Juan Pablo Naranjo
201730006
Gabriel Mateo Mejía
201718064
Maria Fernanda Martinez
201716576

CASO DE ESTUDIO 7
Procesos Estocásticos - IELE 4010

29-10-2021

Punto 1

a)

Se convirtieron los datos de radiación solar a periodos anuales, realizando una sumatoria acumulada para cada mes y dividiendo este total en 12. Obteniendo así un promedio para cada año. Lo anterior, se muestra en el archivo MATLAB adjunto.

b)

Se graficaron los datos de cada una de las tres variables respecto al tiempo, los resultados se muestran a continuación:

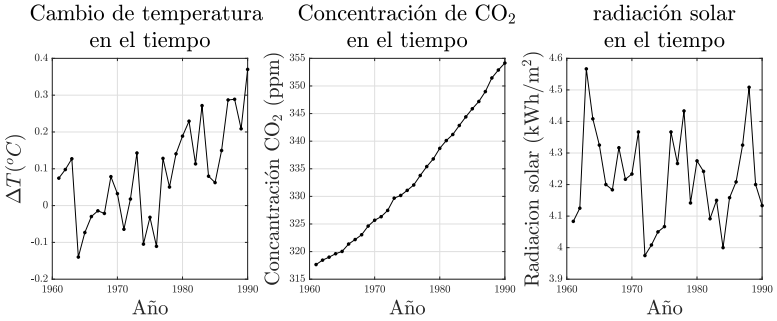


Figura 1: Anomalías de temperatura anual, concentración de CO_2 y radiación solar a través de los años para el espectro de tiempo 1960-1990.

Se puede observar que, aunque hay una componente de ruido importante, las anomalías en la temperatura tienden a aumentar a lo largo de los años. En términos de la concentración de CO_2, se puede observar que ésta aumenta monótonamente lo largo del tiempo con una tendencia exponencial. Finalmente para la radiación solar, no es posible identificar un patrón pues parece tener un comportamiento completamente aleatorio, siendo la más irregular dentro de las 3 variables presentadas.

c)

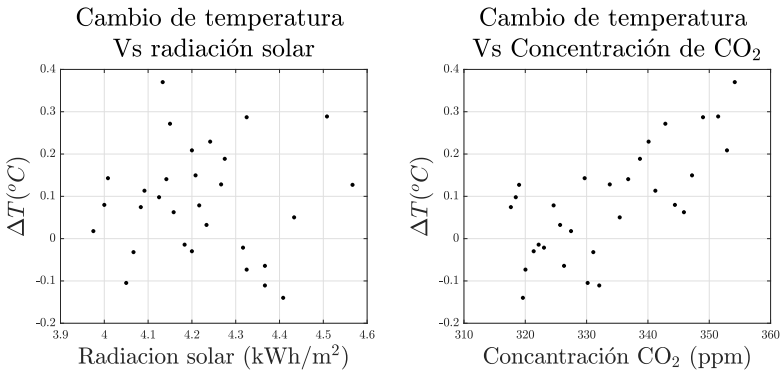


Figura 2: Diagramas de dispersión de las anomalías de temperatura con respecto a la radiación solar y concentración de CO_2 anuales.

Al realizar los diagramas de dispersión entre las anomalías de temperatura y las demás variables del problema (Radiación solar, Concentración de CO_2) como puede observarse en la Fig. 2, son evidentes dos tendencias en los comportamientos conjuntos de las variables: En primera instancia, la radiación solar no parece mostrar ningún patrón aparente de correlación sobre las anomalías de temperatura ya que el diagrama aparece disperso con una distribución aproximadamente circular. Y en segunda instancia, la concentración de CO_2 muestra una correlación positiva con las anomalías de temperatura donde, de forma general, a mayor concentración de CO_2 anual mayor es la variación de temperatura con respecto al modelo. Esto se evidencia en un diagrama aproximadamente elíptico.

d)

Usando primero estimación ML para los vectores de medias se obtiene:

$$\hat{\mu}_{T,Rad} = \begin{bmatrix} 0.085 \\ 4.22 \end{bmatrix} \tag{1}$$

$$\hat{\mu}_{T,CO_2} = \begin{bmatrix} 0.085 \\ 333.41 \end{bmatrix} \tag{2}$$

Luego, usando estimación ML para las matrices de covarianza se obtiene:

$$\hat{Q}_{T,Rad} = \begin{bmatrix} 0.0169 & -0.0003 \\ -0.0003 & 0.0220 \end{bmatrix} \tag{3}$$

$$\hat{Q}_{T,CO_2} = \begin{bmatrix} 0.0169 & 1.0471 \\ 1.0471 & 129.8000 \end{bmatrix} \tag{4}$$

Y usando estos datos para calcular los coeficientes de correlación obtenemos:

$$\hat{\rho}_{T,Rad} = -0.0150 \tag{5}$$

$$\hat{\rho}_{T,CO_2} = 0.7072 \tag{6}$$

Lo cual confirma las observaciones anteriores de una correlación nula entre las anomalías de temperatura y la radiación solar. Junto con una correlación no solo positiva sino muy significativa entre la concentración anual de CO_2 y las anomalías de temperatura.

e)

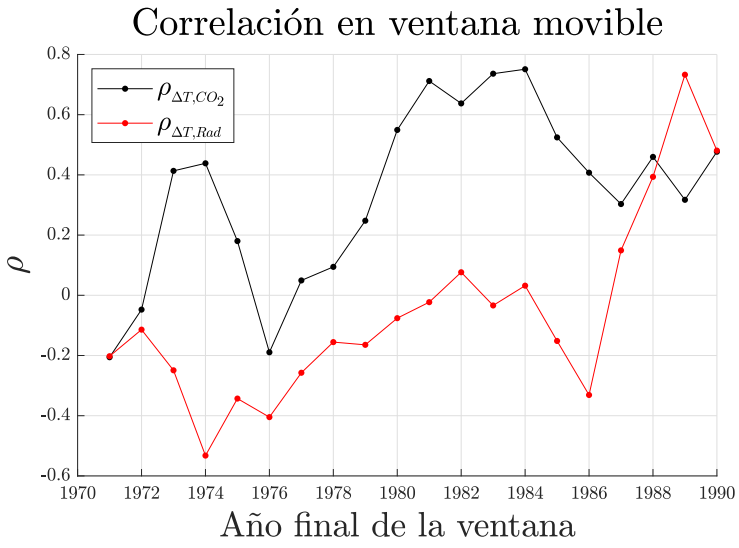


Figura 3: Índice de correlación entre anomalías de temperatura, radiación solar y concentración de CO_2 usando ventanas movibles de 10 años.

Como es evidente en la Fig. 3, la correlación estimada entre las anomalías de temperatura y concentración de CO_2 es positiva para 17 de las 20 ventanas de tiempo estudiadas lo cual permite evidenciar una clara tendencia temporal a pesar del ruido en los datos. Por otra parte el índice de correlación estimado entre las anomalías de temperatura y la radiación solar es en general de magnitud menor y muestra patrones mucho menos definidos con una tendencia débil al aumento en los últimos años.

f)

La pregunta planteada inicialmente es bastante difícil de responder de forma global. Esto debido a que debería definirse de manera formal lo que se entiende por "ciclo natural del sistema". No obstante, asumiendo que esto hace alusión a que que las variaciones de temperatura son explicadas por causas naturales como el aumento de la radiación solar, podemos sacar algunas conclusiones. En

este caso, es claro que no existe correlación alguna entre la radiación solar y las anomalías de temperatura, ni de manera global ni al analizar ventanas movibles de tiempo. Esto nos permite descartar la causalidad de este factor en las anomalías de manera definitiva y afirmar que por lo menos esta variable natural **no** es una causa de las anomalías de temperatura. Por el lado de la concentración de CO_2, es evidente que existe una correlación consistentemente positiva y significativa tanto de manera global como en análisis por ventanas temporales. No obstante, correlación no es sinónimo de causalidad y por ende esta afirmación no puede ser formulada de manera rigurosa ya que ambas observaciones pueden obedecer a una causa global. No obstante, lo que si nos es posible decir es que el aumento de concentración de CO_2 es un candidato plausible para ser la causa de el aumento gradual de anomalías en temperatura. Datos de estudios que propongan modelos mecanísticos acerca de la relación causal entre aumento de concentración de CO_2 y aumento de temperatura serian de vital importancia en este escenario.

Punto 2

Ahora se analizan las mediciones de temperatura T [°C] y humedad relativa H [%] en un invernadero real de un cultivo de rosas, que se modelan como variables aleatorias continuas a través del vector aleatorio $Z = [T, H]^T$. Nos indican que asumir que Z es un vector aleatorio Gaussiano no es suficiente para modelar el comportamiento de la temperatura y la humedad conjuntamente. Se pide estimar una distribución que se ajuste mejor al verdadero comportamiento del invernadero.

a)

En la figura 4 se pueden apreciar las observaciones de temperatura vs. humedad entregadas en el archivo *greenhouseCaso7.txt*.

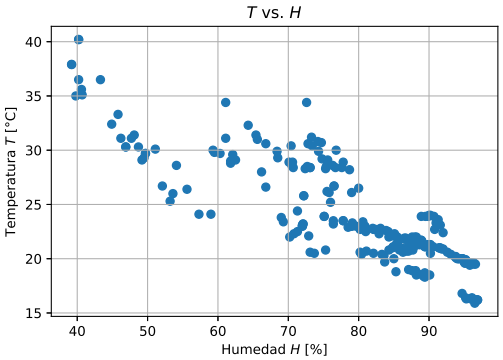


Figura 4: Observaciones de temperatura T vs. humedad relativa H

Se puede notar una gran concentración de datos entre el 80% y 90% de humedad (para temperaturas entre los 15 °C y 25 °C), otra concentración entre el 60% y 80% (temperaturas entre 25 °C y 35 °C) y una última concentración

baja de datos entre el 40% y 60% de humedad (para temperaturas entre los 24 °C y los 40 °C).

b)

Se decidió asumir que el vector Z sí es un vector aleatorio Gaussiano. Teniendo esto en cuenta, se procedió a usar la estimación MLE para determinar los parámetros de este vector aleatorio, que corresponden a su vector de medias μ_Z y matriz de covarianza Q_Z. Usando la estimación MLE, se sabe que estos dos parámetros van a estar dados por

$$\hat{\mu}_Z = \frac{1}{M}\sum_{i=1}^{M} z_i \tag{7}$$

$$\hat{Q}_Z = \frac{1}{M-1}\sum_{i=1}^{M}(z_i - \mu)(z_i - \mu)^T \tag{8}$$

en donde M es el número total de observaciones que se tiene de cada variable aleatoria y z_i es una observación cualquiera de la forma $z_i = [t_i, h_i]^T$. Esto quiere decir que la PDF de este vector aleatorio Gaussiano está dada por

$$f_Z(z) = N(z; \hat{\mu}_Z, \hat{Q}_Z) = \frac{1}{2\pi\sqrt{\det \hat{Q}_Z}} \exp\left(-\frac{1}{2}(z-\hat{\mu}_Z)^T \hat{Q}_Z{}^{-1}(z-\hat{\mu}_Z)\right) \tag{9}$$

donde, realizando los cálculos respectivos:

$$\begin{aligned} \hat{\mu}_Z &= \begin{bmatrix} 23.82 \\ 79.21 \end{bmatrix} \\ \hat{Q}_Z &= \begin{bmatrix} 24.165 & -63.29 \\ -63.29 & 220.64 \end{bmatrix} \end{aligned} \tag{10}$$

En la figura 5 se pueden observar las observaciones de temperatura y humedad nuevamente, pero esta vez con las curvas de nivel de esta PDF traslapadas en la misma gráfica.

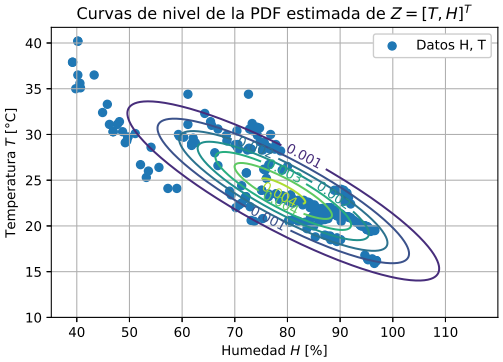


Figura 5: Curvas de nivel de la PDF de Z

Se puede notar que asumir que Z es un vector aleatorio Gaussiano conjunto no es del todo suficiente para dar una buena aproximación de la relación entre temperatura y humedad, pues aunque el centro de la campana de Gauss se ubica en un sitio de alta concentración de datos, hay tres diferentes áreas de alta concentración de datos que solo una función Gaussiana no alcanza a cubrir y les asignaría una baja probabilidad de ocurrencia a datos que en realidad tienen una mayor probabilidad de ocurrencia.

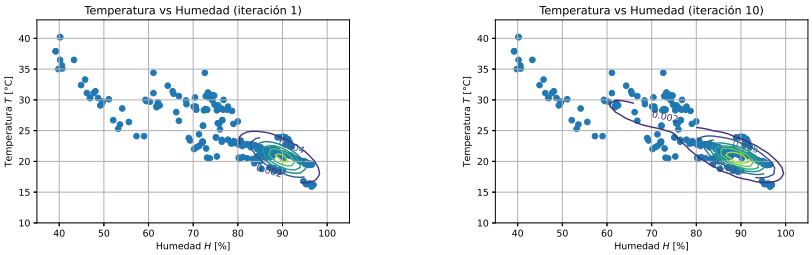
c)

Ahora se implementó el algoritmo EM para tratar de estimar una función de densidad de probabilidad que modele mejor la relación entre temperatura y humedad en el invernadero. Se probó para una combinación de dos PDFs Gaussianas, para así generar una función de la forma

$$f_Z(z) = \sum_{k=1}^{2} \alpha_k N(z; \mu_k, Q_k). \tag{11}$$

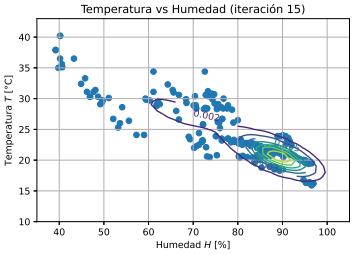
donde α_k es un número real entre 0 y 1 que representa el peso que tiene la k-ésima función Gaussiana sobre la función completa.

A continuación se muestra el comportamiento de las curvas de nivel de esta función para diferentes iteraciones del algoritmo con vectores de medias iniciales en un rango cercano a los datos, matrices de covarianza inicializadas como matrices identidad y α_k inicializados como $1/2$ cada uno.

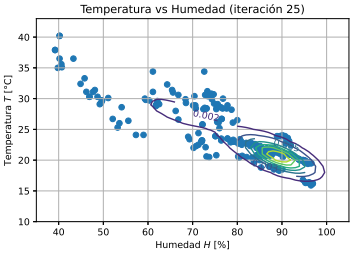


(a) Iteración 1

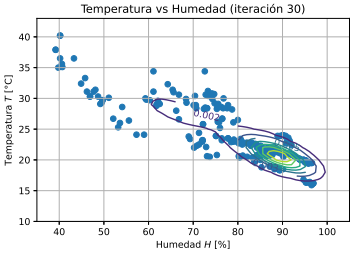
(b) Iteración 10



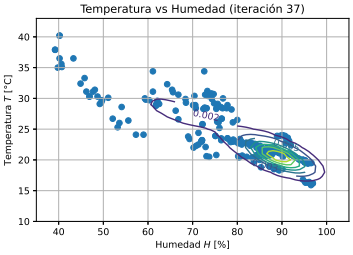
(c) Iteración 15



(d) Iteración 25



(e) Iteración 30



(f) Iteración 37

Figura 6: Algoritmo EM - Combinación de dos Gaussianas para diferentes iteraciones

El algoritmo EM en este caso convergió en la iteración 37 (para una tolerancia de 10^{-5}), como se puede observar en la figura 6f. Se puede notar cómo a medida que avanzan las iteraciones del algoritmo, se ve más claramente cómo se va formando una segunda campana de Gauss alrededor de otra zona del plano con alta concentración de datos, lo que nos indica un funcionamiento adecuado del algoritmo. Cuando el algoritmo convergió, se obtuvieron los siguientes parámetros, en donde el subíndice 1 hace referencia a la primera de las dos Gaussianas que se combinaron y el subíndice 2 hace referencia a la segunda:

- $\alpha_1 = 0.458$
- $\alpha_2 = 0.542$
- $\mu_1 = \begin{bmatrix} 27.66 \\ 67.26 \end{bmatrix}$
- $\mu_2 = \begin{bmatrix} 20.58 \\ 89.29 \end{bmatrix}$

- $Q_1 = \begin{bmatrix} 21.046 & -45.92 \\ -45.92 & 187.80 \end{bmatrix}$
- $Q_2 = \begin{bmatrix} 3.66 & -6.09 \\ -6.09 & 24.75 \end{bmatrix}$

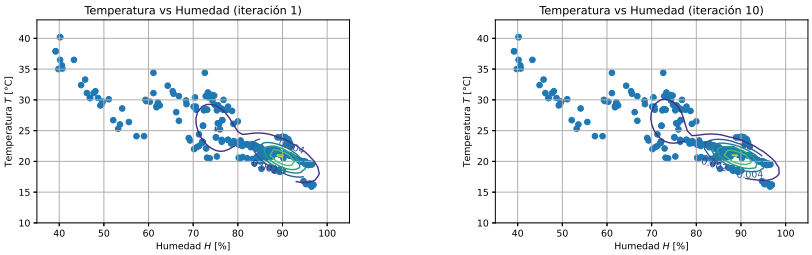
Es claro que esta función, que consiste de una combinación de dos Gaussianas cuyos parámetros se encontraron con ayuda del algoritmo EM y están listados arriba, modela de una mejor manera el comportamiento entre la temperatura y la humedad en comparación con una sola función Gaussiana conjunta, pues esta función le asigna valores más altos de probabilidad de ocurrencia a otro grupo de datos que se encuentran en una zona de alta concentración de datos en el plano. La función Gaussiana única le asignaba valores más bajos de probabilidad a estos datos y por lo tanto no era tan fiable como la combinación de dos Gaussianas. Sin embargo, esta estimación de dos Gaussianas sigue dejando por fuera otro grupo de datos que se encuentran en una tercera zona de alta concentración de datos, alrededor de los 25 °C y 40 °C y el 40% y 60% de humedad relativa.

d)

Se llevó a cabo el mismo procedimiento del literal anterior, pero ahora se combinaron 3 funciones Gaussianas, resultando en

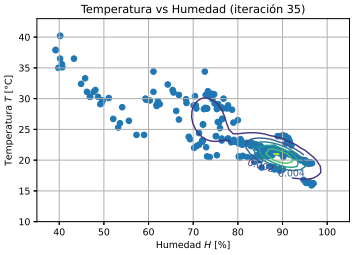
$$f_Z(z) = \sum_{k=1}^{3} \alpha_k N(z; \mu_k, Q_k). \tag{12}$$

Esta vez la inicialización de los vectores de medias también fue en valores en un rango cercano a los datos, matrices de covarianza como matrices identidad y α_k como $1/3$ cada uno. A continuación se muestra el comportamiento de las curvas de nivel de esta función para diferentes iteraciones del algoritmo EM:

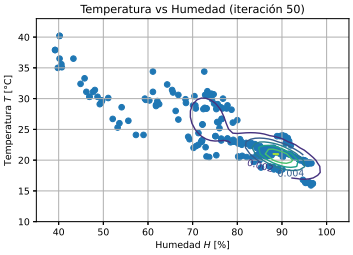


(a) Iteración 1

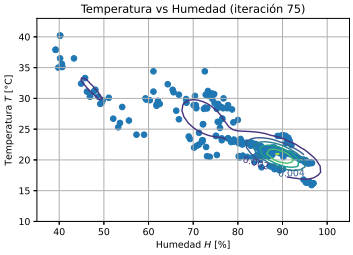
(b) Iteración 10



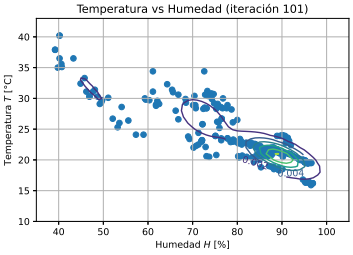
(c) Iteración 35



(d) Iteración 50



(e) Iteración 75



(f) Iteración 101

Figura 7: Algoritmo EM - Combinación de tres Gaussianas para diferentes iteraciones

En este caso, el algoritmo convergió en la iteración 101 (para una tolerancia de 10^{-5}), como se puede observar en la figura 7f. Se puede notar cómo a medida que avanzan las iteraciones del algoritmo, se ve más claramente cómo se van formando la segunda y tercera campana de Gauss alrededor de las diferentes zonas del plano con alta concentración de datos, lo que nuevamente nos indica un funcionamiento adecuado del algoritmo. Cuando el algoritmo convergió, se obtuvieron los siguientes parámetros, en donde el subíndice 1 hace referencia a la primera de las tres Gaussianas que se combinaron, el subíndice 2 hace referencia a la segunda y el subíndice 3 hace referencia a la tercera:

- $\alpha_1 = 0.313$
- $\alpha_2 = 0.572$
- $\alpha_3 = 0.115$
- $\mu_1 = \begin{bmatrix} 26.88 \\ 72.24 \end{bmatrix}$

- $\mu_2 = \begin{bmatrix} 20.59 \\ 89.45 \end{bmatrix}$
- $\mu_3 = \begin{bmatrix} 31.56 \\ 47.27 \end{bmatrix}$
- $Q_1 = \begin{bmatrix} 12.126 & -9.41 \\ -9.41 & 32.53 \end{bmatrix}$
- $Q_2 = \begin{bmatrix} 3.74 & -5.61 \\ -5.61 & 22.83 \end{bmatrix}$
- $Q_3 = \begin{bmatrix} 20.17 & -24.09 \\ -24.09 & 31.96 \end{bmatrix}$

Es claro que esta función, que consiste de una combinación de tres Gaussianas cuyos parámetros se encontraron con ayuda del algoritmo EM y están listados arriba, modela de una mejor manera el comportamiento entre la temperatura y la humedad en comparación con la única función Gaussiana conjunta, y la combinación de dos Gaussianas del literal anterior, pues esta función le asigna valores más altos de probabilidad de ocurrencia a otro grupo de datos que se encuentran en una zona de alta concentración de datos en el plano, alrededor de los 25 °C y 40 °C y el 40% y 60% de humedad relativa.

Se puede afirmar que la combinación de tres Gaussianas genera el mejor ajuste a los datos en comparación con la combinación de dos Gaussianas y solo una (asumir que Z es un vector aleatorio Gaussiano conjunto). Esto se puede notar claramente de la figura 7f, en la que hay una campana de Gauss para cada una de las tres zonas del plano en la que hubo mayor concentración de datos.

e)

Ahora se tienen 50 nuevos datos de temperatura en el archivo *greenhouse50temp.txt* para los cuales se desconoce el valor asociado de humedad relativa.

i)

Con el modelo obtenido en el literal b), es decir, para el modelo de Z como un vector aleatorio Gaussiano, se obtuvo un estimador MMSE sin restricciones para estimar los valores correspondientes de humedad para los 50 nuevos datos de temperatura del archivo *greenhouse50temp.txt*. Como se tiene solo una función Gaussiana conjunta, el estimador MMSE sin restricciones es el mismo estimador lineal.

La PDF conjunta que resulta de estos datos estimados está dada por

$$f_Z(z) = \frac{1}{2\pi\sqrt{\det\{Q_Z\}}} \exp\left(-\frac{1}{2}(z-\mu_Z)^T Q_Z^{-1}(z-\mu_Z)\right) \tag{13}$$

en donde

$$
\begin{aligned}
z &= \begin{bmatrix} t \\ h \end{bmatrix} \\
\mu_Z &= \begin{bmatrix} \mu_T \\ \mu_H \end{bmatrix} = \begin{bmatrix} 23.824 \\ 79.205 \end{bmatrix} \\
Q_Z &= \begin{bmatrix} \sigma_T^2 & \text{cov}(T,H) \\ \text{cov}(T,H) & \sigma_H^2 \end{bmatrix} = \begin{bmatrix} 24.165 & -63.290 \\ -63.290 & 220.637 \end{bmatrix}
\end{aligned} \tag{14}
$$

Se sabe que el estimador MMSE sin restricciones de H basado en observaciones de T está dado por

$$
\begin{aligned}
\hat{H}_{\text{MMSE}}(T) &= \mu_H + \frac{\text{cov}(H,T)}{\sigma_T^2}[T - \mu_T] \\
\hat{H}_{\text{MMSE}}(T) &= -2.619T + 141.602
\end{aligned} \tag{15}
$$

Usando este estimador, se procedió a estimar los valores de humedad relativa correspondientes a los 50 datos de temperatura nuevos del archivo de texto. El resultado de esta estimación para cada uno de los valores conocidos de temperatura se muestra en la figura 8.

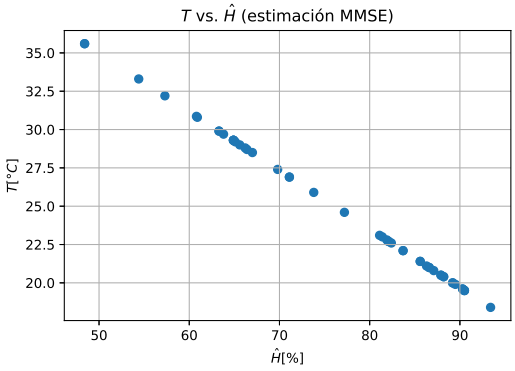


Figura 8: Estimación de valores de humedad con base en valores de temperatura usando el estimador MMSE

ii)

Sea $f_Z(t,h) = \sum_{k=1}^{3} \alpha_k N(t,h;\mu_k,Q_k)$ la distribución encontrada en el literal d). Se procedió a escribir una función en Python que calculara

$$
\hat{H}(t) = E[H|T=t] = \int_{-\infty}^{\infty} h f_H(h|T=t)dh \tag{16}
$$

Para hacer este cálculo, se tuvieron en cuenta las siguientes definiciones:

$$
\begin{aligned}
f_H(h|T=t) &= \frac{f_Z(t,h)}{f_T(t)} \\
f_T(t) &= \int_{-\infty}^{\infty} f_Z(t,h)dh
\end{aligned} \tag{17}
$$

Usando esta función se volvió a encontrar la estimación de la humedad relativa con base en las observaciones de temperatura entregadas en el archivo *datos-greenhouse50temp.txt*. Los resultados de esta estimación se encuentran en la figura 9.

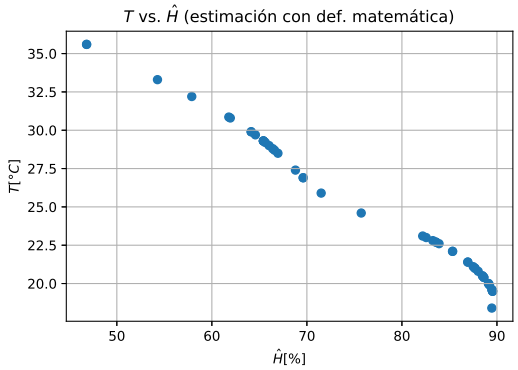


Figura 9: Estimación de valores de humedad con base en valores de temperatura usando el estimador dado por la definición matemática

Como se puede notar, esta estimación a partir de la definición matemática del estimador de la humedad es muy similar a la dada por la estimación MMSE, pues la forma de las estimaciones tiene una forma similar a la de la recta dada por el estimador MMSE.

iii)

Eventualmente se pudieron recuperar las observaciones reales de humedad asociadas a las 50 observaciones de temperatura entregadas previamente. Éstas se encuentran en el archivo *greenhouse50hum.txt*. Usando estos datos reales, se procedió a comparar la precisión de los dos estimadores encontrados previamente en los literales i) y ii) con los datos reales. Esta comparación se hizo a través de la evaluación de la función:

$$error = \frac{1}{50}\sum_{i=1}^{50}\left(h_i - \hat{H}(t_i)\right)^2 \tag{18}$$

donde h_i y t_i se refieren a los valores de humedad y temperatura reales, y $\hat{H}(t_i)$ se refiere a la estimación de h_i dada por cada estimador por separado.

Evaluando la función de error para el estimador MMSE, se obtuvo un valor de función de error dado por

$$error_{\text{MMSE}} = 64.77 \tag{19}$$

mientras que la evaluación de la función de error para el estimador dado por la definición matemática retornó un valor de

$$error_{\text{def.math}} = 64.74 \tag{20}$$

Es claro que el valor de la función de error para ambos estimadores es prácticamente igual, diferenciándose apenas por unas centésimas. Esto nos indica que

ambos estimadores tienen una precisión similar. Sin embargo, el estimador dado por la definición matemática tuvo un valor levemente menor de la función de error comparado con el estimador MMSE. Por otro lado, el tiempo de cómputo del estimador MMSE fue de apenas unas milésimas de segundo. Mientras tanto, el estimador dado por la definición matemática se demoró computando alrededor de 17 minutos, lo que le da una ventaja muy grande al estimador MMSE, dado que las precisiones de ambos estimadores fueron prácticamente las mismas. Por lo tanto, sería más adecuado utilizar el estimador MMSE por su buena precisión y muy bajo tiempo de cómputo, pues hace estimaciones tan buenas como las del estimador dado por la definición matemática en un tiempo mucho menor.

Para visualizar de mejor forma la precisión de ambos estimadores, en la figura 10 se presenta una gráfica con los datos reales (los entregados en los archivos de texto) tanto de temperatura como de humedad, con los datos de humedad estimados por ambos estimadores a partir de los datos de temperatura traslapados.

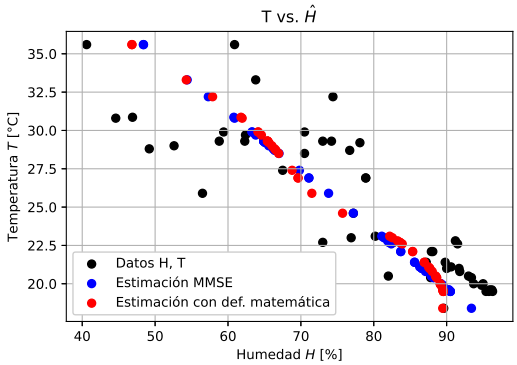


Figura 10: Comparación de estimadores

Se puede notar claramente como la estimación de la humedad dada por ambos estimadores es muy similar, con solo algunas diferencias en la forma de la gráfica que se genera. Además, se puede concluir que ambos estimadores son relativamente buenos con respecto a la distribución de datos reales, pues todas las estimaciones de humedad hechas por cada estimador están dentro del rango de datos reales.