Universidad de Concepción
Facultad de Ingeniería
Departamento de Ingeniería Industrial

"Análisis Estadístico y Exploración de Modelos para Mejorar la Representatividad Nacional en Datos de Conveniencia en Línea"

POR

Cristóbal Alonso Delgado Escobar

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil Industrial

Profesores Guía

Sebastián Astroza Tagle

Carlos Navarrete Lizama

Agosto 2024
Concepción (Chile)

Resumen

Las investigaciones mediante encuestas muestran un incremento en las últimas décadas, utilizándose en diversas áreas de investigación para obtener de manera científica las opiniones de las personas. Con el avance de las tecnologías, especialmente el uso de internet, se realizan numerosas encuestas en línea debido a su rapidez y menor costo comparado con encuestas presenciales. Sin embargo, estas encuestas generalmente producen muestras no probabilísticas, lo que implica que pueden tener sesgos y esto limita la capacidad de realizar inferencias poblacionales con los datos. Aunque en muchos casos esto no representa un problema, si se desean hacer inferencias poblacionales, no se puede utilizar una muestra de este tipo.

Este estudio evalúa varias metodologías aplicables a una muestra no probabilística obtenida de internet, con el objetivo de mejorar la representatividad de la población en dicha muestra y permitir la realización de inferencias poblacionales. Se iguala la distribución de hombres y mujeres en tres grupos etarios utilizando tres metodologías diferentes: factores de expansión, método de sobremuestreo y un enfoque híbrido de sobremuestreo y submuestreo. Cada una de estas metodologías presenta variaciones en su funcionamiento, permitiendo evaluar si alguna de ellas iguala de mejor manera la distribución de la muestra con la de la población.

La muestra utilizada se obtiene de la página Mo2, que contiene datos sobre el uso de tiempo de la población chilena. Tras aplicar estas metodologías, se observa que los tres métodos entregan valores promedios similares de uso de tiempo, aunque en ciertas actividades algunos métodos muestran ligeras diferencias. Si bien todos los métodos logran su objetivo de igualar la distribución de la muestra no probabilística con la de la población, el método SmoteTomek (método híbrido) muestra un rendimiento deficiente al eliminar datos. Por ello, se recomienda sustituir este método por otro que funcione de manera diferente al ya aplicado.

Este estudio destaca la viabilidad de utilizar diversas metodologías para mejorar la representatividad de muestras no probabilísticas obtenidas en línea. Cualquiera de los métodos discutidos puede ser útil para este propósito, proporcionando información valiosa para futuras investigaciones y proyectos de encuestas.

Abstract

Survey-based research has shown an increase in recent decades, being utilized across various research areas to scientifically gather people's opinions. With the advancement of technologies, especially the use of the internet, numerous online surveys are conducted due to their speed and lower cost compared to in-person surveys. However, these surveys typically yield non-probabilistic samples, which may introduce biases and thereby limit the ability to make population inferences with the data. While this may not be problematic in many cases, it precludes the use of non-probabilistic samples when population inferences are desired.

This study evaluates several methodologies applicable to a non-probabilistic sample obtained from the internet, aiming to enhance the sample's representativeness of the population and enable population inferences. The distribution of males and females across three age groups are equalized using three different methodologies: weighting factors, oversampling method, and a hybrid approach of oversampling and undersampling. Each of these methodologies varies in its approach, allowing an assessment of which one best matches the sample distribution with that of the population.

The sample used is obtained from the Mo2 website, which contains data on time use among the Chilean population. After applying these methodologies, it is observed that all three methods provide similar average values for time use, although slight differences are noted in certain activities. While all methods achieve their goal of equalizing the distribution of the non-probability sample with that of the population, the SmoteTomek method (hybrid method) performs poorly in data elimination. Therefore, it is recommended to replace this method with another that operates differently from the one already applied.

This study underscores the feasibility of employing diverse methodologies to enhance the representativeness of non-probabilistic samples obtained online. Any of the discussed methods can serve this purpose effectively, providing valuable insights for future survey research and projects.

Tabla de contenidos

Lista de Tablas

Lista de Figuras

Capítulo 1: Introducción

En este capítulo se expone la motivación detrás del tema de este trabajo, así como el objetivo general y los objetivos específicos que se pretende alcanzar, junto con la información que se presenta en esta memoria de título.

1.1 Motivación

La aplicación de encuestas probabilísticas en poblaciones finitas se formaliza en Neyman (1934). El principal objetivo de estas encuestas es recopilar información pertinente a un grupo de personas. Existen varias formas de obtener este tipo de información, idealmente, se entrevista a todos los miembros de la población, lo cual en ciertos casos se hace, pero resulta inviable para poblaciones numerosas. Es por esto que la alternativa a la aplicación de encuestas a toda la población es aplicarlas a un subgrupo denominado muestra. Con el avance de las tecnologías, muchas de las aplicaciones de estas encuestas se trasladan a internet, ya que es más fácil y rápido que realizar entrevistas a mano. Esto ha llevado a que, en la mayoría de los casos, se obtengan muestras no probabilísticas. Este tipo de muestra no asegura que toda la población esté representada, lo cual en la mayoría de los casos no es un problema, pero cuando se desea inferir respecto a la población, no se puede utilizar una muestra no probabilística.

En esta memoria de título, se utiliza una muestra no probabilística obtenida de una página web llamada Mo2, desarrollada por académicos de la Universidad de Concepción para el proyecto Fondecyt N°1221724. Esta plataforma recopila datos sobre los usos de tiempo de la población chilena y proporciona una muestra por conveniencia (Otzen & Manterola, 2017), ya que los participantes no son escogidos aleatoriamente. Así, solo aquellos que visitan la página pueden responder, por tanto, no existe certeza de que cada sujeto de la muestra represente a la población en general (Walpole, 2012). Al obtener esta muestra no probabilística, se aplican tres métodos diferentes para mejorar la representatividad de la población dentro de la muestra, con el fin de realizar inferencias poblaciones. Algunos de estos métodos requieren información adicional a la que entrega la página, como por ejemplo, información demográfica del país. Es por esto que se buscan bases de datos de acceso público, como el censo de población, que contengan la información requerida para aplicar las distintas metodologías encontradas.

1.2 Objetivos

1.2.1 Objetivo General

Aplicar técnicas estadísticas para obtener una muestra representativa de datos recolectados a través de una página web destinada a caracterizar el uso de tiempo.

1.2.2 Objetivos específicos

1) Realizar una revisión bibliográfica de los métodos existentes para corregir la representatividad de una muestra.
2) Aplicar proceso de extracción, transformación y carga de los datos entregados por la página web.
3) Analizar y describir datos mediante el uso de bibliotecas de Python y/o R.
4) Aplicar las metodologías investigadas para corregir la representatividad.
5) Desarrollar un script generalizado que aplique de manera automática técnicas estadísticas a muestras futuras obtenidas de la página web.

1.3 Estructura del informe

El resto del informe se organiza de la siguiente manera: En el capítulo 2, se presenta una revisión bibliográfica sobre los temas abordados en esta memoria de título, que incluye descripciones de diversos conceptos utilizados para alcanzar las conclusiones pertinentes. En el capítulo 3, se describen tres metodologías que se aplican en este informe, explicando su funcionamiento y componentes. El capítulo 4 detalla los resultados obtenidos tras aplicar estas metodologías y ofrece un análisis general de todos los resultados. Por último, se presentan las conclusiones que se obtienen del trabajo realizado.

Capítulo 2: Antecedentes bibliográficos

En este capítulo se revisa la literatura relevante para comprender el contexto del problema y cómo las soluciones propuestas pueden contribuir a resolverlo.

2.1 Introducción

La investigación por encuestas consiste en establecer reglas que permitan acceder de manera científica a lo que las personas opinan (Leon & Morgado, 1993). Esto se logra realizando diversos tipos de preguntas a un conjunto de individuos, de los que se presume que son representativos de su grupo de referencia, con el fin de conocer sus actitudes respecto al tema objetivo del estudio.

Con el paso de los años, el uso de encuestas en los ámbitos de investigación social y de mercado han ido al alza (Blom et al., 2016; Hays et al., 2015). Idealmente, se espera realizar encuestas a toda la población objetivo del estudio. Esto se puede lograr si la población es pequeña, pero para el caso de poblaciones medianas o grandes, esto se hace inviable principalmente por temas de tiempo y dinero. Es por esto que, generalmente, se entrevista a un subgrupo que sea representativo de la población lo que se denomina muestra. Es importante recalcar que el muestreo a un subgrupo de la población no garantiza ser representativo de población. Para lograr una muestra representativa hay que considerar la forma en la que esta es conseguida, ya que dependiendo del proceso que se lleve a cabo la muestra puede ser de dos tipos: probabilística y no probabilística. Un ejemplo de muestreo probabilístico en estudios de uso de tiempo es la Encuesta Nacional de Uso de Tiempo (ENUT) (Documento Metodológico ENUT 2015, 2015.). Esta encuesta empela un enfoque bietápico y estratificado, considerando tanto la geografía como el tamaño de las viviendas en áreas urbanas, en contraste, la mayoría de las encuestas realizadas en internet se basan en muestras no probabilísticas. A continuación, se detalla con más énfasis los dos tipos de muestreos descritos en esta parte.

2.2 Muestreo probabilístico

Una muestra es probabilística en la medida en que todos los miembros de la población tienen una probabilidad conocida de ser seleccionados distinta de cero (Levy & Lemeshow, 2013). Este tipo de muestreo justifica su exactitud gracias a la teoría del muestreo probabilístico, la cual se basa en un

conjunto de principios matemáticos (Fisher, 1925; Kish, 1965; Neyman, 1934). Gracias a esta sólida base teórica, es posible calcular la precisión de las estimaciones (a través de intervalos de confianza, márgenes de error, entre otros) y otorgar validez universal al método de estimación.

Además, en este tipo de muestreo, los investigadores suelen detallar el proceso mediante el cual se obtiene los datos utilizados, permitiendo a otros investigadores aplicar los ajustes necesarios relacionados con variables como cobertura, sesgo de muestreo y falta de respuesta, para así poder utilizar los datos de esa muestra.

Varios estudios han evaluado la precisión de los muestreos probabilísticos y no probabilísticos, esta evaluación se realiza comparando qué tan parecidas son las muestras obtenidas por estos métodos con la población de estudio. La gran mayoría de los estudios concluye que los muestreos probabilísticos presentan una mayor precisión (Cornesse et al., 2020), lo cual demuestra que, al momento de realizar algún tipo de muestreo, lo ideal es optar por un muestreo probabilístico por sobre uno no probabilístico. Este tipo de muestreo, gracias a su selección aleatoria, hace que la muestra obtenida no esté afectada por juicios subjetivos en la elección de los elementos de la muestra, algunos tipos de muestreos probabilísticos son aleatorio simple, aleatorio sistemático y estratificado.

- Muestreo aleatorio simple: este tipo de muestreo es uno de los más utilizados si la población objetivo es pequeña. Esto se debe a que, para poder llevar a cabo esta técnica de muestreo, se debe asignar un número a todos los elementos de la población. Posteriormente, seleccionar a través de un medio mecánico qué elementos de la población serán parte de la muestra. Esto es posible en poblaciones pequeñas y medianas, siempre y cuando se cuente con una lista de todos los elementos de la población. En el caso de poblaciones grandes, esta tarea es más compleja, ya que se deben tomar otras medidas.

- Muestreo aleatorio sistemático: este tipo de muestreo es parecido al muestreo aleatorio simple. Sin embargo, en este caso, además de enumerar a todos los elementos de la población, se escogen los $i, i+k\ i+2k \ldots (n-1)k$ elementos de la lista, k siendo calculado según la ecuación (1).

$$k = \frac{N}{n} \quad (1)$$

N = tamaño de la población.

n = tamaño de la muestra.

- Muestreo estratificado: este tipo de muestreo divide a la población en subgrupos o estratos, en los cuales cada elemento de la población solo pertenece a un estrato. Estos estratos se originan en función de ciertas características, por ejemplo, puede haber estratos según el sexo, la edad, la profesión, entre otros. Tras obtener todos los estratos, el objetivo es que la muestra tenga elementos en cada uno de ellos. La cantidad de elementos por estrato puede variar según el investigador, pero en general puede haber el mismo número de elementos en cada estrato, puede ajustarse a la distribución de la población o puede ser una mezcla de ambos métodos.

2.3 Muestreo no probabilístico

A diferencia del muestreo probabilístico, el no probabilístico selecciona a los integrantes de las muestras de manera subjetiva en lugar de hacerlo al azar, esto genera dos tipos de sesgo: el sesgo de cobertura, que se produce cuando una parte de los elementos de la población no tienen posibilidad de ser elegidos para participar en el estudio (Weisberg, 2005) y el sesgo de autoselección, que se refiere a la probabilidad diferencial que tienen los elementos poblacionales de sumarse voluntariamente al estudio (Bethlehem & Biffignandi, 2011; Blom et al., 2015). A pesar de los sesgos que genera este tipo de muestreo, a finales de siglo XX y gracias al auge de internet, su popularidad aumentó debido a que era un método rápido y económico para paneles de reclutamiento en línea (Dillman & Bowker, 1999). Por esta razón, gran parte de los datos recopilados en encuestas en línea en todo el mundo hoy en día se basan en muestras no probabilísticas (Callegaro et al., 2014).

Este tipo de muestreo, debido a los sesgos mencionados, no ofrece garantías de que la muestra sea representativa de la población, lo que implica que los resultados obtenidos no son generalizables a la población objetivo. Una generalización con este tipo de muestra solo es viable hacia poblaciones hipotéticas que compartan una distribución similar a la observada en la muestra. Por lo tanto, para evaluar si una muestra no probabilística es buena o no para hacer inferencias, es necesario evaluar qué

tan cercana es la muestra final obtenida a la población en términos de diversas distribuciones, como sexo, rango etario, ubicación, entre otros.

Hasta la fecha, poco se ha propuesto en términos de teoría estadística formal que justifique el uso de muestras no probabilísticas para hacer inferencias poblaciones (Cornesse et al., 2020). No obstante, mientras exista algún grado de justificación respecto a la muestra, procedimientos de ajuste y otros factores, el uso de muestras no probabilística podría ser viable para hacer inferencias poblacionales. Uno de los principales métodos de muestreos no probabilísticos utilizados es el muestro por cuotas. Aunque es un muestreo no probabilístico, este se diseña para garantizar que el conjunto de los encuestados coincida con los parámetros demográficos de la población. Al hacer esto, se puede decir que la muestra obtenida es igual a la distribución de la población objetivo, minimizando así las diferencias sustanciales respecto a una muestra probabilística y permitiendo la realización de inferencias poblaciones sobre los datos obtenidos, gracias a que el diseño de este método neutraliza los sesgo que genera una muestra no probabilística.

Esta idea puede replicarse a todas las muestras no probabilísticas. Si la muestra obtenida no sigue la distribución demográfica de la población, se puede aplicar algún método de ajuste que iguale la distribución de la muestra con la de la población y así facilitar la realización de inferencias poblacionales. Además del método por cuotas mencionado, existen varios otros métodos para obtener muestras no probabilísticas, algunos de estos métodos son muestreo por conveniencia, por juicio y de bola de nieve, entre otros.

- Muestro por conveniencia: este tipo de muestreo consiste en seleccionar elementos convenientes para la investigación. Esta técnica es una de la más económicas y la que menos tiempo toma, ya que el método utilizado para seleccionar a los individuos depende del criterio del investigador. Además, esta técnica prioriza la proximidad de los sujetos a la investigación por sobre si estos sujetos hacen la muestra representativa o no.

- Muestreo por juicio: este tipo de muestreo selecciona sujetos en base al conocimiento y juicio del investigador, es decir, es el encargado de decidir qué sujetos pertenecerán a la muestra, ya que el investigador cree que son representativos de la población en el estudio. Este método es

recomendable de usar cuando la población es pequeña y el responsable conoce estudios anteriores similares que se hayan hecho y en los que la muestra haya sido útil para el estudio.

- Muestreo de bola de nieve: este método de muestreo, tras haber seleccionado a un individuo para la muestra, hace que este deba seleccionar a otro individuo para que sea parte de la muestra, y así sucesivamente, lo cual genera un efecto acumulativo parecido a una bola de nieve (de ahí su nombre). Esta técnica se utiliza en poblaciones en las que no se conoce a los individuos o no se pueden acceder a ellos.

2.4 Representatividad

Para que el muestro probabilístico y no probabilístico tengan validación externa al momento de hacer inferencias poblacionales con los resultados obtenidos, es esencial que sus muestras sean representativas de la población objetivo. La forma en que se seleccionan los sujetos y el tamaño de la muestra son factores fundamentales para determinar en qué medida esa muestra es representativa de la población a la cual refiere. Si la muestra es probabilística, su naturaleza de selección aleatoria debería garantizar su representatividad y ausencia de sesgo, pero para el caso de la muestra no probabilística, esta no es representativa de la población.

El principio de representatividad afirma que cada elemento incluido en una muestra se representa a sí mismo y a un grupo de elementos que no pertenecen a la muestra seleccionada, cuyas características son cercanas a las del elemento incluido en la muestra (Rojas, 2016). Para entender mejor este concepto y su importancia, si se asume que la población objetivo tiene como característica única el género y en la muestra se cuenta con un hombre, este hombre, según el principio de representatividad, representa adecuadamente a otros hombres que no están en la muestra, pero no representa adecuadamente a mujeres que no están en la muestra, ya que cuentan con características relacionadas al género distintas. Si se cuenta con más de una característica, el individuo seleccionado en la muestra representará solo a los otros individuos que comparten estas características (o que tienen característica parecidas, por ejemplo la edad) pero no están seleccionados en la muestra.

El tamaño muestral es un factor relevante para que una muestra sea representativa de la población. Dependiendo de la cantidad de sujetos que componen la población objetivo, varía la cantidad de elementos necesarios en la muestra, para que toda la población esté representada por alguien que se encuentre en la muestra. El tamaño de la muestra cuenta con dos fórmulas, que se aplican para obtener el tamaño muestral ideal para que esta sea representativa. La primera es la fórmula para calcular el tamaño muestral para poblaciones infinitas, la cual se ve afectada por varios elementos, que se pueden apreciar según la fórmula de la ecuación (2).

$$n_{\infty} = \frac{z^2 \text{ x } p \text{ x } q}{e^2} \tag{2}$$

-n_{∞}= tamaño de la muestra buscada.
-*z*= parámetro estadístico que depende del nivel de confianza.
-*p*= probabilidad de que ocurra el evento estudiado.
-*q*= probabilidad de que no ocurra el evento estudiado.
-*e*= error de estimación máximo aceptado.

Tras obtener el valor de la muestra para poblaciones infinitas, hay que realizar un ajuste para estimar el tamaño muestral para poblaciones finitas, este ajuste es calculado según la ecuación (3).

$$n = \frac{n_{\infty}}{1+\frac{n_{\infty}}{N}} \tag{3}$$

-*n*= tamaño de la muestra ajustado.
-*N*= tamaño de la población objetivo.

El valor que se obtiene de esta segunda formula es el tamaño mínimo de la muestra con el cual se puede tener cierto grado de seguridad de que la población se verá representada en la muestra. La fórmula para las poblaciones infinitas incluye varias variables que cuyo valor depende de los objetivos del investigador. En el caso de la variable *z*, esta representa el nivel de confianza que se le asigna al estudio, se expresa en porcentajes y hace referencia a la probabilidad de que una muestra obtenga resultados que coincidan con los de la población. Por lo general, los investigadores utilizan un nivel de confianza del 95% (para aplicar este valor en la fórmula se debe buscar el valor correspondiente en la distribución normal). Las variables *p* y *q* son los equivalentes a la probabilidad de que ocurra o no cierto evento. Cuando no se cuenta con información previa que sirva como guía para determinar

las probabilidades, se suele usar 0,5 tanto para *p* como para *q* (la suma de *p* y *q* debe dar 1). La variable *e* representa el margen de error aceptable para el investigador al realizar el experimento. Este margen de error está relacionado con la diferencia entre la media de la muestra y la media de la población a estudiar. En términos generales, los investigadores comúnmente usan un margen de error de un 5% aunque esto depende del estudio y del investigador.

Tras definir todas las variables mencionadas y teniendo claro cuál es el tamaño de la población objetivo, se puede calcular cuál es el tamaño de la muestra mínimo para que la población completa esté representada.

2.5 Encuestas de uso de tiempo

Las encuestas de uso de tiempo son herramientas metodológicas de recolección de datos sobre las actividades que realizan las personas en un periodo determinado. Este tipo de encuestas comienzan a realizarse a principios del siglo XX (Jara-Díaz & Rosales-Salas, 2015) y buscan proporcionar información sobre cómo las poblaciones asignan su tiempo a actividades como el trabajo remunerado, no remunerado y actividades personales. Existen varios periodos de observación que se pueden aplicar a las encuestas de uso de tiempo, estos tiempos de observación varían desde uno, tres, cinco días, semanas, meses, entre otros. Dependiendo del objetivo de la investigación, cierto periodo de observación será mejor que el otro.

En América Latina, las encuestas sobre el uso del tiempo son fundamentales en el debate sobre el reconocimiento y la redistribución del trabajo no remunerado y permiten orientar la formulación de políticas públicas que atienden las necesidades sociales del cuidado mediante la corresponsabilidad social, trasladando responsabilidades del ámbito familiar al público y al privado (Aguirre & Ferrari, 2014). En el año 2015, se celebra la Octava Reunión de la Conferencia Estadística de las Américas de la Comisión Económica para América Latina y el Caribe. En esta reunión, todos los países miembros adoptaron la Clasificación de Actividades de Uso del Tiempo para América Latina y el Caribe (CAUTAL) como clasificación de actividades de uso de tiempo, con un enfoque de género y adecuada al contexto regional (Comisión Económica para América Latina y el Caribe [CEPAL], 2016). Este hito permite a estos países avanzar hacia la armonización y normalización de las encuestas de uso de

tiempo, similar a como lo hacen lo países de europeos con la Harmonized European Time Use Surveys (HETUS).

En Chile se han realizado tres encuestas para caracterizar el tiempo que destina la población a la realización de diversas actividades. La primera vez fue durante el 2007 con la "Encuesta experimental de Uso del Tiempo en el Gran Santiago" (EUT). Luego el 2015, se lleva a cabo la primera "Encuesta nacional de Uso de Tiempo" (ENUT) y a fines del 2023, se realiza la segunda edición de la ENUT, cuyos resultados aún no se han publicado al momento de la realización de esta memoria de título (Aguirre & Ferrari, 2014). La forma en que se lleva a cabo esta encuesta ha sido consistente a lo largo de los años y similar a otras, contratando y capacitando personal para que recorra diversas manzanas ubicadas en todo el país, realizando esta encuesta en forma presencial. Para la ENUT 2023, la recolección de datos toma un aproximado de 4 meses y tiene un costo de 1.850 millones de pesos aproximadamente para el año 2023 (aún falta el procesamiento de los datos y gastos durante el año 2024) (instituto nacional de estadistica [INE], 2024). Esto demuestra que la aplicación de estos instrumentos requiere de una gran cantidad de tiempo y recursos para poder llevarlas a cabo.

Estas encuestas se realizan en promedio cada ocho años, debido a los costos económicos y a la presunción de que, en el corto plazo no hay cambios significativos en el comportamiento de la población, por lo que no es necesario reducir el periodo entre encuestas. Hay estudios que llevan a cabo encuestas piloto para la recolección de datos de uso en tiempo en línea (Chatzitheochari et al., 2018), pero solo se quedan en eso, estudios piloto, sin avanzar a la implementación de la recolección de datos de uso de tiempo en línea. Esto puede deberse a que, como indican otros estudios (Bethlehem, 2008), las encuestas en línea tienden a sobrerrepresentar o subrepresentar a ciertos grupos de la población. Sin embargo, si se realizan ajustes a la muestra para evitar la sobrerrepresentación o subrepresentación de grupos, entonces una muestra no probabilística podría ser útil.

La realización de este proyecto, es para encontrar métodos que ayuden a ajustar la representatividad de muestras no probabilísticas. En este caso puntual, sería útil para obtener información sobre cómo ha variado el uso de tiempo de la población en un menor periodo y a un costo mucho más reducido. Aunque esta encuesta no tiene como objetivo reemplazar a la ENUT, podría servir como un antecedente valioso para tener una idea sobre los resultados que debería obtener la ENUT.

Capítulo 3: Metodología

En este capítulo se presentan y se definen los tres métodos que se utilizan para mejorar la representatividad de una muestra. Además, se describe el procedimiento de los códigos correspondientes a estos métodos y se proporciona una descripción de la base de datos que se utiliza.

3.1 Introducción

Con el avance de la tecnología, las encuestas realizadas por internet se vuelven más populares, lo cual se refleja en el proyecto Fondecyt N°1221724, sobre el cual se está haciendo esta memoria de título. En este proyecto, el equipo de investigación desarrolla una aplicación web, llamada mo2 (https://mo2.cl), que pregunta a los usuarios por su diario de actividades durante las últimas 24 horas y entrega como recompensa el tipo de uso de tiempo que tiene cada persona, caracterizado con distintos animales.

Cada tipo de uso de tiempo cuenta con su respectiva definición, la cual va asociada con su nombre y animal. Algunos de los ejemplos de tipos de uso de tiempo son Dinociosos, Axolodados, Trabajosos, entre otros. Estos nombres, descripciones e imágenes son creadas por los investigadores de este proyecto.

La muestra generada por esta página web es una muestra no probabilística, específicamente del tipo por conveniencia, ya que solo las personas que conocen la página y deciden participar pueden formar parte de la muestra. Esta muestra, tal y como esta, no sirve para realizar inferencias sobre los resultados y hacerlos representativos de la población chilena, por lo que a estos datos se les debe aplicar diversos métodos que logren replicar la distribución de la población chilena. Para lograr esto, se utiliza como información auxiliar el censo de 2017, ya que hasta la fecha de la realización de esta memoria es el censo más reciente disponible.

3.1.1 Censo

El censo de población y vivienda es el recuento de todas las personas y viviendas que se encuentran en el territorio nacional en un momento determinado. Esta operación estadística es una de las más importantes que se lleva a cabo en el Instituto Nacional de Estadísticas. Esta encuesta participan todos los habitantes del país (Instituto Nacional de Estadistica [INE], 2017) y es la principal fuente de información que se utiliza para diversas políticas públicas y toma de decisiones. El último censo, realizado en el año 2017, indica que la población de Chile corresponde a 17.574.003 personas, de las cuales el 48,9% son hombres (8.601.989) y el 51.1% mujeres (8.972.533).

Gracias a este instrumento se pueden realizar diversas observaciones sobre características específicas de la población. Un ejemplo, en el último censo, se observa que la población chilena está envejeciendo aceleradamente, ya que el grupo eterio de personas de 65 años o más representa el 11,4% de la población, lo cual marca un aumento considerable desde el censo anterior.

Este instrumento se utiliza como base cuando se llevan a cabo encuestas que quieren inferir sus resultados a nivel nacional. Por lo general, se utilizan variables como sexo y rango etario. Cuando se realizan muestras probabilistas, se espera que la muestra se distribuya de manera similar a los datos del censo, reflejando la composición demográfica de la población objetivo. Para el caso de las muestras no probabilísticas, el censo se usa para aplicar diversos métodos estadísticos que permitan replicar la distribución de la población en la muestra obtenida.

3.2 Datos

La página web Mo2 es la encargada de recolectar datos de uso de tiempo de la población chilena. Además, de proporcionar la base de datos que se usa para aplicar los métodos que mejoran la representatividad de una muestra en esta memoria de título, se lanza al público el 8 de mayo del 2024. Se promueve la página a través de redes sociales y medios tradicionales como radios, programas de televisión, entre otros. Esto se realiza con el objetivo de obtener la mayor cantidad de respuestas posibles, es por esto que la muestra utilizada para aplicar los métodos mencionados en este capítulo, es la obtenida hasta el 17 de junio de 2024, los datos recopilados a partir de esa fecha no se consideran en esta memoria de título.

Tras casi un mes y diez días desde el lanzamiento de la página, se reciben un total de 477 respuestas. A esta base de datos se le realizó una limpieza preliminar, resultando en 388 respuestas válidas, las cuales se distribuyen de acuerdo a la figura 3.1 y 3.2. Como se puede apreciar de las figuras, las respuestas recolectadas por la página web provienen principalmente de personas entre 18 a 30 años, este grupo representa al 64% de la muestra, lo que es equivalente a 247 respuestas. En términos de género, la mayoría de las respuestas provienen de hombres, siendo el 52% de estas, es decir, 200 respuestas. Aquí se puede apreciar la principal problemática de obtener muestras no probabilísticas: si alguien quisiera usar los datos tal y como están, no podría hacer inferencias poblacionales representativos de la población chilena.

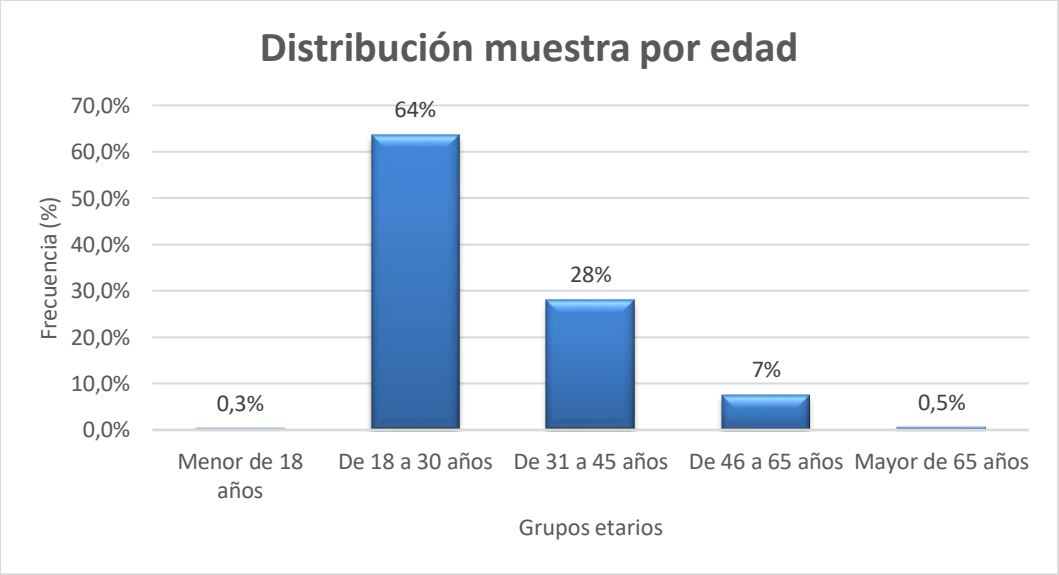


Figura 3.1: Distribución de la muestra por rango etario

Fuente: elaboración propia

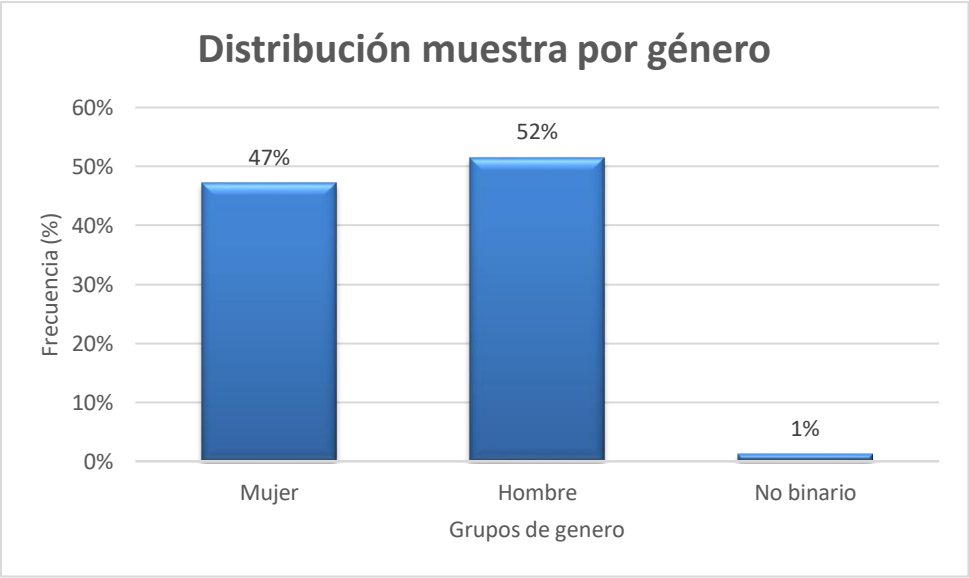


Figura 3.2: Distribución de la muestra por género

Fuente: elaboración propia

Por ello, el objetivo de esta memoria de título es transformar esta muestra no probabilística de tal forma que quede los más parecido posible a la distribución de la población chilena, para así poder hacer inferencias poblacionales. Antes de aplicar los métodos que se mencionan en este capítulo, es necesario realizar algunos ajustes a la muestra. En el ámbito del género, debido a que el censo del 2017 no cuenta con información de personas no binarias, estas son excluidas de la muestra. Para el rango de edad, aunque el censo no entrega de manera explícita la información sobre los grupos etarios utilizados en esta muestra, es posible calcular la distribución de estos grupos. Se excluirá a los menores de 18 años y los mayores de 65 años, ya que no hay respuestas suficientes en estos grupos para aplicar algunos métodos. Tras realizar estos ajustes, la distribución de la muestra se compara con la de la población y se presenta en las Figuras 3.3 y 3.4.

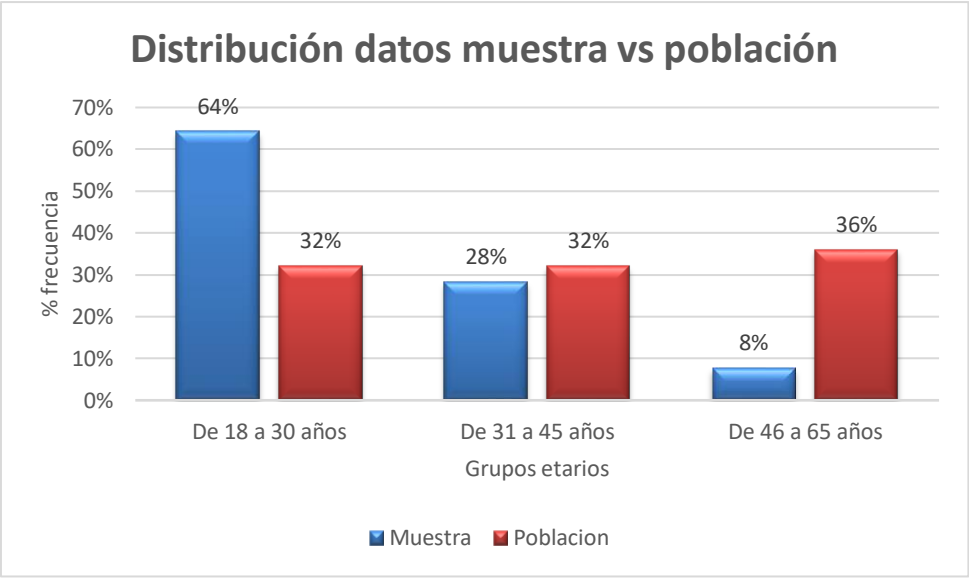


Figura 3.3: Distribución de la muestra por rango etario comparada con la de la población

Fuente: elaboración propia

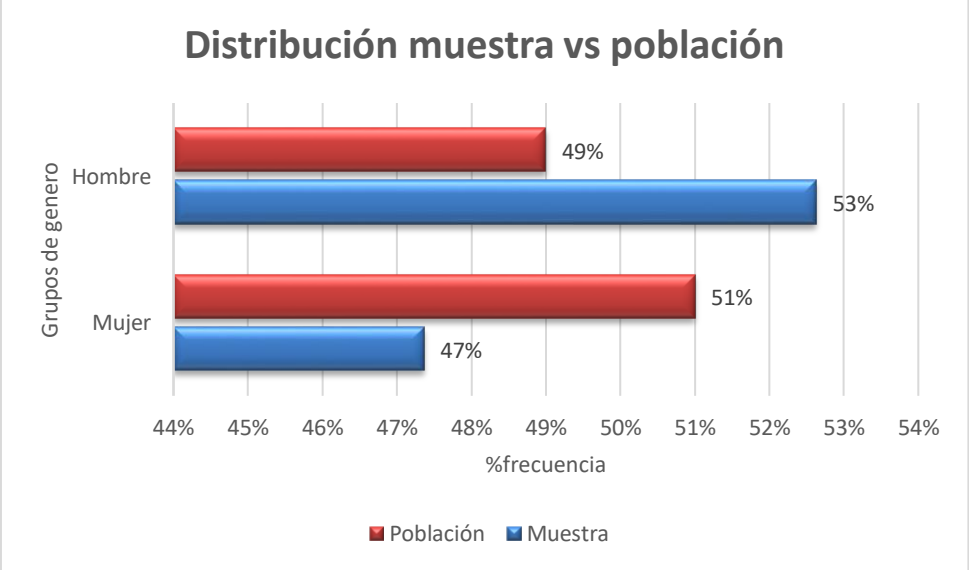


Figura 3.4: Distribución de la muestra por género comparada con la de la población

Fuente: elaboración propia

Tras hacer este último ajuste, la muestra se reduce de 388 a 380 respuestas. Esta cantidad de respuestas es cercana al tamaño muestral mínimo recomendado de 385 respuestas para poblaciones grandes. A pesar de no ser iguales, se podría esperar que gran parte de la población (si no toda) se encuentre

representada dentro de la muestra. Para el caso de los rangos etarios, el grupo de personas de 18 a 30 años se encuentra sobrerrepresentado en la muestra, representando al 64% de las respuestas, mientras que en la población objetivo representan el 32%, lo que implica una diferencia significativa del 32%. Por otro lado, las personas de 31 a 45 y 46 a 65 años están subrepresentados por un 4% y 28% respectivamente. Desde el punto de vista de género, se aprecia que los hombres están sobrerrepresentados en la muestra y las mujeres subrepresentadas por un 4%.

Aunque esta muestra tal y como esta, no sirve para hacer inferencias de la población chilena, no deja de ser interesante conocer cómo distribuyen su tiempo las personas que respondieron la encuesta. Además, contar con esta información será crucial para evaluar si los métodos descritos en este capítulo realmente tienen impacto en cambiar la distribución de la población, ya que si los métodos funcionan adecuadamente, los tiempos promedios de uso de tiempo deberían verse afectados. La información que se presenta sirve como base para corroborar si hubo realmente cambios en la distribución de la muestra o no. Los tiempos promedio de la muestra, tal y como se obtienen de la página web, son presentadas en la tabla 3.1.

Tabla 3.1: Participación, tiempos promedio y distribución del día de la muestra

Actividades	Participación (%)	Tiempo Promedio (en Horas)	Distribución del día
Trabajo Remunerado	264 (69%)	312 (5,2)	22,6%
Trabajo Domestico	280 (74%)	68 (1,1)	4,9%
Trabajo de cuidado	93 (24%)	16 (0,3)	1,1%
Voluntariado	29 (8%)	6 (0,1)	0,4%
Educación	173 (46%)	119 (2,0)	8,6%
Ocio	353 (93%)	230 (3,8)	16,6%
Cuidado Personal	335 (88%)	111 (1,8)	8,0%
Dormir	380 (100%)	468 (7,8)	33,9%
Desplazamiento	280 (74%)	53 (0,9)	3,8%
Total Muestra		1380 (23,0)	100,0%

Fuente: elaboración propia

De la muestra obtenida por la página web, se observa que las actividades con mayor participación son dormir, ocio, cuidado personal y desplazamiento. Estas actividades tienen una participación mínima del 74% de todas las personas que responden la encuesta. A pesar de su alta participación, estas actividades no representan gran parte del uso del tiempo en el día, a excepción de dormir que es la actividad con mayor uso de tiempo promedio con 7,8 horas durante el día. Es importante destacar que, como se aprecia en la Tabla 3.1, el tiempo promedio descrito por el total de la muestra es de aproximadamente 23 horas. Se esperaría que la muestra describiera las 24 horas, pero debido al diseño de la página web, esta no requiere que se describan las 24 horas del día, sino que pide un mínimo de 20 horas, es decir, aproximadamente un 83% del día. Por lo que la mayoría de las personas describen solo 23 horas de su día.

La segunda actividad con mayor uso del tiempo es el trabajo remunerado, en promedio tiene un uso de tiempo de 5,2 horas. Aunque a primera vista este dato puede parecer bajo, esto se debe principalmente a que más de la mitad de las personas que responden esta encuesta son personas que rondan los 18 a 30 años, por lo que la mayoría aún no están trabajando o están recién empezando a trabajar. Por eso, el uso de tiempo en esta actividad es bajo y no debería representar el trabajo de la población chilena. Por último, la tercera actividad a la que se le dedica mayor cantidad del tiempo es el ocio, con un uso de tiempo promedio de 3,8 horas durante el día. Este dato podría parecer alto teniendo en cuenta que la mayor parte de la semana la población le dedica gran parte de su día a trabajar o estudiar más que a otra cosa. Pero, al igual que para el caso del tiempo que se le dedica al trabajo remunerado, esta muestra, al tener el grupo etaria de 18 a 30 años sobrerrepresentado, podría ser la causa de que el tiempo dedicado al ocio sea más grande de lo esperado. Así que, tras aplicar los métodos de ajuste de representatividad que se mencionan más adelante, se espera que los valores de algunas categorías aumenten y otros disminuyan. A continuación, se muestra más información obtenida de la muestra en la Figura 3.5 y 3.6.

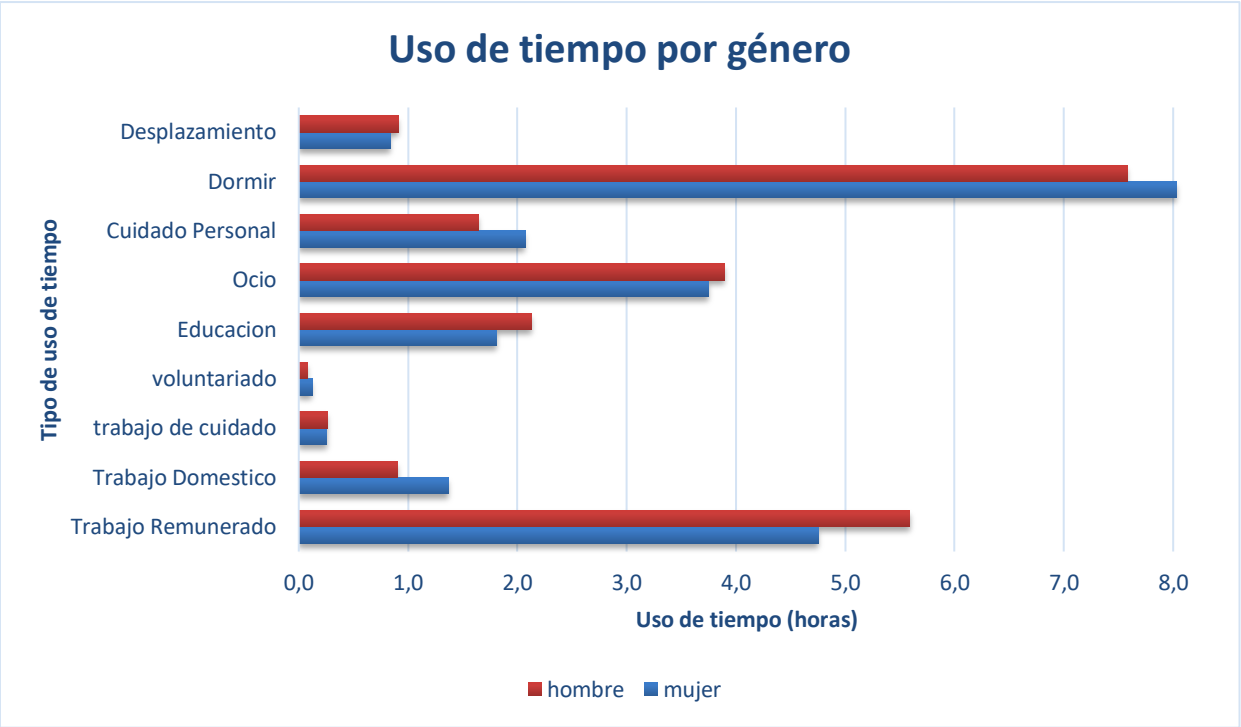


Figura 3.5: Uso de tiempo promedio por género

Fuente: elaboración propia

Desde un punto de vista de género en la figura 3.5, se observa que varias actividades tienen usos de tiempos parecidos. A simple vista, las actividades con mayor diferencia son el trabajo remunerado, donde los hombres dedican aproximadamente 50 minutos más que las mujeres. La segunda actividad con mayor diferencia de uso de tiempo es el trabajo doméstico, en la cual las mujeres le dedican aproximadamente 30 minutos más durante su día que los hombres. Por último, la actividad de dormir, las mujeres dicen dormir aproximadamente 30 minutos más que los hombres.

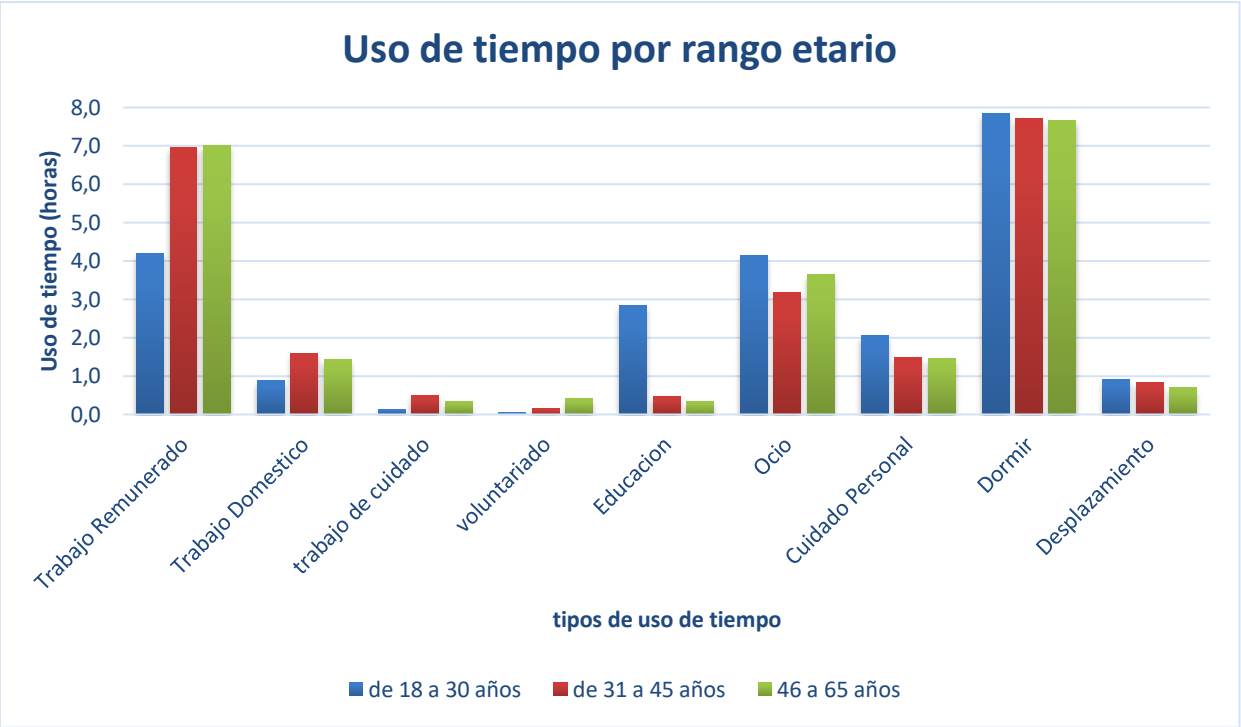


Figura 3.6: Uso de tiempo promedio por rango etario

Fuente: elaboración propia

Desde un punto de vista de rangos etarios en la figura 3.6, las actividades con mayores diferencias son el tiempo dedicado a trabajo remunerado, donde el grupo de 18 a 30 años dedica aproximadamente 3 horas menos que los grupos etarios de 31 a 65 años. Mientras, para el caso de educación, ocurre lo inverso, el grupo de 18 a 30 años le dedica 2,4 horas más que los otros grupos, finalmente, el resto de las actividades tienen usos de tiempo parecidos.

3.3 Factor de Expansión

Los factores de expansión son la medida que tiene cada individuo seleccionado en una muestra para representar el universo del cual forma parte. Esta herramienta es aplicada generalmente en muestras probabilísticas para extrapolar los resultados de una muestra a los de la población de estudio. Para obtener estos factores de expansión, se requiere contar con un instrumento estadístico confiable como el censo, el cual provea de información precisa sobre las características demográficas principales del universo que se desea estudiar.

Para esta memoria de título, que busca hacer inferencias sobre una muestra no probabilística potencialmente sesgada, se propone utilizar el censo para generar los factores de expansión correspondientes. Esto permite reducir al máximo el sesgo con el que cuenta la muestra obtenida originalmente, la fórmula que se utilizara para calcular los factores de expansión según la ecuación (4).

$$w = \frac{N}{n} \tag{4}$$

N = tamaño total de elementos del universo estudiado.

n = número de elementos seleccionados dentro de la unidad de muestreo.

El valor del factor de expansión indica que cada persona de la muestra equivale a w personas de la población, pero este valor no es igual para todos, ya que las proporciones de hombres y mujeres, así como las de las edades, son diferentes en el país. Por tanto, el factor de expansión debe ajustarse según otras características de las personas encuestadas. Se aplica un ponderador al factor de expansión, el cual depende de la distribución demográfica nacional para ajustar el valor de w según las características individuales de cada encuestado, la fórmula del ponderador está dada por la ecuación (5).

$$ponderador\ de\ calibracion = n\ x\ \frac{P_i}{f_i} \tag{5}$$

n= número de respuestas dentro de la muestra.

P_i= proporción de elementos que cumplen con *i* características en la población.

f_i= número de respuestas dentro de la muestra que cumplen con *i* características.

El valor P_i se obtiene utilizando información del censo, que proporciona datos desglosados por variables como sexo, rango de edad, entre otros. En caso de que el censo no proporcione información explícita, se utilizan otros datos disponibles para obtener la información requerida. Una vez obtenidos tanto el factor de expansión como el ponderador, el expansor final es el factor de expansión multiplicado por el ponderador. Este factor se aplica a cada uno de los individuos de la muestra y su valor va variando según las características específicas de cada persona.

3.4 Método Adasyn

Muchas bases de datos enfrentan problemas de desbalanceo de clases, es decir, una clase concentra la mayor cantidad de datos (mayoritarios) respecto a la otra (minoritarios). Con el paso de los años y para diversas aplicaciones, se han creado diversos métodos que realizan un remuestreo de las muestras originales. El remuestreo puede generarse de dos maneras: el sobremuestreo (oversampling) y el submuestreo (undersampling) (Rui, 2019). El segundo método que se plantea aplicar para mejorar la representatividad de la base de datos es el método de sobremuestreo Adaptive Synthetic Sampling o Adasyn (Rui, 2019). La característica general de los métodos de oversampling es que estos generan datos sintéticos a partir de la muestra original para equilibrar la cantidad de datos de todas las clases. El método Adasyn realiza dos pasos para generar nuevos datos. Primero, define el grupo de datos mayoritarios de la base de datos. Luego, para cada elemento de la clase minoritaria, busca los *k* vecinos minoritarios (*k* es el número que se puede elegir y representa la cantidad de vecinos que se va a buscar), cuando se encuentran los *k* vecinos, se aplica la fórmula de la ecuación (6).

$$w_i = \frac{R_i}{V} \tag{6}$$

R_i= número de vecinos de la clase mayoritaria punto minoritario i.
V= *k* vecinos.

El valor obtenido de w_i representa el peso asignado a cada valor de la clase minoritaria, variando de 0 a 1. Es 0 cuando el valor minoritario está rodeado solo por otros valores minoritarios y 1 cuando todos los vecinos del valor minoritario son de la clase mayoritaria. Este proceso aplica para todos los valores minoritarios. Después de asignar un peso a todos los valores minoritarios, se da comienzo al segundo paso. Este consiste en buscar los valores minoritarios con w_i>0, se buscan estos valores minoritarios y se crean *n* datos (siendo *n* un número directamente proporcional con el valor de w_i, mientras más grande sea el peso mayor número de datos se crean) entre el valor y su vecino minoritarios. Así, este proceso se lleva a cabo para todos los datos minoritarios con pesos mayores a 0, hasta que se iguale o se cumpla la proporción dada de datos necesarios por generar.

Adasyn utiliza la interpolación para crear nuevos puntos cerca de los datos mayoritarios, evitando así la generación de datos en áreas con una gran cantidad de datos minoritarios. Esto aumenta la variabilidad dentro de la base de datos y mejora la representación de la clase minoritaria sin replicar datos existentes.

3.5 Método SmoteTomek

Existen métodos de submuestreo que a través de diversas técnicas, eliminan valores de la muestra original para disminuir el desbalanceo de datos. El problema de estos métodos es que al eliminar datos de la muestra original se puede perder información valiosa dentro de la base de datos (Brownlee, 2018). Por ello, este tipo de método se recomienda utilizar en bases de datos que cuentan con muchos datos potencialmente similares.

Como tercer y último método, se utiliza un método híbrido entre sobremuestreo y submuestreo. El método de sobremuestreo a aplicar es distinto al Adasyn para tener variedad de métodos de sobremuestreo y el método de submuestreo se mencionará más adelante.

El método de sobremuestreo escogido para esta tercera y última parte es el método Synthetic Minority Over-sampling Technique, conocido como SMOTE (Brownlee, 2019). Al igual que Adasyn, SMOTE utiliza un proceso de interpolación para generar la cantidad de datos sintéticos necesarios. SMOTE selecciona la primera fila (o una fila aleatoria en ciertos casos) y calcula sus *k* vecinos. Luego, tras tener a los *k* vecinos, de manera al azar selecciona a uno de estos. Tras haberlo seleccionado, la distancia entre el punto minoritario escogido y su vecino es multiplicado por un numero aleatorio entre 0 y 1. Si el número aleatoria es cercano a 0, el dato sintético generado estará cercano al punto minoritario con el que se está trabajando. Si el número es cercano a 1, este nuevo punto sintético estará cerca del vecino escogido.

Este proceso se repite para el punto minoritario seleccionado *N*/100 veces, siendo *N* el porcentaje de sobremuestreo requerido por la clase minoritaria. Si *N*=100, el proceso se lleva a cabo para un solo vecino, en cambio, si *N*=300, el proceso se lleva a cabo para tres vecinos escogidos al azar. Después de completar este paso para el primer punto minoritario, se pasa a la siguiente fila o valor y se repite todo el proceso mencionado hasta completar la cantidad de sobremuestreo solicitada.

Para el submuestreo, se escoge utilizar el método Tomek link. Este método es una modificación de la técnica de submuestreo de vecino más cercano condensado desarrollada por Tomek (1976). El método busca todos los "tomek links" que hay en la base de datos, los cuales son puntos donde la clase mayoritaria y minoritaria tienen la distancia euclidiana más baja entre sí. Tras identificar todos los pares de puntos minoritarios que están muy cerca de puntos mayoritarios, el método elimina de la muestra todos los puntos mayoritarios que tienen tomek link. De esta forma, se obtiene una base de datos donde, en las zonas donde predominan los puntos minoritarios, no se encuentren datos mayoritarios superpuestos.

La combinación del método de sobremuestreo SMOTE con el método de submuestreo Tomek Link es algo que se ha usado previamente en la literatura (Batista et al., 2003; Chawla et al., 2002). A este método híbrido se le denomina SMOTE-TOMEK y ha mostrado mejoras significativas respecto al submuestreo simple.

Los tres métodos definidos en este capítulo son muy distintos entre sí. Primero, se habla de factores de expansión, los cuales utilizan información adicional como la que entrega el censo para poder asignar un peso a las respuestas entregadas por las personas que respondan la encuesta. En segundo lugar, se menciona un método de sobremuestreo que, a través de la interpolación, genera una cantidad de datos sintéticos para que la diferencia entre la cantidad de datos de una clase mayoritaria y minoritaria sea menor. Por último, se describe una combinación entre técnicas de sobremuestreo y submuestreo, que genera datos sintéticos y elimina datos de la muestra original para equilibrar la cantidad de datos que existen entre las clases mayoritarias y minoritarias. Tras aplicar cada uno de los métodos mencionados, se analiza a grandes rasgos si existe una diferencia importante en la aplicación de estos métodos o si los resultados que entregan tienen cierta parecido.

3.6 Desarrollo de Scripts

Tras definir las metodologías que se aplican en la muestra obtenida de Mo2, se utiliza Python para implementarlas. Inicialmente, se investiga si otros desarrolladores han aplicado estas metodologías en Python. Al no encontrar ningún código para el método factor de expansión, se decide desarrollarlo. En cambio, los métodos Adasyn y SmoteTomek se utilizan desde la biblioteca imbalanced-learn.

Se generan archivos Excel o CSV que contienen la información demográfica de las personas que responden la encuesta. La información mínima que debe estar en el archivo es la que se utiliza para dividir a la población. En esta memoria de título, la población se divide por género y rango etario, por lo que todas las respuestas de la muestra cuentan con dos columnas que contienen esta información. Cualquier otra información demográfica que se encuentre en el archivo no se utiliza. Además de la información demográfica, es necesario contar con los valores sobre los cuales se desea hacer inferencias poblacionales. En este caso, como se están analizando los usos de tiempo, todas las respuestas tienen columnas en las cuales se encuentran los nueve tipos de usos de tiempo.

La implementación de factor de expansión, de Adasyn y SmoteTomek tiene en cuenta las características de la muestra entregada por Mo2. Esto significa que, si se quisieran aplicar estos códigos a otras muestras, ya sea de uso de tiempo u otros temas, es necesario realizar ciertas modificaciones para que funcionen correctamente. Todos los código desarrollados y utilizados en la muestra están disponibles en GitHub. El enlace para acceder se encuentra en el anexo A.

3.6.1 Implementación Factor de Expansión

Como primer paso, una vez se agrega al código el archivo con la información demográfica de la muestra, se añade al archivo original tres columnas nuevas denominadas “F.E”, “w” y “p”. Estas columnas almacenan la información que posteriormente genera los valores del factor de expansión final, detallado en el capítulo 3.3. A continuación, el código obtiene la cantidad de respuestas en la muestra y se ingresa el numero correspondiente a la población objetivo que se está estudiando. Con estos dos valores, el código calcula la primera variable, *w*, aplicando la ecuación (4).

Luego, se agregan al código los porcentajes que representan cada grupo de género y grupo etario utilizados para dividir a la población (descritos en el capítulo 3.2). Para cada combinación posible entre los grupos de género y etarios, se calcula el porcentaje que representan dentro de la población objetivo. Este mismo proceso se repite para determinar la cantidad de respuestas que cada uno de estos grupos tiene dentro de la muestra, obteniendo así los valores necesarios para calcular la variable *p* aplicando la ecuación (5).

En la columna "w", se reemplazan los valores originales por los calculados en la variable *w* (en este caso, todas las filas tienen el mismo valor *w*). Los valores de la columna "p" se reemplazan por los obtenidos en las variables *p*, asignando a cada fila el valor correspondiente según las características demográficas de la respuesta. Después de asignar los valores de las columnas "w" y "p" a todas las filas, se calcula y reemplaza el valor de la columna "F.E" a toda la muestra.

Finalmente, como último paso, una vez que se han obtenidos todos los valores de la columna "F.E", este valor se multiplica por los valores en las columnas que contienen información sobre el uso de tiempo de las diferentes actividades. Tras esta multiplicación, el código genera un archivo en el que la distribución de los valores de los usos de tiempo se ajusta para igualar la distribución de la población objetivo.

3.6.2 Implementación Adasyn

Al igual que en el caso del código para el factor de expansión, después de agregar el archivo con la información demográfica de la muestra, se añade una nueva columna al archivo original denominada "clase". Esta columna contiene mayoritariamente valores iguales a 0, excepto en las filas donde la información demográfica corresponda al grupo de hombres o mujeres entre 18 a 30 años, en cuyo caso el valor es igual a 1. Esto se debe a que, como se menciona en el capítulo 3.4, para que esta metodología funcione, es necesario dividir la muestra en una clase mayoritaria y una minoritaria. En este caso, la clase mayoritaria, representada por el numero l, corresponde al grupo mencionado anteriormente. Una vez asignados los valores a la columna "clase", se eliminan del archivo todas las columnas que no contienen únicamente valores números y se rellenan las filas vacías en las columnas restantes, ya que estos son requisitos para que la metodología funcione.

A continuación, el archivo se divide en dos variables necesarias para aplicar la metodología: una es una copia del archivo y la otra contiene únicamente la columna “clase”. Luego, se utiliza la función ADASYN de la biblioteca imbalanced-learn, específicamente la que contiene los métodos de sobremuestreo, para llevar a cabo el remuestreo. Este proceso genera una base de datos con la misma cantidad de datos de la clase mayoritaria, pero con más datos relacionados con la clase minoritaria.

Dado que el archivo resultante contiene más datos que el original, se reutiliza una parte del código del método factor de expansión para obtener la distribución actual de la muestra. Con esta distribución, se calcula cuantas respuestas son necesarias de cada una de las combinaciones posibles entre los grupos de género y etarios para igualar la distribución de la población objetivo con la cantidad de respuestas de la muestra original (en este caso 380 respuestas). Una vez determinado el número de respuestas requerido para cada grupo, se realiza un muestreo aleatorio simple que genera un archivo que refleja la distribución de la población objetivo con la cantidad de respuestas de la muestra original.

3.6.3 Implementación SmoteTomek

La implementación del método SmoteTomek es similar a la del método Adasyn, con la diferencia de que se llama a las funciones correspondientes. En lugar de utilizar la función Adasyn, se emplean las funciones SMOTETomek, TomekLinks y SMOTE, provenientes de las bibliotecas imblearn.combine, imblearn.over_sampling y imblearn.under_sampling respectivamente. El resto de la implementación sigue los mismos pasos descritos para el método Adasyn.

Capítulo 4 Resultados

Este apartado presenta los resultados obtenidos tras aplicar los métodos a la base de datos obtenida de la página Mo2. Como se menciona en el capítulo anterior, la muestra que entrega la página web es preprocesada y se utiliza para corroborar si las metodologías logran tener efecto en la distribución de la muestra y son capaces de igualar su distribución a la de la población. Además, se analizan y comparan los resultados obtenidos con los usos de tiempo promedios que entrega la muestra original.

4.1 Resultados método Factor de Expansión

Tras la aplicación del método de factor de expansión, la base de datos sigue teniendo 380 respuestas, distribuidas de la siguiente forma en las figuras 4.1 y 4.2. Como se aprecia en la Figuras, la distribución de la muestra se iguala a la distribución de la población en términos de grupos de género y etarios. El método de factor de expansión logra su primer objetivo de igualar la distribución de la muestra con la de la población. Luego, se corrobora si la distribución del uso de tiempo también se ve afectada por el cambio de la distribución de la muestra o si este sigue igual. Un punto importante a mencionar es que, como los datos que se obtienen en este método son los mismos que las respuestas originales que entrega la página web, la participación en las actividades es la misma que en la muestra sin la aplicación del factor de expansión. Es importante destacar que esto no ocurre en los otros dos métodos, ya que la creación de datos sintéticos puede influir en la participación de la población en las actividades. Debido a lo anterior, se muestra la información de la participación de las actividades para los otros métodos. A continuación, se presentan los usos de tiempo promedio de la muestra sin la aplicación del factor de expansión y el uso de tiempo tras haber aplicado el factor de expansión en la tabla 4.1.

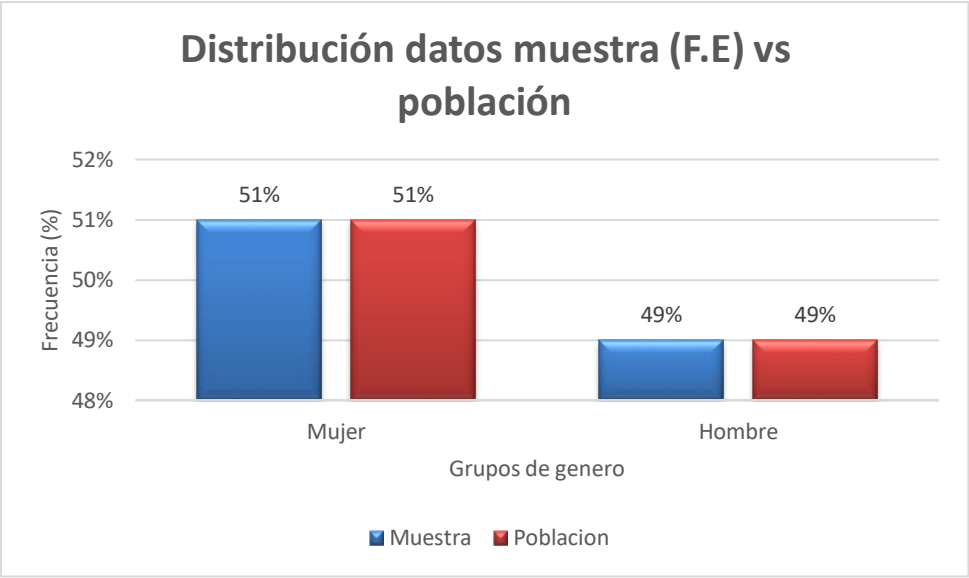


Figura 4.1: Distribución de la muestra (F.E) por género en comparación a la de la población

Fuente: elaboración propia

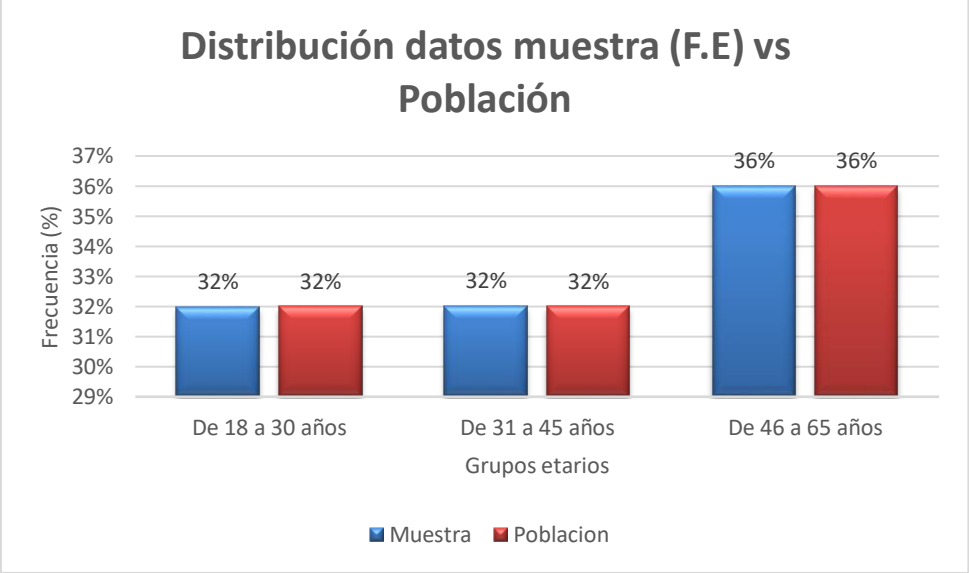


Figura 4.2: Distribución de la muestra (F.E) por rango etario en comparación a la de la población

Fuente: elaboración propia

Tabla 4.1: Comparación usos de tiempo promedio muestra original y tras aplicar F.E

Actividad	Muestra original	Aplicación F.E	
	Tiempo Promedio (horas)	**Tiempo Promedio (horas)**	**Diferencia entre muestras (horas)**
Trabajo Remunerado	312 (5,2)	360 (6,0)	48 (0,8)
Trabajo Doméstico	68 (1,1)	82 (1,4)	14 (0,3)
Trabajo de cuidado	16 (0,3)	20 (0,3)	4 (0,0)
Voluntariado	6 (0,1)	12 (0,2)	6 (0,1)
Educación	119 (2,0)	71 (1,2)	-48 (0,8)
Ocio	230 (3,8)	222 (3,7)	-8 (0,1)
Cuidado Personal	111 (1,8)	100 (1,7)	-11 (0,1)
Dormir	468 (7,8)	465 (7,8)	-3 (0,0)
Desplazamiento	53 (0,9)	49 (0,8)	-4 (0,1)
Total Muestra	1380 (23,0)	1380 (23,0)	0 (0,0)

Fuente: elaboración propia

Como se puede apreciar de la Tabla 4.1, hay varios cambios en los tiempos promedios de uso de tiempo para las actividades. Los tiempos de algunas actividades aumentan y otros disminuyen, se destaca que algunos de los cambios de uso de tiempo promedio son más significativos que otros. La actividad que presenta el mayor cambio es la de trabajo remunerado, pasando de 5,2 horas a 6 horas. Este cambio tiene sentido, considerando que la muestra está sobrerrepresentada por el grupo etario de 18 a 30 años y teniendo en cuenta que la mayoría de la población ronda entre los 31 a los 65 años, no es sorprendente que el uso de tiempo del trabajo remunerado aumente.

Aunque el aumento del tiempo promedio dedicado al trabajo puede no haber sido el esperado o parecer bajo, esto puede deberse a que la recolección de datos se realiza tanto para los días de semana como fines de semana. Aproximadamente un 10% de las respuestas corresponden a tiempos del fin de semana, lo que disminuye el tiempo promedio dedicado al trabajo en comparación con si solo se consideran los días de semana. La segunda actividad con mayor cambio de uso de tiempo promedio es la educación, que pasó de 2,0 horas al día a 1,2 horas, este cambio va muy de la mano con lo que ocurrió con el cambio de uso de tiempo promedio del trabajo remunerado. La tercera actividad con mayor cambio de uso de tiempo promedio es el trabajo doméstico, de 1,1 a 1,4 horas por día. El resto de las actividades, si bien también tienen variaciones en sus tiempos promedios, estos cambios no son tan significativos, en promedio, las variaciones en estas actividades rondan los 6 minutos de aumento

o disminución. La figura 4.3 presenta la distribución de los usos de tiempo por género y por grupo etario.

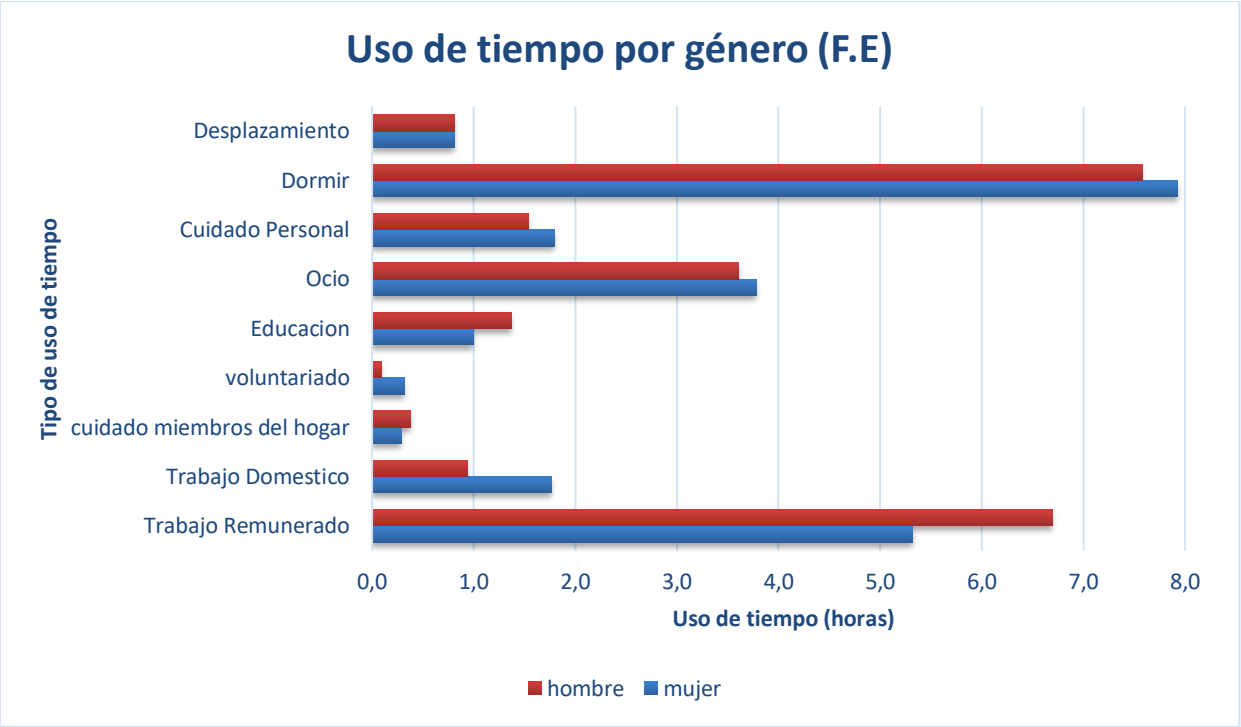


Figura 4.3: Uso de tiempo promedio por género (F.E)

Fuente: elaboración propia

A simple vista, la figura 4.3, al igual que el obtenido de la muestra sin aplicar ninguna metodología, muestra varias actividades a las que hombres y mujeres asignan tiempos muy parecidos durante el día. Las actividades con mayores diferencias de asignación de tiempo son trabajo remunerado, donde los hombres le dedican cerca de 80 minutos más que las mujeres. Mientras, el trabajo doméstico, donde las mujeres le dedican aproximadamente 50 minutos más que los hombres durante el día. Por último, la actividad de educación, donde la diferencia entre géneros es de 25 minutos aproximadamente, donde los hombres le dedican más tiempo que las mujeres. Si se compara con las actividades que tienen mayores diferencias en la muestra original, se aprecia que las actividades que se mantienen en ambos casos son el trabajo remunerado y el trabajo doméstico, pero con tiempos promedios distintos en ambos casos. Tras aplicar el factor de expansión, dormir deja de ser la tercera actividad con mayor diferencia de uso de tiempo promedio, siendo reemplazada por educación. Este cambio de actividades

indica que la metodología no solo afecta la distribución de uso de tiempo de la población en general, sino que también influye en los tiempos promedios desde una perspectiva de vista de género.

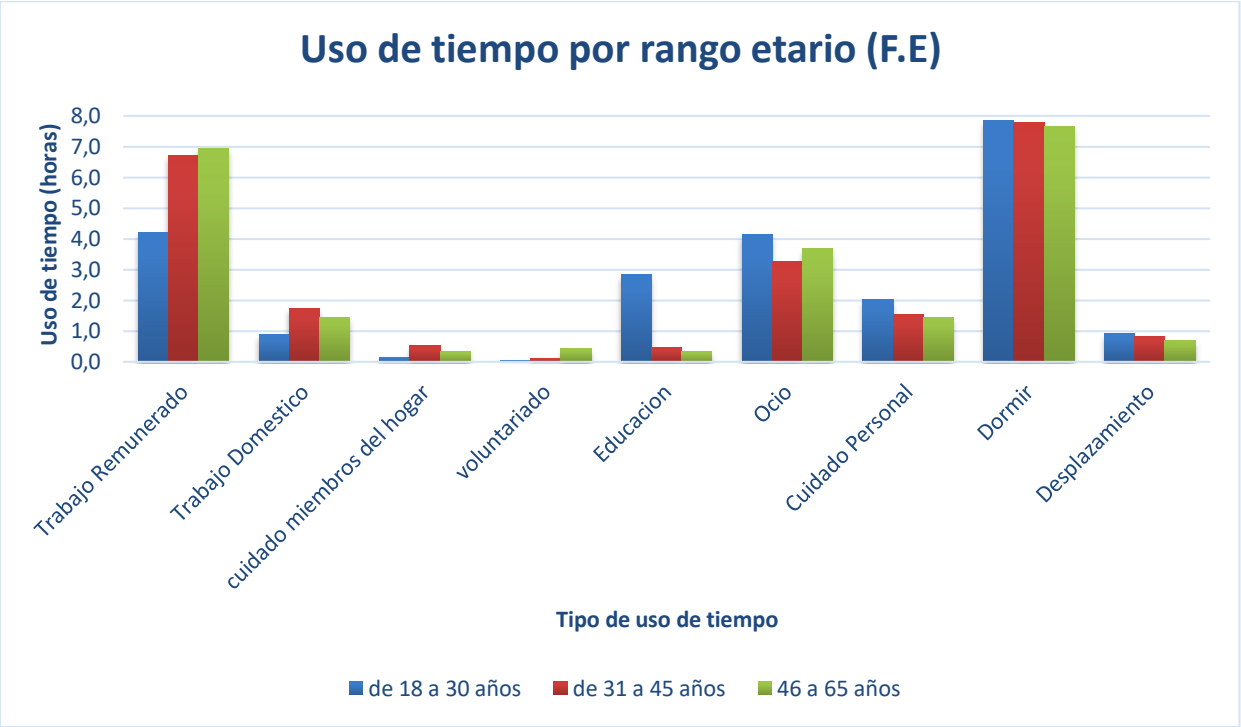


Figura 4.4: Uso de tiempo promedio por rango etario (F.E)

Fuente: elaboración propia

La figura 4.4 presenta las asignaciones de uso de tiempo por rango etario, donde se observa que las diferencias son muy similares a las de la muestra original. Esto puede deberse a la distribución inicial de la muestra, que mantiene las diferencias en el uso de tiempo por edades inalteradas, a pesar de la aplicación del factor de expansión. Alternativamente, aunque esta metodología afecta la distribución del tiempo en la población en general y desde una perspectiva de género, es posible que no logre reflejar un cambio significativo en la distribución del uso de tiempo por edades.

4.2 Resultados método Adasyn

Tras la aplicación del método Adasyn, la base de datos queda con 490 respuestas. Este aumento en la cantidad de respuestas respecto a la muestra original, que tenía 380 respuestas, se debe a la creación de los nuevos datos sintéticos basados en las respuestas existentes en la muestra original. Para comparar los resultados entregados por este método, se realiza un muestreo aleatorio simple dentro de la muestra de 490 respuestas para que quede con 380 respuestas. Pero, cumpliendo con la distribución de los datos obtenidos del censo.

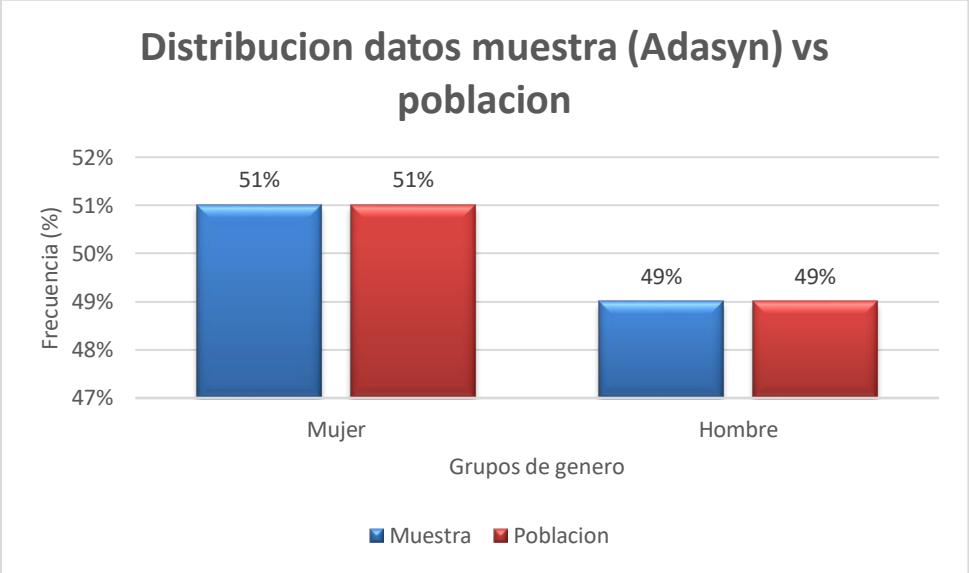


Figura 4.5: Distribución de la muestra (Adasyn) por género en comparación a la de la población

Fuente: elaboración propia

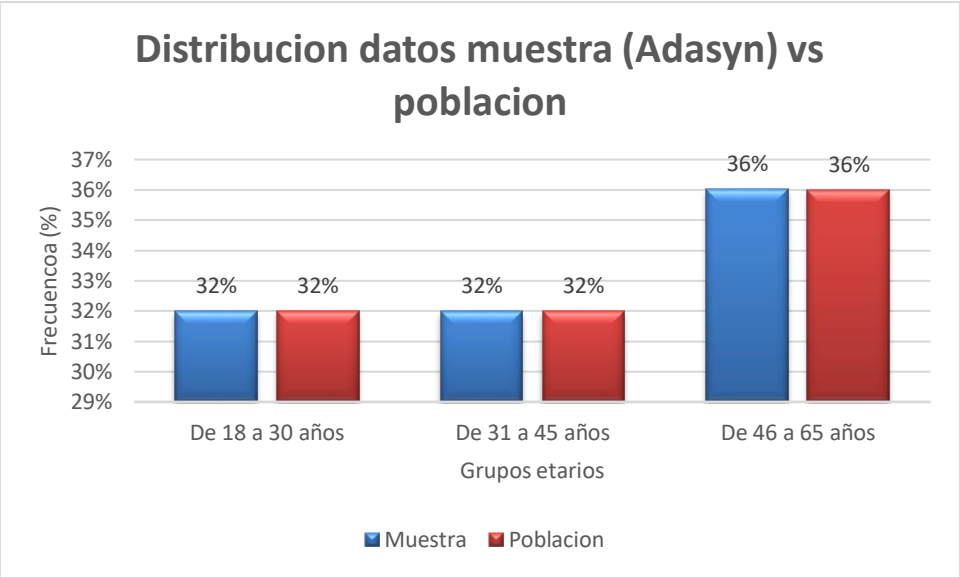


Figura 4.6: Distribución de la muestra (Adasyn) por rango etario en comparación a la de la población

Fuente: elaboración propia

La Figura 4.5 y 4.6 presenta la distribución resultante de los datos de la muestra. A pesar del aumento en las respuestas, el muestro aleatorio simple logra obtener una muestra que se distribuye de igual forma que la población desde un punto de vista de grupos etarios y de género. Un elemento a tener en cuenta respecto a este método dado que se lleva a cabo un muestreo aleatorio simple dentro de la base de datos de 490 respuestas, si se obtiene otra muestra de esta base de datos, algunos de los valores de uso de tiempo promedio, por género o por rango etario podrían llegar a variar entre muestras. Esta diferencia de valores no debería ser significativa, ya que los valores sintéticos generados provienen de los datos originales, por lo que sus distribuciones deben ser similares.

Tabla 4.2: Comparación usos de tiempo promedio muestra original y tras aplicar Adasyn

		Muestra original	Aplicación Adasyn	
Actividad	**Participación (%)**	**Tiempo Promedio (horas)**	**Tiempo Promedio (horas)**	**Diferencia entre muestras (horas)**
Trabajo Remunerado	266 (70%)	312 (5,2)	349 (5,8)	37 (0,6)
Trabajo Doméstico	296 (78%)	68 (1,1)	81 (1,4)	13 (0,3)
Trabajo de cuidado	129 (34%)	16 (0,3)	18 (0,3)	2 (0,0)
Voluntariado	46 (12%)	6 (0,1)	11 (0,2)	5 (0,1)
Educación	134 (35%)	119 (2,0)	74 (1,2)	-45 (0,8)
Ocio	357 (94%)	230 (3,8)	228 (3,8)	-2 (0,0)
Cuidado Personal	338 (89%)	111 (1,8)	100 (1,7)	-11 (0,1)
Dormir	380 (100%)	468 (7,8)	466 (7,8)	-2 (0,0)
Desplazamiento	287 (76%)	53 (0,9)	49 (0,8)	-4 (0,1)
Total Muestra		1380 (23,0)	1375 (22,9)	-5 (0,1)

Fuente: elaboración propia

De la Tabla 4.2 se destaca que, a diferencia del factor de expansión, la implementación del método Adasyn sí afecta la participación en las actividades en el total de las respuestas. Por ejemplo, en la muestra original, la actividad de trabajo remunerado tiene una participación del 69% (es decir, 264 respuestas indican haber dedicado tiempo a esta actividad) y con esta nueva muestra, este porcentaje de participación sube al 70%. Ocurre lo contrario en el caso de la actividad educación, donde en la muestra original se tiene una participación del 46%, pero en esta nueva muestra la participación baja al 35%. Esto pude llegar a sobreestimar o subestimar la participación en algunas actividades de la población.

Analizando los datos entregados por esta muestra, se aprecia que las actividades que tienen mayores variaciones son educación, donde el tiempo promedio dedicado a esta actividad disminuye en 45 minutos. Después viene trabajo remunerado, que tuvo un aumento de 37 minutos y por último, trabajo doméstico, la cual tuvo un aumento de 13 minutos. El resto de las actividades tienen una variación promedio de 4 minutos, los cuales aumentan o disminuyen el tiempo promedio que se le dedica durante el día.

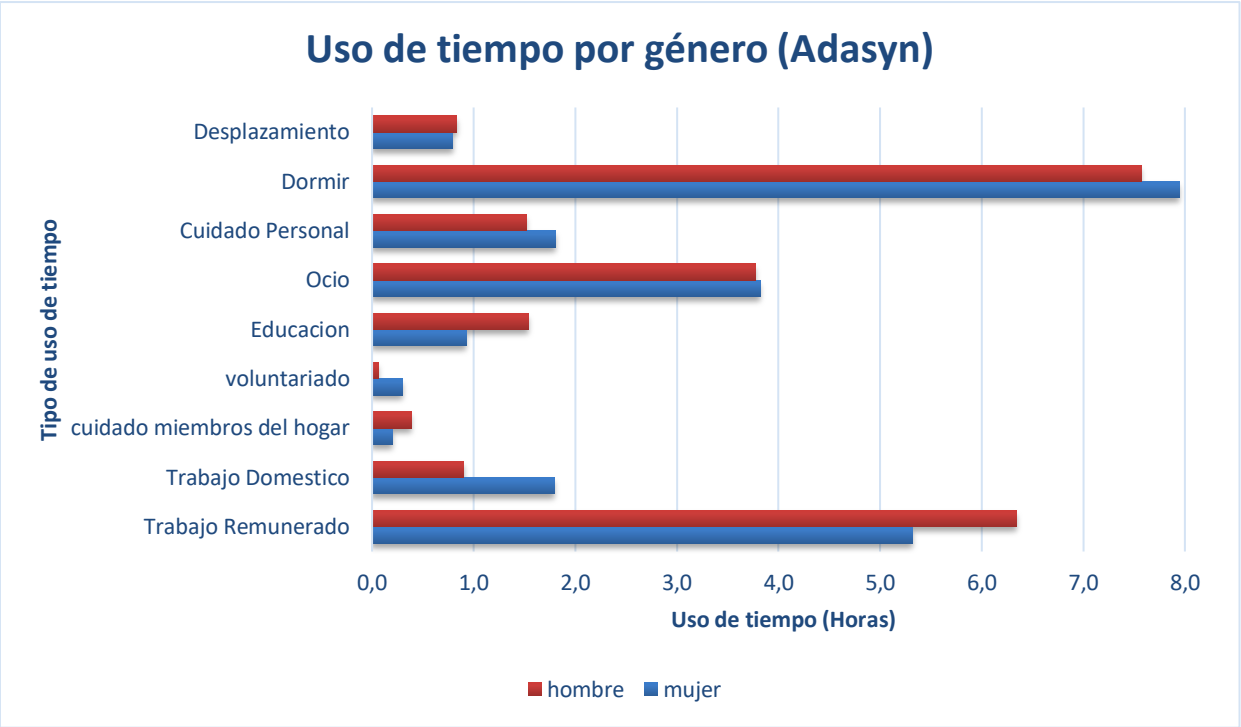


Figura 4.7: Uso de tiempo promedio por género (Adasyn)

Fuente: elaboración propia

En las figuras 4.7 y 4.8 se analizan los usos de tiempo por género y rango etario respectivamente. Desde un punto de vista de género, se aprecia que la actividad con mayor diferencia en uso de tiempo es trabajo remunerado, donde los hombres dedican aproximadamente 1 hora más que las mujeres. La segunda actividad que presenta mayores diferencias es el trabajo doméstico, donde las mujeres le dedican más tiempo durante el día a día que los hombres, con una diferencia de aproximadamente 1 hora. Por último, se tiene la actividad de educación, donde los hombres le dedican alrededor de 35 minutos más que las mujeres. Si se comparan estas diferencias con las obtenidas mediante el método de factor de expansión, se observa que las tres actividades con mayores diferencias de uso de tiempo son las mismas para ambos métodos, aunque hay algunas diferencias en los tiempos promedios. Esto indica que la aplicación de ambos métodos desde un punto de vista de género tiene el mismo efecto en la distribución de los usos de tiempo. Al igual que con la aplicación del método de factor de expansión, la distribución del uso de tiempo por rango etario es muy parecida a la de la muestra original. Si bien cambian algunos tiempos promedios, en algunos rangos etarios estos cambios no son significativos.

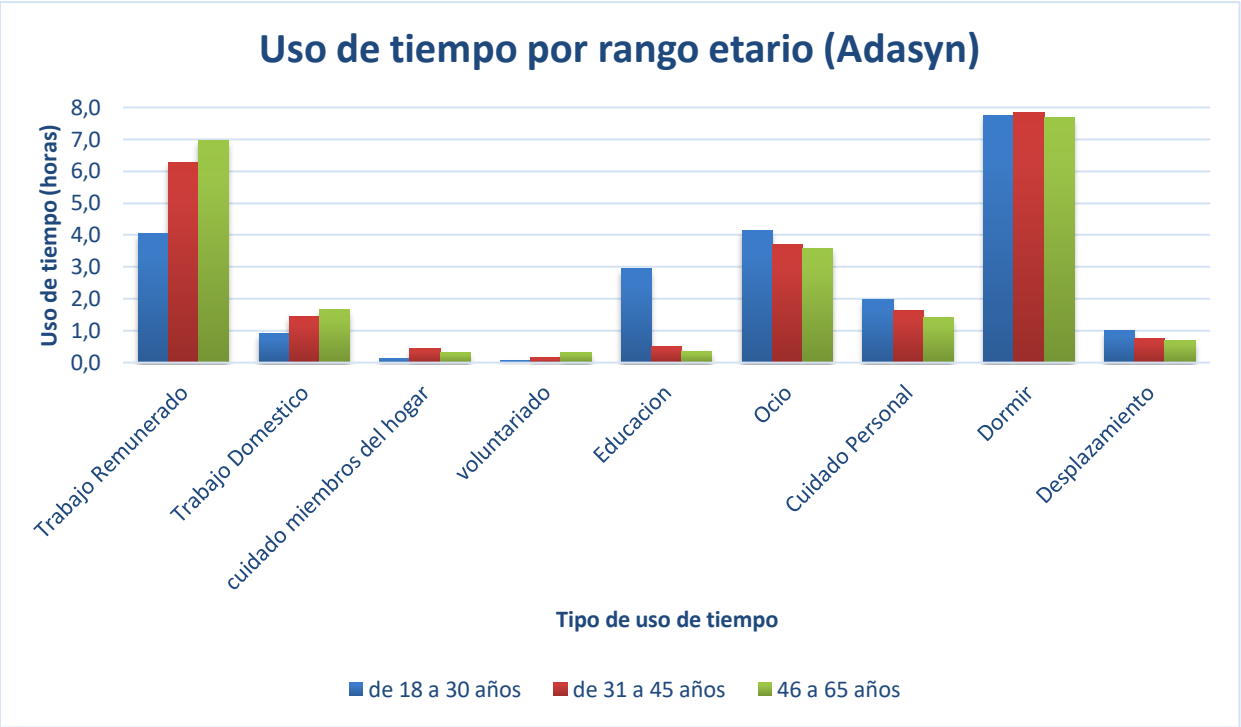


Figura 4.8: Uso de tiempo promedio por rango etario (Adasyn)

Fuente: elaboración propia

4.3 Resultados método SmoteTomek

Como ocurre con la aplicación del método Adasyn, la muestra después de generar nuevos datos sintéticos queda con 439 respuestas. En este método se generan menos respuestas que en Adasyn, y debido a su naturaleza híbrida, además de generar nuevos datos también elimina algunos, tras la eliminación de datos, la muestra queda con 429 respuestas. Se aprecia que la cantidad de datos eliminados es mínima, esto es debido principalmente a la distribución de los datos. Como en Adasyn, se realiza un muestreo aleatorio simple para poder comparar las tres metodologías y obtener una muestra que cumpla con la distribución obtenida del censo, con solo 380 respuestas, equivalentes a la cantidad de datos que hay en la muestra original.

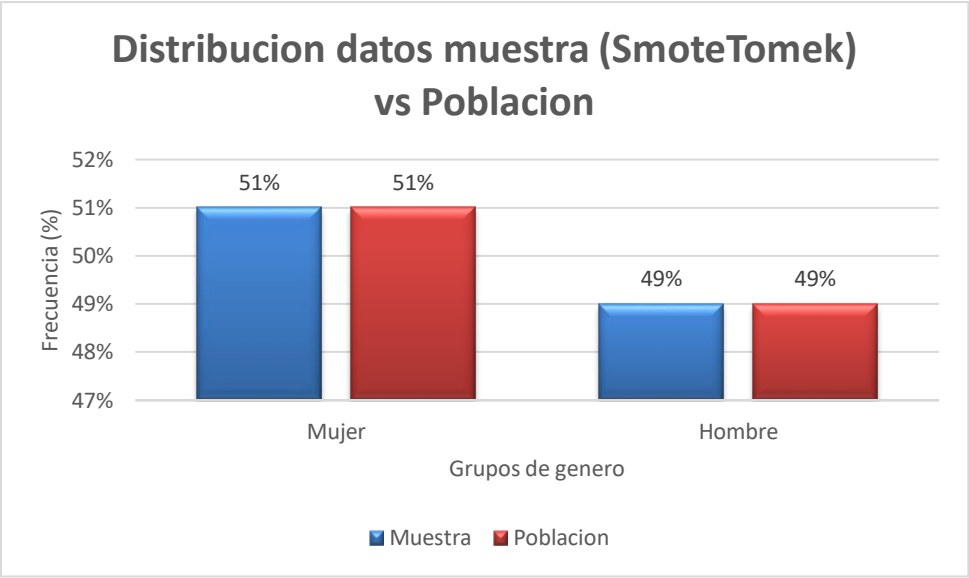


Figura 4.9: Distribución de la muestra (SmoteTomek) por género en comparación a la de la población

Fuente: elaboración propia

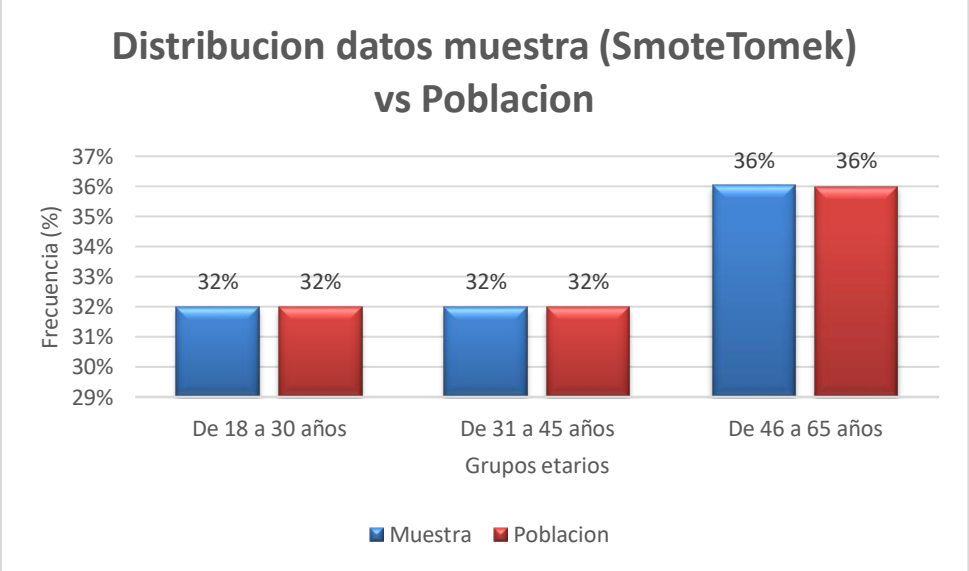


Figura 4.10: Distribución de la muestra (SmoteTomek) por rango etario en comparación a la de la población

Fuente: elaboración propia

Tras este muestreo aleatorio simple, la muestra queda como muestra la Figura 4.9 y 4.10. Al igual que en los otros dos métodos, se aprecia que la distribución de la muestra proporcionada por el método SmoteTomek es igual a la distribución de la población tanto en género como en grupos etarios, permitiendo así que los resultados obtenidos de la muestra se puedan usar para hacer inferencias poblaciones. Los promedios de uso de tiempo quedan de la siguiente manera:

Tabla 4.3: Comparación usos de tiempo promedio muestra original y tras aplicar SmoteTomek

		Muestra original	SmoteTomek	
Actividad	**Participación (%)**	**Tiempo Promedio (horas)**	**Tiempo Promedio (horas)**	**Diferencia entre muestras (horas)**
Trabajo Remunerado	270 (71%)	312 (5,2)	344 (5,7)	32 (0,5)
Trabajo Doméstico	300 (79%)	68 (1,1)	71 (1,2)	3 (0,1)
Trabajo de cuidado	117 (31%)	16 (0,3)	18 (0,3)	2 (0,0)
Voluntariado	48 (13%)	6 (0,1)	12 (0,2)	6 (0,1)
Educación	137 (36%)	119 (2,0)	71 (1,2)	-48 (0,8)
Ocio	352 (93%)	230 (3,8)	243 (4,1)	13 (0,3)
Cuidado Personal	331 (87%)	111 (1,8)	101 (1,7)	-10 (0,1)
Dormir	380 (100%)	468 (7,8)	474 (7,9)	6 (0,1)
Desplazamiento	286 (75%)	53 (0,9)	48 (0,8)	-5 (0,1)
Total Muestra		1380 (23,0)	1382 (23,0)	2 (0,0)

Fuente: elaboración propia

De la Tabla 4.3 se destaca que, al igual que con el método Adasyn, la participación en diversas actividades también se ve afectada tras aplicar el método SmoteTomek. Si se analiza la participación de la actividad trabajo de cuidado, se observa que en esta nueva muestra hay un aumento en la cantidad de respuestas que participan, pasando de un 24% en la muestra original a un 31%. Lo contrario ocurre con la actividad de educación, que disminuye del 46% en la muestra original a un 36% en la nueva muestra.

Las actividades que experimentan mayores cambios en los tiempos promedios dedicados por la población durante el día es educación, con una disminución de 48 minutos. La segunda actividad con mayores diferencias en el uso de tiempo es trabajo remunerado con un aumento de 32 minutos. Por último, la tercera actividad con mayor diferencia en uso de tiempo promedio es el ocio con un aumento de 13 minutos.

Tras analizar las actividades que más cambia sus tiempos promedios después de aplicar las diferentes metodologías, se observa que los tres métodos siguen tendencias parecidas en la mayoría de las actividades: las actividades de educación y trabajo remunerado aparecen como las actividades con mayores diferencias en los tres casos, aunque los tiempos varían. En dos ocasiones se observa la actividad trabajo doméstico y en el último método aparece la actividad de ocio. Este comportamiento probablemente se debe al funcionamiento de los métodos.

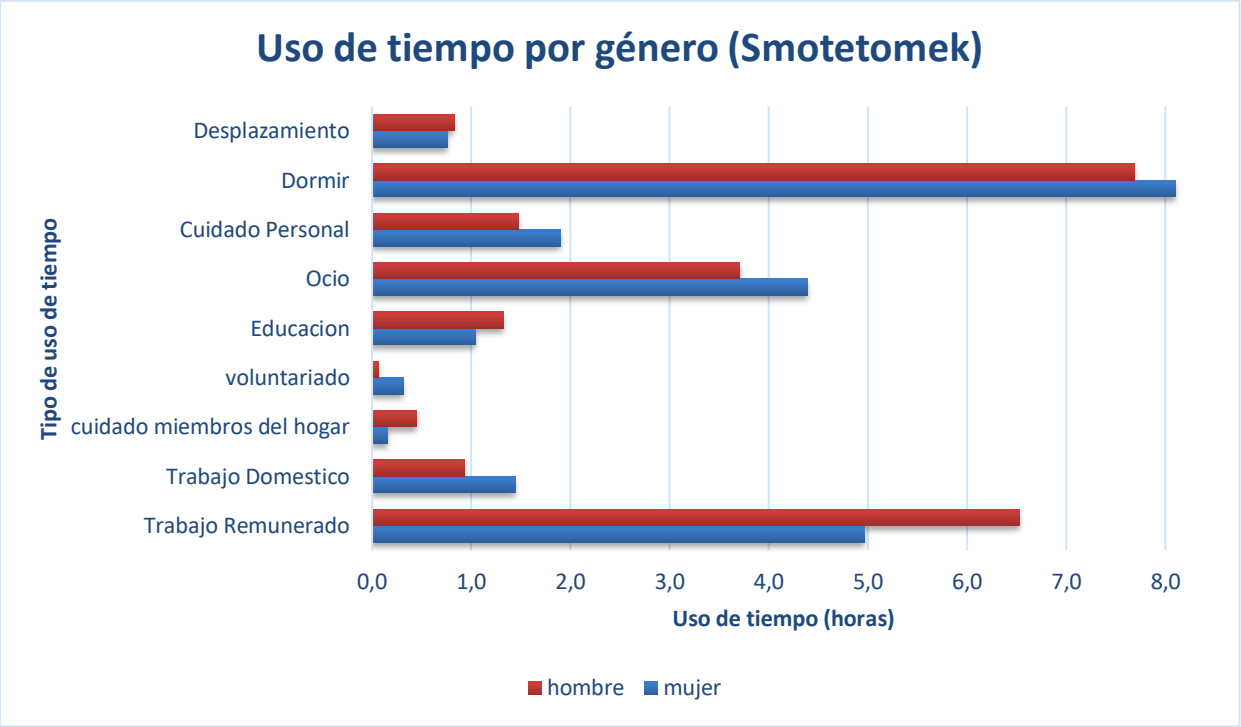


Figura 4.11: Uso de tiempo promedio por género (SmoteTomek)

Fuente: elaboración propia

En las figuras 4.11 y 4.12 se analizan los usos de tiempo por género y rango etario, respectivamente. Desde un punto de vista de género, se observa que la actividad con la mayor diferencia en el uso de tiempo promedio es el trabajo remunerado, con una diferencia aproximada de 95 minutos, siendo los hombres quienes dedican más tiempo a esta actividad. La segunda actividad con mayor diferencia en el uso de tiempo promedio es el ocio, con una diferencia de aproximadamente 40 minutos, siendo las mujeres quienes le dedican más tiempo. Por último, la actividad trabajo doméstico es la tercera

actividad con mayor diferencia en uso de tiempo promedio, donde las mujeres le dedican aproximadamente 30 minutos más que los hombres durante el día.

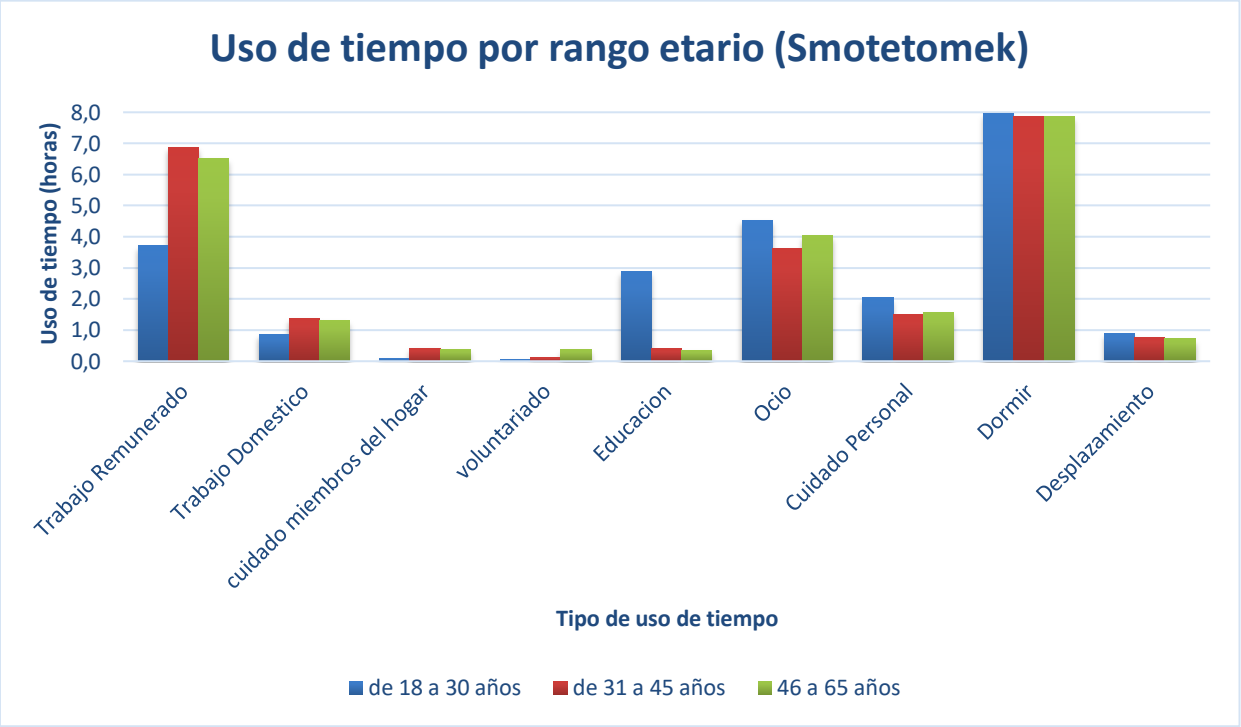


Figura 4.12: Uso de tiempo promedio por rango etario (SmoteTomek)

Fuente: elaboración propia

En la Figura 4.12 se observa que finalmente se cumple el supuesto inicial de que la distribución del uso de tiempo por rango etario es constante para las tres metodologías. Además, a simple vista se puede analizar que las diferencias entre rangos etarios para las diversas actividades son muy parecidas a las ya vistas.

4.4 Comparación de Resultados

Tras haber visto y comparado los resultados que entregan las tres metodologías respecto a la muestra original, en este apartado se realiza un análisis general sobre qué tan diferentes son los resultados entregados por las tres alternativas.

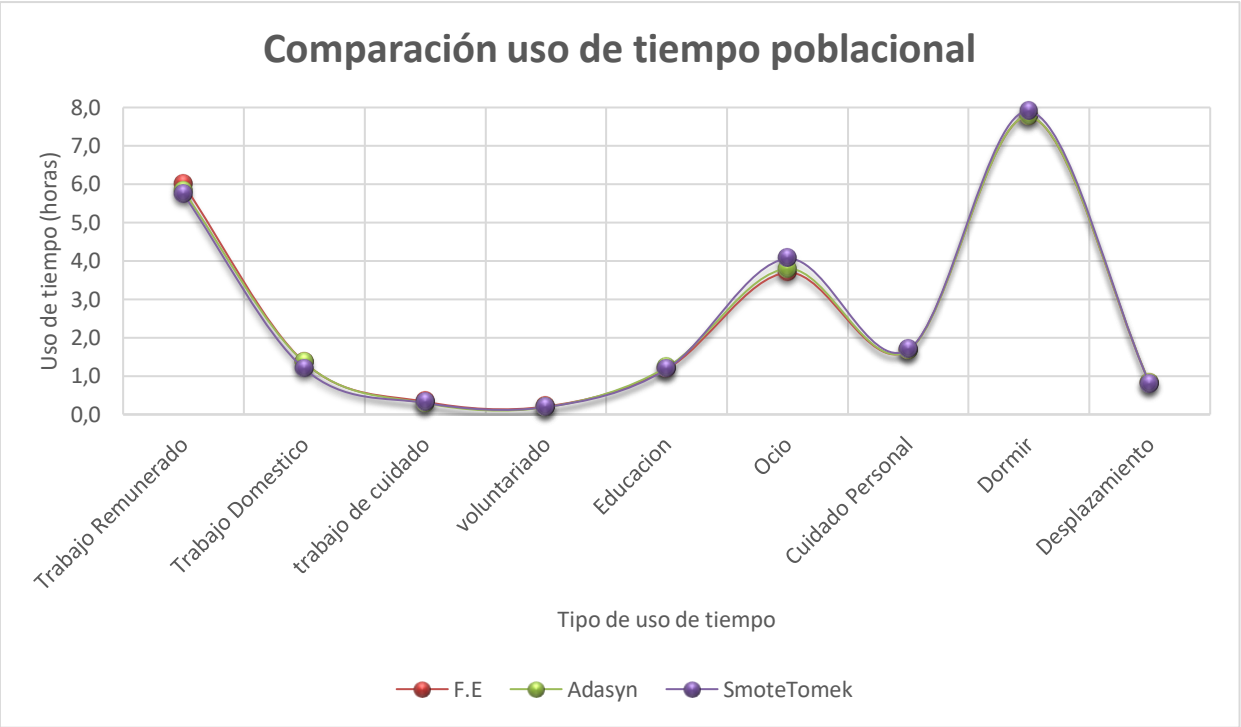


Figura 4.13: Comparación uso de tiempo promedio de la población por metodología

Fuente: elaboración propia

La Figura 4.13 muestra que para las tres metodologías aplicadas a la muestra, en la mayoría de las actividades se obtienen usos de tiempo promedio poblacionales parecidos, en algunos casos uno de los métodos muestra usos de tiempo ligeramente diferentes. Por ejemplo, en la actividad trabajo remunerado, ocio y trabajo doméstico, se observa que uno de los métodos entrega valores medianamente distintos al resto. En el caso de trabajo remunerado, el método factor de expansión entrega un uso de tiempo promedio de la población de 6,0 horas, mientras que Adasyn muestra 5,8 horas y SmoteTomek 5,7 horas.

La diferencia en el uso de tiempo promedio entre el método SmoteTomek y factor de expansión es de 0,3 horas, y entre SmoteTomek y Adasyn es de 0,1 horas. Para el caso de la actividad de ocio, SmoteTomek entrega un uso de tiempo promedio de 4,1 horas, Adasyn con 3,8 horas y factor de expansión con 3,7 horas. Así, la diferencia entre Adasyn y factor de expansión es de 0,1 horas y la diferencia entre Adasyn y SmoteTomek es de 0,3 horas.

Es interesante notar que todos los métodos muestran valores más altos que los otros en alguna actividad. Se podría esperar que el método de factor de expansión tuviera un comportamiento inusual o por su contraparte que los métodos de remuestreo hayan sido los que presentan comportamientos inusuales. Pero no se esperaba que los tres métodos mostraran este comportamiento, lo cual puede ser un indicio de que estos métodos ante ciertas distribuciones o características de los datos son más sensibles al proporcionar valores promedios.

El mismo análisis se realiza respecto a los tiempos promedios por rango etario. El comportamiento fue consistente con el observado en los tiempos promedio de la población, en los anexos B, C y D se encuentran los gráficos que comparan los de uso de tiempo para los grupos etarios de 18 a 30 años, de 31 a 45 años y 46 a 65, respectivamente.

Para finalizar, este análisis se hizo respecto al género, esperando que ambos tuvieran un comportamiento similar al observado hasta ahora. Esto ocurre para los tiempos promedio de las mujeres, de hecho, al revisar el anexo E, se observa que hay más actividades que presentan tiempos notoriamente más altos que otros, la mayoría de estos tiempos dados por el método Adasyn. Para el género masculino, el comportamiento de los tiempos promedio de las actividades para los tres métodos es más consistente, excepto para la actividad de educación y trabajo remunerado, que muestran ligeras diferencias. En términos generales, el comportamiento de los valores promedios para este grupo es bastante más regular que el resto, lo que podría deberse a que la mayoría de las respuestas en la base de datos original eran de hombres y esto puede influir en la sensibilidad de los métodos. La Figura 4.14 muestra el uso de tiempo del género masculino, como se mencionó previamente, se observa un comportamiento más uniforme en los tiempos promedios de las actividades. Sin embargo, esto no se aplica al caso del trabajo remunerado, donde las tres metodologías presentan valores diferentes. Las diferencias entre metodologías son de 0,2 horas, siendo Adasyn el que muestra el menor tiempo promedio con 6,3 horas, y el método Factor de Expansión el que presenta el tiempo más alto con 6,7 horas. En cuanto a la actividad educación, Adasyn presenta un tiempo promedio de 1,5 horas, lo que es ligeramente superior a los otros métodos.

Figura 4.14: Comparación uso de tiempo promedio de los hombres por metodología

Fuente: elaboración propia

Capítulo 5 Conclusión

El uso de muestras no probabilísticas en grupos poblacionales medianos o grandes ha aumentado con el paso de los años. Estas muestras pueden obtenerse de manera tradicional o en línea y son más económicas y rápidas que una muestra probabilística. Si bien son útiles para obtener información específica del grupo de la muestra, hacer inferencias poblacionales con esta puede ser difícil debido a que la distribución de los datos no representa fielmente a la de la población. Si se busca realizar inferencias poblacionales más precisas y se dispone del tiempo y los recursos, es preferible llevar a cabo una muestra probabilista. Sin embargo, son pocos los investigadores que pueden llevar a cabo este proceso, ya que se trata de un proyecto que requiere una considerable inversión de tiempo y dinero, y no todos los investigadores tienen acceso a financiamiento por parte de instituciones ni cuentan con tanto tiempo para realizar este tipo de proyectos.

El objetivo de esta memoria de título fue evaluar diversas metodologías estadísticas para mejorar la representatividad de una muestra no probabilística, que fue recolectada en línea sobre el uso de tiempo, con el fin de realizar inferencias poblacionales sin recurrir a una muestra probabilística.

A lo largo del estudio, se presentaron los resultados obtenidos al aplicar tres metodologías diferentes a una muestra no probabilística por conveniencia obtenida en línea. Cada metodología entrego valores de tiempos promedio muy parecidos para la mayoría de las actividades. El método de factor de expansión, si bien fue el más diferente de los tres, al utilizar solo dato de la muestra original, sin importar cuantas veces se aplique este método, los resultados siempre son los mismos. Esto no ocurre con los métodos Adasyn y SmoteTomek, al generar más datos que la muestra original, los valores promedio de la muestra varían si se aplican varias veces la metodología.

Una limitación encontrada al aplicar los métodos Adasyn y SmoteTomek es que la creación de datos sintéticos afecta la participación en las actividades en el total de las respuestas. Aunque esto no es un problema para esta memoria de título ya que no se enfoca en esta área. Si se analiza esto, podría llevar a la sobreestimación o subestimación de la participación de la población ciertas actividades, esto no ocurre con la aplicación del método factor de expansión. Para trabajos futuros, sería beneficioso realizar una comparación directa entre los resultados de una muestra probabilística y una no

probabilística. Esto permitirá evaluar qué tan cercanos o lejanos son los resultados y así proporcionar un mayor o menor respaldo a la aplicación de estos métodos en muestras no probabilística para realizar inferencias poblacionales.

Las tres metodologías propuestas cumplen su objetivo de mejorar la representatividad de una muestra no probabilística. En todos los casos, las distribuciones de la muestra se igualaron a las de la población, indicando una representación adecuada de la población objetivo. Sin embargo, el método SmoteTomek no funcionó como se esperaba, ya que, aunque igualó la distribución de la muestra a la de la población, su rendimiento en la eliminación de datos fue deficiente, ya que solo eliminó el 2% de estos. Esto lo hizo parecerse al método Adasyn, que solo genera datos sintéticos.

Respecto a las metodologías aplicadas, se puede decir que es recomendable la aplicación de una o varias de estas en muestras no probabilísticas para realizar inferencias poblaciones. Sin embargo, dado el funcionamiento del método SmoteTomek, sería ideal sustituirla por otra metodología híbrida, como la combinación del método SMOTE con la técnica RandomUnderSampler (submuestreo aleatorio) (Brownlee, 2018). Dado que los Tomek Links que utilizan la distancia entre respuestas para eliminar datos, no funcionó bien, se podría probar con una técnica que elimine de manera aleatoria la cantidad de datos necesaria.

En conclusión, aunque las tres metodologías igualan la distribución de una muestra a la de la población y entregan valores promedio similares, los resultados de una muestra no probabilística deben usarse solo como referencia del comportamiento poblacional. Para conocer efectivamente el comportamiento de la población, se recomienda utilizar una muestra probabilística siempre que sea posible.

Capítulo 6 Referencias

Aguirre, R., & Ferrari, F. (2014). Las encuestas sobre uso del tiempo y trabajo no remunerado en América Latina y el Caribe: Caminos recorridos y desafíos hacia el futuro. CEPAL.

Batista, G., Bazzan, A., & Monard, M.-C. (2003). Balancing Training Data for Automated Annotation of Keywords: A Case Study. En The Proc. Of Workshop on Bioinformatics (p. 18).

Bethlehem, J. (2008). Representativity of web surveys: An illusion? SEAPRAP research report: a report of research undertaken with the assistance of an award from the Southeast Asia Population Research Awards Program (SEAPRAP), Institute of Southeast Asian Studies, 19-44.

Bethlehem, J., & Biffignandi, S. (2011). The Problem of Self-Selection. En Handbook of Web Surveys (pp. 303-327). John Wiley & Sons, Ltd. DOI:10.1002/9781118121757.ch9

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., & Krieger, U. (2016). A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe. Social Science Computer Review, 34(1), 8-25. DOI:10.1177/0894439315574825

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. Field Methods, 27(4), 391-408. DOI:10.1177/1525822X15574494

Brownlee, J. (2018). Random oversampling and undersampling for imbalanced classification. Machine Learning Mastery. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

Brownlee, J. (2019). SMOTE oversampling for imbalanced classification. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014). A critical review of studies

investigating the quality of data obtained with online panels based on probability and nonprobability samples1. En Online Panel Research (pp. 23-53). John Wiley & Sons, Ltd. DOI:10.1002/9781118763520.ch2

Chatzitheochari, S., Fisher, K., Gilbert, E., Calderwood, L., Huskinson, T., Cleary, A., & Gershuny, J. (2018). Using New Technologies for Time Diary Data Collection: Instrument Design and Data Quality Findings from a Mixed-Mode Pilot Survey. Social Indicators Research, 137(1), 379-390. DOI:10.1007/s11205-017-1569-5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. DOI:10.1613/jair.953

Comisión Económica para América Latina y el Caribe (CEPAL). (2016). Clasificación de actividades de uso del tiempo para América Latina y el Caribe. Naciones Unidas. Recuperado de https://repositorio.cepal.org/server/api/core/bitstreams/c50dadaf-6fc3-44ec-8dee-b4b6895749fc/content

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. Journal of Survey Statistics and Methodology, 8(1), 4-36. DOI:10.1093/jssam/smz041

Dillman, D., & Bowker, D. (1999). The Web Questionnaire Challenge to Survey Methodologists. Online social sciences, 7, 53-71.

Documento metodológico enut 2015. (2015). Recuperado 10 de agosto de 2024, de https://bit.ly/3SGICl9

Fisher, R. A. (1925). Statistical Methods for Research Workers.

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. Behavior Research Methods, 47(3), 685-690. DOI:10.3758/s13428-015-0617-9

Instituto Nacional de Estadísticas. (2017). ¿Qué es el censo? Recuperado de http://www.censo2017.cl/que-es-el-censo/

Instituto Nacional de Estadísticas. (2024). Informe sobre la Encuesta Nacional sobre Uso del Tiempo (ENUT) 2023 [informe]. Santiago, Chile: INE.

Jara-Díaz, S., & Rosales-Salas, J. (2015). Understanding time use: Daily or weekly data? Transportation Research Part A: Policy and Practice, 76, 38-57. DOI:10.1016/j.tra.2014.07.009

Kish, L. (1965). Survey sampling. John Wiley & Sons.

Leon, O., & Morgado, I. (1993). Diseños de Investigación. Introducción a la lógica de la Investigación en Psicología y Educación. Psicothema, 6(1), 121.

Levy, P. S., & Lemeshow, S. (2013). Sampling of Populations: Methods and Applications. John Wiley & Sons.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97(4), 558. DOI:10.2307/2342192

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. International Journal of Morphology, 35(1), 227-232. DOI:10.4067/S0717-95022017000100037

Rojas, H. A. G. (2016). Estrategias de muestreo: Diseño de encuestas y estimación de parámetros. Ediciones de la U.

Rui. (2019). An introduction to ADASYN with code. Medium. https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa

Tomek, I. (1976). Two Modifications of CNN in IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769-772. DOI:10.1109/TSMC.1976.4309452

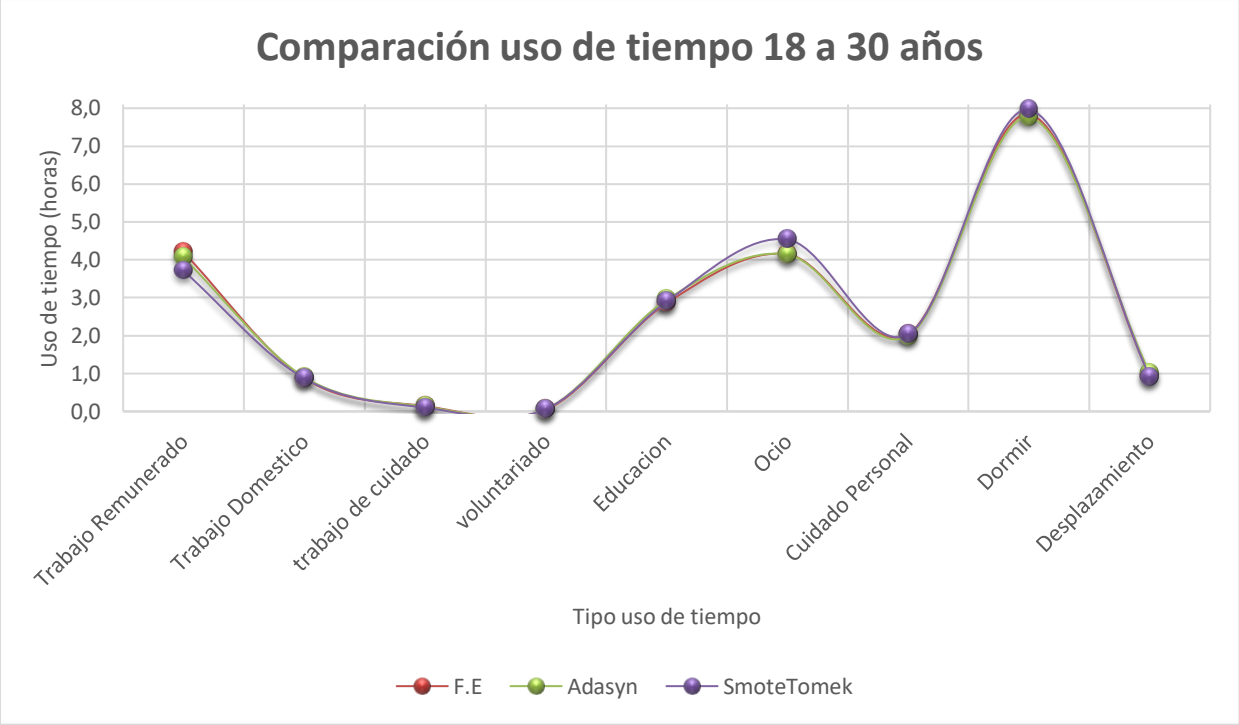
Walpole, R. E. (2012). Probabilidad y estadistica para ingenieria y ciencias. México: Pearson educación.

Weisberg, H. F. (2005). The Total Survey Error Approach. University of Chicago Press.

Capítulo 7 Anexos

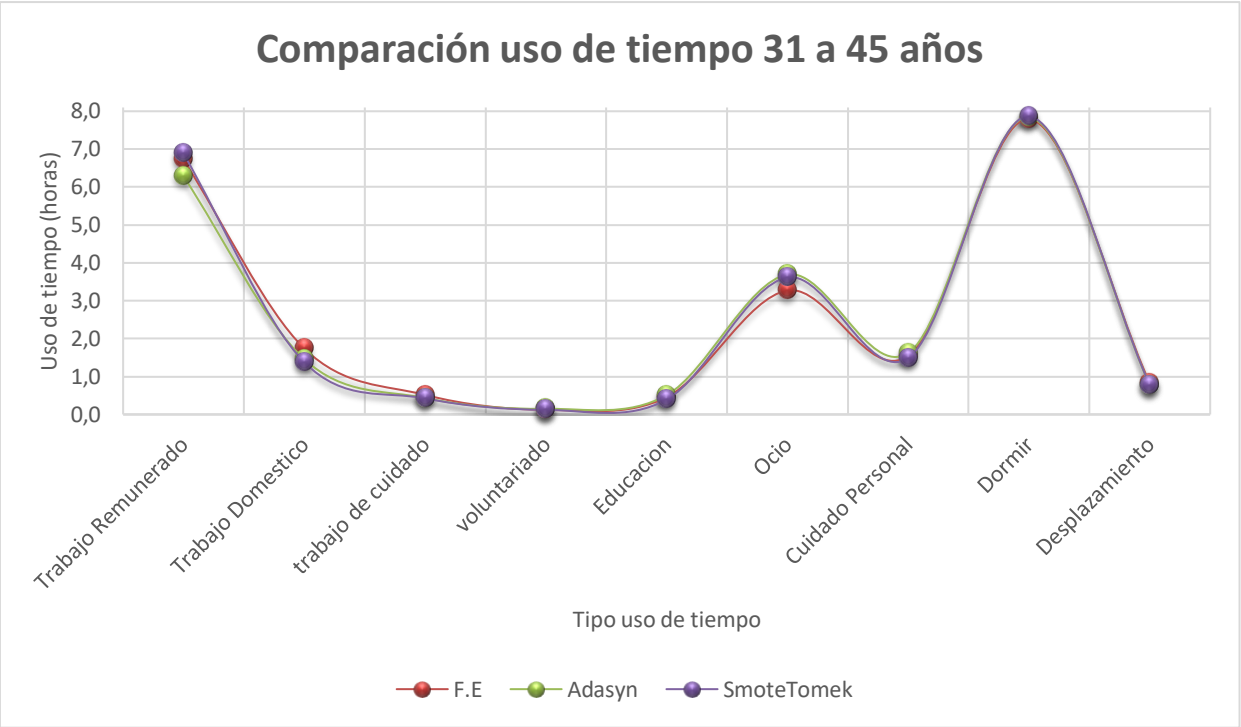
Anexo A: Link de GitHub con los códigos de los tres métodos aplicados en esta memoria de título.

https://github.com/CristobalDelgado/Codigos_MT/tree/master



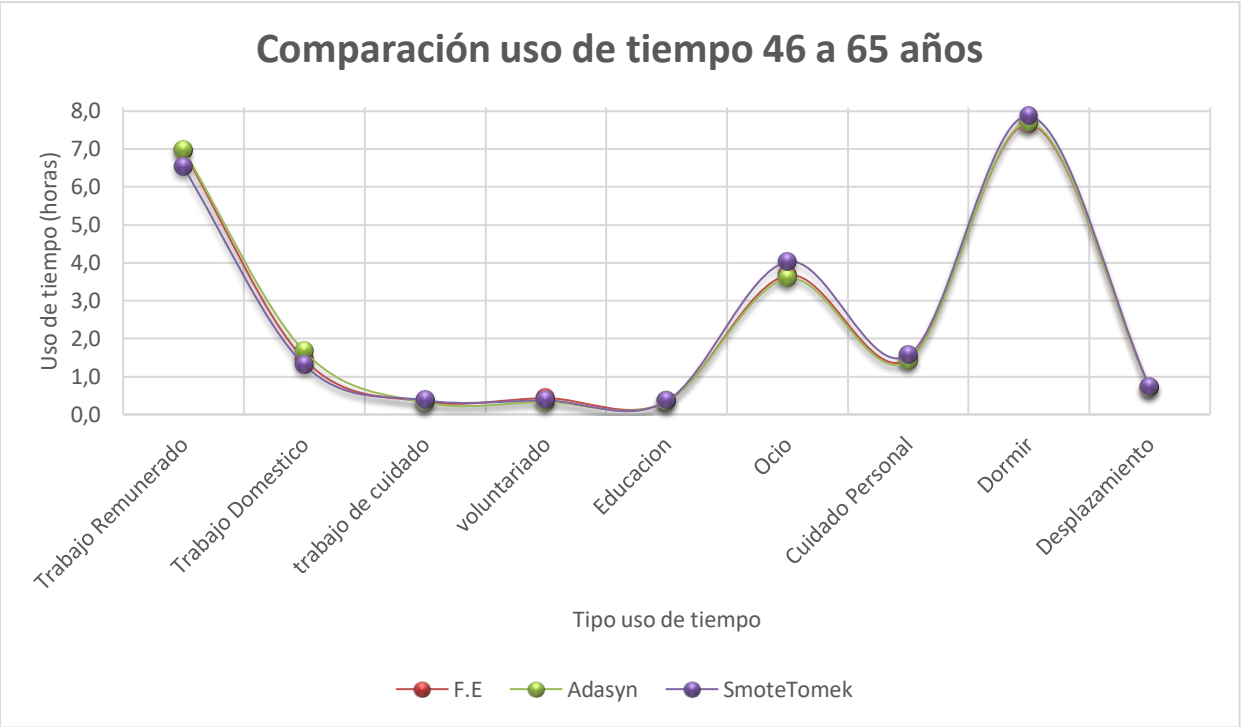
Anexo B: Comparación uso de tiempo promedio de personas de 18 a 30 años por metodología.

Fuente: elaboración propia



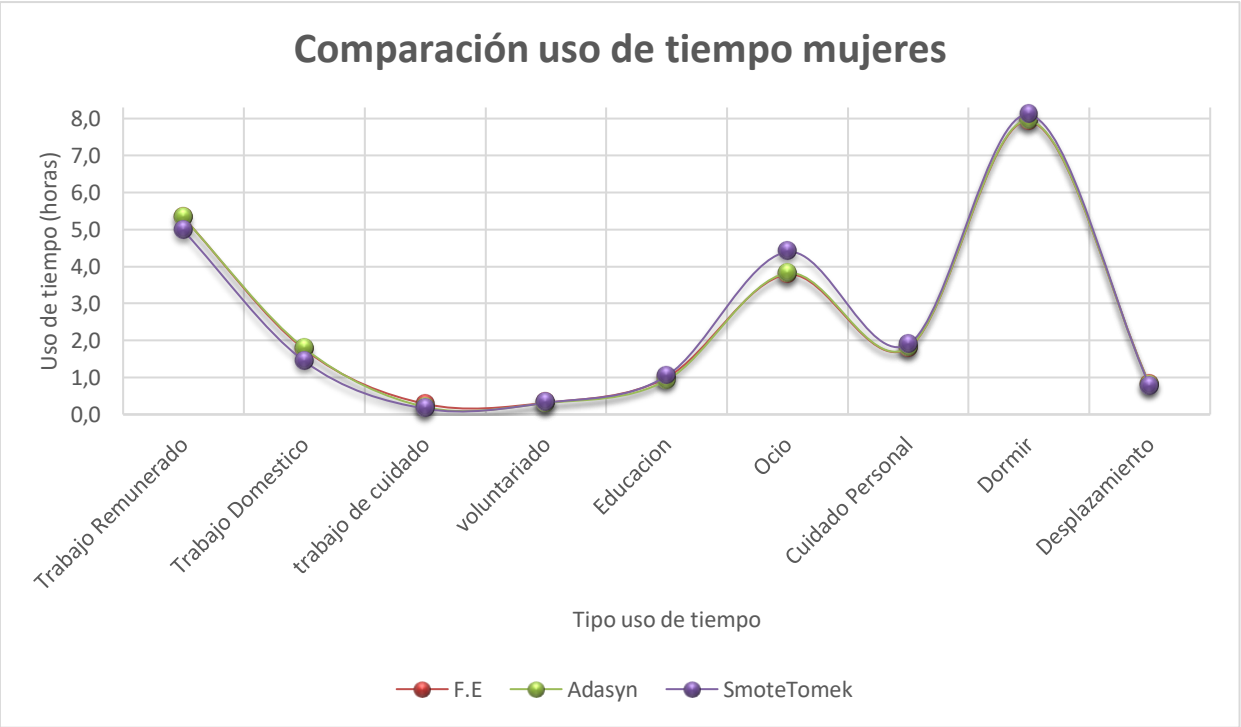
Anexo C: Comparación uso de tiempo promedio de personas de 31 a 45 años por metodología.

Fuente: elaboración propia



Anexo D: Comparación uso de tiempo promedio de personas de 46 a 65 años por metodología.

Fuente: elaboración propia



Anexo E: Comparación uso de tiempo promedio de mujeres por metodología.

Fuente: elaboración propia

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TITULO

Departamento:	Departamento de Ingeniería Industrial
Carrera:	Ingeniería Civil Industrial
Nombre del memorista:	Cristobal Alonso Delgado Escobar
Título de la memoria:	Análisis Estadístico y Exploración de Modelos para Mejorar la Representatividad Nacional en Datos de Conveniencia en Línea
Fecha de la presentación oral:	
Profesor Guía:	Sebastián Astroza Tagle, Carlos Navarrete Lizama
Profesor Revisor:	Carlos Contreras
Concepto:	
Calificación Resumen:	

Resumen

En este estudio propone analizar diversas metodologías que se puedan aplicar a una muestra no probabilística obtenida de internet con el objetivo de mejorar la representatividad de la población dentro de esta para llegar a hacer inferencias poblacionales.

Para lograr esto se debe conseguir que la distribución de la muestra sea igual o muy parecida a la de la población, en este caso se igualaron las distribuciones de los hombres y las mujeres para 3 grupos etarios que van desde los 18 años a los 30 años, 31 años a 45 año y 46 a los 65 años, se aplicaron 3 metodologías distintas, factores de expansión, un método de sobremuestreo y un método hibrido de sobremuestreo y submuestreo.

Tras haber aplicado cada una de estas metodologías se pudo concluir que los 3 métodos lograron igualar la distribución de la muestra a la de la población y entregaron usos de tiempo promedios muy similares para la población por lo que si se quiere realizar inferencias poblaciones usando una muestra no probabilística se puede utilizar 1 o los 3 métodos detallados en este informe.