



Universidad Politécnica
de Madrid



**Escuela Técnica Superior de
Ingenieros Informáticos**

Grado en Administración y Dirección de Empresas

Trabajo Fin de Grado

**Análisis de Uso del Tiempo y el LDI en
España por Comunidades Autónomas**

Autor: Bartomeu Payeras Martorell

Tutor(a): Raúl Gutiérrez Sanchis

Madrid, junio

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

Título: Título del Trabajo, con Mayúscula en Todas las Palabras que no Sean
Conectivas (Artículos, Preposiciones, Conjunciones)

Junio 2024

Autor: Bartomeu Payeras Martorell

Tutor:

Raúl Gutiérrez Sanchis

Departamento de Ingeniería de la Organización, Administración de
Empresas y Estadística

ETSI Informáticos

Universidad Politécnica de Madrid

Resumen

Este Trabajo Fin de Grado (TFG) tiene como objetivo principal realizar un análisis exhaustivo del uso del tiempo en España a nivel de Comunidades Autónomas, utilizando las Encuestas de Empleo del Tiempo (EET) de los periodos 2002-2003 y 2009-2010. Se enfoca en la creación y aplicación del *Lifestyle Dynamics Index* (LDI), un índice desarrollado para cuantificar las dinámicas de vida en diferentes regiones independientemente de la metodología de las encuestas usadas. Este índice se basa en la distribución del tiempo en diversas actividades, ofreciendo una medida sintetizada del dinamismo del estilo de vida.

El análisis se estructura en varias fases. En primer lugar, se calculó el LDI utilizando todos los niveles de categorización de actividades proporcionados por las EET. Esto permitió una evaluación detallada de la distribución del tiempo en cada comunidad y las diferencias entre los periodos estudiados. Además, se realizó una simulación del día promedio para cada comunidad y periodo, proporcionando una visualización clara de las diferencias en las rutinas diarias entre regiones y a lo largo del día.

En una segunda fase, se llevó a cabo un análisis de *clustering* para agrupar las Comunidades Autónomas en función del LDI y otros indicadores socioeconómicos relevantes. Este análisis permitió identificar patrones similares o divergentes entre regiones en términos de estilo de vida y condiciones socioeconómicas. Los resultados del estudio buscan contribuir tanto al conocimiento académico como a la formulación de políticas públicas orientadas a mejorar la calidad de vida en las diferentes regiones de España. Se espera que estos hallazgos proporcionen una comprensión profunda y holística de las dinámicas de vida en el país, destacando la importancia de las diferencias regionales y su impacto en el desarrollo socioeconómico.

El primer objetivo del TFG es analizar exhaustivamente los datos de uso del tiempo y el LDI a nivel regional en España, desglosando la información por Comunidades Autónomas. Esto incluye identificar patrones y tendencias en las dinámicas de estilo de vida a nivel regional. Otro objetivo es realizar comparativas detalladas entre comunidades autónomas en términos de distribución del tiempo, actividades cotidianas, participación en el mercado laboral y otros indicadores relevantes. Estas comparaciones buscan sensibilizar sobre la importancia de comprender las dinámicas regionales y sus efectos en el logro de los Objetivos de Desarrollo Sostenible (ODS), destacando la necesidad de políticas y acciones específicas para abordar los desafíos únicos de cada autonomía.

Finalmente, el estudio pretende relacionar los resultados del análisis con los ODS 8 (Trabajo decente y crecimiento económico), 10 (Reducción de las desigualdades) y 11 (Ciudades y comunidades sostenibles), identificando las disparidades regionales para avanzar hacia un desarrollo equitativo y sostenible.

Abstract

This Final Degree Project (TFG) primarily aims to conduct an exhaustive analysis of time use in Spain at the level of Autonomous Communities, using the Time Use Surveys (TUS) from the periods 2002-2003 and 2009-2010. It focuses on the creation and application of the Lifestyle Dynamics Index (LDI), an index developed to quantify lifestyle dynamics in different regions regardless of the survey methodologies used. This index is based on the distribution of time across various activities, offering a synthesized measure of lifestyle dynamism.

The analysis is structured in several phases. First, the LDI was calculated using all levels of activity categorization provided by the TUS. This allowed a detailed evaluation of the time distribution in each community and the differences between the studied periods. Additionally, a simulation of the average day for each community and period was conducted, providing a clear visualization of the differences in daily routines between regions and over the day.

In a second phase, a clustering analysis was carried out to group the Autonomous Communities based on the LDI and other relevant socioeconomic indicators. This analysis enabled the identification of similar or divergent patterns between regions in terms of lifestyle and socioeconomic conditions. The study's results aim to contribute both to academic knowledge and the formulation of public policies aimed at improving the quality of life in different regions of Spain. These findings are expected to provide a deep and holistic understanding of the lifestyle dynamics in the country, highlighting the importance of regional differences and their impact on socioeconomic development.

The first objective of the TFG is to comprehensively analyse the time use data and the LDI at the regional level in Spain, breaking down the information by Autonomous Communities. This includes identifying patterns and trends in lifestyle dynamics at the regional level. Another objective is to make detailed comparisons between autonomous communities in terms of time distribution, daily activities, labour market participation, and other relevant indicators. These comparisons aim to raise awareness of the importance of understanding regional dynamics and their effects on achieving the Sustainable Development Goals (SDGs), emphasizing the need for specific policies and actions to address the unique challenges of each region.

Finally, the study aims to relate the analysis results to SDGs 8 (Decent Work and Economic Growth), 10 (Reduced Inequalities), and 11 (Sustainable Cities and Communities), identifying regional disparities to advance towards equitable and sustainable development.

Tabla de contenidos

1	Introducción	1
1.1	Objetivos del TFG.....	2
2	Desarrollo	3
2.1	Trabajos previos.....	3
2.1.1	<i>Lifestyle Dynamics Index</i> (LDI)	10
2.2	Datos usados y metodologías de análisis	12
2.2.1	Metodologías en las Encuestas de Empleo del Tiempo.....	12
2.2.1.1	Formato de los microdatos de las EET del INE	15
2.2.2	Indicadores socioeconómicos utilizados en el <i>clustering</i>	19
2.2.3	Metodología y tecnologías para el análisis de los datos	19
2.2.3.1	Cálculo del LDI.....	21
2.2.3.2	Implementación del algoritmo de <i>clustering</i> K-Means	22
2.2.3.3	Interpretabilidad de los resultados del <i>clustering</i>	25
2.3	Análisis de los resultados.....	27
2.3.1	Análisis exploratorio del día promedio en las Comunidades Autónomas.....	27
2.3.2	Resultados del <i>Lifestyle Dynamics Index</i>	31
2.3.2.1	Variabilidad de los resultados del LDI con distintas metodologías	33
2.3.3	Clustering del LDI con otros indicadores socioeconómicos	37
2.3.4	Resultados para el año 2003.....	37
2.3.5	Resultados para el año 2010.....	40
3	Resultados y conclusiones	44
4	Análisis de Impacto.....	45
5	Bibliografía	46
6	Anexos	49
6.1	Anexo I: Datos en bruto del clustering	49
6.2	Anexo II: Valores promedio de cada clúster	50

Índice de figuras

Figura 2.1.1: Tradicionalmente, dados dos productos Y1 e Y2 los individuos en busca de maximizar su utilidad escogen la combinación de Y1 e Y2 en el punto donde su restricción presupuestaria es tangente a las curvas de indiferencia. Para dos restricciones presupuestarias A y A* preferirán A* ya que la utilidad es mayor que en A	6
Figura 2.2.1: Interfaz de la aplicación austriaca STAT_time.....	13
Figura 2.2.2: Extracto del Diario de Empleo del Tiempo	14
Figura 2.2.3: Registro del fichero DHOGAR, periodo 2009-2010.....	17
Figura 2.2.4: Registro del fichero CH, periodo 2002-2003	17
Figura 2.2.5: Registro del fichero DIARIO2, periodo 2009-2010.....	18
Figura 2.2.6: Fragmento de un registro del fichero DI, periodo 2002-2003 ...	18
Figura 2.2.7: Diagrama global de la ETL	21
Figura 2.2.8: Clústeres generados por los algoritmo incluidos en “scikit-learn” sobre distintos conjuntos de datos.....	23
Figura 2.2.9: Aplicación del método del codo sobre los datos del 2010	23
Figura 2.2.10: Aplicación del método de análisis de la silueta sobre los datos de 2003.....	24
Figura 2.2.11: Gráfico de fuerza para Andalucía y el clúster 0 en el año 2010	26
Figura 2.3.1: Porcentaje de personas que realizan una actividad en el transcurso del día para el periodo 2002-2003.....	28
Figura 2.3.2: Porcentaje de personas que realizan una actividad en el transcurso del día para el periodo 2009-2010.....	30
Figura 2.3.3: Resultados del LDI para el periodo 2002-2003	31
Figura 2.3.4: Resultados del LDI para el periodo 2009-2010	32
Figura 2.3.5: Evolución del LDI por Comunidades Autónomas	32
Figura 2.3.6: Comparación de los resultados del LDI para el periodo 2002-2003	33
Figura 2.3.7: Comparación de los resultados del LDI para el periodo 2009-2010	33
Figura 2.3.8: Distribución de los encuestados en el ranking para cada nivel de granularidad de las actividades para el periodo 2002-2003	34
Figura 2.3.9: Distribución de los encuestados en el ranking para cada nivel de granularidad de las actividades para el periodo 2009-2010	34
Figura 2.3.10: Mapa de los clústeres para los datos del 2003.....	37
Figura 2.3.11: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 0 del año 2003.....	38
Figura 2.3.12: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 1 del año 2003.....	39
Figura 2.3.13: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 2 del año 2003.....	40
Figura 2.3.14: Mapa de los clústeres para los datos del 2010.....	40
Figura 2.3.15: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 0 del año 2010.....	41
Figura 2.3.16: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 1 del año 2010.....	42
Figura 2.3.17: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 2 del año 2010.....	42

Índice de ecuaciones

Ecuación 1: Función de utilidad	4
Ecuación 2: Restricción presupuestaria	4
Ecuación 3: Definición de bien - <i>commodity</i> - de Becker	5
Ecuación 4: Función de utilidad de Becker	5
Ecuación 5: Restricción presupuestaria	5
Ecuación 6: Restricción temporal.....	5
Ecuación 7: Forma equivalente de la Ecuación 3	5
Ecuación 8: Unificación de la restricción presupuestaria y temporal	5
Ecuación 9: Terna de actividades y transiciones entre ellas.....	11
Ecuación 10: Fórmula LDI.....	11
Ecuación 11: Función de minimización de la inercia	24

Índice de tablas

Tabla 2.1.1: Número de diarios de uso del tiempo publicados por país y periodo	7
Tabla 2.2.1: Extracto del Diseño de Registro correspondiente al fichero DIARIO2.TXT.....	16
Tabla 2.2.2: Ejemplo de ranking del LDI para 2 actividades y 4 intervalos de tiempo	21
Tabla 6.1.1: Datos en bruto para el <i>clustering</i> del año 2003	49
Tabla 6.1.2: Datos en bruto para el <i>clustering</i> del año 2010	50
Tabla 6.2.1: Valores promedio de cada clúster para cada indicador en 2003	50
Tabla 6.2.2: Valores promedio de cada clúster para cada indicador en 2010	51

Glosario

- **ATUS:** *American Time Use Surveys*, Encuestas de Empleo del Tiempo Americanas.
- **Clustering:** técnica de análisis de datos que tiene como objetivo agrupar un conjunto de datos en subconjuntos homogéneos, denominados clústeres, se busca minimizar las diferencias dentro de cada clúster mientras se maximiza las diferencias entre los clústeres.
- **CTUR:** *Centre for Time Use Research*, Centro para Investigaciones de Empleo del Tiempo.
- **EET:** Encuestas de Empleo del Tiempo.
- **EPA:** Encuesta de Población Activa.
- **ETL:** *Extract, Transform, Load*; proceso en la gestión de datos que implica extraer datos de diferentes fuentes, transformarlos en un formato adecuado y cargarlos en un sistema de almacenamiento o base de datos.
- **Feature importance:** técnica utilizada en *machine learning* que mide la contribución relativa de cada variable en la predicción o clasificación realizada por un modelo, ayudando a interpretar y explicar los resultados del modelo.
- **FEDEA:** Fundación de Estudios de Economía Aplicada.
- **Full income:** concepto económico que incluye tanto el ingreso monetario como el valor del tiempo libre y del trabajo no remunerado, proporcionando una medida más completa del bienestar económico de una persona o un hogar.
- **Full price:** coste total de un bien o servicio, que incluye tanto el precio monetario como el valor del tiempo necesario para adquirir y utilizar dicho bien o servicio.
- **HETUS:** *Harmonised European Time Use Surveys*, Encuestas de Empleo del Tiempo Europeas Armonizadas.
- **ICATUS:** *International Classification of Activities for Time-Use Statistics*, Clasificación Internacional de Actividades para Encuestas de Empleo del Tiempo.
- **IMC:** Índice de Masa Corporal.
- **INE:** Instituto Nacional de Estadística.
- **IPC:** Índice de Precios al Consumo.
- **K-Means:** algoritmo de clustering utilizado en *machine learning* para dividir un conjunto de datos en un número específico de clústeres, asignando cada dato al clúster con el centroide más cercano.
- **LDI:** *Lifestyle Dynamics Index*.
- **Lighthouse:** plataforma central del proyecto Economics @Intelligence, cuyo objetivo es aglutinar todo tipo de estadísticas e indicadores procedentes de distintos organismos oficiales.
- **Machine Learning:** subcampo de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para permitir que las computadoras aprendan de datos y realicen predicciones o decisiones sin ser programadas explícitamente para ello.
- **Microdatos:** término utilizado por el INE para referirse a los datos desagregados y anonimizados de todas las encuestas relacionadas con un indicador.
- **MTUS:** *Multinational Time Use Study*, Estudios Multinacionales del Empleo del Tiempo.
- **ODS:** Objetivos de Desarrollo Sostenible.

- **PIB:** Producto Interior Bruto.
- **Productividad doméstica:** relacionada con la adopción de electrométricos, hace referencia a la eficiencia y eficacia con la que se realizan las tareas y actividades no remuneradas en el hogar, contribuyendo al bienestar general de los miembros del hogar.
- **SHAP:** *SHapley Additive exPlanations*, método basado en los valores de Shapley provenientes de la teoría de juegos cooperativos y que proporciona una forma de descomponer la salida de un modelo de *machine learning* en contribuciones individuales de cada característica, lo que permite entender cómo cada característica contribuye a la predicción final.
- **Taylorismo:** enfoque de gestión científica desarrollado por Frederick Taylor que optimiza la eficiencia laboral mediante la división del trabajo en tareas simples y repetitivas, aumentando la productividad y controlando el tiempo de trabajo.
- **Tidy data:** principio de la ciencia de datos que establece que cada variable debe tener su propia columna, cada observación su propia fila y cada valor su propia celda, facilitando el análisis y la manipulación de los datos.
- **Time poverty:** situación en la que una persona no dispone de suficiente tiempo libre después de cumplir con sus obligaciones laborales, domésticas y personales, afectando negativamente su bienestar y calidad de vida.
- **TUS:** *Time Use Surveys*.
- **UNESCO:** *United Nations Educational, Scientific, and Cultural Organization*, Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.
- **URSS:** Unión de Repúblicas Socialistas Soviéticas.
- **Wellbeing:** estado de satisfacción y buena salud física, mental y emocional de una persona, abarcando factores como la felicidad, la calidad de vida y el equilibrio entre el trabajo y la vida personal.

1 Introducción

Las Encuestas de Empleo del Tiempo (EET), también conocidas como *Time Use Surveys* (TUS) en inglés, representan una herramienta fundamental para comprender cómo las personas distribuyen y emplean su tiempo en actividades diarias. Estas encuestas ofrecen una visión detallada de los patrones de vida, las rutinas y las dinámicas sociales de una población, proporcionando datos valiosos para el análisis y la formulación de políticas públicas.

En el contexto de España, las EET han sido objeto de interés creciente tanto para académicos como para responsables de políticas. Los datos recopilados a través de estas encuestas ofrecen una ventana única para comprender la evolución de los estilos de vida a lo largo del tiempo y entre diferentes regiones del país. En este sentido, el presente Trabajo de Fin de Grado (TFG) se propone realizar un análisis exhaustivo de las EET en España durante los periodos 2002-2003 y 2009-2010, centrándose específicamente en el nivel de las Comunidades Autónomas.

El principal objetivo de este estudio es desarrollar un enfoque integral para analizar los datos de las EET, con un énfasis particular en la creación y aplicación del *Lifestyle Dynamics Index* (LDI). Este índice, desarrollado por los profesores de la Universidad Politécnica de Madrid y la Universidad de Alicante Raúl G. Sanchis y J. Vicente-Pérez, tiene como objetivo cuantificar y comparar las dinámicas de vida entre diferentes regiones, proporcionando una medida sintetizada del estilo de vida en función de la distribución del tiempo en diversas actividades [1].

Para lograr este objetivo, se llevarán a cabo varios análisis. En primer lugar, se calculará el LDI utilizando todos los niveles de categorización de actividades proporcionados por las EET. Este enfoque permitirá una evaluación detallada de cómo se distribuye el tiempo en cada comunidad autónoma y cómo estas distribuciones difieren entre los dos periodos estudiados.

Además del cálculo del LDI, se realizará una simulación del día promedio para cada comunidad autónoma y periodo, con el fin de visualizar de manera clara y concisa las diferencias en las rutinas diarias entre regiones y a lo largo del tiempo. Esta simulación proporcionará una perspectiva intuitiva de cómo las personas dedican su tiempo en actividades específicas en diferentes contextos geográficos y temporales.

En un segundo nivel de análisis, se llevarán a cabo un análisis de *clustering* que consistirá en agrupar las comunidades autónomas en función del LDI y otros indicadores relevantes, como el Producto Interno Bruto (PIB) o la tasa de desempleo, con el objetivo de identificar patrones similares o divergentes entre regiones en términos de estilo de vida y condiciones socioeconómicas.

En resumen, este TFG busca ofrecer una comprensión profunda y holística de las dinámicas de vida en España a través del análisis de las Encuestas de Empleo del Tiempo a nivel de las Comunidades Autónomas. Se espera que los resultados obtenidos contribuyan tanto al avance del conocimiento académico en el campo del estudio del tiempo y las actividades diarias como a la formulación de políticas públicas orientadas a mejorar la calidad de vida de la población en diferentes regiones del país.

1.1 Objetivos del TFG

De forma más concreta, los objetivos finales del Trabajo de Fin de Grado son:

1. Analizar exhaustivamente los datos de uso del tiempo y el LDI a nivel regional en España, desglosando la información por comunidades autónomas.
2. Identificar patrones y tendencias en las dinámicas de estilo de vida a nivel regional, prestando especial atención a factores propios y desafíos socioeconómicos presentes en diferentes regiones (despoblación de áreas rurales, diferencias por nivel de renta o cualquier patrón destacable).
3. Realizar comparativas detalladas entre las comunidades autónomas españolas en términos de distribución del tiempo, actividades cotidianas, participación en el mercado laboral y otros indicadores relevantes como el desarrollo económico de la región.
4. Sensibilizar sobre la importancia de comprender las dinámicas regionales y sus efectos en el logro de los ODS, destacando la necesidad de políticas y acciones específicas para abordar los desafíos únicos enfrentados por cada autonomía.
5. Relacionar los resultados del análisis con el ODS 8 (Trabajo decente y crecimiento económico), el ODS 10 (Reducción de las desigualdades) y el ODS 11 (Ciudades y comunidades sostenibles), identificando y abordando las disparidades regionales hacia el avance a un desarrollo equitativo y sostenible.

2 Desarrollo

2.1 Trabajos previos

Antes de entrar en el análisis detallado de las Encuestas de Empleo del Tiempo (EET) en España y el cálculo del Lifestyle Dynamics Index (LDI) a nivel de las Comunidades Autónomas, es crucial contextualizar este trabajo dentro del marco de investigaciones previas en este campo. A lo largo de los años, numerosos estudios han explorado las dinámicas de vida en diferentes contextos geográficos y temporales. Desde análisis sociológicos hasta investigaciones económicas, la atención se ha centrado en comprender cómo las personas asignan sus recursos temporales y cómo estas elecciones afectan su bienestar y el desarrollo socioeconómico de las regiones. En esta sección, revisaremos con una perspectiva histórica los trabajos previos relacionados con el análisis del uso del tiempo, estableciendo un marco conceptual a nivel español e internacional.

La historia de la investigación sobre el uso del tiempo tiene raíces profundas que se remontan a estudios pioneros realizados antes de 1900 por investigadores rusos del *zemstvo* -gobiernos locales creados durante las reformas liberales del Imperio ruso en la época del zar Alejandro II [2]-, que observaron las actividades de los hogares campesinos. Estos estudios se centraron en describir las condiciones sociales, monitorear la productividad económica y proporcionar información sobre la fuerza laboral [3]. En la década de 1910, la socialista y sufragista británica Maud Pember-Reeve llevó a cabo una investigación que involucró a mujeres londinenses para completar diarios de uso del tiempo, con el propósito de demostrar las desigualdades en la distribución de alimentos relacionadas con la pobreza y la alta mortalidad infantil [4]. Ya en la década de 1920, y de forma casi simultánea, el economista y estadista Stanislav Strumlin en la Unión Soviética y el Departamento de Agricultura de Estados Unidos comenzaron a recolectar diarios de tiempo [5] [6], proporcionando datos tempranos para investigaciones posteriores.

El interés soviético en la asignación del tiempo fue impulsado por la idea de una economía central planificada marxista, donde el valor de los objetos se determinaba por el tiempo de trabajo necesario que incorporaban [7]. En la década de 1920, los soviéticos se interesaron en el taylorismo surgido en Estados Unidos [7], que tenía como objetivo aumentar la productividad mediante la atomización de las tareas del proceso productivo y establecer tiempos predefinidos y cronómetros para racionalizar la producción [8]. A pesar de que la producción en cadena es un invento que comúnmente asociamos al capitalismo más puro, maximización de los beneficios empresariales y alienación del trabajador, dentro de la URSS fue aplicada de igual forma y a nivel teórico su implantación se justificaba con la visión marxista de que en el futuro la tecnología permitiría reducir el tiempo dedicado al trabajo liberando al trabajador de horas de trabajo aumentando su tiempo ocioso y en consecuencia su bienestar [9]. No obstante, es importante subrayar que, si bien hubo un aumento en la producción, este no se tradujo necesariamente en un incremento del tiempo de ocio real para los trabajadores.

Por otro lado, en Estados Unidos el taylorismo, el aumento en la productividad industrial y con ello, el abaratamiento de los productos con cierta sofisticación tecnológica provocó una revolución en la “productividad doméstica” a la mitad del siglo XX, con la introducción de aparatos que permitían ahorrar tiempo en

los hogares (lavadoras, frigoríficos u otros electrodomésticos), lo cual sí llevó a cambios inesperados en el uso del tiempo [7]. En Estados Unidos, los estudios de uso del tiempo se centraron en el papel de las mujeres en la sociedad y los avances en la economía doméstica, y en el Reino Unido, organizaciones como *Mass Observation* y el Departamento de Investigación de Audiencia de la BBC llevaron a cabo estudios similares [7]. A medida que avanzaba el siglo XX, los estudios de uso del tiempo se extendieron a otros países, con grandes estudios basados en diarios en Checoslovaquia, Francia, Hungría, Polonia, Japón y el Reino Unido.

Pero no es hasta la década de 1960 cuando se realizó el primer estudio transnacional armonizado de uso del tiempo para 12 países, financiado por la UNESCO y dirigido por el sociólogo húngaro Alexander Szalai [10]. Este estudio sentó las bases para futuras investigaciones comparativas y estableció un diseño común para los estudios de diarios de tiempo. Pero no es hasta 1965, con la publicación de “*A Theory of the Allocation of Time*” del premio Nobel de economía Gary Becker [11], cuando se sientan las bases matemáticas y económicas del estudio del uso del tiempo. Becker proporcionó un marco teórico basado en la incorporación del factor tiempo a las funciones de utilidad y de la demanda tradicionales, enmarcando su seminal contribución dentro de la corriente “*New Consumer Behavior*”. Esta corriente comenzó con la tesis doctoral de Ironmonger (1959) defendida en la University of Cambridge, y publicada posteriormente como libro en 1972. El modelo económico desarrollado por Becker y otros antecesores se basa fundamentalmente en el trabajo de Ironmonger en su tesis doctoral [12].

Tradicionalmente en microeconomía se entiende que los hogares o los individuos organizan su consumo de forma que tratan de maximizar la Ecuación 1, donde U es la utilidad que obtienen de los y_i bienes comprados en el mercado [13]. Por otro lado, los hogares están sujetos a la restricción de la Ecuación 2, donde p'_i son los precios de esos bienes, I los ingresos, W el sueldo y V otros ingresos no procedentes del trabajo; por lo que se deduce de dicha ecuación que el consumo estará limitado por la dotación presupuestaria de ese hogar.

$$U = U(y_1, y_2, \dots, y_n)$$

Ecuación 1: Función de utilidad. Fuente: [11]

$$\sum p'_i y_i = I = W + V$$

Ecuación 2: Restricción presupuestaria. Fuente: [11]

Becker propone que las decisiones sobre la asignación del tiempo están sujetas a las mismas consideraciones económicas que las decisiones sobre el uso de otros recursos, como el dinero. Por ello, redefine el concepto tradicional de bienes de consumo -*commodities*, llamados *wants* por Ironmonger, y anteriormente por John S. Mill en su “*Utilitarianism*” [14]- (en adelante bienes) como la Ecuación 3, una función dependiente de un vector de bienes de mercado, x_i , y un vector de tiempo, T_i , donde cada elemento del vector de tiempo se corresponde a periodos de tiempo con distinta valoración para el individuo, dado que el autor contempla que los individuos valorarán de forma distinta las horas de la mañana con respecto a las de la noche o los días de entresemana con respecto a los fines de semana [11]. Nótese que f_i sigue la forma de una función de producción incorporando el tiempo como input de dicha función además de las materias primas -bienes de mercado en este caso-, y es que para

Becker los hogares son tanto ofertantes como maximizadores de utilidad, o consumidores, maximizando en este caso la Ecuación 4.

$$Z_i = f_i(x_i, T_i)$$

Ecuación 3: Definición de bien *-commodity-* de Becker. Fuente: [11]

$$U = U(Z_1, \dots, Z_m) \equiv U(f_1, \dots, f_m) \equiv U(x_1, \dots, x_m; T_1, \dots, T_m)$$

Ecuación 4: Función de utilidad de Becker. Fuente: [11]

Otra conclusión importante que podemos sacar de esta definición de bien es que se asemeja más conceptualmente a lo que todos entenderíamos por una actividad más que a un bien de consumo. Por ejemplo, un bien en términos de Becker podría ser, leer un libro donde x_i sería el coste en términos monetarios del libro y T_i el tiempo que el individuo lee ese libro. Por otra parte, difumina la tradicional distinción entre empresas y hogares, definiendo en cierta forma a los hogares como “miniempresas” donde se combinan factores de producción (bienes de capital, materias primas y mano de obra) para la producción de otros bienes [11]. Por ejemplo, cocinarse la cena podría definirse como la combinación del uso de los electrodomésticos (bienes de capital), los ingredientes (materias primas) y el tiempo empleado.

La Ecuación 4 está a su vez sujeta a las restricciones de la Ecuación 5 y la Ecuación 6, donde T_w es el tiempo de trabajo, $\bar{w}$ el salario por unidad de tiempo y T_c el tiempo de consumo [11]; es decir, la Ecuación 5 nos indica que el precio de los bienes de mercado consumidos no puede superar la dotación presupuestaria y la Ecuación 6 nos indica que el tiempo de consumo será igual a la sustracción del tiempo de trabajo al total del tiempo disponible. En realidad, Becker identifica que las dos restricciones se pueden resumir en una única restricción (véase la Ecuación 8) ya que un individuo puede reducir su tiempo de ocio o consumo para trabajar más, lo que terminara resultando en una mayor dotación presupuestaria [11].

$$\sum_1^m p_i x_i = I = V + T_w \bar{w}$$

Ecuación 5: Restricción presupuestaria. Fuente: [11]

$$\sum_1^m T_i = T_c = T - T_w$$

Ecuación 6: Restricción temporal. Fuente: [11]

$$\begin{cases} T_i \equiv t_i Z_i \\ x_i \equiv b_i Z_i \end{cases}$$

Ecuación 7: Forma equivalente de la Ecuación 3. Fuente: [11]

$$\begin{aligned} \sum (p_i b_i + t_i \bar{w}) Z_i &= V + T \bar{w} \\ \begin{cases} \pi_i \equiv p_i b_i + t_i \bar{w} \\ S' \equiv V + T \bar{w} \end{cases} \end{aligned}$$

Ecuación 8: Unificación de la restricción presupuestaria y temporal. Fuente: [11]

A partir de la Ecuación 8, donde b_i y t_i son los vectores coeficientes que indican el peso de los bienes de mercado y el tiempo en la realización de una actividad Z_i , se deducen aspectos importantes. La primera, es que π_i es el coste o precio total (*full price*, en los términos de Becker) de realizar la actividad Z_i [11], donde el primero de los sumandos ($p_i b_i$) se corresponde al precio de los bienes de mercado utilizados y el segundo de los sumandos ($t_i \bar{w}$) es el coste de oportunidad de estar realizando esa actividad en lugar de estar trabajando. La segunda parte de la restricción (S') no es otra cosa que la dotación presupuestaria si todo el tiempo del que dispone el individuo se valora a su remuneración del trabajo por unidad de tiempo ($T\bar{w}$ o *full income*), sumándole los ingresos pasivos (V).

Finalmente, si nos abstraemos de la idea de tiempo de trabajo y tiempo de ocio sin hacer distinción entre ellos, y entendemos al individuo como alguien que lo que busca es maximizar su utilidad (Ecuación 4) del uso de su tiempo independientemente de si es tiempo de trabajo o de ocio, y sujeto a ciertas restricciones de tiempo y dotación presupuestaria, podemos dar explicación a hechos que suceden de forma habitual donde hay gente que renuncia a trabajos con mayores salarios que le resultan insatisfactorios por otros en los que creen que van a ser más felices, a pesar de tener remuneraciones inferiores, pero que maximizan la utilidad global en términos de Becker; es decir, el trabajo peor remunerado como actividad puede ser más “placentero” para el individuo. Esto es algo que no tendría encaje dentro de la concepción de utilidad tradicional donde el individuo siempre tendería a escoger el trabajo que estuviese mejor remunerado ya que le permitiría desplazar su restricción presupuestaria y aumentar su utilidad (mayor capacidad de consumo) (véase Figura 2.1.1).

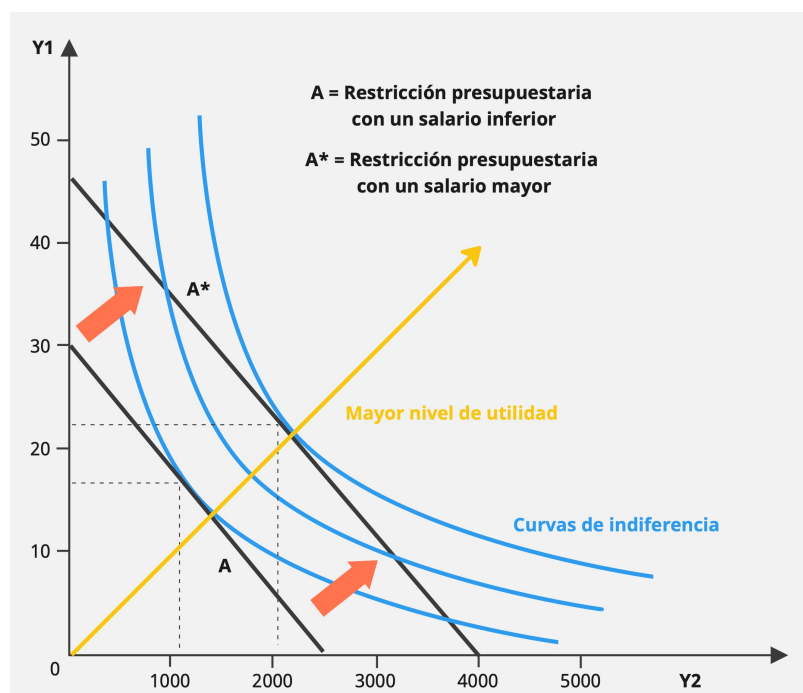


Figura 2.1.1: Tradicionalmente, dados dos productos Y1 e Y2 los individuos en busca de maximizar su utilidad escogen la combinación de Y1 e Y2 en el punto donde su restricción presupuestaria es tangente a las curvas de indiferencia. Para dos restricciones presupuestarias A y A* preferirán A* ya que la utilidad es mayor que en A. Elaboración propia.

En resumen, una de las ideas centrales del artículo de Becker es el concepto de “coste de oportunidad del tiempo”, que se refiere al valor del tiempo empleado

en una actividad en términos de lo que se podría haber hecho en su lugar. Becker argumenta que las personas asignan su tiempo de manera que maximizan su bienestar subjetivo, equilibrando los beneficios y costos de diferentes actividades [11]. Además, el autor concluye su *paper* instando a las instituciones a recolectar datos no solo del gasto doméstico sino también de “gasto de tiempo” para que se tuviera una mejor visión de cómo se reparte el *full income* en los países [11].

Durante los años posteriores del siglo XX y XXI en todo el mundo se empiezan a publicar desde la academia estudios y propuestas relacionadas con el empleo del tiempo y su cuantificación económica; mientras que desde los institutos de estadística oficiales se van publicando los primeros informes y encuestas relacionadas con los diarios del uso del tiempo [7], como se puede ver en la Tabla 2.1.1. No obstante, en España no es hasta principios del siglo XXI cuando se publican los primeros datos del INE referentes a esta cuestión (véase Tabla 2.1.1), aunque en años anteriores desde el organismo homólogo vasco se publicaron las Encuesta de Presupuestos de Tiempo en el País Vasco [15].

Tabla 2.1.1: Número de diarios de uso del tiempo publicados por país y periodo. Fuente: [7]

País	1961-6	1971-5	1979-81	1983-7	1989-92
Austria					25.233
Australia		1.491		3.181	13.806
Bélgica	2.085				
Bulgaria	2.096				
Canadá		2.138	8.727	9.618	25.233
República Checa	2.211				
Alemania	5.078				25.812
Dinamarca	4.174			3.561	
España					
Finlandia			12.038	15.184	
Francia	2.791	4.634			
Hungría					
Israel					4.843
Italia			2.118		38.110
Corea del Sur					
Países Bajos		9.163	19.110	22.841	23.905
Noruega		6.516	6.066		6.129
Perú	777				
Polonia	2.740				
Serbia	1993				
Suecia					7.065

Eslovenia	2.120				
Reino Unido	9.292	20.252		16.316	
Estados Unidos	2.017	7.088		3.339	
Sudáfrica					
Total	37.374	51.282	48.059	48.059	74.040

País	1995-99	1999-04	2005-09	2010-15	Total
Austria					25.233
Australia	14.315		13.617		46.410
Bélgica	2.085				2.085
Bulgaria					2.096
Canadá	10.726		19.597	15.390	91.429
República Checa					2.211
Alemania		35.813			66.703
Dinamarca		6.617			14.352
España		51.813	19.295		71.108
Finlandia	10.074		7.480		44.776
Francia	15.441		27.903		50.769
Hungría			8.391		8.391
Israel					4.843
Italia		51.206	40.940		132.374
Corea del Sur	85.906		40.526		126.432
Países Bajos	22.589	12.691	15.428		125.727
Noruega		7.669			26.380
Perú					777
Polonia		40.292		76.656	119.688
Serbia					1.993
Suecia		7.727			14.792
Eslovenia		12.276			14.396
Reino Unido		20.982		13.538	80.380
Estados Unidos		34.693	64.085	38.182	149.404
Sudáfrica		14.392			14.302
Total	159.051	296.081	257.262	143.766	1.237.051

Ante el progresivo aumento de la publicación de estadísticas y encuestas relacionadas con el uso del tiempo en distintos países, cada uno de ellos con su metodología, con sus propias categorizaciones de las actividades y cuestionarios,

incluso con categorías de actividades cambiantes dependiendo del año en que se realizaba la encuesta dentro del propio país como el caso de España, surge una clara necesidad de estandarización y acuerdo entre los países en la metodología utilizada para poder establecer comparaciones de forma efectiva entre ellos y entre periodos. En este sentido los países que han sido más consistentes en la publicación de estos datos ya desde unos inicios son Canadá, Estados Unidos -a través del ATUS (*American Time Use Survey*, datos publicados por el *Bureau of Labor Statistics* [16])- , Reino Unido y Países Bajos, como se puede observar en la Tabla 2.1.1. Si hablamos de intentos de armonización debemos mencionar los tres más relevantes, *Multinational Time Use Study* (MTUS, por sus siglas en inglés), el *Harmonised European Time Use Surveys* (HETUS, por sus siglas en inglés) y, las guías de las Naciones Unidas.

El primero de ellos, el MTUS es una iniciativa mantenida por el *Centre for Time Use Research* (CTUR, por sus siglas en inglés) de la University College of London. La estandarización que llevan a cabo es *ex post*, es decir después de que la encuestas hayan sido publicadas por los organismos nacionales correspondientes [7]. Los investigadores del CTUR, aún hoy en día, para realizar dicha estandarización siguen como base la metodología utilizada por Szalai en 1972 cuando elaboraba la armonización financiada por la UNESCO [10].

En segundo lugar, el HETUS es una iniciativa de Eurostat, la oficina de estadística de la Unión Europea, en colaboración con organismos nacionales de diferentes países miembro. Esta iniciativa tiene como objetivo la armonización *ex ante* de las encuestas de uso del tiempo en toda Europa [17], lo que implica establecer una metodología común para la recopilación, análisis y presentación de datos sobre cómo las personas distribuyen y utilizan su tiempo en actividades cotidianas. La armonización *ex ante* garantiza la coherencia y comparabilidad de los datos recopilados entre países especialmente en cuanto a categorías de actividades, lo que facilita las comparaciones y el análisis de tendencias a nivel europeo.

Finalmente, la Comisión Económica de las Naciones Unidas para Europa en 2013 creó unas guías para la metodología de las encuestas de uso del tiempo que han de seguir los países y abarcan desde las categorías de actividades hasta la cantidad y formato de los cuestionarios pasando por la frecuencia con la que se deberían realizar estas encuestas [18], aunque cabe decir que estas guías no son de obligado cumplimiento para ningún país y son simplemente recomendaciones. En la elaboración de estas guías participaron miembros del CTUR y Eurostat, además de otros organismos de estadística oficiales como los de Canadá o Australia [18], por lo que las propuestas que aquí se encuentran difieren poco de las existentes en las dos iniciativas anteriores. De forma general, esta guía recomienda la elaboración de encuestas a gran escala cada 10 años y la elaboración de encuestas más pequeñas cada 3 años para evaluar en qué medida han cambiado los hábitos de la sociedad y valorar la necesidad de adelantar de la siguiente encuesta a gran escala si se observan cambios significativos [18]; mientras que para las actividades se recomienda basarse en la Clasificación Internacional de Actividades para Encuestas de Empleo del Tiempo (ICATUS, por sus siglas en inglés) elaborada en 1997 y revisada en 2012 por el grupo de expertos que elaboró las guías [18].

En los últimos años, los datos de uso del tiempo están demostrando ser útiles y de interés en numerosos campos de estudio, además de proporcionar información relevante para la definición de nuevas políticas públicas y revisar el impacto de las existentes en las dinámicas sociales. Algunas de estas áreas son:

- Estimación a nivel económico de ciertas actividades o trabajos no contemplados en la contabilidad nacional tradicional como el Producto Interior Bruto (PIB); siendo el ejemplo más habitual el trabajo realizado por las amas de casa.
- Estudios relacionados con el bienestar de las personas (*wellbeing*), que abarcan temas como la conciliación, ocio y consumo, etc.
- Otros análisis relacionados con aspectos sociales como la medición de la brecha de género dentro y fuera del mundo laboral, pobreza de tiempo (*time poverty*) o aspectos medioambientales (consumo energético, exposición a la contaminación, etc.) [18].

Si ponemos el foco de nuevo en España, las principales investigaciones y estudios se han efectuado siempre en campos relacionados con la brecha de género en el tiempo que se pasa realizando trabajos remunerados frente a los no remunerados como el cuidado de los niños o de personas dependiente, u otras tareas relacionadas con el hogar, habitualmente realizadas por las mujeres [19].

Por otra parte, son frecuentes los trabajos que comparan el uso del tiempo de las personas desempleadas frente a las empleadas. En este sentido, destaca un estudio realizado por FEDEA en 2003 que examina cómo varía el uso del tiempo entre personas empleadas y desempleadas, concluyendo que aunque las personas desempleadas tienden a asignar su tiempo a actividades como el ocio pasivo y el trabajo doméstico en mayor medida que las empleadas [15]; también destinan más tiempo a actividades asociadas al ocio activo y al cuidado de los niños, en comparación con aquellos que no participan en la fuerza laboral [15]. Si bien estas diferencias no pueden atribuirse directamente al desempleo, el estudio sugiere que ofrecen información valiosa sobre la producción doméstica y los costos asociados al desempleo [15], destacando la necesidad de investigaciones más completas para comprender mejor las decisiones de asignación del tiempo y sus implicaciones económicas.

Precisamente, ese es el gran problema de los estudios a nivel español, la falta de datos recientes -desde el 2010 no se han publicado datos del INE- hace que el número de análisis realizados en este ámbito sea cada vez menor por la dificultad de extrapolar los datos y las conclusiones obtenidas a la España actual, 14 años después. Además, si le sumamos el factor de que el INE únicamente publica resultados de forma agregada a nivel nacional y que para realizar estudios por Comunidades Autónomas como el de este TFG es necesario el procesado de los microdatos del INE, entendemos rápidamente porque no existen grandes estudios en esta área. Por ello, es importante recordar que las conclusiones y análisis realizados en este TFG son de aplicación únicamente a los periodos estudiados.

2.1.1 Lifestyle Dynamics Index (LDI)

Una parte importante de este TFG, así como el del grado en ingeniería informática, se centra en el cálculo del LDI, un índice desarrollado por Raúl G. Sanchís, tutor de este TFG, y José Vicente-Pérez, profesor en la Universidad de Alicante.

EL LDI se define como una métrica que captura la dinámica del estilo de vida basada en la información de las encuestas de uso del tiempo; entendiendo el dinamismo de un día no por la actividad física que se realice en él, sino por la variabilidad en las actividades realizadas [1]. El índice se desarrolla a partir de datos de diarios de uso del tiempo incluidos en cualquier Encuesta de Empleo

del Tiempo en el mundo, independientemente del número de actividades y franjas de tiempo utilizadas. En esencia, el LDI sintetiza la información de los cuestionarios de todos los adultos de los hogares la encuesta escogida en un único indicador numérico, respetando los criterios estadísticos originales de cada oficina nacional de estadísticas, dado que es independientemente de la metodología, siempre que se siga la estructura de diario de uso del tiempo para el cuestionario [1].

El cálculo del LDI parte de los datos de los diarios de uso del tiempo, a partir de ellos se elaboran unas ternas o vectores cuyos componentes tienen el siguiente significado:

- El número de actividades diferentes realizadas en el diario (x_d) [1].
- El número de cambios diferentes entre dos actividades distintas, es decir, las transiciones entre actividades sin repeticiones (y_d) [1].
- El número total de cambios entre actividades, incluyendo repeticiones (z_d) [1].

$$(x_d, y_d, z_d)$$

Ecuación 9: Terna de actividades y transiciones entre ellas. Fuente: [1]

De forma general una terna será superior a otra, es decir, más dinámica cuando x_d sea superior; en caso de que x_d sea igual, cuando y_d sea superior; y en caso de que tanto x_d como y_d sean iguales, cuando z_d sea inferior [1].

Una vez calculados estas ternas para todos los diarios en la muestra de la encuesta, se establece un orden que clasifica las ternas entre sí y dentro de un ranking con todas las posibles combinaciones de las ternas dadas N actividades y T el número de franjas temporales [1].

$$LDI = \frac{100}{\delta \kappa} \sum_{d \in \mathcal{D}} p_d$$

Ecuación 10: Fórmula LDI. Fuente: [1]

Finalmente, el LDI se calcula como el promedio de las posiciones de las ternas en el “ranking de dinamismo” establecido previamente [1], como se muestra en la Ecuación 10 donde p_d es la posición de la terna en el ranking; δ el número de cuestionarios, es decir, ternas; y κ el número de filas del ranking, es decir, el número de posibles ternas. Este promedio se normaliza en una escala del 0 al 100 para facilitar su interpretación y comparación entre diferentes contextos [1]. Cabe destacar que el LDI no considera la cantidad de tiempo dedicado a cada actividad, sino más bien la diversidad de las actividades realizadas en un día típico y su dinamismo, entendido como cambios entre actividades.

2.2 Datos usados y metodologías de análisis

En este apartado se presentarán tanto las metodologías seguidas por las encuestas del tiempo como las innovaciones y esfuerzos que se están llevando a cabo, por distintos organismos, en esta materia para digitalizar esas metodologías. Además, también se cubrirá y explicará la metodología utilizada para obtener los distintos análisis que componen este trabajo, partiendo de los microdatos del INE

2.2.1 Metodologías en las Encuestas de Empleo del Tiempo

Tradicionalmente la forma de recolección de datos relacionados con las encuestas de Empleo del Tiempo -TUS, por sus siglas en inglés- se realiza mediante cuestionarios en papel que son rellenados por los encuestados y que entregan posteriormente al organismo que realiza la encuesta correspondiente. Otra alternativa habitual es la recolección de datos a través de encuestas telefónicas donde se pregunta por las actividades que realizó el encuestado en el día anterior, este método es el utilizado por el ATUS [16]. En ambos casos el encargado de la transcripción de los cuestionarios o el entrevistador registra la información proporcionada ajustando las respuestas obtenidas a una de las categorías predefinidas en el marco de la encuesta. Por ejemplo, si un encuestado español respondiera que durante un intervalo de tiempo determinado de su día ha estado “jugando al tenis” esa información sería transcrita y registrada manualmente con el código 614 dentro de las categorías del Instituto Nacional de Estadística (INE), correspondiente a la actividad “Deportes con balón o pelota” [20].

Por otra parte, recientemente se están proponiendo otras metodologías mucho más innovadoras y que gracias las nuevas tecnologías son mucho más automatizadas y fiables. Muchas de estas nuevas metodologías se están proponiendo en proyectos de los distintos institutos de estadística nacionales europeos englobados en el marco del HETUS [17], una iniciativa de la Unión Europea para homogeneizar el formato de las encuestas de Empleo del Tiempo en lo que se refiere al número y tipo de actividades y a la duración de las franjas horarias utilizadas (10 minutos, 1 minuto, etc.).

Las dos principales propuestas, realizadas en 2018, en este ámbito provienen de *Statistik Austria*, el instituto de estadística austriaco, y de una colaboración entre *StatBel* y *DeStatis*, los organismos homólogos de Bélgica y Alemania, respectivamente [21]. En ambos casos el objetivo es aumentar la fiabilidad de las respuestas obtenidas por varias vías:

- Reduciendo el tiempo transcurrido entre que sucede la actividad y el encuestado registra la información, evitando así que el encuestado omita información o dé información errónea de forma inconsciente.
- En el caso de las encuestas telefónicas, evitando que ciertos entrevistados puedan omitir o dar información errónea de forma consciente sugestionados por lo que pueda pensar el entrevistador.
- En el caso de los cuestionarios rellenados en papel, eliminando la necesidad de que las actividades tengan que ser interpretadas y clasificadas por los encargados del estudio sociológico.

El primero de los proyectos, el austriaco, propone el uso de una aplicación para *smartphones* Android (véase Figura 2.2.1) donde los usuarios, es decir los

encuestados, pueden rellenar de forma instantánea, y a lo largo del día objeto de estudio, las actividades que realizan dentro de las categorías predefinidas y contestar el resto de las preguntas vinculadas con el cuestionario [22].

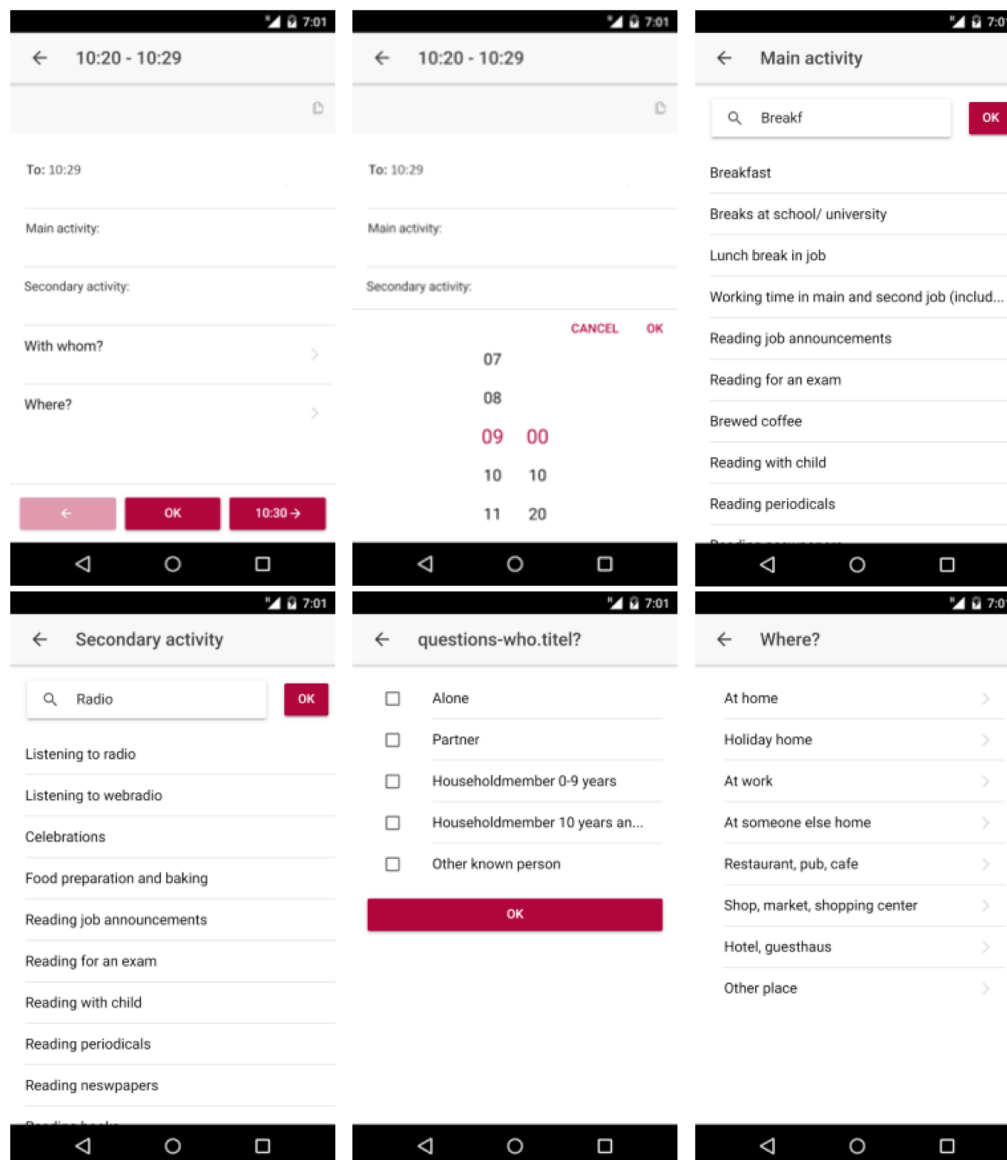


Figura 2.2.1: Interfaz de la aplicación austriaca STAT_time. Fuente: [22]

El segundo, el belga-germano, propone, además de una aplicación para móviles Android y iOS, el uso de tecnologías de reconocimiento de voz y del lenguaje [23]. Además, proponen el desarrollo de *plugins* para los asistentes virtuales de los distintos sistemas operativos (Siri en iOS, Google Assistant en Android y Alexa en los dispositivos de Amazon) para reducir aún más el esfuerzo que tendrían que realizar los encuestados en entrar periódicamente a la aplicación para registrar sus actividades, pudiendo registrar la actividad que están desarrollando, y otra información contextual, directamente desde estos asistentes de voz.

Fuera del ámbito europeo también se están desarrollando proyectos interesantes e innovadores en relación con la recolección de datos. Por ejemplo, en Australia la University of Western Australia en colaboración con el CTUR realizaron un estudio utilizando relojes inteligentes y otros *wearables* para la captura pasiva de datos de forma complementaria a los tradicionales

cuestionarios; los resultados obtenidos fueron que gracias a los relojes inteligentes los episodios relacionados con los desplazamiento se ajustaban mejor a la realidad que los cuestionarios y que además los dispositivos proporcionaban otra información contextual y de comportamiento como los niveles de estrés o satisfacción [24], durante la realización de ciertas tareas, gracias a los datos proporcionados por los distintos sensores de los *smartwatches* que miden desde la frecuencia cardíaca, nivel de oxígeno en sangre o nivel de actividad física.

Volviendo al caso de España, que es el que nos ocupa, las Encuestas de Empleo del Tiempo (EET) son un tipo de encuestas publicadas por el INE con una periodicidad irregular, y tal y como se ha mencionado en apartados anteriores únicamente hay datos de dos encuestas realizadas durante los años 2002-2003 y 2009-2010 [25]. Esta baja frecuencia de publicación de los datos hace que sea bastante complicado plantear el análisis en términos de evolución de los resultados de las encuestas.

En cuanto a la metodología utilizada por el INE en estos dos periodos, en ambos casos se han recogido los datos mediante cuestionarios en papel y con entrevistadores que se desplazaban hacia los domicilios de los encuestados para asegurar que los hogares cumplieran con los requisitos del diseño de la muestra. El formulario se dividía en tres partes: un cuestionario sobre las características del hogar (número de miembros, comunidad autónoma, tamaño del municipio, relación de parentesco entre los miembros del hogar, si disponen de servicio doméstico, etc.), un cuestionario sobre características individuales de cada uno de los encuestados en el hogar (fecha de nacimiento, sexo, nacionalidad, nivel de estudios, actividad económica o puesto de trabajo, etc.) y, por último, el cuestionario del diario [20]. El diario de actividades es un cuestionario tipo agenda o calendario, donde el encuestado escribe las actividades que estaba realizando en intervalos de 10 minutos hasta completar las 24 horas de un día. A continuación, en la Figura 2.2.2 se puede ver un extracto de dicho cuestionario. Finalmente, estos cuestionarios son recogidos y su información es registrada de forma manual.

B. Diario

Hora	(1) ¿Qué estaba haciendo? Escriba su actividad principal para cada intervalo de 10 minutos. Una actividad en cada línea.	Ordenador o Internet	(2) ¿Qué más estaba haciendo? Registre la más importante de las actividades que realizaba al mismo tiempo que la actividad principal.	Ordenador o Internet
	¿A qué hora comenzó esta actividad? hh mm			
06:00–06:10		☐		☐
06:10–06:20		☐		☐
06:20–06:30		☐		☐
06:30–06:40		☐		☐
06:40–06:50		☐		☐
06:50–07:00		☐		☐
07:00–07:10		☐		☐
07:10–07:20		☐		☐
07:20–07:30		☐		☐
07:30–07:40		☐		☐
07:40–07:50		☐		☐
07:50–08:00		☐		☐

Figura 2.2.2: Extracto del Diario de Empleo del Tiempo. Fuente: [26]

Esta metodología presenta muchas de las desventajas mencionadas al principio de este apartado, además de ser especialmente costosa. Según los propios datos del INE, para la encuesta del 2002-2003 se presupuestaron un total de 1.365.680 euros, mientras que para la del 2009-2010 se presupuestaron 4.573.610 euros [27]. Con todo esto, encontramos uno de los motivos que quizás justifiquen la falta de periodicidad e interés en la elaboración de estas encuestas. No obstante, el organismo español para el 2024 ha iniciado una nueva ronda de encuestas en las que el coste presupuesto es claramente inferior, 984.010 euros [27], esto se debe a que finalmente se ha decidido digitalizar y modernizar gran parte del proceso a través de la entrega telemática de los cuestionarios mediante el portal del INE IRIA [28].

2.2.1.1 Formato de los microdatos de las EET del INE

Como se ha mencionado en apartado 2.1 el INE no ofrece estadísticas con el suficiente nivel de desagregación necesario para los análisis que se van a realizar, principalmente para el cálculo del LDI que requiere la información de cada uno de los cuestionarios del diario de actividades. Por ello, se ha partido de los microdatos del INE publicados en su portal web [25]. Estos microdatos son los datos desagregados y anonimizados de todas las encuestas y tratar con los microdatos directamente conlleva ciertos riesgos e inconvenientes, algunos de ellos intrínsecos al manejo de grandes volúmenes de datos y otros relacionados con el formato utilizado por la entidad estadística española.

En INE organiza los microdatos de los diferentes cuestionarios se encuentran en distintos ficheros siguiendo un formato distinto para cada uno de los periodos, lo cual, metodológicamente complica la extracción de información ya que es necesario crear dos procesos distintos para cada uno de los años, duplicando el trabajo necesario para poder conseguir datos en un formato analizable. A continuación, se detalla el contenido de cada uno de los ficheros del periodo 2009-2010:

- DHOGAR.TXT: Contiene información recogida en el cuestionario del hogar (excepto la relativa a cada uno de los miembros que se encuentra en el fichero MHOGAR) y variables derivadas de otras que se encuentran en los cuestionarios (comunidad autónoma, tamaño del municipio, tipo de hogar, etc.). Hay tantos registros como hogares en la muestra, es decir, 9.541 hogares.
- MHOGAR.TXT: Contiene información relativa a cada uno de los miembros de los hogares del fichero DHOGAR. Cada persona del hogar se distingue del resto mediante la variable NPERS. Hay tantos registros como miembros de hogares en la muestra, es decir, 25.895 personas.
- CINDIV.TXT: Contiene información recogida en el cuestionario individual relativa a los miembros de 10 y más años de los hogares del fichero DHOGAR. También contiene variables derivadas de otras que se encuentran en el cuestionario. Hay tantos registros como cuestionarios individuales se han recogido, es decir, 19.295.
- DIARIO1.TXT: Contiene la información complementaria recogida en el diario de actividades relativa a los miembros de 10 y más años de los hogares del fichero DHOGAR. Hay tantos registros como diarios de actividades se han recogido, 19.295, los mismos que cuestionarios individuales.
- DIARIO2.TXT: Contiene información relativa al diario de actividades de los miembros de 10 y más años de los hogares del fichero DHOGAR. La

información se recoge en intervalos de 10 minutos, 144 intervalos por diario de actividades recogido, es decir, en total 2.778.480 registros (144*19.295 diarios).

- HTR1.TXT: Contiene información recogida en el horario de trabajo semanal de como considera la semana de trabajo (habitual, inusual, etc.) el informante de los hogares del fichero DHOGAR y que además ha cumplimentado el horario.
- HTR2.TXT: Contiene información recogida en el horario de trabajo semanal relativa a los miembros de los hogares del fichero DHOGAR y que además han cumplimentado el horario.
- SD.TXT: Contiene información de las condiciones laborales del personal que trabaja en los hogares como servicio doméstico.
- DISEÑOS DE REGISTRO EET 2009 2010.xlsx: Se describe en hojas Excel separadas el diseño de registros de cada uno de los ficheros mencionados anteriormente, incluyendo en cada descripción el nombre de la variable, su tamaño en número de caracteres o posición dentro de cada registro, descripción de la variable y valores válidos (véase Tabla 2.2.1).

Tabla 2.2.1: Extracto del Diseño de Registro correspondiente al fichero DIARIO2.TXT. Fuente: [25]

FICHERO DIARIO2. DIARIO DE ACTIVIDADES (1)			
IDENTIFICACIÓN (1)			
Variable	Posición	Descripción	Valores válidos
IDHOGAR	1-5	Identificador del hogar	
NPERS	6-7	Persona a la que se refiere la información	01-15
TRIM	8	Trimestre	
		1er. Trimestre (de 2010)	1
		2º Trimestre (de 2010).	2
		3er. Trimestre (de 2010)	3
		4º Trimestre (de 2009)	4
DDIASEM	9	Día de la semana de cumplimentación del diario	
		Lunes	1
		Martes	2
		Miércoles	3
		Jueves	4
		Viernes	5
		Sábado	6
		Domingo	7

Para el periodo 2002-2003 los ficheros son:

- CH.TXT: equivalente al DHOGAR, hay 20.603 registros, uno por hogar.
- MH.TXT: equivalente al MHOGAR, hay 60.493 registros, tantos como personas en los hogares de la muestra.
- CI.TXT: equivalente al CINDIV, hay 46.774 registros, uno por cuestionario individual.
- DI.TXT: equivalente al DIARIO1 y DIARIO2
- HT.TXT: equivalente a HTR1 y HTR2.
- Disreg ET 02-03.xls: equivalente a DISEÑOS DE REGISTRO EET 2009 2010.

La información necesaria para los análisis y visualizaciones deseadas son tanto los datos de información de los hogares, para los datos correspondientes a las Comunidades Autónomas, y el diario de actividades; por lo que los ficheros con los que se ha trabajado han sido DHOGAR y DIARIO2 para 2009-2010, y CH y DI para 2002-2003. A continuación, en la Figura 2.2.3, la Figura 2.2.4, la Figura 2.2.5 y la Figura 2.2.6 se muestran los contenidos concretos por cada registro de los ficheros mediante un ejemplo extraído de los propios archivos.

00018161802006 151010017719548462981

Figura 2.2.3: Registro del fichero DHOGAR, periodo 2009-2010. Fuente: [25]

00001011 303076 41041111666666666666103111616
61102610111111166 166 166166666236 6 6 6 6 6 6 6 6...
011100465155265

Figura 2.2.4: Registro del fichero CH, periodo 2002-2003. Fuente: [25]

Como se puede observar los registros tienen formatos distintos en los dos periodos, para la mejor comprensión de los datos comunes se han remarcado del mismo color los campos, que son los siguiente:

- En **rojo**, los cinco primeros caracteres en ambos registros, que se corresponden al identificador del hogar.
- En **verde**, del carácter sexto al séptimo en ambos registros, que se corresponden al identificador de la Comunidad Autónoma; en el caso de la Figura 2.2.3 el código corresponde al País Vasco, mientras que en la Figura 2.2.4 corresponde a Andalucía.
- En **azul oscuro**, el carácter octavo, que se corresponde al tamaño del municipio; en la Figura 2.2.3 y Figura 2.2.4 el código corresponde a capitales de provincia o municipios con más de 100.000 habitantes.
- En **morado**, el carácter noveno en la Figura 2.2.3 y del decimocuarto al decimoquinto en la Figura 2.2.4 el código del tipo de hogar; en el caso de la Figura 2.2.3 el código corresponde a la categoría “Otro tipo de hogar”, mientras que en la Figura 2.2.4 corresponde a dos adultos con un niño dependiente.
- En **naranja**, del carácter décimo al decimoprimer en la Figura 2.2.3 y del decimosegundo al decimotercero en la Figura 2.2.4 el número de personas en el hogar; en el caso de la Figura 2.2.3 dos, mientras que en la Figura 2.2.4 tres.
- En **cian**, el carácter decimoctavo en la Figura 2.2.3 y el octogésimo octavo en la Figura 2.2.4, que se corresponden al código de los ingresos mensuales medios netos del hogar; en el caso de la Figura 2.2.3 el código corresponde a la categoría “NS/NC” (No sabe/No contesta), mientras que en la Figura 2.2.4 corresponde a unos ingresos desde 1.000 hasta 1.499,99 euros.
- En **azul claro**, el carácter decimonoveno en la Figura 2.2.3 y el decimosesto en la Figura 2.2.4, que se corresponden al código que indica si en el hogar se dispone de servicio doméstico; en el caso de la Figura 2.2.3 el hogar si dispone de servicio doméstico, mientras que en la Figura 2.2.4 no.
- En **lila**, del carácter vigésimo al vigesimoprimer en la Figura 2.2.3 y del decimoséptimo al vigesimosexto en la Figura 2.2.4, otros códigos correspondientes a servicio doméstico (número de personas empleadas, residente o no en el hogar, etc.).

- En **beige**, los dieciséis últimos caracteres en la Figura 2.2.3 y los once últimos en la Figura 2.2.4 son el factor de elevación, es decir el inverso de la probabilidad del hogar de ser seleccionado en la muestra.
- Por último, en **amarillo**, se recogen códigos para otra información incluida en los cuestionarios como fuentes de ingresos, ayudas y subsidios recibidos, bienes y servicios con los que cuenta el hogar (TV, Internet, lavavajillas, etc.) y otras cuestiones generales.

0001801450010116 6116666660113500580113784

Figura 2.2.5: Registro del fichero DIARIO2, periodo 2009-2010. Fuente: [25]

0000101No301100066666120110006666612...31161111 1100465155265121752

Figura 2.2.6: Fragmento de un registro del fichero DI, periodo 2002-2003.
Fuente: [25]

De nuevo, como sucede con los ficheros correspondientes a los datos de los hogares los ficheros del diario de actividades siguen formatos distintos, la correspondencia de los campos por colores es la siguiente:

- Los colores **rojo**, **amarillo** y **beige** tienen el mismo significado que en los cuestionarios del hogar.
- En **naranja**, del carácter sexto al séptimo en ambos registros, que se corresponden al identificador de la persona dentro de su hogar.
- En **verde**, el carácter noveno en la Figura 2.2.5 y el décimo en la Figura 2.2.6, que se corresponden al día de la semana al cual hacen referencia los datos de las actividades; en el caso de la Figura 2.2.5 las actividades corresponden a un sábado, mientras que en la Figura 2.2.6 a un miércoles.
- En **azul**, del carácter decimo al decimosegundo en la Figura 2.2.5, que se corresponde con el intervalo de 10 minutos del día al cual hace referencia la actividad del registro; en este caso se trata de la franja de las 6:00 de la mañana a las 6:10. Como se puede ver, la Figura 2.2.6 no cuenta con este campo, ya que para el periodo 2002-2003 todas las actividades del día en referencia a una persona se incluyen en un único registro donde se pone la información correspondiente a cada intervalo de tiempo en bloques de trece caracteres de forma consecutiva hasta hacer los 144 bloques, dando lugar a un registro de un tamaño de 1.913 caracteres mucho mayor que el de 44 del periodo 2009-2010.
- En **verde olivo**, del carácter decimotercero al decimoquinto en la Figura 2.2.5 y del decimoprimer al decimocuarto en la Figura 2.2.6 para el primer intervalo de tiempo, que se corresponden al código de la actividad principal; tanto en el caso de la Figura 2.2.5 como en la Figura 2.2.6 los datos se corresponden al código de la actividad “Dormir”.
- En **rosa**, el carácter decimosexto, el vigésimo y del vigesimotercero al vigesimooctavo en la Figura 2.2.5 y del decimoséptimo al vigesimoprimer en la Figura 2.2.6 para el primer intervalo de tiempo, que se corresponden a los código de la información del contexto en el que se desarrolla la actividad (acompañado o no, uso de ordenador o internet, etc.); tanto en el caso de la Figura 2.2.5 como en la Figura 2.2.6 los códigos indican que no se estaba acompañado por nadie y que no se hacía uso ni del ordenador ni de internet.
- En **gris**, del carácter vigesimoprimer al vigesimosegundo en la Figura 2.2.5 y del vigesimosegundo al vigesimotercero en la Figura 2.2.6 para el primer intervalo de tiempo, que se corresponden al código del lugar en el

que se desarrolla la actividad; en el caso de la Figura 2.2.5 la actividad se desarrolla en la vivienda habitual, mientras que en la Figura 2.2.6 la actividad se desarrolla en la segunda vivienda.

Estos ficheros son los que serán procesados por un técnica informática conocida como ETL (*Extract, Transform and Load*) que tiene como objetivo transformar los datos para que sean manejables y analizables con diversos algoritmos o directamente por humanos. La implementación de este proceso se explica en el apartado 2.2.3, referente a la metodología y las tecnologías empleadas.

2.2.2 Indicadores socioeconómicos utilizados en el *clustering*

Para el *clustering* se han utilizado un total de siete indicadores socioeconómicos, incluyendo el LDI, cuyos valores se muestran en el Anexo I. La fuente de los datos ha sido principalmente el Lighthouse, y para ciertos datos que no es encontraban disponibles en esta plataforma se ha recurrido a obtener los datos directamente del INE. El Lighthouse es la plataforma central del proyecto Economics @Intelligence, cuyo objetivo es aglutinar todo tipo de estadísticas e indicadores procedentes de distintos organismos oficiales y que trabaja de forma mayoritaria con datos del INE para España, entre otras fuentes de datos menos extensas, y con los datos del Banco Mundial para datos a nivel de países [29]. De forma más general, la misión del Lighthouse es facilitar el acceso a cualquier persona a datos oficiales para la aplicación de técnicas estadísticas avanzadas y de inteligencia artificial sobre esos datos para permitir el análisis de estos.

Para el análisis de *clustering* que se plantea en este proyecto se han escogido los siguientes indicadores, además del LDI, calculado por los métodos que se explican en el apartado 2.2.3.1.

- Producto Interior Bruto (PIB) como índice de volúmenes encadenados [30], procedente del INE por falta de datos en el Lighthouse para los dos periodos analizados.
- Tasa de paro a partir de los datos de la Encuesta de Población Activa (EPA) [29], procedente del Lighthouse.
- Índice de Precios al Consumo (IPC) [29], procedente del Lighthouse.
- Edad Media de la población [31], procedente del INE.
- Población que sufre problemas de contaminación y otros problemas ambientales [29], procedente del Lighthouse.
- Tasa de obesidad, es decir el porcentaje de la población con un Índice de Masa Corporal (IMC) igual o superior a los 30 kilogramos por metro cuadrado [29], procedente del Lighthouse.

2.2.3 Metodología y tecnologías para el análisis de los datos

En el desarrollo de este trabajo, se emplearon una serie de metodologías y tecnologías avanzadas para la extracción, transformación y carga (ETL) de datos, con el objetivo de obtener datos en los formatos adecuados para su análisis económico. Se partió de los microdatos del INE, que se transformaron y procesaron siguiendo diversos pasos detallados a continuación, cuyos detalles a nivel de implementación se explican en el TFG de Ingeniería Informática complementario a este.

Para el procesamiento de datos, se utilizó el conjunto de paquetes “tidyverse” [32] de R. Este conjunto de herramientas facilita la manipulación y análisis de

datos basándose en los principios de *tidy data*, que buscan organizar los datos de manera coherente y uniforme [32]. Entre los paquetes más destacados utilizados en las diferentes etapas del proceso se encuentran:

- “ggplot2”: Para la creación de gráficos de alta calidad.
- “dplyr”: Para la manipulación y transformación de datos, facilitando operaciones como filtrar, seleccionar y agregar datos.
- “tidyr”: Para limpiar y organizar datos desordenados.
- “readr”: Para la importación rápida y eficiente de datos desde diferentes tipos de archivos.
- “stringr”: Para la manipulación de cadenas de caracteres.
- “tibble”: Para trabajar con una versión moderna y mejorada de los “data.frames” en R.

Estos paquetes permiten realizar transformaciones de datos de manera eficiente y fiable, convirtiéndose en herramientas con una gran utilidad para todas las etapas del flujo de trabajo de análisis de datos.

El preprocesado de los microdatos del INE, almacenados en archivos de texto plano, fue un desafío debido a su formato no tabular. Estos datos se convirtieron en estructuras de datos manejables en R, como “tibbles”, utilizando funciones de “tidyverse”. Se desarrollaron scripts en R y Python para procesar los datos de diferentes periodos, adaptando las técnicas a las características específicas de cada conjunto de datos.

Para los datos del periodo 2009-2010, se empleó R con scripts que utilizaban las librerías “tibble”, “readr” y “stringr” para cargar, transformar y exportar los datos a archivos CSV. En cambio, para el periodo 2002-2003, se optó por Python debido a la mayor complejidad de los registros y a la capacidad del lenguaje paralelizar tareas con la librería “multiprocessing”, lo que permitió una significativa reducción del tiempo de procesamiento, algo que no es posible conseguir con R.

A nivel global el *pipeline* de datos del proyecto se encuentra resumido en la Figura 2.2.7, donde en primer lugar se extraen los datos del INE, y del Lighthouse del Economics @Intelligence en el caso de algunos datos del *clustering*. A continuación, esos datos se procesan, se realizan los cálculos y se aplican los algoritmos adecuados para obtener los resultados correspondientes a los diversos análisis realizados. Después, se almacenan esos resultados en el formato más adecuado para que sean leídos por el *dashboard* donde se muestran las distintas visualizaciones y gráficos vinculados al análisis. Finalmente, ese *dashboard* es desplegado en el servidor del Economics @Intelligence para que sea accesible por cualquier persona¹.

¹ <http://vps159.cesvima.upm.es:3838/time-use/>

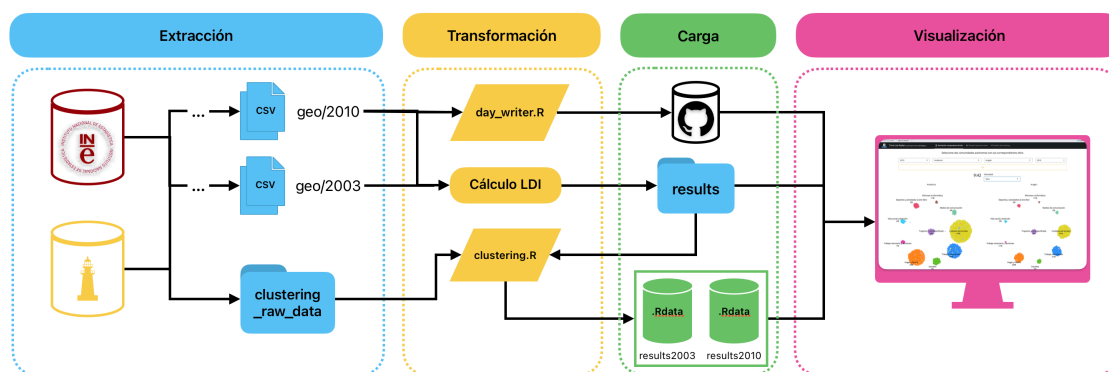


Figura 2.2.7: Diagrama global de la ETL. Elaboración propia

Este enfoque metodológico aseguró que los datos estuvieran limpios, organizados y listos para su análisis, proporcionando una base sólida y fiable para obtener conclusiones económicas precisas y útiles.

2.2.3.1 Cálculo del LDI

El cálculo del LDI también se ha realizado de manera informatizada usando las mismas herramientas que para el preprocesado de los datos, dada la complejidad computacional del algoritmo para su cálculo. Este algoritmo se ha dividido en tres etapas. En primer lugar, se ha computado un ranking con todas las posibles ternas del LDI ordenadas de la menos a la más dinámica. Para hacer esto se definen cuatro funciones auxiliares para calcular el número de combinaciones de ternas posibles dados unos valores para el número de actividades e intervalos de tiempo y a su vez calcular su cardinalidad, es decir su orden. Estas funciones se basan en las funciones del propio artículo del *Lifestyle Dynamics Index* [1], a partir de los resultados de estas funciones se genera el ranking de posibles ternas en un triple bucle anidado donde se escribe cada posible valor de la terna (x_d, y_d, z_d) . Para dar una idea intuitiva del funcionamiento de este proceso en la Tabla 2.2.2 se ofrece cual sería la salida esperada para un ejemplo simplificado en el cual existen tan solo 2 actividades posibles y 4 franjas horarias, ordenando las ternas de la menos dinámica a la más dinámica.

Tabla 2.2.2: Ejemplo de ranking del LDI para 2 actividades y 4 intervalos de tiempo. Elaboración propia

x_d	y_d	z_d
1	0	0
2	1	1
2	2	3
2	2	2

Este proceso se ha repetido un total de 5 veces, ya que se deseaban obtener rankings para los tres niveles de granularidad de actividades que ofrece el INE [20] y en los niveles de mayor grado de detalle y los niveles intermedio el número de actividades es distinto en cada periodo. Siendo las combinaciones de

actividades e intervalos de tiempo para cada periodo y nivel de granularidad las siguientes:

- 10 actividades y 144 intervalos, siendo común a ambos periodos.
- 39 actividades y 144 intervalos, correspondiendo al nivel intermedio de detalle para el periodo 2002-2003.
- 34 actividades y 144 intervalos, correspondiendo al nivel intermedio de detalle para el periodo 2009-2010.
- 192 actividades y 144 intervalos, correspondiendo al nivel más detallado para el periodo 2002-2003. En este caso, en la práctica, para calcular el ranking se usan 144 actividades ya que una persona no puede hacer más actividades que intervalos de tiempo presentes en un día.
- 116 actividades y 144 intervalos, correspondiendo al nivel intermedio de detalle para el periodo 2009-2010.

En segundo lugar, se calculan las ternas de cada uno de los encuestados, procesando sus cuestionarios del diario de actividades, y se obtiene la posición que ocuparía esa terna en el ranking calculado previamente. Por último, se calculan los LDI para cada periodo, nivel de granularidad y Comunidad Autónoma siguiendo la Ecuación 10 y se almacenan los resultados en un formato adecuado para poder producir los gráficos y visualizaciones necesarias para el análisis.

2.2.3.2 Implementación del algoritmo de *clustering* K-Means

Un algoritmo de *clustering* es una técnica de análisis de datos que tiene como objetivo agrupar un conjunto de datos en subconjuntos homogéneos, denominados clústeres [33]. Cada clúster agrupa observaciones que son más similares entre sí en comparación con las observaciones de otros clústeres. La similitud se mide a través de una métrica de distancia, como la distancia euclidiana, y la técnica busca minimizar las diferencias dentro de cada clúster mientras maximiza las diferencias entre los clústeres [33]. En el ámbito de este proyecto se utiliza para identificar patrones entre las Comunidades Autónomas en base a los indicadores socioeconómicos recogidos en el apartado 2.2.2.

Para calcular esos grupos de Comunidades Autónomas similares se ha utilizado la librería “scikit-learn” de Python que ofrece múltiples algoritmos de *clustering* como se muestra en la Figura 2.2.8. Para este análisis se ha optado por el algoritmo K-Means, ampliamente utilizado por su sencillez y velocidad de cómputo.

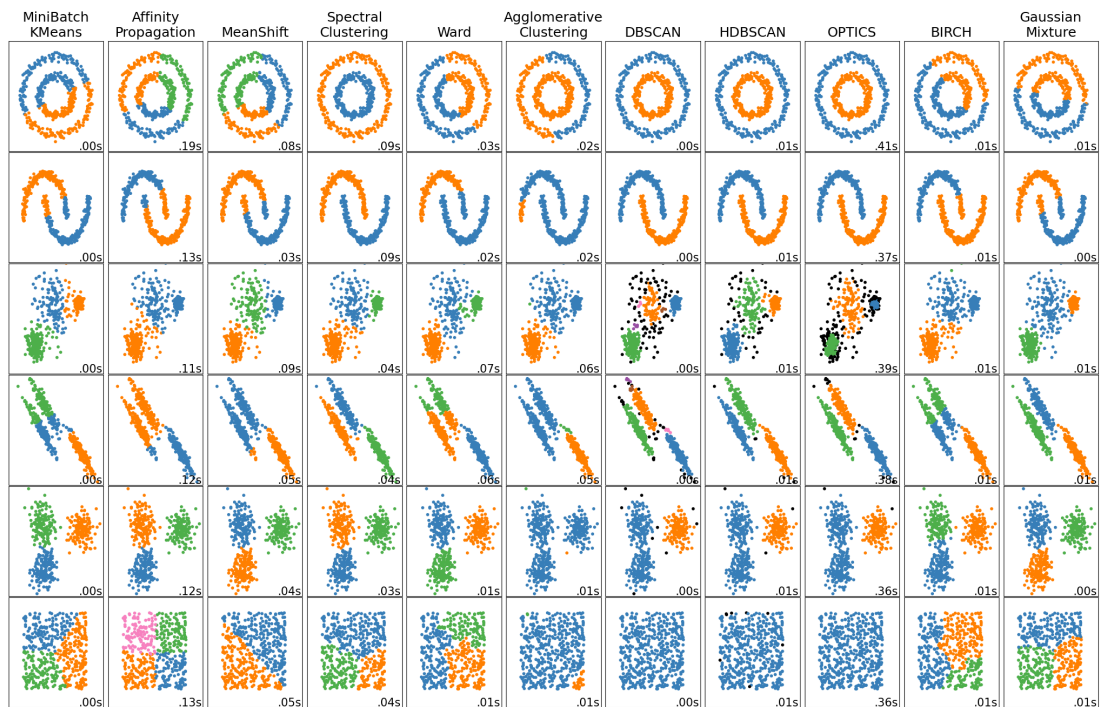


Figura 2.2.8: Clústeres generados por los algoritmo incluidos en “scikit-learn” sobre distintos conjuntos de datos. Fuente: [33]

La elección de este algoritmo hace que se tengan que predefinir el número de clústeres que queremos, algo que no sucede con otros algoritmos. Para ello, existen una gran variedad de métodos, siendo los más comunes el método del codo y el método del análisis de la silueta. El método del codo permite identificar gráficamente el número de clústeres óptimo. En un gráfico, como el de la Figura 2.2.9, se muestra la reducción de la inercia en eje de ordenadas frente al número de clústeres en el eje de las abscisas [34]. El punto donde la disminución de la inercia se vuelve menos pronunciada indica el número ideal de clústeres, formando el codo en el gráfico que da nombre al método [34].

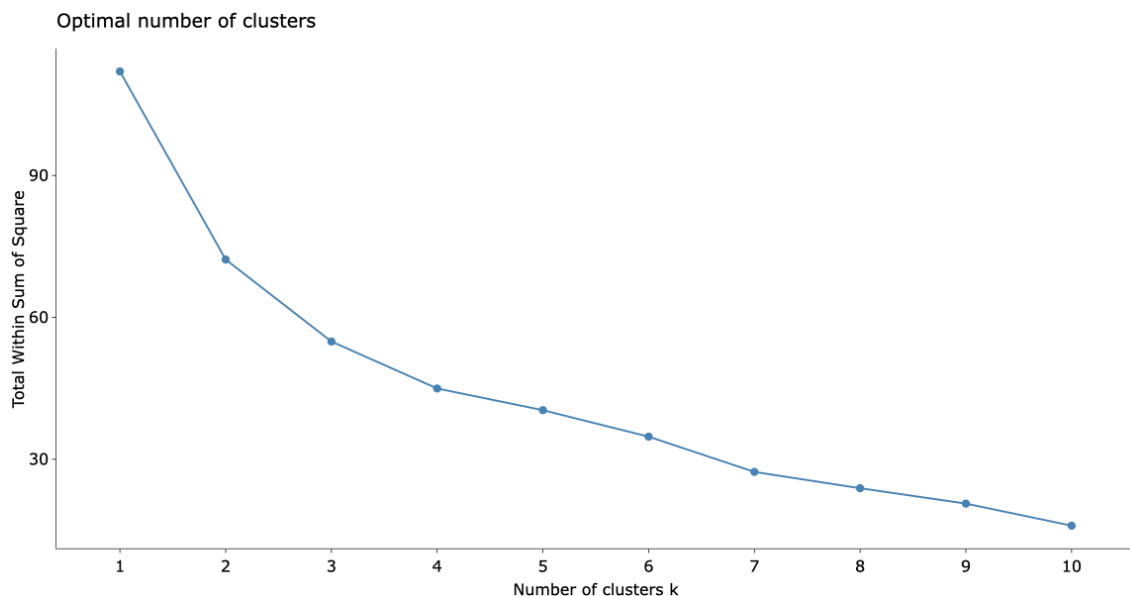


Figura 2.2.9: Aplicación del método del codo sobre los datos del 2010

Por otra parte, el método del análisis de la silueta mide la similitud de una observación con su grupo asignado en comparación con otros clústeres. Valores altos (cercanos a 1) del coeficiente de silueta indican que la observación está bien asignada a su clúster, dado que minimiza la distancia intra-clúster y maximiza la distancia inter-clúster [34]. En el gráfico de la Figura 2.2.10 se muestra la evolución del coeficiente medio de la silueta para diferentes números de clústeres con los datos correspondientes al año 2003, donde se selecciona el número de clústeres con el coeficiente máximo, en este caso dos.

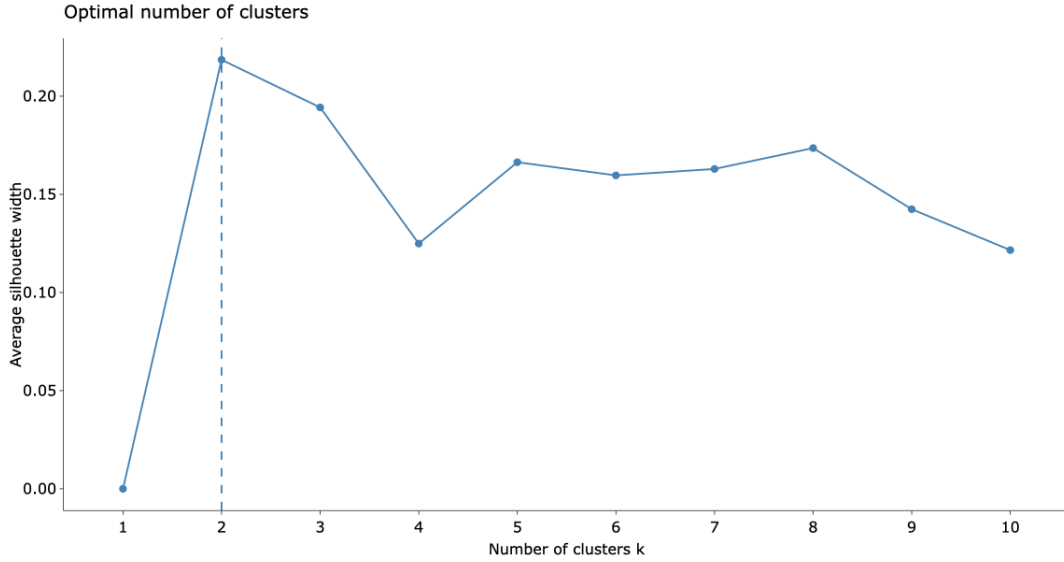


Figura 2.2.10: Aplicación del método de análisis de la silueta sobre los datos de 2003

Estos métodos, aunque útiles gracias a que son muy visuales, tienen desventajas. Además de la subjetividad en la selección del número adecuado de clústeres, requieren múltiples iteraciones del mismo algoritmo para diferentes números de clústeres, lo que implica un alto coste computacional. Por no mencionar el hecho de que pueden producir resultados muy dispares. Para abordar estas discrepancias, el paquete “NbClust” [35] de R ejecuta varios algoritmos y selecciona el número de clústeres que se repite con mayor frecuencia entre los diferentes métodos. Teniendo en cuenta tanto los resultados producidos por “NbClust” como los del método del codo se ha concluido que el número óptimo de clústeres para ambos periodos es tres.

El algoritmo K-Means tiene como objetivo crear un número n de clústeres disjuntos (C), cada uno caracterizado por un punto central denominado centroide (μ_j), que representa la media de las observaciones en ese clúster [33]. Inicialmente, los centroides se establecen aleatoriamente a partir de los valores del conjunto de datos de entrada [33]. En las iteraciones sucesivas, las observaciones se asignan al centroide más cercano, y se recalculan los centroides como la media de las observaciones asignadas en cada iteración [33]. El proceso continúa hasta que los centroides no cambian significativamente. La eficacia del *clustering* puede evaluarse mediante la inercia (suma de las distancias al cuadrado de cada observación a su centroide):

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Ecuación 11: Función de minimización de la inercia. Fuente: [33]

Cuanto menor sea la inercia, mejor formados estarán los clústeres, ya que se busca minimizar la distancia intra-clúster y maximizar la distancia inter-clúster. Sin embargo, la inercia es una medida relativa y no absoluta, por lo que solo es comparable con otros resultados en el mismo contexto y con datos de entrada similares. Evaluar el rendimiento de algoritmos de aprendizaje no supervisado como K-Means puede ser complicado debido a la falta de etiquetas reales con las cuales comparar los resultados, por ello se han desarrollado técnicas específicas para facilitar esa evaluación y la interpretabilidad de los resultados como se explica en el apartado 2.2.3.3.

Finalmente, con los resultados obtenidos se elaboran toda una serie de gráficos para su análisis que se encuentran disponibles en la aplicación web y a lo largo de este TFG.

2.2.3.3 Interpretabilidad de los resultados del *clustering*

Cuando se trabaja con algoritmo de *machine learning*, en especial con algoritmos de aprendizaje automático como el *clustering*, a menudo se tiene la sensación de estar ante una especie de caja negra donde se conocen las entradas y salidas, pero los procesos intermedios que llevan a una decisión específica son desconocidos. Esta falta de transparencia puede ser un obstáculo significativo para la adopción de la Inteligencia Artificial en campos donde la legislación y la ética son cruciales, ya que la interpretabilidad de los resultados es tan importante como los mismos resultados. Este fenómeno se conoce como *feature importance*, y consiste en dotar de cierta explicabilidad de las salidas obtenidas *a posteriori*.

Entre las metodologías más modernas y ampliamente utilizadas para mejorar la interpretabilidad del aprendizaje automático se encuentra el método SHAP (*SHapley Additive exPlanations*). Este método, implementado en el *dashboard*, se basa en los valores de Shapley provenientes de la teoría de juegos cooperativos y puede utilizarse para explicar la salida de cualquier modelo de *machine learning* [36]. Este método proporciona una forma de descomponer la salida de un modelo en contribuciones individuales de cada característica, lo que permite entender cómo cada característica contribuye a la predicción final [36]. La idea detrás de SHAP es asignar una “importancia” a cada característica, indicando cuánto contribuye esa característica a la predicción del modelo para una instancia de datos específica, un territorio en el caso del este análisis. SHAP tiene la ventaja de ser un enfoque aditivo, lo que significa que las contribuciones individuales de cada característica suman la clasificación del modelo para una instancia de datos dada [36]. Esto proporciona una explicación global y coherente de cómo se llega a un resultado específico, lo que puede ser crucial para interpretar y comprender el modelo, especialmente en entornos donde se requiere transparencia y explicabilidad [36], como en el caso de este análisis.

A nivel de implementación se ha recurrido a la librería de Python “shap” para los cálculos y a la librería de R “shapviz” [37] para generar gráficos como la de la Figura 2.2.11 de la cual podemos extraer que para el año 2010 todas las variables contribuían de forma positiva para que Andalucía formase parte del clúster 0, siendo más importante la contribución de la Tasa de obesidad seguida del LDI.

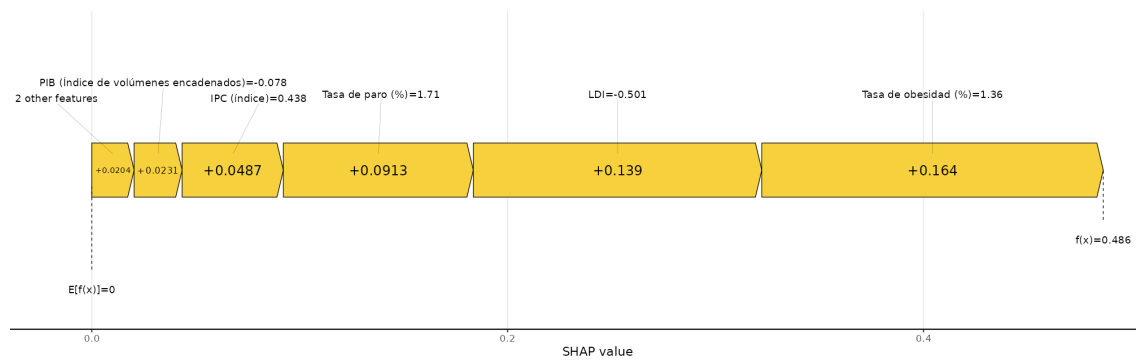


Figura 2.2.11: Gráfico de fuerza para Andalucía y el clúster 0 en el año 2010

2.3 Análisis de los resultados

2.3.1 Análisis exploratorio del día promedio en las Comunidades Autónomas

Antes de proceder al análisis del LDI y del *clustering* es conveniente realizar un análisis exploratorio de los datos para dar algo de contexto y tener una idea inicial de como son los datos para cada uno de los periodos, viendo a nivel intuitivo que Comunidades Autónomas tienen un mayor nivel de dinamismo en su población. Este trabajo se realiza en la aplicación web mediante una pestaña con una simulación donde se muestran dinámicamente como se van moviendo los encuestados por los diferentes grupos de actividades a lo largo del día. No obstante, para este documento la información se muestra condensada en dos gráficos, un formato estático más adecuado para documentos.

La Figura 2.3.1 muestra un gráfico de barras apiladas que representa los porcentajes de personas que realizan diversas actividades en las Comunidades Autónomas de España, basados en los resultados de las encuestas de empleo del tiempo del INE para el periodo 2002-2003. A nivel nacional las actividades más frecuentes son “Cuidados personales”, donde se incluye dormir, realizada por el 100% de la población; seguida de “Medios de comunicación”, “Trayectos y tiempo no especificado” y “Hogar y familia” en el entorno del 80%. Por último, le siguen la categoría de “Vida social y diversión” con un 66,8% y el resto de las actividades con porcentajes inferiores al 50%.

A nivel de las Comunidades Autónomas, podemos analizar cuáles son las más dinámicas, definiendo en este caso dinamismo como los porcentajes de población que realiza una actividad agregados (es decir, sumados). Aunque esta definición del dinamismo no es la misma que en LDI, el analizar estos valores puede dar una aproximación de cuales van a ser los resultados del LDI previamente a realizar el cálculo de este.

En este sentido, los territorios cuya población realiza un mayor número de actividades con respecto a los valores medios para España en un día, es decir que tienen un día más dinámico, son en primer lugar Castilla y León, seguida de la Comunidad Foral de Navarra, La Rioja y las Illes Balears.

En el caso de Castilla y León las actividades cuyo porcentaje de población es superior al promedio son “Medios de comunicación” con un porcentaje del 89%, “Hogar y familia” con un 82%; pero, sobre todo, son las actividades de “Aficiones e informática”, teniendo el nivel más alto de todos los territorios con un 22,2%, “Deporte y actividades al aire libre” con un 50,3%, solo por detrás del País Vasco y Navarra.

En los casos de la Comunidad Foral de Navarra y La Rioja las actividades que más destacan con respecto al promedio nacional son “Trabajo remunerado”, con un 37,5% y 37,1% respectivamente, quedando solo por detrás de Baleares cuyo porcentaje es bastante superior al promedio con un 39,7%. Además, las dos comunidades del norte antes mencionadas, también destacan en el uso de “Medios de comunicación” con un 89,8% y 87,8%, respectivamente. Por último, los navarros, en porcentaje, son los españoles que más reuniones tienen y realizan más trabajos voluntarios con un 18,2%, 5,8 puntos porcentuales por encima de Total Nacional.

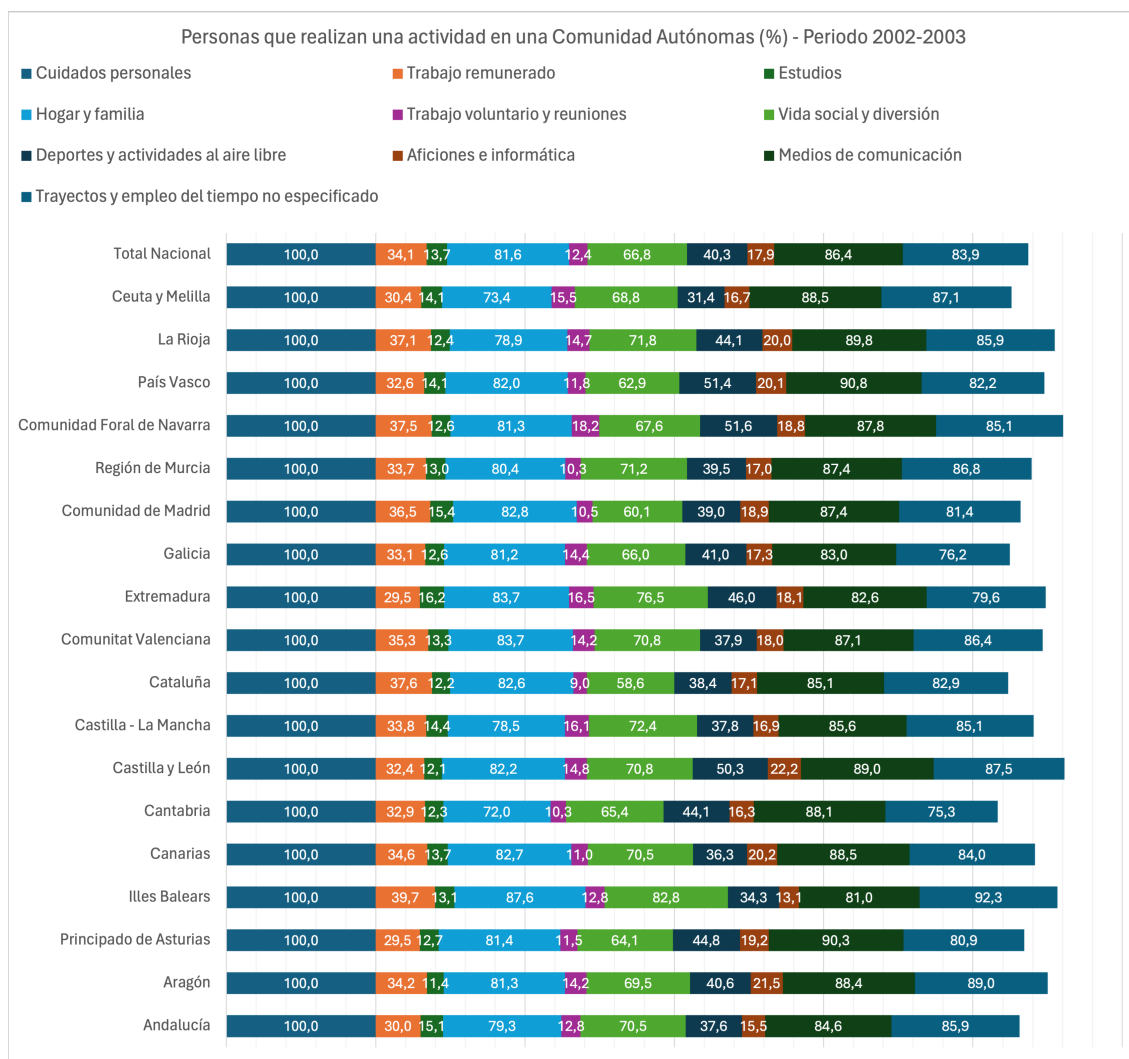


Figura 2.3.1: Porcentaje de personas que realizan una actividad en el transcurso del día para el periodo 2002-2003. Elaboración propia con datos del INE [25]

El caso de Baleares es especialmente particular, pues a pesar de situarse en términos generales de dinamismo claramente por encima del resto de territorios, en algunas actividades tienen los porcentajes de población más elevados, mientras que en otras sus porcentajes son de los más bajos comparativamente con el resto de las Comunidades Autónomas. Las islas, destacan positivamente en “Trabajo remunerado”, “Hogar y familia”, “Vida social y diversión” y “Trayectos y tiempo no especificado” con porcentajes del 39,7%, 87,6%, 82,8% y 92,3%, respectivamente; teniendo los porcentajes más elevados de entre todas las autonomías. Por otra parte, las Baleares tienen el valor más bajo en “Aficiones e informática” con un 13,1%, en “Medios de comunicación” con un 81%, y en “Deportes y actividades al aire libre” con un 34,3% solo se sitúa por delante de Ceuta y Melilla.

En el otro lado de la moneda tenemos a las comunidades menos dinámicas, encabezadas por Cantabria y seguida por Cataluña, Galicia y, Ceuta y Melilla. Cantabria, destaca por tener el porcentaje de personas más bajo en las actividades “Hogar y familia”, un 72%, y “Trayectos y empleo del tiempo no especificado”, un 75,3%. Por otro lado, Cataluña, aunque con valores bastante cercanos al promedio en la mayoría de las actividades, en dos de ellas su

porcentaje es el más bajo de todas las autonomías. Estas son: “Trabajos voluntarios y reuniones” y “Vida social y diversión” con porcentajes del 9% y 58,6%, respectivamente. Una diferencia de 3,8 puntos porcentuales en el caso de los trabajos voluntarios y reuniones, y de 11,9 puntos porcentuales en el caso con respecto al Total Nacional. Por último, Galicia destaca por tener valores por debajo de la norma en los trayectos y Ceuta y Melilla, como ya se ha mencionado, en lo relacionado a deportes.

Si comparamos el dinamismo de la población española entre los periodos 2002-2003 y 2009-2010 (véase la Figura 2.3.2) observaremos un aumento de ese dinamismo; pasando de 537,1 puntos porcentuales, obtenidos sumando los porcentajes de cada actividad, a 541,1 puntos. Aunque esto nos puede llevar a pensar que la población tenía un estilo de vida más activo en el 2010 que en el 2003, si profundizamos en las categorías cuyo número de personas crece y en cuales decrece se observa un aumento de actividades más sedentarias como “Medios de comunicación” y un decrecimiento en actividades como “Deporte y actividades al aire libre”.

Otra de las tendencias claras con respecto al periodo 2002-2003 es que en 2009-2010 se experimenta un aumento de la categoría “Aficiones e informática”, pasando de un 17,9% a un 29,7%; simultáneamente, se puede ver como la categoría “Vida social y diversión” decrece en igual medida desde un 66,8% a un 57,7%. Este movimiento puede encontrar su justificación muy posiblemente en la adopción generalizada de Internet por parte de la población, lo que ha repercutido en que los españoles pasarán menos tiempo realizando actividades de tipo social y que conlleven contacto físico con otras personas, a realizar actividades más individuales y en línea. Aunque es algo fuera del alcance de este proyecto, un área de estudio interesante sería ver cómo ha impactado este cambio no solo en los estilos de vida sino también en aspectos como el bienestar (*wellbeing*) de la población; pudiendo incluir datos de periodos más actuales una vez sean publicados por el INE para ver el impacto de, por ejemplo, las redes sociales y el *smartphone* en los hábitos de los españoles.

El último patrón a nivel nacional que se observa es una disminución en todos los territorios de las personas que realizan actividades en la categoría “Trabajo remunerado”, la explicación de esta tendencia de los datos esta necesariamente relacionada con las elevadas tasas de paro de ese periodo, 2009-2010, a causa de la crisis en la que nuestro país se encontraba sumergido.

A nivel autonómico, las Illes Balears, Castilla y León, La Rioja y la Comunidad Foral de Navarra siguen siendo los cuatro territorios más dinámicos, en ese orden. Las islas siguen siendo un caso particular, pues en las actividades “Vida social y diversión” y “Trayectos y empleo del tiempo no especificado” tiene los valores más altos, con un 67,5% y un 91,8% respectivamente; mientras que en “Trabajo voluntario y reuniones” tiene el valor más bajo con un 9,9%. Comparativamente con el periodo 2002-2003, destaca en Baleares que pasa de ser la comunidad con menos personas en la actividad “Aficiones e informática” a situarse por encima de los valores medios, 13,1% en 2002-2003 a 33,6% en 2009-2010.

Siguiendo con los riojanos y navarros, aunque no tienen el máximo porcentaje en ninguna de las actividades si tienen porcentajes superiores a la norma en la mayoría de las actividades, lo cual es indicio de estilos de vida más dinámicos en términos generales. Con Castilla y León sucede algo similar, aunque también se observa un aumento considerable de las personas que realizan actividades de la categoría “Deportes y actividades al aire libre” pasando del 50,3% al 53,6%,

lo cual sumado al hecho de los porcentajes del País Vasco y Navarra han disminuido, convierten a Castilla y León al territorio con más personas realizando esta actividad.

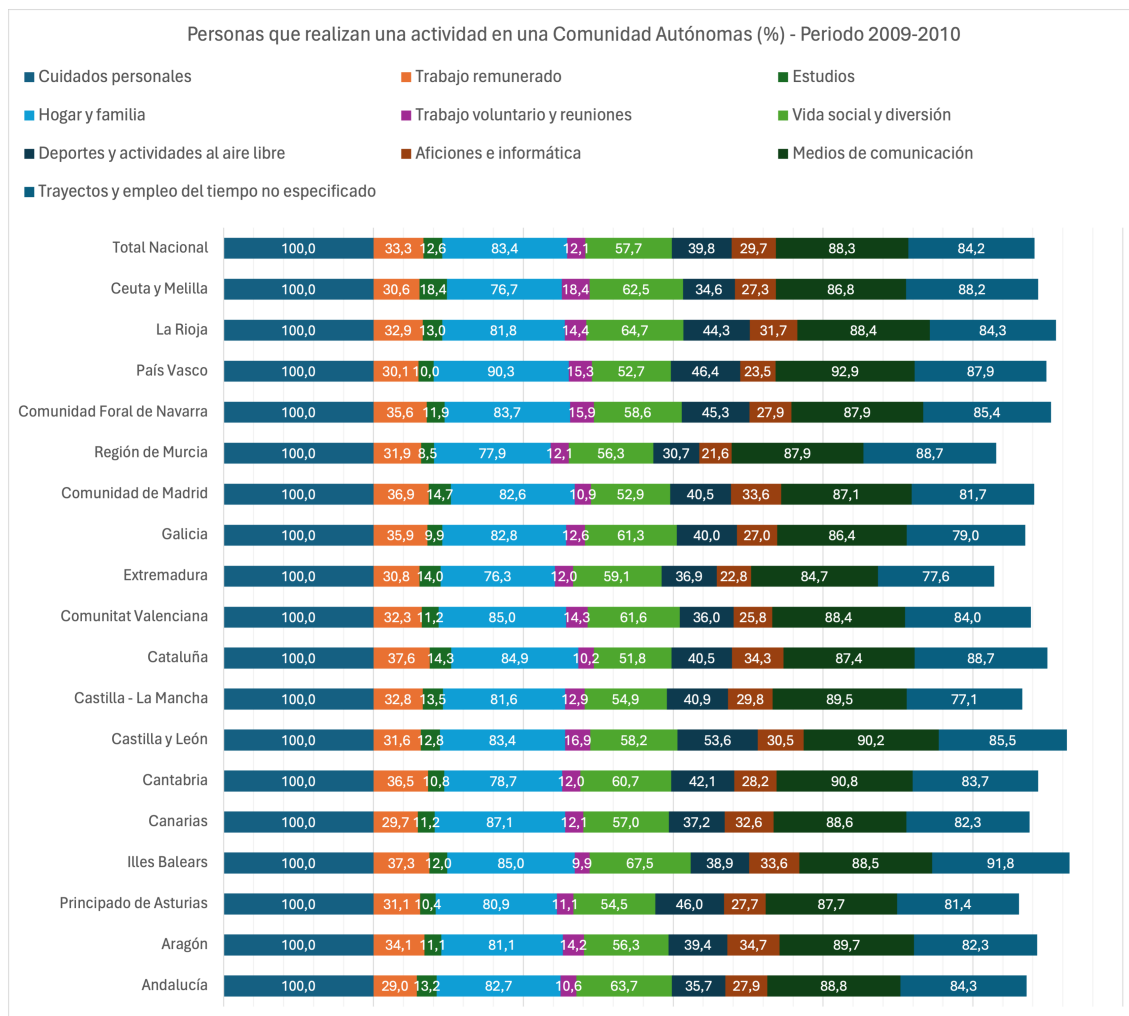


Figura 2.3.2: Porcentaje de personas que realizan una actividad en el transcurso del día para el periodo 2009-2010. Elaboración propia con datos del INE [25]

Por otro lado, las dos comunidades que claramente tiene estilos de vida menos dinámicos para el periodo 2009-2010 son la Región de Murcia y Extremadura. El caso murciano es especialmente llamativo, pues han pasado de tener un dinamismo superior al promedio a estar muy por debajo; los murcianos son los españoles que en promedio menos realizan las actividades “Estudios”, “Deportes y actividades al aire libre” y “Aficiones e informática” con porcentajes del 8,5%, 30,7% y 21,6%, respectivamente. Mientras que en Extremadura sucede lo mismo para las actividades “Hogar y familia” con un 76,3% y “Medios de comunicación” con un 84,7%.

Otro dato llamativo es que el periodo 2009-2010 había 9 territorios con valores de dinamismo, porcentajes de cada actividad sumados, superiores al Total Nacional, mientras que en el 2002-2003 eran 11 territorios. Si tenemos en cuenta que el Total Nacional fue superior en 2009-2010 que en 2002-2003, esto nos indicaría que las autonomías siguen tendencias divergentes en cuanto a dinamismo del estilo de vida, pudiendo dar lugar a desigualdades.

En resumen, el análisis de las actividades en las Comunidades Autónomas muestra una tendencia uniforme en actividades como “Medios de comunicación”, “Trayecto y empleo del tiempo no especificado” y, por supuesto, en “Cuidados personales”, mientras que el “Trabajo remunerado”, “Vida social y diversión” y “Deporte y actividades al aire libre” presentan variaciones significativas entre las distintas comunidades. Todo este análisis, también permite obtener una primera aproximación del dinamismo en los territorios de España y otras conclusiones a nivel intuitivo.

2.3.2 Resultados del *Lifestyle Dynamics Index*

Hasta el momento únicamente se ha realizado un análisis exploratorio pormenorizado de los datos, pero como ejercicio para aterrizar ese análisis y sintetizar la información que se ha visto hasta el momento se ha procedido a calcular el LDI. Si realizamos el análisis de dinamismo en términos de LDI debemos tener en cuenta que, tal y como se ha expuesto en el apartado 2.1.1, no se analizará únicamente el número de actividades que realizan los encuestados, sino también las transiciones entre esas actividades y el número de ellas.

Aunque los resultados que intuitivamente se han obtenido durante el análisis exploratorio difieren poco de los resultados del LDI, existen ciertas diferencias. De esta forma, podemos decir que el LDI nos permite obtener con una metodología fiable el dinamismo a nivel de cada territorio, mientras que el análisis exploratorio de los datos nos permite explicar qué actividades son las que contribuyen más a ese dinamismo en el caso de cada Comunidad Autónoma.

En la Figura 2.3.3 se muestran los resultados del *Lifestyle Dynamics Index* para el periodo 2002-2003. En ella se puede ver, tal y como se ha anticipado en durante el análisis exploratorio que las comunidades más dinámicas son Castilla y León y la Comunidad Foral de Navarra, mientras que las menos dinámicas son Cataluña y Cantabria.

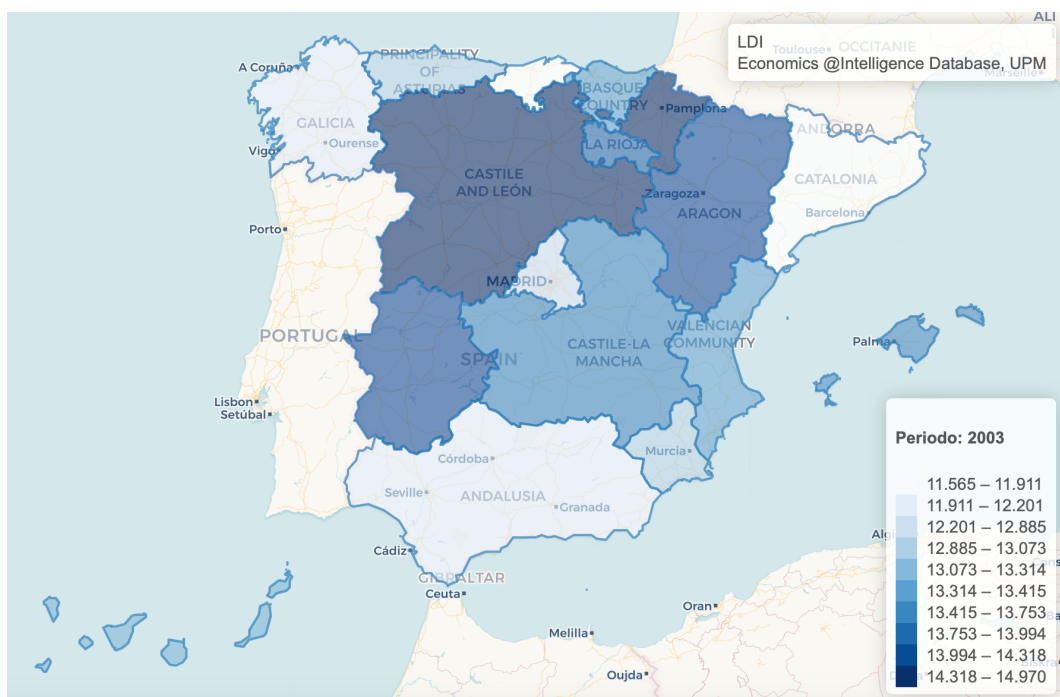


Figura 2.3.3: Resultados del LDI para el periodo 2002-2003

Las Comunidades Autónomas de La Rioja e Illes Balears, que *a priori* parecían ser las más dinámicas solo por detrás de Castilla y León y Navarra, tras el cálculo de LDI se observa que son superadas por Aragón y Extremadura; esto podría indicar que estas dos primeras autonomías tienen rutinas con transiciones más monótonas, es decir que tienen un menor número de transiciones entre actividades que Aragón y Extremadura o que realizan las mismas transiciones de forma más repetida.

Por otro lado, en la Figura 2.3.4 se pueden observar los resultados para el periodo 2009-2010, donde se puede ver que los territorios líderes en términos de dinamismo son Castilla y León y Baleares, seguidas de La Rioja y Aragón. En el otro extremo se encuentran Extremadura y Murcia.

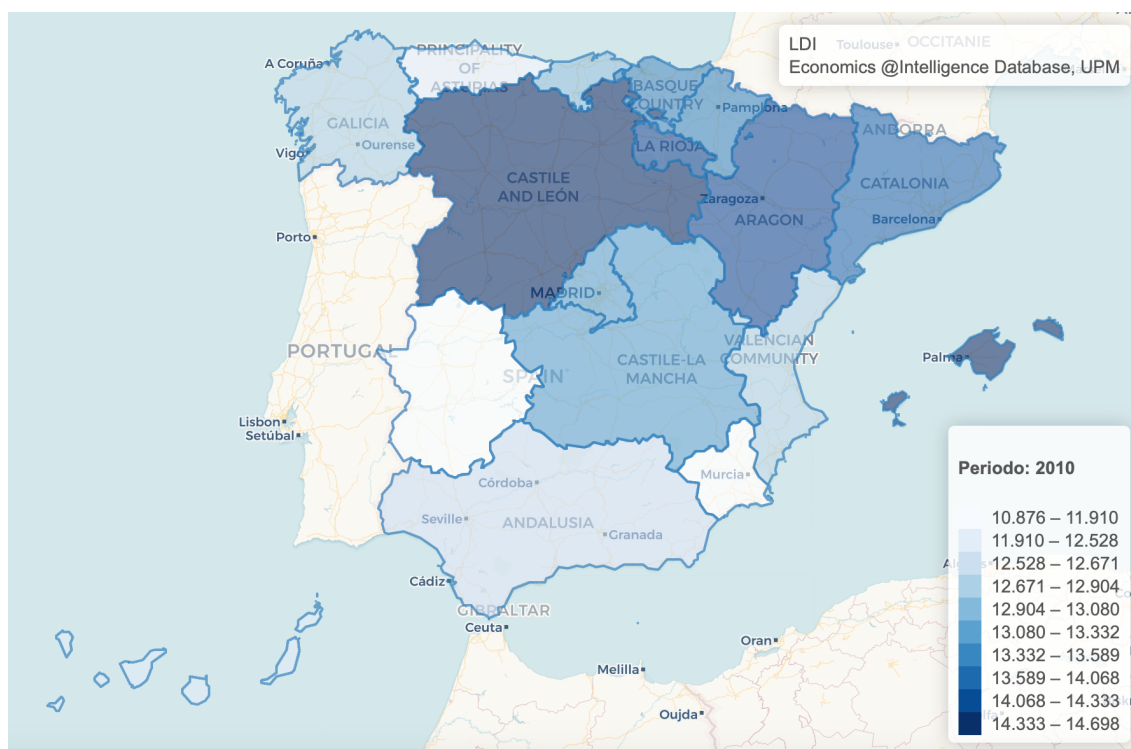


Figura 2.3.4: Resultados del LDI para el periodo 2009-2010

Si comparamos el cambio entre un periodo y otro, sirviéndonos de la Figura 2.3.5, se puede ver como las comunidades cuyo indicador crece más con respecto a los valores del periodo anterior son Cantabria, Cataluña y Ceuta y Melilla. Por otra parte, la mayor caída del LDI es la representada por Extremadura, seguida de la Región de Murcia, la Comunidad Foral de Navarra, la Comunidad Valenciana, el Principado de Asturias y Canarias.

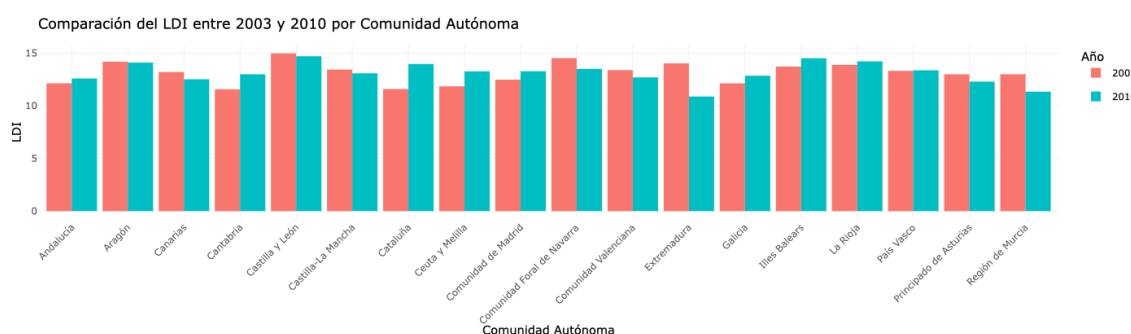


Figura 2.3.5: Evolución del LDI por Comunidades Autónomas

También crecen, pero en menor medida, Illes Balears, Comunidad de Madrid, Galicia y La Rioja. El resto de los territorios experimentan variaciones en el nivel de dinamismo de un periodo a otro poco significantes.

2.3.2.1 Variabilidad de los resultados del LDI con distintas metodologías

Una de las características más importantes del *Lifestyle Dynamics Index* es que puede calcularse con cualquier metodología de encuestas de uso del tiempo, es decir para cualquier número de actividades e intervalos de tiempo. No obstante, aunque se podría creer que esto implica que los resultados que produce son comparables, esto no es así; pues las salidas que produce el algoritmo o la fórmula del LDI son dependientes del número de actividades y del número de intervalos de tiempo utilizados por los motivos que se exponen en este apartado.

En primer lugar, se puede demostrar usando los propios datos del INE que en principio, si el LDI produjera resultados comparables entre metodologías, los resultados del LDI deberían ser iguales o muy similares para una misma Comunidad Autónoma y periodo estudiado con distintos niveles de granularidad en el número de actividades. Sin embargo, en las Figura 2.3.6 y Figura 2.3.7 se puede observar cómo esto no se cumple.

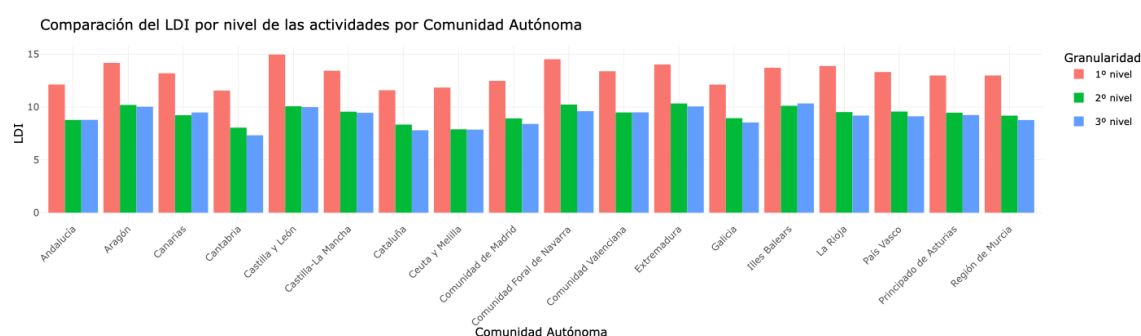


Figura 2.3.6: Comparación de los resultados del LDI para el periodo 2002-2003

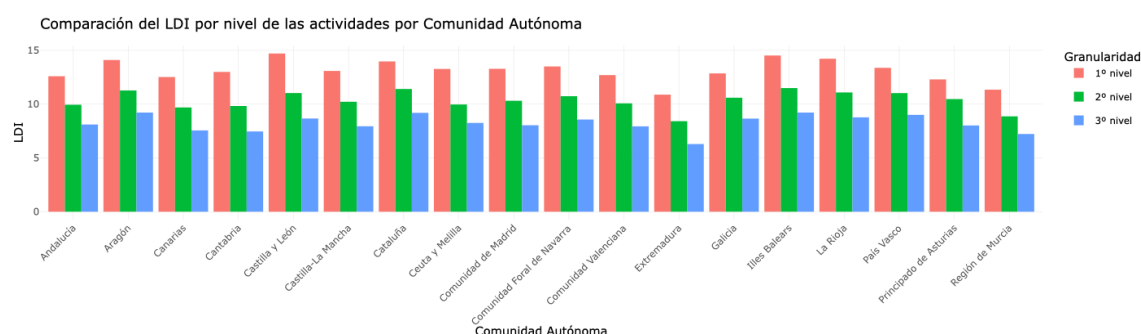


Figura 2.3.7: Comparación de los resultados del LDI para el periodo 2009-2010

Una forma de indagar en lo que está provocando esto es representar la distribución de los encuestados dentro del ranking que se genera para calcular el LDI, es decir, ver que posiciones del ranking son más frecuentes en cada caso y metodología y donde se concentran estas posiciones. Esto se muestra en las Figura 2.3.8 y Figura 2.3.9, donde el 0 representa la terna menos dinámica, (1, 0, 0) donde solo se realiza una actividad por día.

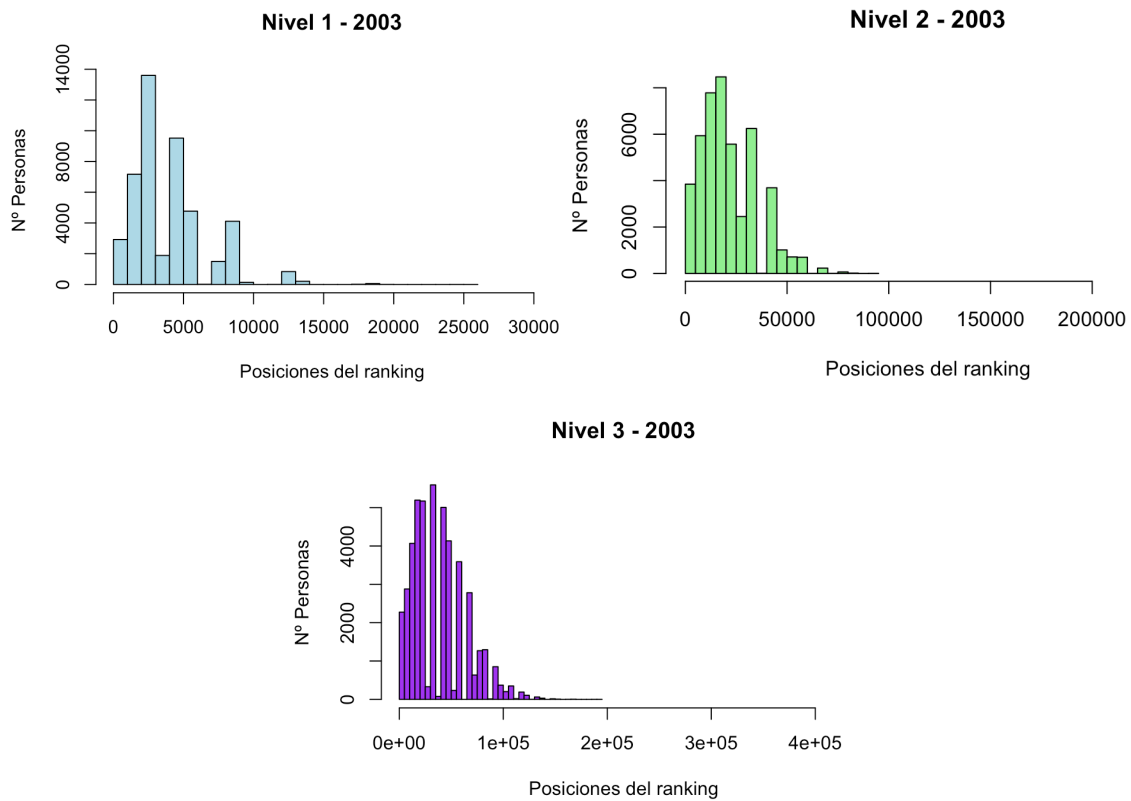


Figura 2.3.8: Distribución de los encuestados en el ranking para cada nivel de granularidad de las actividades para el periodo 2002-2003

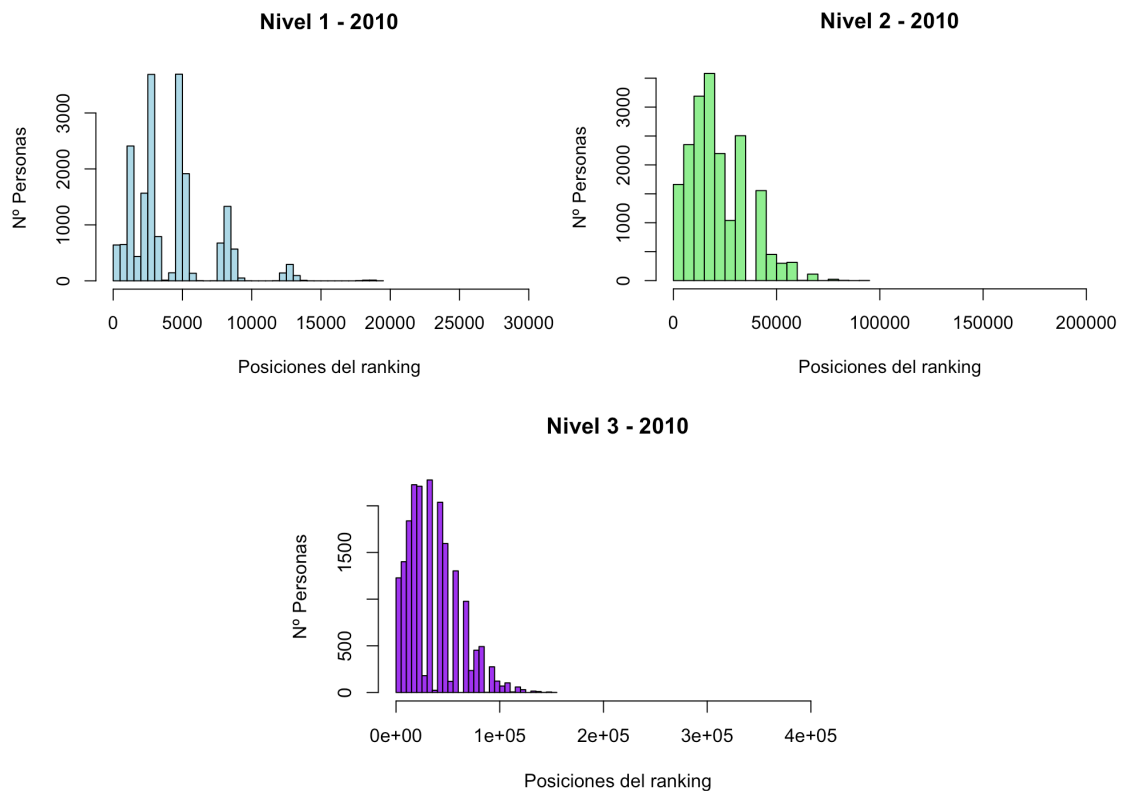


Figura 2.3.9: Distribución de los encuestados en el ranking para cada nivel de granularidad de las actividades para el periodo 2009-2010

Analizar estos gráficos con los puestos del ranking agregados a nivel español resulta revelador para comprender porque no son comparables los resultados del LDI con distintas metodologías. En el primer nivel de actividades, el de menor número de actividades, en el periodo 2002-2003 hay observaciones hasta la posición 25.261 sobre los 31.216 posibles puestos en el ranking; en el periodo 2009-2010 hay observaciones hasta el puesto 19.120 sobre un total de 31.216 posibles puestos del ranking. Se recuerda que un número más elevado significa una posición superior en el ranking.

En el segundo nivel, las posiciones más altas del ranking están en 92.949 y 92.211 de un total de 238.693 posibles puestos y 209.793 posibles puestos para el 2002-2003 y el 2009-2010, respectivamente. En el tercer nivel, el de mayor número de actividades, para el periodo 2002-2003 las posiciones más altas de la muestra del INE se sitúan en el 194.063 de un total de 431.718 posibles posiciones; mientras que para el periodo 2009-2010 las posiciones más altas de la muestra del INE se sitúan en el puesto 152.524 de un total de 428.036.

En definitiva, lo que se puede ver es que el número de posibles ternas aumenta, así como aumenta el número de actividades, pero que la distribución de las posiciones que ocupan los individuos dentro de ese ranking no tiende a mantenerse estable, sino que a mayor número de actividades las distribuciones tienden hacia una mayor asimetría positiva, es decir con los datos desplazados más a la izquierda en el gráfico, en posiciones menos dinámicas relativamente.

Esto también se puede ver con la posición promedio que en 2002-2003 para el nivel con menos actividades era la 3.969,69 (el top 12,72% menos dinámico en términos relativos), para el nivel intermedio era de 21.704,24 (el top 9,09% menos dinámico en términos relativos) y para el nivel con más actividades era de 38.156,36 (el top 8,88% menos dinámico en términos relativos). Para el periodo 2009-2010 en el nivel con menos actividades la posición promedio se situaba en 4.105,81 (el top 13,15% menos dinámico en términos relativos), en el nivel intermedio en 21.815,85 (el top 10,39% menos dinámico en términos relativos) y en el nivel con más actividades en 35.415,1 (el top 8,27% menos dinámico en términos relativos).

Esto sucede porque, aunque aumenten el número de actividades posibles que a nivel teórico los encuestados pueden realizar, lo cierto es que las personas rara vez realizan tantas actividades en su día a día; piense el lector en el caso del nivel con mayor detalle de actividades del INE con 192 y 116 categorías en los periodos 2002-2003 y 2009-2010 es tremendamente improbable, por no decir imposible, que un encuestado realice tantas actividades en un día. Esto se puede aplicar también al número de transiciones, a nivel teórico con la metodología del INE una terna del tipo (2, 2, 143) sería posible (dos actividades, dos transiciones distintas entre ellas y 143 transiciones totales entre ellas), pero en la realidad ningún ser humano realiza una rutina de este tipo; esto explicaría que las distribuciones de la Figura 2.3.8 y la Figura 2.3.9 sean tan similares, con barras que concentran un gran número de observaciones seguidas de barras más bajas, indicando que hay ciertas ternas que son mucho más improbables que se produzcan en una muestra real.

Este razonamiento también aplica a si se producen cambios en el número de los intervalos de tiempo entre metodologías. Por ejemplo, en una metodología con 1440 intervalos de tiempo, uno por minuto, a nivel teórico y a efectos del cálculo del LDI los encuestados podrían hacer 10 veces más actividades que si se usa una metodología donde se toman intervalos de 10 minutos como la del

INE; pero lo cierto es que el número de actividades en la práctica que realizarían las personas sería, cuando no fuese el mismo, solo ligeramente mayor.

Con todo esto, teniendo en cuenta la Ecuación 10 del cálculo LDI que normaliza por el número de filas del ranking, se tiene que para metodologías con un mayor número de actividades o más intervalos de tiempo se obtendrán valores menores del LDI, pues al normalizar por el número de posiciones en el ranking las metodologías que generen rankings con más posiciones incluirán un mayor número de esas ternas que en la práctica son extremadamente improbables. Esto no supone un problema si se comparan datos usando las mismas metodologías, pues este efecto se aplicará a todos los datos por igual; por ello, en este trabajo al comparar los datos del periodo 2002-2003 y 2009-2010 se está haciendo servir el nivel con menos actividades, dado que es el único con el mismo número de actividades e intervalos de tiempo para ambos periodos.

Si se desearán comparar datos con metodologías distintas sería necesario realizar una estandarización del número de actividades usando una metodología como la seguida por el CTUR para los datos del MTUS, basada en la metodología de Szalai, como se mencionó en el apartado 2.1, que implican el reagrupamiento de ciertas actividades en actividades más globales [38]. Otra alternativa sería modificar el índice para incluir ponderaciones para minorar el peso de esas ternas cuya probabilidad de que se den en la realidad es casi nula o directamente establecer un límite máximo de actividades distintas transiciones posibles para tener un ranking estable, independientemente del número de actividades e intervalos de tiempo, sin esas ternas que en la práctica no existen y que afectan al resultado del índice. No obstante, esta segunda alternativa solo se podría llevar a cabo si se realiza un estudio previo del máximo de encuesta del uso del tiempo posibles y se encuentra que existe un patrón, como el detectado en la Figura 2.3.8 y la Figura 2.3.9, que lleve a un “número mágico” que determine un máximo de actividades reales que se realizan en el día basándose en la evidencia empírica de esas encuestas.

Según los datos solo de estas encuestas del INE ese “número mágico” del máximo de actividades se situaría en el entorno de las 10 actividades, que son el número de actividades correspondientes a la terna de la posición promedio calculada con anterioridad para el nivel con más actividades y el nivel intermedio. No obstante, para el nivel con un menor grado de detalle en el número de actividades, el nivel 1, el número de actividades de la terna en la posición promedio es 5, lo que indica que en la vida real de los encuestados se producen transiciones reales entre actividades que no se reflejan con un nivel tan bajo de detalle haciendo que esas distintas actividades se registren en una única categoría a nivel del cuestionario.

Aplicar ese “número mágico” también haría que el LDI dejara de ser compatible con todas las metodologías, pues para metodologías cuyo número de posibles actividades fuera inferior o muy cercano a ese hipotético “número mágico” no podrían obtenerse resultados comparables, haciendo también necesario muy probablemente un número mínimo de posibles actividades que fuera superior a ese “número mágico”. Por una parte, se ganaría la capacidad de comparar los resultados entre la mayoría de las metodologías, pero por otra se perdería la capacidad de calcular el índice para cualquier metodología; lo cual *a priori* no sería muy problemático, pues la mayoría de las encuestas ofrecen los datos según varios niveles de detalle de las actividades y únicamente se tendrían que escoger los datos con las actividades más detalladas. En cualquier caso, implementar este “número mágico” como parte del LDI para poder comparar los resultados independientemente de la metodología podría conllevar otros

problemas no contemplados, por lo que sería necesario un análisis más profundo para ver todas las posibles implicaciones.

2.3.3 Clustering del LDI con otros indicadores socioeconómicos

En este apartado, se parte de los datos del LDI calculados anteriormente y de los indicadores socioeconómicos que se mencionan en el apartado 2.2.2 para realizar dos *clustering* con los datos de cada periodo. A continuación, se exponen los resultados para el año 2003 y 2010 que permite agrupar las Comunidades Autónomas con los valores más similares de esos indicadores.

2.3.4 Resultados para el año 2003

En la Figura 2.3.10 se muestran los resultados del *clustering* con los datos del 2003. En ella, se observa un primer clúster con la etiqueta 0 con formado por las regiones del mediterráneo, Madrid, Cantabria, Galicia y las Islas Canarias. Un segundo grupo conformado por las comunidades del centro peninsular (Aragón, La Rioja y las dos castillas), habitualmente referidas en los últimos años en los medios de comunicación como la “España vaciada”, junto con Asturias, Euskadi y Navarra. El segundo clúster, con etiqueta 2, está conformado por las Comunidades Autónomas de Extremadura y Andalucía.

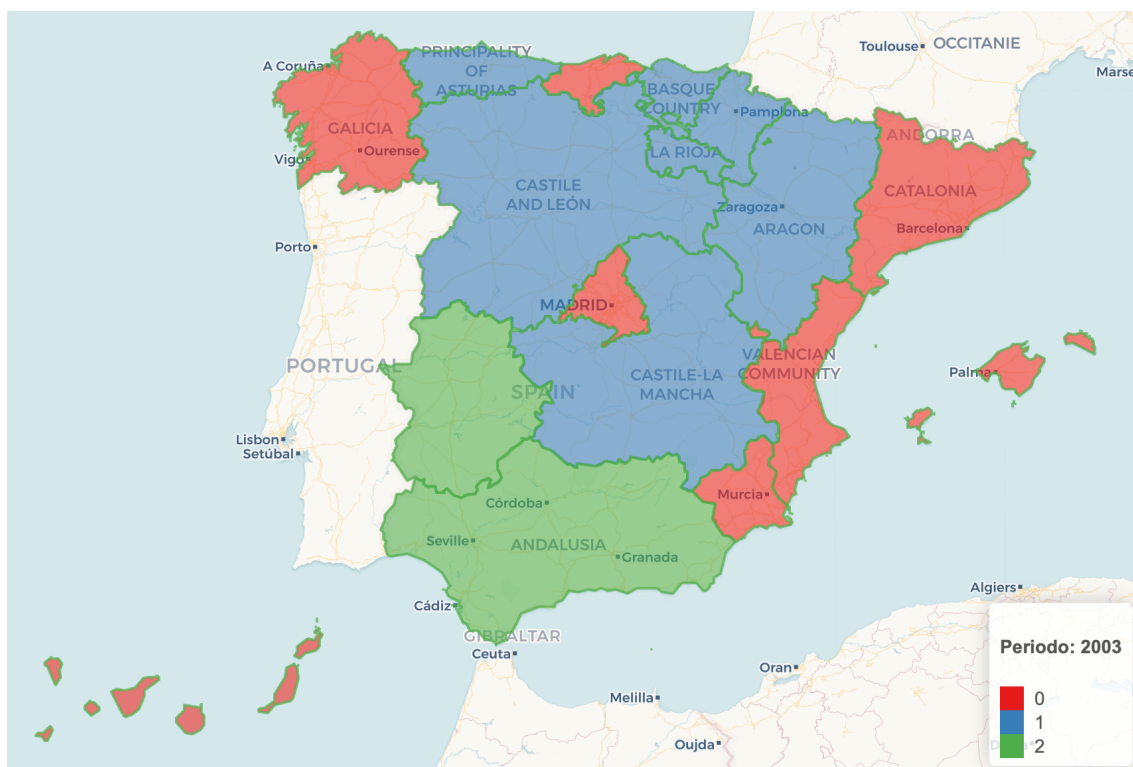


Figura 2.3.10: Mapa de los clústeres para los datos del 2003

Para explorar las características de los territorios dentro de un grupo específico, se aprovecharán gráficos de interpretabilidad global generados con el método SHAP, tal y como se ha explicado en el apartado 2.2.3.3. Estos gráficos permitirán identificar y medir los indicadores que tienen la mayor influencia en la asignación de las autonomías a un determinado clúster. Así pues, en la

Figura 2.3.11 se ofrece información sobre que indicadores han contribuido a la formación del clúster con etiqueta 0 en mayor medida. El clúster 0 se caracteriza por valores unos porcentajes de población que sufre problemas ambientales y de contaminación altos, con un valor promedio del 17,35% de la población y llegando al 20,7% en el caso de la Comunidad de Madrid. Otra característica diferencial es que tienen valores del LDI comparativamente bajos y una población más joven, con una edad promedio de 39,7 años. También se caracteriza por unos tasas de obesidad y de paro bajas o medias, con una tasa de obesidad en el entorno del 13% y unas tasas de paro a nivel promedio del 10,5%. Además, en menor medida, también contribuyen a la formación de este clúster unos valores del IPC y del PIB de bajos a moderados.

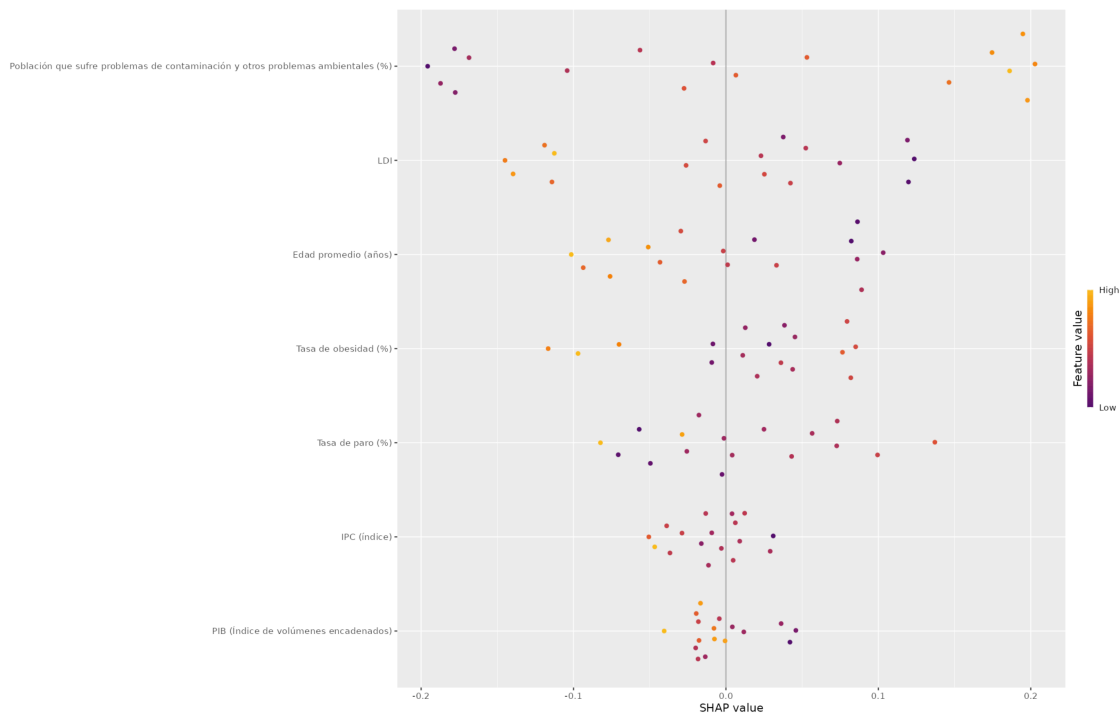


Figura 2.3.11: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 0 del año 2003

En segundo lugar, la Figura 2.3.13 muestra la importancia global de los indicadores que contribuyen en mayor medida a la formación del grupo con etiqueta 1. Este grupo se caracteriza por unos porcentajes de población con problemas de contaminación y ambientales bajos, situados en el 10%, destacando Castilla-La Mancha donde solo un 6,4% de la población sufre este tipo de problemas. Además, este clúster se caracteriza por una población más envejecida con un promedio de 42,3 años. También, se caracteriza por estilos de vida más dinámicos, es decir, con valores del LDI superiores y las tasa de paro más bajas de los tres clústeres situadas en un 8,6%. Otras características que contribuye en menor medida a la formación de este clúster son unas tasas de obesidad bajas y unos valores del PIB relativamente altos, destacando en este caso dentro del propio clúster las regiones del norte.

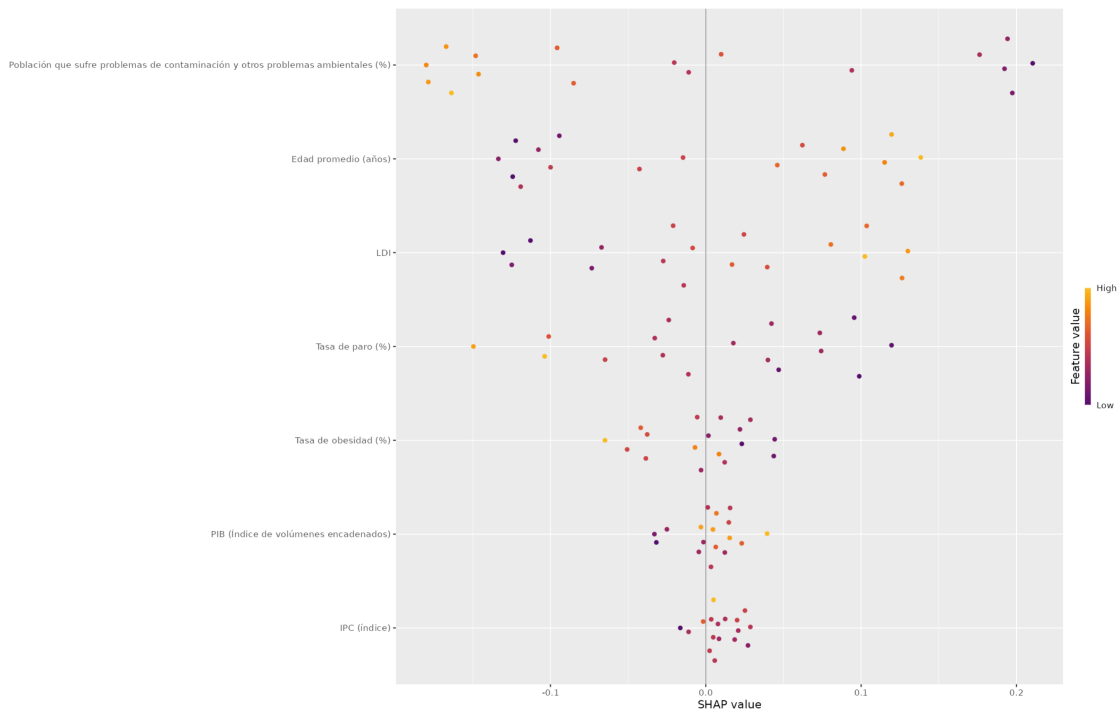


Figura 2.3.12: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 1 del año 2003

En tercer lugar, se tiene el clúster conformado por Andalucía y Extremadura, si se observa la Figura 2.3.13 se puede ver los factores más determinantes en la formación de este clúster son tasas de paro y obesidad inusualmente altas. El paro se sitúa en el 18,35% en Andalucía y el 17,14% en Extremadura, mientras que la obesidad se sitúa en el 16,92% y el 18,96% de la población en Andalucía y Extremadura, respectivamente. También contribuyen en menor medida a la formación de este clúster una población joven y unos niveles de inflación moderadamente altos.

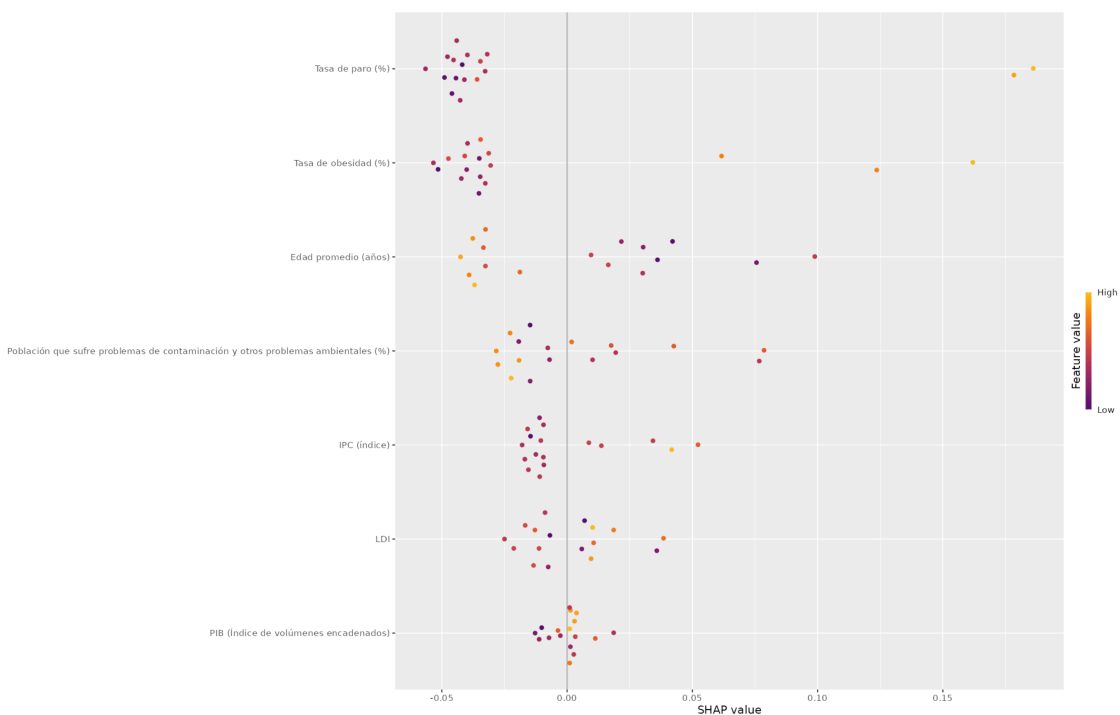


Figura 2.3.13: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 2 del año 2003

Todo esto, muestra que ya en 2003 existían diferencias entre las Comunidades Autónomas más rurales del centro de la península ibérica y las de la regiones de costa y la capital, teniendo por otra parte el caso atípico de Extremadura y Andalucía. Siendo las principales diferencias el porcentaje de personas afectadas por problemas de contaminación, la edad promedio de los habitantes y el dinamismo del estilo de vida en términos del LDI.

2.3.5 Resultados para el año 2010

Si se da un salto en el tiempo y se ven los resultados del *clustering* para el año 2010 (véase la Figura 2.3.14) se observa una clara diferenciación entre las regiones del norte y del sur. El primer clúster, con etiqueta 0, engloba las comunidades del sur de España, así como Galicia y el Principado de Asturias. El segundo clúster, con etiqueta 1, es el formado por Castilla y León, La Rioja, Comunidad Foral de Navarra y Aragón, que se podrían categorizar como las regiones del norte de España con un menor grado de industrialización si se las compara con el tercero de los clústeres, con la etiqueta 2, formado por territorios del norte de España más industrializados: Madrid, Cataluña País Vasco, Cantabria y Baleares. Para ver que caracteriza a cada uno de estos grupos de Comunidades Autónomas se hará uso de los siguientes gráficos de con los valores SHAP de importancia global.

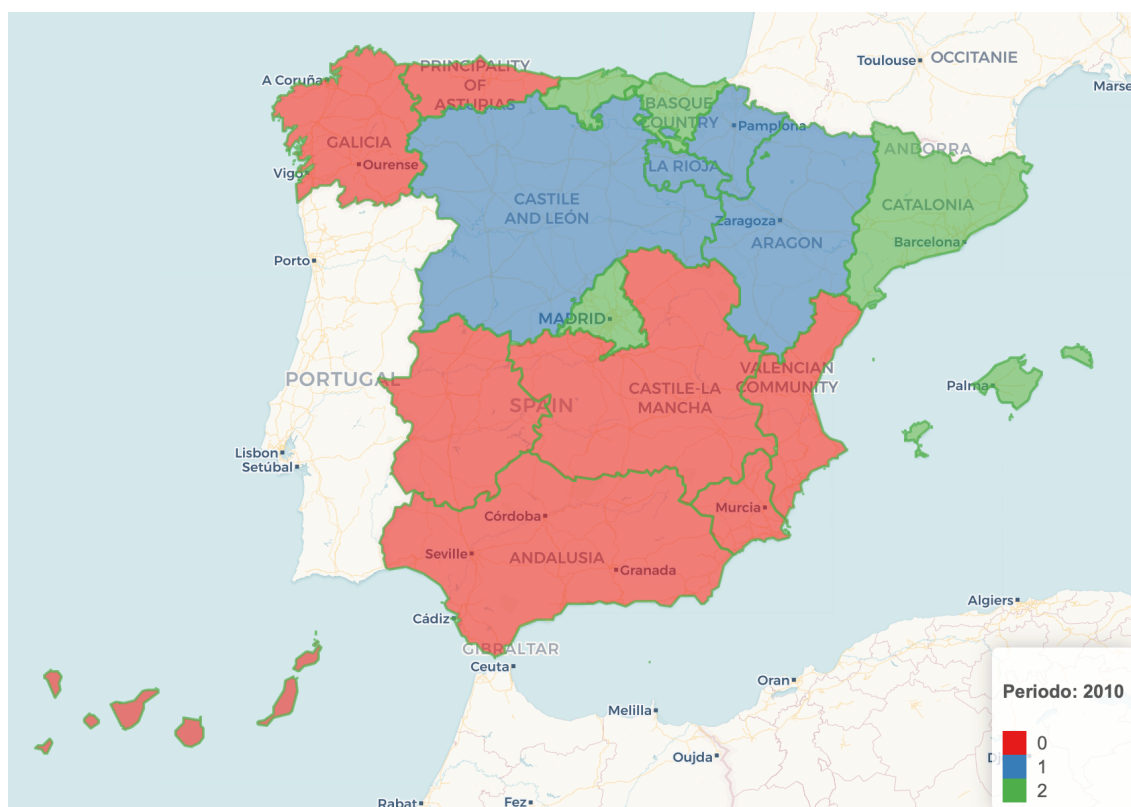


Figura 2.3.14: Mapa de los clústeres para los datos del 2010

En primer lugar, el clúster con etiqueta 0 (véase la Figura 2.3.15) se caracteriza por unas tasas de obesidad elevadas, con un 19,6% de la población obesa, y

valores del LDI bajos, indicando estilos de vida poco dinámicos en los territorios de este grupo; tanto en el sentido del LDI, pocas actividades y variabilidad baja entre ellas, y en el sentido del movimiento físico vinculado a esas actividades. Otra característica de estas autonomías son unas tasas de paro elevadas que se sitúan en promedio en el 22,7%, llegando al 28,8% en el caso de Canarias. Luego, también contribuyen a la formación de este clúster en menor medida unos niveles de PIB medianos y una edad promedio baja.

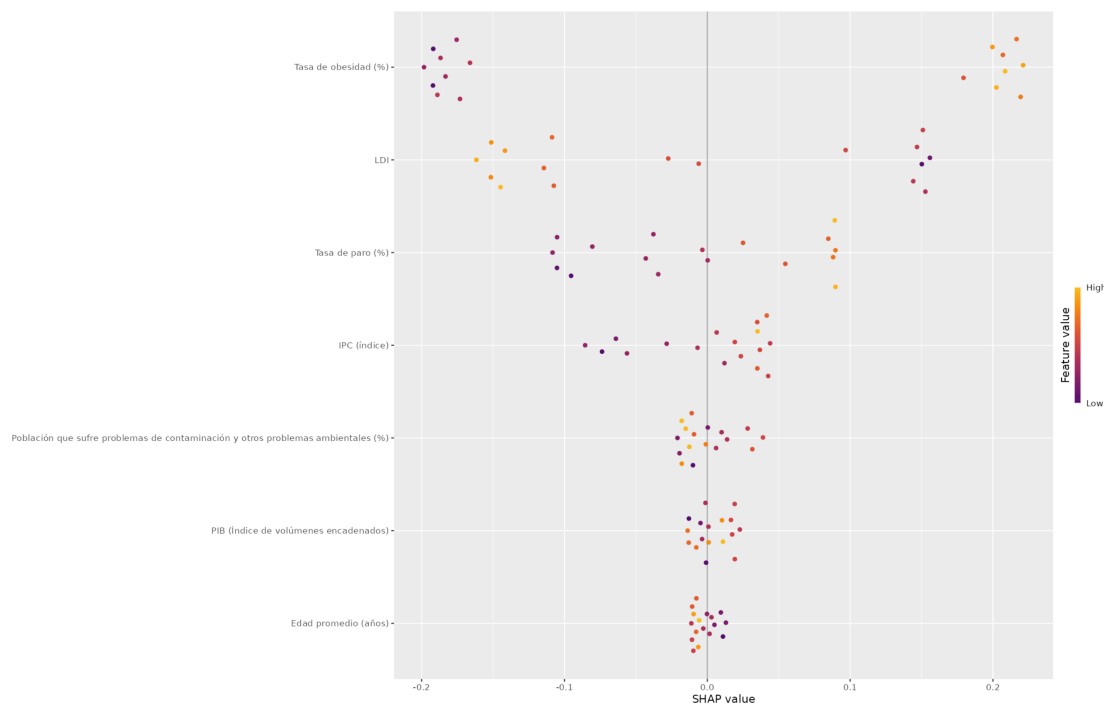


Figura 2.3.15: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 0 del año 2010

En segundo lugar, el clúster con etiqueta 1 representado en la Figura 2.3.16 se caracteriza por unas tasas de obesidad bajas, del 13,9% de la población en promedio, y por valores del LDI elevados; mostrando unos estilos de vida más dinámicos. Otra característica relevante de este grupo de autonomías es el bajo porcentaje de personas afectadas negativamente por la contaminación y otros problemas medioambientales, un 4,9% de la población en términos promedio. Por último, se caracterizan por tener tasa de paro bajas, del 14,9% en media, y unos niveles de PIB y IPC medios.

En tercer lugar, el clúster con etiqueta 2, como se muestra en la Figura 2.3.17, por unos tasas de obesidad bajas, del 14% en promedio, y unos valores del LDI medios; mostrando unos estilos de vida más dinámicos que el clúster 0, pero menos dinámicos que el clúster 1. Otra característica diferencia con respecto al clúster 1 es que los territorios del clúster 2 tiene porcentajes de población que sufre problemas de contaminación y otros problemas ambientales mayores, situados en el 12,7% de la población en promedio. Además, el clúster 2 se caracteriza por unos niveles de inflación bajos y tasas de paro bajas. En menor medida, también contribuyen a la formación de este clúster un PIB relativamente alto y una edad promedio baja.

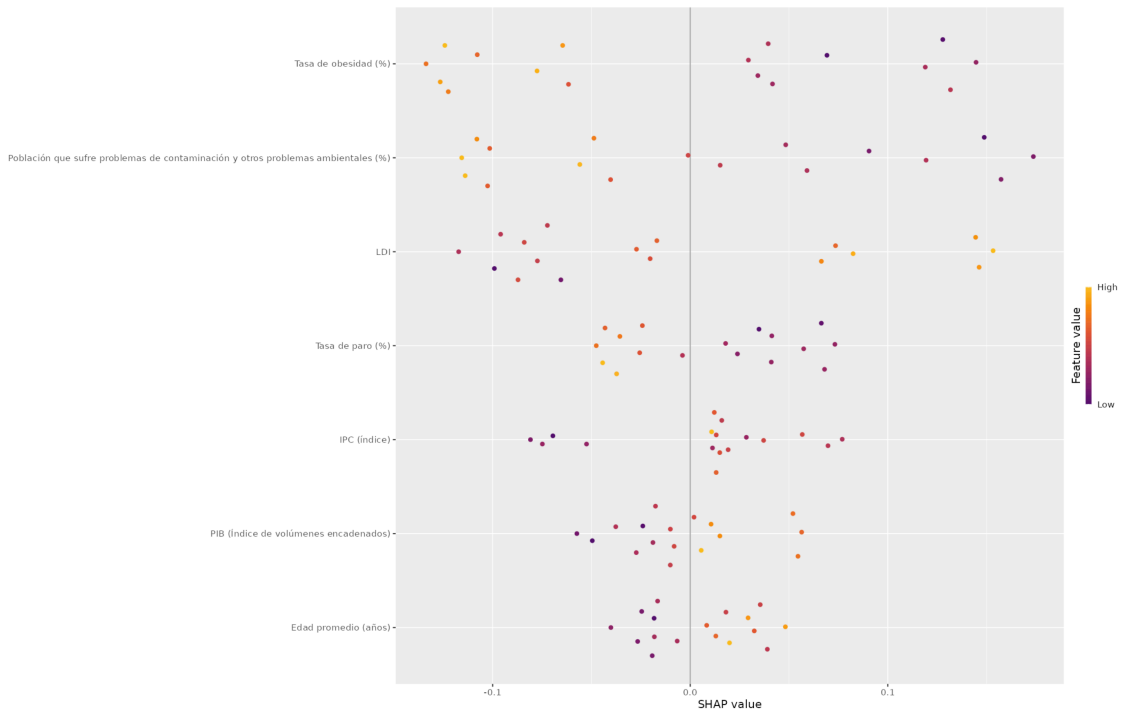


Figura 2.3.16: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 1 del año 2010

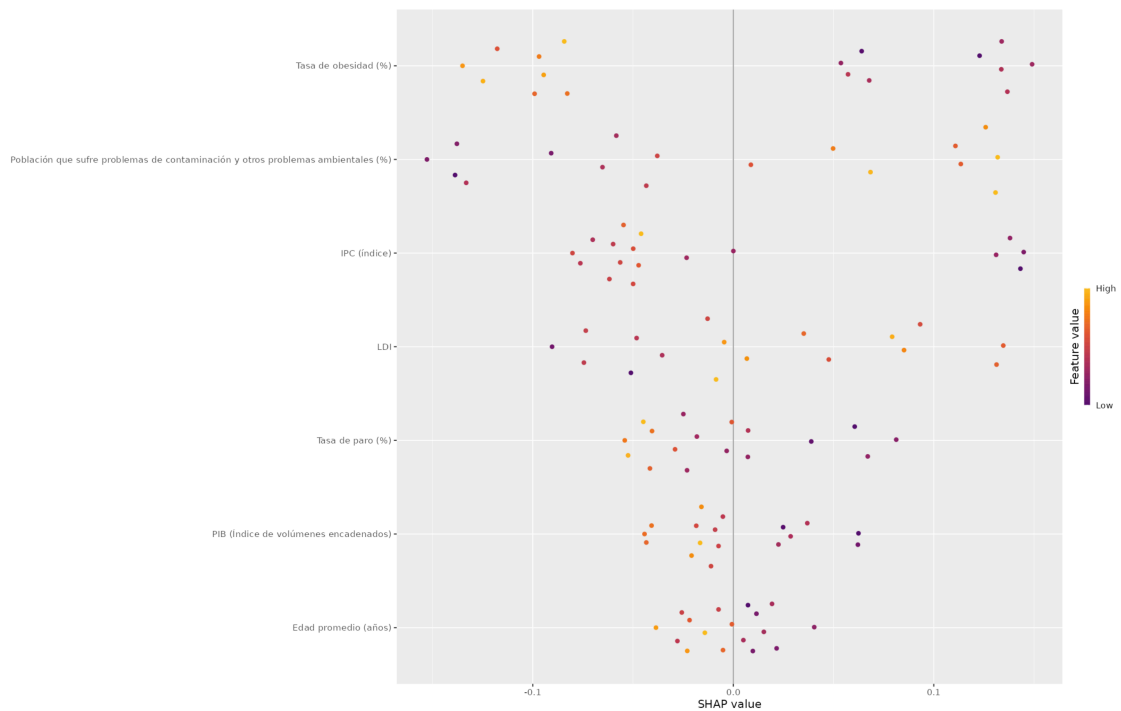


Figura 2.3.17: Gráfico de los valores de SHAP para la interpretabilidad global del clúster 2 del año 2010

En definitiva, lo que se puede concluir comparando los resultados de ambos *clusterings* es que el del año 2003 saca a relucir las diferencias entre los territorios más rurales, apodados como la “España vaciada”, y los territorios más industrializados, más turísticos y con una población más joven y urbanita. Mientras que el *clustering* del año 2010 muestra las diferencias norte-sur dentro de los distintitos territorios españoles, donde los territorios del norte son más

dinámicos con respecto a los del sur tanto a nivel económico, con tasa de paro más bajas, niveles de inflación moderados en caso del clúster 1 y bajos en el clúster 2 y un mayor PIB en el clúster 1, como a nivel del estilo de vida, con tasas de obesidad inferiores y valores del LDI superiores.

3 Resultados y conclusiones

Este Trabajo Fin de Grado (TFG) ha llevado a cabo un análisis detallado del uso del tiempo en España, enfocado en las Comunidades Autónomas, mediante las Encuestas de Empleo del Tiempo (EET) de los periodos 2002-2003 y 2009-2010. El principal aporte del estudio es la creación y aplicación del *Lifestyle Dynamics Index* (LDI), un índice diseñado para cuantificar las dinámicas de vida en diferentes regiones. Este índice se calculó utilizando todos los niveles de categorización de actividades proporcionados por las EET, lo que permitió evaluar la distribución del tiempo en cada comunidad y observar las diferencias entre los periodos estudiados, así como concluir que el LDI no produce resultados comparables entre metodologías.

En la segunda fase de análisis, se realizó un *clustering* utilizando el LDI y otros indicadores socioeconómicos relevantes. Este análisis permitió agrupar las comunidades en función de sus similitudes, revelando patrones tanto de similitud como de divergencia entre las diferentes regiones. Por ejemplo, en el año 2003, se identificaron tres clústeres distintos: uno formado por regiones del Mediterráneo, Madrid, Cantabria, Galicia y las Islas Canarias; otro por las comunidades del centro peninsular y el norte de España; y un último clúster compuesto por Extremadura y Andalucía. En el año 2010, los clústeres reflejaron diferencias más pronunciadas entre el norte y el sur de España. Las regiones del norte mostraron mayores niveles de dinamismo económico y estilos de vida más saludables, con tasas de obesidad y desempleo más bajas, así como un PIB más alto.

Las conclusiones del estudio subrayan la importancia de analizar el uso del tiempo y las dinámicas de vida para comprender mejor las diferencias regionales y su impacto en el desarrollo socioeconómico. Los resultados muestran que existen disparidades significativas entre las Comunidades Autónomas en términos de distribución del tiempo y condiciones socioeconómicas.

El análisis del LDI ha demostrado ser una herramienta eficaz para cuantificar el dinamismo del estilo de vida y comparar las diferentes regiones. La identificación de clústeres ha permitido agrupar comunidades con características similares, que hubieran podido facilitar la formulación de políticas públicas más específicas y dirigidas a abordar las necesidades únicas de cada región.

4 Análisis de Impacto

Este apartado examina el impacto potencial de los resultados del Trabajo Fin de Grado (TFG) en diversos contextos, incluyendo el personal, empresarial, social, económico, medioambiental y cultural, estableciendo una relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030.

A nivel personal, los resultados pueden aumentar la conciencia sobre el uso del tiempo, empoderando a las personas para tomar decisiones más informadas. Sin embargo, también podrían causar estrés al revelar ineficiencias en la gestión de su tiempo. En el ámbito empresarial, los hallazgos pueden optimizar los recursos humanos y facilitar la implementación de políticas de flexibilidad laboral, aunque esto puede conllevar costes adicionales y resistencia interna. Socialmente, una mejor gestión del tiempo puede mejorar la calidad de vida promoviendo estilos de vida más dinámicos y saludables, y promover la igualdad de género en términos de uso del tiempo especialmente con las actividades relacionadas con el hogar y la familia, aunque existe el riesgo de exacerbar desigualdades regionales si no se implementan políticas equitativas.

Económicamente, la optimización del tiempo puede aumentar la productividad y contribuir al desarrollo regional equilibrado. Medioambientalmente, una mejor gestión del tiempo puede fomentar actividades sostenibles y reducir la necesidad de desplazamientos diarios, lo que disminuye la huella de carbono. Culturalmente, una mayor sensibilización del uso que se realiza del tiempo podría provocar cambios en las rutinas y estilos de vida de las personas, aunque algunos sectores podrían resistirse a cambiar sus hábitos establecidos.

Este estudio contribuye directamente a los ODS 8 (Trabajo decente y crecimiento económico), ODS 10 (Reducción de las desigualdades) y ODS 11 (Ciudades y comunidades sostenibles). En conjunto, este TFG proporciona una base sólida para entender y mejorar las dinámicas de vida en España, promoviendo un desarrollo más equitativo y sostenible. Aunque, para que estos conocimientos fueran de aplicabilidad hoy en día sería necesario que el INE siguiera publicando las EET para tener datos más actualizados.

5 Bibliografía

- [1] R. G. Sanchis and J. Vicente-Pérez, "Lifestyle dynamics index: Worldwide results and the case of the United States of America," 2021. [Online]. Available: <https://doi.org/10.1016/j.seps.2021.101172>. [Accessed 15 April 2024].
- [2] P. P. Gronsky, "The Zemstvo System and Local Government in Russia, 1917-1922," 1923. [Online]. Available: <https://www.jstor.org/stable/2142478?seq=1>. [Accessed 13 April 2024].
- [3] C. AV, "Peasant farm organization. The theory of peasant economy," 1966. [Online]. Available: http://www.ciesas.edu.mx/publicaciones/clasicos/00_CCA/Articulos_CCA/CCA_PDF/035_CHAYANOV_On_the_theory.pdf . [Accessed 13 April 2024].
- [4] Maud Pember-Reeve, «A survey carried out by Fabian Society's Women's Group of families living on an income of 1978,» de *Round about a pound a week*, 1913, pp. 18-26.
- [5] S. Strumlin, «Time Budgets of Russian Workers in 1923–1924,» *Planovoe khozyaistvo*, vol. VII, 1925.
- [6] H. Kneeland, «Woman's economic contribution in the home.,» *Ann Am Acad Pol Soc Sci*, vol. 143, n° 1, p. 33–40, 1929.
- [7] A. Bauman, M. Bittman y J. Gershuny, «A short history of time use research; implications for public health,» 2019. [Online]. Available: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6760-y>. [Accessed 13 April 2024].
- [8] F. W. Taylor, «The principles of scientific management,» 1911. [Online]. Available: <https://archive.org/details/principlesofscie00taylrich>. [Accessed 13 April 2024].
- [9] D. A. Spencer, «Cambridge Journal of Economics,» 2023. [Online]. Available: <https://doi.org/10.1093/cje/bead046>. [Accessed 13 April 2024].
- [10] A. Szalai, "The use of time: daily activities of urban and suburban populations in twelve countries first edition,," 1972. [Online]. Available: <https://www.cabidigitallibrary.org/doi/full/10.5555/19821888623>. [Accessed 13 April 2024].
- [11] G. Becker, "A theory of the allocation of time," 1965. [Online]. Available: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/eoj.12157>. [Accessed 13 April 2024].
- [12] D. Ironmonger, *New commodities and consumer behaviour*, 1972.
- [13] H. Varian, «La utilidad,» de *Microeconomía intermedia*, Antoni Bosch, editor, 2010, pp. 58-74.
- [14] J. S. Mill, " Utilitarianism," in *Seven masterpieces of philosophy*, Routledge, 2006, pp. 329-375.

- [15] N. Ahn, J. F. Jimeno and A. Ugidos, "“Mondays at the sun”: Unemployment, Time Use, and Consumption Patterns in Spain," 2003-18. [Online]. Available: <https://documentos.fedea.net/pubs/dt/2003/dt-2003-18.pdf>. [Accessed 13 April 2024].
- [16] U.S. Bureau of Labor Statistics, "American Time Use Survey," [Online]. Available: <https://www.bls.gov/tus/>. [Accessed 13 April 2024].
- [17] Eurostat, «HARMONISED EUROPEAN TIME USE SURVEYS (HETUS),» 2024. [Online]. Available: <https://ec.europa.eu/eurostat/web/time-use-surveys/information-data/projects#expand-at-1712073719847>. [Accessed 14 April 2024].
- [18] UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, "Guidelines for Harmonizing Time-Use Surveys," 2013. [Online]. Available: https://unece.org/DAM/stats/publications/2013/TimeUseSurvey_Guidelines.pdf. [Accessed 14 April 2024].
- [19] G. Meil and J. Rogero, "Las Encuestas de Empleo del Tiempo como instrumento para analizar el reparto del trabajo remunerado y no remunerado," April 2012. [Online]. Available: <http://www.revistaindice.com/numero51/p20.pdf>. [Accessed 15 April 2024].
- [20] Instituto Nacional de Estadística, "Encuesta de Empleo del Tiempo 2009-2010. Metodología," Julio 2011. [Online]. Available: <https://www.ine.es/metodologia/t25/t25304471.pdf>. [Accessed 8 April 2024].
- [21] Eurostat, «Eurostat grant projects,» [Online]. Available: <https://ec.europa.eu/eurostat/web/time-use-surveys/information-data/projects#expand-at-1712073719847>. [Accessed 8 April 2024].
- [22] S. Ghassemi-Bönisch, L. Fürst and F. Csaicsich, "Statistics Austria," Diciembre 2019. [Online]. Available: https://ec.europa.eu/eurostat/documents/9986036/10970802/GRANT_2018_HETUS_AT_2019_STAT_time_UserGuide.pdf. [Accessed 8 April 2024].
- [23] K. Sabbe, E. Nagel, B. Lenuweit and J. Minnen, "Eurostat," April 2020. [Online]. Available: https://ec.europa.eu/eurostat/documents/9986036/10970806/847218_BE_final+report_Speech_Recognition.pdf. [Accessed 8 April 2024].
- [24] T. Harms, C. Pattison y D. Olaru, «USING WEARABLE TECHNOLOGY TO ANALYSE DAILY TRAVEL BEHAVIOUR,» 2019. [Online]. Available: <https://www.timeuse.org/sites/default/files/2020-11/Active%20Travel%20Study%20Summary.pdf>. [Accessed 8 April 2024].
- [25] Instituto Nacional de Estadística, "Encuesta de empleo del tiempo. Resultados," [Online]. Available: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176815&menu=resultados&idp=1254735976608. [Accessed 9 April 2023].

- [26] Instituto Nacional de Estadística, "Diario de Empleo del Tiempo," [Online]. Available: https://www.ine.es/daco/daco42/empleo/cues_dactividad10.pdf. [Accessed 8 April 2024].
- [27] Instituto Nacional de Estadística, «Encuesta de Empleo del Tiempo, Ficha histórica: 30447,» [Online]. Available: <https://www.ine.es/dyngs/IOE/es/fichaHistorica.htm?cid=1259930838691>. [Accessed 9 April 2024].
- [28] Instituto Nacional de Estadística, «IRIA,» [Online]. Available: <https://iria.ine.es/iriaPortal/contents/index.jsf>. [Accessed 8 April 2024].
- [29] Economics @Intelligence, "Lighthouse," [Online]. Available: <http://vps159.cesvima.upm.es/lighthouse/>. [Accessed 12 May 2024].
- [30] Instituto Nacional de Estadística, «Contabilidad Regional de España,» [Online]. Available: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736167628&menu=ultiDatos&idp=1254735576581. [Accessed 12 May 2024].
- [31] Instituto Nacional de Estadística, "Indicadores de Estructura de la Población," [Online]. Available: <https://www.ine.es/jaxiT3/Tabla.htm?t=3198&L=0>. [Accessed 12 May 2024].
- [32] Posit, "Tidyverse packages," [Online]. Available: <https://www.tidyverse.org/packages/>. [Accessed 11 May 2024].
- [33] scikit-learn, "2.3. Clustering," [Online]. Available: <https://scikit-learn.org/stable/modules/clustering.html>. [Accessed 18 May 2024].
- [34] M. Natarajan, "Deciphering Optimal Clusters: Elbow Method vs. Silhouette Method," 17 October 2023. [Online]. Available: <https://medium.com/@megha.natarajan/deciphering-optimal-clusters-elbow-method-vs-silhouette-method-7e311c604201>. [Accessed 18 May 2024].
- [35] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, "NbClust: Determining the Best Number of Clusters in a Data Set," 2 May 2022. [Online]. Available: <https://cran.r-project.org/package=NbClust>. [Accessed 18 May 2024].
- [36] "Welcome to the SHAP documentation," [Online]. Available: <https://shap.readthedocs.io/en/latest/index.html>. [Accessed 18 May 2024].
- [37] M. Mayer and A. Stando, "shapviz: SHAP Visualizations," 12 January 2024. [Online]. Available: <https://cran.r-project.org/package=shapviz>. [Accessed 18 May 2024].
- [38] K. Fisher, "Time Use Research, History and Current Archives," Centre for Time Use Research, University of Oxford, 29 June 2016. [Online]. Available: https://www.mtusdata.org/mtus/resources/linked_docs/workshop2016/02_History.pdf. [Accessed 25 May 2024].

6 Anexos

6.1 Anexo I: Datos en bruto del *clustering*

Tabla 6.1.1: Datos en bruto para el *clustering* del año 2003. Elaboración propia con datos del INE [30] [31] y el Lighthouse [29]

	PIB (Índice de volúmenes encadenados)	Tasa de paro (%)	IPC (índice)	Edad promedio (años)	Población que sufre problemas de contaminación y otros problemas ambientales (%)	Tasa de obesidad (%)	LDI
Andalucía	92.2	18.35	74.052	37.87	15.5	16.92	12.1315991799077
Aragón	92.2	6.53	73.446	42.88	14.9	12.45	14.181829830856
Canarias	95.5	12.04	78.418	37.07	18.4	14.55	13.1983598154792
Cantabria	95.9	10.42	72.961	42.04	18.8	11.97	11.5645822655049
Castilla y León	95.5	11.02	73.113	43.78	11.1	11.74	14.9698872373142
Castilla-La Mancha	87	9.8	73.724	40.54	6.4	17	13.4354177344951
Cataluña	89	9.72	70.433	40.58	18	11.43	11.5902101486417
Comunidad de Madrid	83	7.16	73.82	39	20.7	9.23	12.4775756022553
Comunidad Foral de Navarra	87.4	6.16	74.29	41.02	8.6	13.63	14.5213992824193
Comunidad Valenciana	93.5	10.68	73.869	39.8	18.6	14.21	13.3841619682214
Extremadura	88.5	17.14	75.465	40.29	12.4	18.98	14.0216555612506
Galicia	87.3	13.46	73.288	43.18	12.2	14.32	12.1251922091235
Illes Balears	86.8	10.01	73.424	38.68	16.7	12.54	13.7205279343926
La Rioja	90	6.75	72.93	41.73	9	10.22	13.8839056893901
País Vasco	89	9.93	72.204	42.07	11.8	10.4	13.3136852895951
Principado de Asturias	97.2	10.14	74.325	44.29	10.1	12.86	12.9869297796002
Región de Murcia	85.2	10.76	73.935	37.2	15.4	15.38	12.9901332649923

Tabla 6.1.2: Datos en bruto para el *clustering* del año 2010. Elaboración propia con datos del INE [30] [31] y el Lighthouse [29]

	PIB (Índice de volúmenes encadenados)	Tasa de paro (%)	IPC (índice)	Edad promedio (años)	Población que sufre problemas de contaminación y otros problemas ambientales (%)	Tasa de obesidad (%)	LDI
Andalucía	100.8	28.04	88.314	39.2	10	21.15	12.5864941055869
Aragón	102.9	16.29	88.257	42.92	4.9	15.65	14.0953357252691
Canarias	101.1	28.78	90.314	39.18	6.6	18.22	12.5128139415684
Cantabria	104.5	14.67	87.156	43.03	12.9	11.12	12.9869297796002
Castilla y León	103.4	15.78	87.161	44.98	7.3	13.84	14.6975909789851
Castilla-La Mancha	104.5	21.72	87.933	40.64	4.5	20.59	13.0798308559713
Cataluña	99.9	17.86	85.998	40.79	14.9	14.25	13.9607893388006
Comunidad de Madrid	97.2	15.54	88.23	39.9	10.5	14.28	13.2784469502819
Comunidad Foral de Navarra	99.2	11.77	87.841	41.45	2.6	11.15	13.4962839569452
Comunidad Valenciana	101.3	22.57	88.148	40.77	14.7	17.33	12.6922091235264
Extremadura	101.4	23.95	88.676	41.81	8.1	21.64	10.8758329062019
Galicia	100.4	15.62	87.341	44.83	8.9	19.18	12.8491799077396
Illes Balears	95.9	21.78	87.091	39.35	10.5	15.09	14.511788826243
La Rioja	103.2	15.58	87.604	41.98	4.7	14.88	14.213864684777
País Vasco	99.6	10.98	86.749	43.36	14.8	15.26	13.3681445412609
Principado de Asturias	106.2	16.61	88.469	45.82	7.2	18.65	12.2917734495131
Región de Murcia	96	24.45	88.811	38.09	12.1	20.32	11.3371348026653

6.2 Anexo II: Valores promedio de cada clúster

Tabla 6.2.1: Valores promedio de cada clúster para cada indicador en 2003

Promedio Cluster	Promedio PIB (Índice de volúmenes encadenados)	Promedio Tasa de paro (%)	Promedio IPC (índice)	Promedio Edad promedio (años)	Promedio Población que sufre problemas de contaminación y otros problemas ambientales (%)	Promedio Tasa de obesidad (%)	Promedio LDI
0	89.525	10.531	73.768	39.694	17.35	12.954	12.631
1	91.186	8.619	73.433	42.33	10.271	12.614	13.899
2	90.35	17.745	74.758	39.08	13.95	17.95	13.077

Tabla 6.2.2: Valores promedio de cada clúster para cada indicador en 2010

Promedio Cluster	Promedio PIB (Índice de volúmenes encadenados)	Promedio Tasa de paro (%)	Promedio IPC (índice)	Promedio Edad promedio (años)	Promedio Población que sufre problemas de contaminación y otros problemas ambientales (%)	Promedio Tasa de obesidad (%)	Promedio LDI
0	101.463	22.718	88.501	41.293	9.012	19.635	12.278
1	102.175	14.855	87.716	42.832	4.875	13.88	14.126
2	99.42	16.166	87.045	41.286	12.72	14	13.621

Este documento esta firmado por



Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
Fecha/Hora	Fri May 31 20:14:46 CEST 2024
Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
Numero de Serie	561
Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)