UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA

http://dyna.medellin.unal.edu.co/

Article formatting example

Ejemplo del formato de un artículo

Carlos Mario Henao Garrido

Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. cmhenaoga@unal.edu.co

Recibido: 09 de diciembre de 2022.

Abstract
Based on the need to have highly detailed precipitation data and good representativeness of the phenomena that occur inside an interest zone, this document presents the analysis carried out on various meteorological, topographic and location variables, in the objective of performing down-scaling of satellite precipitation data. The above applying knowledge of machine learning techniques and making use of the Python 3 programming language.

Keywords: machine learning, down scaling, unsupervised methods, supervised methods, database construction.

Resumen
Partiendo de la necesidad de contar con datos de precipitación de alto detalle y buena representatividad de los fenómenos que suceden al interior de una zona de interés, en el presente documento se presenta el análisis realizado a diversas variables meteorológicas, topográficas y de localización, en el objetivo de realizar sub-escalamiento de datos de precipitación satelital. Lo anterior aplicando conocimientos de técnicas propias de machine learning y haciendo uso del lenguaje de programación Python 3.

Palabras clave: machine learning, down scaling, métodos no supervisados, métodos supervisados, construcción de bases de datos.

1 Introducción

En geociencias y concretamente en el desarrollo de modelos hidrológicos e hidráulicos, es cada vez más común la utilización de modelos distribuidos en busca de lograr un mayor conocimiento y representatividad sobre los fenómenos que se desarrollan sobre el área de interés, sin embargo, resulta común que la falta de datos de precipitación genere grandes dificultades para la obtención de dichos modelos; y, si bien la información satelital aparece como una opción, típicamente es insuficiente por temas de detalle para los requisitos del modelo demandado.

Así, en este artículo se propone la construcción de modelos de reducción de escala aplicando métodos de machine learning, partiendo de relaciones entre la precipitación tomada satelitalmente, asumiendo que esta guarda bien la variabilidad espacial de la precipitación e información ambiental relevante como la topografía y condiciones meteorológicas medidas mediante estaciones en tierra, estás últimas serán las condiciones de humedad, temperatura y precipitación.

2 Zona de estudio

La zona de estudio, (Figura 1) es la cuenca Aburrá, ubicada en el departamento de Antioquia, Colombia y que comprende a los municipios de Caldas, Envigado, Itagüi, Medellín, Bello, Copacabana y Barbosa, entre otros. Cuenta con una extensión total de 1207 km², siendo esta una zona montañosa con un valle encajado en el centro de la cuenca.

Esta cuenca ha sufrido un fuerte proceso de antropización pues habitan en ella más de 4 millones de personas, siendo el municipio de Medellín la segunda ciudad más poblada del país. Dicha antropización y consecuente impermeabilización de los suelos, adicional a los efectos de cambio climático ha generado que los eventos extremos sean cada vez más frecuentes, por que haya que replantearse los análisis hidrológicos e hidráulicos realizados hasta el momento.

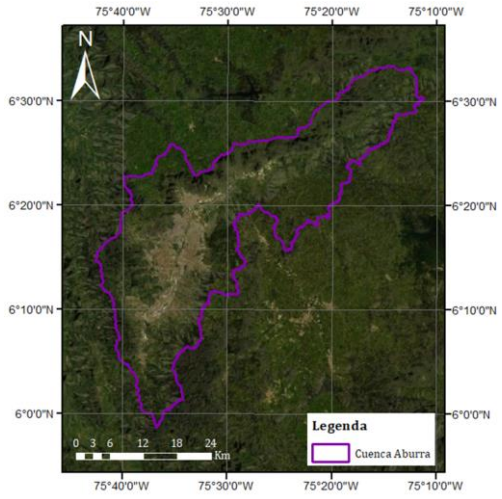

Figura 1. Delimitación de la zona de interés.
Fuente: Elaboración propia.

Se escoge esta cuenca dado que cuenta con una red de monitoreo altamente sofisticada, manejada por el SIATA, que capta información de reflectividad horizontal (dBZ), la cual puede ser transforma, mediante ecuaciones ya validadas para la zona en concreto, en intensidad de precipitación (R), siendo esta la variable objetivo 'Y'. Por otra parte, se cuenta con una red de estaciones relativamente amplia que registran información meteorológica desde tierra, contando así con todos los insumos desde los que parte esta propuesta.

3 Base de datos

La conformación de una adecuada base de datos es uno de los mayores retos a afrontar. Primero se debe tener en cuenta que los métodos que se desea implementar exigen un formato concreto para la base de datos. Según este, cada columna representa una variable descriptiva y cada fila una observación de la variable, tal como se muestra en la Tabla 1, se debe garantizar, además, que la base de datos no tenga datos tipo NaN, y preferiblemente eliminar tantas inconsistencias como sea posible, evitando se reflejen en el futuro modelo.

Tabla 1. Formato requerido para la base de datos.

Observación	X_1	X_2	…	Xn
1	x_{11}	x_{21}	…	x_{n1}
2	x_{12}	x_{22}	…	x_{n2}
⋮	⋮	⋮		⋮
n	x_{1n}	x_{2n}	…	x_{nn}

Fuente: Elaboración propia.

Pensando en esto y dado que los datos a trabajar varían tanto espacialmente como temporalmente, pues se tiene un mapa de dimensiones n×m para cada fecha por cada variable. Se realiza una vectorización "Ravel()" de los mapas y se acoplan de tal forma que i-ésimos n×m datos conforman el mapa de datos de la fecha i. Se aclara que se una variable no cambia temporalmente, como es el caso de la elevación topográfica, esta se acopla de forma repetida, respetando así el formato mencionado para la base de datos.

Ahora bien, se pasa a revisar la variables descriptivas.

3.1 Precipitación obtenida desde el SIATA, 'Y'

Desde la entidad SIATA se entrega información de reflectividad horizontal (dBZ), 5 minutal y con distribución espacial de 125 m. que debe ser transformada por el usuario en precipitación, y hacia la escala espacio-temporal deseada.

La ecuación para transformar reflectividad horizontal (dBZ) en intensidad de precipitación horaria (R), es la ecn. (1) presentada a continuación.

$$Z = 250 \cdot R^{1.2} \ \wedge \ Z = 10^{dBZ/10} \quad (1)$$

Donde Z tiene unidades de (mm^6 / m^3) y R es la precipitación diaria en (mm / hr).

Así para obtener la precipitación diaria se debe realizar la transformación a intensidad indicada y dividir por 12 para obtener intensidad 5 minutal (frecuencia de los datos). Finalmente, sumando la información del día, se obtiene la precipitación deseada, en mm/día.

3.2 Precipitación satelital obtenida desde ERA5

El ERA5 entrega información de precipitación horaria, por lo que basta con sumar los datos del día correspondiente para obtener la precipitación en escala diaria. El reto en este caso es la resolución de la información, motivación de este trabajo, pues la precipitación viene dada en celdas de 0.25°, mientras que se desea información en celdas de 0.01° (≈ $1km^2$).

Luego lo que se hace es simplemente crear un mapa con pixeles del tamaño deseado, donde el valor de estos será igual al del píxel de mayor tamaño (0.25°) que los contenga, es decir quedarán muchos pixeles de menor tamaño con el mismo valor.

3.3 Humedad, Temperatura y Precipitación

Para las variables de humedad, temperatura y precipitación, dado que se cuenta con información de estaciones puntuales, en principio no se cuenta con un mapa que describa el comportamiento de dichas variables en la extensión total de la cuenca. Es así, que con ayuda de la librería 'Pykrige', se decide interpolar los datos de las estaciones, haciendo uso de la teoría de variogramas, y asumiento un ajuste Gaussiano, obteniendo mapas de distribución en la cuenca para cada variable, este proceso es similar a lo que realiza el conocido método de ArcGIS 'Krigging'.

3.4 Elevación y Localización

Por último, para completar la base de datos, se añaden las variables Elevación, Coordenada X y Coordenada Y, las cuales, como se había mencionada antes, se repetirán para todas las fechas pues no cambian en el tiempo.

La variable elevación fue obtenida del DEM de 30 metros suministrado por el recurso cartográfico El ALOS PALSAR DEM, y reescalada a la resolución deseada 0.1°.

Las variables de Coordenadas X y Y, si bien no son de carácter meteorológico ni topográfico, son incluidas dentro de la base de datos pensando en que puedan describir procesos diferenciados de precipitación entre los extremos Sur-Norte y Este-Oeste de la cuenca.

4 Metodología

4.1 Análisis exploratorio de los datos

Del análisis exploratorio de los datos es posible conocer de forma detallada el comportamiento de las variables con que se trabaja, hablando de su media, varianza o si se comportan de forma normal, su asimetría e importancia de los valores extremos en la distribución de los datos, entre muchas otras características. Sin embargo, para las intenciones de este trabajo lo primordial es el comportamiento de las variables en su conjunto, entender cómo se relacionan unas con otras, siendo esto de lo que nos habla la matriz de correlación (Figura 2).

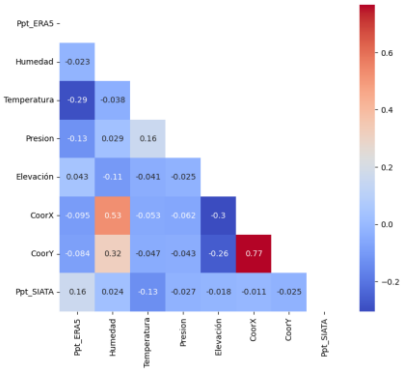

Figura 2. Matriz de correlación.
Fuente: Elaboración propia.

De la Figura 2 es posible apreciar como hay una relación considerable entre la precipitación obtenida desde el SIATA (variable objetivo, 'Y'), y la precipitación aportada por el ERA5, lo cual refuerza la hipótesis de que los datos satelitales, reflejan bien la variabilidad de la precipitación en la cuenca. Se observa también una correlación significativa entre la precipitación SIATA y la temperatura, mostrándose la temperatura como una variable descriptora de gran relevancia.

Se comenta además, que se realizó un análisis por componentes principales 'PCA', encontrando que las variables de Coordenadas aportan una significativa varianza al conjunto de datos, es decir, información nueva no contenida por las demás variables.

4.2 Modelos

Con miras a encontrar un modelo que, en base a los datos disponibles, realice un sub-escalamiento representativo de los procesos que se dan en la cuenca de interés, se evalúan 5 métodos de machine learning, los cuales son Regresión Lineal, K-nearest neighbors (KNN), Support Vector Machine (SVM), Redes Neuronales y Métodos ensamblados.

Se aclara que las métricas obtenidas, resultan tras un proceso de validación cruzada (cv) de 5. A continuación, se describe brevemente el proceso llevado a cabo en cada uno de los métodos, donde la principal librería de apoyo es 'sklearn'.

4.2.1 Regresion Lineal

Se realiza tanto un análisis de regresión univariado como multivariado.

4.2.1.1 Análisis de regresión univariado

Se comienza por realizar un análisis de regresión univariado, buscando entender más de las relaciones individuales entre la precipitación asumida como real y las variables explicativas, más únicamente se logra concluir que el error residual no se distribuye de forma normal, además de que una única variable resulta insuficiente para describir de forma acertada la precipitación.

4.2.1.1 Análisis de regresión multivariado

Para el análisis multivariado se hace uso de la librería 'statsmodel' que permite obtener un resumen muy últil con información detallada sobre la implementación y evaluación del modelo.

Como primer aspecto, se encontró que la presión no es una buena variable descriptiva, por lo que se deja de tener en cuenta dicha variable, luego se realizaron varias combinaciones, aplicando transformaciones lineales y no lineales a las variables, encontrando, que el mejor desempeño se presentaba al tener en cuenta todas las variables (menos presión), sin realizar combinaciones entre ellas y eliminando el intercepto. Este último aspecto es polémico dado que, aunque mejoren las métricas, carece de sentido físico construir un modelo cuyo intercepto es cero, más por el bajo rendimiento de las demás combinaciones se decidió seleccionarlo.

Se aclara que se opta por no hacer uso de los métodos de regularización Lasso y Ridge, dado que tras realizar la selección de hiperparámetros se obtuvo que el mejor parámetro es $\alpha = 0$, lo que es equivalente hacer uso de regresión lineal, careciendo de sentido implementar dichos métodos.

4.2.2 K-nearest neighbors

De manera iterativa se encuentra que el mejor resultado para KNN se obtiene haciendo uso de un parámetro n_neighbors ≈ 30, sin embargo no se logra obtener un buen desempeño y puede estar relacionado con un hecho visto mediante análisis por Cluster y es que los datos no son fácilmente separables, por lo que el método de vecino cercano no resulta ser una buena alternativa, a pesar de lo anterior a este método también se le presentaran sus resultados.

4.2.3 Support Vector Machine

Para el método de Support Vector Machine, no se lograron obtener métricas mínimamente aceptables con las combinaciones de parámetros ensayadas, por lo que para el presente método, en la sección de resultados, únicamente se presentarán sus métricas, más no los resultados de la ejecución del modelo.

4.2.3 Redes Neuronales

En el caso de redes neuronales se logró estabilizar las métricas haciendo uso de 5 capas oculta y 7 neuronas por capas, además de fijar un máximo de 500 iteraciones, pues valores mayores a los mencionados ya no producían una mejora apreciable en el desempeño del modelo.

4.2.3 Métodos Ensamblados

Por último, para la implementación de métodos ensamblados se logró una buena respuesta haciendo uso de un número de iteraciones igual a 50, y una validación cruzada para los métodos de 3, lo que arroja un total de 150 modelos de ajuste. Es posible que con mayor número de combinaciones se logre mejorar lo obtenido, más se requiere de alta capacidad computacional, además de que el desempeño del modelo con los parámetros mencionados es realmente notable.

5 Resultados y discusión

En la Tabla 2 se presenta de forma resumida las métricas obtenidas para cada uno de los métodos empleados.

Tabla 2. Desempeño de los modelos analizados.

	Reg. Lineal	KNN	SVM	Redes Neuronales	Métodos Ensamblados
Score R2	0.219	0.028	-0.075	0.154	0.218
Std	…	0.003	0.001	0.009	0.010

Fuente: Elaboración propia.

En función de lo que dicen las métricas el mejor método sería regresión lineal, sin ser ninguna métrica realmente aceptable, más en este caso, es realmente importante observar que están devolviendo como producto cada uno de los modelos construidos, como representan la variabilidad espacial de la precipitación al interior de la cuenca. Resultados que se presentan en Figura 3.

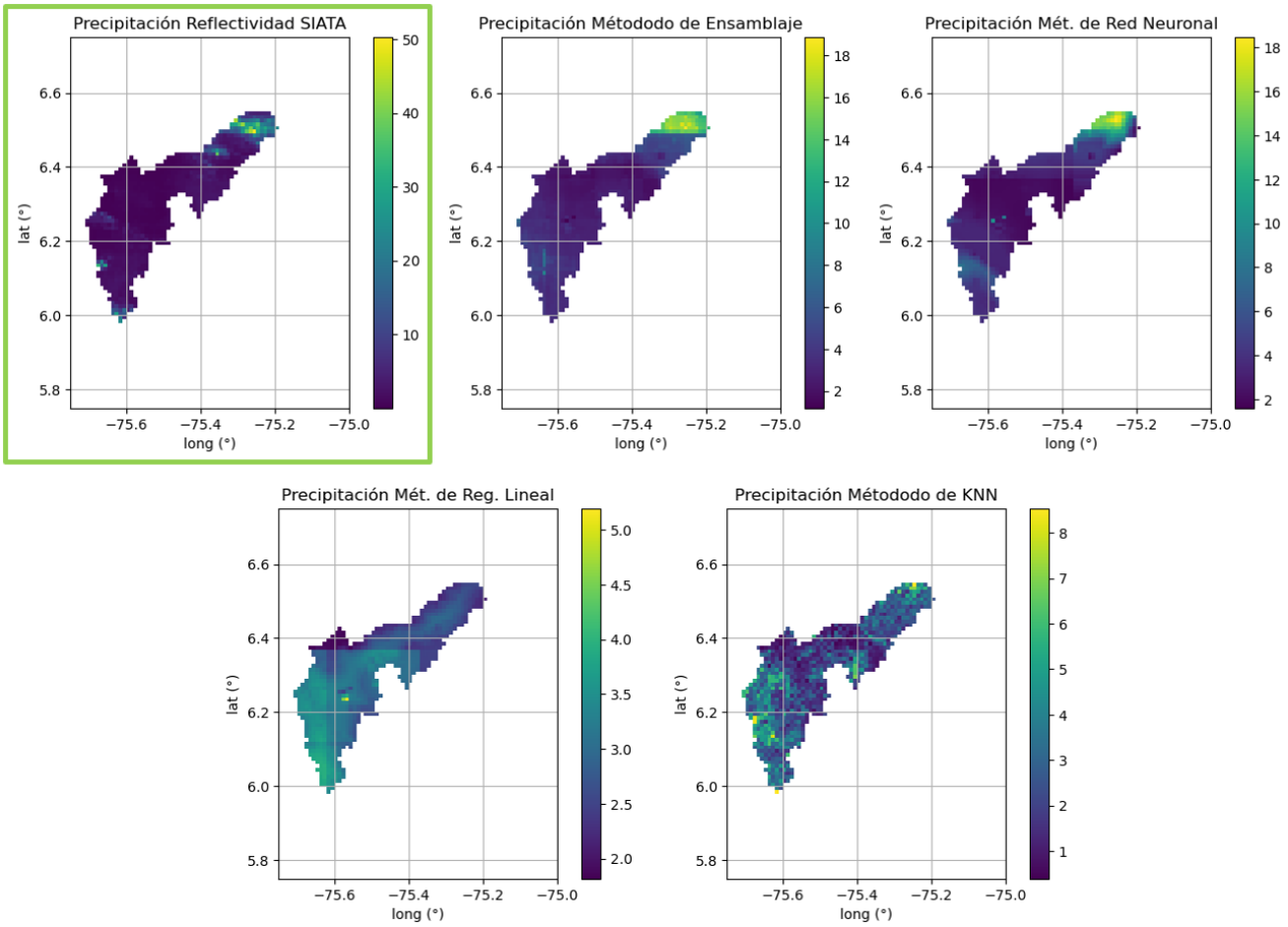

Figura 3. Precipitación predicha por modelos empleados.
Fuente: Elaboración propia.

Apoyado por los resultados gráficos es posible decir que efectivamente el método de KNN, es el peor de los expuestos en el objetivo de hacer sub-escalamiento de la precipitación. Pero también que aunque las métrica de regresión lineal sea mejor a las obtenidas para los métodos de redes neuronales y ensamblaje, estos dos métodos logran una representación del fenómeno mucho más acertada, capturando de mejor manera valores extremos, mínimos y máximos, pues al modelo obtenido mediante regresión lineal le cuesta mucho alejarse de la media de los datos, probablemente provocando un aumento en sus métricas más no resultando en el comportamiento deseado.

Es propicio decir, que si se analizan los resultados obtenidos para diferentes fechas es posible apreciar como los resultados obtenidos por el método de ensamblaje son sumamente representativos del comportamiento de la precipitación en la cuenca, por lo cual, a pesar de ser un método computacionalmente muy demandante, parece ser también una excelente alternativa para dar solución a este problema de reducción de escala espacial.

5 Conclusiones

- A partir de técnicas de machine learning es posible dar solución a problemas de sub-escalado de información en base a variables descriptivas.
- El método de ensamblaje, aunque es muy costoso computacionalmente, es sumamente poderoso en la resolución de problemas de variabilidad espacial.
- El uso de información satelital puede convertirse en una alternativa viable, para aportar variabilidad espacial a bases de datos limitadas.
- Se presentó un buen comportamiento en torno a la hipótesis de que la información de precipitación tomada satelitalmente captura bien la variabilidad espacial de la precipitación real.

6 Referencias

[1] Fang, J., Du, J., Xu, W., Shi, P., Li, M., & Ming, X. (2013). Spatial downscaling of TRMM precipitation data based on the orographical effect and meteorological conditions in a mountainous area. Advances in Water Resources, 61, 42-50.