# Documento en construcción

Este documento contiene algunos fragmentos de la documentación del proyecto dasolidar de la Junta de Castilla y León (elaborada para uso interno). Se refiere a contenidos que exceden lo que hemos subido a OpenCayle, (más capas y herramientas que se usan en el entorno corporativo).

No obstante, la ponemos aquí a disposición pública (de forma provisional) para facilitar la interpretación de la información colgada en OpenCayle. Está pendiente de adaptar este documento a la versión publicada en Opencayle.

Fecha prevista de adaptación: marzo de 2026

## Directorio dasoLidar

### Contenidos

Las capas dasoLidar son capas ráster de 10 m de píxel que cubren toda Castilla y León y representan métricas Lidar y variables dasométricas. Son archivos TIFF de una sola banda (un fichero por cada métrica o variable).

Cada métrica o variable tiene una definición que está vinculada a un determinado tamaño de píxel o celda[1], por lo que esta referencia espacial, los 10 x 10 m, forma parte de su definición.[2]

DasoLidar está organizado en cinco directorios:

- doc
- PNOA1_2010-2014
- PNOA2_2017-2021
- PNOA2_vs_PNOA1
- varios

Los productos derivados del PNOA2, los del PNOA1 y los obtenidos comparando métricas de ambos se han separado en directorios distintos. Además, se ha incluido un directorio de documentación, en el que está este manual (doc) y otro (varios) que incluye información auxiliar.

La comparativa PNOA2_vs_PNOA1 no debe interpretarse necesariamente como un cambio en la cubierta vegetal, es simplemente una primera aproximación consistente en restar valores PNOA2 – PNOA1 sin más. Las diferencias se deben en parte a cambios reales en la cubierta vegetal y en parte a diferencias en las características de los vuelos (sensor, configuración del sensor, altura de vuelo y, sobre todo, estado fenológico de la vegetación). El sensor, su configuración, y la altura y velocidad del vuelo condicionan la densidad, el tamaño de la huella, la penetrabilidad, la sensibilidad para identificar retornos, entre otros factores. Por lo tanto, las capas del directorio PNOA2_vs_PNOA1 son solo una pista sobre lo que puede haber pasado entre las fechas de uno y otro vuelo.

[1] Usamos estos dos términos indistintamente.

[2] Hay una excepción que es la altMax_r02m, que registra la altura del retorno más alto sobre el suelo en cada celda de 2x2 m. Como esta resolución da lugar a ficheros grandes, esta métrica se ha repartido en cinco ficheros, correspondientes a los cinco 'cuadrantes' de vuelo de Castilla y León (ver más adelante información sobre los cuadrantes de vuelo). Para reducir el tamaño de los ficheros esta variable está almacenada en centímetros (int 16 bit en lugar de float 32 bit; cada fichero ronda los 4-6 GB).

## Métricas Lidar (ML) y variables dasométricas (VD)

*Aclaraciones previas*

Cuando hablamos de capas con **métricas Lidar**, nos referimos a capas ráster con valores generados exclusivamente a partir de las nubes de puntos Lidar. En cambio, las capas con **variables dasométricas** se estiman con apoyo de datos de parcelas (IFN en el caso de dasolidar) y hacen referencia a variables que se pueden medir/estimar en campo.

---

*Las métricas Lidar no requieren parcelas de campo, mientras que las variables dasométricas sí.*

---

***Métricas Lidar***

Algunas métricas lidar tienen una correspondencia muy estrecha con variables dasométricas y, de hecho, son magnitudes más precisas y objetivas que las variables dasométricas a las que apuntan. Podemos considerarlas proxy de esas variables dasométricas y usarlas en sustitución de ellas, pero teniendo en cuenta no son lo mismo.

Es el caso de las métricas alt95, cob3m y cob5m que también las podemos referir como altura dominante Lidar y cobertura arbolada Lidar.

- **Altura dominante lidar (alt95)**: percentil 95 de las alturas de los retornos sobre el suelo (primeros retornos). Si está bien obtenida[3], es una métrica precisa que se corresponde plenamente con la denominada "altura dominante de parcela", que se definió referida a una superficie de 10x10 m y, de hecho, es la mejor forma de calcular la altura dominante de un rodal (promediando "alturas dominantes de parcela"). Esta medida puede considerarse como variable cierta con la que se pueden comparar otras formas de estimar la altura dominante (y no al revés).
- **Cobertura Lidar**: porcentaje de retornos que están a más de cierta altura sobre el suelo (primeros retornos). Si consideramos que las hojas y ramas situados a más de 5 m sobre el suelo constituyen las copas de los árboles, la métrica **cob5m** es una expresión de la cobertura de las copas, asumiendo que las copas no son opacas sino parcialmente permeables. En consecuencia, esta métrica no se corresponde con la fracción de cabida cubierta (**FCC**), que se estima habitualmente considerando las copas de los árboles opacas. De hecho, la FCC es siempre superior a la correspondiente cobertura Lidar. Una masa con plena cobertura de copas, pero con copas parcialmente permeables a la luz (depende mucho de la especie), puede tener una cob5m=70% y considerarse con FCC=100%.

  Dependiendo del tipo de masa, puede resultar útil trabajar con la referencia de 5 m (p. ej., para pinares adultos), con 3 m (p. ej., para montes bajos de frondosas) o con ambos (p. ej., para identificar diferentes estructuras):
  - **Cobertura Lidar 3m (cob3m)**: porcentaje de retornos que están a más de 3 m sobre el suelo (primeros retornos).
  - **Cobertura Lidar 5m (cob5m)**: porcentaje de retornos que están a más de 5 m sobre el suelo (primeros retornos).

Dasolidar se centra en ofrecer métricas y variables que expresan la estructura del arbolado (y la vegetación en general) sin entrar en su interpretación, dejando esta tarea en el lado del usuario o de herramientas complementarias de dasolidar.

Una herramienta muy útil para trabajar en la interpretación selvícola de las métricas es silvilidar que, en sus nuevas versiones, funciona de forma más sencilla y rápida apoyándose en capas ya generadas dentro del proyecto dasolidar.

***Variables dasométricas***

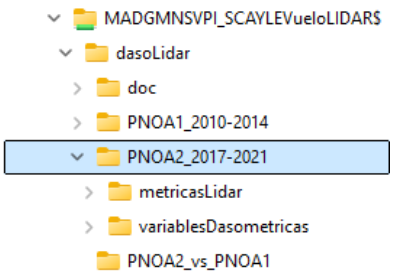
---

[3] Con el suelo bien capturado y obtenida con hoja, es decir durante el periodo vegetativo si es una especie caducifolia.

Con carácter general, las variables dasométricas pueden ser de árbol individual (diámetro normal, altura total, altura de fuste, altura hasta la primera rama viva, volumen de fuste, biomasa aérea del árbol, etc.) o de masa (altura dominante, densidad, área basimétrica, fracción de cabida cubierta, volumen en pie con corteza, volumen de leñas, biomasa aérea, etc.).

Las capas del proyecto dasolidar, que se encuentran en lidarData son siempre variables de masa (estimadas utilizando parcelas de campo y métricas Lidar).

Pte: detallar las variables dasométricas que se usan en dasolidar.



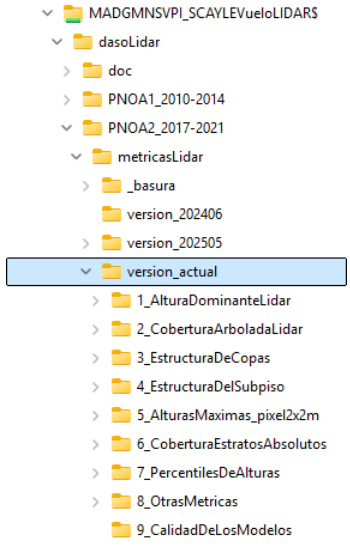

***Lidar-PNOA1 y Lidar-PNOA2***

Se han generado métricas lidar para los vuelos Lidar-PNOA1 y Lidar-PNOA2, pero sólo se han estimado variables dasométricas para el **PNOA2** utilizando los datos de las parcelas del IFN4 y la cartografía MFE25.

Pte: explicar PNOA1, PNOA2, PNOA3.

*Lidar-PNOA2*

Listado de métricas Lidar




La carpeta principal de métricas lidar es la que tiene la **versión actual**. El resto son versiones anteriores, que se dejan por si se quiere consultar qué cambios o mejoras ha habido respecto a ellas. La fecha de referencia de la versión actual se anota en el leeme.txt de esta carpeta.

Listado de capas (feb 2026) organizado por carpetas:

```
+---1_AlturaDominanteLidar
|       Alt95_cm_PNOA2.tif
|       Alt95_m_PNOA2.tif
|
+---2 CoberturaArboladaLidar
|       Cob3m_PRT_PNOA2.tif
|       Cob5m_PRT_PNOA2.tif
|
+---3_EstructuraDeCopas
|
Alt20_tlr_sobre2m_filtradoAlt95_4m_baseDeCopa_cm_PNOA2.tif
|       Cob_midHD_TopHD_PNOA2.tif
|       RazonDeCopaInf_4m_PNOA2.tif
|       RazonDeCopaSup_4m_PNOA2.tif
|
+---4_EstructuraDelSubpiso
|       RazonEstr_050cm_200cm_rptoMidTop_tlr_PNOA2.tif
|       RazonEstr_200cm_midHD_rptoMidTop_tlr_PNOA2.tif
|
+---5_AlturasMaximas_pixel2x2m
|       AltMaxSobreMdk_CE_cm_PNOA2_int16.tif
|       AltMaxSobreMdk_NE_cm_PNOA2_int16.tif
|       AltMaxSobreMdk_NW_cm_PNOA2_int16.tif
|       AltMaxSobreMdk_SE_cm_PNOA2_int16.tif
|       AltMaxSobreMdk_SW_cm_PNOA2_int16.tif
|
+---6_CoberturaEstratosAbsolutos
|       CobEstrDe0025a0050cm_prt_PNOA2.tif
|       CobEstrDe0050a0150cm_prt_PNOA2.tif
|       CobEstrDe0150a0250cm_prt_PNOA2.tif
|       CobEstrDe0250a0300cm_prt_PNOA2.tif
|       CobEstrDe0300a0500cm_prt_PNOA2.tif
|       CobEstr_0025_0050_tlr_PNOA2.tif
|       CobEstr_0050_0150_tlr_PNOA2.tif
|       CobEstr_0150_0250_tlr_PNOA2.tif
|       CobEstr_0250_0300_tlr_PNOA2.tif
|       CobEstr_0300_0500_tlr_PNOA2.tif
|       CobEstr_0500_9999_tlr_PNOA2.tif
|       PorcentajeRetornos050cm_250cm_rptoRetornosSub250cm_tlr.tif
|       PorcentajeRetornos050cm_250cm_rptoRetornosSub250cm_tlr_new.tif
|       PorcentajeRetornos050cm_250cm_rptoRetornosSub250cm_tlr_new2.tif
```

```
|       PorcentajeRetornos_sub250cm_tlr_new.tif
|
+---7_PercentilesDeAlturas
|       Alt100SobreMdb prt sinUmbral cm PNOA2.tif
|       Alt35SobreMdb_tlr_filtradoAlt95_4m_conUmbral2m_cm_PNOA2.tif
|       Alt50SobreMdb_prt_sinUmbral_cm_PNOA2.tif
|       Alt65SobreMdb_prt_sinUmbral_cm_PNOA2.tif
|       Alt80SobreMdb_prt_sinUmbral_cm_PNOA2.tif
|       Alt95SobreMdb_prt_sinUmbral_cm_PNOA2.tif
|
+---8_OtrasMetricas
|   |   Cob200cm_prt_PNOA2.tif
|   |
|   +---extras
|           CobEstrDe0050a0200cm_TLR_PNOA2.tif
|
+---9_CalidadDeLosModelos
```

Significado de algunos códigos (cada capa se explica más adelante, estos códigos son para facilitar la interpretación del nombre sin buscar en la documentación):

```
Alt95, alt95: Percentil 95 de alturas de los retornos sobre el suelo.
Alt20 .. Alt100: Percentiles 20 a 100 (100 = alt max)
Cob3m, Cob5m: porcentaje de retornos por encima de 3 o 5 m.
PRT, prt: primeros retornos
TLR, tlr: todos los retornos (no solo primeros)
midHD: Mitad de la altura dominante Lidar (alt95)
TopHD. Altura dominante Lidar (alt95)
Mdb, Mdk: códigos internos referentes a la metodología de estimación del suelo.
```

Descripción de las capas:

| **Métricas lidar** | |
|---|---|
| **1_AlturaDominanteLidar** | |
| Alt95_cm_PNOA2.tif | Percentil 95 de alturas de los retornos sobre el suelo en m (primeros retornos) |
| Alt95_m_PNOA2.tif | Percentil 95 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| **2_CoberturaArboladaLidar** | |
| Cob3m_PRT_PNOA2.tif | % de primeros retornos que están a 3 o más metros sobre el suelo |
| Cob5m_PRT_PNOA2.tif | % de primeros retornos que están a 5 o más metros sobre el suelo |
| **3_EstructuraDeCopas** | |
| Alt20_tlr_sobre2m_filtradoAlt95 _4m_baseDeCopa_cm_PNOA2.tif | Percentil 20 de alturas sobre el suelo (todos los retornos) tras eliminar los retornos que están a menos de 2 m sobre el suelo, filtrado a valores > 4m (~base de copa). |
| Cob_midHD_TopHD_PNOA2.tif | % de retornos que están encima de la mitad de la Alt95 (todos los retornos). |
| RazonDeCopaInf_4m_PNOA2.tif | Porcentaje de la altura de base de copa respecto a la alt95 |
| RazonDeCopaSup_4m_PNOA2.tif | Porcentaje de la altura de base alta de copa respecto a la alt95[4] |
| **4_EstructuraDelSubpiso** | |
| RazonEstr_050cm_200cm _rptoMidTop_tlr_PNOA2.tif | Proporción de los retornos entre 50 cm y 2 m sobre el suelo referido a los retornos por encima de la mitad de la Alt95 (multiplicado x 100). Todos los retornos. |
| RazonEstr_200cm_midHD _rptoMidTop_tlr_PNOA2.tif | Proporción de los retornos entre 2m sobre el suelo y la mitad de la Alt95 referido a los retornos por encima de esa altura (multiplicado x 100). Todos los retornos. |
| **5_AlturasMaximas_pixel2x2m** | |
| AltMaxSobreMdk_CE_cm _PNOA2_int16.tif | Altura en cm del retorno más alto en cada celda de 2x2 m |
| AltMaxSobreMdk_NE_cm _PNOA2_int16.tif | Altura en cm del retorno más alto en cada celda de 2x2 m |
| AltMaxSobreMdk_NW_cm _PNOA2_int16.tif | Altura en cm del retorno más alto en cada celda de 2x2 m |
| AltMaxSobreMdk_SE_cm _PNOA2_int16.tif | Altura en cm del retorno más alto en cada celda de 2x2 m |
| AltMaxSobreMdk_SW_cm _PNOA2_int16.tif | Altura en cm del retorno más alto en cada celda de 2x2 m |
| **6_CoberturaEstratosAbsolutos** | |
| CobEstrDe0025a0050cm_prt_PNOA2.tif | % de retornos que están entre 25 y 50 cm sobre el suelo (primeros retornos) |

[4] Base alta de copa: equivalente a la base de copa pero percentil 40 en vez de 20.

| | |
|---|---|
| CobEstrDe0050a0150cm_prt_PNOA2.tif | % de retornos que están entre 50 cm y 1,5 m sobre el suelo (primeros retornos) |
| CobEstrDe0150a0250cm_prt_PNOA2.tif | % de retornos que están entre 1,5 y 2,5 m sobre el suelo (primeros retornos) |
| CobEstrDe0250a0300cm_prt_PNOA2.tif | % de retornos que están entre 2,5 y 3,0 m sobre el suelo (primeros retornos) |
| CobEstrDe0300a0500cm_prt_PNOA2.tif | % de retornos que están entre 3 y 5 m sobre el suelo (primeros retornos) |
| CobEstr_0025_0050_tlr_PNOA2.tif | % de retornos que están entre 25 y 50 cm sobre el suelo (todos los retornos) |
| CobEstr_0050_0150_tlr_PNOA2.tif | % de retornos que están entre 50 cm y 1,5 m sobre el suelo (todos los retornos) |
| CobEstr_0150_0250_tlr_PNOA2.tif | % de retornos que están entre 1,5 y 2,5 m sobre el suelo (todos los retornos) |
| CobEstr_0250_0300_tlr_PNOA2.tif | % de retornos que están entre 2,5 y 3,0 m sobre el suelo (todos los retornos) |
| CobEstr_0300_0500_tlr_PNOA2.tif | % de retornos que están entre 3 y 5 m sobre el suelo (todos los retornos) |
| CobEstr_0500_9999_tlr_PNOA2.tif | % de retornos que están a más de 5 m sobre el suelo (todos los retornos) |
| PorcentajeRetornos050cm_250cm _rptoRetornosSub250cm_tlr.tif | % de retornos que están entre 50 cm y 2,5 m referido al número de retornos por debajo de 2,5 m (todos los retornos) – versión antigua (pte revisar) |
| PorcentajeRetornos050cm_250cm _rptoRetornosSub250cm_tlr_new.tif | % de retornos que están entre 50 cm y 2,5 m referido al número de retornos por debajo de 2,5 m (todos los retornos) – versión nueva1 (pte revisar) |
| PorcentajeRetornos050cm_250cm _rptoRetornosSub250cm_tlr_new2.tif | % de retornos que están entre 50 cm y 2,5 m referido al número de retornos por debajo de 2,5 m (todos los retornos) – versión nueva2 (pte revisar) |
| PorcentajeRetornos _sub250cm_tlr_new.tif | % de retornos que están por debajo de 2,5 m sobre el suelo (todos los retornos) (pte revisar) |
| **7_PercentilesDeAlturas** | |
| Alt100SobreMdb_prt _sinUmbral_cm_PNOA2.tif | Altura del retorno más alto sobre el suelo en m. |
| Alt35SobreMdb_tlr_filtradoAlt95 _4m_conUmbral2m_cm_PNOA2.tif | Percentil 35 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| Alt50SobreMdb_prt _sinUmbral_cm_PNOA2.tif | Percentil 50 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| Alt65SobreMdb_prt _sinUmbral_cm_PNOA2.tif | Percentil 65 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| Alt80SobreMdb_prt _sinUmbral_cm_PNOA2.tif | Percentil 80 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| Alt95SobreMdb_prt _sinUmbral_cm_PNOA2.tif | Percentil 95 de alturas de los retornos sobre el suelo en cm (primeros retornos) |
| **8_OtrasMetricas** | |
| Cob200cm_prt_PNOA2.tif | % de primeros retornos que están a 2 o más metros sobre el suelo |
| CobEstrDe0050a0200cm_TLR_PNOA2.tif | % de retornos que están entre 50 cm y 2,0 m sobre el suelo (todos los retornos) |

Cada capa tiene su correspondiente archivo de estilos (.qml). Recomendamos usar esta representación de las variables, que mantendremos en las diferentes capas y versiones.

### Variables dasométricas

La carpeta principal de variables dasométricas es la que tiene la **versión actual**.

Además de esa carpeta, hay otras que contienen:

- Versiones anteriores, que se dejan por si se quiere consultar qué cambios o mejoras ha habido respecto a ellas.
- Versiones obtenidas por diferentes métodos:
  - Modelo lineal: hay varias versiones con diferente número de variables explicativas
  - *Random forest*: hay varias versiones con diferentes hiperparámetros
  - Redes neuronales: hay varias versiones con diferentes hiperparámetros

La fecha de referencia de la versión actual se anota en el leeme.txt de esta carpeta.

| Variables dasométricas | |
|---|---|
| dasoLidar_VCC_lr_cyl.tif | Volumen en pie con corteza en m3/ha |
| dasoLidar_Abas_lr_cyl.tif | Área basimétrica en $m^2$/ha |
| dasoLidar_Npies_lr_cyl.tif | Numero de pies mayores por hectárea (mayores: Dn > 7,5 cm) |
| dasoLidar_DCM_lr_cyl.tif | Diámetro cuadrático medio (diámetros normales con corteza) en cm[5] |
| dasoLidar_IAVC_lr_cyl.tif | Incremento anual en volumen en pie con corteza en m3/ha.año |
| dasoLidar_BA_lr_cyl.tif | Biomasa aérea en toneladas por hectárea |

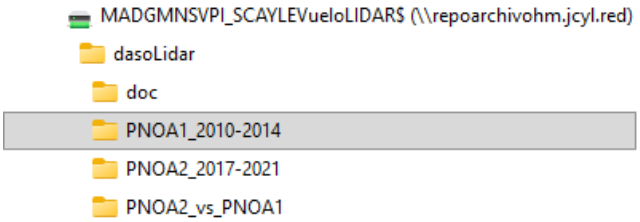
[5] Abas, Npies y DCM están relacionados: (Abas = NPies · π · $DCM^2$ / 40000)

| | |
|---|---|
| dasoLidar_VLE_lr_cyl.tif | Volumen de leñas en m3/ha |

Cada capa tiene su correspondiente archivo de estilos (.qml). Recomendamos usar esta representación de las variables, que mantendremos en las diferentes capas y versiones.

*PNOA1*

Por cuestiones de espacio, en lidarData sólo están los ficheros de nubes de puntos del Lidar-PNOA2.

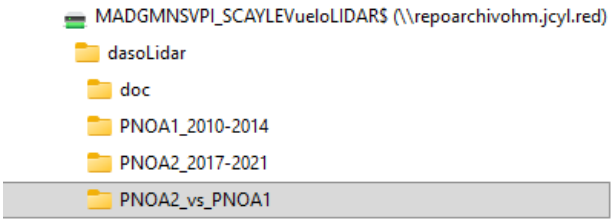



No obstante, sí hemos dejado dos capas dasolidar con métricas Lidar:

- PNOA1_2010-2014
  - **Alt95_m_PNOA1**.tif Percentil 95 de altura sobre el suelo (primeros retornos)
  - **Cob3m_PRT_PNOA1**.tif % de primeros retornos que están a 3 o más metros sobre el suelo

*Comparativa PNOA1-PNOA2*

La obtención de estas capas Alt95 y Cob3m correspondientes al Lidar-PNOA1 no siguió los mismos procedimientos que las del Lidar-PNOA2, por lo que la interpretación de cualquier comparativa debe hacerse a la luz de las diferencias metodológicas y de sensores usados.




No obstante, a modo ilustrativo, se ha incluido en la carpeta dasolidar una de comparativa que contiene la resta de estas capas:

- PNOA2_vs_PNOA1
  - **Alt95_cm_PNOA1_PNOA2**.tif Resta de Alt95 del PNOA2 menos PNOA1 (en cm)
  - **Cob3m_PNOA2-PNOA2**.tif Resta de Cob3m del PNOA2 menos PNOA1

# Directorio PNOA2

## COPC

```
├──────.aux
└──────lazcopc
    ├──────ce
    ├──────ne
    ├──────nw
    ├──────se
    ├──────sw
```

## Cuadrantes de vuelo

En este directorio se encuentran los archivos Lidar propiamente dichos, con las nubes de puntos. Están organizados en cuadrantes de vuelo y, dentro de cada cuadrante, en bloques de 2x2 km. Cada bloque se identifica por las coordenadas UTM de su esquina superior izquierda (en miles de km).

Estos ficheros se han obtenido mediante vuelos realizados entre 2017 y 2021 (Lidar-PNOA2), dentro del Plan Nacional de Ortofotografía Aérea (https://pnoa.ign.es/). En Castilla y León participan CNIG, JCYL e ITACYL:

CNIG -> https://www.cnig.es -> Centro Nacional de Información Geográfica

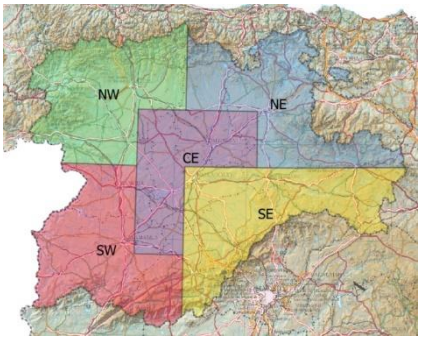
JCYL -> https://www.jcyl.es -> Junta de Castilla y León

ITACYL -> https://www.itacyl.es -> Instituto Tecnológico Agrario de Castilla y León

Los vuelos se han licitado en cinco "cuadrantes". Las empresas adjudicatarias de los vuelos han sido:

SE, CE y NW -> SERVICIOS POLITECNICOS AÉREOS, S.A. (SPASA)

NE y SW -> UTE TOPCAD INGENIERIA - PRIMUL MERIDIAN

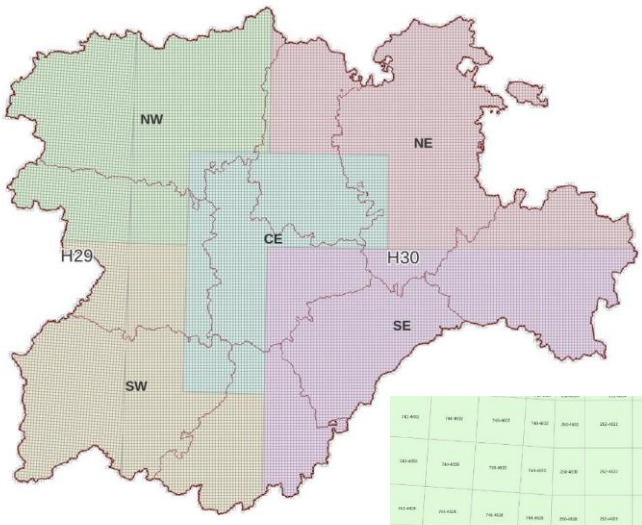


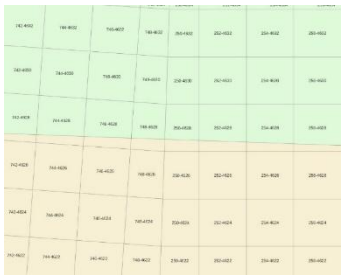

Ver las fechas de vuelo en el apartado 0 Fechas de vuelo.

El detalle de los cuadrantes puede consultarse en el proyecto lidarQgis.qgz (antiguo LidarPNOA2.qgz). Los lazFiles están organizados por cuadrantes de vuelo, de forma que cada directorio incluye los lazFiles correspondientes a un vuelo. Esos ficheros cubren una superficie que excede ligeramente la delimitación de cada cuadrante, por lo que hay cierto solape, necesario para garantizarla cobertura total licitada. En los bordes de las zonas reales de vuelo los lazFiles son parcialmente incompletos (no hay puntos en todo el bloque de 2x2 km), pero esto no es un inconveniente porque eso ocurre siempre fuera del ámbito estricto del cuadrante (el licitado).

En consecuencia, en las franjas cercanas a los límites de los cuadrantes de vuelo hay bloques 2x2 km en dos versiones, voladas en fechas distintas y posiblemente por emperesas distintas y con sensores distintos.

Los cuadrantes NW y SW incluyen zonas en huso 29 (además del 30 que es el del resto de Castilla y León). En esas zonas del oeste de León, Zamora y Salamanca, las coordenadas correspondientes a la proyección en el huso 29 es diferente a las que se obtienen proyectando en el 30 extendido. Los lasFiles que están en el ámbito del huso 29 están proyectados en dicho uso. No obstante, cuando se cargan en Qgis, en un proyecto creado en el huso 30, se reproyectan al vuelo y aparecen reubicadas en las coordenadas correspondientes al huso 30. Se ha incluido información del sistema de proyección en los lazFiles para que todo esto ocurra de forma automática al cargarlos en Qgis, sin intervención del usuario.



Sin embargo, todas las capas dasoLidar generadas están proyectadas en huso 30 para su visualización en una sola capa, como esta adjunta de las mallas 2x2 km, que incluye tanto la zona del huso 30, como la del 29 proyectada en el 30 extendido.

El nombre de los ficheros lidar (lazFiles) incluye las coordenadas de su esquina superior izquierda (en miles de km); dichas coordenadas están en el huso original del bloque (29 o 30). En todo caso, se puede deducir el huso de un bloque a la vista de las coordenadas ya que la coordenada X de los bloques 2x2 de Castilla y León correspondientes al huso 30 son siempre inferiores a 610.000 (códigos de bloque 240-YYYY a 600-YYYY) y las que están en el huso 29 son siempre superiores a 650.000 (códigos de bloque 656-YYYY a 754-YYYY).

El número de cuadrados de 2x2 km que cubren todo Castilla y León es de 24.481 bloques. Debido a los solapes entre las zonas voladas, el número total de bloques 2x2 km es de 26.589.

En la siguiente tabla se indican los sensores utilizados, así como la precisión y las densidades de pulsos licitadas:

| Cuadrante | Inicio vuelo | Fin vuelo | Pulsos/m2 | RMSE xy | RMSE z | Sensor principal |
|---|---|---|---|---|---|---|
| CYL SE | nov-17 | dic-18 | 1 | 0,2 | 0,15 | LEICA ALS80 |
| CyL-NW | jun-21 | oct-21 | 2 | 0,3 | 0,15 | OPTECH GALAXY T2000 |

| Cuadrante | Inicio vuelo | Fin vuelo | Pulsos/m2 | RMSE xy | RMSE z | Sensor principal |
|---|---|---|---|---|---|---|
| CyL SW | sep-19 | jul-20 | 4 | 0,3 | 0,15 | RIEGL LMS-Q1560 |
| CyL NE | oct-19 | oct-21 | 4 | 0,3 | 0,15 | RIEGL LMS-Q1560 |
| CyL C | sep-19 | oct-19 | 1 | 0,3 | 0,15 | LEICA ALS80 |

Se pueden consultar más detalles, incluidas las especificaciones de cada licitación en: https://pnoa.ign.es/web/portal/pnoa-lidar/segunda-cobertura. Aviso: si se consulta esta página desde un PC del trabajo es posible que no proporcione información actualizada debido a que el navegador de internet muestra una versión cacheada de esta web, obtenida hace bastante tiempo, antes de que el IGN volcara toda la información de la segunda cobertura (Lidar-PNOA2). Se recomienda acceder a esta ruta desde el móvil o desde un PC fuera de la red de la Junta.

## Fechas de vuelo

Los vuelos Lidar se planifican normalmente mediante un barrido en pasadas este-oeste con cierto solape entre ellas. Los puntos de cada pasada se registran en un intervalo pequeño de tiempo, por ejemplo, entre 10 y 25 minutos (si la pasada de vuelo es de 50 o 100 km y el avión vuela entre 250 y 300 km/h).

Al componer los bloques de 2x2 km se incluyen todos los puntos volados en ese cuadrado, que normalmente incluye más de una pasada. Las pasadas pueden ser del mismo día, de días distintos o incluso de años distintos, con lo que no se puede asignar una única fecha a cada bloque 2x2, sino un intervalo de fechas que va del primer retorno al último.

El formato ASPRS incluye en su cabecera un campo para registrar el año de generación del fichero ".las" pero no un campo para almacenar el rango de fechas de vuelo. Para cada punto, la propiedad "time" si contiene esta información, pero no está sintetizada en la cabecera.

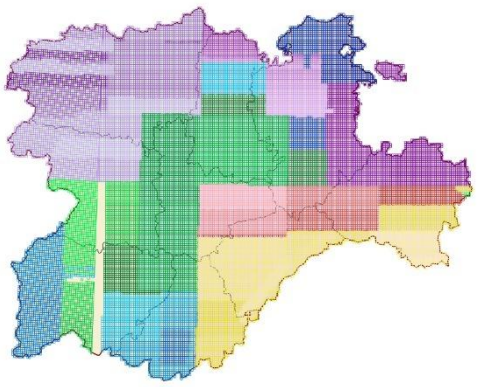
Para solventar este inconveniente se ha utilizado el campo System Identifier (sysid) para anotar esta información. Las especificaciones ASPRS crearon este campo para recoger el dispositivo de captura (modelo de sensor Lidar, normalmente un Leica, Optech o Riegl), pensando en ficheros ".las" de salida de los dispositivos de captura. Pero como esos ficheros se postprocesan antes de ofrecerlos al público, la ASPRS optó por recoger también en este campo el tipo de procesado ("MERGE", "MODIFICATION", "EXTRACTION", "TRANSFORMATION" y "OTHER").

La realidad es que los ficheros son objeto de varios procesos y no tiene mucho sentido mencionar únicamente el último de ellos, razón por la cual, se ha obviado esta especificación ASPRS y se ha utilizado este campo para recoger el intervalo de fechas de vuelo con el siguiente formato:

FechasDeVuelo_AAAAMMDD_AAAAMMDD

El primer AAAMMDD se refiere a año, mes día del primer retorno y el segundo al último.

La distribución de fechas de vuelo es diversa (los distintos colores reflejan distintos intervalos de fechas y la intensidad de color el día concreto; para ver datos concretos, consultar el proyecto de trabajo LidarQgis.qgz (antiguo LidarPNOA2plus.qgz):



Para tener una información más detallada de las fechas de vuelo se podría hacer un histograma de fechas de cada fichero 2x2 km. Esto permitiría un mejor criterio de uso de esta información. Sin embargo, por el momento no se ha realizado. Si se desea incluir esta información en el formato ".las", sería necesario utilizar un VLR

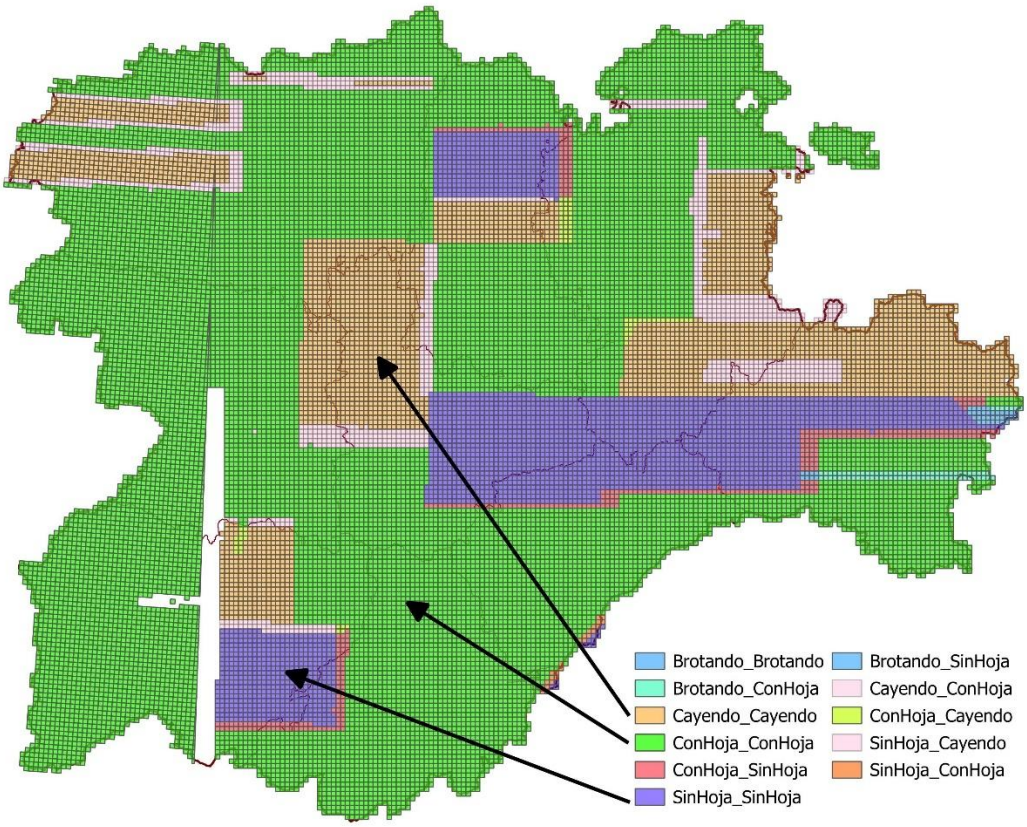
A modo de resumen, estas son las fechas de vuelo de los distintos cuadrantes:

- Cuadrante CE: volado todo en 2019 (3810 bloques)

- Cuadrante NE:  volado principalmente en 2020 y 2021 (595 bloques terminados en 2019, 1540 bloques en 2020 y 3665 bloques en 2021)
- Cuadrante NW:  volado todo en 2021 (5715 bloques)
- Cuadrante SE:  volado principalmente en 2017 y 2018 (2105 bloques terminados en 2017, 3418 bloques en 2018 y 123 bloques en 2019; supongo que en 2019 los que solapan con el cuadrante CE)
- Cuadrante SW: volado en 2019 y 2020 (1993 bloques terminados en 2019 y 3617 terminados en 2020)

| Numero de bloques terminados de volar por cuadrante y año | | | | | | |
|---|---|---|---|---|---|---|
| **Cuadrante** | **2017** | **2018** | **2019** | **2020** | **2021** | **Total** |
| **CE** | | | 3.810 | | | **3.810** |
| **NE** | | | 595 | 1.540 | 3.665 | **5.800** |
| **NW** | | | | | 5.715 | **5.715** |
| **SE** | 2.105 | 3.418 | 123 | | | **5.646** |
| **SW** | | | 1.993 | 3.617 | | **5.610** |
| **Total** | **2.105** | **3.418** | **6.521** | **5.157** | **9.380** | **26.581** |

Basándonos en las fechas inicial y final de cada bloque podemos deducir si el vuelo se ha hecho con hoja o sin hoja, con el siguiente resultado:




---

*La propiedad time de los puntos lidar*

---

*La propiedad time es el número de segundos desde las 0 horas del 6 de enero de 1980 (GPS time) menos 1E9. Para obtener la fecha correspondiente al valor time de un punto concreto, se puede convertir a timestamp (referida a 1 de enero de 1970) sumándole 1315964800 segundos y usar un*

*conversor on-line (o con código) de timesatmp a hora UTC u hora local (p. ej. https://timestamp.online/). En realidad, a día de hoy, el resultado tiene un desfase de 18 segundos respecto a la UTC (por los segundos intercalares con que se ha ido corrigiendo la UTC). También se puede convertir a GPS time (sumando 1E9) y usar un conversor de GPS time a UTC on-line (p. ej. https://gwosc.org/gps/) y, si el conversor tiene en cuenta los segundos intercalares, (como es el caso) el resultado es UTC exacto. Para obtener la hora local de Europa central (y Madrid) hay que sumar una o dos horas, según si la fecha cae dentro del horario de invierno o verano.*

## Los ficheros Lidar (lazFiles)

Los archivos Lidar están en formato laz (lazFiles), que es la versión comprimida de los lasFiles (ASPRS: versión las-1.4, formato de punto 8). Están indexados en formato copc con pdal para una carga rápida en QGIS.

El nombre de los ficheros incluye esta información:

- AAAA (año de vuelo correspondiente al último punto volado en el bloque): esta información no se ha incluido en el nombre en el cuadrante SW (la versión de los lazFiles del cuadrante SW es provisional y el año que figura en el nombre es, provisionalmente, 2023, hasta tanto generemos un dataset completo definitivo). Ver más detalles en el apartado 0 Fechas de vuelo.
- VV (valores CE/NE/NW/SE/SW): cuadrante de vuelo
- XXX, YYYY: coordenadas UTM de la esquina superior izquierda del bloque 2x2 km.
- ORT: alturas ortométricas
- CLR: clasificación de puntos revisada
- RGBI: coloreado RGBI (incluye 4 bandas: rojo, verde, azul e infrarrojo)

Directorios que contienen los cuadrantes y el formato de nombre de fichero[6]:

- ce -> PNOA_AAAA_CYL_CE_XXX-YYY_ORT_CLR_RGBI.copc.laz (3810 files, 140 GB)
- ne -> PNOA-AAAA-CYL-NE-XXX-YYYY-ORT-CLR-RGBI.copc.laz (5800 files, 1,64 TB)
- nw -> PNOA_AAA_CYL_NW_XXX-YYYY_ORT_CLR_RGBI.copc.laz (5715 files 741 GB)
- se -> PNOA_AAA_CYL_SE_XXX-YYYY_ORT_CLR_RGBI.copc.laz (5646 files 277 GB)
- sw -> Pendiente

Como a fecha 1 de octubre de 2024 no disponemos de una versión definitiva del cuadrante SW, hemos incluido una versión provisional:

- sw_semiDepurado_20231009 -> PNOA-2023-CYL-SW-XXX-YYYY-ORT-000-RGBI_LF14PF8.copc.laz (5618 files 1,32 TB)

[6] Como se puede comprobar, no se ha unificado el criterio de uso de guión medio y bajo en el nombre de los ficheros. Este asunto está pendiente.

# Regresiones Lidar

Estos datos y figuras corresponden a la versión de 2024. Pendiente actualizar a los modelos usados en 2025.

En 2025 se han generado también capas obtenidas con random forest y redes neuronales. Esas versiones están disponibles en repoarchivohm.jcyl.red pero no se han subido, por el momento, a OpenCayle, con lo que no se detalla aquí la metodología.

## El modelo lineal

Las variables que vamos a estimar con el modelo lineal (variables explicadas) son:

- VCC: Volumen con corteza ($m^3$/ha)
- DCM: Diámetro cuadrático medio (cm)
- NPIES: Número de pies por hectárea (n/ha)
- ABAS: Área basimétrica ($m^2$/ha)
- IAVC: Incremento anual en volumen con corteza ($m^3$/ha.año)
- VLE: Volumen de leñas ($m^3$/ha)
- BA: Biomasa aérea (t/ha)

Sus estadísticas descriptivas se recogen en el apartado **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**.

Los datos con los que se construyen las regresiones proceden de las parcelas del IFN4. Ver detalles en **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**.

Las variables explicativas usadas en las regresiones varían de una a otra porque en cada regresión se lleva a cabo una selección de las variables explicativas que más aportan a la misma, usando para ello el método de forward stepwise con el criterio AIC (ver apartado 0. Selección de variables explicativas).

El punto de partida es una lista de 30 variables

Lista de variables explicativas:

| Variable | Min | Med-2σ | Media | Med+2σ | Max |
|---|---|---|---|---|---|
| alt100PrtMdb | 101.21 | 265.37 | 1103.30 | 1941.23 | 4784.43 |
| alt095PrtMdb | 8.50 | 148.19 | 956.12 | 1764.05 | 3832.75 |
| alt080PrtMdb | 3.25 | -18.26 | 773.61 | 1565.48 | 3604.12 |
| alt065PrtMdb | 1.25 | -131.27 | 638.53 | 1408.33 | 3396.94 |
| alt050PrtMdb | 0.00 | -216.96 | 514.48 | 1245.91 | 3186.00 |
| alt035_2TlrMdb | | | | | |
| alt020_2TlrMdb | 0.00 | 15.96 | 555.70 | 1095.44 | 2789.56 |
| cob050Prt | 0.21 | 20.40 | 57.80 | 95.20 | 100.00 |
| cob200Prt | 0.21 | 12.50 | 51.76 | 91.01 | 100.00 |
| cob300Prt | 0.14 | 6.74 | 47.62 | 88.50 | 99.68 |
| cob500Prt | 0.00 | -6.09 | 38.56 | 83.21 | 99.58 |
| cobPrtMdbRango0025_0050 | 0.00 | -2.82 | 2.41 | 7.63 | 64.00 |
| cobPrtMdbRango0050_0150 | 0.00 | -4.61 | 4.01 | 12.63 | 52.33 |
| cobPrtMdbRango0150_0250 | 0.00 | -3.90 | 3.64 | 11.18 | 53.00 |
| cobPrtMdbRango0250_0300 | 0.00 | -2.32 | 1.94 | 6.20 | 29.67 |
| cobPrtMdbRango0300_0500 | 0.00 | -6.95 | 9.01 | 24.96 | 78.00 |
| cobTlrMdbRango0025_0050 | 0.00 | -1.90 | 2.31 | 6.52 | 49.00 |
| cobTlrMdbRango0050_0150 | 0.00 | -3.42 | 3.30 | 10.02 | 40.67 |
| cobTlrMdbRango0150_0250 | 0.00 | -2.43 | 2.76 | 7.94 | 34.00 |
| cobTlrMdbRango0250_0300 | 0.00 | -1.31 | 1.36 | 4.03 | 18.00 |
| cobTlrMdbRango0300_0500 | 0.00 | -3.28 | 6.39 | 16.06 | 53.50 |

| | | | | | |
|---|---|---|---|---|---|
| cobTlrMdbRango0500_9999 | 0.00 | -6.97 | 26.67 | 60.31 | 95.42 |
| CobTodosRet_RangoDe0025a0150cm | 0.00 | -4.16 | 5.61 | 15.39 | 54.33 |
| CobTodosRet_RangoDe0250a0500cm | 0.00 | -4.06 | 7.75 | 19.55 | 69.33 |
| PropTlrMdbEstr050cm_200cm_rptoMidTop | 0.00 | -12.91 | 20.20 | 53.30 | 126.00 |
| PropTlrMdbEstr200cm_midHD_rptoMidTop | 0.00 | -12.91 | 20.20 | 53.30 | 126.00 |
| CobTodosRet_Rango_050cm_200cm | 0.00 | -2.33 | 4.50 | 11.33 | 34.02 |
| CobTodosRet_Rango_200cm_midHD | 0.00 | -2.33 | 4.50 | 11.33 | 34.02 |
| CobTlrMdbEstrmidHD_TopHD | 0.00 | 3.97 | 30.64 | 57.31 | 94.58 |
| mdb10 | 185.50 | 633.64 | 1007.78 | 1381.92 | 2036.57 |
| | | | | | |
| RazonDeCopaInf | -0.52 | 0.24 | 0.42 | 0.60 | 1.00 |
| RazonDeCopaSup | | | | | |
| Alt100**1.7 | 2563.80 | -47435.03 | 171067.61 | 389570.26 | 1801818.86 |
| Alt095**1.7 | 38.02 | -53640.40 | 138285.88 | 330212.15 | 1235855.90 |
| Alt080**1.5 | 5.86 | -11444.73 | 25184.79 | 61814.30 | 216371.36 |
| Alt080**1.7 | 7.42 | -64771.18 | 103475.12 | 271721.41 | 1113165.50 |
| Alt065**1.5 | 1.40 | -13465.25 | 19916.81 | 53298.88 | 197984.57 |
| Alt050**1.5 | 0.00 | -14527.66 | 15392.61 | 45312.88 | 179832.70 |
| Alt035**1.5 | | | | | |
| Alt020**1.5 | 0.00 | -7340.13 | 15022.00 | 37384.13 | 147334.40 |
| Cob0m**1.5 | 0.10 | 59.73 | 471.50 | 883.26 | 1000.00 |
| Cob2m**1.5 | 0.10 | -8.34 | 409.37 | 827.07 | 1000.00 |
| Cob3m**1.5 | 0.05 | -52.35 | 370.39 | 793.13 | 995.27 |
| Cob3m**1.7 | 0.04 | -206.12 | 852.57 | 1911.26 | 2498.42 |
| Cob5m**1.5 | 0.00 | -135.38 | 294.62 | 724.62 | 993.69 |
| Cob5m**1.7 | 0.00 | -381.75 | 675.26 | 1732.27 | 2493.93 |
| CobSup**1.7 | 0.00 | -160.88 | 402.83 | 966.54 | 2284.81 |
| CobSup**2 | 0.00 | -785.40 | 1254.88 | 3295.15 | 8945.18 |
| inv_CobTodosRet_RangoDe0250a0500cm | 0.01 | 0.03 | 0.07 | 0.10 | 0.10 |

Una de las premisas de la regresión es la de hacer las estimaciones con el mínimo margen de error posible. Eso pasa por:

- Obtener datos fiables para construir los ajustes.
- Elegir adecuadamente el modelo y el proceso de ajuste
- Que el proceso de ajuste incluya la selección de variables explicativas más adecuadas.

## Planteamiento estadístico

Algunas variables dasométricas de masa se pueden estimar a partir de métricas Lidar, asumiendo que responden a un modelo lineal:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Que se aplica en forma matricial:

$$y=X^T\beta+\epsilon \qquad \mathbf{y} = \mathbf{X}^T\beta + \epsilon$$

Donde:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ 1 & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_n \end{bmatrix}$$

- La variable explicada ($y_i$, y) es una variable dasométrica medida en una parcela de campo que, en este caso, es una parcela del IFN4 (ver **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**)

- Las variables explicativas ($x_{i1}$, $x_{i2}$ ... $x_{ip}$, X) son métricas Lidar (originales o transformadas) obtenidas para esas mismas parcelas de campo (ver apartado **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**). El modelo puede incluir otros estimadores no Lidar para mejorar su capacidad predictiva.
- Los coeficientes o parámetros ($\beta_0$, $\beta_1$, $\beta_2$ ... $\beta_p$) los estimamos en el ajuste.
- ($\epsilon_i$, $\epsilon$) es el error o componente aleatoria no explicada por el modelo.

## La clave de casi todo: el error residual

### *Qué es el error cuadrático medio residual*

La componente aleatoria es una parte importante del modelo y queda expresada por los errores residuales ($\epsilon i$, $\epsilon$). Si retiramos esa componente del modelo, lo que obtenemos es el comportamiento de la media poblacional.

El parámetro que mejor nos permite saber qué tal funciona una regresión Lidar es la raíz del error cuadrático medio residual (se suele representar como rmse: *root mean square error*)[7]. Refleja cuanto distan, en promedio, los valores reales concretos del valor estimado por la regresión (residuos). Está expresado en las mismas unidades que la variable dasométrica explicada:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

El RMSE es la raíz cuadrada del error cuadrático medio (ECM o MSE)

Para estimar este error usamos el RMSE ajustado con el número de coeficientes de la regresión. Se obtiene multiplicando el RMSE por $\sqrt{(n/n-p-1)}$, siendo n el tamaño de la muestra con la que se construye el modelo y p el número de variables explicativas del modelo (el número de coeficiente es igual al número de variables más uno, siendo este último el coeficiente $\beta_0$).

El RMSE ajustado también se puede representar como:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

donde RSS es la suma de residuos elevados al cuadrado.

### *Consideraciones prácticas*

Una parte importante del proceso de ajuste es la de conseguir que los residuos cumplan unas condiciones, especialmente que no haya sesgos en determinados rangos de las variables explicativas (y explicadas) o en determinadas zonas geográficas. Un adecuado análisis del error residual permite establecer intervalos de confianza para las estimaciones y esta es una de las ventajas fundamentales del modelo lineal frente a otros métodos numéricos basados en machine learning.

Algunas reflexiones prácticas relacionadas con el error residual:

- Si nos entregan un modelo ajustado y detectamos casos en los que éste sobreestima o infraestima nuestras mediciones de campo, hay varias cosas que podemos hacer para hacer frente a esta aparente discordancia:
  - Analizar la fiabilidad de nuestras mediciones de campo y verificar si se han obtenido con las mismas ecuaciones de cubicación o crecimiento (en su caso) que los datos usados en la construcción del modelo. La estimación del volumen de una parcela de campo, que pudiéramos dar por cierta, en realidad conlleva siempre el uso de uno o dos modelos de regresión que

[7] También se puede denominar "raíz de la Desviación Cuadrática Media" (RDCM). A veces se abrevia a "error cuadrático medio".

pueden ser imprecisos o sesgados: la ecuación de cubicación que da el volumen de un árbol a partir de su altura y diámetro normal y, con frecuencia, también la curva diámetros-alturas, que da la altura de un árbol a partir de su diámetro. Ésta última debe ser siempre específica de un monte o tipo de gestión concretos.
    - Analizar si las diferencias entre nuestras mediciones y las estimaciones del modelo son consistentemente positivas o negativas. Debe tenerse en cuenta que, si consideramos todo el ámbito de aplicación del modelo, el proceso de ajuste tiene como consecuencia que prácticamente no haya sesgos. Si se detectan sesgos, lo más probable es que ocurran en determinados ámbitos (geográfico o de rango de valores de las variables) y deben identificarse para evitar en lo posible este tipo de sesgos.
    - Para saber si nuestras mediciones son coherentes o discordantes con las estimaciones del modelo hay que conocer la fiabilidad de éste y evaluar si nuestros valores son anormalmente diferentes de la predicción del modelo. Para ello es necesario conocer los intervalos confianza de sus estimaciones (para valores individuales y para la media poblacional).
- Debemos entender el significado del error residual: la capacidad predictiva del modelo de regresión está siempre limitada por la capacidad predictiva de los datos con los que se construye. Y el error residual refleja el conjunto de circunstancias que influyen en la variable explicada pero no está contemplada en ninguna variable explicativa. Si bien es cierto que cuanto mejor sea la información de campo, mayor va a ser el partido que podemos sacar de ella, la fiabilidad de la inferencia siempre estará limitada por la capacidad predictiva de las variables disponibles, siendo ese límite infranqueable por mucho que se aumente la cantidad de datos usados en la construcción de las regresiones.
- Las variables dasométricas no pueden ser estimadas sin error; siempre hay un cierto margen de error cuya cuantificación es parte del proceso de ajuste. Incluso las mediciones en campo de las variables dasométricas tienen sus respectivos errores asociados, especialmente aquellas que se obtienen de forma indirecta como el volumen de los árboles o su crecimiento. Incluso algunas mediciones más o menos directas como las alturas, tienen sus errores de medición.

## Estimación de los coeficientes

Los estimadores de los parámetros β son los que minimizan los residuos:

$$\hat{\beta} = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}^T\beta)^2$$

El estimador de la y se obtiene con esos estimadores de los parámetros:

$$\hat{\boldsymbol{y}}_i = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x}_{i1} + \hat{\beta}_2 \boldsymbol{x}_{i2} + \ldots + \hat{\beta}_p \boldsymbol{x}_{ip}$$

Estimador de su varianza:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-p} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p}$$

Error medio residual:

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

Porcentaje de varianza explicada por el modelo:

$$1 - \frac{\sum(\hat{y_i} - y_i)^2}{\sum(y_i - \bar{y})^2}$$

$$R^2 = \frac{\text{Suma de cuadrados totales - Suma de cuadrados residuales}}{\text{Suma de cuadrados totales}} =$$

$$R^2_{ajustado} = 1 - \frac{SSE}{SST} x \frac{n-1}{n-p-1} = R^2 - (1-R^2)\frac{n-1}{n-p-1} = 1 - \frac{SSE/df_e}{SST/df_t}$$

Para saber si un modelo es significativo se puede usar el F test:

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)}$$

Este contraste responde a la pregunta de si el modelo en su conjunto es capaz de predecir la variable respuesta mejor de lo esperado por azar, o lo que es equivalente, si al menos uno de los predictores que forman el modelo contribuye de forma significativa. Si utilizamos las métricas Lidar habituales los modelos son siempre significativos: en este caso, este test tiene interés para valorar si la inclusión de una nueva variable explicativa mejora o no la capacidad predictiva del modelo. También se puede analizar individualmente cada coeficiente estimado para valorar si es significativamente distinto de cero.

## Los coeficientes del modelo: significado y fiabilidad

Los coeficientes $\beta j$ expresan el efecto promedio que tiene sobre la variable respuesta el incremento en una unidad de la variable predictora $xj$, manteniéndose constantes el resto de las variables

La magnitud de cada coeficiente parcial de regresión depende de las unidades en las que se mida la variable predictora a la que corresponde, por lo que su magnitud no está asociada con la importancia de cada predictor.

Para poder comparar el impacto que tienen en el modelo cada una de las variables, se emplean los coeficientes parciales estandarizados, que se obtienen al estandarizar (sustraer la media y dividir entre la desviación estándar) las variables predictoras previo ajuste del modelo. En este caso, β0. se corresponde con el valor esperado de la variable respuesta cuando todos los predictores se encuentran en su valor promedio, y βj el cambio promedio esperado de la variable respuesta al incrementar en una desviación estándar la variable predictora xj, manteniéndose constantes el resto de las variables.

Para cada uno de los coeficientes de la ecuación de regresión lineal (βj) se puede calcular su significancia y su intervalo de confianza. La prueba estadística más empleada es el t-test de significancia para los coeficientes (βj) del modelo lineal, que considera como hipótesis:

- H0: el predictor xj no contribuye al modelo (βj=0), en presencia del resto de predictores.
- Ha: el predictor xj sí contribuye al modelo (βj≠0), en presencia del resto de predictores. En el caso de regresión lineal simple, se puede interpretar también como que sí existe relación lineal entre ambas variables por lo que la pendiente del modelo es distinta de cero βj≠0.

**Cálculo del estadístico T y del *p-value*:**

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Donde

$$SE(\hat{\beta}_j)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ji} - \bar{x})^2}$$

## Hipótesis de partida del modelo lineal: que hay que chequear (además del *rmse*) para saber si nuestra regresión va a funcionar como se espera

Una vez seleccionado el mejor modelo que se puede crear con los datos disponibles, se puede comprobar su capacidad prediciendo nuevas observaciones que no se hayan empleado para ajustarlo. De este modo se verifica si el modelo se puede generalizar. Una estrategia comúnmente empleada es dividir aleatoriamente los datos en dos grupos, ajustar el modelo con el primer grupo y estimar la precisión de las predicciones con el segundo.

Esta estrategia es imprescindible cuando se trabaja con otros métodos numéricos que no tienen el potente respaldo estadístico del modelo lineal. En nuestro caso, preferimos seguir el camino de la ortodoxia estadística y hemos optado por analizar los requisitos del modelo lineal, que apunta a la raíz de los posibles problemas, más que hacer contrastes a ciegas sin buscar las causas de posibles problemas de generalización.

Las hipótesis de partida que dan confianza a la inferencia que se hace con el modelo lineal son:

- Buena distribución de los residuos: normalidad, homocedasticidad y medias cercanas a cero en todo el rango de valores de las variables, como confirmación de la linealidad de la relación entre las variables explicativas y explicadas.
- Baja colinealidad entre variables explicativas.
- Independencia entre los elementos que componen la muestra.
- No existencia de valores atípicos (*outliers* que pueden ser debidos a errores).
- Tamaño muestral suficiente .

A estas hipótesis de partida hay que añadir otra comprobación importante y es que la inferencia se haga dentro del rango de validez de las variables explicativas.

### *Normalidad en la distribución de los residuos*

Esta hipótesis es siempre la primera que se menciona en el modelo lineal, pero, en mi opinión, es la que menos riesgos conlleva cuando no se cumple de forma estricta. El hecho de que las medias muestrales (que es lo que importa) tiendan a la distribución normal hace que no dediquemos mucho esfuerzo a analizar este requisito. Además, su incumplimiento tiene como principal consecuencia una ligera sobreestimación o infraestimación de los intervalos de confianza y, en realidad, la cuantificación de los intervalos de confianza tiene otros problemas más importantes, que se abordan en el correspondiente apartado.

### *Relación lineal entre la variable explicada y las explicativas*

Una cuestión relacionada con la normalidad es que la media de los residuos sea cercana a cero en todo el rango de valores de las variables implicadas. Esto es lo mismo que decir que el modelo es adecuado: si la relación entre la variable explicada y las explicativas no es lineal sino, por ejemplo, cuadrática, es imposible cumplir este requisito y esto hace que el modelo funcione mal en determinados rangos de valores (aquellos en los que la media de los residuos dista más del modelo ajustado).

En alometría es bien conocido que la relación entre el volumen de fuste o la biomasa de un árbol no tiene una relación lineal ni con el diámetro ni con la altura, sino con estas variables elevadas a cierta potencia que depende del número de entradas de la ecuación de cubicación o ecuación alométrica. En las ecuaciones de dos entradas es habitual que el diámetro deba elevarse a una potencia ligeramente inferior a 2 y la altura a una potencia ligeramente inferior a 1. En ecuaciones de una sola entrada (diámetro) éste debe elevarse a potencias entre 2 y 3.

### Linealidad entre cada variable explicativa por separado y la explicada

Se ha analizado la linealidad entre la variable explicada y las explicativas y se ha detectado falta de linealidad clara para algunas de las métricas. Para solucionar este problema se han incluido en el modelo variables explicativas que se obtienen a partir de las variables con problemas de linealidad, elevándolas a las potencias que favorecen esta linealidad.

Es el caso de los percentiles de alturas y porcentajes de cobertura usados como métricas en las regresiones, que mejoran su relación lineal elevándolos a potencias ligeramente inferiores a 2 (cambia según los casos), razón por la cual se han incluido variables explicativas consistentes en dichas métricas elevadas a potencias de 1.5, 1.7 y 2.0. Esto se ha usado para las métricas que presentan relaciones más estrechas con las variables explicadas (variables dasométricas) en las que es más fácil detectar esta falta de linealidad. Para el resto de las métricas no se han incluido este tipo de variables derivadas.

Para mostrar el tipo de relación entre cada variable explicada y las explicativas se han generado gráficos de dispersión que se recogen en el directorio:

\dasoLidar\varios\modelos&ajustes\graficos\explicadas_vs_explicativas

De la ubicación de red:

\\repoarchivohm.jcyl.red\MADGMNSVPI_SCAYLEVueloLIDAR$

Los gráficos se han generado para cada variable explicada considerando el conjunto de las parcelas (10.116) y por grupos según:

- Especie (con dos códigos de colores: por provincia y por altitud).
- Especie & provincia (código de colores por altitud).

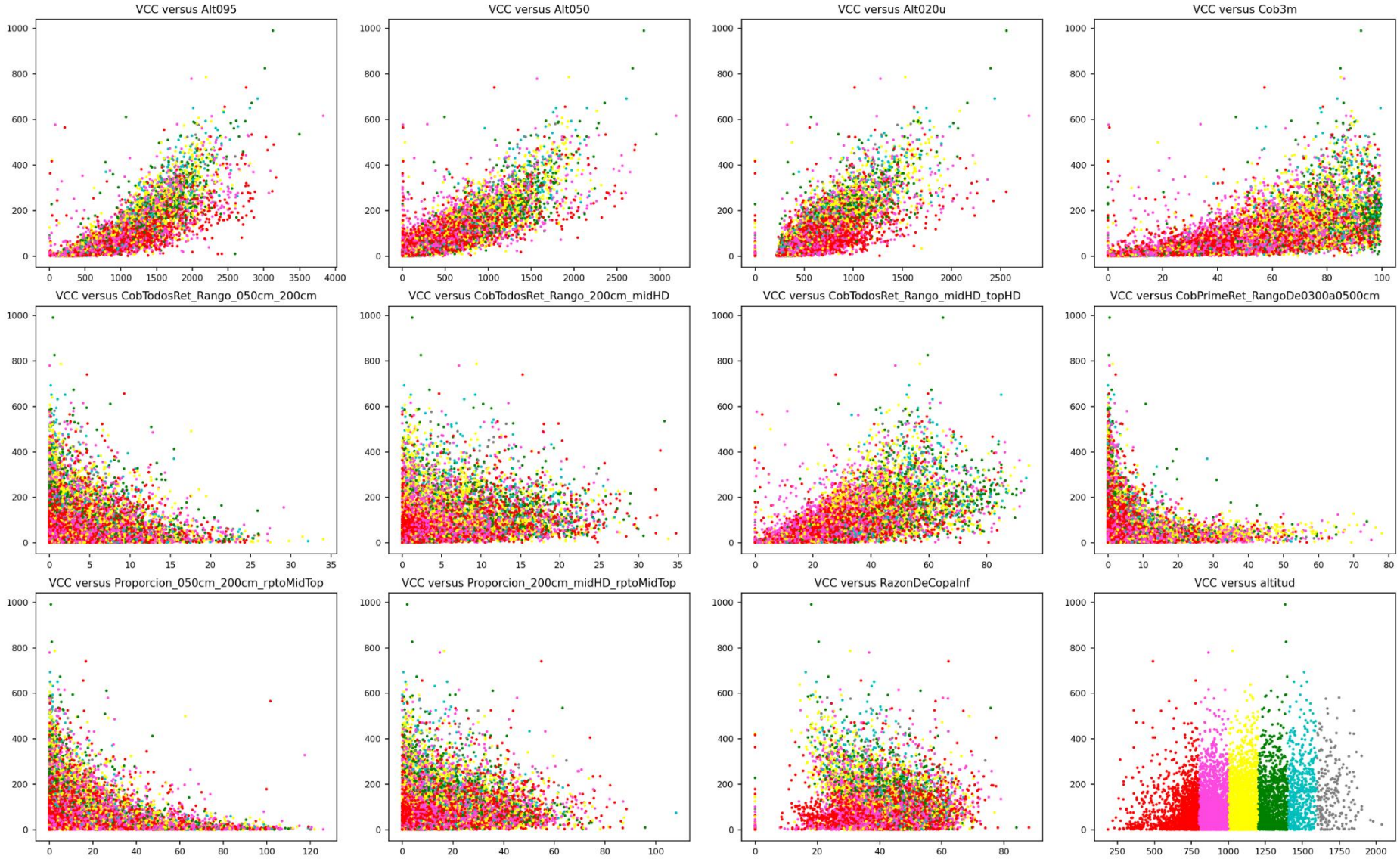
Las 5.952 figuras generadas (496 ficheros con 12 gráficos de dispersión cada uno) permiten ver la relación entre cada variable explicada y 12 explicativas principales para las distintas especies y provincias. A continuación, se muestra la correspondiente

### Análisis de los residuos de la regresión frente a cada variable explicativa

Aunque el anterior apartado da pistas sobre la linealidad de la relación entre las variables explicativas y la explicada, esa linealidad se puede verificar con mejor criterio representando los residuos del modelo de regresión frente a cada una de las variables explicativas. Este análisis también es orientativo, ya que, en rigor, debería verificarse si la relación es lineal cuando el resto de las variables explicativas se mantienen constantes. Este análisis de los residuos de las regresiones queda pendiente para la próxima versión de dasoLidar.

Relación de la variable explicada VCC con 12 explicativas principales con 10116 parcelas

*Los colores indican la altitud. <800: rojo; 800-1000: rosa; 1000-1200: amarillo; 1200-1400: verde; 1400-1600: azul; >1600: gris.*



### *Homocedasticidad en la distribución de los residuos*

La homocedasticidad es la constancia en la varianza de los residuos en todo el rango de validez de las variables. En la versión 1.0 de dasoLidar se ha optado por no hacer ninguna transformación de las variables explicadas, de forma que esta cuestión se valorará en futuras versiones. La transformación logarítmica resuelve en buena medida el tipo de heterocedasticidad que tienen este tipo de regresiones y que se manifiesta en las figuras comentadas en el anterior apartado, especialmente para las variables explicativas más relacionadas con las explicadas como es el percentil 95 de alturas.

Se ha optado por no hacer la mencionada transformación de la variable porque ello conlleva algunos inconvenientes y porque se ha optado por dar prioridad a otras cuestiones más importantes.

La principal consecuencia de la heterocedasticidad existente es que los intervalos de confianza se sobreestiman en los rangos bajos de las variables explicadas y se subestiman en los altos. En esta versión de dasoLidar se ha optado por llegar a intervalos de confianza orientativos que se traducen en reglas generales para los rangos de valores más habituales de las variables dasométricas (valores medios ± σ)., para los cuales no se presentan los mencionados problemas de sub- o sobre-estimación de los intervalos de confianza.

Entre los inconvenientes, dejando de lado los asociados el hecho de que la suma de los logaritmos no es igual al logaritmo de la suma y eso tiene consecuencias en la forma en que se procesan los datos, el más importante es que la transformación logarítmica incrementa en exceso el peso de los residuos correspondientes a los valores pequeños de las variables explicadas y esto fuerza excesivamente que los modelos ajustados se adapten a esta zona de las nubes de puntos en detrimento de la zona opuesta que es la que, en términos no logarítmicos, tiene más efectos cuantitativos sobre los resultados.

### *Colinealidad*

En las regresiones es conveniente que las variables explicativas no presenten colinealidad entre ellas (no deben estar linealmente relacionadas entre sí). La principal consecuencia de la colinealidad es que impide identificar de forma precisa el efecto individual que tiene cada variable explicativa sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto de que resulta imposible establecer su significancia estadística.

Los efectos sobre la variable explicada son menores, siempre y cuando se no trabaje en zonas marginales del rango de valores de las variables explicativas, que es donde la inferencia es más incierta y lo es tanto más cuanto menos precisos sean los estimadores de los coeficientes, que es uno de los problemas de la colinealidad. Cuando hay colinealidad, pequeños cambios en una de las variables estimadoras pueden generar cambios importantes e inesperados. Esto afecta principalmente a los estimadores de los coeficientes en el ajuste, pero también tiene efectos en la inferencia de la variable predicha.

### Coeficientes de correlación para el conjunto de las métricas Lidar

Se ha analizado la colinealidad de las métricas utilizadas en dasoLidar para el conjunto de las parcelas usadas en dasoLidar, con el resultado de la siguiente página. En esa tabla se recoge el coeficiente de correlación entre las variables explicativas consideradas por pares. Este análisis no aborda la colinealidad que puede haber en de forma conjunta entre tres o más variables explicativas (para ello sería necesario analizar si cada variable explicativa puede expresarse cono una combinación lineal del resto de variables). En todo caso, vale como primera aproximación para darnos cuenta de que existe colinealidad entre algunas de las variables y es conveniente limitar los efectos de esta seleccionando y reduciendo las variables explicativas que intervienen en cada regresión (se ha limitado a cinco seleccionadas mediante el método *forward stepwise*; ver apartado 0 Selección de variables explicativas).

Se ha elaborado esta misma tabla por grupos de parcelas:

- Para cada especie
- Para cada estrato de cada especie

El resultado se recoge en los ficheros

- CoefCorr_VCC_especies.xlsx
- CoefCorr_VCC_estratos.xlsx

Que están en la ruta:

\dasoLidar\varios\modelos&ajustes\correlaciones\EntreVariablesExplicativas_todas

dentro de la ubicación de red de dasoLidar:

\\repoarchivohm.jcyl.red\MADGMNSVPI_SCAYLEVueloLIDAR$

El patrón general de correlaciones se repite entre especies y estratos, como es el caso de la correlación positiva entre algunas métricas principales. El caso más previsible es el de los percentiles de alturas sobre el suelo de primeros retornos entre sí (V01 a V07) y el de los porcentajes de primeros retornos que están por encima de determinadas alturas de referencia entre sí (V08 a V11).

Por otro lado, hay algunas diferencias entre especies, como es el caso de la correlación entre esas métricas principales (V1 a V11) y los porcentajes de retornos en estratos determinados intermedios o inferiores (V12 a V24). Estas correlaciones son negativas de forma consistente en la mayor parte de las especies (excepto V23) y adopta los valores más negativos en las especies más atlánticas (haya, roble). Se salen de este patrón las especies más mediterráneas, como la encina o la sabina, en las que apenas hay correlación entre estos dos grupos de métricas. En el caso de las choperas de producción estas correlaciones también son particularmente bajas.

Las bajas correlaciones juegan en favor de la inclusión de métricas de ambos grupos en las regresiones.

### Coeficientes de correlación para las variables usadas en cada regresión

Las regresiones obtenidas entre las variables dasométricas y las métricas Lidar incluyen un proceso de selección de variables explicativas (métricas Lidar), de forma que éstas cambian de unas a otras regresiones. Por ello también se ha cuantificado el coeficiente de correlación entre las variables explicativas seleccionadas en cada regresión, para el espacio muestral de cada regresión. El resultado se recoge en los ficheros CoefCorr_regresiones_XXX_CyL.xlsx (donde XXX indica la variable explicada elegida en cada regresión -variable dasométrica-) que están en:

\dasoLidar\varios\modelos&ajustes\correlaciones\EntreVariablesExplicativas_deCadaRegresion

dentro de la ubicación de red de dasoLidar:

\\repoarchivohm.jcyl.red\MADGMNSVPI_SCAYLEVueloLIDAR$

Estos ficheros contienen una hoja para cada regresión. El máximo de variables explicativas incluidas en cada regresión es 5. La inclusión de un número elevado de variables explicativas aumenta el riesgo de incluir variables que están correlacionadas entre sí con las consecuencias negativas que ello conlleva. Esta es la principal razón para restringir el número de variables explicativas por regresión.

**Coeficientes de correlación entre variables explicativas para el conjunto de las parcelas (10.064 registros)**

| | V01 | V02 | V03 | V04 | V05 | V06 | V07 | V08 | V09 | V10 | V11 | V12 | V13 | V14 | V15 | V16 | V17 | V18 | V19 | V20 | V21 | V22 | V23 | V24 | V25 | V26 | V27 | V28 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **V01** | 0,79 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **V02** | 0,82 | 0,98 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **V03** | 0,83 | 0,95 | 0,98 | | | | | | | | | | | | | | | | | | | | | | | | | |
| **V04** | 0,83 | 0,91 | 0,95 | 0,98 | | | | | | | | | | | | | | | | | | | | | | | | |
| **V05** | 0,82 | 0,86 | 0,90 | 0,95 | 0,98 | | | | | | | | | | | | | | | | | | | | | | | |
| **V06** | 0,80 | 0,89 | 0,91 | 0,89 | 0,86 | 0,82 | | | | | | | | | | | | | | | | | | | | | | |
| **V07** | 0,79 | 0,85 | 0,87 | 0,86 | 0,84 | 0,81 | 0,99 | | | | | | | | | | | | | | | | | | | | | |
| **V08** | 0,51 | 0,58 | 0,62 | 0,69 | 0,73 | 0,74 | 0,39 | 0,37 | | | | | | | | | | | | | | | | | | | | |
| **V09** | 0,58 | 0,65 | 0,69 | 0,76 | 0,80 | 0,81 | 0,50 | 0,48 | 0,96 | | | | | | | | | | | | | | | | | | | |
| **V10** | 0,63 | 0,71 | 0,75 | 0,81 | 0,84 | 0,85 | 0,59 | 0,56 | 0,91 | 0,98 | | | | | | | | | | | | | | | | | | |
| **V11** | 0,71 | 0,80 | 0,84 | 0,88 | 0,89 | 0,89 | 0,73 | 0,70 | 0,78 | 0,87 | 0,94 | | | | | | | | | | | | | | | | | |
| **V12** | -0,22 | -0,22 | -0,25 | -0,28 | -0,28 | -0,28 | -0,24 | -0,24 | -0,22 | -0,36 | -0,36 | -0,33 | | | | | | | | | | | | | | | | |
| **V13** | -0,29 | -0,28 | -0,31 | -0,32 | -0,31 | -0,31 | -0,34 | -0,35 | -0,03 | -0,31 | -0,37 | -0,38 | 0,56 | | | | | | | | | | | | | | | |
| **V14** | -0,36 | -0,42 | -0,42 | -0,40 | -0,38 | -0,36 | -0,52 | -0,51 | 0,03 | -0,18 | -0,35 | -0,50 | 0,16 | 0,54 | | | | | | | | | | | | | | |
| **V15** | -0,36 | -0,44 | -0,43 | -0,39 | -0,36 | -0,34 | -0,52 | -0,51 | 0,07 | -0,07 | -0,26 | -0,49 | 0,06 | 0,30 | 0,82 | | | | | | | | | | | | | |
| **V16** | -0,37 | -0,44 | -0,42 | -0,38 | -0,35 | -0,33 | -0,54 | -0,53 | 0,15 | 0,07 | -0,06 | -0,41 | -0,01 | 0,14 | 0,49 | 0,73 | | | | | | | | | | | | |
| **V17** | -0,14 | -0,13 | -0,16 | -0,17 | -0,17 | -0,17 | -0,20 | -0,21 | -0,04 | -0,18 | -0,20 | -0,20 | 0,93 | 0,54 | 0,19 | 0,10 | 0,05 | | | | | | | | | | | |
| **V18** | -0,24 | -0,21 | -0,24 | -0,24 | -0,24 | -0,23 | -0,31 | -0,32 | 0,09 | -0,18 | -0,25 | -0,29 | 0,54 | 0,96 | 0,51 | 0,31 | 0,19 | 0,58 | | | | | | | | | | |
| **V19** | -0,35 | -0,38 | -0,38 | -0,36 | -0,34 | -0,31 | -0,51 | -0,52 | 0,12 | -0,10 | -0,26 | -0,43 | 0,17 | 0,54 | 0,95 | 0,79 | 0,53 | 0,22 | 0,58 | | | | | | | | | |
| **V20** | -0,36 | -0,41 | -0,41 | -0,37 | -0,34 | -0,32 | -0,53 | -0,53 | 0,13 | -0,01 | -0,19 | -0,43 | 0,07 | 0,32 | 0,79 | 0,94 | 0,74 | 0,13 | 0,37 | 0,84 | | | | | | | | |
| **V21** | -0,37 | -0,39 | -0,39 | -0,35 | -0,31 | -0,29 | -0,54 | -0,54 | 0,22 | 0,14 | 0,02 | -0,31 | 0,00 | 0,16 | 0,44 | 0,64 | 0,93 | 0,08 | 0,24 | 0,54 | 0,75 | | | | | | | |
| **V22** | -0,41 | -0,44 | -0,47 | -0,48 | -0,48 | -0,46 | -0,52 | -0,53 | -0,20 | -0,42 | -0,51 | -0,56 | 0,49 | 0,74 | 0,68 | 0,47 | 0,25 | 0,47 | 0,71 | 0,67 | 0,47 | 0,24 | | | | | | |
| **V23** | 0,00 | 0,22 | 0,18 | 0,14 | 0,11 | 0,08 | -0,10 | -0,18 | 0,35 | 0,30 | 0,27 | 0,19 | 0,01 | 0,07 | 0,08 | 0,06 | 0,18 | 0,10 | 0,17 | 0,21 | 0,19 | 0,31 | 0,10 | | | | | |
| **V24** | -0,23 | -0,18 | -0,19 | -0,19 | -0,19 | -0,18 | -0,35 | -0,37 | 0,23 | 0,01 | -0,11 | -0,25 | 0,30 | 0,64 | 0,66 | 0,55 | 0,43 | 0,39 | 0,74 | 0,75 | 0,63 | 0,49 | 0,73 | 0,36 | | | | |
| **V25** | 0,20 | 0,41 | 0,40 | 0,40 | 0,39 | 0,39 | 0,10 | 0,02 | 0,62 | 0,61 | 0,59 | 0,50 | -0,13 | -0,06 | -0,04 | -0,02 | 0,11 | 0,00 | 0,07 | 0,12 | 0,12 | 0,29 | -0,10 | 0,84 | 0,29 | | | |
| **V26** | 0,61 | 0,67 | 0,72 | 0,78 | 0,82 | 0,85 | 0,59 | 0,58 | 0,87 | 0,92 | 0,92 | 0,87 | -0,31 | -0,28 | -0,24 | -0,16 | -0,08 | -0,17 | -0,17 | -0,16 | -0,10 | 0,02 | -0,43 | 0,13 | -0,03 | 0,50 | | |
| **V27** | -0,11 | 0,07 | 0,04 | 0,00 | -0,02 | -0,03 | -0,18 | -0,27 | 0,29 | 0,22 | 0,17 | 0,06 | 0,09 | 0,15 | 0,18 | 0,18 | 0,28 | 0,17 | 0,23 | 0,29 | 0,28 | 0,39 | 0,20 | 0,81 | 0,40 | 0,65 | 0,07 | |
| **V28** | -0,23 | -0,06 | -0,12 | -0,16 | -0,17 | -0,17 | -0,32 | -0,38 | 0,18 | 0,09 | 0,03 | -0,10 | 0,16 | 0,23 | 0,27 | 0,28 | 0,36 | 0,22 | 0,30 | 0,36 | 0,36 | 0,45 | 0,30 | 0,71 | 0,45 | 0,52 | -0,04 | 0,93 |

**Variables explicativas**

| | | |
|---|---|---|
| V01 | alt100 | Percentil 100 de altura sobre el suelo (primeros retornos) |
| V02 | alt095 | Percentil 95 de altura sobre el suelo (primeros retornos) |
| V03 | alt080 | Percentil 80 de altura sobre el suelo (primeros retornos) |
| V04 | alt065 | Percentil 65 de altura sobre el suelo (primeros retornos) |
| V05 | alt050 | Percentil 50 de altura sobre el suelo (primeros retornos) |
| V06 | alt035u | Percentil 35 de altura sobre el suelo de retornos excluyendo los que están a menos de 2 m sobre el suelo |
| V07 | alt020u | Idem alt035u, pero percentil 20 |
| V08 | cob050 | % de primeros retornos que están a 0,5 o más m s/suelo |
| V09 | cob2m | % de primeros retornos que están a 2 o más m s/suelo |
| V10 | cob3m | % de primeros retornos que están a 3 o más m s/suelo |
| V11 | cob5m | % de primeros retornos que están a 5 o más m s/suelo |
| V12 | cobPrt_0025_0050 | % de primeros retornos entre 25 y 50 cm s/suelo |
| V13 | cobPrt_0050_0150 | % de primeros retornos entre 50 y 150 cm s/suelo |
| V14 | cobPrt_0150_0250 | % de primeros retornos entre 150 y 250 cm s/suelo |

**Variables explicativas**

| | | |
|---|---|---|
| V15 | cobPrt_0250_0300 | % de 1º retornos 2, 5 - 3,0 m s/suelo |
| V16 | cobPrt_0300_0500 | % de 1º retornos 3,0 - 5,0 m s/suelo |
| V17 | cobTlr_0025_0050 | % de retornos 25 - 50 cm s/suelo |
| V18 | cobTlr_0050_0150 | % de retornos 50 -150 cm s/suelo |
| V19 | cobTlr_0150_0250 | % de retornos 150 - 250 cm s/suelo |
| V20 | cobTlr_0250_0300 | % de retornos 2,5 - 3,0 m s/suelo |
| V21 | cobTlr_0300_0500 | % de retornos 3,0 - 5,0 m s/suelo |
| V22 | PropTl_050cm_200cm_rptoMidTop | Proporción del nº ret. del estrato 50 - 200 cm rpto. al nº ret. que hay por encima de ½ Alt095 |
| V23 | PropTlr_r200cm_midHD_rptoMidTop | Ídem para el estrato2 m - ½ Alt095 |
| V24 | CobTlr_Rango_050cm_200cm | % retornos entre 50 y 200 cm |
| V25 | CobTlr_Rango_200cm_midHD | % retornos entre 2 m y ½ Alt095 |
| V26 | CobTlr_rmidHD_TopHD | % retornos situados a más de ½ Alt095 |
| V27 | RazonDeCopaInf | (alt095 - alt020u) / alt095 |
| V28 | RazonDeCopaSup | (alt100 - alt035u) / alt100 |

*Sesgos y representatividad de la muestra*

Si bien es cierto que no es imprescindible que la muestra con la que construimos un modelo de regresión sea representativa de la población en lo que se refiere a los valores de las variables explicadas (variables dasométricas) si debe serlo en lo que se refiere a la relación entre estas y las variables explicativas.

Con carácter general, una muestra representativa de lo primero suele serlo también de lo segundo y, de hecho, esa es la forma más sencilla de conseguir el objetivo que se pretende. Sin embargo, cuando hay limitaciones en el número de parcelas, hay que dar prioridad al objetivo de que las parcelas permitan modelar la relación entre variables explicativas y explicadas en todo el rango de valores de éstas, es decir, representar suficientemente todo el rango de variación de las métricas Lidar (todo el rango de alturas, espesuras, etc.), eso sí, sin caer en sesgos.

Es importante evitar sesgos que afecten de forma explícita o implícita, a la relación entre variables explicativas y explicadas como los siguientes:

- Si las parcelas se ubican en zonas particularmente transitables, donde hay menos sotobosque, esto puede conllevar una relación entre métricas Lidar y variables dasométricas no representativa de la población que se pretende modelizar.
- Algo parecido ocurre si la selección de las parcelas tiene en cuenta aspectos fisiográficos como la pendiente, altitud o la posición dentro de la ladera.
- Ídem, si la muestra selecciona preferentemente zonas en las que se han hecho (o no) tratamientos selvícolas,
- Igualmente, no debe haber sesgos relacionados con ubicar las parcelas en zonas de densidad particularmente baja o alta, ya que esto condiciona la relación entre métricas Lidar y variables dasométricas.

En dasoLidar se han usado las parcelas del IFN4, que están ubicadas en la malla UTM de 1 km, con lo que la objetividad y representatividad de la muestra está bastante garantizada. De hecho, esta muestra no solo da garantías de ausencia de sesgos, sino que también aporta representatividad respecto a la población que se modela. Esto da un colchón de seguridad para paliar algunas posibles deficiencias de las regresiones.

*Rangos de validez de las variables explicativas*

Ver ficheros auxiliares

*Independencia de la muestra*

Cada elemento de la muestra debe elegirse de forma independiente a todos los demás.

Problema en las parcelas: desplazamientos (ya que estoy aquí hago otras parcelas cerca: conglomerados de parcelas)-> cuidado en la inferencia, especialmente en los intervalos de confianza.

## Inferencia (predicción, estimación, etc.)

Valor estimado para cada X:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Diferenciar entre la varianza (dispersión) en la estimación de:

- Intervalo de confianza para las predicciones puntuales:

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

Nota: esta expresión es para una sola variable explicativa; para p variables: donde dice n-2, debe decir n-p-1:

$$\hat{y}_i \pm t_{\alpha/2,n-p-1} \cdot \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \sum_{j=1}^{p} \frac{(x_{ij} - \bar{x}_j)^2}{S_{xj}^2}\right)}$$

Donde:

- *y*^*i* es la predicción en el punto i: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$
- (xi1,xi2,…,xip)(*xi*1,*xi*2,…,*xip*).
- tα/2,n−p−1*tα*/2,*n*−*p*−1 es el valor crítico de la distribución t de Student con n−p−1*n*−*p*−1 grados de libertad.
- σ^2*σ*^2 es la estimación de la varianza del error.
- n*n* es el número de observaciones.
- x¯j*x*¯*j* es la media de la variable explicativa xj*xj*.

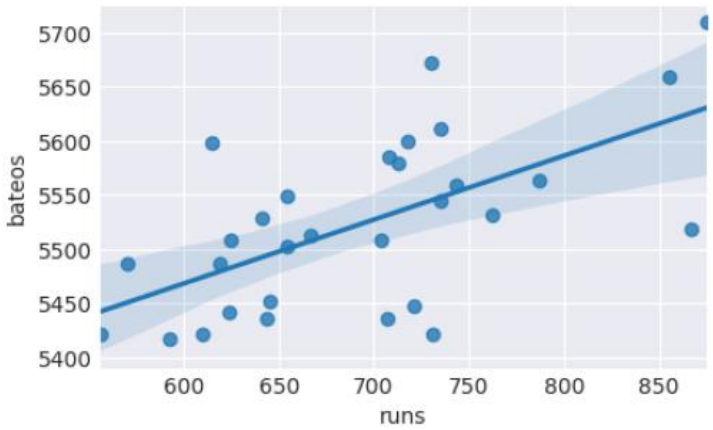
También se puede expresar matricialmente a partir de la matriz X referida anteriormente, con la que se ajusta el modelo) y $x_i$, correspondiente a la lista de valores de las variables explicativas para el ejemplo i

$$\hat{y}_i \pm t_{\alpha/2,\, n-p-1} \cdot \sqrt{[\sigma^2 \cdot (1 + \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i)]}$$

- Intervalo de confianza para estimación de la media poblacional

$$\hat{y} \pm t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x}^2)}\right)}$$

Los intervalos se amplían conforme nos alejamos de las medias muestrales de X:




## Validación del modelo

*Partición de la muestra train/test*

Partimos la muestra en dos grupos: uno para construir el modelo y otro para validarlo. Normalmente 80%/20%.

Puede ayudar a detectar sesgos o cuestiones que se nos han escapado porque no tenemos buena base estadística o no verificamos algunas hipótesis en las que se basa la predicción.

Como interpretamos la validación? Significación estadística. Puede ayudar a volver atrás para revisar errores o posibles malas prácticas.

*Outliers*

Hay que analizarlos, uno por uno y ver si son errores o no, y en este caso pueden aportar información relevante para interpretar el modelo

## Ámbito de aplicación – espacio muestral

Se adopta como hipótesis que esta relación es válida para un determinado ámbito de aplicación, que queda definido por:

- La especie principal de la parcela
- Un ámbito territorial y de tipo de gestión dentro de cada especie
- El rango de valores de validez de las variables explicativas

Por lo tanto, es necesario definir los distintos ámbitos de aplicación, con el objeto de llevar a cabo, en cada uno de ellos, los correspondientes ajustes de regresión de forma independiente: cada ámbito de aplicación constituye la población en la cual se obtiene una muestra con la que se construye el modelo lineal.

## Selección de variables explicativas

Mientras que las variables explicadas son las mismas para todos los casos (ver **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**), las variables explicativas seleccionadas para el modelo lineal cambian de uno a otro.

Aunque solo sea por azar, cualquier variable que incluyamos en el modelo mejora el porcentaje de varianza explicada, pero eso no significa que sea adecuado incluirla.

Hay que evitar el vicio de incluir variables de más.

La selección de variables explicativas (entre las disponibles, que se enumeran en el apartado **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**) se lleva a cabo por el método forward stepwise usando el criterio de Akaike (AIC: Akaike Information Criterion)[8] y restringiendo el número máximo de variables explicativas a 5 (número máximo de parámetros estimados: 6).

Métodos de selección de variables:

- Selección paso a paso
  - Forward
  - Backward
  - Mixto

Criterios de selección de variables:

- Cp
- AIC
- BIC
- R2ajustado.

Yo uso el método Akaike (AIC), que tiende a ser más restrictivo e introducir menos predictores que el R2-ajustado. En todo caso pongo tope de antemano. Tengo pendiente el tratamiento adecuado de los valores nulos y noData de algunas variables de explicativa.

[8] Ver ejemplo de uso y referencias en http://verso.mat.uam.es/~amparo.baillo/MatEstII/RegMultVarSel.html

## Significación de los estimadores de los parámetros

*Cuantía*

Que sea pequeño (0.000041) no indica que sea poco significativo xq depende de la magnitud empleada para las variables explicativas

*Significación estadística*

Indica si es significativamente distinto de cero, es decir, si la correspondiente variable aporta algo (influye, o explica parte de la variación de y)

-> t de Studet -> significación (p value) -> Intervalos de confianza para los parámetros

$$\hat{\beta}_j \pm t_{df}^{\alpha/2} SE(\hat{\beta}_1)$$

## Explicaciones para alumnos

*Planteamientos*

Objetivo de la regresión:

- Ver cómo influyen un conjunto de variables explicativas (predictoras, regresoras, independientes, de entrada, features, X) en una variable explicada (respuesta, predicha, dependiente, de salida, y): si las variaciones de la variable respuesta se pueden explicar o relacionar con las explicativas. Calcular en qué medida afecta cada x a la y si esa afección, relación o influencia es significativa.
- Calcular el valor de Y a partir de las X.
  - Inferir (estimar) valores concretos
  - Inferir (estimar) medias

En toda regresión hay una parte no explicada por el modelo: componente que a veces llamamos aleatoria (no lo es en la práctica, pero sí en el modelo teórico). Es el residual y lo caracterizamos con el ECMR (mse). El error estándar (su raíz cuadrada) es una magnitud fundamental y se expresa en las unidades de la variable explicada (p. ej. m3/ha)

Cuando decimos que medimos el volumen de un árbol, no es cierto, lo estimamos y eso tiene un margen de error: error de regresión que expresamos con el ECMR.

Atención porque en los cálculos de volumen de un árbol hay otra fuente de error que es muy importante y que normalmente no tenemos bien acotada: la relación altura diámetro -> Ver conclusiones de los cálculos de Zamora.

Otra forma de expresar cómo de buena es la regresión es la $R^2$: porcentaje de varianza explicada por el modelo. El resto de la varianza se debe a otras causas no incluidas en el modelo (no es realmente aleatoria). Como el $R^2$ es un porcentaje parece más fácil de interpretar que el ECMR, pero lo que más nos interesa es el error estándar.

Las variables X que nos ayudan a explicar las variaciones de y (estimar el valor de y para un caso concreto) pueden ser varias -> regresión múltiple X: x1, x2, x3, etc.

Si tenemos muchas candidatas a variable explicativa corremos el riesgo de pensar que cuantas más, mejor. Pero no es así: el exceso puede ser más peligroso que el defecto por efectos indeseados.

En regresión hay una fase que es construir el modelo (elegirlo y calcular los estimadores de los parámetros y otra que es la inferencia (es más automática, pero cuidado con las variables explicativas que tienen margen de error).

La regresión lineal simple es fácil de representar e inspeccionar visualmente. La múltiple lo es un poco menos, pero también se puede inspeccionar visualmente, especialmente los residuos frente a cada una de las variables explicativas.

No se debe esconder una mala gestión estadística detrás de los números, coeficientes y parámetros estadísticos. Una vestimenta estadística elegante puede esconder una mala base conceptual y práctica.

*Regresión versus muestreo*

Dos campos de la estadística, que tienen sus planteamientos, sus bases estadísticas, sus procedimientos, etc. Pero reflexionando sobre esta cuestión en el caso de las IOFs me ha llevado a integrar los dos campos, porque hay veces que jugamos a hacer modelos de regresión y en realidad estamos flirteando con el muestreo:

*regresiones débiles se apañan si la muestra con la que se construye la regresión es representativa*

En una regresión hay una componente que es el intercept y que es muy importante en estos casos, y que adquiere una interpretación concreta si trabajamos con variables explicativas normalizadas (o calculamos la y para los valores medios de las X): podemos vestir un muestreo de regresión y eso no es bueno, al menos si no se hace explícitamente.

Gráfica de nube de puntos poco relacionada con la X.

La interpretación puede ser válida si la aplico a todo el ámbito con el que he construido la regresión, pero no a un rodal concreto.

*El espacio muestral: población y muestra*

Más importante que el número es la representatividad.

Media poblacional y media muestral (estimadora de aquella)

Varianza poblacional y varianza muestral (estimadora de aquella)

*Muestreo en ámbito territorial*

Uno de los requisitos del muestreo es que las muestras sean aleatorias o por lo menos no estén relacionadas entre sí: en el territorio las muestras son más parecidas cuanto más cerca estén y eso es difícil de cuantificar y de incorporar al modelo.

Una muestra muy grande pero concentrada en una zona o en un tipo de lugares tiene menos valor que una más pequeña, pero sin esos sesgos.

*Elección de la muestra en regresión y en muestreo*

Elección - aleatoriedad – representatividad

- En regresión es importante que la variable explicada no condicione la muestra porque eso introduce sesgos, sin embargo, no lo es tanto para las variables explicativas (salvo que juguemos a hacer regresión cuando en realidad estamos haciendo muestreo).
- En muestro, es muy importante la representatividad, que se consigue con aleatoriedad

Importante es espacio muestral: tiene que estar bien definido y debemos muestrearlo en todo su ámbito (territorial o dasométrico) **para ambos casos**.

## Hay que definir bien la variable que estimamos

La variable no queda definida sólo indicando el parámetro, sino tb u resolución espacial y demás detalles de cómo se obtiene.

## Escala (resolución espacial) y representatividad del “caso” individual

Cuando hablamos del valor de una variable dasométrica en un punto es una falacia: siempre es en una porción de terreno.

No es lo mismo que la variable explicada sea el volumen por hectárea en una parcela de 10 m de radio que en una de 25. La esperanza matemática (la media poblacional o su estimador que es la

muestral) pueden ser iguales pero sus varianzas son muy diferentes y si queremos acotar el error de estimación hay que concretar esta cuestión y ser consecuentes con ello.

En un ráster dasométrico, un pixel no es un punto sino una celda: ¿Cuánta heterogeneidad espacial queremos integrar? Eso influye en el tamaño del pixel que más nos interesa.

Visualizarlo en Qgis

## Como hacemos una regresión

Se puede hacer con una aplicación específica de estadística (de pago: spss, sas, statgraphics, etc.; libres: jasp, jamovi, etc.), con código (paquetes para python, R, Matlab) o, simplemente, con Excel.

No es mejor una regresión si se hace con un software más potente: el modelo ajustado debe ser el mismo en todos: cambia el uso, la interfaz, la información que aporta, etc.

## Relación lineal entre variables

La linealidad se refiere a los parámetros

En regresión múltiple podemos chequear la linealidad visualizando los residuos frente a las variables o con determinados estadísticos.

*Transformación de las variables explicativas*

Sin problema en regresión

*Transformación de las variables explicada*

Cuidado: factor de corrección para evitar sesgos

*Linealización del modelo*

y = A xB -> log(y) = log(A) + B log(x) -> YY = AA + B XX

Cuidado con los cálculos cuando se deshace la linealización -> sesgos