



Trabajo Visualización de Datos

Máster en Ingeniería y Ciencia de Datos (UNED)

Jorge Pablo Ávila Gómez

Enlaces importantes:

Link a la visualización:

<https://jorgeavilag.github.io/vespa-velutina-trabajo-visualizacion-datos/>


Repositorio del trabajo en github:

<https://github.com/JorgeAvilaG/vespa-velutina-trabajo-visualizacion-datos>

Fuente original de los datos:

Vespa Velutina * Kopuru

Las especies exóticas invasoras (EEI) constituyen la segunda causa de

 <https://kopuru.com/desafio/vespa-velutina/>



Introducción a la fuente de datos y objetivo de la visualización

Para este trabajo se ha decidido usar los datos de la competición de Kopura sobre la Vespa Velutina (Avispa Asiática). Es una competición que consiste en la predicción sobre la cantidad de nidos que se van a retirar en cada municipio de Bizkaia en 2020.

Datos: <https://kopuru.com/desafio/vespa-velutina/>

El avispón asiático o avispa negra se introdujo en Europa sobre el año 2004, se considera una especie exótica invasora, con importantes consecuencias tanto ecológicas, como económicas. Esta competición tiene como objetivo entender mejor como afecta esta especie a los hábitats, que consecuencias tiene y poder predecir la evolución de su población en Bizkaia. Se considera interesante este tema por ser un problema de actualidad, donde se van a poder aplicar los conocimientos adquiridos en la asignatura.

Los datos constan de:

- Datos meteorológicos. Diversos csv con datos de diferentes estaciones meteorológicas de 2016 a 2019.
- Datos sobre la distribución y cantidad de colmenas.
- Datos sobre la distribución de árboles frutales.
- Datos sobre la distribución de la masa forestal de la zona, con datos históricos de 2016 a 2020.
- Datos históricos sobre la distribución de los nidos de avispas desde 2017.

El objetivo principal de la visualización será estudiar diferentes parámetros que afectan a la distribución y proliferación de la avispa asiática. Entre los diferentes aspectos se pueden destacar:

- Estudio de la influencia de los tipos de vegetación que favorecen la proliferación de la avispa.
- La influencia de colmenas cercanas.
- Localizaciones geográficas donde abunda la avispa asiática.
- Influencia de árboles frutales cercanos.
- Estudio de otros aspectos meteorológicos que puedan ser de interés en el ciclo de vida de la avispa.

- Análisis históricos de los datos para encontrar posibles patrones, como, por ejemplo, mayor proliferación de la avispa si en meses anteriores hubo o no grandes lluvias.

La variedad de los datos va a permitir utilizar un gran número de gráficos diferentes, intentando utilizar siempre los tipos más adecuados. Unas de las visualizaciones más interesantes será el uso de mapas para representar datos que dependa de su geolocalización.

Localización de los documentos del trabajo y la visualización

Todos los datos usados en el trabajo y los diferentes notebooks empleados para el tratamiento de los datos y generación de las visualizaciones se encuentran accesibles en un repositorio de github público. Se puede acceder para su descarga mediante el siguiente enlace:

<https://github.com/JorgeAvilaG/vespa-velutina-trabajo-visualizacion-datos>

La visualización es accesible a través de internet con el enlace:

<https://jorgeavilag.github.io/vespa-velutina-trabajo-visualizacion-datos/>

Se ha empleado la funcionalidad de pages de github que simplifica la generación de páginas webs de los proyectos almacenados en github.

Para ello se ha generado un archivo de tipo markdown con todas las gráficas e información que se desea que aparezca en la visualización, admitiendo también código en html. Luego este archivo markdown es renderizado por github en una página web funcional, sencilla y elegante.

Preprocesamiento de los datos

Las principales tareas de preprocesado se han realizado en el notebook preprocessing.ipynb dentro de la carpeta notebooks. El lenguaje de programación utilizado es Python, usando las estructuras del módulo pandas, principalmente los dataframes.

En este notebook se realiza una pequeña exploración de los datos, pero su principal función es adaptar, limpiar y transformar las variables de manera que se puedan explorar mejor los datos y realizar visualizaciones útiles.

Entre las diferentes tareas que se realizan tenemos tareas de limpieza, como la eliminación de columnas con información duplicada en euskera, se actualizan los

nombres de las columnas a inglés y se corrigen tipos de variables como pasar a tipo datetime la variable con la información de la fecha.

Utilizando la columna con la fecha se realizaron tareas de generación de nuevas variables, como el cálculo del número de la semana que corresponde la fecha, o el día de la semana, de lunes a domingo en número.

```
In [9]: wasp_hives["date_close"] = pd.to_datetime(wasp_hives["date_close"], errors="coerce")
wasp_hives["date_close_day"] = wasp_hives["date_close"].dt.day
wasp_hives["date_close_week"] = wasp_hives["date_close"].dt.isocalendar().week
wasp_hives["date_close_month"] = wasp_hives["date_close"].dt.month
wasp_hives["date_close_year"] = wasp_hives["date_close"].dt.year
wasp_hives["date_close_day_of_year"] = wasp_hives["date_close"].dt.dayofyear
wasp_hives["date_close_day_of_week"] = wasp_hives["date_close"].dt.dayofweek
```

Un apartado importante en el preprocesamiento de los datos fue la transformación de las variables de localización, northing y easting, a longitud y latitud que son las unidades más usadas y las que se utilizan en los principales paquetes para visualizar localizaciones.

Para ello, hubo que estudiar como funcionaba ese sistema de coordenadas el cual divide el globo terráqueo en zonas. En nuestro caso se encontró que la zona donde está el País Vasco es la zona 30, y utilizando la funcionalidad de la siguiente web, pudimos transformar las coordenadas en latitud y longitud.

https://www.engineeringtoolbox.com/utm-latitude-longitude-d_1370.html

Se tuvo que extraer y transformar la información necesaria con el formato adecuado para poder introducirla en la web, y que los datos obtenidos se lean en el notebook y se añadan de nuevo al dataframe principal.

```
coordinates = pd.read_csv("../data/external/beehives_coordinates.csv")[1:-1]
coordinates
```

	northing	easting	zone	longitude	latitude
1	4777436.519	532124.0816	30.0	-2.604924	43.148996
2	4779715.773	531000.5783	30.0	-2.618614	43.169567
3	4779313.024	532159.4404	30.0	-2.604380	43.165892
4	4777072.989	532457.6901	30.0	-2.600843	43.145708
5	4777964.231	531388.7069	30.0	-2.613938	43.153779
...
6786	NaN	NaN	30.0	NaN	NaN
6787	NaN	NaN	30.0	NaN	NaN
6788	4794126.682	504669.9582	30.0	-2.942425	43.299954
6789	NaN	NaN	30.0	NaN	NaN
6790	NaN	NaN	30.0	NaN	NaN

Otra transformación de variable importante fue pasar de coma a punto en algunas columnas con información numérica. El problema es que los datos en español usa la coma como separador decimal y esto hay que corregirlo, ya que la mayoría de los módulos están escritos de manera que entienden que el separador decimal es el punto.

Para poder unificar información de diferentes archivos era importante tener una clave que permita unir los datos de diferentes fuentes. Se decidió utilizar diferentes columnas que pudiesen servir como claves a la hora de unir la información. Por un lado, tenemos los códigos postales, y el código del pueblo, este último no se conoce bien que significa, pero aparecía en diversos archivos y sirve para identificar a los diferentes pueblos. Por el otro lado, había archivos que no incluían ninguno de los dos códigos, para ello se escribió una función que normalizaba los nombres de los pueblos, (es decir, quitaba mayúsculas, espacios y otros caracteres), muy útil si aparecía el mismo pueblo en dos archivos diferentes, pero con algunas variaciones, como por ejemplo había un caso en el que estaba escrito entero en mayúsculas. Además, se añadió una columna final con el nombre de los pueblos escrito de manera estándar y legible, es decir con la primera letra en mayúsculas y el resto en minúsculas. Esta columna será útil a la hora de crear visualizaciones en las que aparezca el nombre de los pueblos, ya que quedará más legible y elegante.

Por último, se ha procesado información incluida en otros documentos, en particular, datos sobre los diferentes tipos de cultivos y en que pueblos se encuentran, y datos sobre la localización de las colmenas de abejas. Toda esta información fue

procesada, manipulada y transformada de manera que tuviese un formato semejante al resto de la información que se tiene y de esta forma poder unirla con facilidad y explorar relaciones entre ellas.

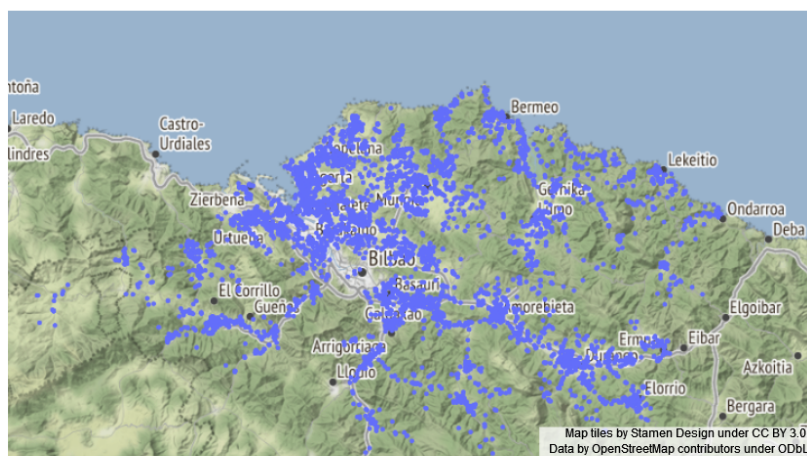
Al final del notebook se pueden encontrar algunas celdas con otras tareas como filtering o casting, y finalmente, los datos obtenidos se guarda en formato comprimido para poder ser accedidos y utilizados desde otros medios.

Procesamiento y análisis

Las tareas de procesamiento y análisis se han desarrollado en un notebook independiente denominado `processing.ipynb`. En este notebook se ha seguido utilizando el módulo Pandas para la manipulación de los datos utilizando principalmente los dataframes. También se han utilizado los modulos altair, para las representaciones en general, y plotly para realizar representaciones en mapas.

Altair es una librería para realizar visualizaciones que utiliza una gramática de programación de tipo declarativo, basada en Vega y Vega-Lite. Permite obtener de manera sencilla visualizaciones atractivas, configurables fácilmente y con posibilidad de incluir interactividad.

Plotly in en especial Plotly Express es otra librería de python para realizar visualizaciones. En este caso, Plotly Express es una API de alto nivel que permite obtener visualizaciones muy atractivas. En particular, ha interesado la funcionalidad para obtener mapas interactivos con localizaciones de lugares. Con otras librerías usar un mapa a color puede no ser gratis o complejo de configurar, con esta librería se han podido obtener mapas interactivos funcionales sin demasiados problemas.



Durante el proceso de análisis de los datos el notebook ha pasado por muchos ciclos de reescritura en el que ha ido avanzando y evolucionando a su forma final donde se concentra el código que genera las visualizaciones finales. Aun así, durante el proceso de evolución se han explorado otras muchas posibles visualizaciones y tipos de gráficos que han servido para conocer mejor los datos, pero o no se han incluido en el documento final o simplemente no se han considerado necesarias a la hora de exponer la visualización de problema.

Durante el procesamiento se ha estudiado si existe algún tipo de evolución temporal de los datos, se ha encontrado que hay una dependencia a lo largo del año y se ha decidido exponerla con los casos que aparecen mes a mes. Otras visualizaciones temporales como los nidos recogidos cada día de la semana no se han considerado porque realmente no aportan información acerca de la distribución y comportamiento de la avispa asiática.

También se ha estudiado la preferencia de las avispas en tener nidos en zonas urbanas o rurales, y también mediante mapas topológicos la distribución espacial de los nidos. Tenemos mapas en los que aparecen todos los nidos recogidos, también animaciones con los nidos recogidos que pertenecen a cada pueblo, o una animación temporal que va marcando los nidos que se han recogido día a día.

Por otro lado, se ha estudiado la relación entre la presencia de nidos de avispas con colmenas de abejas o la presencia de diferentes tipos de cultivos. Además, en los casos más interesantes se ha estudiado el mapa 2D de las distribuciones de estas relaciones.

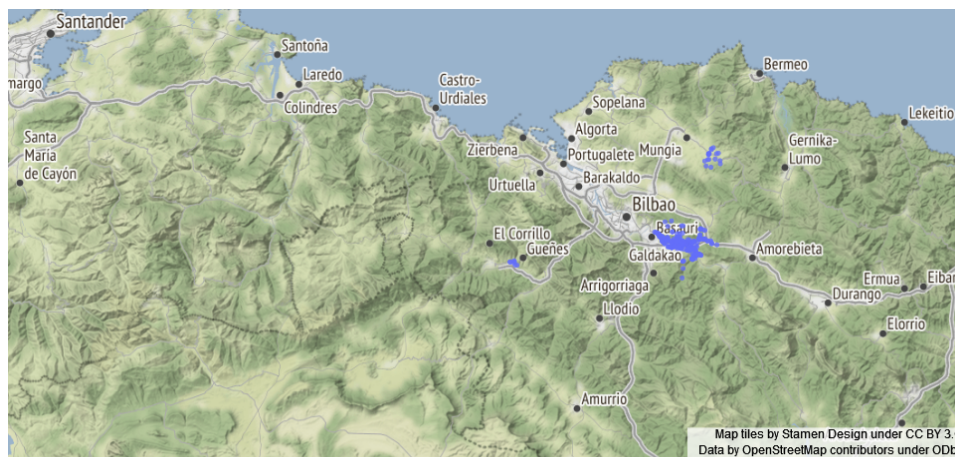
Análisis y comentarios sobre la visualización

En la visualización desarrollada en este trabajo se ha intentado contestar principalmente a la pregunta de "¿Cómo se distribuye la avispa asiática en el País Vasco?". Para ello se han usado diferentes gráficos.

La primera parte consta de dos mapas de la zona del País Vasco en cuestión donde aparecen todos los nidos de avispa asiática que se han recogido. En el primero aparecen toda la información junta y en el segundo se puede explorar la distribución de los nidos por pueblos. Gracias a que la representación es dinámica se puede hacer zoom y mover el mapa para explorar las zonas de interés de cada uno, y además, si se pasa el ratón por encima de los puntos se obtiene información complementaria como el nombre de la localidad en la que se encuentra.

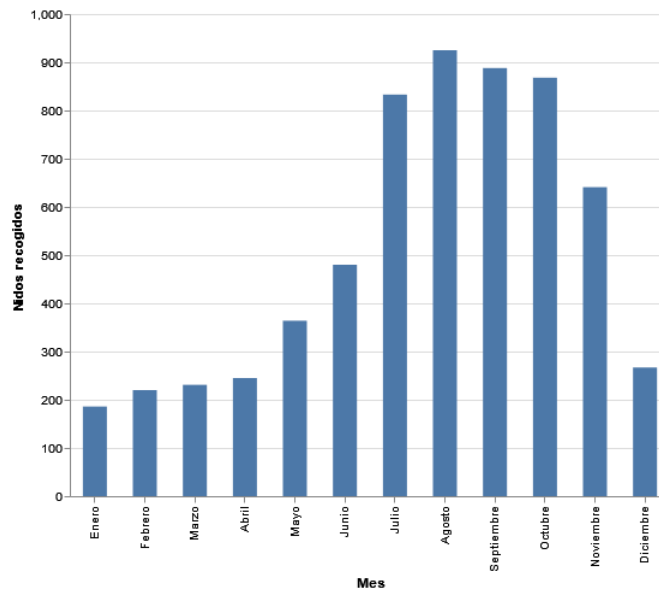
Explorando los datos se puede observar que las localizaciones presenta fallos y hay pueblos que parecen que están mezclados, de todas formas no podemos estar

seguro de cuál sería las localizaciones verdaderas, ya que son errores del etiquetado original de los datos.



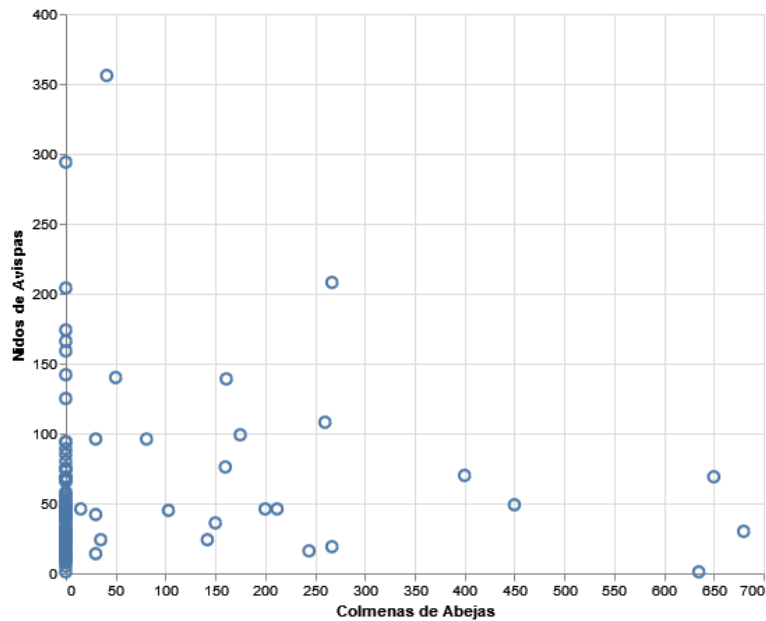
Se continúa, en este caso explorando la distribución temporal de los nidos. Se utiliza una gráfica de barras para ver la cantidad de nidos que se recogen cada vez y se ve que hay diferencias según la época del año. Estas diferencias deben estar muy relacionadas con el ciclo de vida de la avispa asiática, la cual tiene preferencia para crear nuevos nidos y reproducirse en verano, mientras que en invierno permanece bajo hibernación.

Además se añade un mapa interactivo que muestra la localización de los nidos recogidos cada día de los últimos dos años. Si nos movemos a los meses de verano se puede apreciar una mayor densidad de puntos.

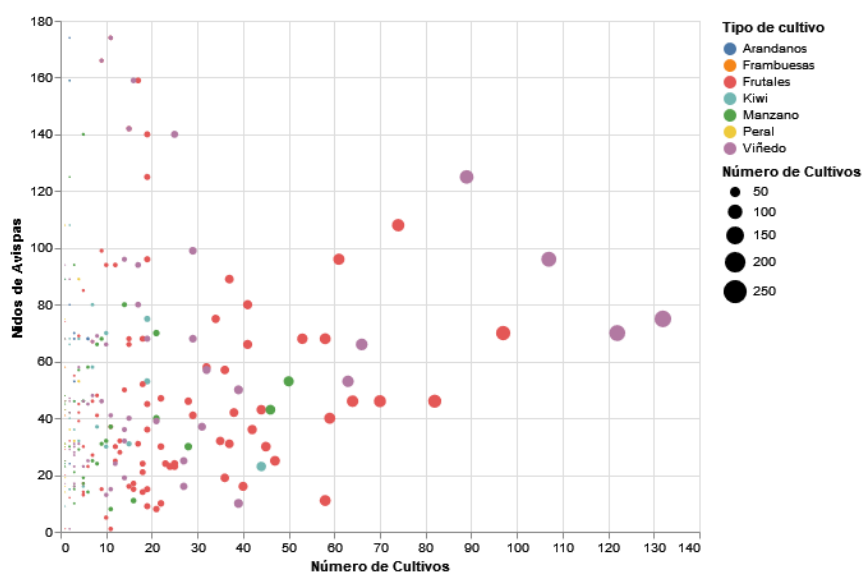


En el siguiente apartado se comenta las diferencias entre nidos recogidos en zonas urbanas o zonas rurales. Se ha utilizado un gráfico de barras para poder comparar las cantidades fácilmente, además pasando el ratón por encima se pueden ver las cantidades exactas. Es interesante observar que aunque el número total de nidos cambia de un año a otro, las proporciones se mantienen, es decir, se recoge siempre aproximadamente el mismo número de nidos en zonas urbanas que en zonas rurales. Parece que las poblaciones son estables y no se está realizando trasvase o crecimiento selectivo en alguno de los dos entornos.

Al estudiar la influencia de la presencia de panales de abejas en la cercanía de los nidos de avispas se descubrió que sorprendentemente no parece que haya una dependencia fuerte, a pesar de que las abejas sean una presa común para este tipo de avispas. En la gráfica que aparece en la visualización, cada pueblo es representado por un punto, y se observa claramente que en pueblos donde se han recogido un número grande de nidos, a veces tienen muy pocos panales o incluso ninguno.



Por último, se estudió si la presencia de diferentes tipos de cultivos podría influir en la abundancia de avispas asiáticas en la zona. En la gráfica presentada se puede apreciar una leve tendencia lineal positiva. En el que un mayor número de cultivo se relaciona con un mayor número de nidos recogidos en ese pueblo. Además, los diferentes tipos de cultivos se representaron utilizando diferentes colores así podemos identificar fácilmente que los viñedos y los frutales son los más influyentes. Como apunte extra, se hizo que el tamaño del punto dependiese de la cantidad que hubiese de ese cultivo, a pesar de que se está duplicando la información, ya que es lo mismo que aparece en el eje x, le añade un carácter visual más intuitivo a la presentación.



Para terminar se añade un mapa en el que aparecen los nidos recogidos y la localización de los cultivos más influyentes, viñedos y frutales. Se puede observar que en las zonas donde hay presencia de esto estos cultivos suelen estar acompañadas de numerosos nidos de avispa asiática recogidos.

Conclusiones

En este trabajo se ha realizado un estudio sobre la distribución de la avispa asiática en el País Vasco, ilustrando los resultados obtenidos en una visualización. Por un lado, se ha tenido que tratar con un gran volumen de datos reales. La parte del preprocesado ha sido la más tediosa y larga del trabajo posiblemente por la naturaleza real de los datos, los cuales ha presentado anomalías, datos incorrectos, correcciones necesarias y un gran número de transformaciones y adaptaciones para poder comparar y analizar de manera efectiva los datos. Se ha realizado un intenso proceso de análisis para intentar comprender mejor los datos y extraer conclusiones interesantes. De los análisis se han creado una serie de gráficas lo más atractivas e informativas posibles incluyendo interactividad en la mayoría de los gráficos, gracias a las modernas librerías de Python utilizadas, siempre intentando seguir los conocimientos desarrollados en la asignatura. Finalmente se ha desarrollado una visualización, accesible a través de internet, que incluye las gráficas y conclusiones obtenidas.