**DIAGNÓSTICO DEL ESTADO ACTUAL Y DE RESULTADOS DE CONTROLES DE CALIDAD DE DATOS DE ESTACIONES METEOROLÓGICAS AUTOMÁTICAS, VARIABLES PRESIÓN ATMOSFÉRICA, HUMEDAD RELATIVA, TEMPERATURA Y HUMEDAD DEL SUELO Y EVAPORACIÓN**

**Paola Andrea Álvarez Betancourt**

Profesional Contratista

Contrato No. 196 de 2024

Grupo de Gestión de Datos y Red Meteorológica

Subdirección de Meteorología

Instituto de Hidrología, Meteorología y Estudios Ambientales

IDEAM

**Entregable:**

“Informe sobre el diagnóstico actual y de resultados de controles de calidad de las series de datos procedentes de estaciones automáticas, específicamente, de las variables: presión atmosférica, humedad relativa, temperatura y humedad del suelo y evaporación, última variable cuyo diagnóstico tendrá fines exploratorios”

Bogotá, D.C., \_\_ de diciembre de 2024

LISTA DE CONTENIDOS

[1. OBJETIVOS 5](#_Toc170999345)

[1.1. OBJETIVO GENERAL 5](#_Toc170999346)

[1.2. OBJETIVOS ESPECÍFICOS 5](#_Toc170999347)

[2. METODOLOGÍA 5](#_Toc170999348)

[3. RESULTADOS 6](#_Toc170999349)

[3.1. PRESIÓN ATMOSFÉRICA 6](#_Toc170999350)

[3.1.1. Exploración inicial de datos crudos 6](#_Toc170999351)

[3.1.2. Análisis descriptivo 6](#_Toc170999352)

[3.1.3. Análisis diagnóstico 8](#_Toc170999353)

[3.2. Humedad relativa 8](#_Toc170999354)

[3.2.1. Exploración inicial de datos crudos 8](#_Toc170999355)

[3.2.2. Análisis descriptivo 8](#_Toc170999356)

[4. ANEXOS 10](#_Toc170999357)

[5. BibliograFÍA 10](#_Toc170999358)

[6. HISTORIAL DE CAMBIOS 11](#_Toc170999359)

**LISTA DE FIGURAS**

[Figura 1 Metodología para análisis diagnóstico de datos de EMA de variables meteorológicas de interés. 6](#_Toc170999360)

[Figura 2 Periodos y cantidades de datos de presión atmosférica de EMA analizados en el contrato 196 de 2024. 6](#_Toc170999361)

[Figura 3 Estaciones de atención por lo obtenido en estadísticos descriptivos, presión atmosférica. 7](#_Toc170999362)

[Figura 4 Periodos y cantidades de datos de variable humedad relativa de EMA analizados en el contrato 196 de 2024. 8](#_Toc170999363)

[Figura 5 Estaciones de atención por lo obtenido en estadísticos descriptivos, humedad relativa. 9](#_Toc170999364)

[Figura 5 Estaciones de atención por lo obtenido en estadísticos descriptivos, humedad relativa. 10](#_Toc170999365)

**LISTA DE TABLAS**

[Tabla 1 Descripción de la información exploratoria o descriptiva de los datos crudos de presión atmosférica provenientes de EMA. 5](#_Toc170999366)

[Tabla 2 Descripción de la información exploratoria o descriptiva de los datos crudos de humedad relativa provenientes de EMA. 7](#_Toc170999367)

**LISTA DE SIGLAS, ACRÓNIMOS Y ABREVIATURAS**

|  |  |
| --- | --- |
| CDMS | Sistema de gestión de datos climáticos (siglas en inglés) |
| EMA | Estación(es) meteorológica(s) automática(s) |
| GGD | Grupo de Gestión de Datos y Red Meteorológica |
| IDEAM | Instituto de Hidrología, Meteorología y Estudios Ambientales |
| OMM | Organización Meteorológica Mundial |
| QC | Quality Control, ‘Controles de calidad’, en español. |
| QCS | Quality Control System, ‘Sistema de controles de calidad’, en español |

**DIAGNÓSTICO DEL ESTADO ACTUAL Y DE RESULTADOS DE CONTROLES DE CALIDAD DE DATOS DE ESTACIONES METEOROLÓGICAS AUTOMÁTICAS, VARIABLES PRESIÓN ATMOSFÉRICA, HUMEDAD RELATIVA, TEMPERATURA Y HUMEDAD DEL SUELO Y EVAPORACIÓN**

El análisis de datos, según Dibekulu (2020) sean los cualitativos o cuantitativos, puede consistir en describir y resumir los datos, identificar relaciones entre variables, compararlas, identificar sus diferencias y pronosticar resultados; es un proceso de inspección, limpieza, transformación y modelado de datos con el objetivo de descubrir información útil, sugerir conclusiones y apoyar la toma de decisiones; ante los millones de datos que se pueden obtener, el objetivo de analizar datos es obtener información que se pueda usar y que sea útil. Así, Cote (2021) expone que el análisis de datos se usa para responder preguntas, identificar tendencias y extraer ideas; posee 4 tipos clave: el análisis descriptivo (responde la pregunta ¿qué pasó?), el diagnóstico (¿por qué?) el predictivo y prescriptivo (¿qué podría pasar y qué se debe hacer?, respectivamente), últimos dos que no se abordarán en este informe.

El análisis de diagnóstico es el proceso de utilizar datos para determinar las causas de las tendencias y las correlaciones entre variables; puede verse como el siguiente paso lógico después de utilizar análisis descriptivos para identificar tendencias (Cote, 2021). Comúnmente, comprende la ***evaluación de hipótesis***, que en este caso puede ser “es posible que los datos provenientes de estaciones meteorológicas automáticas (EMA) de las variables de interés tengan datos erróneos”; la ***correlación versus causalidad,*** enmarcada en variables meteorológicas que posiblemente estén correlacionadas y que puedan significar o no causalidad de una con respecto a la otra; el ***análisis de regresión diagnóstica*** que permiten ver relaciones (en casos que pueden no ser tan evidentes) entre dos variables mediante regresión lineal única o múltiple (Cote, 2021).

Con lo anterior, se pretende dilucidar al lector los contenidos que observará en este documento y comprender más fácilmente sus objetivos. Este análisis diagnóstico de datos provenientes de EMA de la red de observación en superficie IDEAM y de otras entidades, prosigue al análisis descriptivo que se mencionará al inicio de cada sección por variable, denominado como diagnóstico inicial, para luego evidenciar el por qué mediante el análisis diagnóstico a los controles de calidad que se apliquen a las series temporales de cada variable.

La OMM (2021) establece que los controles de calidad (QC, por sus siglas en inglés) pueden llevarse a cabo en el punto de observación, las bases de datos anteriores, el punto de captación y en el ‘sistema de gestión de datos climáticos’ (CDMS, por sus siglas en inglés) o “en modo diferido”, es decir, aquellos conjuntos de datos o series temporales ya alojadas en los sistemas de información o repositorios de los Servicio Meteorológico e Hidrológico Nacional (SMHN), en este caso, del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM); es decir, los QC aplicados a las series de datos provenientes de los CDMS. Los QC aplicados a los últimos conjuntos de datos son el alcance de las actividades del contrato 196 de 2024.

Los QC consisten en “aplicar pruebas de calidad y asignar un banderín o etiqueta a cada dato” (OMM, 2021), según se superen o no las pruebas de los QC o del sistema de los mismos (QCS) propuesto por la contratista autora del presente informe, para el IDEAM. En la ejecución de actividades de su contrato, el 196 de 2024, cuyo objeto es “Prestar los servicios profesionales para la implementación los criterios de validación en estaciones automáticas a través de la elaboración y suministro de los códigos fuentes (scripts) en las series de las variables meteorológicas articularlos con el sistema DHIME”, el segundo producto entregable es el presente informe de diagnóstico actual y de resultados de controles de calidad de las series de datos procedentes de EMA específicamente, de las variables: presión atmosférica, humedad relativa, temperatura y humedad del suelo y evaporación, última variable cuyo diagnóstico tendrá fines exploratorios.

Como diagnóstico inicial (análisis descriptivo), se obtuvo, mediante scripts en lenguaje Python, la información respectiva a: cálculo de estadísticos descriptivos de las series temporales de cada estación (valor mínimo, máximo, media, desviación estándar, varianza); fecha inicial y fecha final de la serie temporal por estación; conteo total de datos por variable; conteo de datos por estación de cada una por año, año-mes, año-mes-día; gráficos de longitud y completitud. Como análisis diagnóstico, se generan estadísticos de los resultados de las pruebas del QCS planteados, desarrollados, también a través de rutinas en el lenguaje de programación Python a las series temporales provenientes de EMA, de mediciones de las variables de interés ya nombradas.

# OBJETIVOS

## OBJETIVO GENERAL

Generar análisis descriptivos y diagnósticos del comportamiento en relación a la calidad de los datos de las series temporales procedentes de EMA de la red de observación de superficie del IDEAM y otras entidades, de las mediciones de las variables de presión atmosférica, humedad relativa, temperatura y humedad del suelo y evaporación como variables meteorológicas de interés

## OBJETIVOS ESPECÍFICOS

* Construir rutinas para la obtención masiva de información exploratoria y descriptiva que permita reconocer el estado inicial de las series temporales de las variables meteorológicas de interés en la ejecución del contrato de prestación de servicios profesionales número 196 de 2024 a IDEAM.
* Desarrollar scripts para análisis diagnóstico (comprobación de hipótesis de existencia de errores, correlaciones o causalidades y relaciones entre variables -regresión-) de los resultados obtenidos al aplicar las pruebas del QCS a las series temporales de las variables meteorológicas de interés, teniendo en cuenta la categoría de etiquetad.
* Concluir el análisis diagnóstico con la exhibición de información útil, sugerencia de conclusiones y herramientas propicias para la toma de decisiones en el IDEAM en relación a la calidad de los datos de EMA de las variables meteorológicas de interés.

# METODOLOGÍA

Para el cumplimiento de los objetivos del presente informe, se planteó la metodología que se muestra en la Figura 1

Figura 1 Metodología para análisis diagnóstico de datos de EMA de variables meteorológicas de interés.

Fuente: Elaboración propia.

# RESULTADOS

## PRESIÓN ATMOSFÉRICA

### Exploración inicial de datos crudos

Se obtuvieron los conjuntos de datos a través del repositorio de Big Data, ‘Cassandra’ de IDEAM, por cada EMA de la variable presión atmosférica, de la red de observación de superficie de IDEAM y otras entidades. Se encontraron en el panorama inicial, el periodo de datos y cantidades que se observan en la Figura 2

Figura 2 Periodos y cantidades de datos de presión atmosférica de EMA analizados en el contrato 196 de 2024.

Fuente: elaboración propia.

### Análisis descriptivo

En relación con el estado inicial de los datos crudos, se obtuvo mediante rutinas de código la información que describe la Tabla 1 junto con el nombre de la carpeta/archivo que le corresponde. Es importante tener en cuenta que esta se ubica en anexos ya que al ser datos estadísticos descriptivos y de exploración obtenidos para 422 estaciones diferentes, se requiere generar un anexo.

Tabla 1 Descripción de la información exploratoria o descriptiva de los datos crudos de presión atmosférica provenientes de EMA.

|  |  |  |
| --- | --- | --- |
| **No. Anexo** | **Información obtenida** | **Carpeta/Archivo** |
| 1 | Gráficos de longitud y completitud de datos por estación, dividido en grupos de 50 estaciones o menos. | An1\_Figs\_LongyCompletDatos\_Patm (Directorio con 9 gráficos). |
| 2 | Fecha inicial y final de transmisión de cada estación | An2\_FechaInicialFinal\_Patm.csv |
| 3 | Estadísticos descriptivos | An3\_EstadDescript\_Patm\_Raw.csv |
| 4 | Cantidad de datos por cada día, cada día-mes, cada día-mes-año, para facilidad de análisis. | An4\_ConteoDatos\_AnioMesDia\_Patm.csv |

Fuente: elaboración propia.

En relación con los estadísticos descriptivos calculados de las series temporales de **datos crudos de presión atmosférica**, se encontró que existen 70 estaciones cuyo dato mínimo es inferior a 0.0 hPa. Por otro lado, según se indagó en la bibliografía, muchos SMHN, a partir del análisis de datos en sus CDMS, observaron como límite rígido/físico/duro superior para la presión atmosférica 1.100 hPa, por lo cual, es de especial atención lo observado para las series temporales de presión atmosférica alojadas en sistemas de información IDEAM pues 87 estaciones tienen como dato máximo, cifras superiores a 1.100 hPa, siendo 63 de ellas las mismas que están fuera del umbral físico inferior de 0.0 hPa. Los códigos de estas estaciones se muestran a continuación:

Figura 3 Estaciones de atención por lo obtenido en estadísticos descriptivos, presión atmosférica.

Se observan en rojo en el bloque de estaciones de la derecha, aquellas que están en común en ambos grupos.

Fuente: Elaboración propia.

Estaciones cuya transmisión corresponde únicamente de 1 a 5 datos:

21185090, 21237020, 21255160, 24027070, 26155270, 32067030, 35017020, 35027510, 35037100, 54017040, 2120500204, 2120700162

### Análisis diagnóstico

**Comprobación de hipótesis**

Inicialmente, con respecto a la **prueba de transmisión**, se encontró que existen estaciones cuya transmisión presentó errores dentro de los registros de la serie temporal, siendo un total de 58 del total de 422, que se enuncian a continuación:

11105020, 11135030, 12015100, 13055040, 13085050, 15085050, 17025020, 17035010, 21015050, 21195170, 21206930, 21206940, 21206950, 21206990, 23035030, 23065505, 23085270, 24015340, 24035390, 24035410, 24035430, 24055070, 25025350, 26015030, 26055110, 26055120, 26085170, 26145090, 28025120, 28035060, 29035000, 32105080, 35035100, 35075070, 35075080, 35080050, 35165000, 35185010, 35195020, 35260050, 36015020, 37015030, 44015060, 46015030, 51025060, 52045080, 52055090, 52055150, 53045040, 56125730, **57015010**, 57025020, 88112901, 1501500054, 2319500043, 3526500201, **5311500147**, **5311500149.**

Se encontró que existen un total de 49 estaciones con toda su serie temporal que no superó las pruebas de controles de calidad, de las 424 estaciones que se han mencionado que existen de la variable presión atmosférica, correspondientes a los códigos:

21085030, 21185090, 21197430, 21205506, 21205509, 21205513, 21205514, 21205515, 21205516, 21206940, 21237020, 21237060, 21255160, 22065040, 23065160, 23067020, 23195040, 24027070, 24030350, 24035360, 24035390, 25025002, 26155270, 26185020, 26200120, 27015300, 28015110, 29045150, 29065140, 32067030, 35017020, 35027510, 35037100, 35105050, 35195020, 35237040, 51025080, 54017040, 54025010, 1501500054, 1602500063, 2120500204, 2120700162, 2202500051, 2401500052, 2603500127, 2612500124, 3507500133, 3701500117

## Humedad relativa

### Exploración inicial de datos crudos

Como se ha aclarado previamente, las series temporales por estación son extraídas del repositorio de Big Data, ‘Cassandra’, de IDEAM, correspondientes a datos de la red de observación de IDEAM y de otras entidades. Las cantidades y fechas generales se enuncian en la Figura 4

Figura 4 Periodos y cantidades de datos de variable humedad relativa de EMA analizados en el contrato 196 de 2024.

Fuente: elaboración propia.

### Análisis descriptivo

Se continuó con el análisis descriptivo ya planteado para la variable humedad relativa. Los estadísticos, información relevante y gráficos obtenidos, se pueden revisar en los anexos que se enuncian en la Tabla 2

Tabla 2 Descripción de la información exploratoria o descriptiva de los datos crudos de humedad relativa provenientes de EMA.

|  |  |  |
| --- | --- | --- |
| **No. Anexo** | **Información obtenida** | **Carpeta/Archivo** |
| 5 | Gráficos de longitud y completitud de datos por estación, dividido en grupos de 50 estaciones o menos. | An5\_Figs\_LongyCompletDatos\_HR (Directorio con 13 gráficos). |
| 6 | Fecha inicial y final de transmisión de cada estación | An6\_FechaInicialFinal\_Patm.csv |
| 7 | Estadísticos descriptivos | An7\_EstadDescript\_Patm\_Raw.csv |
| 8 | Cantidad de datos por cada día, cada día-mes, cada día-mes-año, para facilidad de análisis. | An8\_ConteoDatos\_AnioMesDia\_HR.csv |

Fuente: elaboración propia.

Los estadísticos descriptivos calculados de las series temporales de **datos crudos de la variable humedad relativa,** permitieron evidenciar que hay 139 (de 620) estaciones cuyo dato mínimo es inferior a 0.0 % y 320 (de 620) superiores a 100 %, rango fuera del cual, físicamente no es posible que se comporte esta variable. Tales códigos, se enuncian a continuación:

Figura 5 Estaciones de atención por lo obtenido en estadísticos descriptivos, humedad relativa.

Fuente: Elaboración propia.

Figura 5 Estaciones de atención por lo obtenido en estadísticos descriptivos, humedad relativa.

Fuente: Elaboración propia. Se observan en rojo en el bloque de estaciones de la derecha, aquellas que están en común en ambos grupos.

## Temperatura del suelo -10CM

### Exploración inicial de datos crudos

Los datos extraídos de la base de datos Big Data, ‘Cassandra’, de IDEAM, correspondientes a datos de la red de observación de IDEAM y de otras entidades, tienen las siguientes fechas iniciales y finales, cantidades de datos y estaciones Figura 7

Figura 7 Periodos y cantidades de datos de variable humedad relativa de EMA analizados en el contrato 196 de 2024.

Fuente: elaboración propia.

### Análisis descriptivo

En este análisis, los estadísticos descriptivos, información relevante y gráficos obtenidos, se pueden revisar en los anexos que se enuncian en la Tabla 2

Tabla 3 Descripción de la información exploratoria/descriptiva de los datos crudos de temperatura del suelo -10cm provenientes de EMA.

|  |  |  |
| --- | --- | --- |
| **No. Anexo** | **Información obtenida** | **Carpeta/Archivo** |
| \_\_ | Gráficos de longitud y completitud de datos por estación, dividido en grupos de 50 estaciones o menos. | An16\_Figs\_LongyCompletDatos\_TS10 (Directorio con 2 gráficos). |
| \_\_ | Fecha inicial y final de transmisión de cada estación | FechaInicialFinal\_TS10.csv |
| \_\_ | Estadísticos descriptivos | EstadDescript\_TS10\_Raw.csv |
| \_\_ | Cantidad de datos por cada día, cada día-mes, cada día-mes-año, para facilidad de análisis. | ConteoDatos\_AnioMesDia\_TS10.csv |

Fuente: elaboración propia.

Los estadísticos descriptivos calculados de las series temporales de **datos crudos de la variable temperatura del suelo a -10cm,** mostraron que hay 46 de las 85 estaciones cuyo dato mínimo es inferior -15°C, conocido como umbral inferior posible de esta variable. Asimismo, se encontraron 66 de las 85 estaciones cuyo dato máximo es superior a 45°C, excediendo el límite máximo conocido como posible (más un margen de aumento) en el suelo del territorio nacional. En la\_\_\_ se dan a conocer los códigos respectivos:

Figura 8 Estaciones que exceden límites físicos conocidos de la variable de temperatura del suelo -10°C.

Se observan en rojo en el bloque de estaciones de la derecha, aquellas que están en común en ambos grupos.

Fuente: elaboración propia.

**Transmisión**

***Buena:*** 26157190, 35095120

**Luego de limrig:** 21115010 (corta pero homogénea)

# ANEXOS

Ver carpeta [\_\_\_\_\_](https://shorturl.at/tNOT9), cuyos contenidos se han relacionado en el cuerpo de este documento.

# BibliograFÍA

OMM. (2021). *OMM No. 1269 Directrices para el control de la calidad y el aseguramiento de la calidad de los datos de estaciones de observación en superficie para aplicaciones climáticas.* Genève 2: Organización Meteorológica Mundial. Recuperado el 17 de Abril de 2023

# HISTORIAL DE CAMBIOS

|  |  |  |
| --- | --- | --- |
| **VERSIÓN** | **FECHA** | **DESCRIPCIÓN** |
| -- | \_\_/\_\_/2024 | Elaboración y presentación inicial del documento. |

|  |  |  |
| --- | --- | --- |
| **ELABORÓ:**  Paola Andrea Álvarez Betancourt  Profesional Contratista, CTO 196 de 2024  Grupo Gestión de Datos y Red Meteorológica  Subdirección de Meteorología | **REVISÓ:**  Ruth Leonor Correa Amaya  Coordinadora Grupo Gestión de Datos y Red Meteorológica  Subdirección de Meteorología  Supervisora de Contrato | **APROBÓ:** |