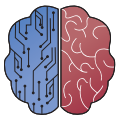
UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



INGENIERÍA
DE LA SALUD

TFG del Grado en Ingeniería de la Salud

título del TFG
Documentación Técnica

Presentado por nombre alumno
en Universidad de Burgos

22 de mayo de 2025

Tutores: nombre tutor – nombre tutor 2

Índice general

Índice de figuras

Índice de tablas

Apéndice A

Plan de Proyecto Software

A.1. Introducción

bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla bla.

Ojo [1]

A.2. Planificación temporal

Planificación económica

Viabilidad legal

[1]Los anexos deben de tener su propia bibliografía, eso es tan fácil como utilizar referencias igual que en la memoria [**?**]

Apéndice B

Documentación de usuario

B.1. Requisitos software y hardware para ejecutar el proyecto.

B.2. Instalación / Puesta en marcha

B.3. Manuales y/o Demostraciones prácticas

Apéndice C

Manual del desarrollador / programador / investigador.

C.1. Estructura de directorios

Descripción de los directorios y ficheros entregados.

C.2. Compilación, instalación y ejecución del proyecto

En caso de ser necesaria esta sección, porque la compilación o ejecución no sea directa.

C.3. Pruebas del sistema

Esta sección puede ser opcional.

Puede tratarse de validación de la interfaz por parte de los usuarios, mediante escuestas o similar o validación del funcionamiento mediante pruebas unitarias.

C.4. Instrucciones para la modificación o mejora del proyecto.

Instrucciones y consejos para que el trabajo pueda ser mejorado en futuras ediciones.

Apéndice D

Descripción de adquisición y tratamiento de datos

Tablas, imágenes, señales, secuencias de ADN...

D.1. Descripción formal de los datos

Los datos empleados para la elaboración del trabajo provienen de la plataforma Our World in Data (OWD). Las variables consideradas son la prevalencia de la enfermedad de Parkinson, la tasa de mortalidad por contaminación del aire Número estimado de muertes atribuidas a diferentes tipos de contaminación atmosférica, la tasa de carga de morbilidad por exposición al plomo, las muertes atribuidas a fuentes de agua insalubres, el uso de plaguicidas y la precipitación anual.

D.1.1. Prevalencia de la enfermedad del parkinson (Variable dependiente)

- **Definición y unidad de medida:** Esta variable se define como el numero estimado de personas con enfermedad del Parkinson, cuya unidad de medida se expresa por cada 100.000 habitantes.
- **Estrucutura de los datos:** Los datos se encuentra organizados por país y año, con un rango temporal que abarca desde 1990 hasta 2021.
- **Decripción:** es la variable dependiente en este trabajo, ya que con el estudio de esta se busca entender como factores como la contaminación,

el uso de pepticidas u otras variables pueden estar relacionadas con la prevalencia de la enfermedad del Parkinson.

D.1.2. Variables independientes

Las variables independientes son aquellas que se consideran factores que pueden influir o tener un impacto sobre la prevalencia de la enfermedad de Parkinson.

1. **Tasa de mortalidad por contaminación del aire**

   - **Definición y unidad de medida:** Representa el numero estimado de muertes atribuidas a diferentes tipos de contaminación del aire por cada 100.000 habitantes.
   - **Estructura:**Los datos están disponibles por país y año desde 1990 hasta 2021.
   - **Descripción:** Esta variable mide el impacto de la contaminación del aire en la mortalidad. A través de esta variable, se puede evaluar como la exposición a ciertos contaminantes como las partículas PM2.5, podría estar relacionada con la prevalencia de la enfermedad.

2. **Tasa de carga de enferemdad por exposición al plomo**

   - **Definición y unidad de medida:** Numero estimado de años de vida ajustados por discapacidad (AVAD) debido a la exposición al plomo, estandarizados por edad, provenientes de todas las causas, por cada 100.000 habitantes.
   - **Estructura:** Los datos se encuentran organizados por pais y año, con un rango temporal que abarca desde 1990 hasta 2021.
   - **Descripción:** Los años de vida ajustado por discapacidad (AVAD) miden la carga total sobre la salud de la población, considerando los años de vida perdidos por muertes prematuras y los años vividos con discapacidad. En este caso, la exposición al plomo se asocia con diversos problemas de salud que afectan a la calidad de vida y la mortalidad. La carga total se calcula sumando todos los efectos de salud relacionados con esta exposición, sin especificar las causas exactas de las muertes o discapacidades.

3. **Muertes atribuidas a fuentes de agua inseguras**

- **Definición y unidad de medida:** Se define como el número total de muertes causadas por fuentes de agua no seguras.
- **Estructura:** Los datos están organizados por país y año, con un rango temporal que abarca desde 1990 hasta 2021.
- **Descripción:** Esta variable mide el impacto del consumo de agua no segura en la mortalidad, sumando todas las muertes que pueden estar relacionadas con el agua insalubre, como enfermedades transmitidas por el agua o infecciones gastrointestinales. Se considera el total de muertes atribuidas a esta causa, sin especificar cada enfermedad o condición que causó la muerte.

4. **Uso total de pesticidas**

- **Definición y unidad de medida:** Se define como el uso total de pesticidas medido en toneladas.
- **Estructura:** Los datos se encuentran organizados por país y año, con un rango temporal que cubre desde 1990 hasta 2022.
- **Descripción:** Los pesticidas totales, incluyen los insecticidas, fungicidas y bactericidas, herbicidas, reguladores de crecimiento de las plantas, rodenticidas, desifenctantes entre otros.

5. **Precipitaciones anuales**

- **Definición y unidad de medida:** Se define como las precipitaciones anuales totales (lluvia y nieve), calculada como la suma de los promedios diarios y expresada como la profundidad del agua que cae a la superficie de la Tierra, excluyendo la niebla y el rocío.La variable se mide por milímetros de precipitación.
- **Estructura:**Los datos están organizados por país y por año, con un rango temporal que abarca desde 1940 hasta 2024.
- **Descripción:** Esta variable representa la cantidad total de precipitación que ocurre en un área durante un año, incluyendo tanto la lluvia como la nieve derretida. La medida se expresa en milímetros, indicando la profundidad del agua que caería sobre la superficie terrestre si se recogiera toda la precipitación. Los valores no incluyen fenómenos como la niebla o el rocío, que no aportan agua de manera significativa al suelo.

D.2. Descripción clínica de los datos.

En esta sección se presenta la prespectiva clíncia de las variables consideradas para el estudio, el objetivo de esto es contextualizar de que manera estoas variables pueden influir en la prevalencia de la enfermedad del Parkinson.

D.2.1. Prevalencia de la enfermedad del Parkinson

La enfermedad de Parkinson es un trastorno neurodegenerativo progresivo que afecta principalmente al sistema motor, causado por la pérdida de neuronas dopaminérgicas en la sustancia negra del cerebro. Clínicamente, se manifiesta con síntomas como temblores en reposo, rigidez muscular, bradicinesia (lentitud de movimientos) y alteraciones posturales. Su prevalencia aumenta con la edad y puede estar influenciada por factores ambientales y genéticos.[Instituto Nacional sobre el Envejecimiento (NIA), 2022] .

D.2.2. Tasa de mortalidad por contaminación del aire

La exposición prolongada a contaminantes del aire como las partículas finas ($PM_{2,5}$), dióxido de nitrógeno (NO_2) y ozono (O_3) se ha asociado con un mayor riesgo de enfermedades cardiovasculares y neurodegenerativas. Estudios recientes sugieren que la contaminación del aire puede inducir estrés oxidativo e inflamación sistémica, lo que podría contribuir a la neurodegeneración observada en enfermedades como el Parkinson.[Kirrane et al., 2015]

D.2.3. Carga de enfermedad por exposición al plomo

El plomo es un neurotóxico conocido que puede acumularse en el cerebro y alterar funciones neurológicas. En adultos, la exposición crónica al plomo ha sido relacionada con una mayor incidencia de deterioro cognitivo y enfermedades neurodegenerativas. Desde una perspectiva clínica, su asociación con el Parkinson se explica por el daño oxidativo y la disfunción mitocondrial inducida por este metal pesado.[Pyatha et al., 2022]

D.2.4. Muertes atribuidas a fuentes de agua inseguras

Aunque las enfermedades derivadas del consumo de agua contaminada no tienen una relación directa con el Parkinson en todos los casos, la exposición a ciertos contaminantes químicos presentes en el agua, como pesticidas y metales pesados, ha sido asociada con efectos neurotóxicos. Varios estudios

indican que la exposición prolongada a contaminantes del agua, como el tetracloroetileno (TCE) y otros productos químicos, puede estar relacionada con un mayor riesgo de desarrollar enfermedades neurodegenerativas, incluida la enfermedad de Parkinson.[Pacheco Moisés et al., 2011, inf, 2023, ken, 2022].

D.2.5. Uso total de pesticidas

El uso de pesticidas, especialmente herbicidas como el paraquat y fungicidas como el maneb, ha sido consistentemente asociado con un mayor riesgo de desarrollar la enfermedad de Parkinson. Estos compuestos pueden inducir estrés oxidativo y afectar la función mitocondrial, contribuyendo al daño neuronal característico de la enfermedad. Varios estudios han encontrado que la exposición prolongada a estos pesticidas aumenta significativamente el riesgo de desarrollar Parkinson, particularmente en áreas agrícolas donde su uso es elevado.[Pearce et al., 2013, Tanner et al., 2011, Starks et al., 2013]

D.2.6. Precipitaciones anuales

Aunque las precipitaciones no influyen directamente en la salud humana, pueden actuar como moduladores del entorno, afectando la dispersión de contaminantes o el uso agrícola de pesticidas. Desde un punto de vista clínico, su relevancia radica en su potencial para modificar la exposición a factores ambientales vinculados con la neurotoxicidad.[America, 2023]

Apéndice E

Manual de especificación de diseño

Si es necesario.

Planos (Si procede) Diseño arquitectonico (Si procede) Diagrama de clases, diagrama de despliegue

E.1. Planos

Si procede

E.2. Diseño arquitectónico

Si procede.

Diagramas de clases, diagramas de despliegue ...

Apéndice F

Especificación de Requisitos

Si procede.

F.1. Diagrama de casos de uso

F.2. Explicación casos de uso.

Se puede describir mediante el uso de tablas o mediante lenguaje natural.

Una muestra de cómo podría ser una tabla de casos de uso:

F.3. Prototipos de interfaz o interacción con el proyecto.

CU-1	**Ejemplo de caso de uso**
Versión	1.0
Autor	Alumno
Requisitos asociados	RF-xx, RF-xx
Descripción	La descripción del CU
Precondición	Precondiciones (podría haber más de una)
Acciones	1. Pasos del CU 2. Pasos del CU (añadir tantos como sean necesarios)
Postcondición	Postcondiciones (podría haber más de una)
Excepciones	Excepciones
Importancia	Alta o Media o Baja...

Tabla F.1: CU-1 Nombre del caso de uso.

Apéndice G

Estudio experimental

G.1. Cuaderno de trabajo.

Enumeración de todos los métodos probados con resultados positivos o no.

Con el fin de evaluar la relación entre las variables ambientales y la prevalencia del Parkinson a escala mundial, se ha llevado a cabo un análisis exploratorio preliminar.Esta etapa tuvo como objetivo detectar posibles problemas de multicolinealidad entre las variables independientes y orientar adecuadamente la fase de modelado.

Para ello, se construyó una matriz de correlación (véase Figura G.1), a partir de la cual se observó que ninguna de las variables presentaba coeficientes de correlación superiores a 0.7. Esto permitió descartar redundancia estadística y justificar el uso conjunto de todas ellas en los modelos predictivos.

Posteriormente, se estudiaron las relaciones individuales entre cada variable independiente y la variable objetivo (casos de Parkinson). Para ello, se realizaron gráficos de dispersión con líneas de tendencia ajustadas, lo que permitió observar que algunas relaciones eran no lineales.

La Figura G.2 muestra la relación entre Muertes por agua contaminada y los casos de Parkinson, donde se puede observar una curvatura inicial que luego se estabiliza.

De manera similar, las variables Exposición al plomo y Contaminación del aire también mostraron un comportamiento curvado.(ver Figura G.3 y Figura G.4).

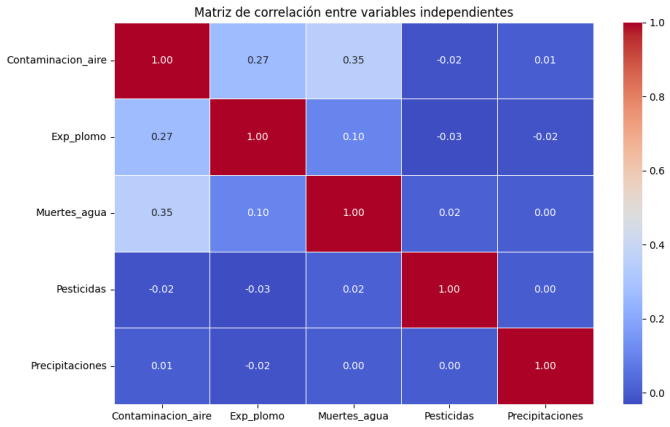


Figura G.1: Matriz de correlación entre variables independientes.

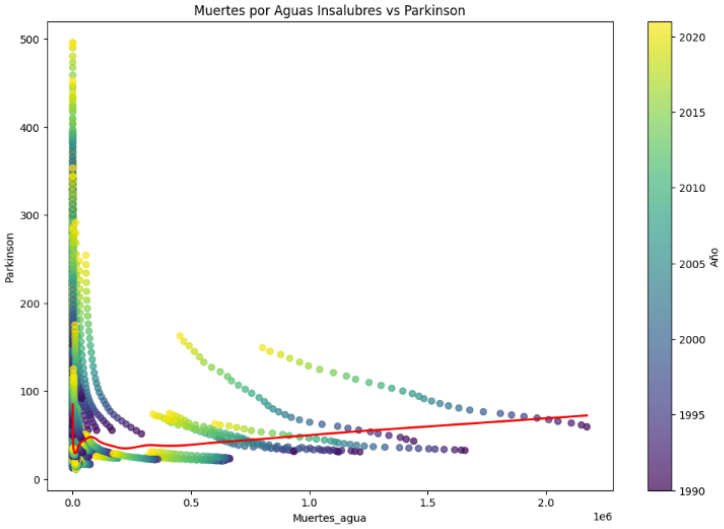


Figura G.2: Relación entre el Parkinson y las Muertes atribuidas a fuentes de agua inseguras.

Por otro lado, la variable Pesticidas mostró un patrón logarítmico con picos, lo que sugiere que enfoques de modelado que contemplen relaciones no lineales o funciones logarítmicas podrían resultar más apropiados para capturar su comportamiento. La Figura G.5 ilustra este comportamiento logarítmico.

En cuanto a Precipitaciones, la relación con los casos de Parkinson fue lineal,lo que sugiere que modelos lineales podrían ser adecuados para este predictor, como se muestra en Figura G.6.

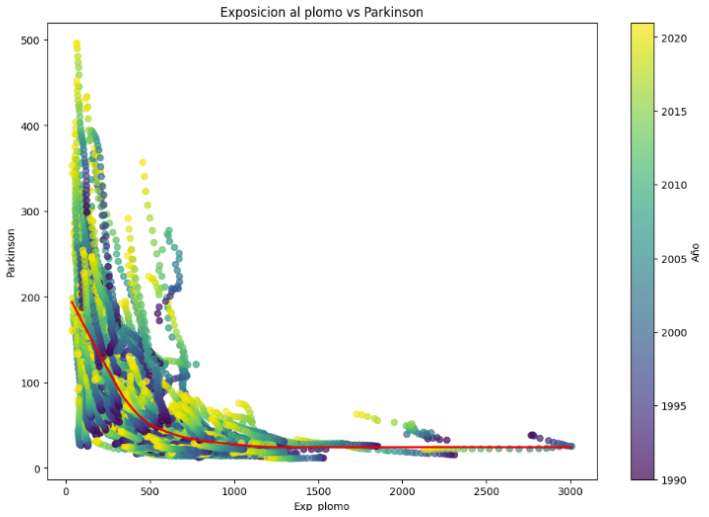


Figura G.3: Relación entre el Parkinson y la Tasa de carga de enferemdad por exposición al plomo.

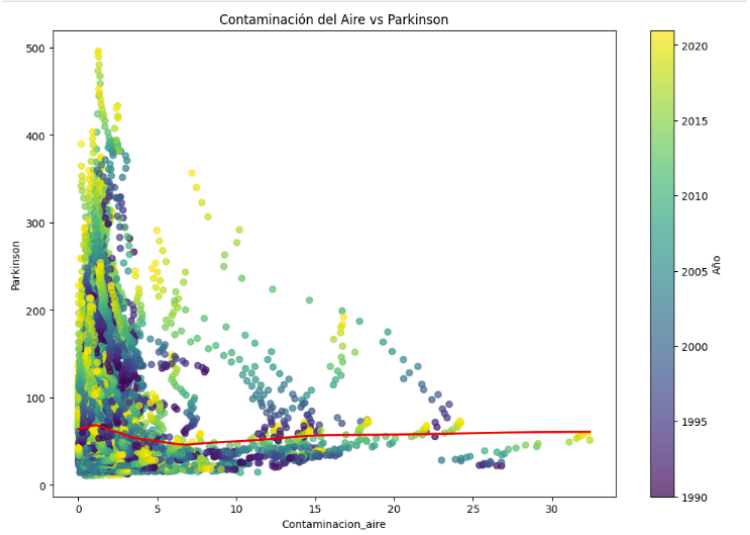


Figura G.4: Relación entre el Parkinson y la Tasa de mortalidad por contamiación de aire.

Este análisis preliminar no implicó la transformación directa de las variables, sino que se enfocó en identificar la naturaleza de sus relaciones con el objetivo de orientar la elección y formulación de modelos adecuados en etapas posteriores. De este modo, los modelos se adaptarán a los datos y

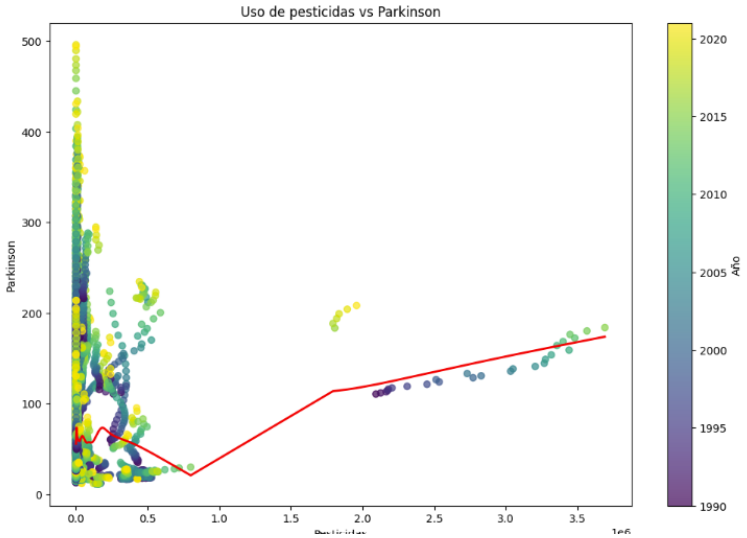


Figura G.5: Relación entre el Parkinson y el uso de pesticidas.

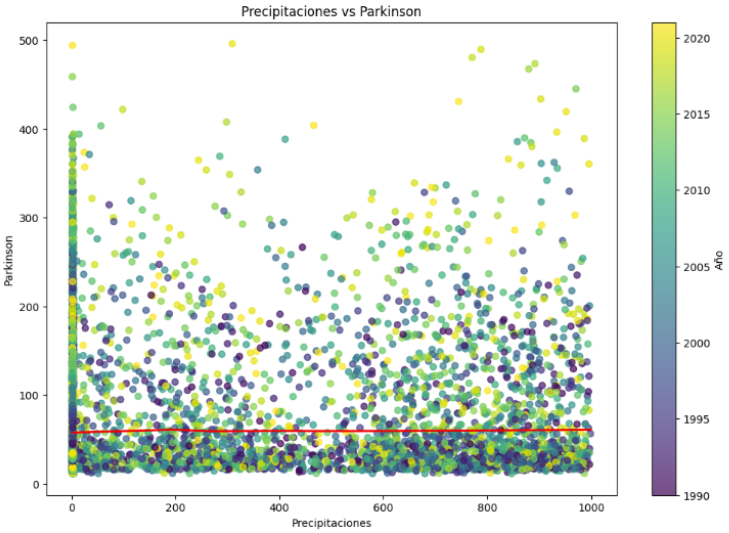


Figura G.6: Relación entre el Parkinson y el uso de pesticidas.

no al revés, respetando sus patrones inherentes y optimizando la capacidad predictiva de cada enfoque.

G.1.1. Elección de modelos

Una vez realizado el análisis preliminar de los datos, se estudian que modelos son los más afines a utilizar para la predicción según las características

que presentan los datos. El objetivo es aplicar una combinación de modelos de diferentes familias ya que ningún modelo por sí solo es capaz de capturar completamente la complejidad de las relaciones presentes en los datos. Cada tipo de modelo tiene características y capacidades particulares que lo hacen más adecuado para ciertos patrones de datos.

Dado que la variable objetivo en este estudio es el número estimado de casos de Parkinson,es decir, un conteo que representa el número de casos en diferentes países, se requiere un enfoque que se adapte específicamente a variables de recuento. Los modelos seleccionados para esta tarea son aquellos que son capaces de manejar correctamente datos con características no lineales, distribuciones sesgadas o complejas y relaciones no evidentes entre las variables predictoras. A continuación, se explica por qué se eligieron los siguientes modelos:

1. **Generalized Linear Model (GLM) con distribución Binomial Negativa**

   - **Motivo y aplicación:** La familia GLM es una opción adecuada para modelar variables de recuento, ya que permite ajustar la distribución de la variable objetivo según la naturaleza de los datos. En este caso, se seleccionó la distribución Binomial Negativa en lugar de Poisson, ya que la varianza de los recuentos era mayor que la media, y este modelo supone que ambas son iguales.

     A pesar de probar con el modelo Cuasi-Poisson (Figura G.7), los resultados obtenidos fueron insatisfactorios, con un Pseudo R-squared de 1, lo que indicaba un sobreajuste de los datos. Al comparar la verosomilitud entre ambos modelos y al obtener resultados similares, el modelo seleccionado fue el modelo Binomial Negativo (Figura G.8).

   - **Estructura del modelo:** El modelo GLM se rige por la siguiente fórmula:

     $$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

     donde $\mathbb{E}[Y]$ representa el numero estimado de casos de Parkinson y X_i son las variables ambientales consideradas. Esta forma permite capturar efectos multiplicativos, lo que resulta adecuado para datos de recuento.

- **Adaptación del modelo a los datos:** A partir de los resultados del análisis exploratorio preliminar, se observó que algunas variables presentaban relaciones no lineales con la variable objetivo. Para reflejar estas dinámicas sin transformar directamente los datos, se adaptó la fórmula del modelo GLM incorporando diferentes formas funcionales según la naturaleza observada de cada variable. En particular, para las variables que mostraban curvatura inicial, se evaluaron términos polinómicos hasta grado 3 (X, X^2, X^3), no se utilizaron polinomios de grado superior a 3 porque pueden generar sobreajuste. Además, dado que la variable objetivo representa un número de casos de una enferemdad, no tiene sentido aplicar funciones más complejas que podrían dar lugar a predicciones poco realistas.

  El uso de términos hasta grado 3 permite capturar la curvatura observada sin perder sentido práctico ni interpretabilidad. En el caso de *Pesticidas*, se incorporó el término $\log(1 + X)$ al reflejar un comportamiento logarítmico, mientras que *Precipitaciones* se mantuvo en forma lineal, al no requerir ajustes adicionales.

  La inclusión final de cada término se basó exclusivamente en su significancia estadística, conservando solo aquellos que aportaban valor explicativo real al modelo. Este enfoque permitió ajustar la forma funcional del modelo a la naturaleza de los datos, mejorando su capacidad predictiva y manteniendo la estructura propia del GLM.

Modelo Cuasi-Poisson:
Generalized Linear Model Regression Results

Dep. Variable:	Parkinson	No. Observations:	4323
Model:	GLM	Df Residuals:	4313
Model Family:	Poisson	Df Model:	9
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-57750.
Date:	Thu, 10 Apr 2025	Deviance:	90197.
Time:	18:05:00	Pearson chi2:	1.06e+05
No. Iterations:	6	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.1322	0.002	1913.827	0.000	4.128	4.136
Contaminacion_aire	0.2118	0.005	44.254	0.000	0.202	0.221
Exp_plomo	-1.5512	0.005	-310.105	0.000	-1.561	-1.541
Muertes_agua	-0.3351	0.009	-36.244	0.000	-0.353	-0.317
Pepticidas	-0.0057	0.002	-3.703	0.000	-0.009	-0.003
Precipitaciones	0.0024	0.002	1.448	0.148	-0.001	0.006
Contaminacion_aire_2	-0.0715	0.005	-15.303	0.000	-0.081	-0.062
Muertes_agua_2	0.2830	0.008	34.030	0.000	0.267	0.299
Exp_plomo_2	0.9234	0.005	175.366	0.000	0.913	0.934
Pepticidas_log	0.0535	0.002	29.559	0.000	0.050	0.057

Error Cuadrático Medio (RMSE): 47.40631071307246
Error Absoluto Medio (MAE): 31.832376321300544

Figura G.7: Modelo Cuassi-poisson.

Modelo Binomial Negativo:
Generalized Linear Model Regression Results

Dep. Variable:	Parkinson	No. Observations:	4323
Model:	GLM	Df Residuals:	4313
Model Family:	NegativeBinomial	Df Model:	9
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-22292.
Date:	Thu, 10 Apr 2025	Deviance:	1150.3
Time:	18:05:01	Pearson chi2:	1.56e+03
No. Iterations:	13	Pseudo R-squ. (CS):	0.4104
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.1466	0.015	269.926	0.000	4.116	4.177
Contaminacion_aire	0.1897	0.039	4.896	0.000	0.114	0.266
Exp_plomo	-1.4937	0.043	-34.496	0.000	-1.579	-1.409
Muertes_agua	-0.2855	0.050	-5.722	0.000	-0.383	-0.188
Pepticidas	-0.0057	0.017	-0.341	0.733	-0.039	0.027
Precipitaciones	0.0027	0.015	0.177	0.860	-0.027	0.033
Contaminacion_aire_2	-0.0577	0.037	-1.546	0.122	-0.131	0.015
Muertes_agua_2	0.2609	0.049	5.354	0.000	0.165	0.356
Exp_plomo_2	0.9005	0.043	21.178	0.000	0.817	0.984
Pepticidas_log	0.0653	0.017	3.841	0.000	0.032	0.099

Error Cuadrático Medio (RMSE): 47.63369176286732
Error Absoluto Medio (MAE): 32.00714819047614

Figura G.8: Modelo-Binomial Negativo

2. **Random Forest**

   - **Motivo y aplicación:** El modelo Random Forest es una técnica basada en la combinación de múltiples árboles de decisión, lo que le permite capturar relaciones complejas y no lineales entre las variables. Se adapta bien a datos con ruido y a relaciones no evidentes, lo que lo convierte en una buena opción para el tipo de datos utilizados en este estudio.

     Se empleó el algoritmo RandomForestRegressor dado que la variable objetivo es numérica (número estimado de casos de Parkinson). Este modelo permite obtener predicciones precisas sin necesidad de asumir una forma funcional específica entre las variables independientes y la variable objetivo.

3. **XGBoost**

   - **Motivo y aplicación:** XGBoost es un modelo de boosting basado en árboles que destaca por su alta precisión y capacidad para capturar relaciones no lineales y complejas entre variables. Dado que el análisis exploratorio mostró patrones no lineales en las relaciones entre las variables ambientales y los casos de Parkinson, XGBoost resultó adecuado para modelar este tipo de datos.

     Aunque es ampliamente utilizado en tareas de clasificación, XGBoost también dispone de una versión para regresión (XGBRegressor), que fue la empleada en este trabajo, ya que la variable objetivo (número estimado de casos de Parkinson) es de tipo continuo y de recuento. Este modelo es eficaz incluso en presencia de ruido y relaciones complejas difíciles de capturar por modelos lineales.

4. **Support Vector Regression (SVR)**

   - **Motivo y aplicación:** El modelo SVR es adecuado para capturar relaciones complejas entre variables, incluso cuando estas no siguen patrones lineales. Esto se logra gracias al uso de funciones núcleo (kernel), que permiten proyectar los datos a espacios de mayor dimensión, donde las relaciones no lineales pueden ser modeladas mediante una función lineal en ese nuevo espacio.

     En este trabajo se utilizó el kernel radial (RBF), que es especialmente útil para detectar patrones no lineales suaves. A diferencia de modelos basados en transformaciones explícitas de las variables, como el GLM, el SVR incorpora la no linealidad de forma implícita a través del kernel.

Aunque el SVR se emplea comúnmente en problemas de regresión continua, puede aplicarse también en contextos de datos de recuento si la escala y la naturaleza de la variable objetivo lo permiten. En este caso, se modeló el número estimado de casos de Parkinson con buenos resultados en cuanto a precisión y generalización, empleando la implementación del modelo disponible en `scikit-learn`.

5. **K-Nearest Neighbors Regression (KNN):**

   - **Motivo y aplicación:** El modelo KNN es un algoritmo basado en instancias que realiza predicciones en función de la similitud entre observaciones. Su principal ventaja es que no requiere asumir una forma funcional específica entre las variables predictoras y la variable objetivo, lo que lo hace especialmente útil para modelar patrones locales o relaciones complejas que varían en distintas regiones del espacio de características.

     Para este estudio, se utilizó la implementación `KNeighborsRegressor` de `scikit-learn`. Debido a su sensibilidad a la escala y a la dispersión de los datos, las variables fueron estandarizadas previamente. Aunque KNN no realiza ninguna inferencia paramétrica ni captura directamente relaciones no lineales generales, sí puede adaptarse bien a estructuras no lineales locales presentes en los datos.

     Este modelo resultó útil para capturar tendencias locales en el número estimado de casos de Parkinson, particularmente en combinaciones de variables donde se observaban patrones heterogéneos o no globalmente lineales.

6. **Multi-Layer Perceptron (MLP):**

   - **Motivo y aplicación:** El MLP es una red neuronal de tipo *feedforward* que permite modelar relaciones altamente complejas y no lineales entre las variables. Gracias a su arquitectura basada en capas ocultas y funciones de activación no lineales, posee una gran capacidad de aprendizaje y es especialmente útil cuando los patrones subyacentes no pueden ser capturados adecuadamente por modelos más simples.

     Se utilizó la implementación `MLPRegressor` de `scikit-learn`. Aunque el MLP puede aprender transformaciones internas complejas, se realizó una estandarización previa de las variables

predictoras para mejorar la estabilidad numérica y acelerar la convergencia del modelo durante el entrenamiento.

Este modelo fue capaz de capturar interacciones no evidentes entre variables y mostró un buen rendimiento predictivo. No obstante, su principal desventaja radica en la menor interpretabilidad en comparación con modelos lineales o basados en reglas.

G.2. Configuración y parametrización de las técnicas.

Con el fin de garantizar un ajuste óptimo de los modelos predictivos empleados en el estudio, se llevó a cabo un proceso de configuración y ajuste de hiperparámetros específico para cada técnica. Esta etapa es fundamental, ya que permite optimizar el rendimiento de cada modelo en función de las características de los datos.

G.2.1. Generalized Linear Model (GLM - Binomial Negativo)

Para el ajuste del GLM, se utilizó la librería statsmodels, empleando la familia Binomial Negativa (NegativeBinomial), adecuada para variables de recuento con sobredispersión, es decir, cuando la varianza excede a la media.

La tabla G.1 resume la configuración y parametrización aplicada al modelo Generalized Linear Model (GLM) utilizando la distribución Binomial Negativa.

Tabla G.1: Configuración aplicada al modelo GLM (Binomial Negativa)

Aspecto	**Descripción**
Distribución elegida	Binomial Negativa
Función de enlace	Logarítmica
Estandarizacion	Aplicada
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE

G.2.2. Modelo Random Forest

A continuación se presenta la tabla G.2 con la configuración y los parámetros aplicados al modelo Random Forest utilizado en este análisis. Los

parámetros descritos incluyen configuraciones clave como el número de árboles, la profundidad de los árboles, y otras opciones relacionadas con el proceso de entrenamiento y la división de los datos. Estos parámetros fueron seleccionados con el objetivo de optimizar el rendimiento del modelo, asegurando que se capture la complejidad de los datos sin alcanzar el sobreajuste.

Tabla G.2: Configuración aplicada al modelo Random Forest

Aspecto	**Descripción**
Número de estimadores	1000
Profundidad máxima	Ningún límite (None)
Mínimo de muestras para dividir	2
Mínimo de muestras por hoja	1
Características por división	sqrt
Valor de semilla	42
Estandarización	No aplicada (Random Forest no lo requiere)
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE, R^2

Descripción de los campos:

- **Número de estimadores (`n_estimators`):** Este parámetro determina cuántos árboles se van a construir en el modelo. Un número mayor de árboles aumenta la precisión y la estabilidad del modelo, ya que cada árbol contribuye a reducir el error total. Sin embargo, un número muy alto también puede aumentar el tiempo de entrenamiento.
- **Profundidad máxima (max_depth):** Define la profundidad máxima de cada árbol. Es decir, cuántos niveles puede tener el árbol desde la raíz hasta las hojas. Una profundidad mayor permite capturar relaciones más complejas entre las características, pero si es demasiado alta, puede llevar a un sobreajuste (overfitting).
- **Mínimo de muestras para dividir (min_samples_split):** Este parámetro especifica el número mínimo de muestras requeridas para

dividir un nodo. Si se establece un valor alto, el modelo se vuelve más conservador y evita crear divisiones en nodos con pocos datos. Esto puede ayudar a reducir el sobreajuste, aunque a su vez limita la capacidad del modelo para aprender patrones más complejos.

- **Mínimo de muestras por hoja (min_samples_leaf):** Controla el número mínimo de muestras que debe haber en un nodo hoja. Este parámetro es importante porque asegura que las hojas del árbol contengan una cantidad significativa de datos, lo que ayuda a evitar que el modelo aprenda demasiado de los ruidos o las fluctuaciones pequeñas de los datos.
- **Características por división (max_features):** Este parámetro controla cuántas características se consideran para la división en cada nodo. Si se utiliza una fracción más pequeña de las características, se introduce mayor aleatoriedad, lo que puede ayudar a reducir el sobreajuste y hacer el modelo más robusto. Usar todas las características puede llevar a un modelo más específico para los datos de entrenamiento, pero puede ser propenso a sobreajustarse.
- **Valor de semilla (random_state):** Se utiliza para fijar la aleatoriedad del modelo, garantizando que los resultados sean reproducibles. Si no se establece un valor de semilla, los resultados pueden variar en cada ejecución debido a la selección aleatoria de muestras y características.
- **Estandarización:** No es necesaria, ya que Random Forest no depende de las escalas de las variables.
- **División de datos:** Se utiliza un 80 % de los datos para entrenamiento y un 20 % para prueba.
- **Evaluación:** Se emplean métricas como RMSE, MAE y R^2 para evaluar el rendimiento del modelo.

G.2.3. Modelo XGBoost

En la Tabla G.3 se detallan los principales hiperparámetros utilizados para entrenar el modelo XGBoost. Estos valores fueron seleccionados con el objetivo de optimizar el rendimiento del modelo y evitar el sobreajuste. Cabe destacar que la división de los datos se realizó en un 80 % para entrenamiento y un 20 % para prueba, y que el subsampling corresponde a una técnica interna de XGBoost que selecciona aleatoriamente una fracción del conjunto de entrenamiento en cada iteración para mejorar la generalización.

Tabla G.3: Configuración aplicada al modelo XGBoost

Aspecto	**Descripción**
Número de estimadores	1000
Tasa de aprendizaje	0.05
Profundidad máxima	7
Peso mínimo por hoja	5
Submuestreo	80 % de los datos por árbol
Proporción de características por árbol	100 % (colsample_bytree = 1.0)
Valor de semilla	42
Estandarización	No aplicada (escalado interno)
División de datos	80 % entrenamiento / 20 % prueba
Evaluación	RMSE, MAE, R^2

Descripción de los campos:

- **Tasa de aprendizaje (learning_rate):** La tasa de aprendizaje controla cuánto cambia el modelo con cada árbol. Un valor bajo (como 0.05) hace que el modelo aprenda más lentamente, lo que ayuda a evitar el sobreajuste y mejora la generalización. Sin embargo, valores bajos también requieren más árboles para alcanzar un buen rendimiento.
- **Peso mínimo por hoja (min_child_weight):** Indica el número mínimo de instancias que deben estar en una hoja del árbol. Un valor de 5 asegura que un nodo hoja contenga una cantidad significativa de datos, lo que previene que el modelo se ajuste a pequeñas fluctuaciones o ruidos en los datos. Si se establece demasiado bajo, el modelo podría aprender patrones no representativos.
- **Submuestreo (subsample):** Especifica el porcentaje de datos que se utilizarán para entrenar cada árbol. Con un valor del 80 %, solo una parte del conjunto de entrenamiento se usa para cada árbol. Este submuestreo introduce variabilidad en el modelo y ayuda a prevenir el sobreajuste, ya que no todos los datos se usan en cada árbol.
- **Proporción de características por árbol (colsample_bytree)**:Controla la fracción de las características que se usan para entrenar cada árbol.

Con un valor del 100 % , el modelo utiliza todas las características disponibles en cada árbol. Si se reduce este valor, se puede introducir más aleatoriedad y reducir el riesgo de sobreajuste.

- El número de estimadores, la profundidad máxima, el valor de semilla, el mínimo de muestras para dividir, la estandarización, la división de datos y la evaluación tienen la misma definición que la explicada en el modelo Random Forest.

G.2.4. Modelo SVR

El modelo Support Vector Regression (SVR) requiere la configuración de varios hiperparámetros clave que impactan su rendimiento. Para optimizarlos, se utilizó GridSearchCV con validación cruzada de 5 particiones (cv=5), lo que mejora la estimación del rendimiento y ayuda a evitar el sobreajuste. La métrica de optimización empleada fue el error cuadrático medio negativo (neg_mean_squared_error), ya que esta penaliza los errores grandes, lo que favorece un modelo preciso.

Los parámetros escogidos fueron el resultado de aplicar transformaciones.(Véase La tabla G.4)

Tabla G.4: Parámetros utilizados en el modelo SVR con variables transformadas

Parámetro	**Valor**
C	1000
ϵ	1
γ	1
`kernel`	`rbf`

A continuación, se describen los efectos generales de cada parámetro y cómo influyen en el comportamiento del modelo:

- **C (parámetro de regularización):** Este parámetro controla el equilibrio entre la maximización del margen y la minimización de los errores de predicción. Un valor alto de `C` penaliza fuertemente los errores, lo que permite al modelo ajustarse bien a los datos de entrenamiento. Sin embargo, un valor muy alto puede llevar al sobreajuste, ya que el modelo se adapta demasiado a los detalles específicos del conjunto de entrenamiento.

- **Epsilon:** Este parámetro define un margen de tolerancia dentro del cual los puntos de datos no afectan el modelo. Un valor pequeño de `epsilon` indica que el modelo tratará de minimizar todos los errores, incluso los pequeños, lo que puede hacer que el modelo sea más sensible y propenso al sobreajuste. Un valor mayor proporciona mayor margen de error y, por tanto, un modelo menos sensible a fluctuaciones pequeñas.

- **Gamma:** Este parámetro controla la influencia de cada punto de datos en el modelo. Un valor bajo de `gamma` implica que los puntos de datos lejanos tienen mayor influencia, mientras que un valor alto hace que solo los puntos cercanos tengan impacto, lo que puede permitir capturar relaciones complejas, pero también puede aumentar el riesgo de sobreajuste si el modelo se ajusta demasiado a las pequeñas variaciones en los datos.

- **Kernel:** El kernel define la función que transforma los datos en un espacio de mayor dimensión, lo que permite al modelo encontrar patrones no lineales. El kernel `rbf` (Radial Basis Function) es una opción común debido a su capacidad para modelar relaciones complejas entre las variables. Este kernel es especialmente útil cuando las relaciones en los datos no son lineales y requiere que el modelo aprenda patrones de forma flexible.

G.2.5. Modelo K-Nearest Neighbors (KNN)

El modelo K-Nearest Neighbors (KNN) fue optimizado mediante una búsqueda de hiperparámetros utilizando lo mismo mencionado en el modelo SVR.

Estas combinaciones de hiperparámetros fueron seleccionadas para obtener el mejor rendimiento del modelo, y los valores de **MAE** y **RMSE** se utilizaron como criterios de evaluación para la calidad de las predicciones realizadas por el modelo.

Finalmente los hiperaprámetros escogidos para el modelo fueron que se encuentran en la G.5

Los siguientes parámetros son esenciales para el funcionamiento del modelo KNN y afectan su rendimiento al determinar cómo se calculan las predicciones. A continuación, se describen los efectos generales de cada parámetro:

Tabla G.5: Parámetros utilizados en el entrenamiento del modelo KNN

Parámetro	**Valor**
n_neighbors	15
weights	'distance'
algorithm	'auto'
metric	'manhattan'

- **`n_neighbors:`** Este parámetro especifica el número de vecinos más cercanos que se considerarán al hacer una predicción. En este caso, el modelo utilizará 3 vecinos. Un valor más bajo de `n_neighbors` puede hacer que el modelo sea más sensible al ruido, mientras que un valor más alto puede hacer que el modelo sea más suave y menos sensible a los detalles locales de los datos.
- **weights:** El parámetro `weights` define la forma en que se ponderan los vecinos en la predicción. El valor `'distance'` significa que los vecinos más cercanos tendrán más peso en la predicción, lo que puede mejorar la precisión en áreas donde los puntos de datos son más densos.
- **algorithm:** Este parámetro especifica el algoritmo utilizado para calcular los vecinos más cercanos. El valor `'auto'` permite que el modelo elija automáticamente el algoritmo más adecuado según el número de muestras y las características del conjunto de datos. Los algoritmos disponibles son `'ball_tree'`, `'kd_tree'`, `'brute'`, y `'auto'`.
- **metric:** El parámetro `metric` define la métrica de distancia utilizada para calcular la proximidad entre los puntos de datos. En este caso, se utiliza `'manhattan'`, que mide la distancia entre puntos sumando las diferencias absolutas de sus coordenadas. Esta métrica es útil cuando los datos tienen características discretas o si los patrones de distancia tienen una forma lineal.

G.2.6. Modelo Perceptrón Multicapa (MLPRegressor)

El modelo **Perceptrón Multicapa** (MLPRegressor) fue optimizado mediante una búsqueda de hiperparámetros al igual que en los modelos anteriores.

Aunque la combinación los hiperparámetros resultantes proporcionó los mejores resultados, el **MSE** seguía siendo relativamente alto, lo que

indicaba que, a pesar de los esfuerzos de optimización, el modelo no logró una predicción precisa. En este punto, se exploraron alternativas para la búsqueda de otros hiperparámetros que minimizaran el MSE (*RandomizedSearchCV*).

Debido al timepo de cómputo necesario para la obtención de hiperparámetros que produjeran mejores resultados se optó por mantener los mejores hiperparámetros obtenidos de la combinación de variables y ajustar manualmente algunos parámetros para observar su impacto en la reducción del error. Este proceso se repitió hasta minimizafr el máximo posible el error.

Al final, los parámetros finales seleccionados para el modelo fueron los que proporcionaron el **mejor rendimiento** sin necesidad de un proceso de búsqueda exhaustiva debido al tiempo de cómputo elevado.(Tabla G.6)

Tabla G.6: Parámetros utilizados en el entrenamiento del modelo MLP

Parámetro	**Valor**
`hidden_layer_sizes`	(256, 128)
`activation`	`relu`
`max_iter`	10000
`alpha`	0.01
`random_state`	42

A continuación, se describen los efectos generales de cada parámetro:

- **hidden_layer_sizes:** Este parámetro define la arquitectura de las capas ocultas de la red neuronal. En este caso, tiene dos capas ocultas, una con 256 neuronas y otra con 128. El número y el tamaño de las capas ocultas afectan directamente la capacidad del modelo para aprender representaciones complejas de los datos. Un mayor número de neuronas o capas permite que el modelo capture patrones más complejos, pero también puede aumentar el riesgo de sobreajuste si no se ajusta adecuadamente.

- **activation:** El parámetro de activación define la función utilizada en las neuronas de las capas ocultas. En este caso, se utiliza `ReLU` (Rectified Linear Unit), que es una de las funciones de activación más comunes y eficientes. La función ReLU introduce no linealidad en el modelo, permitiendo que aprenda representaciones complejas de los datos. Además, es menos propensa a problemas de desvanecimiento del gradiente en redes profundas, lo que facilita el entrenamiento de redes grandes.

- **max_iter:** Este parámetro establece el número máximo de iteraciones (o épocas) para entrenar el modelo. En este caso, se fijó en 10,000. A mayor número de iteraciones, el modelo tiene más oportunidades para aprender de los datos, lo que puede mejorar el rendimiento. Sin embargo, un número demasiado alto puede llevar a un tiempo de entrenamiento innecesariamente largo, especialmente si el modelo ya ha convergido.

- **alpha:** El parámetro `alpha` controla la regularización L2, que es una técnica para prevenir el sobreajuste penalizando los pesos grandes. Un valor pequeño de `alpha` significa que la regularización tiene menos impacto, permitiendo que el modelo se ajuste más estrechamente a los datos de entrenamiento. Un valor más grande aumenta la regularización, lo que puede ayudar a generalizar mejor el modelo, pero puede reducir su capacidad para ajustarse a los detalles específicos del conjunto de entrenamiento.

- **random_state:** Este parámetro se utiliza para establecer la semilla aleatoria para la inicialización de los pesos y la división de los datos. Fijar un valor para `random_state` asegura que los resultados sean reproducibles. Si no se establece, cada ejecución del modelo puede resultar en diferentes configuraciones, lo que puede afectar la consistencia de los resultados.

G.3. Detalle de resultados.

Apéndice H

Anexo de sostenibilización curricular

H.1. Introducción

Este anexo incluirá una reflexión personal del alumnado sobre los aspectos de la sostenibilidad que se abordan en el trabajo. Se pueden incluir tantas subsecciones como sean necesarias con la intención de explicar las competencias de sostenibilidad adquiridas durante el alumnado y aplicadas al Trabajo de Fin de Grado.

Más información en el documento de la CRUE https://www.crue.org/wp-content/uploads/2020/02/Directrices_Sosteniblidad_Crue2012.pdf.

Este anexo tendrá una extensión comprendida entre 600 y 800 palabras.

Bibliografía

[ken, 2022] (2022). Parkinson's disease and camp lejeune contaminated water claims. *Ken Allen Law*.

[inf, 2023] (2023). Sustancia química que permanece en el agua puede aumentar un 70 *InfoSalus*.

[America, 2023] America, P. N. (2023). Los pesticidas y el cambio climático: Un círculo vicioso. Accedido: 2025-04-10.

[Instituto Nacional sobre el Envejecimiento (NIA), 2022] Instituto Nacional sobre el Envejecimiento (NIA) (2022). La enfermedad de parkinson: causas, síntomas y tratamientos. Consultado el 10 de abril de 2025.

[Kirrane et al., 2015] Kirrane, E. F., Bowman, C., Davis, J. A., Hoppin, J. A., Blair, A., Chen, H., Patel, M. M., Sandler, D. P., Tanner, C. M., Vinikoor-Imler, L., et al. (2015). Associations of ozone and pm2.5 concentrations with parkinson's disease among participants in the agricultural health study. *Journal of Occupational and Environmental Medicine*, 57(5):509–517.

[Pacheco Moisés et al., 2011] Pacheco Moisés, F. P. et al. (2011). Toxicidad de plaguicidas y su asociación con la enfermedad de parkinson. *Archivos de neurociencias*, 16(1):33–39.

[Pearce et al., 2013] Pearce, N. et al. (2013). Paraquat and parkinson's disease: A systematic review and meta-analysis of observational studies. *Environmental Health Perspectives*, 121(5):704–709.

[Pyatha et al., 2022] Pyatha, S., Kim, H., Lee, D., and Kim, K. (2022). Association between heavy metal exposure and parkinson's disease: A review of the mechanisms related to oxidative stress. *Antioxidants*, 11(12):2467.

[Starks et al., 2013] Starks, Z. et al. (2013). Pesticide exposure and parkinson's disease: The potential role of environmental factors. *Journal of Clinical Neuroscience*, 20(6):794–799.

[Tanner et al., 2011] Tanner, C. M. et al. (2011). Pesticide exposure and parkinson's disease: A review of the literature. *Environmental Health Perspectives*, 119(6):823–827.