MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS / MASTER IN DATA SCIENCE

FACULTAD DE CIENCIAS.

Curso 2024-2025

ASIGNATURA: CICLO DE VIDA DE LOS DATOS:

PROPUESTA DE PROYECTO

ADAPTACIÓN DE CULTIVOS CONDICIONADO POR LA PRECIPITACIÓN

PRESENTADO AL PROFESOR:

FERNANDO AGUILAR

EQUIPO BLANCO

INTEGRANTES:

ACEVEDO GARCIA, ADRIAN

COMBITA NIÑO, JOHANA PATRICIA

GONZALEZ PEREZ, EDUARDO

NASTA, MIHAI CRISTIAN

SAINZ-AJA GUERRA, JOSE ADOLFO

VILLA VIARD, OSCAR SANTIAGO

Fecha de entrega: 07/02/2025

RESUMEN

El cambio climático representa un desafío crítico para el sector primario en España, con impactos significativos en la producción agrícola debido a las alteraciones en los patrones de precipitación y la mayor frecuencia de eventos climáticos extremos. Este proyecto tiene como objetivo analizar la relación entre las precipitaciones históricas y la producción agrícola en el territorio español, con el fin de diseñar estrategias de adaptación que permitan garantizar la sostenibilidad del sector.

Para ello, se emplearán datos históricos desde 1992, evaluando la influencia de las precipitaciones en la productividad de cultivos clave y proyectando posibles escenarios futuros bajo distintos modelos climáticos. A través del desarrollo de modelos predictivos y visualizaciones informativas, se identificarán regiones vulnerables y se propondrán medidas de adaptación específicas, como la optimización del uso del agua y la selección de cultivos más resilientes.

Los resultados esperados incluyen un informe detallado sobre la relación entre precipitaciones y producción agrícola, la elaboración de mapas de vulnerabilidad climática, el desarrollo de modelos predictivos, recomendaciones prácticas para agricultores y formuladores de políticas, así como la generación de bases de datos abiertas y publicaciones científicas. Este proyecto contribuirá a la toma de decisiones estratégicas para la mitigación del impacto del cambio climático en la agricultura española, favoreciendo la resiliencia y sostenibilidad del sector primario.

Contenido

1 DESCRIPCIÓN DEL PROBLEMA

El cambio climático representa uno de los desafíos más críticos de nuestra era, con efectos que trascienden el ámbito ambiental y afectan directamente a sectores económicos y sociales fundamentales. Entre ellos, el sector primario, que incluye la agricultura, la ganadería, la pesca y la silvicultura, se posiciona como uno de los más vulnerables. En España, donde este sector tiene una relevancia histórica, económica y cultural significativa, los efectos del cambio climático ya están generando repercusiones directas y preocupantes.

El aumento de la temperatura global, las alteraciones en los patrones de precipitación, la mayor frecuencia de eventos climáticos extremos y el incremento de la desertificación son algunas de las manifestaciones más evidentes del cambio climático. Según informes internacionales, como los del IPCC, estas transformaciones están intensificándose y exigen respuestas inmediatas para mitigar su impacto.

El sector primario desempeña un papel crucial en la economía española, representando una fuente significativa de empleo en muchas regiones rurales y una contribución clave al PIB. Además, es fundamental para garantizar la seguridad alimentaria del país y constituye una pieza esencial en la exportación de productos como el vino, el aceite de oliva, las frutas y hortalizas, entre otros. Por otra parte, el sector primario también mantiene un estrecho vínculo con el patrimonio cultural y la identidad de muchas regiones de España.

El cambio climático está afectando al sector primario de múltiples maneras. Las alteraciones en las precipitaciones han modificado los ciclos agrícolas y las disponibilidades hídricas, lo que afecta tanto a la cantidad como a la calidad de las cosechas. La creciente incidencia de olas de calor, sequías prolongadas e inundaciones extremas está comprometiendo la producción de cultivos tradicionales, como los viñedos y los olivos, pilares de la economía agrícola española. En la ganadería, el estrés térmico y la reducción de pastos están mermando la productividad, mientras que, en la pesca, el calentamiento de las aguas está desplazando especies y alterando ecosistemas.

En España, las particularidades climáticas y geográficas exacerban los efectos del cambio climático. Regiones como el sureste peninsular ya enfrentan un grave problema de escasez de agua, mientras que otras áreas tradicionalmente productivas, como Castilla-La Mancha o Andalucía, ven comprometida su viabilidad agrícola. El impacto en el sector primario no solo amenaza la economía rural, sino que también tiene implicaciones para el desarrollo sostenible, la cohesión territorial y la lucha contra la despoblación.

Frente a este panorama, es urgente identificar estrategias adaptativas que permitan al sector primario español mantenerse competitivo y sostenible. Este proyecto, centrado en analizar la relación entre las precipitaciones y la producción agrícola,

busca generar conocimiento clave para anticiparse a los desafíos del cambio climático y proponer soluciones que garanticen la supervivencia de este sector vital.

El proyecto se centrará en analizar la relación entre las precipitaciones y la producción agrícola en España, cubriendo la totalidad del territorio nacional. Esta cobertura geográfica permitirá identificar variaciones regionales significativas en la relación entre las precipitaciones y la productividad, fundamentales para diseñar estrategias adaptativas específicas.

En cuanto a la cobertura temporal, se utilizarán datos históricos que abarcan desde 1992. La extensión temporal se limitará por la calidad y la consistencia de los datos disponibles, priorizando aquellas series temporales que ofrezcan suficiente continuidad y detalle para un análisis fiable.

La elección de la cobertura geográfica y temporal ha estado condicionada por la disponibilidad de datos. En particular, se priorizarán las regiones y cultivos para los que se cuente con información más completa y precisa tanto de precipitaciones como de producción agrícola. Sin embargo, estas limitaciones también representan una oportunidad para identificar vacíos de datos que podrían ser abordados en futuros estudios.

2 OBJETIVOS

2.1 Objetivo General

Diseñar una estrategia de adaptación de cultivos en el sector agrícola mediante un estudio de las relaciones entre precipitaciones históricas y la producción para focalizar acciones que permitan una adaptación y supervivencia de los principales cultivos en España.

2.2 Objetivos específicos

Para alcanzar el objetivo principal, se definen una serie de objetivos específicos:

- Analizar las relaciones entre precipitaciones históricas y la producción agrícola de cultivos clave en España, identificando patrones relevantes y variaciones geográficas.
- Evaluar el impacto de la variación de la precipitación en la productividad agrícola y estimar cambios futuros bajo diferentes escenarios de cambio climático a partir de proyecciones multi-modelo.
- Identificar áreas geográficas prioritarias para la implementación de medidas de adaptación y mitigación frente al cambio climático basado en la correlación entre precipitaciones y producción agrícola.
- Definir recomendaciones para agricultores, investigadores y formuladores de políticas públicas, enfocada en prácticas agrícolas resilientes y sostenibles.

3 RESULTADOS ESPERADOS

Los resultados esperados de este proyecto se han definido en nueve puntos clave, cada uno orientado a generar conocimiento práctico, herramientas útiles y estrategias de adaptación que contribuyan a la sostenibilidad del sector primario en el contexto del cambio climático.

- ✓ **Informe detallado sobre la relación entre precipitaciones y productividad agrícola**

Elaboración de un informe exhaustivo que analice cómo los niveles de precipitación afectan la productividad agrícola, desglosado por región y tipo de cultivo, identificando patrones clave e implicaciones prácticas.

- ✓ **Visualizaciones informativas**

Se desarrollarán herramientas visuales intuitivas, como mapas, gráficos de tendencia y diagramas de correlación, para representar la relación entre precipitaciones y producción agrícola de manera clara y accesible. Como parte de este esquema, se elaborará un mapa de vulnerabilidad agrícola que identifique las regiones más sensibles a las variaciones climáticas, facilitando la detección de áreas prioritarias para la implementación de estrategias de adaptación y mitigación.

- ✓ **Modelo predictivo para estimar la producción agrícola**

Se diseñará un modelo predictivo capaz de estimar la producción agrícola en función de escenarios climáticos históricos y proyectados, incorporando variables clave para un análisis preciso y fundamentado. A partir de este modelo, se generarán simulaciones que integren tendencias del cambio climático, permitiendo proyectar con mayor certeza su impacto en la producción agrícola y facilitando la planificación de estrategias de adaptación a corto y largo plazo.

- ✓ **Recomendaciones prácticas para la gestión de cultivos**

Propuestas concretas para optimizar la gestión de cultivos en función de los escenarios de precipitación, incluyendo estrategias adaptativas como riego eficiente, selección de cultivos resilientes y planificación regional.

Guía para formuladores de políticas y transferencia de conocimiento

Desarrollo de un documento dirigido a formuladores de políticas con recomendaciones específicas sobre sostenibilidad agrícola, complementado con actividades de difusión como talleres y foros para transferir los hallazgos a agricultores, investigadores y tomadores de decisiones.

- ✓ **Publicaciones científicas y técnicas**

Publicación de artículos que documenten los hallazgos del proyecto, contribuyendo al conocimiento científico en sostenibilidad agrícola y adaptación al cambio climático.

- ✓ **Base de datos enriquecida y abierta**

Creación de una base de datos consolidada que combine datos históricos de precipitaciones y producción agrícola, disponible para análisis futuros y accesible a la comunidad científica y técnica.

4 REQUISITOS Y REQUERIMIENTOS TÉCNICOS

Para llevar a cabo el presente proyecto de análisis de la relación entre precipitaciones históricas y la producción agrícola en España, se han identificado los siguientes requisitos y requerimientos técnicos, agrupados en función de los aspectos clave del desarrollo:

- **Datos**

Se requiere acceso a bases de datos históricas que incluyan registros detallados de precipitaciones en España, desglosadas por región y año, con una resolución espacial y temporal adecuada para el análisis. Asimismo, es indispensable disponer de registros históricos de producción agrícola por tipo de cultivo (viñedos, olivos, arroz, centeno, entre otros) y región. Adicionalmente, sería deseable contar con datos sobre otros factores climáticos relevantes, como temperatura media, humedad relativa y eventos extremos (sequías o inundaciones), así como datos socioeconómicos relacionados con la actividad agrícola, en caso de que puedan aportar valor al análisis.

- **Herramientas y Software**

El análisis de datos y el desarrollo de modelos predictivos se llevará a cabo utilizando herramientas como Python, con librerías especializadas (Pandas, NumPy, Scikit-learn, TensorFlow) o R, según sea necesario. Para la generación de visualizaciones claras y efectivas, se emplearán herramientas como Tableau, Power BI, o librerías de visualización en Python (Matplotlib, Seaborn, Plotly). Además, se utilizarán sistemas de información geográfica (SIG) como QGIS o ArcGIS para la creación de mapas temáticos. La gestión y almacenamiento de datos se realizará mediante bases de datos SQL o NoSQL, dependiendo del volumen y la estructura de los datos disponibles.

- **Recursos Humanos**

El equipo multidisciplinar necesario para el proyecto estará compuesto por:

- Especialistas en análisis de datos y machine learning para el desarrollo de los modelos predictivos y el análisis estadístico.
- Un ingeniero agrónomo o experto en agricultura que aporte conocimiento específico para la interpretación de los datos y validación de resultados.
- Un experto en cambio climático que contextualice los resultados obtenidos en términos de escenarios futuros.
- Un diseñador gráfico o especialista en visualización para la elaboración de mapas, gráficos y otras herramientas de comunicación visual.
- Técnico GIS para la gestión de bases de datos alfanuméricas y de cartografía digital, mediante programas diseñados para la implantación de sistemas de información geográfica.

- Project Management para la planificación, coordinación, monitoreo y control del proyecto.

➢ **Recursos Materiales**

Para garantizar un desarrollo fluido del proyecto, se necesitará una infraestructura informática adecuada que permita procesar los datos y desarrollar los modelos predictivos. Esto incluye ordenadores con capacidad suficiente o acceso a servicios de computación en la nube como AWS, Google Cloud o Azure. También será necesario disponer de acceso a publicaciones científicas y bases de datos académicas para obtener referencias y estudios previos relacionados.

➢ **Requisitos Organizativos**

El proyecto requerirá la definición de un cronograma claro que incluya hitos específicos y entregables a lo largo del desarrollo. Además, será crucial asignar roles y responsabilidades dentro del equipo de trabajo para garantizar una ejecución eficiente. Finalmente, se debe contar con un presupuesto detallado que contemple los costos asociados al acceso a datos, licencias de software, servicios en la nube y contratación de personal especializado.

5 DESCRIPCIÓN DE LAS FUENTES DE DATOS

Para el desarrollo del proyecto, se emplearán tres fuentes principales de datos: datos de precipitación acumulada histórica, proyecciones de cambio climático, y datos de producción de cultivos. A continuación, se describen en detalle cada una de estas fuentes, su procedencia, el formato en el que se encuentran disponibles y su relevancia para el estudio.

5.1 Datos de Precipitación Acumulada Histórica

Los datos de precipitación histórica se obtendrán de la base de datos Reanálisis ERA5, un conjunto de datos climáticos producido por el European Centre for Medium-Range Weather Forecasts (ECMWF) en el marco del Servicio de Cambio Climático de Copernicus (C3S).

ERA5 es una base de datos de reanálisis, lo que significa que combina observaciones meteorológicas históricas con modelos numéricos atmosféricos para generar una reconstrucción detallada del clima pasado. Este proceso permite corregir inconsistencias en los registros históricos y proporcionar información homogénea en el tiempo y el espacio.

Características de los datos:

- Periodo de cobertura: desde 1940 hasta la actualidad.
- Resolución temporal: desde datos horarios hasta agregados mensuales.
- Resolución espacial: 0.25° x 0.25° con cobertura global.
- Formato de descarga: archivos en formato GRIB (GRIdded Binary), un estándar utilizado para el almacenamiento de datos meteorológicos en mallas de grilla.
- Fuente de descarga: disponible a través de la plataforma Climate Data Store (CDS) de Copernicus, accesible mediante interfaz web y API.

5.2 Datos de Proyecciones de Cambio Climático

Las proyecciones de cambio climático se obtendrán del Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al., 2016), una iniciativa internacional coordinada por el World Climate Research Programme (WCRP).

CMIP6 es un esfuerzo colaborativo de la comunidad científica para mejorar la modelización del clima global. A través de múltiples modelos climáticos desarrollados por distintas instituciones, se generan simulaciones del clima futuro bajo distintos escenarios de emisiones de gases de efecto invernadero.

Características de los datos:

- Escenarios climáticos considerados:
  Histórico: utilizado como referencia.

SSP2-4.5: un escenario de emisiones intermedio.
SSP5-8.5: un escenario de altas emisiones y calentamiento extremo.

- Resolución temporal: mensual.
- Formatos de descarga: NetCDF (Network Common Data Form), un formato optimizado para el almacenamiento y manejo de grandes volúmenes de datos multidimensionales.
- Fuente de descarga: disponible a través de las plataformas oficiales de CMIP6 mediante API.

Una vez descargados los datos de todos los modelos disponibles, se realizará una evaluación de su calidad y coherencia. A partir de esta evaluación, se seleccionará un subconjunto representativo para el análisis.

5.3 Datos de Producción de Cultivos

Los datos de producción agrícola en España se obtendrán del Anuario de Estadística del Ministerio de Agricultura, Pesca y Alimentación (MAPA, (Escudero Población et al., 2024). Esta base de datos consiste en una recopilación anual de información estadística sobre distintos aspectos del sector agrícola en España, incluyendo datos de producción, superficies cultivadas, rendimientos y otros indicadores relevantes.

Características de los datos:

- Periodo de cobertura: desde 1999 hasta la actualidad.
- Nivel de desagregación: información disponible a nivel de comunidad autónoma y nacional.
- Formatos de descarga: archivos en CSV (Comma-Separated Values) y XLSX (Excel Spreadsheet).
- Fuente de descarga: accesible a través de la plataforma oficial del MAPA.

Estos datos permitirán evaluar la relación entre la producción agrícola y las condiciones climáticas, proporcionando un contexto fundamental para la interpretación de los resultados del análisis.

6 PLANIFCIACIÓN DEL PROYECT

6.1 Work Breakdown Structure (WBS)

Para el desarrollo del proyecto y alcanzar los objetivos propuestos, se han definido las siguientes fases y paquetes de trabajos, que se representan de manera estructurada en el WBS.

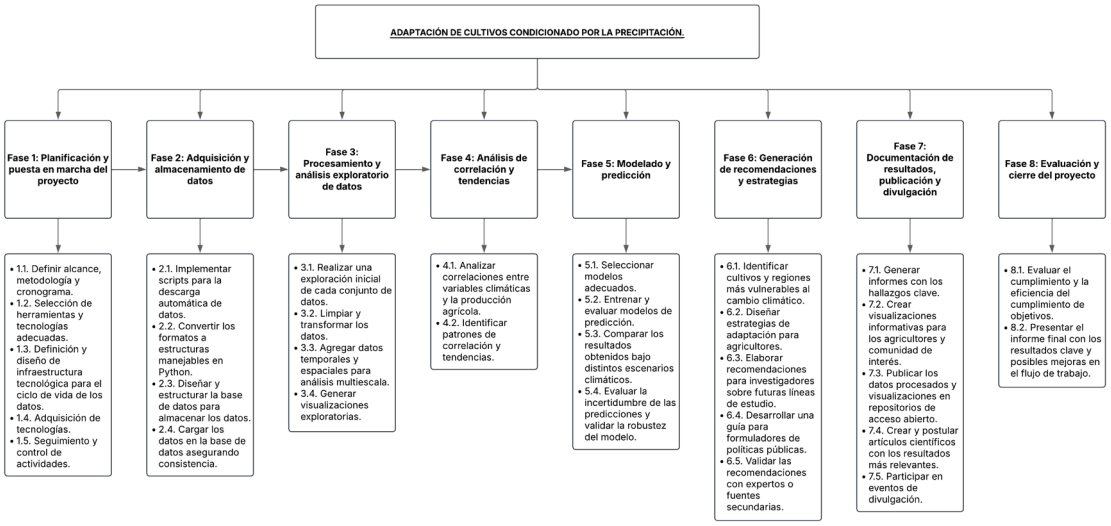

Imagen 1. Work Breakdown Structure por fases.

Fase 1. Planificación y puesta en marcha del proyecto

Esta fase tiene como objetivo concretar el alcance, metodología y herramientas a utilizar en el proyecto. Partiendo del project charter se planifica el proyecto con mayor detalle, implementando herramientas de seguimiento como Project y tableros Kanban para control de actividades.

Paquetes de trabajo:

1.1. Definir alcance, metodología y cronograma.

1.2. Selección de herramientas y tecnologías adecuadas.

1.3. Definición y diseño de infraestructura tecnológica para el ciclo de vida de los datos.

1.4. Adquisición de tecnologías.

1.5. Seguimiento y control de actividades

Fase 2. Adquisición y almacenamiento de datos.

El objetivo de esta fase es automatizar la descarga de datos y almacenarlos en una base de datos. La base de este proyecto radica en la recopilación de datos meteorológicos provenientes del programa Copernicus y datos agrícolas del

Ministerio de Agricultura, Pesca y Alimentación (MAPA). Estos datos serán obtenidos mediante APIs y procesados utilizando herramientas como Python.

Paquetes de trabajo:

2.1. Implementar scripts para la descarga automática de datos desde:

- Climate Data Store (CDS) de Copernicus (ERA5, formato GRIB).

- CMIP6 (proyecciones climáticas, formato NetCDF).

- Anuario del MAPA (producción de cultivos, formato CSV/XLSX).

2.2. Convertir los formatos a estructuras manejables en Python.

2.3. Diseñar y estructurar la base de datos para almacenar los datos.

2.4. Cargar los datos en la base de datos asegurando consistencia.

Fase 3. Procesamiento y análisis exploratorio de datos

Los datos recolectados deben ser sometidos a un proceso de curación para garantizar su calidad y coherencia. Se llevarán a cabo tareas de normalización, detección y tratamiento de valores faltantes, y validación de integridad de los registros. Estas actividades serán realizadas principalmente con Python (pandas, NumPy) y R. Además, un análisis exploratorio de los datos permitirá tener una visión inicial de los datos y su consistencia.

Paquetes de trabajo:

3.1. Realizar una exploración inicial de cada conjunto de datos.

3.2. Limpiar y transformar los datos (manejo de valores nulos y atípicos, conversión de unidades y normalización de variables).

3.3. Agregar datos temporales y espaciales para análisis multiescala.

3.4. Generar visualizaciones exploratorias.

Fase 4. Análisis de correlación y tendencias

Esta fase busca determinar la relación entre variables climáticas y la producción agrícola. A partir de un análisis exploratorio, buscar patrones y correlaciones significativas entre la precipitación y la producción agrícola. Se elaborarán representaciones gráficas como scatter plots e histogramas para visualizar las relaciones entre variables.

Paquetes de trabajo:

4.1. Analizar correlaciones entre variables climáticas y la producción agrícola.

4.2. Identificar patrones de correlación y tendencias.

Fase 5. Modelado y predicción.

Esta fase consiste en aplicar técnicas de modelado para evaluar posibles escenarios futuros. Para esto, se pondrán a prueba modelos estadísticos y redes neuronales, con validación cruzada y análisis de métricas de rendimiento.

Paquetes de trabajo:

5.1. Seleccionar modelos adecuados.

5.2. Entrenar y evaluar modelos de predicción del impacto del clima en la producción agrícola.

5.3. Comparar los resultados obtenidos bajo distintos escenarios climáticos (CMIP6).

5.4. Evaluar la incertidumbre de las predicciones y validar la robustez del modelo.

Fase 6. Generación de recomendaciones y estrategias

Esta fase busca traducir los hallazgos en acciones concretas para distintos actores. A partir de la socialización de los principales resultados con los grupos de interés, construir acciones y estrategias que permitan adaptar diferentes sectores agrícolas acorde a los escenarios de cambio climático, variables metereológicas y condiciones geográficas.

6.1. Identificar cultivos y regiones más vulnerables al cambio climático.

6.2. Diseñar estrategias de adaptación para agricultores (fechas óptimas de siembra según cambios en patrones de lluvia, uso de variedades de cultivos más resistentes, prácticas agrícolas para mejorar la retención de agua y reducir el impacto de sequías)

6.3. Elaborar recomendaciones para investigadores sobre futuras líneas de estudio.

6.4. Desarrollar una guía para formuladores de políticas públicas (medidas de apoyo a la agricultura sostenible, políticas de incentivos para cultivos resilientes, estrategias de gestión del agua en zonas agrícolas afectadas)

6.5. Validar las recomendaciones con expertos o fuentes secundarias.

Fase 7. Documentación de resultados, publicación y divulgación.

Esta fase busca documentar y compartir los datos, análisis y modelos generados, a partir de diferentes medios como visualizaciones informativas (mapas, gráficos, mapa de vulnerabilidad, entre otros), artículos científicos, y presentación de resultados.

Paquetes de trabajo:

7.1. Generar informes con los hallazgos clave.

7.2. Crear visualizaciones informativas para los agricultores y comunidad de interés.

7.3. Publicar los datos procesados y visualizaciones en repositorios de acceso abierto.

7.4. Crear y postular artículos científicos con los resultados más relevantes.

7.5. Participar en eventos de divulgación.

Fase 8. Evaluación y cierre del proyecto

Esta fase busca validar el cumplimiento de objetivos y extraer aprendizajes para mejoras futuras.

Paquetes de trabajo:

8.1. Evaluar el cumplimiento y la eficiencia del cumplimiento de objetivos.

8.2. Presentar el informe final con los resultados clave y posibles mejoras en el flujo de trabajo.

6.2 Diagrama de Gantt

Se adjunta documento Excel con la información del cronograma por fases y actividades, donde es posible identificar la duración del proyecto, la consecución de actividades y precedencias.

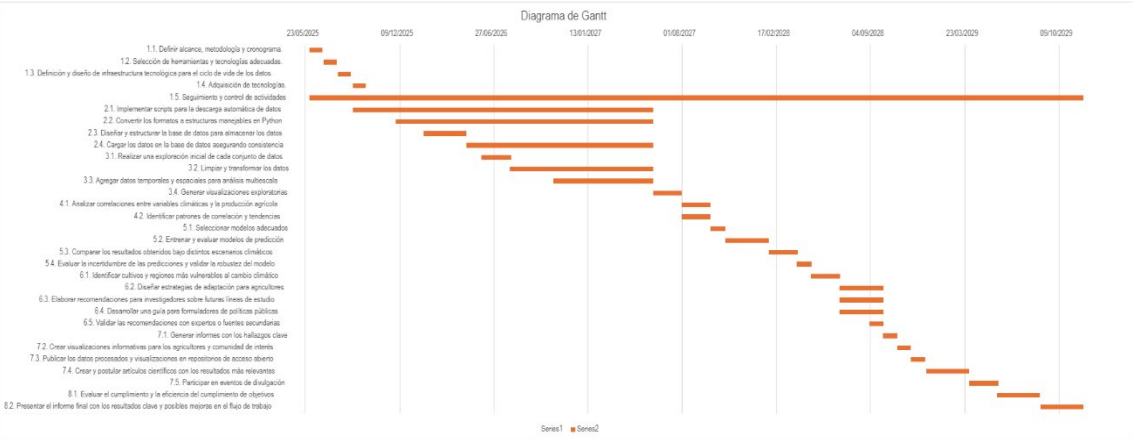

Imagen 2. Diagrama de Gantt.

6.3 Hitos y entregas

En la siguiente tabla se comparten los hitos del proyecto con la fecha en los cuales se espera obtener los avances y entregables principales.

Hito	**Fecha**
Implementación de la base de datos con datos brutos	01/06/2027
Datos limpios y listos para análisis	01/08/2027
Resultados de relación entre precipitaciones y producción agrícola	01/10/2027
Modelos de predicción entrenados y evaluados	01/05/2028
Estrategias de adaptación y recomendaciones definidas	01/10/2028
Presentación de informe final	01/12/2029

Tabla 1. Hitos del proyecto.

6.4 Presupuesto

Se adjunta documento Excel con la información del presupuesto por fases y actividades, donde es posible identificar los recursos a implementar (humanos y tecnológicos) con su presupuesto.

7 PLAN DE PRESERVACIÓN

El plan de preservación de datos busca garantizar la integridad, disponibilidad y accesibilidad de los datos generados y utilizados durante la ejecución del proyecto, Además, dado el alcance del proyecto para consulta y acciones futuras, se considera un plan que permita reutilizar y referenciar los datos, ya sea por investigadores, agricultores, entes públicos, y personal interesado, garantizando un enfoque robusto para la sostenibilidad de la información. A continuación, se presentan los aspectos clave de almacenamiento físico, almacenamiento de datos, gestión de datos, respaldos y almacenamiento en la nube.

7.1 Almacenamiento Físico.

En este proyecto, la cantidad de datos climáticos y de producción agrícola, así como las necesidades de procesamiento, requieren soluciones robustas de almacenamiento físico:

Dispositivos SSD con interfaz NVMe: Estos dispositivos se utilizarán para el almacenamiento temporal de los datos en análisis intensivos debido a su alta velocidad de lectura/escritura. Esto es crucial para manejar modelos predictivos complejos y grandes volúmenes de datos climáticos. Los NVMe ofrecen tasas de transferencia de hasta 5000 MB/s, optimizando el tiempo de procesamiento durante los cálculos del modelo.

Discos HDD con interfaces SAS o SATA III: Se emplearán para almacenar copias locales de datos que no necesiten acceso inmediato. Estos discos son adecuados para datos menos críticos, como backups históricos del proyecto, permitiendo una solución económica para preservar datos importantes a largo plazo, combinando capacidad con costo.

RAID 10 y Dynamic Disk Pooling (DDP): Estas configuraciones aseguran redundancia y protección contra fallos de hardware, manteniendo la disponibilidad de los datos críticos. RAID 10 se utilizará en servidores principales para análisis activos, mientras que DDP optimizará la recuperación de datos en casos de fallo.

Protección física: Mantener los dispositivos de almacenamiento en ambientes controlados en cuanto a temperatura y humedad. Además, tendrá una protección contra accesos no autorizados mediante mecanismos de bloqueo físico y seguridad perimetral.

7.2 Almacenamiento de Datos

El almacenamiento efectivo de los datos recolectados y procesados es esencial para su gestión y análisis continuo, por lo cual se plantea lo siguiente:

Sistemas de archivos avanzados (Ext4 y Ceph):

- **Ext4:** Este sistema será implementado en los discos físicos locales por su alta compatibilidad y rendimiento.
- **Ceph:** Se usará para almacenamiento unificado en entornos distribuidos. Ceph permite manejar almacenamiento en bloque, archivo y objeto, adaptándose a los distintos tipos de datos del proyecto.Es ideal para un proyecto que maneja múltiples fuentes de datos y necesita escalabilidad a largo plazo.

Virtualización de almacenamiento mediante LVM (Logical Volume Manager): LVM permitirá dividir y gestionar particiones de discos físicos de manera flexible, facilitando el ajuste a necesidades cambiantes durante el desarrollo del proyecto. Proporciona una forma eficiente de reorganizar el almacenamiento si los volúmenes de datos crecen más de lo esperado.

Organización de los datos: Implementar una estructura de carpetas jerárquica clara con convenciones de nombres consistentes. Además, establecer una política de control de versiones para registrar cambios en los datos y evitar duplicados innecesarios.

7.3 Gestión de Datos

La gestión adecuada de los datos incluye su seguridad y acceso, por lo cual se determina los siguientes controles:

- **Control de acceso basado en roles (RBAC):** Los datos serán protegidos mediante permisos detallados a nivel de archivo, asegurando que solo usuarios autorizados puedan acceder o modificar información crítica. De esta forma se protege la confidencialidad de datos sensibles relacionados con el impacto climático en cultivos específicos. Utilizar autenticación de dos factores (2FA) para mayor seguridad.

7.4 Respaldo de datos (Backup)

Como en todo proyecto, es importante manejar un sistema de backup que asegure la continuidad del proyecto y su sostenibilidad, por lo cual se plantea las siguientes medidas:

- **Política de respaldos:** Se realizarán respaldos diarios incrementales y semanales completos utilizando herramientas como **Veeam Backup** o **Bacula**, y se almacenarán tanto localmente como en la nube. De esta manera, se garantiza la recuperación de datos en caso de fallos técnicos, accidentes o ataques cibernéticos.
- **Uso de cintas LTO (Linear Tape Open):** Las cintas LTO serán el medio principal para el almacenamiento a largo plazo de datos históricos una vez finalizado el proyecto. Estas ofrecen alta capacidad de almacenamiento

(hasta 12 TB por cinta) con un costo por GB competitivo, asegurando la preservación de datos durante décadas.

- **Redundancia de los respaldos:** Mantener al menos tres copias de los datos: una en almacenamiento local, una en un dispositivo externo y otra en la nube. Además, comprobar regularmente la integridad de las copias de seguridad para garantizar que los datos puedan ser recuperados correctamente.

7.5 Almacenamiento en la Nube

El almacenamiento en la nube complementará las soluciones físicas, proporcionando flexibilidad y redundancia, por lo cual se plantea lo siguiente:

Plataformas como AWS S3 y Google Cloud Storage: Estas plataformas permitirán almacenar datos de forma distribuida y accesible desde cualquier ubicación. Además, facilitan la colaboración entre los miembros del equipo, especialmente en el intercambio de grandes volúmenes de datos climáticos y agrícolas. Importante destacar que permiten cifrado de datos y escalabilidad para manejar aumentos inesperados en el volumen de datos.

Por otra parte, Amazon S3 u otros servicios similares permitirán asociar metadatos extensos a los archivos, mejorando la organización y recuperación de información.

7.6 Estrategia de Preservación a Largo Plazo

El proyecto necesita asegurar la sostenibilidad de los datos más allá de su finalización:

- **Migración periódica:** Los datos serán revisados y migrados a formatos actualizados para evitar la obsolescencia tecnológica.
- **Publicación en repositorios abiertos:** Los conjuntos de datos se subirán a plataformas como **Zenodo** o **Figshare**, haciendo los datos accesibles para investigadores y agricultores interesados. De esta manera, se facilita la transferencia de conocimiento y fomenta el uso continuo de los datos para estrategias agrícolas adaptativas.

8 DESARROLLO DE ANÁLISIS DE DATOS PRELIMINAR

8.1 Curar/Limpiar datos

Para el presente estudio fue necesario acceder a la información climática de derivada de ERA5 (Hersbach et al., 2020) y titulada Essential climate variables for assessment of climate variability from 1979 to present (Hersbach et al., 2018). Este conjunto de datos ya presenta la información curada y limpiada para su uso, por la cual se accede a través de una API para conectar con la información necesaria para el estudio.

Es importante destacar que a este conjunto de datos ya está curado y se le ha aplicado un proceso previo de curación y limpieza de datos, debido a que, como todo proceso de recolección o generación de datos, es normal encontrar faltantes, valores atípicos, inconsistencias y heterogeneidad de los datos.

Este proceso de curado que se le ha aplicado a los datos, comienza desde antes de la ejecución del modelo que se utiliza en la generación de los datos de ERA5. ERA5 es un reanálisis atmosférico, esto quiere decir que asimila datos instrumentales de todo tipo que proceden de dispositivos electrónicos ubicados en estaciones meteorológicas, globos meteorológicos, mediciones de satélite e incluso mediciones visuales. En todo este proceso puede haber fallos en la obtención de los registros o en las mediciones, huecos recurrentes, sesgos asociados a los aparatos de medidas, entre otros. Por lo que algunos de los pasos clave realizados de forma previa a la asimilación de los datos instrumentales serían los siguientes:

1. Validación de integridad y consistencia de los datos: Identificación de registros incompletos, comparando series temporales para detectar discontinuidades. Además, la identificación de registros duplicados que generen redundancia y sesgo. Los registros incompletos y duplicados, dependiendo de su relevancia, son eliminados o tratados con técnicas específicas.
2. Normalización de formatos: Dado que existen múltiples estaciones meteorológicas y fuentes de datos, es necesario realizar un chequeo y unificación de unidades para facilitar la comparabilidad. De igual manera, se estandarizan los formatos de los registros temporales (hora y fecha) para asegurar un manejo uniforme de la información.
3. Detección y manejo de valores atípicos: implementación de métodos estadísticos como la desviación estándar para identificar los valores extremos. Posteriormente, evaluar su naturaleza consultando registros históricos y expertos en la materia. Los valores atípicos identificados como errores son corregidos o eliminados, mientras que los fenómenos extremos reales se conservan como casos de interés.
4. Eliminación de datos irrelevantes: Filtrado de registros fuera de áreas geográficas de interés, y la eliminación de columnas que no aportan valor.

Para ello se habrán empleado herramientas específicas para el manejo de esta información, su limpieza y procesamiento, todo esto puede haberse hecho con herramientas como Python, empleando las librerías de Pandas y Numpy para la gestión de datos masivos y técnicas de limpieza, y Xarray para trabajar con datos multidimensionales de manera eficiente.

Tras esto se pasa a la fase de modelado, en el cual mientras se ejecuta el modelo, se van asimilando las diferentes fuentes de datos instrumentales, tras la fase de modelado viene una segunda fase de curado de datos que puede consistir en los siguientes puntos:

- Revisión de la continuidad de los datos, realizando un chequeo de los resultados consistente en la búsqueda de huecos en las diferentes variables, inhomogeneidades o inestabilidades que puedan haber sucedido durante la ejecución.
- Comparativa entre variables y verificación de la coherencia de los resultados.
- Validación de los resultados y estimación de su calidad mediante la comparativa con conjuntos de datos instrumentales que preferiblemente no hayan sido utilizados en el proceso de asimilación de datos y que además sean adecuados para ello.
- Aplicación de filtros para detectar valores anómalos, una vez detectados decidir qué acciones tomar, eliminar, modificar o relanzar la fase de modelado en dicho periodo.
- Postprocesado de toda la información, transformando los resultados a un formato manejable y que pueda ser fácilmente consultados o extraídos por otras personas. Esto incluye el postprocesado a diferentes escalas temporales como datos diarios o mensuales.

Tras todo lo anterior, ya estarían los conjuntos de datos tal como se han extraído y han sido utilizados en el estudio.

Cabe destacar que estos últimos pasos también se han tenido que realizar para los datos de proyecciones de precipitación que tienen en cuenta los cambios debidos al cambio climático, en estas simulaciones no se asimilan datos instrumentales a lo largo de la ejecución, pero una vez que se generan los resultados estos también tienen que ser curados de cierta manera, para luego ser puestos a disposición del resto de usuarios dentro de la plataforma de descarga de datos del proyecto CMIP6.

Por otro lado, en el estudio se trabajó con la base de datos de Estadística digital del Ministerio de Agricultura, Pesca y Alimentación (Segunda fuente). De igual forma, en este repositorio de datos la información se encuentra curada y lista para análisis, pero a diferencia de la primera fuente esta presenta un sistema de visualización preliminar y permite la descarga de la información a través de un formato CSV.

En el caso de esta segunda fuente, la recolección de la información iimplica un proceso de centralización de los datos obtenidos de cada una de las comunidades de España, relacionando año, cantidad producida y las diferentes clasificaciones de los productos, por lo cual también es necesario una serie de pasos adicionales para la consolidación de estos datos:

1. Validación de estructura de datos y formato: Validación de que las columnas y variables sigan una estructura. Adicionalmente, la corrección de discrepancias en los nombres de cultivos o las comunidades. Finalmente, la unificación de unidades y la optimización en el tipo de dato (categóricas, numéricas, etc.).
2. Detección de valores atípicos y faltantes: la información preliminar se obtiene de los resultados de producción mensual, donde es normal encontrar estacionalidad, faltantes y otros comportamientos que se deben analizar para identificar si se debe a la naturaleza del producto, son registros incompletos o se presentan valores atípicos. Es importante contar con expertos y estadísticos para poder hacer ese tipo de análisis y se pueda consolidar la información de producción anual por producto.
3. Filtrado de datos irrelevantes: eliminación de columnas que no brindan información importante, filtrado de áreas que no correspondían a cultivos o zonas agrícolas específicas.

Para el manejo de esta información, su limpieza y procesado de los registros es posible implementar herramientas como OpenRefine para la detección de inconsistencias y corrección de errores de entrada manual, como nombres duplicados, mal escritos, tipo de datos, eliminación de columnas, etc. Por otro lado, la herramienta de Power Query (herramienta integrada en Power BI) permite transformar, limpiar y combinar datos desde diferentes fuentes y archivos para la integración de la información de las diferentes comunidades del país.

Estas técnicas aseguran que la información extraída de ambas fuentes sea precisa, confiable y adecuada para los objetivos del estudio, maximizando la utilidad de los datos en los análisis climáticos y agrícolas.

8.2 Obtención de datos y post-procesado.

8.2.1 Obtención de datos

- Identificación de fuentes fiables de datos meteorológicos (AEMET, NOAA).

Se han buscado fuentes homogéneas, continuas y sin huecos de datos de precipitación media mensual acumulada, en un alcance inicial se han buscado en base a ello diferentes conjuntos de datos de diferentes organismos como la Agencia Estatal de Meteorología (AEMET), el ECMWF o la NOAA. Tras diferentes consultas se han considerado un dataset de datos de precipitación del Climate Data Store (CDS) de Copernicus y generado por el ECMWF que son datos históricos de precipitación media acumulada mensual del reanálisis atmosférico ERA5 (Hersbach et al., 2020).

Un reanálisis atmosférico es un producto generado mediante la combinación de observaciones meteorológicas históricas y modelos numéricos del clima para proporcionar una representación coherente y detallada de las condiciones atmosféricas en el pasado. Es una herramienta fundamental en meteorología y climatología, utilizada para estudiar el comportamiento de la atmósfera a lo largo del tiempo.

Se han utilizado los datos del reanálisis ERA5 por sus ventajas frente a las mediciones in-situ que son las siguientes:

- Proporcionan datos en todas las regiones del mundo, incluyendo áreas remotas donde no hay estaciones in situ, como océanos, desiertos, montañas o regiones polares.
- Utilizan un modelo numérico único y técnicas de asimilación de datos, lo que garantiza una representación consistente del sistema atmosférico a lo largo del tiempo, incluso si las observaciones cambian
- Proveen una amplia gama de variables atmosféricas y climáticas (como viento, temperatura, humedad, radiación, etc.) a nivel global, muchas de las cuales no se miden directamente en estaciones terrestres
- Ofrecen datos en resoluciones espaciales y temporales uniformes (por ejemplo, cada 1 hora y con celdas de 25 km x 25 km), permitiendo análisis detallados y comparables en cualquier parte del mundo.
- Utilizan modelos y técnicas de asimilación de datos para llenar los vacíos donde faltan observaciones, proporcionando una imagen más completa y continua del sistema atmosférico.
- Los modelos numéricos utilizados en los reanálisis consideran procesos físicos y dinámicos de la atmósfera, lo que permite estimar variables no medidas directamente (como flujos de energía o radiación) y simular interacciones entre componentes del sistema climático.

Estas ventajas cubren ciertos aspectos problemáticos de las medidas instrumentales in-situ como que tienen huecos, no cubren largos periodos, pueden contener inhomogeneidades debidas a cambios en el aparato de medida o que solo dan información puntual y no espacial.

Por otro lado, se han considerado dichos datos porque por un lado tienen una API específica para su descarga y por su homogeneidad y cobertura espacial.

Dentro del proyecto, se buscará la ampliación de esta tarea en los siguientes aspectos:

- Búsqueda de conjuntos de datos de precipitación históricos de mejor calidad o de mayor resolución espacial o temporal.
- Extender el periodo de análisis a futuro con la utilización de proyecciones de precipitación a futuro, para estimar su variabilidad debida al cambio climático. Para ello se utilizarían datos generados por diferentes modelos

climáticos y bajo diferentes escenarios futuros dentro del proyecto CMIP6 (Coupled Model Intercomparison Project Phase 6). Estos escenarios futuros serán el ssp245 y el ssp585 que son los más utilizados para las estimaciones de cambios debidos al cambio climático.

La incorporación de dichos conjuntos implicaría ampliar el tiempo requerido en el procesamiento de datos.

- Descarga y almacenamiento de datos meteorológicos.

Para la descarga de los datos de precipitación históricos, se utiliza la API especifica en Python del Climate Data Store (CDS), esta API descarga los datos en un formato grib desde 1979 hasta la actualidad, dichos datos son postprocesado posteriormente para facilitar su uso y su análisis.

En el caso de los datos adicionales de proyecciones se utilizarán las APIs propias que tiene disponible el servicio de descarga de datos del proyecto CMIP6 rin(Eyring et al., 2016) en sus diferentes nodos de descarga de datos.

- Obtención de datos sobre producción de cultivos por comunidad autónoma.

Esta información de la producción de los diferentes cultivos por comunidad autónoma se obtiene a través del Anuario de Estadística del Ministerio de Agricultura, Pesca y Alimentación (MAPA), el acceso se realiza a través del siguiente enlace.

Esta fuente proporciona información detallada sobre la distribución general del suelo, las superficies cultivadas y las producciones anuales de los principales cultivos en España. Dentro de dicho anuario se pueden descargar diferentes estadísticas asociadas a los diferentes cultivos en diferentes años, agrupados a nivel nacional y por comunidad autónoma en formato Excel o csv, entre otros.

En dicho conjunto de datos, clasifica los cultivos en 5 niveles diferentes, de los cuales solo consideramos los dos primeros niveles:

- Nivel1: Distingue entre Leñosos y herbáceos
- Nivel 2: Diferente para cada grupo del nivel 1:
- Leñosos: Bayas, Cítricos, Frutales Carnosos, Frutales de Hueso, Frutales de Pepita, Frutales secos, Olivar, Otros cultivos leñosos y Viñedo.
- Herbáceos: Aprovechamientos, Cereales, Forrajeros, Hortalizas, Industriales, Leguminosas y Tubérculos.

El resto de los niveles no se consideran porque hay variabilidades que hay que pueden no depender directamente de la precipitación, si no de otros condicionantes como puede ser la aplicación de técnicas de rotación de cultivos, cambio de tipo de cereal o cambio de tubérculo que se cultiva, entre otros factores.

8.2.2 Postproceso de la información.

Cada fuente de datos tiene un formato diferente y diferente características por lo que el postproceso es distinto, tenemos

En el caso de los datos históricos de precipitación estos están en formato grib y cubren todo el mundo, por lo que dichos datos son post-procesados y pasados a un formato en el cual permita su fácil manejo y que agrupe los datos por regiones. Para ello se realiza el siguiente post-procesado:

- Convertir los datos del formato grib a un formato NetCDF con el programa CDO (Climate Data Operators)
- Recortar los datos de forma que solo contengan información en el territorio español con Python y las librerías pertinentes.
- Agrupar la información de precipitación a nivel de España y por comunidades autónomas para calcular las medias anuales de precipitación acumulada.
- Guardado de la información en ficheros CSV, adecuado para su integración en herramientas de análisis de datos y sistemas de modelado.

Para los datos de producción agrícola, se han descargado en formato xlsx y se ha visto que dicho formato es operativo a nivel de lectura con las diferentes herramientas de análisis, además de esta forma se evita la perdida de información. Con dicha información el post-procesado se aplica una vez leídos los datos y es sumar agrupar los datos en función de las diferentes categorías de nivel 2, ya que se la producción en toneladas viene asociada al nivel de menor entidad existente. Por ejemplo, para un año en la producción de cereales se ha sumado la producción de Cebada, Trigo, Maíz, Centeno, etc. Esto se ha hecho para todas las categorías de nivel 2 y para cada año.

Para otros conjuntos de datos de precipitación histórica de mejor calidad que se puedan encontrar, en una ampliación del estudio se contempla un post-procesado similar al aplicado al conjunto de datos de precipitación utilizado.

En el caso de datos de proyecciones de precipitación a futuro se tiene en cuenta el siguiente post-procesado:

- Recortar la información de los ficheros NetCDF descargados para que solo contengan la península ibérica y las islas de los archipiélagos canario y balear.
- Agrupar la información de precipitación a nivel de España y por comunidades autónomas para calcular las medias anuales de precipitación acumulada.
- Cálculo de las medias anuales en los diferentes escenarios de cambio climático considerados a futuro dentro de cada uno de los modelos climáticos y también en el periodo histórico de dichos modelos climáticos desde 1979 hasta el último año disponible de dicho periodo histórico.
- Análisis de la calidad de los diferentes modelos climáticos en base al análisis de las climatologías en este periodo pasado.

- Analizar la consideración o no de la aplicación de una corrección de sesgo a cada uno de los modelos climáticos en base a los datos del reanálisis utilizado ERA5.
- Cálculo de los cambios a futuro de la precipitación con respecto al periodo histórico de cada uno de los modelos de cambio climático en base a los diferentes escenarios.

8.3 Creación de metadatos

Tras generar los ficheros que contienen toda la información, se guardarán en un formato fácilmente accesible y se publicarán, en esta publicación se incluirán todos los metadatos pertinentes siguiente Dublin Core.

Como ejemplo en un proceso de exploración ya se han publicado en el repositorio de Zenodo Sandbox los datos de precipitación que se han descargado y postprocesado (Acevedo García et al., 2025).

Los metadatos de los datos generados se incluyen en un fichero .xml, utilizando el formato Dublin Core y adjuntos a la entrega, en la siguiente imagen se muestra el contenido:

```
<?xml version="1.0" encoding="UTF-8"?>
<dublin_core xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
    <dc:title>Annual accumulated precipitation in Spain and CCAA from 1979 to 2024</dc:title>
    <dc:creator>Acevedo García, Adrián</dc:creator>
    <dc:creator>Cómbita Niño, Johana</dc:creator>
    <dc:creator>Pérez González, Eduardo</dc:creator>
    <dc:creator>Nasta, Mihai Cristian</dc:creator>
    <dc:creator>Sainz Guerra, Jose Adolfo</dc:creator>
    <dc:creator>VILLA VIARD, OSCAR SANTIAGO</dc:creator>
    <dc:subject>Spain</dc:subject>
    <dc:subject>Annual</dc:subject>
    <dc:subject>Mean</dc:subject>
    <dc:subject>Atmospheric precipitation</dc:subject>
    <dc:description>Annual mean accumulated precipitation in Spain and CCAA from 1979 to 2024 calculated from ERA5 Reanalysis</dc:description>
    <dc:publisher>Zenodo</dc:publisher>
    <dc:contributor>Instituto de Hidráulica Ambiental de la Universidad de Cantabria "IH Cantabria"</dc:contributor>
    <dc:contributor>University of the Coast</dc:contributor>
    <dc:contributor>Universidad de Cantabria</dc:contributor>
    <dc:date>2025-02-05</dc:date>
    <dc:type>Dataset</dc:type>
    <dc:format>text/csv</dc:format>
    <dc:identifier>https://sandbox.zenodo.org/records/162490</dc:identifier>
    <dc:language>en</dc:language>
    <dc:rights>Creative Commons Attribution 4.0 International</dc:rights>
</dublin_core>
```

Imagen 3. Creación de metadatos.

8.4 Análisis de datos

De los resultados obtenidos en el análisis exploratorio inicial se observa que la categoría cítricos a nivel España presenta la correlación más alta (0.579) con respecto al resto de categorías de productos a analizar, seguidos por viñedo (0.498), cereales (0.485) y olivar (0.433). Las correlaciones positivas indican que un aumento en la precipitación podría estar asociado con un incremento en la producción, sin embargo, el coeficiente no mide causalidad, por lo que habría que profundizar.

En la gráfica de cítricos, se observa una relación moderada con tendencia lineal. Sin embargo, parece haber dispersión en ciertos puntos, lo que sugiere que otras variables no incluidas podrían estar influyendo. Es importante recordar que la precipitación es solo un factor. Es necesario incluir otras variables como temperatura, tipo de suelo, fertilización, prácticas agrícolas o incluso eventos extremos (sequías, inundaciones) para un análisis más robusto. Podrían trabajarse con modelos no supervisados para identificar las variables más importantes y su relación, para posteriormente integrar al modelo predictivo.

En una fase exploratoria 2, se parte de las categorías más representativas del análisis anterior (cítricos, viñedo y cereales) para hacer un análisis a nivel de comunidades. De esta manera, se podrían identificar cuáles son las regiones que mayor aporte dan a la producción de estos productos y si su correlación se mantiene o aumenta. Además, se ampliarán las variables independientes, incorporando factores como temperatura media anual, radiación solar, calidad del suelo, políticas agrícolas regionales, entre otras, utilizando modelos de regresión múltiple para evaluar la interacción entre diferentes variables y su impacto combinado en la producción. Esto permitirá identificar cuáles son las comunidades autónomas con mayor contribución a la producción, para ver si la correlación entre precipitación y producción varía regionalmente.

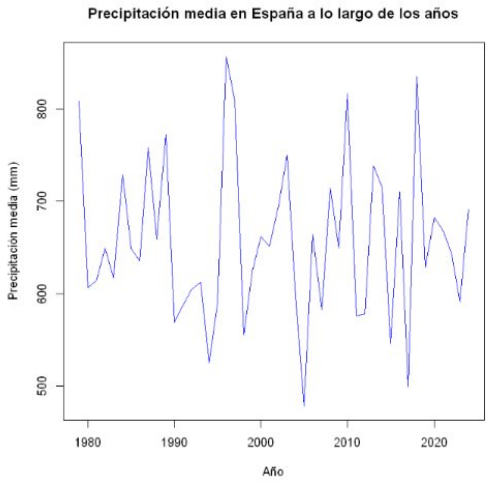

Imagen 4. Producción media de España a lo largo de los años.

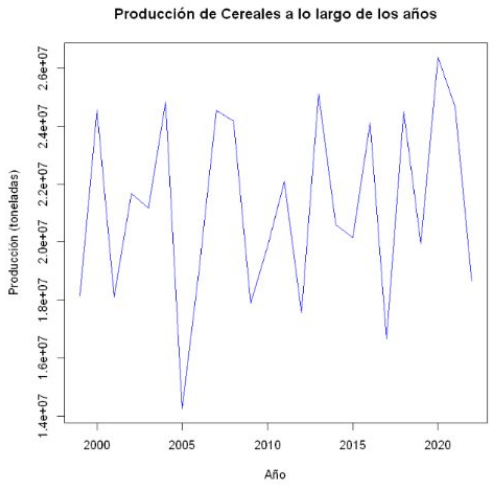

Imagen 5. Producción de Cereales a lo largo de los años.

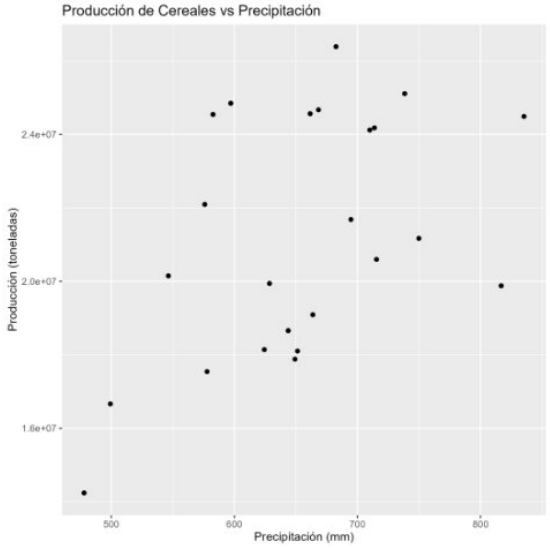

Imagen 6. Producción de Cereales Vs Precipitación.

Categoría	**Correlación**
Forrajeros	0,091
Cereales	**0,485**
Aprovechamientos	0,061
Olivar	**0,433**

Categoría	Correlación
Industriales	-0,173
Viñedo	**0,498**
Hortalizas	-0,095
Cítricos	**0,579**
Tubérculos	-0,141
Frutales de hueso	-0,107
Frutales de pepita	-0,069
Frutales carnosos	0,132
Frutales secos	0,179
Leguminosas	0,388
Otros cultivos leñosos	0,065
Bayas	-0,056

Tabla 2. Correlación por categoría.

9 PLAN DE GESTIÓN DE DATOS (DMP)

9.1 Información General

El presente Plan de Gestión de Datos (DMP) se enmarca dentro del proyecto titulado “Adaptación de cultivos condicionado por la precipitación” Este proyecto tiene como objetivo principal analizar cómo las variaciones en las precipitaciones históricas han afectado la producción agrícola, proponiendo estrategias que permitan a los agricultores adaptarse a los efectos del cambio climático. A continuación, se detalla el enfoque integral para gestionar los datos recopilados, procesados y utilizados durante este proyecto.

9.2 Tipos de Datos y Formatos

Se trabajará con diferentes tipos de datos para alcanzar los objetivos del proyecto:

- **Datos meteorológicos:**
  - **Origen:** Climate Data Store (Copernicus).
  - **Variables:** Precipitación acumulada, temperaturas y niveles de humedad.
  - **Formato:** CSV y NetCDF, seleccionados por su compatibilidad con herramientas de análisis avanzado.
- **Datos agrícolas:**
  - **Origen:** Ministerio de Agricultura, Pesca y Alimentación de España (MAPA).
  - **Variables:** Producción agrícola desglosada por región y tipo de cultivo (viñedos, olivos, cereales, etc.), incluyendo jerarquías detalladas como las presentes en "EjemploDatos.xlsx".
  - **Formato:** CSV y Excel, optimizados para integración con software de modelado estadístico.
- **Datos complementarios:**
  - **Origen:** Informes sobre sistemas de riego, calidad del suelo y eventos climáticos extremos (Ministerio de Agricultura, AEMET). Adicionalmente, se emplearán gráficos y correlaciones como las provistas en los archivos "scatter_plot" y "results_correlation".
  - **Formato:** CSV, JSON, PNG, y notebooks Jupyter.

9.3 Documentación y Metadatos

Para garantizar la trazabilidad y utilidad de los datos, implementaremos un sistema robusto de documentación y metadatos:

- **Metadatos descriptivos:**
  - Capturarán información sobre el origen, la estructura, las variables incluidas, las unidades empleadas y los intervalos temporales de los datos.
- **Estandarización:**

  - Se utilizará el modelo Dublin Core extendido, asegurando que los datos sean comprensibles para usuarios humanos y sistemas automatizados.
- **Generación automatizada:**
  - Se emplearán herramientas como Python, con librerías específicas para extraer y registrar automáticamente metadatos.

9.4 Políticas de Acceso, Uso y Reúso

Promoveremos la accesibilidad y reutilización de los datos, priorizando la transparencia y la seguridad:

- **Datos abiertos:**
  - Publicaremos los datos procesados en repositorios como Zenodo y la plataforma de la Universidad de Cantabria (UNICAN), asegurando su disponibilidad para estudiantes, investigadores y terceros interesados.
- **Protección de datos sensibles:**
  - Implementaremos técnicas de anonimización y cifrado para garantizar la seguridad de los datos personales o sensibles.
- **Licencias:**
  - Se utilizarán licencias Creative Commons BY-SA, fomentando el uso compartido y la atribución adecuada.

9.5 Almacenamiento y Seguridad

Estableceremos infraestructuras seguras y redundantes para el almacenamiento de datos:

- **Infraestructura local:**
  - Dispositivos SSD para almacenamiento temporal durante el análisis intensivo y HDD para copias de seguridad a largo plazo.
  - Configuraciones RAID garantizarán la protección frente a fallos.
- **Almacenamiento en la nube:**
  - Plataformas como AWS S3, Google Cloud Storage y copias adicionales en Google Drive.
- **Respaldo de datos:**
  - Se realizarán copias de seguridad incrementales diarias y completas semanales, almacenadas en local, en la nube y en dispositivos externos.
- **Seguridad de acceso:**
  - Implementaremos control de acceso basado en roles (RBAC) y autenticación de dos factores (2FA).

9.6 Preservación y Reutilización a Largo Plazo

El proyecto garantizará que los datos sean útiles más allá de su finalización:

- **Migración de formatos:**
  - Los datos se revisarán y convertirán periódicamente a formatos actualizados para evitar la obsolescencia tecnológica.
- **Repositorios públicos:**
  - Se utilizarán plataformas como Zenodo y UNICAN para alojar datos y documentos asociados, asegurando su conservación y accesibilidad.
- **Identificadores persistentes:**
  - Cada conjunto de datos se asociará a un DOI único para facilitar su citación y seguimiento.

9.7 Costos y Financiamiento

El presupuesto total estimado del proyecto para la gestión de datos asciende a **170,000 €**, distribuidos de la siguiente forma:

- **Infraestructura física y en la nube:**
  - Servidores dedicados y almacenamiento en la nube (AWS, Google Cloud): **20,000 €**.
  - Dispositivos SSD y HDD: **5,000 €**.
- **Software y licencias:**
  - Licencias de Python (librerías especializadas), R, ArcGIS y Tableau: **10,000 €**.
- **Personal especializado:**
  - Dos analistas de datos durante 24 meses (2,500 €/mes): **120,000 €**.
- **Formación y capacitación:**
  - Talleres y cursos sobre buenas prácticas de gestión de datos: **5,000 €**.
- **Difusión y publicación:**
  - Publicación en revistas científicas y participación en conferencias internacionales: **10,000 €**.

9.8 Cronograma

El proyecto se desarrollará durante 3 años, con las siguientes etapas:

Fase	Duración
Recolección de datos	Todo el proyecto
Limpieza y curación	Meses 1-12
Análisis exploratorio	Meses 13-16
Modelado predictivo	Meses 17-30

Fase	Duración
Desarrollo estratégico	Meses 31-33
Publicación de resultados	Meses 34-36

Durante toda la duración del proyecto se continuará recopilando datos adicionales para su incorporación y validación de los modelos. Esto implica que, aunque las fases de análisis, modelado y visualización tengan sus propios periodos, puede ser necesario revisitar tareas de limpieza y procesamiento al integrar nuevos datos.

La fase de desarrollo estratégico consta de creación de una hoja de ruta

9.9 Ética y Cumplimiento Legal

Cumpliremos con las normativas legales y éticas aplicables:

- **Cumplimiento regulatorio:**
  - Se respetará el RGPD y otras regulaciones pertinentes.
- **Consulta ética:**
  - Se realizarán revisiones regulares con comités éticos para garantizar la integridad del proyecto.
- **Consentimiento informado:**
  - En caso de utilizar datos personales, se obtendrá el consentimiento explícito de las partes involucradas.

9.10 Impacto y Beneficios

Los resultados de este proyecto tendrán un impacto significativo:

- **Avances científicos:**
  - Los datos contribuirán al entendimiento de los efectos del cambio climático en la agricultura.
- **Herramientas prácticas:**
  - Los modelos y visualizaciones desarrollados serán útiles para agricultores y formuladores de políticas públicas.
- **Futuro reutilizable:**
  - Los datos estarán disponibles para investigaciones futuras, fomentando nuevas soluciones sostenibles.

Este DMP se revisará y actualizará periódicamente para alinearse con las necesidades emergentes del proyecto y maximizar su impacto.

10 CONCLUSIONES

La adaptación anticipada del sector primario al cambio climático es un factor crucial para asegurar la sostenibilidad y eficiencia de la producción agrícola en España. Ante la creciente incertidumbre climática, disponer de pautas claras sobre qué tipo de cultivo plantar según las condiciones climáticas esperadas resulta vital para optimizar los rendimientos y reducir los riesgos asociados a fenómenos extremos, como sequías o lluvias excesivas. Este trabajo pone de manifiesto la importancia de explorar vínculos entre las precipitaciones y los tipos de cultivos, ayudando a los agricultores a tomar decisiones informadas que permitan anticiparse a los cambios climáticos y adaptarse a ellos.

En cuanto a lo aprendido, los resultados obtenidos en el análisis preliminar revelan que sí existe una correlación entre las precipitaciones y la producción de ciertos cultivos. De hecho, la categoría de cítricos a nivel nacional muestra la correlación más alta (0.579) con respecto al resto de los productos analizados, seguida de viñedo (0.498), cereales (0.485) y olivar (0.433). Aunque estas correlaciones positivas sugieren que un aumento en las precipitaciones podría estar asociado con un incremento en la producción, es importante destacar que el coeficiente de correlación no mide causalidad, lo que implica que no se pueden sacar conclusiones definitivas sin un análisis más profundo. Este hallazgo nos permite avanzar hacia la identificación de patrones y potenciales tendencias, pero también resalta la necesidad de seguir explorando más variables para obtener un panorama completo.

Sin embargo, las limitaciones del estudio son evidentes. Dado que se trata de un análisis preliminar, se han omitido numerosas variables que influyen de manera significativa en la producción agrícola, como el tipo de suelo, la temperatura, las prácticas agrícolas y fenómenos meteorológicos extremos como sequías o inundaciones. La disponibilidad de datos limitados (solo unos pocos años) también ha restringido el alcance del análisis, lo que subraya la necesidad de contar con una mayor cantidad de información a lo largo del tiempo para validar y refinar los resultados.

En cuanto al ciclo de vida de los datos, uno de los aprendizajes clave ha sido la importancia de una gestión adecuada y continua de los datos. Este trabajo ha mostrado cómo un análisis exploratorio bien planteado puede proporcionar una visión inicial valiosa, pero también ha quedado claro que las bases de datos deben mantenerse actualizadas y enriquecidas con el tiempo, incorporando nuevas variables y ajustando los modelos a medida que se recopilan más datos. Este enfoque dinámico permite mejorar los modelos predictivos y garantizar que las pautas generadas para los agricultores sean lo más precisas y adaptadas posible a los cambios climáticos futuros.

Se recomienda ampliar el estudio para incluir un mayor número de cultivos y un periodo de tiempo más largo. Además, debería trabajarse con modelos que

permitan identificar las variables más influyentes y su relación con las precipitaciones. De forma paralela, una exploración a nivel regional con variables adicionales permitirá afinar aún más las recomendaciones para los agricultores y mejorar la capacidad de adaptación del sector primario a los cambios climáticos.

11 BIBLIOGRAFÍA

Acevedo García, A., Cómbita Niño, J., Pérez González, E., Nasta, M. C., Sainz Guerra, J. A., & Villa Viard, O. S. (2025). *Annual accumulated precipitation in Spain and CCAA from 1979 to 2024*. https://doi.org/10.5072/ZENODO.162490

Escudero Población, A., Mancheño Losa, S., & López Pérez, J. J. (2024). *Anuario de estadística 2023*. https://cpage.mpr.gob.es/

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/GMD-9-1937-2016

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/QJ.3803

Hersbach, H., Muñoz Sabater, J., Nicolas, R., I., S., Vamborg, F., A., B., B., B., P., B. G., Buontempo, C., Horányi, A. , J., Peubey, C., Radu, R., Schepers, D., Soci, C., Dee, D., & Thépaut, J.-N. (2018). *Essential climate variables for assessment of climate variability from 1979 to present*. https://cds.climate.copernicus.eu/datasets/ecv-for-climate-change?tab=overview

ANEXO 1

Resumen

1.1 Breve descripción del resultado de la investigación descrito

1.1.1 ¿Qué tipo de resultado de investigación estás describiendo? Este proyecto describe datos relacionados con precipitaciones meteorológicas y producción agrícola en España. Se trata de conjuntos de datos geoespaciales, estadísticos y visualizaciones.

1.1.2 ¿Es físico o digital? El output obtenido es completamente digital.

1.1.3 ¿Estás generándolo o reutilizándolo? Los datos incluyen tanto datos generados (modelos predictivos y análisis) como reutilizados (bases de datos de Copernicus y MAPA).

1.1.4 ¿Cuál es el tipo de conjunto de datos descrito? El conjunto de datos combina datos secundarios (limpiados y compartidos por Copernicus y MAPA) con datos primarios generados mediante análisis propios.

1.1.5 ¿Cuál es su formato? Los datos estarán en formatos CSV, Excel, JSON y NetCDF, mientras que las visualizaciones se generarán como archivos PNG.

1.1.6 ¿Cuál es su tamaño esperado? Se estima que el tamaño total de los datos será de aproximadamente 10 TB, incluyendo datos originales, procesados y visualizaciones.

1.1.7 ¿Por qué estás recolectando/generando o reutilizando estos datos? Los datos se utilizan para comprender la relación entre las precipitaciones y la producción agrícola, con el fin de desarrollar modelos predictivos y estrategias adaptativas frente al cambio climático. El motivo que nos incita a utilizar datos ya recolectados es la disposición de una infraestructura ya disponible, con la posibilidad de obtener datos de precipitaciones de largos periodos, ayudando a crear conclusiones más robustas.

1.1.8 ¿Cuál es su origen o procedencia? Los datos provienen de Copernicus (Climate Data Store) y MAPA (producción agrícola). Los datos generados se derivan de análisis y modelos propios.

1.1.9 ¿A quién podrían serle útiles? Estos datos serán útiles para investigadores, formuladores de políticas, agricultores e inversores interesados en agricultura sostenible y energía renovable.

2. Enlaces entre resultados

Los datos y análisis respaldarán resultados científicos descritos en publicaciones. Los datasets procesados estarán vinculados a gráficos y modelos predictivos.

3. Prácticas FAIR

3.1 Haciendo que los datos y otros resultados sean localizables, incluyendo disposiciones para metadatos

3.1.1 ¿Qué tipo(s) de identificadores persistentes usas para el conjunto de datos/resultado descrito? Se asignarán DOIs a través de repositorios como Zenodo y la plataforma UNICAN. Los metadatos estarán detallados según el modelo Dublin Core.

3.1.2 ¿Proveerás metadatos para el conjunto de datos/resultado descrito? Sí, se incluirán metadatos detallados según el modelo Dublin Core.

3.1.5 Proporciona URL o descripción de los vocabularios usados El vocabulario empleado se basará en estándares de datos meteorológicos y agrícolas, como el NetCDF Climate and Forecast (CF) Metadata Conventions.

3.2 Haciendo que los datos y otros resultados sean accesibles abiertamente

3.2.1.1 ¿En qué repositorio se depositará el conjunto de datos/resultado? Zenodo y el portal de Datos Abiertos de la Universidad de Cantabria (UNICAN).

3.2.1.2 ¿Es el repositorio seleccionado una fuente confiable? Sí, ambos repositorios son confiables y cumplen estándares internacionales.

3.2.1.4 ¿Qué acuerdos se han hecho con el(los) repositorio(s) donde se depositará el conjunto de datos/resultado? Se garantizará la accesibilidad mediante acuerdos de uso abierto y licencias específicas.

3.2.1.5 ¿El(los) repositorio(s) asigna(n) identificadores persistentes a los conjuntos de datos/resultados? Sí, Zenodo y UNICAN generan DOIs para cada conjunto de datos.

3.2.1.7 ¿El repositorio admite versionado de datos? Sí, ambos repositorios permiten la versionado de datos.

3.2.2 Datos

3.2.2.1 ¿Cuál es el título del conjunto de datos/resultado descrito? "Relación entre precipitaciones y producción agrícola en España: Análisis y modelos predictivos."

3.2.2.2 ¿Cómo se compartirá el conjunto de datos/resultado? Será compartido como datos abiertos a través de Zenodo y UNICAN.

3.2.2.5 ¿Se necesitan métodos o herramientas específicas para acceder al conjunto de datos/resultado? Los datos estarán disponibles en formatos estándares y accesibles, con herramientas comunes como Python y R, para el público. Los resultados obtenidos, así como el desarrollo de los metodos de analisis serán accesibles bajo licencia Creative Commons BY-SA.

3.2.2.8 ¿El conjunto de datos/resultado está respaldado por un comité de acceso a datos? No, pero contará con medidas de acceso controlado si es necesario.

3.2.2.9 ¿Cómo se accederá al conjunto de datos/resultado durante y después de finalizar el proyecto? Durante el proyecto, los datos estarán disponibles en repositorios privados. Posteriormente, se migrarán a plataformas públicas.

3.2.2.10 ¿Por cuánto tiempo después del final del proyecto estará accesible el conjunto de datos/resultado? Los datos estarán disponibles indefinidamente en los repositorios seleccionados.

3.2.3 Metadatos

3.2.3.1 ¿Proveerás metadatos aunque no puedas compartir abiertamente el conjunto de datos/resultado descrito? Sí, los metadatos estarán disponibles incluso si los datos no son accesibles.

3.2.3.2 ¿Bajo qué licencia se proporcionarán los metadatos? Creative Commons BY-SA.

3.2.3.3 ¿Los metadatos proporcionan información sobre cómo acceder al conjunto de datos/resultado descrito? Sí, incluirán instrucciones detalladas de acceso.

3.2.3.4 ¿Permanecerán disponibles los metadatos después de que el conjunto de datos/resultado ya no esté accesible? Sí, los metadatos se mantendrán disponibles en los repositorios públicos.

3.3 Haciendo que los datos y otros resultados sean interoperables

3.3.1 ¿Tu (meta)datos usan vocabulario controlado? Sí, se emplearán vocabularios controlados como los estándares NetCDF y CF.

3.3.5 ¿Qué metodología sigues? La metodología se basará en análisis estadísticos y modelos predictivos reproducibles.

3.4 Incrementando la reutilización de los datos y otros resultados

3.4.1 ¿Qué licencia internacionalmente reconocida usarás para tu conjunto de datos/resultado? Creative Commons BY-SA.

3.4.4 ¿Asegurarás la reutilización por terceros después de que tu proyecto termine? Sí, el proyecto fomentará la reutilización mediante datos abiertos y documentación completa.

4. Asignación de recursos

4.1.1 ¿Cuál será el costo de hacer que el resultado sea FAIR? Se estima un costo de 170,000 €.

4.1.2 ¿Cómo se cubrirá este costo? A través de financiación europea, pública y subvenciones.

5. Seguridad

5.1.1 ¿Qué medidas de seguridad se siguen? Cifrado, autenticación multifactorial y respaldos periódicos.

6. Aspectos éticos

6.1.1 ¿Existen cuestiones éticas o legales que puedan afectar la compartición del conjunto de datos/resultado descrito? No, los datos son reutilizados y no contienen información personal sensible.

ANEXO 2

Horizon Europe

Data Management Plan Template

Version 1.0

19 January 2025

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	**19.01.2025**	● **Initial version**

Action Number: Project0119012025

Action Acronym: ALCBORD

Action title: Adaptation of crops conditioned by precipitation in Spain

Date: 19 of January 2025

DMP version:

https://docs.google.com/document/d/1aHuZtt2Fwzr6oDY0MEGdp_leR3vEE88Y6np4lB34U9I/edit?tab=t.0

The Horizon Europe Model Grant Agreement requires that a data management plan (‘DMP’) is established and regularly updated.

The use of this template is recommended for Horizon Europe beneficiaries. In completing the sections of the template the

requirements for research data management of Horizon Europe as described in article 17 and analysed in the Annotated Grant

Agreement, article 17, must be addressed.

1. Data Summary

The data used in this study is extracted from official sources such as **Essential climate variables for assessment of climate variability from 1979 to present** from the European Union Programme, obtained from Copernicus, as well as data collected by the Spanish Ministry of Agriculture, Fisheries, and Food in its study **“Crop Areas and Productions”**. Therefore, these are reused data.

The purpose of this data analysis is to generate insights based on the collected rainfall data in the Spanish region, enabling comparisons with regional agricultural production to draw various conclusions in this area.

These findings could be useful for future agricultural investors to determine whether their planned investments are appropriately distributed based on the location. Focusing on other economic sectors, specifically within renewable energies, this study can aid decision-making regarding where to invest—or not—in solar energy.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

The data will be hosted in a trusted repository, specifically the repository of the University of Cantabria. These datasets will be accessible to anyone who wishes to use them, considering they are derived from official sources. Access to the data will be available through the UNICAN platform, making them accessible to students, alumni, researchers, and anyone interested in utilizing them.

The data will remain accessible to any interested user for an indefinite period, depending on the university platform's management. The datasets will be provided in CSV format to enhance accessibility.

2.3. Making data interoperable

The vocabulary present in the data is specific to meteorological data and economic variables, such as agricultural production in the study area. For this reason, it will not include specific ontologies.

These datasets will include references to the sources from which they were obtained.

2.4. Increase data re-use

The data used for the analysis comes from public sources, making it mandatory to reference the original authors. This facilitates their update and reuse by other users interested in the domain.

Since the data will be available on a public platform, namely that of the University of Cantabria, their accessibility and free availability are reiterated, encouraging reuse by third parties.

3. Other research outputs

As previously mentioned, the open availability of the data can benefit small or large agricultural investors who wish to expand their exploitation areas based on the specific characteristics of each analyzed region.

On the other hand, in the field of renewable energy, various solar energy producers might find these results valuable, as they highlight climates with low rainfall, which are more favorable for solar energy production.

4. Allocation of resources

Since these data will be published on a public platform, the associated costs will be covered by the platform itself, funded through local grants. The data will be available in the database managed by the University of Cantabria's Library.

5. Data security

On one hand, these data will be available on the University of Cantabria's platform, and additional measures will be taken, such as maintaining a backup copy on Google Drive. This ensures the data can be restored in any circumstance, guaranteeing its long-term security.

6. Ethics

These data will be available without any ethical concerns, as they are reused from European and national sources, ensuring their authenticity and accuracy.

7. Other issues

Additionally, efforts will be made to access various national grants related to the digitization of bibliographic heritage and its dissemination and preservation through repositories. This will support the preservation of the curated data and enhance its accessibility to the public.