FEDERACION ESPAÑOLA DE MUNICIPIOS Y PROVINCIAS

ORIENTACIÓN AL DATO PARA EL DESARROLLO DE TERRITORIOS INTELIGENTES

GRUPO DE TRABAJO”

COORDINADOR

Francisco Javier Delgado Álvarez

OA CIPSA – Diputación de Salamanca

INTEGRANTES

(Por orden alfabético)

Juan Alfaro Márquez	Diputación de Huelva
Fernando Álvarez García	Ayuntamiento de Madrid
Julio Cerdá Díaz	Ayuntamiento de Arganda del Rey
Fernando Gallego García	Ayuntamiento de Valencia
Manuel González Maya	Ayuntamiento de Sant Feliu de Llobregat
Patricia Moreno Atanasio	Ayuntamiento de Viladecans
Juan Jesús Muñoz	Ayuntamiento de Madrid
Javier Peña Alonso	Diputación de Burgos
Esther Serrano Fernández	Diputación de Ciudad Real
Víctor Solla Bárcena	Ayuntamiento de Málaga

RESPONSABLE FEMP

Pablo Mª Bárcenas Gutiérrez

Secretario de la Comisión de Sociedad de la Información, Innovación Tecnológica y Agenda Digital

FEDERACION ESPAÑOLA DE MUNICIPIOS Y PROVINCIAS

TABLA DE CONTENIDO

1. INTRODUCCIÓN

La Agenda Digital Española presentó, a mediados de 2017, la estrategia de Territorios Inteligentes, que dio continuación al actual Plan Nacional de Ciudades Inteligentes, y que impulsa la aplicación de las Tecnologías de la información y la Comunicación al funcionamiento de las ciudades, el turismo, las zonas rurales y los servicios públicos. Con el objetivo de mejorar la eficiencia y el desarrollo económico, social y ambiental de las regiones españolas, municipios y sus agrupaciones, islas, territorios transfronterizos, etc. esta estrategia consolida las iniciativas ya emprendidas y anticipa el objetivo de avanzar hacia un nuevo modelo de país inteligente

Motivada por la mencionada Agenda, La Federación Española de Municipios y Provincias publicó en 2017 y 2019 sendas guías para fomentar la utilización de datos abiertos en administraciones locales, impulsando la publicación de grupos de datos de interés común. Estas iniciativas constituían los primeros pasos encaminados a avanzar en la explotación del dato en las administraciones locales, como herramienta imprescindible para el avance de la sociedad.

Por otro lado, tanto la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, modificada por el Real Decreto-ley 24/2021, de 2 de noviembre, así como el Reglamento (UE) 2018/1807 del Parlamento Europeo y del Consejo de 14 de noviembre de 2018 relativo a un marco para la libre circulación de datos no personales en la Unión Europea, o más recientemente, el Reglamento de Ejecución (UE) 2023/138 de la Comisión de 21 de diciembre de 2022 por el que se establecen una lista de conjuntos de datos específicos de alto valor y modalidades de publicación y reutilización, pretenden fomentar la creación de cadenas de valor basadas en el tratamiento de datos, para lo cual dichos datos deben estar disponibles y localizables, tanto para las

administraciones como para el sector infomediario que pretenda llevar a cabo su explotación.

La presente iniciativa se centra en la puesta en valor de los distintos conjuntos de datos tratados por las administraciones públicas y la identificación de métodos y algoritmos que permitan extraer de los mismos información y conocimiento que sirvan para avanzar en un desarrollo sostenible del territorio. Se responde así a una propuesta concreta formulada por la Comisión de Sociedad de la Información, Innovación Tecnológica y Agenda Digital, en su sesión de 21 de marzo de 2022. En dicha sesión se propuso la creación de un Grupo de Trabajo, que colaborase con la Comisión de Asuntos Económicos y Transformación Digital del Senado, con la intención de dar respuesta a los importantes desafíos que se afrontan en las zonas rurales, en especial la despoblación y la pérdida de servicios prestados a los ciudadanos. Este Grupo de Trabajo no constituye una duplicidad con relación a la Red Española de Ciudades Inteligentes (RECI) o a la Comisión de Desarrollo Rural y Pesca, ya que sus objetivos son distintos y se dirigen a un grupo de entidades de características marcadamente diferentes.

La mencionada Comisión de Sociedad de la Información, Innovación Tecnológica y Agenda Digital, en su sesión telemática del día 27 de junio, acordó formalmente la creación del pretendido Grupo de Trabajo y se procedió a concretar los objetivos del nuevo **Grupo de Trabajo de Orientación al Dato para el Desarrollo de Territorios Inteligentes**, así como el establecimiento de una metodología específica para el desarrollo de las diferentes tareas a llevar a cabo por sus miembros. Como ya se ha indicado, sus trabajos se centrarán en la puesta en valor de los distintos conjuntos de datos tratados por las entidades locales, así como la identificación de diferentes métodos y algoritmos que permitan extraer de los mismos información y conocimiento que sirvan para avanzar en un desarrollo sostenible del territorio.

2. ESTRUCTURA DE LOS TRABAJOS

Los trabajos que inicialmente se ha previsto sean llevados a cabo en el Grupo de Trabajo se desglosan en los siguientes hitos:

1. **Identificar desafíos** existentes en el mundo local que aparenten indiciariamente que puedan ser analizados, estudiados, atendidos y, en su caso, solucionados, desde una perspectiva basada en datos. En este apartado se formularán los desafíos, y sus enunciados, que parezca puedan ser abordados desde una perspectiva basada en datos.

2. **Obtener un conjunto de indicadores** de Estrategia de Territorio, aplicables al mundo local y especialmente al ámbito provincial, a partir, por ejemplo, de los Objetivos de Desarrollo Sostenible (ODS) y basándose en indicadores previamente establecidos por entidades tales como Naciones Unidas, UE, JRC [1], Estados o Ciudades. Se establecerán en este objetivo de manera inicial 2 o 3 ODS como objetivos de alto impacto en una primera fase, con una estimación de un máximo de 10 indicadores por objetivo. Esto resultará en un catálogo de entre 20 y 30 indicadores de Estrategia de Territorio, con sus definiciones, aplicables al mundo local, y especialmente al provincial, que aparenten tener repercusión en los desafíos formulados en el punto anterior.

3. **Identificar posibles fuentes de datos** asociadas al listado de indicadores. Se pretende la utilización de fuentes de datos ya existentes, pertenecientes a los denominados "datos abiertos", pero no obligatoriamente, que preferiblemente formen parte de conjuntos de datos de relevancia contrastada, como puedan ser el listado compilado por la FEMP o el elaborado para su aplicación en ciudades inteligentes, siguiendo, en todo caso, la estrategia promovida por el

[1] *Joint Research Centre*. Servicio de la UE para aportar asesoramiento científico independiente.

portal de datos abiertos "datos.gob.es". En caso de no existir en ese momento disponibilidad de los conjuntos de datos necesarios para el trabajo propuesto, se investigarán posibles conjuntos de datos alternativos o brókeres de datos que pudiesen facilitar esos datos necesarios. Obtendremos como resultado posibles fuentes de datos para la obtención de los indicadores que forman parte del catálogo correspondiente a los indicadores del apartado precedente.

4. **Analizar la aplicación de posibles herramientas, procedimientos, algoritmos o modelos** que, utilizando los conjuntos de datos identificados y provisionados en el apartado precedente, puedan ser de aplicación al análisis y resolución de las problemáticas o grandes desafíos establecidos en el punto primero. Esto nos permitirá evaluar herramientas, procedimientos, algoritmos o modelos que pudieran ser aplicados a los datos e indicadores para su análisis y, quizá, resolución de los desafíos, así como posibles casos de uso.
5. **Formular posibles propuestas de acción** específicas en el mundo local, donde la utilización de los datos y técnicas establecidas en los objetivos anteriores genere un alto impacto en el territorio, aportando beneficios para las personas que en el mismo residen, así como para el medio ambiente y restantes seres vivos que lo comparten. Finalmente se enumerarán posibles actuaciones que puedan concluirse de los resultados obtenidos y que puedan tener efectos positivos en la posible solución de las problemáticas o grandes desafíos establecidos en el apartado inicial. Es evidente que estas propuestas serán generales, ya que las posibles actuaciones concretas en cada territorio dependerán de los resultados obtenidos en los análisis realizados en cada caso, sin que, seguramente, existan recetas válidas para todos los casos que puedan considerarse.

3. HITO #1: IDENTIFICACIÓN DE DESAFÍOS

Los posibles desafíos cuyo análisis podría abordarse mediante el estudio de los datos asociados a los mismos pueden ser muy variados. En realidad, nos encontramos en un mundo cada vez más sometido a los datos, a los algoritmos que los datos alimentan, y cuyos resultados sirven, o debieran servir, de soporte al análisis y a las decisiones que se adoptan por nuestros responsables políticos. Se han detectado algunos problemas en nuestros municipios que ya se están intentando resolver mediante el uso de datos: desde la gestión del tráfico de vehículos en las ciudades, hasta la formulación de rutas óptimas para la recogida de residuos, pasando por la reducción de consumos energéticos, optimización del riego de jardines, estudios sobre la afluencia de personas, etc. Sin embargo, en aras a simplificar el trabajo a realizar por el Grupo de Trabajo, y ofrecer una respuesta alineada con el objetivo propuesto que sirvió de argumento a la creación del grupo, se ha resuelto que la opción idónea para los desafíos a abordar bien pudiera circunscribirse a los problemas inicialmente formulados por la Comisión, estos son, la **despoblación** y la **pérdida de servicios en las zonas rurales**.

3.1. DESAFÍO #1: LA DESPOBLACIÓN

La despoblación tiene importantes efectos en el territorio que la experimenta. En primer lugar, y como bien sabemos, la participación de las administraciones en los tributos recaudados por el estado, que en gran medida constituye la principal fuente de ingresos de muchas entidades locales, está vinculada a la población, al igual que sucede con algunas subvenciones. Esta disponibilidad de recursos económicos condiciona en gran medida los presupuestos de la entidad local correspondiente, así como las políticas que pueda llevar a cabo para intentar revertir los efectos de la despoblación, entrando así en un ciclo perverso con difícil

salida. A nivel del sector privado, la despoblación hace perder el interés de las empresas por establecerse en el territorio, al encontrarse con un reducido mercado potencial para su posible tráfico comercial.

En los últimos años se ha intensificado la movilización social en determinadas zonas de España exponiendo el grave problema de su despoblación. Inicialmente manifestado a través de los movimientos "Teruel Existe" y "Soria ¡Ya!", este activismo ha traspasado las fronteras provinciales y autonómicas, convirtiéndose en un grito social de socorro a lo largo de todo el territorio.

Se considera, a efectos objetivos, que forman parte de la denominada "España Vacia[da]" aquellas provincias que en las elecciones al Congreso eligen a 5 o menos diputados, lo que en la actualidad se corresponde con aquellas provincias con una población inferior a 700.000 habitantes[2]. Formarían parte de esta "España Vacia[da]" las siguientes 27 provincias (por orden alfabético): Álava, Albacete, Ávila, Badajoz, Burgos, Cáceres, Cantabria, Castellón, Ciudad Real, Cuenca, Guadalajara, Huelva, Huesca, Jaén, La Rioja, León, Lérida, Lugo, Navarra, Orense, Palencia, Salamanca, Segovia, Soria, Teruel, Valladolid y Zamora. Estas provincias representan aproximadamente el 21% de la población total de España, habiendo descendido desde el 23% que representaban hace 25 años. No obstante, en estos últimos 25 años, solo 12 de esas 27 provincias han experimentado un descenso en su población. Donde esta reducción se ha producido llega a representar en algunos casos casi el 20% de la población que tenían en 1996, como sucede en Zamora, con una bajada del 18,7% en su población.

[2] Fuente: Wikipedia "España Vaciada" [Consultado el 15 de octubre de 2022]

Si descendemos a nivel municipal, aproximadamente dos de cada tres municipios españoles han perdido población en los últimos 25 años. La práctica totalidad de ellos tienen menos de 20.000 habitantes, un 95% tiene menos de 5.000 habitantes y casi el 80% tienen menos de 1.000 habitantes, evidenciando así que el problema de la despoblación tiene mayor prevalencia en este grupo de municipios escasamente poblados.

Esta tendencia a la despoblación puede verse también claramente si analizamos el crecimiento del número de municipios muy pequeños: así, entre 1996 y 2021 el número de municipios de menos de 10 habitantes se ha multiplicado aproximadamente por seis y el de municipios de entre 11 y 100 habitantes se ha incrementado en ese mismo período en aproximadamente el 60%.

Todo esto, por lo que respecta a la evolución de la población en los últimos 25 años, pero si observamos la evolución prevista para los próximos 15 años[3], los resultados no son tampoco muy halagüeños. Albacete, Badajoz, Burgos, Cáceres, Ciudad Real, Jaén, León, Lugo, Orense, Palencia, Salamanca, Valladolid y Zamora, dentro del grupo de esa España Vaciada, experimentarán descensos de población, que en el caso de Zamora se estima aproximadamente en el 14%, siendo de nuevo este el caso más desfavorable de entre todas las provincias enumeradas. Alguna otra provincia, fuera inicialmente de ese grupo, tampoco presenta buenas expectativas. Así, Ceuta tendrá un descenso aproximado del 9% y Asturias otro descenso cercano al 7%. Para mayor diferenciación con la tendencia global en el estado, todos estos descensos están encuadrados dentro de un incremento estimado de población de casi el 9% para el total nacional en España durante en ese mismo periodo.

[3] Fuente: INE – Proyecciones de población, Resultados por provincias, Serie 2022-2037

En resumen, la despoblación de amplias zonas del territorio nacional se plantea como un importante desafío que las diferentes administraciones afrontan ya, o deben afrontar con urgencia. Las estrategias puestas en juego hasta la fecha no parecen haber tenido un efecto significativo para amortiguar estos descensos, toda vez que las series de población y las proyecciones a futuro siguen mostrando una continuada caída que no aparenta cambiar de tendencia. Dado el grado de capilaridad, o desagregación, y fiabilidad de los datos poblacionales disponibles de manera pública, este desafío se contempla como claramente abordable para el método de trabajo que se ha propuesto sea utilizado en el Grupo de Trabajo.

3.2. DESAFÍO #2: PERDIDA DE SERVICIOS EN LAS ZONAS RURALES

Con bastante probabilidad relacionado con el desafío anterior, nos encontramos la pérdida de servicios en las zonas rurales. Tal y como ya incidíamos en el punto anterior, la despoblación tiene como efecto más negativo la disminución de recursos públicos disponibles y la pérdida de interés en dichos territorios para las empresas. Ambas situaciones conllevan la ausencia de múltiples servicios que sí se encuentran disponibles en ciudades, incluso de tamaño medio. Aun siendo conscientes de que mejoras en la despoblación traerían sin mayores esfuerzos adicionales el incremento en la oferta disponible de múltiples servicios, no deja de ser menos cierto que la relación inversa puede ser igualmente formulada, y la disponibilidad de dichos servicios podría facilitar el asentamiento de población en el territorio. Afirmamos, pues, que entre ambos desafíos puede existir una clara correlación, siendo posible que la causalidad pueda tener amparo en cualquiera de los dos posibles sentidos. Por lo tanto, pretendemos explorar la pérdida de

servicios en las zonas rurales como un desafío independiente del poblacional, al menos, en un principio.

La primera cuestión que debemos plantearnos es a qué servicios nos referimos, considerando tanto aquellos suministrados por el sector público como los facilitados por el sector privado. Claramente en el primer caso, el del sector público, los ejemplos iniciales que podríamos exponer serían los de salud y educación, pero existen otros quizá menos evidentes, y con su relativa importancia, tales como la recogida de residuos y bibliotecas, entre otros, o con una provisión mixta, público-privada, como por ejemplo el transporte público. En el apartado privado tiene en la actualidad gran importancia la cobertura de redes de telecomunicación, tanto para telefonía fija y móvil, como para acceso a Internet a través de redes de banda ancha, igualmente fija o móvil. Pero tampoco podemos olvidar otros servicios más habituales, tales como, comercios, entre los que se encontrarían supermercados y tiendas de alimentación, peluquerías y tiendas de ropa, entre otros, o establecimientos de ocio, tales como bares, restaurantes, etc. Sin olvidar unos de los servicios que en los últimos años están experimentando un retroceso en su disponibilidad, como son los servicios bancarios, tanto de oficina presencial, como de cajeros automáticos. En resumen, se trataría de asimilar la amplia oferta de servicios disponibles en las ciudades que se encuentran fuera del ámbito rural.

A priori no parece difícil la obtención de datos sobre muchos o incluso de todos los servicios antes mencionados, bien a través de fuentes abiertas o también de la información tratada directamente por las administraciones, y que pudiese ser incorporada a nuestro estudio tras su adecuada anonimización, si procediese.

4. HITO #2: IDENTIFICACIÓN DE INDICADORES

En esta segunda fase identificaremos posibles indicadores de Estrategia del Territorio asociados a los dos desafíos que hemos identificado anteriormente. Ciertamente la lista de indicadores que propondremos estará muy lejos de poder considerarse cerrada, ya que cada organización/administración podrá valorar la inclusión de indicadores diferentes o su eliminación, en función de sus condicionantes particulares.

Los indicadores que buscamos podrán ser de dos tipos. En primer lugar, indicadores asociados directamente a nuestros desafíos, esto es, indicadores en gran medida descriptivos de dichos desafíos, lo que en estadística serían consideradas como variables dependientes. En realidad, estos indicadores serán los más "sencillos" de identificar, dada la concreción en las definiciones de los desafíos y la condición latente que hemos impuesto de que dichas problemáticas puedan reflejarse en datos. En segundo lugar, buscaremos indicadores que hipotéticamente puedan estar asociados de alguna forma a nuestros desafíos, esto es, indicadores cuya variación pensemos que pueda tener algún efecto sobre los desafíos, lo que en estadística se denominan variables independientes. Este segundo catálogo será el más complejo de establecer y el que estará más abierto a circunstancias locales. Quizá, y debido a esta variabilidad en el establecimiento de estos conjuntos de datos, una de las cuestiones más importantes que deberíamos abordar en primer lugar, sería determinar las características que deben presentar dichos indicadores, para facilitar su identificación por las diferentes organizaciones/administraciones y la posterior consecución de los objetivos que hemos establecido para nuestro estudio.

Obviamente, los datos han de tener suficiente "calidad" para considerar su incorporación a nuestro análisis. Este concepto de "calidad de los

datos", como característica esencial de los mismos, es multifactorial, contemplando, al menos, los siguientes aspectos:

- Datos **accesibles**, se encuentran disponibles y en un formato útil. Aquí podría ser de aplicación la clasificación en cinco estrellas de Tim Berners Lee[4], que en nuestro caso se traduciría en que los datos que buscamos dispongan de, al menos, dos estrellas (dato estructurado legible automáticamente por ordenador) siendo recomendable que dispusiese de tres estrellas (igual que antes, pero en formato no propietario). Volveremos sobre este punto en el apartado relativo a las fuentes de datos.
- Datos **precisos y exactos**, representando los valores verdaderos (actualizados) del indicador, sin sesgos, ni errores que puedan afectar a los resultados.
- Datos **coherentes**, que permitan combinarlos con otros datos relevantes de manera precisa y acorde a nuestros objetivos.
- Datos **completos**, que no haya datos perdidos, o faltantes, que nos obliguen a aplicar estrategias para resolver dicha problemática que puedan afectarnos en la aplicación del análisis o condicionarnos el mismo al uso de herramientas que permitan sortear esa situación, y que los datos sean suficientes para abordar el problema.
- Datos **consistentes**, que concuerden entre diferentes fuentes, o no induzcan a pensar la existencia de errores o sesgos.
- Datos claramente **definidos**, en los que sepamos qué significa cada campo, sin ninguna duda al respecto.
- Datos **relevantes** para nuestros desafíos, aunque aquí, en muchos casos, solo podremos tener indicios de esa relación, más o menos

[4] https://5stardata.info y https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

obvios, que tendremos que confirmar, o refutar, a partir de nuestros análisis. De hecho, ese podría ser el objetivo final del trabajo al completo, identificar variables cuya variación afecte a los desafíos que hemos formulado.

- Datos **fidedignos**, principalmente su confianza y credibilidad, que en gran medida se vinculará a su procedencia, siendo conveniente habilitar mecanismos de trazabilidad tanto para su origen como para su posterior tratamiento, especialmente importante en el caso de datos sintéticos.
- Datos **recientes**, que se encuentren disponibles en un corto intervalo de tiempo desde su adquisición, o que dicho tiempo sea razonable. Si los datos forman parte de una serie temporal, será importante que su resolución o periodicidad temporal se adapte al resto de conjuntos de datos utilizados y a nuestra metodología de análisis del desafío concreto que abordemos.
- Datos con **capilaridad, o desagregación,** geográfica suficiente para nuestro estudio, disponiendo en nuestro caso, al menos, de datos de ámbito provincial, pero probablemente fuese conveniente disponer de datos a nivel municipal. Igualmente deberemos considerar datos que se encuentren disponibles en todo el territorio, con objeto de poder analizar efectos y causas de sus variaciones y así determinar las posibles actuaciones a llevar a cabo.

4.1. INDICADORES PARA EL DESAFÍO #1: LA DESPOBLACIÓN

La identificación de indicadores para el desafío de la despoblación es aparentemente el caso más sencillo, al menos para los indicadores que hemos denominado dependientes. Es evidente que la población, o alguna variable derivada de la misma, será el indicador que deberemos

elegir. Pero claro, ¿cuál es el valor de la población, o su variación, que consideramos representativa de que esté sucediendo una despoblación en un territorio? Hasta ciudades de decenas de miles de habitantes sufren la pérdida de población, al igual que municipios de tan solo cien habitantes. Así pues, lo más lógico parece ser que debamos considerar no la "población" en sí misma, sino la "variación de la población", en términos absolutos, relativos, o en tasa [por unidad de tiempo], o incluso la densidad de población, igualmente pudiendo considerar en este caso su variación absoluta, relativa, o en tasa. También se podría utilizar para esta caracterización algún indicador más complejo, como por ejemplo algún índice de envejecimiento de la población, considerando para este supuesto alguna medida de tendencia central, media o mediana, calculada sobre la edad de la población del territorio considerado.

La selección de los indicadores que hemos denominado independientes es algo mucho más abierto. Así, se podrían considerar variables que se relacionarían directamente con la variación de la población, tales como, nacimientos, defunciones o migraciones, por ejemplo. Sin embargo, parece, a priori, poco recomendable la inclusión adicional de estas variables en el estudio: ya sabemos que estas variables están relacionadas con la variación de la población y, sin embargo, su inclusión quizá no nos aporte mayor capacidad de actuación para abordar el problema al que nos enfrentamos. O sí, si detectamos, por ejemplo, que la emigración constituye uno de los fenómenos que más afecta al descenso de población, y consecuentemente se puedan arbitrar medidas públicas para revertir ese proceso o mitigarlo. No obstante, no parece que estas tres variables correspondan con las que hemos denominado independientes, que son, recordemos, aquellas sobre las que se dispone de capacidad para actuar y que presumimos están relacionadas de forma teóricamente causal con el desafío formulado.

Este conjunto de datos sobre los que deberíamos tener capacidad de actuación será más complejo de determinar y muy probablemente,

además, de obtener. Ciertamente en una primera aproximación podríamos recurrir directamente a los indicadores que se formulasen con objeto de caracterizar nuestro segundo desafío, los servicios en las zonas rurales, dando por sentado que la pérdida de población es consecuencia (y probablemente también la causa) de la falta de servicios en esas mismas poblaciones. No obstante, el problema que esa conexión nos origina es que no tenemos, en principio, capacidad de influencia directa sobre esos servicios, que por lo general corresponderán al sector privado. Lo que buscamos en ese punto son variables sobre las que las diferentes administraciones puedan actuar, más o menos de forma directa, con objeto de revertir o, al menos, mitigar la tendencia negativa de la población. Podríamos pensar aquí en un ejemplo básico: una forma de paliar el descenso de la población bien podría ser potenciar la natalidad en el territorio, y una política que se puede llevar a cabo desde las administraciones es fomentar el nacimiento de bebes en el territorio a través de ayudas públicas, por ejemplo. Para poder introducir dicha variable en nuestro modelo, como ya se ha indicado, necesitaríamos disponer de los recursos dispuestos por las administraciones con ese objetivo en todo el territorio nacional, y así poder evaluar su efecto. Esta disponibilidad podría ser complicada de conseguir, pero este punto será objeto de posterior consideración. Algo similar podría aventurarse de otros posibles conjuntos datos, como, por ejemplo, los recursos públicos puestos en juego para facilitar el asentamiento de familias, a través de la disponibilidad de vivienda pública, de ayudas al alquiler, o de subvenciones para fomentar la contratación de personas en empresas, entre otras estrategias posibles para intentar fijar población en el territorio.

Así pues, en este primer desafío hemos identificado tres posibles tipos de datos para su inclusión en nuestro estudio: datos asociados directamente al desafío, variables asociadas directamente al problema sobre las que podríamos actuar mediante políticas públicas, y otras variables que se encuentran embebidas precisamente en el segundo desafío propuesto,

la disponibilidad de servicios, cuya caracterización y posibles actuaciones estudiaremos a continuación.

4.2. INDICADORES PARA EL DESAFÍO #2: PERDIDA DE SERVICIOS EN LAS ZONAS RURALES

En este segundo desafío, los indicadores dependientes estarán más abiertos que en el caso anterior. Podremos considerar, por ejemplo, la disponibilidad en el territorio de servicios relacionados con la salud (ambulatorios/CAP, hospitales privados, farmacias), con la educación (bibliotecas, escuelas infantiles y guarderías, colegios, institutos, universidades), con el transporte (trenes, autobuses, aeropuertos/puertos, aparcamientos, carreteras/autovías), con el comercio (supermercados/hipermercados, comercios al por menor), con la industria (empresas), con la banca (oficinas bancarias, cajeros automáticos), con el turismo (hoteles, campings, casas rurales), con el ocio (bares, restaurantes, pistas deportivas, piscinas), con la cultura (museos, cines, teatros, BICs, fiestas interés cultural, festivales, aulas/centros culturales, yacimientos arqueológicos) o con las telecomunicaciones (telefonía móvil, banda ancha fija/móvil), entre otras muchas posibilidades. En algunos casos puede ser conveniente disponer de métricas relacionadas con la distancia, o el tiempo necesario, para llegar al punto de prestación del servicio correspondiente que se encuentre más cercano, en el caso de que no exista en la localidad considerada.

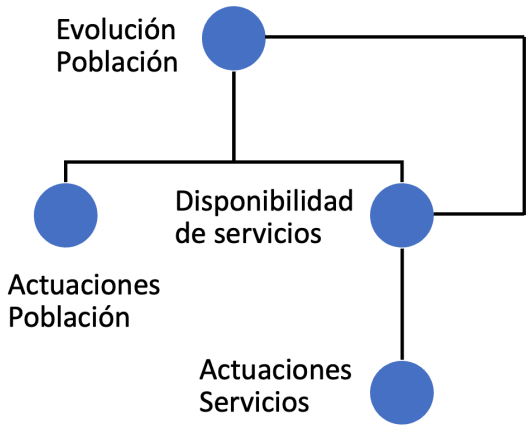

Figura: Posibles relaciones entre desafíos e indicadores

Es igualmente obvio, como ya hemos propuesto, que es perfectamente posible, y quizá hasta recomendable, que estos indicadores dependientes puedan formar parte de los indicadores independientes correspondientes al primer desafío, en una suerte de segundo nivel de posibles puntos de actuación política. Recordemos, de nuevo, que ya se hipotetizó anteriormente que ambos desafíos podrían estar interrelacionados entre sí. De hecho, sería conveniente considerar, sin prejuzgar los resultados que se puedan obtener posteriormente en cuanto a su interdependencia, que la población es una de las variables independientes de este segundo desafío. Como ejemplo simple de dicha vinculación, solo recordar que la presencia de colegios, un servicio prestado mayoritariamente por el sector público, está directamente relacionado con la existencia de un determinado número potencial de alumnos que puedan asistir al centro educativo considerado.

Variables dependientes	Variables independientes
Despoblación	
Variación de la **población** Variación de la **densidad de población** [Nacimientos, mortalidad, migraciones]	Actuaciones sobre fomento de la **natalidad** Actuaciones sobre disponibilidad de **vivienda** Actuaciones sobre **alquiler** de vivienda Actuaciones sobre fomento de la **contratación** … ¿Disponibilidad de **servicios** en el territorio?
Disponibilidad de servicios	
Salud (ambulatorios/CAP, hospitales, farmacias) **Educación** (bibliotecas, escuelas infantiles y guarderías, colegios, institutos, universidades) **Transporte** (trenes, autobuses, puertos y aeropuertos, aparcamientos, carreteras/autovías) **Comercio** (supermercados, comercios…) **Industria** (empresas) **Banca** (oficinas y cajeros automáticos) **Turismo** (hoteles, campings, casas rurales) **Ocio** (bares, restaurantes, pistas deportivas, piscinas) **Cultura** (museos, cines, teatros, BICs, fiestas, festivales, centros culturales, yacimientos arqueológicos)	Actuaciones sobre servicios de **salud** Actuaciones sobre servicios **educativos** Actuaciones sobre medios de **transporte** Actuaciones para el fomento del **comercio** Actuaciones para el asentamiento de **industrias** Actuaciones para habilitar **servicios bancarios** Actuaciones sobre promoción del **turismo** Actuaciones para la mejora de actividades de **ocio** Actuaciones en promoción de aspectos **culturales** Actuaciones para la mejora de la cobertura de servicios de **telecomunicación** …

Variables dependientes	Variables independientes
Telecomunicaciones (telefonía y banda ancha, servicios fijos y móviles) [**Disponibilidad** o **distancia** al punto de prestación]	¿Número de **habitantes** en el territorio? ¿Medida de tendencia central de la **edad**?

Tabla resumen de posibles conjuntos de datos sobre despoblación y disponibilidad de servicios

Por otro lado, y al igual que nos ocurría en el caso anterior, las variables independientes que deberíamos identificar en esta fase deberán estar relacionadas con las posibles actuaciones que se pudieran promover políticamente con objeto de incrementar la disponibilidad en el territorio de los servicios objeto de análisis, o, alternativamente, facilitar su accesibilidad por los ciudadanos, acercándolos a través de la habilitación de otros servicios, como, por ejemplo, el transporte público o la mejora de las carreteras existentes.

5. HITO #3: POSIBLES FUENTES DE DATOS

Una vez identificados, en el punto anterior, los posibles conjuntos de datos, o categorías de datos, que pudiesen ser útiles para los desafíos que se han propuesto, corresponde la localización de estos entre las posibles fuentes de datos que puedan estar disponibles.

La primera cuestión por dilucidar es sí esas fuentes pueden ser internas, de la propia administración, o deben ser externas. Lógicamente la respuesta no será única, y dependerá del conjunto de datos considerado, pero también de la extensión deseada para el mismo. Así, por ejemplo, es posible, y lógico, que conjuntos de datos relativos a actuaciones de la propia administración, en los campos enumerados como variables independientes en el punto anterior, ya obren en poder de la administración actuante, principalmente si se trata de su ámbito territorial, pero, si se desea una mayor extensión, habrá que recurrir bien a otras administraciones, para que nos los cedan o con las que intercambiemos dichos datos, o bien a conjuntos de datos abiertos, ya publicados, que los contemplen, que serán estos últimos probablemente la opción mayoritaria. Para averiguar de qué conjuntos de datos podemos disponer es posible acceder a catálogos publicados, entre los cuales destaca sobremanera el Catálogo Nacional de Datos Abiertos[5], en el que, en realidad, no hay datos, sino puntos de acceso a los datos publicados por otras entidades. También se podría considerar, llegado el caso, obtener la información que se precisase de fuentes de datos de

[5] https://datos.gob.es

pago, tipo "Statista"[6], o del tipo de la Agencia de Datos de Europa Press[7] o Expansión[8], por ejemplo. No obstante, hemos de tener en mente que es posible llevar a cabo comparativas deseadas sobre el efecto de las actuaciones ejecutadas para intentar resolver nuestros desafíos, considerando solo nuestro ámbito territorial de actuación, por lo tanto, utilizando solo datos internos. Aunque disponer de datos de, por ejemplo, todo el territorio nacional puede aportarnos más información sobre qué efectos tienen unas y otras actuaciones; pero para ello deberemos obtener unos conjuntos de datos coherentes entre todas las administraciones cesionarias de los mismos, con objeto de evitar datos faltantes en nuestros análisis. En este sentido, ahondando en la normalización de los conjuntos de datos publicados por las diferentes administraciones, es necesario traer a colación el relevante documento publicado por la FEMP[9] con los 40 conjuntos de datos que las entidades locales debieran publicar en abierto. Muchos de estos conjuntos de datos podrían ser aplicables en nuestros desafíos, como, por ejemplo, aparcamientos públicos; censo de locales, actividades y terrazas; instalaciones deportivas; lugares de interés turístico; transporte público; entre otros, cuya disponibilidad echaremos en falta para poderlos utilizar en nuestro estudio.

[6] https://es.statista.com Aunque está más enfocado a estudios de mercado que a datos en general.

[7] https://www.epdata.es , que dispone de datos agregados, por ejemplo, del gasto en servicios públicos esenciales por comunidad autónoma (desagregado a ese nivel autonómico)

[8] https://datosmacro.expansion.com, que dispone, por ejemplo, de datos de renta y deuda desagregados a nivel municipio, con datos provenientes del INE y del Ministerio de Hacienda y Administraciones Públicas ya procesados.

[9] "DATOS ABIERTOS FEMP 2019. 40 conjuntos de datos a publicar por las Entidades Locales", FEMP, 2018

Otro problema que aparece en este punto sobre la información interna es si efectivamente disponemos de dicha información sobre actuaciones llevadas a cabo por nuestra administración debidamente territorializada, o geolocalizada, con el nivel de capilaridad, o desagregación, que deseemos utilizar en nuestro estudio. Esta nueva explotación de los datos seguramente no siempre habrá sido contemplada por nuestra administración en el pasado y por lo tanto ese dato de localización no habrá sido debidamente informado en nuestros procesos, pero ahora deberíamos considerar que fuese incorporado.

Una fuente de datos interna, que quizá también podría aportarnos información sobre los desafíos que hemos planteado, proviene de la sensorización del territorio, esto es, de las *Smart cities* o "ciudades inteligentes"; si bien debemos considerar el término "ciudad" en nuestro caso en un sentido más amplio, como "territorio", y hablar en general de "territorio inteligente", tal y como consta en el nombre del Grupo de Trabajo. Existen a lo largo y ancho de España múltiples proyectos sobre "territorios inteligentes", aplicados tanto a ciudades, en su sentido más literal, como al medio rural. Ciertamente el volumen de datos que generan este tipo de proyectos puede ser notable, por lo que será muy recomendable analizar las ventajas e inconvenientes de su posible participación en nuestros desafíos, tanto como datos "locales", internos de la administración considerada, como aportados por otras entidades. Si bien en la lista de indicadores sugeridos en el apartado anterior, ninguno de ellos se corresponde con una fuente de datos proveniente de la sensorización del territorio, no existe ningún problema para su inclusión, si consideramos interesante para nuestros análisis la presencia de algún tipo de datos procedente de esta fuente.

Como ya hemos indicado, solo en algunos casos dispondremos de una fuente única para incorporar datos de todo el territorio bajo estudio. En los casos restantes, los datos provendrán de fuentes múltiples, diversas o distribuidas, que deberemos consolidar convenientemente. Como

ejemplo de esta situación podemos pensar en todos aquellos conjuntos de datos relativos a competencias de las diferentes comunidades autónomas y que se publican individualmente por cada una de ellas (sanidad y educación, por ejemplo y entre otras). Si queremos una visión global de toda España, deberemos procesar todos esos conjuntos de datos de diferente procedencia, para unificarlos en un solo juego de datos. Y eso no siempre será sencillo, como bien sabemos, atendiendo, por ejemplo, a las diferentes situaciones acaecidas durante la pandemia de Covid-19, durante la cual, las comunidades autónomas facilitaban datos en tiempos diferentes y con definiciones de los mismos igualmente distintas, obligando a un trabajo previo de normalización, que no siempre obtenía resultados satisfactorios en el proceso de consolidación. Estas diferencias en los datos aportados hacían muy complicado, o incluso imposible, modelar correctamente la evolución de la pandemia en aquellos momentos.

No obstante, esta labor de preprocesado de los datos que posteriormente intervendrán en nuestros análisis será, por lo general, un paso previo obligado. De hecho, esta fase dispone de "personalidad" propia y ha dado lugar a herramientas específicas para facilitar su realización[10]. En efecto, la "Extracción" de datos de su fuente original, su posterior "Transformación" para posibilitar su tratamiento, y la "Carga" final mapeando las variables en los correspondientes campos de la base de datos o de la herramienta de analítica utilizada, se denomina por su acrónimo en inglés, ETL (*Extraction-Transform-Load*). Un ejemplo sencillo de la necesidad de este paso previo podemos observarlo si consideramos la incorporación a nuestros análisis de fuentes de datos no estructuradas, como pueden ser alguno de los múltiples estudios realizados, por ejemplo, por el Banco de España[11], y publicados en formato PDF, conteniendo

[10] En http://dfkoz.com/ai-data-landscape se puede encontrar un listado de herramientas para ETL.

[11] Por ejemplo, https://repositorio.bde.es/bitstream/123456789/17531/1/do2122.pdf

tablas que deben ser extraídas del documento, procesadas, para disponer de una tabla numérica al uso y finalmente incorporados estos datos en nuestro análisis. Otra alternativa en este caso de datos no estructurados es acceder, si se citan en el documento, directamente a las fuentes originales de los datos[12]. Pero como decíamos, este paso previo puede ser también necesario para obtener un conjunto de datos consolidado proveniente de fuentes múltiples, aunque se encuentren todas ellas en un único formato ya legible automáticamente por ordenador. Otro posible caso de uso en el que las herramientas de ETL justifican su existencia es la necesidad de tratar diferentes tipos de datos, nominales/ordinales, continuos/discretos, directos/sintéticos, que precisan una normalización o tratamiento previo a su incorporación en los análisis de datos, o incluso simplemente porque los tipos de ficheros de origen de los datos no son coincidentes entre sí, o con el exigido en nuestra herramienta de análisis.

Es necesario determinar los formatos específicos en los que estarán disponibles los datos en función de la naturaleza de dicha información. Siempre que sea posible y no implique un esfuerzo desproporcionado, se facilitarán los documentos en formatos abiertos y legibles por máquina, conjuntamente con sus metadatos, y en el mayor grado posible de precisión y desagregación. Tanto el formato como los metadatos seguirán normas formales y estándares abiertos[13].

[12] En el caso del documento citado del BdE, los datos provienen de Kompil, M., C. Jacobs-Crisioni, L. Dijkstra y C. Lavalle (2019). «Mapping accessibility to generic services in Europe: A market-potential based approach», Sustainable Cities and Society, vol. 47, 101372.

[13] *Cf.* Artículo 5.3 Ley 37/2007

Se han establecido los siguientes niveles de reutilización, en gran medida en paralelo con la clasificación en 5 estrellas de Tim Berners Lee anteriormente citada.

- **Nivel 1**: Se publican los datos en la web en cualquier formato y bajo una licencia abierta. Se pueden ver, imprimir, guardar localmente, cambiar y compartir. Para la administración es simple publicarlo, sin embargo, los datos están atrapados en un documento y es difícil trabajar con ellos. Por ejemplo, un archivo en formato PDF.
- **Nivel 2**: Se publican como datos estructurados, por ejemplo, un archivo Microsoft Excel. Se puede hacer todo lo que se permite en el Nivel 1 más procesar los datos directamente con software propietario y exportarlos a otro formato estructurado. Es un formato ampliamente utilizado para el intercambio de datos debido a que ofrece buenas oportunidades de mantener la estructura de los datos y la forma en que los archivos son construidos, y permite a los desarrolladores escribir partes de la documentación con los datos sin interferir en la lectura de ello.
- **Nivel 3**: Se usa formato no propietario, por ejemplo, un CSV[14] o JSON[15] o XML[16]. Se puede hacer todo lo que se permite en el Nivel 2 y manipular los datos de cualquier forma, sin limitación de características o de uso de algún tipo de software en particular.

[14] CSV Comma Separated Values. Archivos de Valores Separados por Comas. Los archivos CSV pueden ser un formato útil, debido a que son compactos y por lo tanto adecuados para transferir grandes conjuntos de datos con la misma estructura

[15] JSON: es un formato de archivo muy fácil de leer por cualquier lenguaje de programación. Su simplicidad significa que generalmente es fácil de procesar para las computadoras a comparación de otros, como lo es XML

[16] XML: eXtensible Markup Language. Es un lenguaje de etiquetas parecido en al HTML de la páginas web.

- **Nivel 4**: Se utilizan URIs[17] para identificar y localizar datos. Una forma de representar los datos es utilizar RDFs[18]. Se puede hacer todo lo que se permite en el Nivel 3 y enlazarlos desde cualquier otro sitio, marcarlo como favorito, reutilizar parte de los datos y reutilizar herramientas. Recomendado por W3C[19], permite presentar datos en una forma que facilita la combinación de información de diferentes fuentes. Los datos RDF pueden ser almacenados en XML y JSON, entre otras serializaciones. RDF promueve el uso de URLs[20] como identificadores, lo que brinda una manera conveniente de interconectar iniciativas open data existentes.

- **Nivel 5**: Se enlazan con otros datos externos para proveer contexto (*Linked Open Data*). Se puede hacer todo lo que se permite en el Nivel 4 y además se pueden descubrir más datos relacionados, conocer directamente el esquema de datos e incrementar el valor de los datos.

Los procesos de ETL, tal y como se han expuesto anteriormente, han experimentado últimamente cambios en la forma de realizarse, dando lugar a nuevos procesos más concretos. Así, por ejemplo, aparecen procesos ELT, en el que se cargan los datos sin prestar inicialmente demasiada atención a su formato de origen; o procesos ETLT, en los que se realiza un preprocesado previo ligero, como por ejemplo sucedía en el caso antes descrito de documentos en formato PDF para extraer sus tablas de datos, que luego serían transformadas de nuevo de manera más específica como paso previo a su tratamiento. Estas variantes de los

---

[17] URI: *Uniform Resource Identifier*. Identifica un recurso de manera única. Una URL (dirección de Internet) es un ejemplo de URI, pero los hay con otros formatos e informaciones incluida en ellos.

[18] RDF: Resource Description Framework. Método para la descripción o modelado de datos.

[19] W3C: World Wide Web Consortium. Consorcio constituido para desarrollar las normas que rigen la WWW.

[20] URL: Uniform Resource Locator. Es la denominación técnica de una dirección única de un recurso en la red Internet.

procesos de ETL permiten, en algunos casos, una carga de datos en bruto más rápida en los sistemas de información, así como transformar posteriormente solo los datos que realmente vamos a necesitar y no emplear recursos de computación en procesar datos inútiles. Ciertamente hay que tener algunas precauciones en este último caso, en los que almacenamos datos que no necesitamos para nuestros tratamientos, si operamos con datos personales, pero esto no sucederá en los tratamientos objeto de nuestro estudio, en los que las operaciones se ejecutan sobre datos anónimos, o debidamente anonimizados, sobre los que no aplican las regulaciones de protección de datos personales. Hay disponibles múltiples documentos sobre métodos para anonimizar información que inicialmente contiene datos de carácter personal (y por lo tranto protegidos), destacando, entre otros, de la Agencia Española de Protección de Datos[21,22] (AEPD), del Comité Europeo de Protección de Datos [23] (CEPD), de la Agencia de la Unión Europea para la Ciberseguridad (ENISA) [24] y del portal nacional de datos abiertos datos.gob.es[25], por ejemplo, todos ellos disponibles *on-line*.

No obstante, la obtención de los posibles juegos de datos que dieran soporte a los indicadores seleccionados en el hito anterior presenta diversos problemas, que podrían ser probablemente generales y no circunscritos a los indicadores elegidos. En primer lugar, la dificultad que podríamos considerar obvia: la disponibilidad de los propios datos. Esta

[21] "Orientaciones y garantías en los procedimientos de anonimización de datos personales", AEPD, 2016

[22] "Guía básica de anonimización. Elaborada por la Autoridad Nacional de Protección de Datos de Singapur (PDPC)", y herramienta de anonimización asociada. AEPD, 2022

[23] "Dictamen 05/2014 sobre técnicas de anonimización", CEPD, 2014

[24] "*Data pseudonymisation: advanced techniques & use cases*", ENSISA, jan-2021

[25] "Introducción a la anonimización de datos. Técnicas y casos prácticos", Datos.gob.es y otros, nov-2022

disponibilidad puede venir de la ya comentada falta de consolidación de los mismos, algo que sucede, por ejemplo, con los datos de "Centros Escolares", para los que existe una solicitud pendiente de resolver en el portal "datos.gob.es" del 25 de marzo de 2019, con más de 50 adhesiones a la misma, y sin culminar [consultado el 14 de marzo de 2023]. Pero también puede estar causada por la falta de los juegos de datos propiamente dicha, como sucede, por ejemplo, con la existencia de una base de datos de comercios (aunque es cierto que en este caso sí que quizá pudiera obtenerse dicha información, o similar, a través de plataformas de datos de pago), o una base de datos de estaciones de bus/tren, de la que (aparentemente) el Ministerio de Transportes, Movilidad y Agenda Urbana no dispone, al menos públicamente, como tampoco se ha localizado una base de datos de Museos o de Oficinas de Correos, por ejemplo. Esta ausencia del conjunto de datos deseado puede obligarnos a variar ligeramente el indicador seleccionado[26] para disponer de dicha información.

También relacionada con la disponibilidad, pero esta vez a nivel "técnico", alguna de las fuentes de datos existente no ha estado disponible para su consulta durante algunos períodos de tiempo (en este caso ocurrió con el Catastro). Otra de las dificultades que hemos encontrado son limitaciones a la descarga de información, teóricamente disponible, esto nos ha sucedido con la Base de Datos Nacional de Subvenciones (BDNS), que hemos pretendido utilizar para obtener información relativa a actuaciones de las diferentes administraciones, y que "solo" permite la descarga simultánea de 10.000 registros de los más de 300.000 de que dispone. En este caso es probable que pueda ser viable la solicitud directa de la base de datos completa a su tenedor.

[26] Por ejemplo el Catastro dispone de una base de datos de edificaciones por tipo de uso que puede resolver la ausencia de alguno de los juegos de datos no localizados.

Aun considerando todo lo expuesto anteriormente, se han identificado 14 fuentes de datos públicas que permiten recopilar la práctica totalidad de los indicadores seleccionados en el hito precedente, o alternativos, y cuyo detalle puede consultarse en el Anexo I, salvo las ausencias ya anunciadas.

Confiamos que, a partir de la convocatoria de subvenciones publicada en el BOE el 1 de septiembre pasado[27], destinadas a la transformación digital y modernización de las administraciones de las entidades locales, en el marco del Plan de Recuperación, Transformación y Resiliencia, y en concreto de su "Línea estratégica 3. Gobierno del dato" que con el objetivo de "democratizar el acceso a los datos por parte de ciudadanos, empresas y empleados y empleadas públicos, permitiendo el libre acceso a la información y su replicabilidad, así como el uso de los datos para el desarrollo de servicios digitales de alto valor añadido orientados al ciudadano" y con posibles actuaciones subvencionables de creación de una "plataforma de datos integrada" que define como el "Desarrollo de una plataforma del dato integrable, bajo la perspectiva de la federación de datos, con el Data lake de la Administración General del Estado, que permita habilitar proyectos interoperables entre la Administración General del Estado y otras administraciones públicas" y la consiguiente disponibilidad de fondos, permita avanzar en una mayor disponibilidad de más conjuntos de datos reutilizables.

[27] (14391) Orden TER/836/2022, de 29 de agosto, por la que se aprueban las bases reguladoras de subvenciones destinadas a la transformación digital y modernización de las administraciones de las entidades locales, en el marco del Plan de Recuperación, Transformación y Resiliencia, y se efectúa la convocatoria correspondiente a 2022. BOE nº 201 de 1 de septiembre de 2022.

6. HITO #4: POSIBLES HERRAMIENTAS DE ANÁLISIS, PROCEDIMIENTOS, ALGORITMOS O MODELOS

Tras la culminación del Hito #3, dispondremos de los datos obtenidos para nuestros desafíos en forma de una matriz, o tensor, previsiblemente tridimensional, con dimensiones "territorio" (municipios, por ejemplo), "variables" (cada indicador considerado) y "tiempo/condiciones" (anualidad o condiciones existentes, por ejemplo). Esta tabla de tres entradas, o de tres vías, como se denomina en estadística, será el elemento sobre el que pretenderemos aplicar diferentes herramientas para obtener información que permita "resolver" nuestros desafíos.

Antes de poder someter nuestro "cubo de datos" a análisis estadísticos concretos, hemos de resolver algunas problemáticas que, sin duda, puede presentar y que, en su caso, dificultarían la aplicación de muchos de los métodos estadísticos que podrían ser elegibles.

6.1. TRATAMIENTO PARA DATOS FALTANTES

En este primer punto vamos a tratar la problemática que se nos plantea con los "datos faltantes", o no existentes. Esta situación puede venir derivada de, principalmente, dos contextos diferentes. El primero sería el más obvio, la no disponibilidad del indicador en cuestión, o variable, para un territorio, o para varios territorios. Las alternativas disponibles de actuación en este caso, y salvo mejor opinión, son limitadas.

Como primera opción, podríamos considerar el dejar la matriz de datos tal cual, con datos faltantes, limitándonos posteriormente a aplicar métodos u algoritmos de análisis capaces de operar con datos faltantes.

Esto es posible porque existen métodos no solo aplicables a matrices con datos faltantes, sino incluso a matrices escasamente pobladas, denominadas matrices *sparse*. No obstante, esta situación de datos faltantes nos limitará en gran medida el catálogo de los métodos que podremos utilizar posteriormente.

La segunda alternativa, y de la que tendremos que valorar su aplicación, sería eliminar de nuestro estudio el territorio del que no tengamos un valor de algún indicador. Esta segunda alternativa tiene como otra posibilidad, más drástica aún que la eliminación del territorio, la eliminación de la variable que presenta dichos datos faltantes, o bien buscar una alternativa al indicador que nos introduzca una información similar en el análisis y del que sí dispongamos de valores para todos los territorios considerados.

El segundo contexto posible, en el que también nos tendremos que enfrentar a datos faltantes, provendría de la existencia de diferentes períodos de recogida de los datos. Así, por ejemplo, si un indicador ofrece valores obtenidos trimestralmente y otro dispone de datos con una periodicidad anual, estaremos intentando comparar series numéricas con desiguales longitudes y datos faltantes en determinados puntos de muestreo. En este caso, ciertamente, las posibilidades de tratamientos que tenemos son más interesantes.

En primer lugar, el caso trivial, consistente en proceder como se ha propuesto anteriormente, y eliminar el "exceso" de datos en la serie con mayor frecuencia de captura, dejando todas las series numéricas con igual longitud e idénticos puntos temporales de muestreo, a costa, eso sí, de una pérdida considerable de información en aquellos indicadores que disponían de más valores que los realmente considerados tras este submuestreo.

La segunda opción, más atractiva, consistiría en interpolar la serie de la variable con datos faltantes para completarla hasta igualar la frecuencia de la serie con menor periodo de captura. De manera evidente esto seguramente introduzca un sesgo en nuestro análisis, pero es una forma de no perder la información contenida en las series con mayor número de datos. El método de interpolación puede ser seleccionado entre los muchos disponibles, desde una interpolación polinómica, como, por ejemplo, una simple interpolación lineal, considerando los elementos situados entre los dos puntos disponibles en los extremos del intervalo a completar, o con polinomios de potencias más elevadas; o utilizar otros métodos más sofisticados, utilizando interpolaciones más complejas, por ejemplo, trigonométricas, con desarrollos de Fourier. En estos dos últimos casos se considerarían más puntos de la serie a completar, que simplemente los extremos del intervalo que acogen los datos faltantes y que se utilizarían en una simple interpolación lineal.

De cualquier modo, tampoco se ha de tener una especial "obsesión" en nuestro caso con disponer de datos con una capilaridad temporal muy elevada. Efectivamente, y dado que los desafíos que estamos afrontando presentan variaciones a largo plazo, no será necesario disponer de una elevada resolución temporal en los datos. Por el contrario, a nivel geográfico sí que puede ser interesante disponer de una capilaridad a nivel municipal, ya que, si bien a nivel provincial también se presumen variaciones en los indicadores, se entiende que la sensibilidad será mayor a nivel municipal.

6.2. NORMALIZACIÓN DE LOS DATOS: CENTRADO Y ESCALADO

Una vez que disponemos de una matriz de indicadores completa, procederemos a abordar la siguiente problemática a la que tendremos que enfrentarnos con una alta probabilidad: que los indicadores participantes en el estudio presenten rangos muy diferentes y difícilmente comparables. Por ejemplo, la población del territorio considerado seguramente se moverá en órdenes de magnitud de centenares o de miles de habitantes; sin embargo, la dotación de los diferentes servicios educativos muy probablemente oscilará en un rango que no superará la decena, o quizá simplemente la unidad (presencia/ausencia). Esta diferencia notable, en los rangos que presentan las distintas variables, puede ocasionar por sí sola perturbaciones en los resultados de los análisis que realicemos, dificultando la obtención de resultados válidos. Para evitar este efecto pernicioso, la solución consiste en normalizar la matriz de datos de forma que se conserven los distintos perfiles de las variables, o indicadores, mientras los rangos se tornan comparables entre sí.

La normalización de un conjunto de datos de tres vías es más complicada que en el caso de una matriz bidimensional. Por ejemplo, en un conjunto de datos de tres vías, el proceso de normalización se puede llevar a cabo por cualquiera de las tres vías posibles, por combinaciones de dos de ellas, o por las tres simultáneamente. No obstante, en la literatura se alerta de que, en este último caso, optar por normalizar el conjunto de datos por sus tres vías simultáneamente, se pueden perder características estructurales de los datos y por ello no es recomendable.

No obstante, el primer paso será seleccionar el método de normalización a aplicar, entre todos los posibles. Consideramos que todos nuestros datos son numéricos, esto es no son ordinales, ni nominales y es indiferente que sean números reales o enteros. Cualitativamente el resultado más evidente de la normalización es la transformación de una variable con unidades en otra sin unidades, esto es, la población, por ejemplo, una vez normalizada ya no corresponde, al menos directamente, con los habitantes en el territorio considerado. Hay múltiples métodos de

normalización disponibles en la literatura, principalmente en función del resultado que deseemos obtener de la aplicación de este procedimiento.

Estos métodos de normalización se pueden descomponer en dos procedimientos diferentes, el centrado de los datos y su escalado que, además, como ya hemos indicado, se pueden aplicar por los diferentes modos que presentan nuestro conjunto de datos. Para exponer esto más claramente veamos un ejemplo. Consideremos centrar nuestros datos en uno de los modos, sea en este caso el correspondiente a las variables o indicadores. Procedemos a desplegar nuestra matriz tridimensional en una bidimensional yuxtaponiendo las matrices indicador vs. tiempo de los diferentes territorios, como se muestra a continuación. Una vez desplegada la matriz, restaremos a los diferentes valores de indicadores el valor medio de cada fila, correspondiente a un indicador a lo largo de todo el tiempo para todos los territorios. De esta forma, la media de cada indicador, a lo largo del tiempo, será nula con lo que dicho indicador habrá sido centrado. El centrado en el conjunto de la matriz está desaconsejado por comprometer, como se ha indicado, la estructura subyacente en los datos.

Territorio-1	Tiempo-1	…	Tiempo-T	…	**Territorio-K**	Tiempo-1	…	Tiempo-T
Indicador-1		…		…	Indicador-1		…	
Indicador-2		…		…	Indicador-2		…	
…		…		…	…		…	
Indicador-J		…		…	Indicador-J		…	

Esquema de centrado de los datos a través del modo, o dimensión, "indicadores"

Procederíamos de igual manera, yuxtaponiendo las matrices bidimensionales correspondientes, si, por ejemplo, deseásemos escalar los valores de los indicadores para cada punto temporal de muestro, por ejemplo, en este caso aplicaríamos la fórmula correspondiente a las columnas de la matriz bidimensional anteriormente obtenida, si es que dicho escalado tuviese sentido en nuestro análisis. El escalado conjunto de la matriz o de múltiples modos se torna más complejo, ya que, una vez realizado en uno de los modos, afecta a los modos restantes, por lo que debe ser realizado iterativamente y la progresión no siempre será convergente.

Aunque este procedimiento de normalización pueda parecer complicado, en realidad en la práctica no lo será tanto, ya que suele realizarse de una manera más intuitiva: será habitual, por ejemplo, normalizar, aplicando un centrado y escalado de los datos, a través de cada uno de los indicadores y a lo largo del tiempo. También será común, y una opción interesante, la normalización para cada una de las series temporales correspondiente a cada territorio y para cada indicador.

Tras esta fase de preprocesado, disponemos de nuestros datos organizados en una tabla de tres vías, completa y con rangos comparables, eso es, lista para poder aplicar sobre ella múltiples técnicas estadísticas que nos permitan extraer información de interés para el abordaje de nuestros desafíos, que es nuestro objetivo.

6.3. APLICACIÓN DE MÉTODOS GRÁFICOS

En general, para la extracción de información de los indicadores tradicionalmente se utilizan cuadros de mando, con representaciones

más o menos imaginativas, que permiten visualizar los datos y dejan al ojo humano el trabajo de extracción de información. Y esta no es una mala idea ya que la visión humana es una excelente herramienta para el análisis exploratorio de datos, estando muy cualificada para la detección de patrones. No obstante, conforme aumenta el número de datos en la representación, esta se oscurece, dificultando la visión y se deben utilizar otras herramientas que, en algunos casos, sintetizan las representaciones gráficas en diferentes indicadores numéricos. Pero los indicadores numéricos utilizados de forma aislada no siempre son buenos aliados en el análisis exploratorio de datos, siendo el ejemplo habitual de estos "malos amigos" el denominado "Cuarteto de Anscombe"[28]. En ese ejemplo se constata como cuatro conjuntos de datos, con estadísticas descriptivas prácticamente iguales, presentan distribuciones gráficas totalmente distintas, poniendo de manifiesto la importancia de las representaciones gráficas en el análisis exploratorio de datos. Como prueba de la importancia de las representaciones gráficas en el análisis de datos cabe destacar la existencia de una rama de estudio propia, denominada "analítica visual" o "*visual analytics*"[29,30].

En todo caso, y considerando los posibles problemas que la utilización de estos métodos visuales pueda conllevar, hemos de limpiar, de alguna manera, la representación que disponga de un número de datos tan elevado, que dificulte su interpretación. Una técnica habitualmente aplicada para este fin es la reducción de la dimensionalidad de los datos. Estas técnicas de reducción de la dimensionalidad, aplicadas en nuestro

---

[28] https://es.wikipedia.org/wiki/Cuarteto_de_Anscombe y F. J. Anscombe, «Graphs in Statistical Analysis», The American Statistician, vol. 27, n.o 1, p. 17- 21, 1973.

[29] J. J. Thomas y K. A. Cook, *Illuminating the path: the research and development agenda for visual analytics*, 1st ed. Los Alamitos, CA: IEEE, 2005, ISBN 0-7695-2323-4

[30] E. R. Tufte, The visual display of quantitative information, 2.a ed. Cheshire, Conn.: Graphics Press, 2001.

estudio, deben tener en cuenta que nos encontramos en presencia de conjuntos de datos de tres vías.

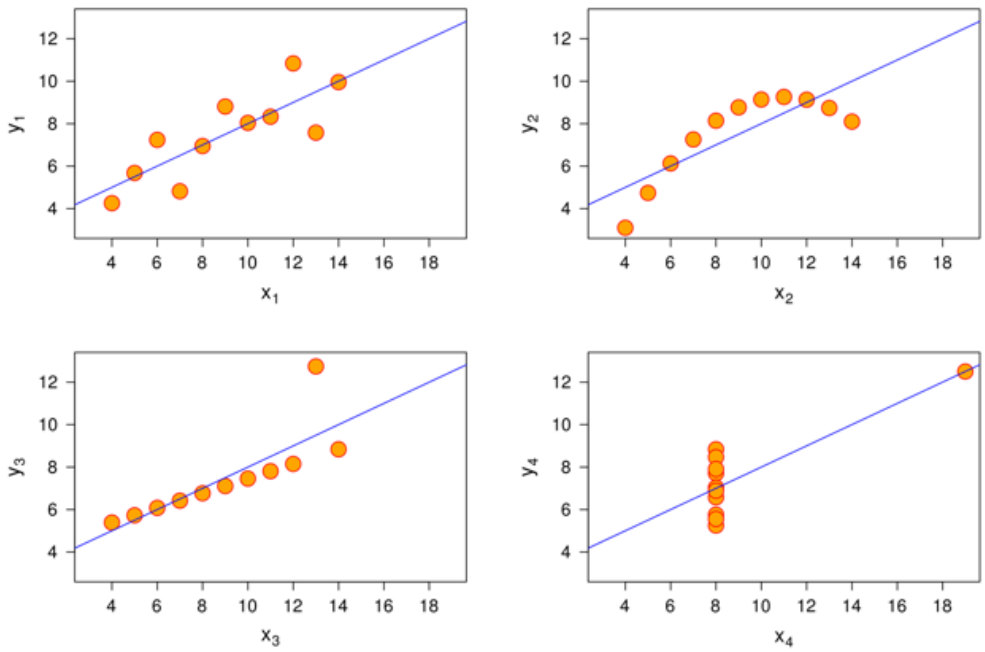

Figura: "Cuarteto de Anscombe" [31]

6.4. APLICACIÓN DE MÉTODOS ANALÍTICOS

La primera cuestión que tenemos que resolver en este punto es establecer cuál será el objetivo de nuestro análisis, algo que en los métodos gráficos se dejaba un poco a la interpretación del observador a la vista de las representaciones gráficas obtenidas. Este objetivo bien podría ser el detectar la interacción entre las diferentes variables consideradas, con objeto de delimitar cuáles de todas ellas afectan a los

[31] Fuente: GPL, https://commons.wikimedia.org/w/index.php?curid=863306

indicadores de referencia de nuestros dos desafíos. Si dicha interacción se verifica, desearíamos obtener un indicador que refleje cuáles de las variables independientes, sobre las que se dispone de capacidad de actuación, son más idóneas para influir en el comportamiento de los indicadores de referencia de nuestros dos desafíos. Como objetivos latentes de nuestro estudio se pueden considerar, por ejemplo, la determinación de grupos de territorios y/o de variables que presenten comportamientos similares.

Un aspecto que debemos tener en cuenta es que, a pesar de que nuestro conjunto de datos es de tres vías, es posible utilizar para su análisis métodos de dos vías. Así es posible realizar un desplegamiento del tensor de tres vías, o matrización, que no es más que la yuxtaposición de las matrices de dos vías que lo componen, como vimos anteriormente, y aplicar métodos destinados específicamente a operar con datos de dos vías. En efecto, la matriz tridimensional que responde a nuestros datos puede ser "cortada" a través de cualquier de sus tres ejes, formando una colección de matrices bidimensionales que puede ser analizada mediante métodos de dos vías. Por ejemplo, partimos de nuestra matriz de tres vías, territorios, indicadores, tiempo (año o condiciones aplicadas, por ejemplo). Podemos obtener una matriz bidimensional para cada período temporal o condición aplicada, con los territorios por filas, y los indicadores por columnas, que podemos someter al análisis de dos vías que deseemos. También podemos disponer de una matriz bidimensional para cada territorio, con la anualidad por filas y los indicadores por columnas. Finalmente, podemos analizar el conjunto de matrices para cada indicador, que tiene por filas la anualidad, o las condiciones, y por columnas el territorio. Evidentemente será ya labor posterior la extracción de información a partir de cada análisis de dos vías en comparación con la vía a partir de la que se ha segmentado, o desplegado, nuestro conjunto de datos. Tampoco hemos de pensar que los métodos de tres vías resuelven completamente este escenario ya que algunos de los métodos ofrecen respuestas diferentes en función del orden que establezcamos para las vías en el análisis concretamente aplicado.

A continuación, presentaremos algunos de los métodos de análisis de datos multivariantes que pueden resultar, a priori, más prometedores para extraer información relevante de nuestro conjunto de datos, con el objetivo de resolver los desafíos propuestos. Todos ellos disponen de una interpretación geométrica, que, junto con la posibilidad de representar gráficamente los resultados, nos facilitará su posterior interpretación.

1. **Representación en coordenadas paralelas[32]:** Estamos acostumbrados a representaciones gráficas bidimensionales, en las que vectores de dos componentes se representan como un punto en un plano cartesiano, dando lugar a lo que conocemos como un gráfico de dispersión. Esta situación bidimensional puede trasladarse al espacio 3D, con tres ejes perpendiculares entre sí, pero no a espacios con un mayor número de dimensiones. Las coordenadas paralelas vienen a resolver este problema. En esta representación cada una de las variables a representar dispone de un eje propio, como sucedía en las representaciones cartesianas, pero en este caso, en lugar de situarse perpendicularmente entre sí, lo hacen en paralelo. Cada vector que representar, en lugar de constituir un punto en el gráfico, se representa como una línea que une cada eje paralelo según el valor de la componente correspondiente del vector. Si bien puede parecernos una representación un poco extraña, por lo infrecuente de la misma, seguramente estaremos más familiarizados con representaciones denominadas "radar" o "tela de araña", que no son más que representaciones en coordenadas paralelas en las que los ejes parten de un mismo punto central. Si bien es cierto que un elevado número de vectores a representar bajo cualquiera de ambos métodos, coordenadas paralelas o "radar", puede enmarañar la representación obtenida hasta hacer imposible la extracción de

[32] Inselberg, A. (1985). The plane with parallel coordinates. The Visual Computer.

información, hay muchas situaciones en las que este tipo de gráficos pueden ayudar en la exploración de nuestros datos y facilitar así el extraer conclusiones. La representación en coordenadas paralelas también puede utilizarse con los resultados obtenidos de otros métodos que normalmente se representan en 2D o 3D, para aumentar la dimensión de la representación, a costa, eso sí, de reducir la "interpretabilidad" de la misma.

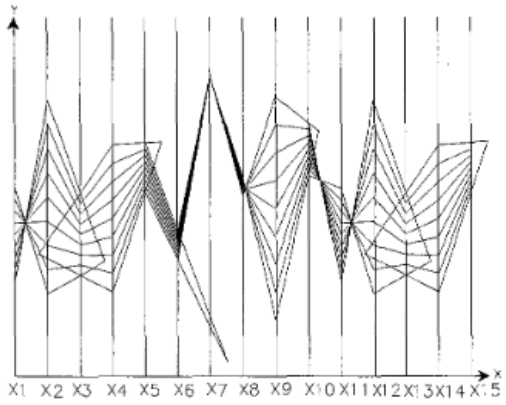

Figura: Representación en coordenadas paralelas [32]

2. **Métodos de clustering**[33]: También denominados "de conglomerados". Se trata de un método para analizar tablas de dos vías, que permite identificar *clusters*, agrupaciones, o conglomerados, de individuos similares según sus atributos. Esta técnica se incluye entre los métodos de aprendizaje no supervisado y casi como un incipiente método de "inteligencia artificial", en el sentido de que permite clasificar elementos de manera automática sin intervención humana. El método en sí es muy intuitivo: cada individuo (recordemos que en nuestro caso se corresponderían con territorios) se identifica con un vector compuesto por los valores de cada uno de sus atributos (en nuestro caso, los indicadores). Este vector se corresponde con un

[33] Jain, A. and Dubes, R. (1988) Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs.

punto en el espacio multimensional de los atributos/indicadores, espacio en el que se pueden situar geométricamente el resto de los individuos/territorios. Pues bien, es posible que se puedan identificar agrupaciones de individuos/territorios que se encuentren "próximos" entre sí. El concepto de proximidad, tal y como lo conocemos, es fácil de comprender, si bien existen varias formulaciones matemáticas para el mismo. Este es, de hecho, uno de los parámetros que deberemos concretar antes de aplicar el método de clustering seleccionado: la "distancia matemática" vamos a utilizar. El concepto que habitualmente utilizamos como distancia es la conocida como distancia euclídea, pero existen otras muchas definiciones de distancia a nuestra disposición, como, por ejemplo, la distancia Manhattan, Mahalanobis, y Bray-Curtis, entre otras. Igualmente se han formulado múltiples métodos de clustering en la literatura. Algunos parten del total de individuos y van segmentando los mismos según su proximidad, y otros métodos parten de un individuo y le van agregando otros situados en su proximidad. Por citar algunos, destacamos los métodos SLINK, CLINK, Ward y el popular *K-means,* o K-medias. El resultado de un análisis de cluster puede ser un dendrograma, que algunos autores también denominan dendograma, y que se trata de una forma de representar el resultado de un análisis de clustering jerárquico. Esta representación es un árbol cuyos nudos se sitúan en los niveles de proximidad que ocasionan la agrupación correspondiente. Otro modo de visualizar los resultados de un análisis de clustering puede ser simplemente la agrupación de los diferentes individuos en subconjuntos. Lógicamente, al final le corresponde al analista la interpretación de estas agrupaciones y la posterior extracción de las conclusiones del análisis obtenido.

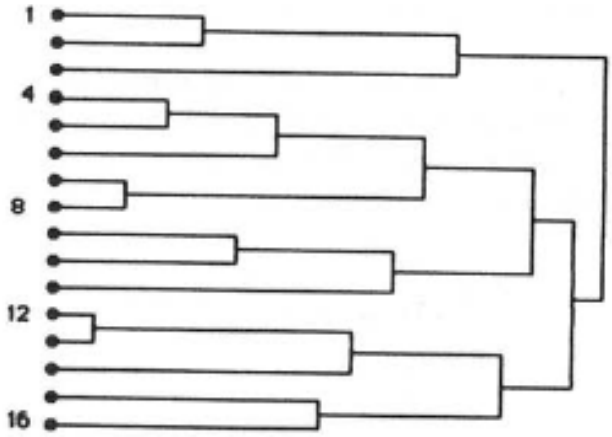

Figura: Análisis de cluster. Dedrograma [33]

3. **Métodos de regresión lineal:** la regresión lineal constituye un método de modelado de datos para datos de dos vías muy utilizado y estudiado desde niveles educativos básicos. Esto no es óbice para que pueda aplicarse en numerosos casos prácticos, con buenos resultados. Recordemos en este punto, de nuevo, el "Cuarteto de Anscombe" y los problemas que la regresión lineal puede ocasionarnos, si no somos precavidos en su uso. Considerando el conjunto de datos de que disponemos, no podremos aplicar un modelo de regresión lineal "simple", sino que tendremos que recurrir a un modelo de regresión lineal múltiple. Es poco probable que nuestros datos cumplan todas las hipótesis que requiere la utilización de este modelo, no obstante, tampoco estamos buscando un modelo de alto nivel de predicción, solo estamos intentando conocer sobre qué variables independientes podemos influir, que afecten en mayor medida a nuestros indicadores. Los métodos de regresión disponen de criterios de bondad de ajuste que nos pueden facilitar el conocer información acerca de lo bien que está ajustando/operando el modelo. Disponen también de procedimientos para "filtrar" las variables que estamos utilizando en la obtención de los parámetros constitutivos del modelo, facilitando la eliminación de aquellas variables que son irrelevantes para el resultado. En definitiva, a pesar de su supuesta sencillez, el modelo de regresión múltiple merece ser

considerado en nuestro estudio, ya que puede aportarnos valiosa información sobre nuestros datos, poniendo en evidencia posibles interacciones entre nuestros indicadores/variables independentes y dependientes.

4. **Análisis de Componentes Principales (ACP)**[34]: Como en el caso de la regresión lineal, se trata de un método clásico de análisis exploratorio de datos, también dirigido a datos de dos vías. En este caso el método no considera variables dependientes e independientes, como hemos establecido en nuestro problema, sino que todas las variables son iguales. Las condiciones que precisan cumplir los datos para ser utilizado este método son menores que el caso de la regresión lineal. El Análisis de Componentes Principales, o ACP, permite reducir el número de variables observadas a un número más pequeño de variables (denominadas componentes principales) que retengan la mayor parte posible de la información (variabilidad) contenida en los datos originales y que son el resultado de una combinación lineal de las variables de partida. Estas nuevas variables, que son las componentes principales, se pueden ordenar según su importancia en la capacidad de explicación de los datos originales, lo que nos permite reducir la dimensión de los mismos. El propósito general es reducir la dimensión de los datos con el fin de poder interpretar las similitudes entre individuos. El ACP está íntimamente relacionado con la Descomposición en Valores Singulares (DVS/SVD) de la matriz de datos. El reto del método realmente consiste en descubrir el significado contenido en estas nuevas variables, lo que se consigue mediante el estudio de las correlaciones con las variables originales. En la representación gráfica que se puede obtener, las distancias entre individuos, en nuestro caso probablemente serán los territorios, se interpretan en términos de similitud entre ellos. Y también se pueden extraer conclusiones en función de la posición con relación a las

[34] Jolliffe, I. (2002). Principal component analysis. Aberdeen: Springer.

componentes principales obtenidas y la interpretación que se haya dado a las mismas. El método es aplicable, con una interpretación similar a la expuesta, a series temporales o numéricas, permitiendo extraer de estas sus componentes comunes y detectar posibles anomalías, por ejemplo.

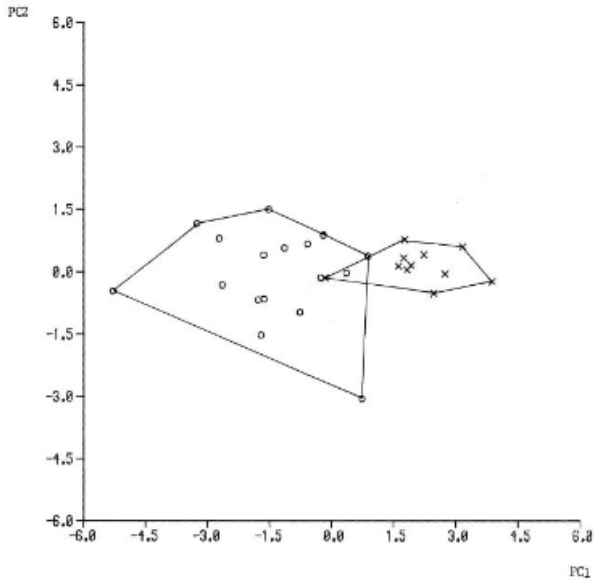

Figura: Análisis de Componentes Principales[34]

5. **Análisis Factorial (AF)[35]:** Se trata de un método muy parecido al Análisis de Componentes Principales. Como el ACP, el Análisis Factorial intenta explicar un conjunto de variables observables, lo que serían nuestros indicadores, mediante un número reducido de variables hipotéticas, denominadas factores. También como en el ACP, todas las variables observables juegan idéntico rol, sin establecerse una diferenciación entre variables dependientes o independientes. Algunas de las diferencias con el ACP es que las variables que resultan del Análisis Factorial son nuevas, sin que puedan relacionarse

[35] Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. The American Journal of Psychology, 15(2), 201–292.

mediante una combinación lineal de las variables de partida, como ocurría en el ACP. De hecho, en el AF sucede a la inversa, las variables de partida se relacionan mediante una combinación lineal de los factores obtenidos. Otra diferencia es que el ACP explica la variabilidad total, mientas que el AF se centra más en las correlaciones entre variables, entendida como la variabilidad común a todas ellas. En principio, el Análisis Factorial parece más adecuado a nuestros objetivos, dado que persigue identificar los constructos que subyacen en los datos. Algunos autores entienden que ambos métodos no tan solo no son excluyentes, sino que son realmente complementarios: realizado en primer lugar, el ACP puede extraer determinada información que resulte de utilidad al efectuar el AF sobre ese mismo conjunto de datos.

6. **Métodos biplot[36]:** Un biplot es una representación gráfica de datos multivariantes de dos vías, que representa una matriz mediante un vector para cada una de sus filas y otro vector para cada columna. A estos vectores se les denomina marcadores, así se establecen marcadores fila y marcadores columna. El método de cálculo de dichos marcadores, también mediante la DVS/SVD, facilita la obtención de una aproximación de menor rango, posibilitando la proyección de dichos vectores en un espacio bidimensional más fácilmente interpretable. De la representación gráfica de dichos marcadores puede extraerse información sobre relaciones entre marcadores fila, entre marcadores columna, y entre marcadores fila y columna. Es habitual disponer en filas los individuos, en nuestro caso probablemente serán los territorios, y en columnas, las variables, en nuestro caso los denominados indicadores. Así, en la representación bidimensional, la proximidad entre los marcadores fila (territorios) se

[36] Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58, 453-467. Galindo, M. P. (1986). Una alternativa de representación simultánea: HJ-biplot. Questiió, 10(1),13-23.

relaciona con la similitud entre territorios, los ángulos que forman los marcadores de las columnas, con la correlación entre indicadores y, finalmente, las proyecciones ortogonales de los marcadores fila sobre los marcadores columna hacen referencia a los valores que los territorios tienen sobre los indicadores. Por si toda esta información no fuese suficiente, el método permite disponer de indicadores que reflejan no solo la calidad de la representación global, sino incluso la calidad de representación de cada marcador de forma individual.

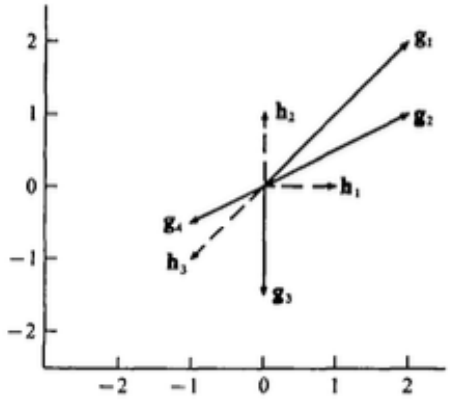

Figura: Análisis Biplot[36]

7. **Desplegamiento multidimensional (*unfolding*):** Hay versiones encuadradas en la familia de los métodos *unfolding* para datos de dos y tres vías[37]. Es una variante del Escalado Multidimensional (MDS) tradicional que representa en un mismo sistema de referencia los individuos (territorios) y las alternativas (indicadores) como puntos. La distancia entre individuos (territorios) se interpreta como similitud, la distancia entre alternativas (variables) como similitud entre alternativas. Por último, la distancia de un individuo a una alternativa

[37] DeSarbo, W. S, *et al* (2009). A three-way clusterwise multidimensional unfolding procedure for the spatial representation of context dependent preferences. Computational Statistics and Data Analysis, 53, 3217–3230

se interpreta en términos de preferencia. Tiene también una interpretación "ecológica" en la que la abundancia de una especie (por ejemplo, un servicio) puede ser considerada como una medida de la preferencia por un lugar (territorio) con unas determinadas características (indicadores), en el caso del *unfolding* de tres vías.

8. **Análisis Canónico de Correspondencias (ACC)[38]:** En la literatura se denomina a veces como análisis (directo) del gradiente, o se engloba entre esas técnicas. Se trata, como sucedía con el desplegamiento multidimensional, de un método con aplicaciones en ecología, que podría considerarse de tres vías, pero vistas estas de otra manera que en casos anteriores. Consiste en analizar dos matrices de datos, una que contiene información sobre la distribución de especies (abundancia/frecuencia) en varios lugares de muestreo, y otra matriz que contiene información ambiental (variables climáticas, variables de suelo, etc.) recogidas en esos mismos lugares de muestreo. El objetivo es conocer qué combinación de variables ambientales es responsable de la distribución de las especies. Es evidente la potencial aplicabilidad del método a nuestro problema: disponemos de unos territorios (los lugares de muestreo) en las que podemos medir la población, nacimientos, defunciones, migraciones, o los diferentes servicios disponibles (la distribución de especies) y las actuaciones que se realizan y/o la situación concreta de cada territorio (la información ambiental). El paralelismo de los escenarios es obvio. A diferencia de otros métodos anteriores, el ACC no busca explicar la variabilidad de los datos de la comunidad, sino la variabilidad que es explicada por las variables ambientales consideradas en el estudio. Deseamos, como no podría ser de otra manera, detectar las relaciones existentes entre la distribución de las especies (población y servicios) y la

[38] Ter Braak, C. J. F. (1986). Canonical Correspondence Analysis: A new Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology, Ecological Society of America, 67(5), 1167-1179.

información ambiental (actuaciones o variables/indicadores dependientes), los lugares de muestreo concretos tienen en este caso un rol secundario. El método obtiene como resultado varias representaciones gráficas que permiten extraer interpretaciones sobre los efectos de las variables dependientes sobre las independientes, tal y como buscamos en nuestros desafíos. Al no contemplar este método un eje temporal, la comparación entre intervalos temporales deberá realizarse a partir del análisis de los diferentes resultados obtenidos por separado para cada uno de dichos intervalos. Pero recordemos que el eje "temporal" puede interpretarse tan solo como un índice secuencial para las diferentes condiciones existentes cada momento.

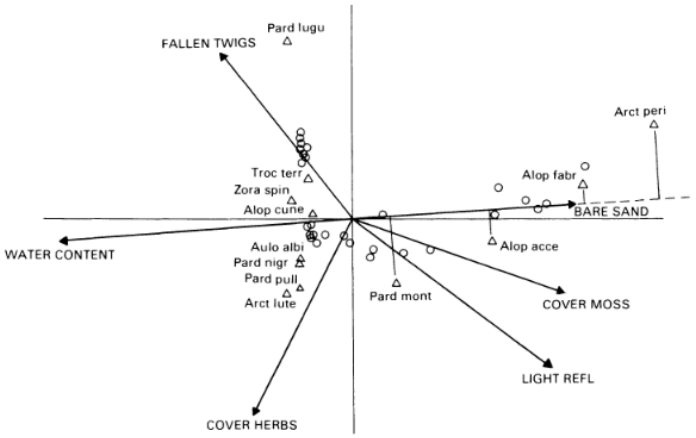

Figura: Análisis Canónico de Correspondencia[38]

9. **Función de Correlación Cruzada Muestral (FCCM):** Hemos expuesto métodos que no tienen en consideración explícitamente la componente temporal de nuestros datos. En este caso vamos a trabajar precisamente solo con dicha componente temporal. Intentaremos extraer información de la correlación de las series numéricas que componen nuestro juego de datos o, más propiamente, vamos a aplicar la Función de Correlación Cruzada Muestral. Las series numéricas sobre las que se aplica dicha función normalmente corresponderán con series temporales, pero no tiene

por qué ser siempre así. Si las series numéricas/temporales se encuentran normalizadas, la correlación tendrá el valor unitario para un ajuste perfecto, pero, es más, en el caso de tratarse de una serie temporal, el punto en el que dicho valor se obtenga nos aportará información sobre el posible desfase existente entre dichas series. El modo de proceder sería, por ejemplo, calcular una matriz cuadrada en la que cada elemento sería la función de correlación cruzada muestral entre la variable correspondiente a cada fila y cada columna. La diagonal principal obviamente se correspondería con el ajuste perfecto, al tratarse de la correlación consigo misma, o autocorrelación, de la serie numérica tratada al efecto. También podemos obtener esta matriz de series como la FCCM solo entre las variables/indicadores dependientes e independientes, para poner de manifiesto las relaciones entre ambos tipos de indicadores.

10. **CANDECOMP/PARAFAC (CP)[39]:** Este método se puede interpretar como un Análisis de Componentes Principales extendido para datos multivía/multimodo, en nuestro caso con tres modos. Como ocurría en el ACP, también se ha de decidir cuántas componentes retendremos en nuestro análisis, con el objeto de reducir la dimensionalidad de los datos, manteniendo la mayor parte de la información/variabilidad. En este caso el número de componentes elegido tiene que ser el mismo para todos los modos estudiados. Los pesos obtenidos por el modelo, que se relacionan con las variables de partida, solo vinculan una componente de un modo con otra componente de otro modo. Es importante tener en consideración que CANDECOMP/PARAFAC responde al principio denominado de "perfiles paralelos proporcionales", lo que quiere decir, someramente, que los efectos

[39] Carroll, J.D.; Chang, J.J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika 1970, 35, 283–319. Harshman, R.A. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis. UCLA Work. Pap. Phon. 1970, 16, 1–84.

sobre las variables de las diferentes condiciones tienen un cierto paralelismo: los perfiles que presentan las variables en un estudio son los mismos que en otros estudios, pero desplazados, en función de las diferentes condiciones experimentales. Al contrario de lo que sucedía con el ACP, en este caso los resultados obtenidos no están directamente relacionados con las correlaciones entre las variables y las componentes. Sin embargo, sí que es posible extraer información sobre la importancia que las distintas componentes tienen en el modelo, aunque esta importancia no esté relacionada con la desviación estándar, como sucedía en el ACP. El modelo obtenido para un número determinado de componentes es único, lo que convierte a CANDECOMP/PARAFAC en un método muy interesante para conseguir una representación de los patrones estructurales existentes en los datos. Por otro lado, no siempre será posible obtener la solución al método para cualquier conjunto de datos, pudiendo resultar incluso en soluciones degeneradas o inestables. En este supuesto suele ser habitual recurrir a otros modelos diferentes, como por ejemplo TUCKER3.

11. **TUCKER3** [40] **:** A diferencia de CANDECOMP/PARAFAC, el modelo obtenido por el método TUCKER3 proporciona vinculación entre todas las componentes de todos los modos entre sí, lo que hace el modelo más complejo, pero más flexible para ajustar los datos. Como en casos anteriores, hemos de seleccionar el número de componentes que se retendrán en nuestro modelo en el espacio de dimensión reducida. A diferencia de lo que sucedía con el método CANDECOMP/PARAFAC, se puede establecer que cada modo retenga un número diferente de componentes, lo que proporciona de nuevo una gran flexibilidad. Además, TUCKER3 no responde al principio de "perfiles proporcionales

[40] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3), 279–311. Tucker, L. (1972). Relations between multidimensional scaling and three-mode factor analysis. Psychometrika, 37(1), 3-27.

paralelos", como sucedía en CANDECOMP/PARAFAC. El resultado del método TUCKER3 no es único, pudiendo aplicarse rotaciones, lo que lo convierte en menos eficaz que CANDECOMP/PARAFAC para obtener estructuras más simples y fácilmente interpretables. No obstante, si lo que buscamos, como en nuestro caso, es descubrir los patrones que vinculan las variables originales, las entidades y las condiciones existentes, cuando disponemos de poca información a priori, más que la búsqueda de entidades latentes (como sucedía en el ACP o CP), la libertad de rotación de ejes es una ventaja que considerar. TUCKER3 será también una opción destacable si CANDECOMP/PARAFAC no obtiene resultados satisfactorios, bien sean degenerados o sin convergencia, ya que TUCKER3 no presenta ese tipo de problemas con las soluciones, la convergencia del algoritmo o con dificultades de cálculo.

12. **Métodos STATIS[41]:** Los comúnmente denominados métodos STATIS constituyen herramientas exploratorias específicas para el análisis de datos de tres vías, como el caso en el que nos encontramos. STATIS es el acrónimo de la expresión francesa "*Structuration des Tableaux à Trois Indices de la Statistique*". Esta técnica también es conocida bajo el acrónimo de "ACT", proveniente de la también expresión francesa "Analyse Conjointe de Tableaux". En el fondo STATIS/ACT no es más que una generalización del Análisis de Componentes Principales (ACP). El objetivo de los métodos STATIS/ACT es analizar varios conjuntos de variables (no necesariamente las mismas) medidas en el mismo juego de observaciones (individuos u objetos), o varios juegos de observaciones (no necesariamente las mismas) medidas sobre mismo juego de variables. La idea principal es comparar diferentes tablas de datos (matrices) obtenidas bajo varias condiciones experimentales, pero conteniendo el mismo número de filas y/o columnas. Del análisis STATIS/ACT se obtienen dos tipos de resultados

[41] L'Hermier des Plantes, H. (1976). Structuration des tableaux à trois indices de la statistique. Université de Montpellier II. Y Lavit C., *et al*, The ACT (STATIS method), Computational Statistics & Data Analysis, Volume 18, Issue 1, 1994, Pages 97-119.

diferentes. El primero consiste en un análisis de la similaridad entre las estructuras de los diferentes juegos de datos (matrices) y que se denomina análisis de la "interestructura". Este primer análisis proporciona un juego de pesos que serán utilizados para obtener el segundo resultado, que consiste en el análisis de la "intraestructura" o estructura interna de las diferentes tablas o matrices que componen nuestro juego de datos.

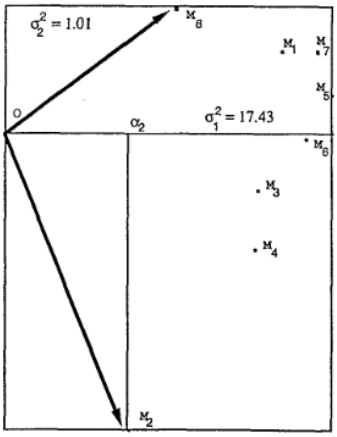

Figura: Interestructura de matrices en análisis STATIS [41]

Hemos expuesto, muy brevemente, algunos de los métodos de análisis de datos multivariante más populares en la literatura para datos de dos y tres vías, entre otros muchos posibles que pueden considerarse. Una vez aplicados tentativamente a nuestros juegos de datos, y analizados los resultados obtenidos, podremos, quizá, disponer de indicaciones más específicas sobre la aplicabilidad de otros métodos existentes que pudieran ser más adecuados a nuestros objetivos, si bien los citados anteriormente parecen bastante prometedores en ese sentido.

7. HITO #5: ELABORAR PROPUESTAS DE ACCIÓN RESULTANTES DEL ANÁLISIS

A lo largo del desarrollo de los trabajos que han dado lugar a este documento, se han detectado algunas estrategias que es posible adoptar por las entidades locales con el objetivo de incrementar las probabilidades de éxito en la aplicación futura de las diversas técnicas propuestas.

#1 - Disponer de un liderazgo sensible a la aplicación de estrategias de "orientación al dato"

La aplicación de técnicas de análisis orientadas al dato pasan por un liderazgo que apoye este tipo de iniciativas. No será tanto por las inversiones económicas necesarias, como por los recursos humanos que deben participar en el proyecto y su necesaria capacitación.

Lógicamente, como paso previo a las iniciativas expuestas, los responsables políticos de las diferentes administraciones deben ser conscientes de las posibilidades que los datos tienen en la aplicación de estrategias públicas. Y como paso inicial, la toma en consideración por estos mismos responsables de que el dato constituye un bien público, en este caso, digital[42], que deben gestionar con los mismos principios que regirían la gestión de cualquier otro bien público tangible.

[42] https://datos.gob.es/es/blog/los-datos-abiertos-como-bienes-digitales-publicos

Una figura que puede resultar clave para liderar esta "orientación al dato" es la denominada CDO (*Chief Data Officer*), que se puede traducir libremente como "jefatura de datos". Este cargo ejerce como responsable de los proyectos de datos en la organización y dispone de "asiento" entre los restantes perfiles directivos. Se trata de una figura de nueva adopción en la administración pública y, como veremos, la Administración General del Estado nombró su CDO en julio del año 2021.

Ciertamente, existen claras reticencias, por los propios responsables políticos y por el resto de la ciudadanía, a la utilización generalizada de los datos por las administraciones, tanto para la extracción de información, como para la puesta a disposición pública de los mismos. Para sortear estos obstáculos, la transparencia en las actuaciones que se lleven a cabo juega un papel fundamental, como ampliaremos posteriormente, junto con la capacidad de poder demostrar el cumplimiento de la legislación de protección de datos y la adecuada anonimización de los datos publicados por parte de las administraciones.

#2 - Constituir un equipo de trabajo multidisciplinar para abordar el proyecto

Se deben configurar equipos multidisciplinares, apoyándose en el liderazgo del CDO, compuestos por personal informático, analistas de datos o científicos de datos, expertos en privacidad y, por supuesto, expertos en los desafíos concretos propuestos, que conozcan el escenario global en el que se desarrolla el estudio, entre otros posibles perfiles. Estos equipos deberán instar a otras unidades a facilitar conjuntos de datos para su tratamiento, proceder a su posterior publicación y, por supuesto, someterlos a los oportunos análisis para la obtención de información de interés.

Los datos juegan, y jugarán, un papel tan relevante en la administración pública que, siguiendo los pasos del sector privado, se creó en el año 2020 la "Oficina del Dato" de la administración general del estado[43]. Entre sus funciones se incluye "El desarrollo de un Centro de Competencia de analítica avanzada de datos que defina las metodologías y mejores prácticas y que asegure que se desarrollan las competencias tecnológicas y las herramientas necesarias para la toma de decisiones basadas en datos por parte de las Administraciones Públicas, permitiendo el desarrollo de políticas basadas en evidencia." Es más que previsible que, tras los pasos de la AGE, tanto administraciones autonómicas como locales aborden la creación de diferentes "Oficinas del dato", si es que no lo han hecho ya. Además, la Administración General del Estado puso al frente de su "Oficina del Dato" en 2021 a un CDO (*Chief Data Officer*), cuyo nombramiento seguramente también acometerán en el futuro otros niveles administrativos.

Como ya hemos citado anteriormente, se convocaron en 2022 líneas de subvenciones para la transformación digital y modernización de las entidades locales, entre cuyas finalidades podía incardinarse la creación de estas "Oficinas del Dato", por lo que suponemos que algunas administraciones habrán presentado proyectos en este sentido.

#3 - Fomentar la publicación de datos abiertos en el sector público y privado

Como parte imprescindible de la estrategia de orientación al dato, es evidente la necesidad de publicación de datos, siguiendo las

[43] Orden ETD/803/2020, de 31 de julio, por la que se crea la División Oficina del Dato y la División de Planificación y Ejecución de Programas en la Secretaría de Estado de Digitalización e Inteligencia Artificial.

recomendaciones emanadas de múltiples instituciones. Esta disponibilidad facilitará a todas las administraciones públicas la puesta en valor de los datos que obran en su poder. La publicación de datos abiertos debe llevarse a cabo con las debidas características que faciliten su reutilización y con garantías de privacidad.

Todas estas actuaciones no se conciben sin un mayor compromiso por parte de las administraciones con la publicación de datos, que no debe ser algo accesorio, o accidental en el tiempo, sino que debe estar embebido en la filosofía de trabajo la propia administración. Una piedra angular de toda esta filosofía es la Reutilización de la Información del Sector Público (RISP) que consiste en poner la información del sector público disponible, en formatos estándar abiertos, facilitando su acceso y permitiendo su reutilización tanto a particulares como a empresas, para fines comerciales o no. Y es que tener acceso a los datos de la Administración garantiza la transparencia y fomenta la igualdad. Además, estos datos abiertos tienen el poder de transformar la interacción entre los gobiernos, las empresas y la sociedad, permitiendo desarrollar nuevas aplicaciones para resolver problemas y obteniendo así beneficios sociales, económicos o medioambientales para toda la sociedad[44].

La finalidad última, es que esos datos de los que dispone la Administración y que son de dominio público, excluyendo datos personales o comprometidos, sean expuestos para que ciudadanos y empresas puedan beneficiarse haciendo uso de ellos. A las personas físicas o jurídicas que reutilizan la información se les conoce como agente

[44] Para más información consultar la Guía Aporta sobre reutilización de la información del sector público, en el marco establecido en la Ley 37/2007.

https://datos.gob.es/es/acerca-de-la-iniciativa-aporta

reutilizador o infomediario[45], dedicadas a analizar y tratar información procedente del sector público y privado, para crear productos de valor añadido destinados a terceras empresas o a la ciudadanía en general. Son, por tanto, uno de los agentes principales a la hora de convertir la información en un activo económico cuantificable.

Los datos abiertos pueden aportar distintos beneficios a los gobiernos, las empresas y las personas. Sirven para ayudar a mejorar los servicios, impulsar el crecimiento de las economías y proteger nuestro planeta.

La apertura de datos públicos contribuye a revalorizar la actividad pública. También contribuye a mejorar la calidad de los datos y la monitorización y ajuste de políticas públicas. Fomenta nuevos servicios para el ciudadano y nuevas oportunidades de negocio en el sector de la reutilización de datos. Facilita la participación, colaboración y transparencia de cara a favorecer unos servicios públicos mejores y más eficientes.

Los datos que obran en manos de las Administraciones Públicas, por su propio funcionamiento, tienen un gran potencial para la generación de beneficios económicos, sociales, medioambientales e incluso para la propia administración [46]. Pero para que los datos abiertos por las administraciones lleguen a producir beneficios se precisa de toda una estrategia activa que implique a los distintos agentes del ecosistema de la reutilización.

[45] https://www.idee.es/resources/presentaciones/JIIDE16/2016/PDFpresentaciones/41_ASEDIE_ReutilizacionInformacionValorSectorInfomediario.pdf y https://datos.gob.es/es/noticias-tags/sector-infomediario

[46] https://datos.gob.es/es/documentacion/la-reutilizacion-de-datos-publicos-en-su-papel-transformador

#4 - Identificar nuevos desafíos que puedan ser objetivables mediante datos

Hemos establecido, entre otras posibilidades, que los dos desafíos que pretendemos sean abordados inicialmente utilizando los datos como herramienta desde las administraciones públicas, serían la despoblación y la pérdida de servicios en el medio rural. No obstante, sería igualmente posible identificar por cada organización otros desafíos, también abordables desde una orientación al dato, que pudiesen resultarles más interesantes.

Los dos desafíos planteados han motivado múltiples actuaciones por parte de diferentes administraciones públicas, que han pretendido resolverlos mediante acciones, más o menos estructuradas, que fijasen población y habilitasen o mantuviesen la disponibilidad de servicios en el medio rural. En este último desafío, los servicios, las acciones llevadas a cabo se han visto claramente facilitadas en determinados casos debido a que algunos de los servicios son prestados por las propias administraciones públicas, como sucede con aquellos relativos a la salud y la educación.

#5- Aplicar metodologías reconocidas en la gestión del proyecto

Es importante que los proyectos que impliquen tratamientos de datos para la obtención de información que ayude en el diseño de políticas públicas, sean abordados mediante la aplicación de metodologías avaladas por expertos. Entre todas ellas resaltaríamos dos metodologías fácilmente aplicables y que incrementarán las posibilidades de éxito de

nuestros proyectos de puesta en valor de los datos. Nos referimos a las técnicas ágiles y a la metodología CRISP-DM.

Si bien los Principios del Manifiesto Ágil[47] se llevan aplicando desde hace años en el desarrollo de programas informáticos, no es menos cierto que sus principios caben en la gestión de prácticamente cualquier tipo de proyecto. La entrega continua de resultados, el facilitar los posibles cambios en los objetivos del proyecto y el trabajo conjunto del equipo de proyecto, son algunos de los doce principios que constituyen el Manifiesto Ágil. Ciertamente los proyectos que precisan de dilatados periodos de tiempo para disponer de resultados tangibles corren grave riesgo de quedar relegados, en un tiempo actual en el que la sociedad demanda inmediatez en la obtención de respuestas a sus demandas. Tampoco la rigidez en los objetivos es algo que se pueda tener a gala en el escenario actual. La velocidad a la que cambia nuestra sociedad, que hemos podido sentir en los últimos tiempos con la pandemia de Covid-19 y la guerra de Ucrania, por poner dos ejemplos, precisa que los proyectos que abordemos dispongan de los mecanismos necesarios para facilitar el cambio rápido de nuestros objetivos, manteniendo las estructuras que apoyan el desarrollo de los proyectos. Finalmente, el trabajo colaborativo entre expertos de diversas ramas del conocimiento es imprescindible hoy en día para que todo proyecto llegue a buen fin. Ningún experto en analítica, o en tecnologías de la información, o en despoblación, o en provisión de servicios, por ejemplo, va a disponer de todo el conocimiento necesario para abordar individualmente los problemas que surgirán en el desarrollo de cualquier proyecto basado en la "orientación al dato".

[47] https://agilemanifesto.org/iso/es/principles.html

La utilización de metodologías específicas para la aplicación del análisis de datos a la resolución de problemas, como por ejemplo CRISP-DM[48] (*Cross Industry Standard Process for Data Mining*), nos puede servir para iluminar un camino pautado hacia la obtención de respuestas a nuestros desafíos. CRISP-DM es una metodología consolidada, que ya tiene más de 20 años de aplicación, y que está enfocada explícitamente al desarrollo de proyectos de minería de datos. El modelo de referencia de CRISP-DM comprende seis fases: comprensión del negocio (en nuestro caso, del problema/desafío), comprensión de los datos, preparación de los datos, modelado, evaluación y, finalmente, distribución. Este modelo es flexible y puede personalizarse fácilmente en función de la organización que lo aplica y del objetivo establecido en cada proyecto. La documentación disponible del método es abundante y muy detallada, incluyendo tanto el propio modelo de referencia, como una guía para el usuario. El modelo de referencia comprende una descripción de las distintas fases de la metodología, así como las entradas y salidas de cada fase, y las tareas a llevar a cabo en las mismas. Ciertamente CRISP-DM no es la única metodología que podemos aplicar en nuestros trabajos de minería de datos, por lo que sugerimos explorar otras alternativas existentes en la literatura[49,50] para seleccionar aquella que consideremos idónea para nuestro proyecto.

Sobre la publicación de datos abiertos existen igualmente directrices que es posible seguir para maximizar nuestras probabilidades de éxito en este apartado. La armonización de principios para los datos abiertos es un esfuerzo que ha dado como resultado la *Open Data Charter*[51], la Carta

---

[48] *Chapman, P. et al. "CRISP-DM 1.0: Step-by-step data mining guide." (2000).*

[49] Moine, J.M. *et al*. Estudio comparativo de metodologías para minería de datos. https://core.ac.uk/download/pdf/301040544.pdf [Enlace verificado el 31/01/23]

[50] Marco de referencia de la DAMA (*Data Management Association*) https://dama.org

[51] http://opendatacharter.net/principles-es/

Internacional de Datos Abiertos, con principios y buenas prácticas para la publicación de datos gubernamentales que aspiran a ser considerados como datos abiertos. Como se ha expuesto ya anteriormente, se definen los Datos abiertos como datos digitales que son puestos a disposición con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar.

#6 - Profundizar en la aplicación de la analítica avanzada de datos al estudio de los datos disponibles

Es evidente que los datos sin herramientas de análisis son como los ladrillos sin cemento. Es imprescindible avanzar en la identificación de métodos matemáticos que puedan ayudarnos en la extracción de conocimiento sobre los retos establecidos. Para ello se necesita disponer de herramientas informáticas que posibiliten el procesado de los datos y su análisis. Las disponibilidades presupuestarias no constituyen ningún problema para ello, ya que, en gran medida, todas estas herramientas se pueden obtener de forma gratuita, al tratarse en casi todos los casos de desarrollos realizados sobre software libre. No obstante, sí que es preciso disponer de personal con los conocimientos apropiados para la utilización de estas herramientas, bien a través de la captación de nuevo talento o bien capacitando a personal propio para ese fin.

En el ejemplo seguido, corresponde a las administraciones públicas identificar las posibles actuaciones que puedan realizar y que afecten positivamente a los indicadores de población o de la disponibilidad de servicios, algo que no es sencillo. Y ese es el fin con el que se propone la utilización de herramientas matemáticas. Esta aplicación debe ser muy cuidadosa en la aplicación de los algoritmos propuestos, para evitar posibles sesgos en los resultados. La necesaria transparencia en los algoritmos deberá facilitar su estudio por los agentes participantes, como

garantía de su correcta aplicación. Esto es algo de lo que ya existen iniciativas en curso[52], incluida entre ellas la creación del Centro Europeo para la Transparencia Algorítmica[53], con sede en Sevilla.

#7 – Explorar las posibilidades existentes para construir una plataforma que facilite la aplicación de estrategias orientadas al dato

Una posible vía para facilitar que las administraciones locales puedan recurrir a estrategias de orientación al dato, en general, y en particular a la aplicación de herramientas de análisis para la extracción de información de los datos que posean, bien podría ser la construcción de una plataforma que aglutine todos los datos aportados por las entidades locales y disponga de algoritmos de análisis listos para su aplicación sobre dichos datos. Todos los conocidos sistemas hiperescalares ya provisionan entre sus servicios este tipo de plataformas de análisis de datos, listas para su utilización. La propuesta que se realiza en este punto va más en el camino de poder recurrir a la iniciativa europea Gaia-X[54], cuyo objetivo es precisamente la creación de una infraestructura de datos abierta y segura, que habilite una soberanía de datos europea. De facto, dentro del proyecto Gaia-X ya se aboga, entre otros muchos proyectos, por la creación de un Hub de Ciencia de Datos y Aprendizaje Automático[55].

El planteamiento propuesto consistiría en el despliegue de una plataforma que aglutinase un portal común para residenciar los

[52] https://www.algorithmregister.org/

[53] https://algorithmic-transparency.ec.europa.eu

[54] https://www.gaiax.es

[55] https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Artikel/UseCases/data-science-and-machine-learning-hub.html

conjuntos de datos facilitados por las administraciones locales, que podría estar abierto al público, o no, en función de los condicionantes marcados por las entidades que aportasen los conjuntos de datos. Incluso esa apertura podría determinarse a nivel de conjunto de datos, o de variables específicas de cada conjunto de datos. Por otro lado, se dispondrían de algoritmos de análisis de datos, y de métodos de representación, que operasen directamente con los datos almacenados en la plataforma, con el objeto de facilitar la extracción de información y resultados. De este modo se ofrecería de manera conjunta todas las herramientas necesarias para que una entidad local, independientemente de su tamaño, pudiese aplicar una estrategia de orientación al dato en su gestión y en la adopción y evaluación de sus políticas públicas.

Quizá se podría estudiar si no sería conveniente que la FEMP se adhiriese a la Asociación Gaia-X España[54] con objeto de disponer de una posición estratégica a la hora de valorar el posible desarrollo de proyectos bajo dicho marco, en una estrategia similar a la seguida de aproximación con otros servicios hiperescalares, en este caso comerciales.

8. CONCLUSIONES

La sociedad, en general, y las administraciones públicas, en particular, crean cada día ingentes cantidades de datos. Datos que, en muchos casos, no son puestos en valor adecuadamente y que quedan almacenados sin más aportación a la sociedad. Las ciudades inteligentes, o los territorios inteligentes, son una de las posibles fuentes de datos más importantes en la actualidad, por volumen y por relevancia. Los dispositivos desplegados en el territorio, sensores y actuadores, generan grandes volúmenes de datos, que pueden aportar información muy notable. Pero también existen otras fuentes de datos, fácilmente accesibles, que pueden facilitarnos nuevas estrategias para afrontar viejos problemas, de la mano de la aplicación de métodos matemáticos avanzados.

Como conclusión de los trabajos realizados en el Grupo, es interesante resaltar que las ventajas que presenta la aplicación de una orientación al dato en la resolución de problemas públicos solo serán aprovechadas si se involucran todos los agentes concernidos en esa tarea. Se trata de una labor colectiva. Poco importa que unas pocas administraciones públicas se tomen en serio el reto, y publiquen e intenten explotar los datos de que disponen, propios o ajenos, si el resto de las administraciones no hacen lo mismo. En primer lugar, porque si otras administraciones no publican sus datos, las capacidades de análisis pueden verse reducidas significativamente. Y, en segundo lugar, y no menos importante, porque de la búsqueda de soluciones y la prueba de herramientas que realice cada administración, si difunde sus resultados, se beneficiarán otras administraciones públicas y la sociedad en general.

El objetivo de este documento, elaborado por el Grupo de Trabajo sobre "Orientación al dato para el desarrollo de territorios inteligentes", es tan solo esbozar unas breves ideas sobre las posibilidades y alternativas que

la analítica de datos más tradicional presenta para la resolución de problemas en la administración pública. No hemos abordado otras posibles técnicas, como pueden ser el *big data* o la inteligencia artificial. En el primer caso por no disponer nuestros datos de las características que hacen necesaria la aplicación de dichas técnicas tan específicas. En el segundo supuesto, porque la inteligencia artificial, ciertamente excepto en el caso de la inteligencia artificial explicable, no nos proporciona información sobre el porqué de las predicciones realizadas, y es que, en nuestro caso, buscamos precisamente los motivos por los que nuestros desafíos son tan complejos de resolver.

Si tuviésemos que sintetizar las actuaciones que se deberían afrontar para un desarrollo del territorio basado en la "estrategia del dato", estas estarían reflejadas en las propuestas de acción descritas en el Hito #5 :

1. Disponer de un liderazgo sensible a la aplicación de estrategias de "orientación al dato"
2. Constituir un equipo de trabajo multidisciplinar para abordar el proyecto
3. Fomentar la publicación de datos abiertos en el sector público y privado
4. Identificar nuevos desafíos que puedan ser objetivables mediante datos
5. Aplicar metodologías reconocidas en la gestión del proyecto
6. Profundizar en la aplicación de la analítica avanzada de datos al estudio de los datos disponibles
7. Explorar las posibilidades existentes para construir una plataforma que facilite la aplicación de estrategias orientadas al dato

Debido a las limitaciones temporales existentes para el desarrollo de los trabajos del Grupo, cuyas actuaciones efectivas no comenzaron hasta mediados del mes de septiembre pasado, no ha sido posible proceder a

la aplicación de los métodos que se han propuesto, ni siquiera sobre un conjunto de datos de prueba. Queda, por lo tanto, pendiente para fases posteriores la materialización de los análisis comentados y la tentativa de extraer de ellos indicaciones sobre las posibles actuaciones a ejecutar para, al menos, intentar paliar los efectos de los desafíos propuestos, a partir de la evidencia que proporcione el enfoque orientado al dato.

Confiamos, por todo lo anteriormente expuesto, que este trabajo, meramente prospectivo, tenga continuación en un futuro relativamente próximo, una vez tenga lugar el proceso electoral del próximo mes de mayo y se constituya una nueva FEMP, con sus Comisiones y respectivos Grupos de Trabajo, entre los que esperamos se encuentre aquel que prosiga nuestras incipientes investigaciones.

ANEXO I: CONJUNTOS DE DATOS Y POSIBLES FUENTES

Variables dependientes	Variables independientes
Despoblación	
Variación de la **población*** Variación de la **densidad de población*** [Nacimientos, mortalidad, migraciones]	Actuaciones sobre fomento de la **natalidad** Actuaciones sobre disponibilidad de **vivienda** Actuaciones sobre **alquiler** de vivienda Actuaciones sobre fomento de la **contratación** **…** ¿Disponibilidad de **servicios** en el territorio?
Disponibilidad de servicios	
Salud* (ambulatorios/CAP, hospitales, farmacias) **Educación*** (bibliotecas, escuelas infantiles y guarderías, colegios, institutos, universidades) **Transporte** (trenes, autobuses, puertos y aeropuertos, aparcamientos, carreteras/autovías) **Comercio** (supermercados, comercios…) **Industria*** (empresas) **Banca*** (oficinas y cajeros automáticos) **Turismo*** (hoteles, campings, casas rurales) **Ocio** (bares, restaurantes, pistas deportivas, piscinas) **Cultura*** (museos, cines, teatros, BICs, fiestas, festivales, centros culturales, yacimientos arqueológicos)	Actuaciones sobre servicios de **salud** Actuaciones sobre servicios **educativos** Actuaciones sobre medios de **transporte** Actuaciones para el fomento del **comercio** Actuaciones para el asentamiento de **industrias** Actuaciones para habilitar **servicios bancarios** Actuaciones sobre promoción del **turismo** Actuaciones para la mejora de actividades de **ocio** Actuaciones en promoción de aspectos **culturales** Actuaciones para la mejora de la cobertura de servicios de **telecomunicación** … ¿Número de **habitantes** en el territorio?

Variables dependientes	Variables independientes
Telecomunicaciones* (telefonía y banda ancha, servicios fijos y móviles) [**Disponibilidad** o **distancia** al punto de prestación]	¿Medida de tendencia central de la **edad**?

FUENTES

1. Dataset: Datos de población.

   Desagregación: municipal.

   Fuente: www.ine.es

2. Dataset: Establecimientos turísticos

   Desagregación: provincial

   Fuente: www.ine.es

3. Dataset: Empresas

   Desagregación: municipal

   Fuente: www.ine.es (explotación datos Directorio Central Empresas -DIRCE-)

4. Dataset: Distribución geográfica de oficinas bancarias

   Desagregación: municipal

   Fuente: www.bde.es

5. Dataset: Registro centros, servicios y establecimientos sanitarios

   Desagregación: municipal

   Fuente: http://regcess.mscbs.es/regcessWeb/inicioBuscarCentrosAction.do

6. Dataset: Paro registrado

   Desagregación: municipal.

   Fuente: https://sepe.es/HomeSepe/que-es-el-sepe/estadisticas/datos-estadisticos.html

7. Dataset: Actuaciones AAPP (subvenciones)

   Desagregación: municipal

   Fuente: https://www.infosubvenciones.es

8. Dataset: Bienes inmuebles por usos (http://www.catastro.minhap.gob.es/esp/estadistica_1.asp)

   Desagregación: municipal

   Fuente: https://www.catastro.meh.es/esp/estadisticas.asp?var=menuleft3

9. Dataset: Cobertura banda ancha fija y móvil (2013-2020)

   Desagregación: Entidad Singular de Población (ESP)

   Fuente: https://avancedigital.mineco.gob.es/banda-ancha/cobertura/paginas/informacion-cobertura.aspx

10. Dataset: CULTURABase

    Desagregación: provincial

    Fuente: https://www.culturaydeporte.gob.es/servicios-al-ciudadano/estadisticas/cultura/mc/culturabase/portada.html

11. Dataset: Registro Estatal de Centros Docentes no Universitarios (RCD)

    Desagregación: municipal

    Fuente: https://www.educacion.gob.es/centros/selectaut.do

12. Dataset: Agenda Urbana Española (varios)

    Desagregación: municipal

    Fuente: https://www.aue.gob.es/

13. Dataset: Urban Data Platform Plus

    Desagregación: varios niveles

    Fuente: https://urban.jrc.ec.europa.eu/ y

    https://urban.jrc.ec.europa.eu/rel2018/#/en/download

14. Dataset: LUISA Territorial Modelling Platform

    Desagregación: varios niveles

    Fuente: https://joint-research-centre.ec.europa.eu/luisa_en

Catálogo Nacional de Datos Abiertos: https://datos.gob.es/es/

ANEXO II: MARCO LEGISLATIVO Y NORMATIVO

- **Datos Abiertos y Reutilización de la información:**

  - Directiva UE 2019/1024, del Parlamento Europeo y del Consejo de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público.

  - Reglamento (UE) 2018/1807 del Parlamento Europeo y del Consejo de 14 de noviembre de 2018 relativo a un marco para la libre circulación de datos no personales en la Unión Europea

  - Reglamento de Ejecución (UE) 2023/138 de la Comisión de 21 de diciembre de 2022 por el que se establecen una lista de conjuntos de datos específicos de alto valor y modalidades de publicación y reutilización.

  - Real Decreto-ley 24/2021, de 2 de noviembre, transpuso de urgencia varias directivas de la Unión Europea, entre la que se encuentra la Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo, de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la información del sector público.

  - Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público

  - Real Decreto 1495/2011, de 24 de octubre, por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal.

- **Transparencia:**

  - Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno.

- **Protección de Datos:**

  - Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)

- Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales