

---

## Informe Final: Análisis Exploratorio de Datos

Docente: [Ana María Cuadros Valdivia](#)

---

### ANEXO

Este es el formato sugerido, puede agregar secciones pero no puede omitir las sugeridas.

#### INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS .....

##### 1. Hipótesis iniciales:

###### 1.1. Motivación:

La bicicleta compartida se ha consolidado como una alternativa sostenible para la movilidad urbana. En Madrid, el sistema BiciMAD permite a miles de usuarios desplazarse diariamente. Sin embargo, a pesar de su éxito, existen indicios de desigualdad en la cobertura y uso del sistema, especialmente cuando se analizan los viajes según la hora del día, el día de la semana y el tipo de entorno urbano donde comienzan.[1]

Actualmente, la toma de decisiones sobre dónde instalar estaciones o redistribuir bicicletas no siempre se basa en datos espacio-temporales integrados con el entorno urbano real. Este vacío limita la eficacia del servicio y puede afectar negativamente la equidad territorial, dejando barrios residenciales o zonas periféricas con menor acceso o cobertura en horas críticas.[2]

Este proyecto se enfoca en utilizar técnicas de Visual Analytics espacio-temporal para detectar estas desigualdades con precisión, analizando viajes de BiciMAD y cruzándolos con datos oficiales del uso del suelo (SIOSE) y (HILUCS). A partir de esto, se busca generar visualizaciones interactivas y mapas de calor que permitan detectar zonas con sobreuso o infrarrepresentación del servicio en función del contexto urbano.[1][4]

###### 1.2. Exprese sus hipótesis en forma de pregunta (sea claro y conciso)

Hipótesis 1:

Los viajes realizados en bicicletas suelen usar con mayor frecuencia zonas urbanas.

Hipótesis 2:

Las estaciones de partida o llegada están relacionadas con el uso del suelo de las zonas urbanas.

Hipótesis 3:

Los viajes realizados en bicicletas suelen usar con mayor frecuencia zonas rurales.

1.3. Plan de análisis:

Describe qué pasos siguió para investigar las hipótesis.

2. **Fuente de datos:**

2.1. **Fuente:**

El archivo **trips\_febrero\_2023.csv** extraído de la base de datos BICIMAD que nos da información sobre viajes de bicicletas con sus respectivos lugares de salida y llegada, ubicación, fechas.

El archivo **suelo.csv** extraído de la base de datos **SIOSE** que se encarga de etiquetar el uso que se le da a los suelos en España, mediante coordenadas se determina el área y lugar del suelo.

2.2. **Descripción:**

Describe el conjunto de datos:

a) **A nivel de atributos:**

b) **A nivel de registros:**

En **trips\_febrero\_2023.csv**, cada fila representa un viaje realizado por una bicicleta en una fecha y hora específicas.

En **suelo.csv**, cada fila representa un polígono de tierra con una combinación específica de uso del suelo.

Ambos tienen un nivel de granularidad espacial: **trips\_febrero\_2023.csv** a nivel de punto (inicio y fin), **suelo.csv** a nivel de área.

El archivo **trips\_febrero\_2023.csv** no está etiquetado para clasificación; se puede derivar etiquetas (p. ej., zona urbana, natural, etc.) cruzando con **suelo.csv**.

c) **Relación entre atributos:**

Se puede realizar un overlay espacial entre geolocation\_unlock/lock (puntos) de **trips\_febrero\_2023.csv** y geometry de **suelo.csv** (polígonos que determinan el área y lugar que ocupa determinado suelo) para determinar qué tipo de suelo se utiliza al iniciar/finalizar un viaje.

Existe posible correlación inversa entre SELLADO y FCC (zonas urbanas vs. naturales).

A nivel de trips, se puede estudiar correlaciones entre duración (trip\_minutes) y variables como tipo de flota (fleet), estación de desbloqueo, zonas urbanas/rurales (desde suelo).

### **HILUCS (Hierarchical INSPIRE Land Use Classification System)**

Representa una clasificación jerárquica del uso del suelo basada en funciones socioeconómicas. Cada código es un número entero (ej. 500) que se traduce en categorías como 5\_ResidentialUse o 3\_1\_CommercialServices. Esta clasificación es útil para interpretar el propósito funcional del terreno en términos de actividades humanas o económicas.

- Diccionario: hilucs\_dict
- Columna relacionada: HILUCS\_CODE o extraído de SIOSE\_CODE

### **CODIIGE (Clasificación española de coberturas del suelo)**

Este sistema nacional codifica la cobertura física observada en la superficie del suelo, como bosques, carreteras, cultivos, zonas urbanas, etc. Los códigos (como 311, 140, 121) están asociados a descripciones como “Bosque de frondosas” o “Instalación agrícola”.

- Diccionario: codiige\_dict
- Columna relacionada: CODIIGE o parte del SIOSE\_CODE

### **COBERTURAS**

Representa una descripción simplificada (alfanumérica en mayúsculas) de los elementos físicos del suelo, que incluyen edificaciones (EDF), zonas verdes artificiales (ZAU), cultivos (CHL, LOL), cuerpos de agua (AUC), etc. Estas abreviaturas se encuentran frecuentemente concatenadas dentro del campo SIOSE\_CODE.

- Diccionario: COBERTURAS
- Columna relacionada: SIOSE\_CODE, SIOSE\_XML, o variable derivada

**Realice un cuadro resumen de la descripción de los atributos.**

#### **2.3. Formato:**

El archivo **trips\_febrero\_2023.csv** extraído de la base de datos **BICIMAD** que tiene la información de los viajes en bicicleta tiene los datos de esta forma:

Campo	Descripción
date	La fecha en que se realizó el viaje.
idbike	Identificador único de la bicicleta utilizada para el viaje.

fleet	Flota a la que pertenece la bicicleta utilizada. Existían dos diferentes.
trip_minutes	Duración del viaje en minutos.
geolocation_unlock	Coordenadas geográficas del punto de inicio del viaje.
address_unlock	Dirección postal donde se desbloquea ( <b>dirección inicial del viaje</b> ) la bicicleta.
unlock_date	Fecha y hora exacta en que comenzó el viaje.
locktype	Estado de la bicicleta antes del viaje ( <b>si está en la estación o en uso</b> ).
unlocktype	Estado de la bicicleta después del viaje ( <b>si está en la estación o en uso</b> ).
geolocation_lock	Coordenadas geográficas del punto final del viaje.
address_lock	Dirección postal donde se bloqueó ( <b>dirección final del viaje</b> ) la bicicleta.
lock_date	Fecha y hora exacta en que finalizó el viaje.
station_unlock	Número de estación donde estaba guardada la bicicleta antes del viaje (si aplica).
dock_unlock	Muelle de la estación donde estaba fondeada la bicicleta antes del viaje (si aplica).
unlock_station_name	Nombre de la estación de desbloqueo( <b>estación donde se recoge la bicicleta para su uso</b> ) (si aplica).
station_lock	Número de estación donde quedó anclada la bicicleta tras el viaje (si aplica).
dock_lock	Muelle de la estación donde quedó fondeada la bicicleta tras el viaje (si aplica).
lock_station_name	Nombre de la estación de bloqueo ( <b>estación donde se deja la bicicleta para su uso</b> ) (si aplica).

El archivo **suelo.csv** extraído de la base de datos **SIOSE** tiene los datos de esta forma que describe el uso que se le da a una área determinada mediante posiciones gps que llamaremos “**POLIGONO**”:

Campo	Descripción
ID_POLYGON	Identificador Universal Único del polígono (UUID), con namespace URN. Es único para cada polígono.
SIOSE_CODE	Rótulo SIOSE, según el documento “Descripción del Modelo de Datos y Rótulo SIOSE”. <b>(CODIFICA MEDIANTE NÚMERO QUE REPRESENTA PORCENTAJE DEL TOTAL DE USO DEL SUELO, ABREVIACIONES EN LETRAS QUE ES EL USO DLE SUELO Y EN MINÚSCULAS LO MISMO, PERO APRA LOS ATRIBUTOS DEL SUELO)</b>
SIOSE_XML	Información completa de las coberturas del suelo asociadas a cada polígono, junto con sus atributos, en formato XML.
SUPERF_HA	Superficie del polígono en hectáreas, con precisión de 4 decimales. Obtenida sobre la proyección UTM correspondiente a cada comunidad autónoma.
CODIIGE	Código del Consejo Directivo de la Infraestructura de Información Geográfica de España (uso del suelo en España). <b>(NÚMERO QUE CON ÉL SE PUEDE BUSCAR UNA DESCRIPCIÓN DEL SUELO ACCEDIDO EN ESPAÑA.)</b>
HILUCS	Clasificación del uso del suelo según la nomenclatura europea INSPIRE (HILUCS). <b>(NÚMERO QUE CON ÉL SE PUEDE BUSCAR UNA DESCRIPCIÓN DEL SUELO ACCEDIDO EN EUROPA Y TAMBIÉN DE FORMA INTERNACIONAL.)</b>
SELLADO	Porcentaje de superficie sellada del polígono, si tiene presencia de clases artificiales (valor entre 0 y 100%).
FCC	Fracción de cabida cubierta: porcentaje de cobertura de arbolado forestal en el polígono (valor entre 0 y 100%).
CODBLQ	Código numérico del INE correspondiente a la comunidad autónoma.

geometry	Geometría del polígono: representa el área espacial que contiene la tierra donde se usa el suelo.
----------	---

#### 2.4. Transformaciones:

El formato de las fechas de unlock\_date y lock\_date **que se refieren a las fechas de inicio del viaje en bicicleta y final del viaje en bicicleta** no estaban en el formato para usarlas en python

ANTES	DESPUÉS
2023-02-01T00:00:10	2023-02-01 00:00:10

El formato de las localizaciones geolocation\_unlock y geolocation\_lock **que se refieren a las ubicaciones GPS en la que da inicio el viaje en bicicleta y final del viaje en bicicleta** La localización de inicio y llegada estaba en un objeto que no se podía usar directamente.

El archivo **suelo.csv** extraído de la base de datos **SIOSE** tiene los datos de esta forma en la columna geometry:

ANTES	{'type': 'Point', 'coordinates': [-3.708834, 40.411274]} (EJEMPLO)			
DESPUÉS	lat_unlock	lon_unlock	lat_lock	lon_lock
	40.413280	-3.695618	40.411274	-3.708834

En el archivo que se consigue de **SIOSE un organismo en España que cataloga el uso de los pisos en ese país** tenemos la columna geometry se refiere a los puntos de polígono que al graficarlos en un mapa se verá el área correspondiente a esa área del suelo.

ANTES	<POLYGON ((468026.713 4459234.371, 468026.713 4459086.6, 468026.713 4458967....> (EJEMPLO)
-------	--

DESPUÉS	lat_unlock	lon_unlock	lat_lock	lon_lock
	40.413280	-3.695618	40.411274	-3.708834

Las columnas **CODIIGE**, **HILUCS**, **SIOSE\_CODE** tienen especificaciones del uso del suelo cada más específico que el anterior.

Rótulo SIOSE (documento “Descripción del Modelo de Datos y Rótulo SIOSE”)

CODIIGE Consejo Directivo de la Infraestructura de Información Geográfica de España

HILUCS Hierarchical INSPIRE Land Use Classification System

	ANTES	DESPUÉS
CODIIGE	112	Ensanche
HILUCS	500	5_Residential_Use
SIOSE_CODE	UEN(70EDFem_15ZAU_15VAP)	70% Edificacion con edificio entre medianeras, 15% de zona verde artificial y arbolado urbano, 15% vial aparcamiento o zona peatonal sin vegetación

## 2.5. Limpieza de datos:

El archivo **trips\_febrero\_2023.csv** extraído de la base de datos **BICIMAD** tiene los datos de esta forma:

FILA 1	FILA 2	FILA 3	FILA 4	FILA 5	.....	FILA M
DATOS	DATOS	DATOS	DATOS	DATOS	DATOS	DATOS
DATOS	DATOS	DATOS	DATOS	DATOS	DATOS	DATOS
DATOS	DATOS	DATOS	DATOS	DATOS	DATOS	DATOS

DATOS						
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
DATOS						

Después de la eliminación de las filas completamente nulas nos queda todavía en el archivo **trips\_febrero\_2023**:

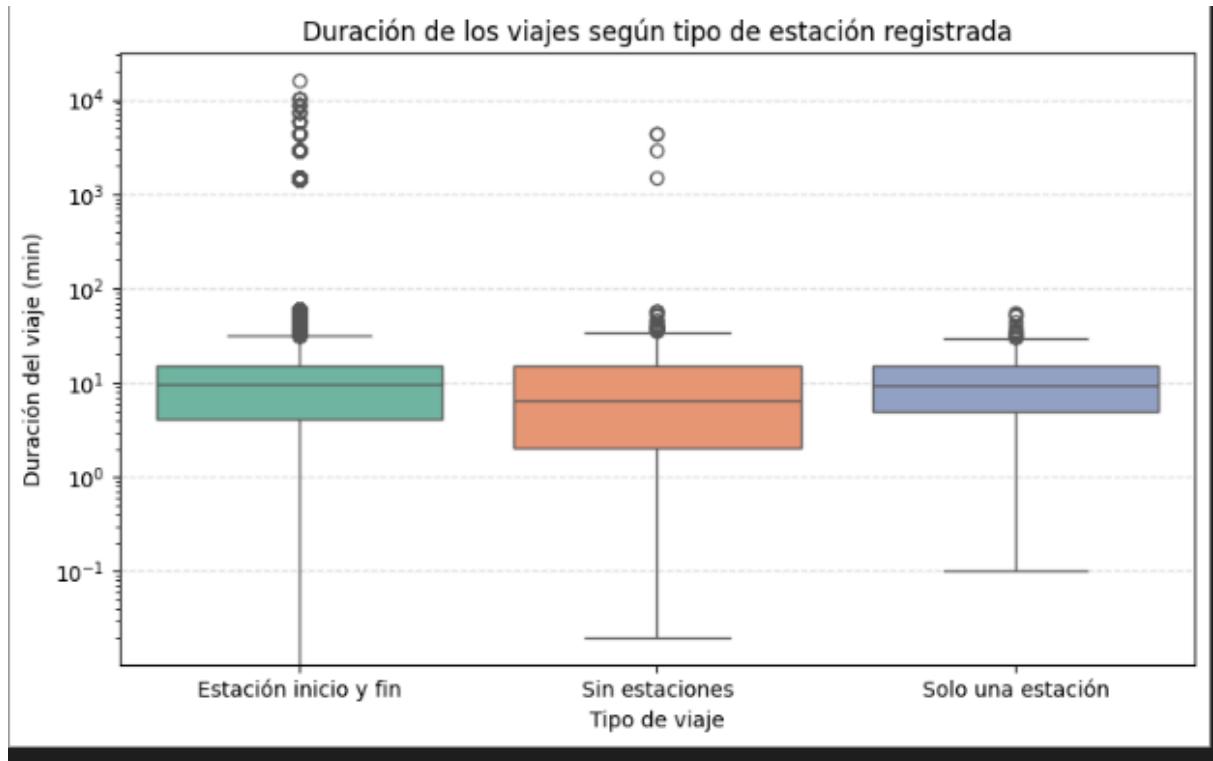
Cantidad de nulos por columna	
<b>station_unlock</b>	<b>380</b>
<b>dock_unlock</b>	<b>380</b>
<b>unlock_station_name</b>	<b>380</b>
<b>station_lock</b>	<b>476</b>
<b>dock_lock</b>	<b>476</b>
<b>lock_station_name</b>	<b>476</b>

Verificando los valores nulos de las columnas **station\_unlock**, **dock\_unlock**, **unlock\_station\_name** estan en la misma fila y las columnas **station\_lock**, **dock\_lock**, **lock\_station\_name** también.

Grupo de columnas	Nulos completos	¿Qué significa?
station_unlock, dock_unlock, unlock_station_name	380	Hay 380 viajes que no tienen información sobre la estación de inicio del viaje.
station_lock, dock_lock, lock_station_name	476	Hay 476 viajes que no tienen información sobre la estación de fin del viaje.
Ambos grupos sin datos	308	Hay 308 viajes que no tienen estación de inicio ni de fin — es decir, no tienen puntos de anclaje registrados

### 3. Exploración:

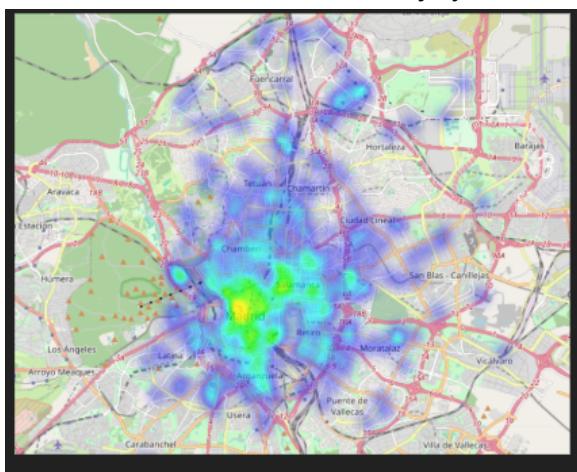
Los viajes registrados con sus estaciones de inicio a fin con lo que tiene más valores atípicos.



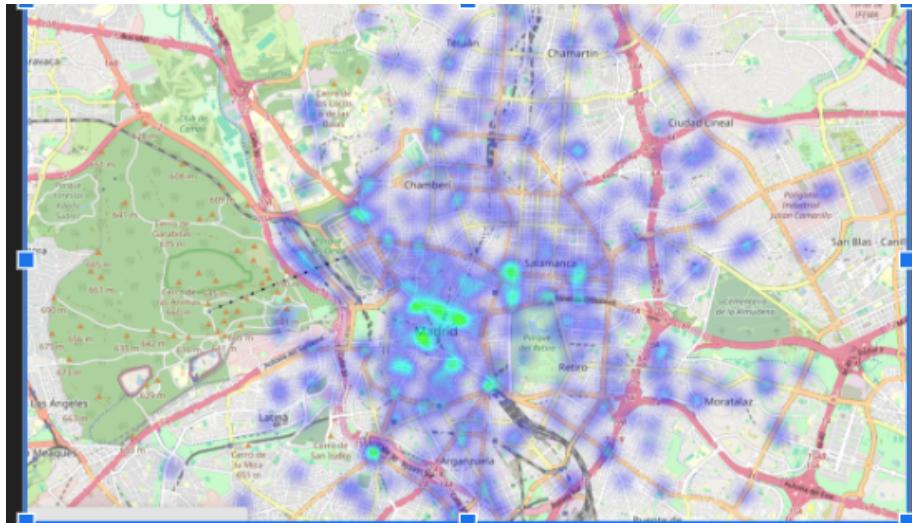
Los datos de **SIOSE\_CODE** como por ejemplo UEN(70EDFem\_15ZAU\_15VAP) explican de forma explícita la **división del suelo en porcentajes** de acuerdo a su uso los números representan el porcentaje respecto a 100, las **mayúsculas** uso en formato acortado del suelo y las letras **minúsculas** los atributos del uso del suelo.

La columna geometry sus datos los guardaba en el orden **longitud, latitud** y para poder usarlo en python debíamos darle la forma de **latitud, longitud**.

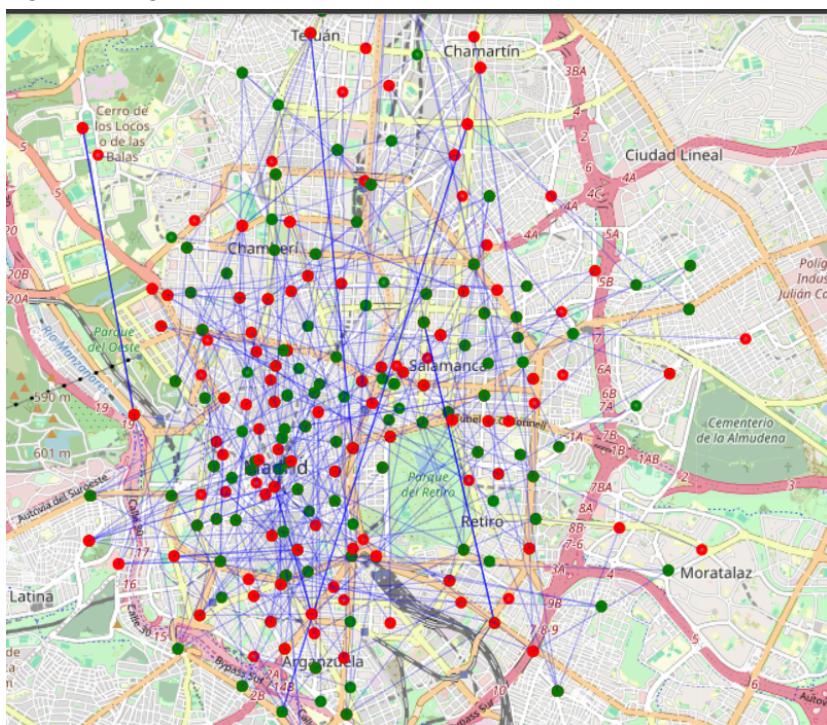
Uso en el desbloqueo(**inicio del lugar del viaje**) para el uso de las bicicletas relacionado a los minutos de viaje y un radio de 10



Bloqueo(**fin del lugar del viaje**) para la devolución de las bicicletas relacionado su posición y un radio de 10



Los primeros 500 trazo de las rutas de ida y vuelta. Verde lugar de partida y rojo lugar de llegada (devolucion).



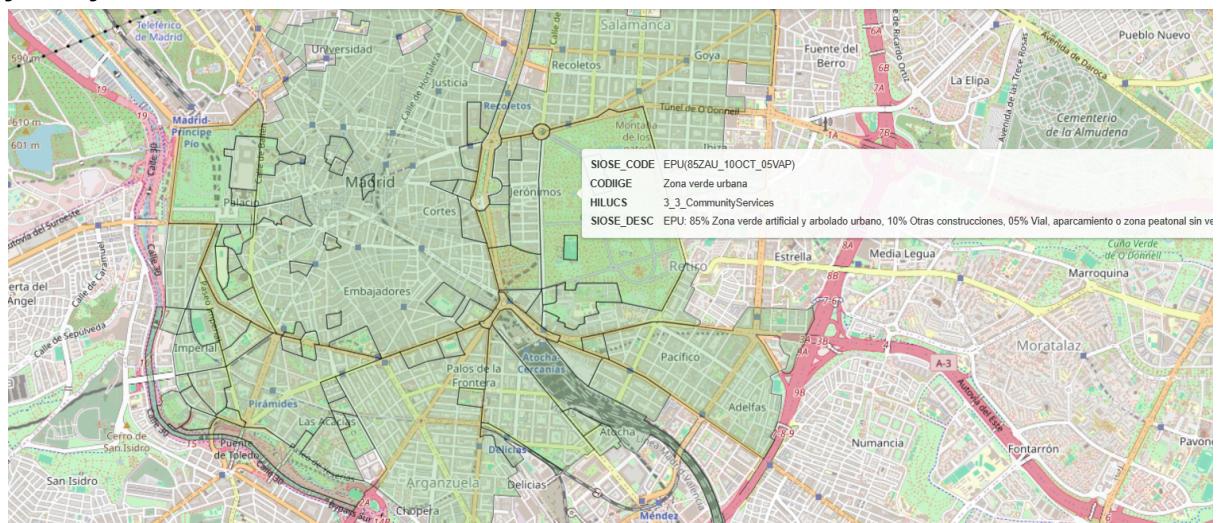
Filtramos por el orden en el que salgan los suelos y sus respectivas áreas [40.4168, -3.7038] y mostramos los 1000 primeros de los suelos dibujados y la respectiva área que ocupan.



Filtramos por zona geográfica en el centro de Madrid [40.4168, -3.7038]. Área a evaluar:

**xmin, xmax = -3.72, -3.68**

**ymin, ymax = 40.40, 40.42**

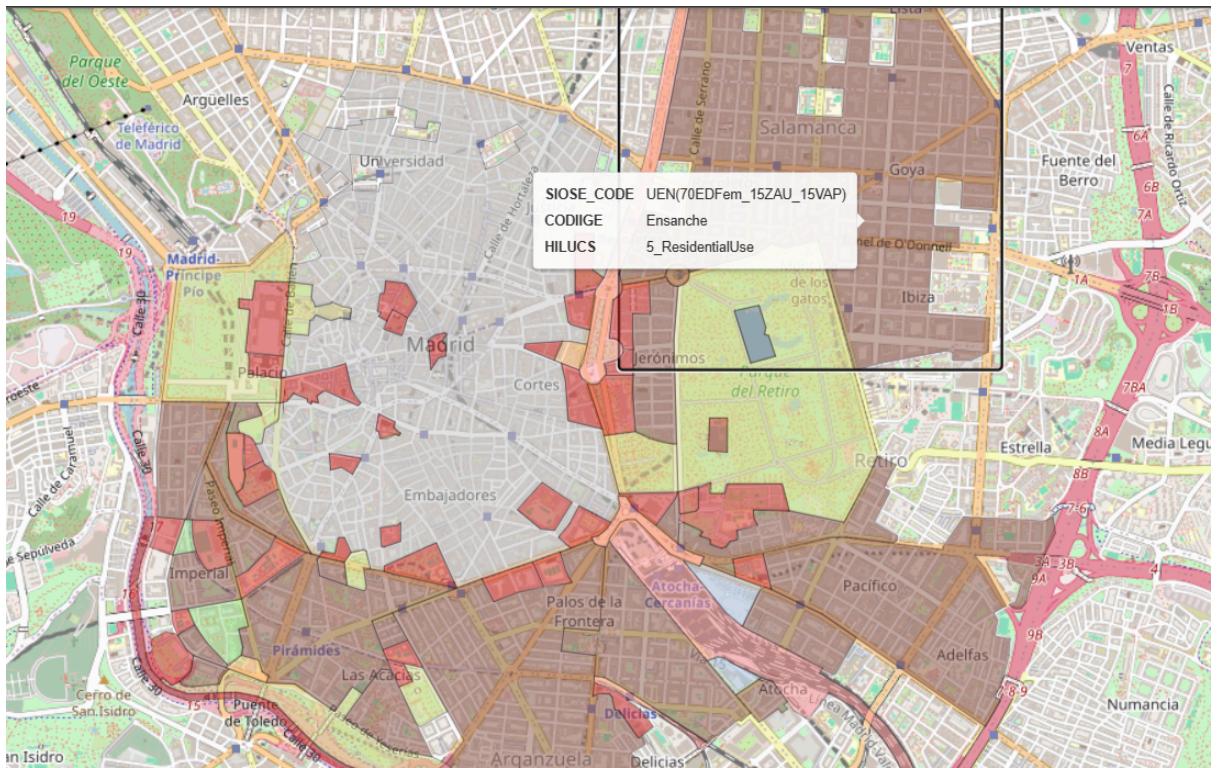


Filtramos por zona geográfica en el centro de Madrid [40.4168, -3.7038]. Área a evaluar:

**xmin, xmax = -3.72, -3.68**

**ymin, ymax = 40.40, 40.42**

Y se agrupa del mismo color tipo de suelo que usa esa área correspondiente.

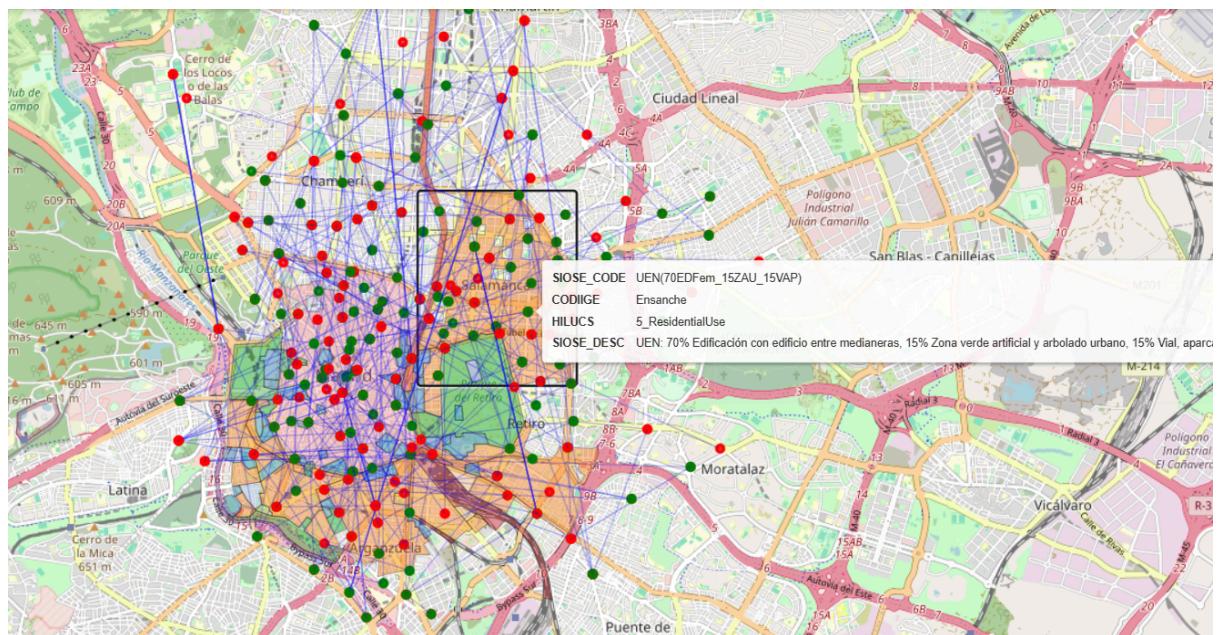


Filtramos por zona geográfica en el centro de Madrid [40.4168, -3.7038]. Área a evaluar:

**xmin, xmax = -3.72, -3.68**

**ymin, ymax = 40.40, 40.42**

Y se agrupa del mismo color tipo de suelo que usa esa área correspondiente. También los viajes que se realizan en las áreas de suelo correspondiente.



#### 4. Conclusión:

Qué conocimiento obtuvo del análisis exploratorio de datos.

Para cada hipótesis incluya sus conclusiones intermedias y finales.

- Como el suelo determina de manera decisiva del uso de rutas específicas dependiendo del lugar.
- Las zonas más usadas son las zonas urbanas, sobre todo en las zonas céntricas.
- Las zonas rurales y de edificaciones ayudan a ver asu alrededor un uso común de las zonas urbanas por los ciclistas.
- Los ciclistas que registran su estación de inicio y fin son los que en determinados casos más usan las bicicletas teniendo comportamientos atípicos.

Anexos:

- Pdf colab :  
(Etapas del ciclo de vida de Ciencia de Datos (autores) ))  
Data wrangling (autores)

Referencias

**1. Artículo científico principal (Visual Analytics + micromovilidad)**

Escribano, A., Jiménez, F., & Ruiz, M. (2023). *Uncovering spatiotemporal micromobility patterns through the lens of space-time cubes and GIS tools*. *Journal of Geographical Systems*.

<https://doi.org/10.1007/s10109-023-00418-9>

**2. Documento técnico sobre la base de datos SIOSE**

Instituto Geográfico Nacional. (2021). *Estructura y contenido de la base de datos SIOSE v3.0. Sistema de Información sobre Ocupación del Suelo en España (SIOSE)*.

[https://www.siose.es/SIOSEtheme-theme/documentos/pdf/Estruc\\_Cons\\_Bas\\_dat\\_SIOSE\\_v3.pdf](https://www.siose.es/SIOSEtheme-theme/documentos/pdf/Estruc_Cons_Bas_dat_SIOSE_v3.pdf)

**3. Lista de códigos HILUCS del estándar INSPIRE (clasificación europea del uso del suelo)**

INSPIRE. (n.d.). *HILUCSValue — Hierarchical INSPIRE Land Use Classification System*. INSPIRE Thematic Codelists.

<https://inspire.ec.europa.eu/codelist/HILUCSValue>

**4. Prestifilippo, G., Ballatore, A., et al. (2024). Visual Analytics for Sustainable Mobility: Usability Evaluation and Application in UrbanFlow Milano. *Smart Cities*, 4(4), 41. <https://doi.org/10.3390/smartcities4040041>**