificadores (QI), variables que a primera vista no parecen mostrar información relevante, pero que combinadas pueden llevar a identificar a un individuo, por ejemplo: edad, género, ciudad. iii) atributos sensibles (SA), variables que muestran información sensible.

### Técnicas utilizadas para anonimizar un dato

Las tres técnicas más comunes para cambiar un dato, Ferreira y Bernandino (2020) son reemplazar, modificar o remover un atributo o registro.

1. **Remover o suprimir atributos.**- Utilizada cuando un atributo o variable no es relevante para el análisis, una vez removido, es imposible recuperar la información, la Tabla 1 y 2 describe un data set antes y después de ser anonimizado esta técnica:

**Tabla 1:** Antes de anonimización

Nombre	Tutor	Puntaje
Ana	Roger	54
Paúl	Seda	65
Juan	Seda	72
José	Roger	80

Elaborado por: Dirección de Registros Administrativos

**Tabla 2:** Después de anonimización

Tutor	Puntaje
Roger	54
Seda	65
Seda	72
Roger	80

Elaborado por: Dirección de Registros Administrativos



2. **Reemplazo de carácter.**- Los caracteres de un atributo o valores de un dato son sustituidos con un término o símbolo predefinido, por ejemplo x o \*, por ejemplo: el código de cantón 170150 es reemplazado por 17xxxx.
3. **Permutación.**- El data set es reorganizado de forma randómica, esta técnica es aplicada cuando se analiza un atributo y no es necesario relacionarlo con otros, por ejemplo, el análisis de la masa salarial en una región determinada y la permutación no tiene influencia en los resultados.
4. **Adición de ruido.**- Una forma de aplicar esta técnica es por ejemplo, añadiendo o sustrayendo días o meses a una fecha determinada.
5. **Generalización.**- Consiste en la generalización de los atributos al cambiar la escala o la magnitud. Un ejemplo es reemplazar el atributo (dd/mm/aaaa) por el atributo año. Existen dos técnicas consideradas de generalización: K-Anonimato y L-Diversidad.

5.1 K- Anonimato.- Esta técnica consiste en agrupar los registros de K individuos dentro de una categoría, agrupándolos bajo la misma combinación (Ferreira y Bernardino, 2020, pág. 237). En el ejemplo a continuación, los atributos identificados son edad y tipo de discapacidad, con K=3 el data set anonimizado tendrá al menos tres registros para cada combinación de los atributos identificados. Si se tienen dos individuos (Tabla 3) con sus valores individuales, posterior se utiliza la técnica de k-anonimato (Tabla 4).

**Tabla 3:** Sin aplicar k anonimato

Nombre	Edad	Tipo discapacidad
Paul	24	Auditiva
Michael	40	Visual

Elaborado por: Dirección de Registros Administrativos

**Tabla 4:** Aplicando k anonimato

Rango de edad	Tipo discapacidad
21-30	Auditiva
21-30	Auditiva
21-30	Auditiva
31-40	Visual
31-40	Visual
31-40	Visual

Elaborado por: Dirección de Registros Administrativos



5.2 L- Diversity.- Similar al k -anonimato, en el cual al menos L distintos valores deben existir para cada grupo equivalente y los atributos sensibles. El objetivo de esta técnica es limitar la ocurrencia de equivalencia de clase con baja variabilidad de los atributos (Machanavajjhala et al., 2007).

Como resumen se presentan las técnicas de anonimización y el riesgo de re-identificación para cada una:

**Tabla 5:** Técnicas de anonimización vs. Riesgo de re-identificación

Técnica	Permite re-identificación
Supresión	No
Reemplazo de caracteres	Si
Permutación	Si
Adición de ruido	Si
K-Anonimato	No (mínimo)
L-Diversidad	No

Fuente: Adaptado de (Ferreira y Bernandino, 2020)

De acuerdo a Díaz y López (2022, pág. 2), 3 son las técnicas o algoritmos de anonimización que ayudan en la verificación del grado de anonimización de un data set, considerado un listado de quasi-identificadores (QI) y atributos sensibles (SA). Estas técnicas previo su implementación toman en cuenta el término equivalencia de clase, la cual consiste en particionar una base de datos de tal manera que los quasi-identificadores tengan el mismo valor.

Algoritmos de anonimización más utilizados:

1. K-anonimato (k-anonymity)<sup>1</sup>.- Para cada fila de la base de datos existen al menos k-1 filas que no se pueden distinguir con respecto a los quasi-identificadores, note que  $k \geq 1$  es siempre verificado.
2. L-diversity.- En el caso de un atributo sensible, este se satisface si por cada equivalencia de clase, existen al menos l distintos valores para S, note que  $l \geq 1$  es siempre verificado.
3. t-closeness.- Una base de datos con un atributo sensible S es verificado con t-closeness, si todas las equivalencias de clase son verificadas, es decir, si la distribución de los valores de S están a una distancia no cercana que la distribución de los atributos sensibles en todo el conjunto de datos.

La reducción del riesgo de re identificar personas ha sido implementada en entornos para el lenguaje integrado de programación, como RStudio. Uno de estos paquetes es el sdcmicro "Statistical Disclosure Control para micro datos" (Templet et al., 2015 pp. 4-

<sup>1</sup> La descripción en inglés corresponde a las técnicas aplicadas en el lenguaje de programación Python, bajo la librería pyCANON.



14). El mismo que se basa en el concepto para anonimizar micro datos, requiriendo de clasificar las variables de la siguiente manera:

- **Identificadores directos.**- son variables que identifican de forma precisa una unidad estadística, por ejemplo: nombres de personas, direcciones, número único de identificación, número de la seguridad social.
- **Variables clave.**- variables que consideradas juntas permiten identificar a una unidad estadística, por ejemplo: al combinar variables como el sexo, edad, región, y ocupación un individuo puede ser identificado. Estas también son conocidas como **quasi- identificadores** o **identificadores implícitos**.
- **Variables no confidenciales.**- variables que no son identificadores directos o variables clave.
- **Variables sensibles.**- son utilizadas para métodos específicos tales como l-diversity, por ejemplo: la descripción de enfermedades.

Una vez que se han clasificado los atributos de la base de datos con las variables anteriormente mencionadas, se plantean los siguientes pasos para llevar a cabo la anonimización:

1. Remover de la base de datos las variables que contienen identificadores directos.
2. Seleccionar las variables clave categóricas que formarán parte de los **quasi- identificadores**, y almacenarlas dentro del vector **keyvar**.
3. Identificar las variables numéricas y almacenarlas dentro del vector **numvar**.

Al crear un objeto **sd**, se calcula el tipo de riesgo individual, mismo que representa la probabilidad de riesgo del registro a nivel de los **quasi-identificadores** en referencia a la población evaluada del REBPE (Templet et al., 2015, pág. 14).

El valor de **k** anonimato según este método, está representado por **fk**, el mismo que indica el número de registros u observaciones que comparten valores iguales e indistinguibles en los atributos o variables seleccionadas como **quasi-identificadores**, por ejemplo, en la Tabla 5 se observa que existen 7 individuos que contienen valores indistinguibles en sus atributos correspondientes a los **quasi-identificadores**, con una probabilidad de riesgo de ser re identificados del 0.0016.

**Tabla 6: Nivel de Riesgo y fk**

Riesgo	fk
0.001663	7
0.000555	19
0.002493	5

Elaborado por: Dirección de Registros Administrativos





La técnica del k-anonimato requiere que cada combinación de valores de los quasi-identificadores aparezca al menos k veces en la base de datos, por lo que técnicas como l-diversity y t-closeness proveen un nivel más fuerte de anonimización (Technische Universität München, 2013, pág. 41).

L-diversity requiere de requisitos adicionales para las tuplas o grupos de k en el conjunto de datos. La base de datos debe contener al menos l valores diferentes para las variables o atributos sensibles.

T-closeness es una técnica de anonimización superior a l-diversity (Li N. and Li T., 2007), de tal manera que generaliza cada conjunto de datos de tal forma que la distribución de los atributos sensibles de los diferentes grupos de k difieren lo mínimo posible.



## Metodología

### Enfoque

Considerando lo mencionado en el marco teórico y las características de la base de datos del REBPE se ha considerado como la técnica más adecuada para la anonimización el k-anonimato Díaz y López (2022), misma que, permitirá verificar los valores resultantes de la anonimización: fk, y Risk (riesgo), a partir del set de quasi-identificadores "keyvar".

Adicionalmente se aplicará la técnica l-diversity y t-closeness a la base de datos para obtener un reporte más robusto al respecto de los valores de los atributos sensibles.

### Objetivo General

Determinar un método que permita la anonimización de la base estadística del REBPE acorde a la realidad de los atributos que dispone, considerando de las 3 técnicas como prioritarias el k anonimato.

### Algoritmo de Anonimización del REBPE utilizando *k-anonimato*

1. Cargar la base de datos del REBPE.
2. Generar una variable clave única o keyvar a partir de los quasi-identificadores: sexo, cant\_res, e\_civil, es\_madre, es\_padre, padres\_fall, tiene\_disc, tipo\_disc, es\_jubilado, edad, anios\_viud, edad\_viud.
3. Reemplazar el valor NA de los atributos o variables que lo contienen por "-".
4. Generar un conteo de la clave única, para identificar el valor de fk por registro u observación.
5. Crear una nueva tabla con los conteos del paso anterior por clave única (fk).
6. Vincular la columna fk con la tabla original del REBPE.
7. Obtener una frecuencia de los casos cuyo valor de fk sea igual a 1, 2 y 3; es decir, aquellos casos con alto riesgo de ser re identificados.
8. Aplicar las reglas para reducir el riesgo de vulnerabilidad sobre las nuevas variables o atributos creados para aplicar la técnica de anonimización.

**Tabla 7: Frecuencia k anonimato de Riesgo alto**

fk	Frecuencia	Porcentaje de Riesgo
1	355.644	2.09%
2	83.696	0.49%
3	38.813	0.23%
Total	478.153	2.69%

Elaborado por: Dirección de Registros Administrativos

La Tabla 6 indica los casos cuyo valor de k anonimato es menor igual a 3 y que han sido identificados con mayor riesgo de re identificación.

Una vez que se ha determinado el valor del riesgo de re identificación en toda la base de datos cercano al 3%, es necesario aplicar reglas que protejan la privacidad de los individuos que se encuentran en estos grupos fk, considerados de alta vulnerabilidad.

### Reglas para reducir la vulnerabilidad de re identificación

Para determinar las reglas de reducción de re identificación se tomará en cuenta los siguientes parámetros:

- ✓ quasi-identificadores: sexo, cant\_res, e\_civil, es\_madre, es\_padre, padres\_fall, tiene\_disc, tipo\_disc, es\_jubilado, edad, anios\_viud, edad\_viud,
- ✓ Técnica "Reemplazo de caracter" para reducir el grado de vulnerabilidad al exponer o publicar la base de datos del REBPE.
- ✓ Valores de fk con alto riesgo (Risk), es decir fk=1, 2 y 3.

Una vez que se haya segmentado o filtrado el grupo de alto riesgo de la base de datos del REBPE, se aplicará la técnica reemplazo de caracter a la base de datos, en el siguiente orden:

1. Si fk=1, 2 y 3; y, cant\_res es diferente de NULL, reemplazar los dos últimos dígitos de cant\_res por \*\*.
2. Si edad >99, enmascarar los dos últimos dígitos de cant\_res por \*\*, y las categorías de tipo\_disc por \*.
3. Si edad\_viud >56, enmascarar los dos últimos dígitos de cant\_res por \*\*.

A pesar de que la variable sexo en la categoría indeterminado cuenta con una población significativamente reducida, se descarta el enmascaramiento de la variable cant\_res en sus dos últimos dígitos, debido a que el valor con el que cuenta este grupo en esta variable es 9999.



## Reporte de anonimización: Propiedades de anonimato *l-diversity* y *t-closeness*

El resultado de aplicar *l-diversity* a una base de datos, significa que por cada quasi-identificador, es decir la combinación de atributos por los que potencialmente se puede identificar a un individuo, debería existir al menos *l* valores bien representados para las variables o atributos sensibles.

Para el caso de la base de datos del REBPE la variable o atributo sensible considerada fue el tipo de discapacidad (*tipo\_disc*). En la Ilustración 1 el resultado de aplicar la técnica *l-diversity* resultó igual a 1, lo que indica que la base de datos no satisface la propiedad de *l-diversity*, en otras palabras, representa que por al menos un quasi-identificador, existe solo un único valor por cada atributo sensible.

Por lo tanto, para mejorar la preservación de la privacidad de la base de datos se aplicó la técnica *t-closeness*, cuyo resultado obtenido de 0.989 indica que el conjunto de la base de datos satisface el requisito de privacidad, corroborando lo explicado en el marco teórico, de que esta última técnica provee un nivel de protección de privacidad más fuerte. Con el valor de *t-closeness* significa que la distancia entre las distribuciones de los atributos o variables sensibles en cada equivalencia de clase y en la base de datos completa es menor o igual a 0.989, lo cual sugiere que la información del atributo sensible *tipo\_disc* está bien protegida, en vista de que las distribuciones locales dentro de cada equivalencia de clase son muy similares a la distribución global. El valor de 0.989 para *t-closeness* es bastante cercano al valor máximo 1, lo cual indica que las distribuciones de los atributos sensibles son exactamente los mismos entre la equivalencia de clase local y el conjunto de datos general, en otros términos, la base de datos del REBPE satisface el requisito de privacidad para la técnica *t-closeness*, suministrando un nivel de protección fuerte para el atributo sensible (*tipo\_disc*).

**Ilustración 1:** Resultados de anonimato utilizando pyCANON

Anonymity technique	Value(s)
k-anonymity	k = 1
( $\alpha$ ,k)-anonymity	$\alpha = 1.0$ and k = 1
l-diversity	l = 1
Entropy l-diversity	l = 1
(c,l)-diversity	c = nan and l = 1
Basic $\beta$ -likeness	$\beta = 3342.934049368758$
Enhanced $\beta$ -likeness	$\beta = 8.11490325193139$
t-closeness	t = 0.9894910709251118
$\delta$ -disclosure privacy	$\delta = 8.11490325193139$

Elaborado por: Dirección de Registros Administrativos



## Referencias

- Díaz J. y López Á. (2022) pyCANON: A Python library to check the level of anonymity of a dataset. Instituto de Física de Cantabria (IFCA). España.
- Ferreira J. and Bernardino J. (2020) Analysis of Data Anonymization Techniques. 12th International Conference on Knowledge Engineering and Ontology Development (KEOD), pág. 235-241.
- Li N. and Li T. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. International Conference on Data Engineering (ICDE), (2).
- Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, a. M. (2007). L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 24-24.
- Templet M. et al. (2015) Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software. Vol. 67, Issue 4.
- Technische Universität München (2013). Proceedings of the Seminars Future Internet (FI), Innovative Internet Technologies and Mobile Communications (IITM), and Autonomous Communication Networks (ACN). Network Architectures and Services NET. Munich, Germany.



@InecEcuador



@ecuadorencifras



@ecuadorencifras



INECEcuador