

Trabajo Práctico Especial

Teoría de la Información 2025

Integrantes

-Etcheverría, Victoriano

vetcheverria@alumnos.exa.unicen.edu.ar

-Rapela, Guadalupe

grapela@alumnos.exa.unicen.edu.ar

-Russo, Dolores Magalí

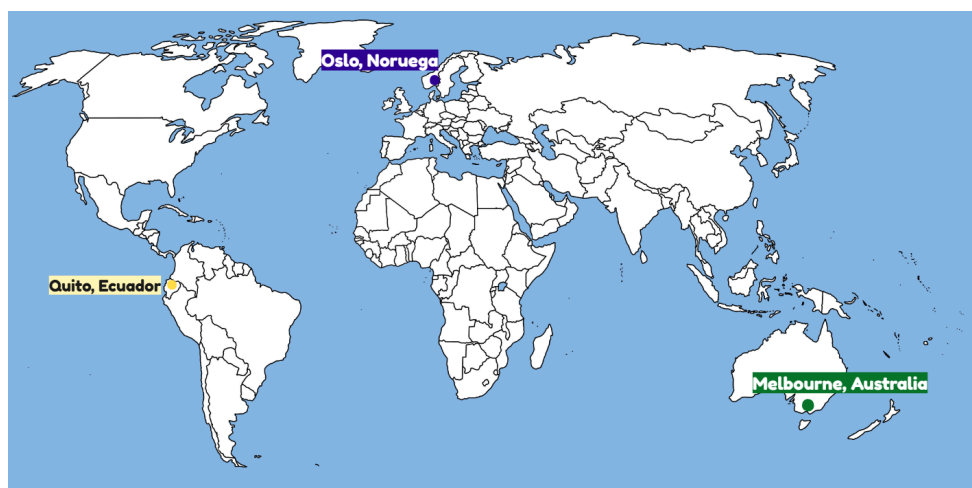
drusso@alumnos.exa.unicen.edu.ar

Introducción.....	3
Desarrollo y análisis.....	3
Preparación de los datos.....	3
Parte 1: Estadísticas.....	4
Cálculo de la media aritmética.....	4
Cálculo de la desviación estándar.....	4
Interpretación de resultados.....	4
Factor de Correlación Cruzada.....	5
Parte 2: Modelado de las fuente Markovianas.....	6
Matriz de transición.....	6
Probabilidad estacionaria [$p(F)^*$, $p(T)^*$, $p(C)^*$].....	7
Tiempo medio de 1ra recurrencia.....	8
Parte 3: Entropía, Huffman y Compresión.....	9
Análisis de códigos de Huffman.....	9
Quito.....	9
Oslo.....	9
Melbourne.....	10
Parte 4: Canal Melbourne ruidoso.....	10
Conclusión.....	11

Introducción

El presente trabajo aplica los contenidos aprendidos en la cátedra de Teoría de la Información 2025. La consigna se basa en aplicar técnicas de análisis, codificación y muestreo computacional sobre tres datasets que contienen las temperaturas promedio diarias de las ciudades de Quito, Melbourne y Oslo durante un mismo lapso de tiempo. Cada dataset muestra una secuencia de números enteros que representa la temperatura promedio diaria en °C a lo largo de aproximadamente 16 años (5.845 días). Las ciudades de las que provienen las muestras se encuentran muy distantes entre sí, con Quito ubicado casi sobre la línea del ecuador, Oslo en el extremo norte de Europa y Melbourne situada en el hemisferio sur, al sudeste de Australia. Esta información se considera relevante para la interpretación de los resultados.

Los objetivos del trabajo apuntan a que podamos implementar en máquina las herramientas adquiridas en la materia con el fin de procesar, comprender y concluir de forma asertiva sobre datos reales en función del contexto dado.



Desarrollo y análisis

Preparación de los datos

Realizamos una preparación de los datasets con el objetivo de mejorar la calidad de los datos antes de aplicar cualquier análisis. Para ello, identificamos valores atípicos (outliers) utilizando el método del rango intercuartílico (IQR), que permite detectar datos que se alejan significativamente del comportamiento típico de la muestra. Los valores que se encontraban por fuera de los límites establecidos ($Q1 - 1.5 * IQR$ y $Q3 + 1.5 * IQR$) fueron considerados posibles outliers. Analizamos utilizando este criterio el dataset de las distintas ciudades.

En el dataset correspondiente a Oslo se observaron valores de temperatura como -21°C , -22°C y -23°C , que resultan coherentes con el clima de la región. Sin embargo, también se detectaron registros extremos, como -73°C , que son físicamente imposibles, lo que

sugiere la presencia de datos atípicos o errores en la recolección o carga de los mismos. Esto último también sucede en el dataset de Melbourne.

En el caso de Quito, la baja variabilidad de los datos genera un IQR reducido, haciendo que se destaquen como atípicos valores que en realidad no lo son. Por esta razón, dichos registros no serán considerados outliers.

Para los valores considerados imposibles, se procedió a realizar una imputación por media en cada ciudad. Esta estrategia permite conservar la estructura general de los datos sin introducir sesgos significativos, asegurando así que los análisis posteriores no se vean mayormente afectados.

Parte 1: Estadísticas

Lo primero que se realizó fue el cálculo de las medidas de tendencia central (media y desviación estándar) de cada señal para conocer su comportamiento general. Luego, se estudió la correlación entre cada par de ciudades para ver si existe alguna relación entre sus patrones de temperatura, o si varían de forma totalmente independiente.

Cálculo de la media aritmética

Definición: Valor promedio de temperatura que se esperaría observar en cada ciudad si se repitiera el experimento aleatorio una gran cantidad de días.

Cálculo de la desviación estándar

Definición: Medida de dispersión que indica cuán dispersos están los valores de temperatura de cada ciudad respecto a la media.

	Media	Desvío Estándar
Oslo	4.8217	8.7057
Quito	13.6035	1.3016
Melbourne	17.7986	4.2474

Interpretación de resultados

En Oslo, la temperatura media se mantiene baja, cercana a los 5 °C, lo cual es coherente con su clima frío. La desviación estándar es alta, lo que indica una gran variabilidad de temperatura. Estas variaciones son típicas de regiones con estaciones bien definidas, con inviernos severos y veranos templados o frescos.

En Quito, la temperatura media es moderadamente cálida y se mantiene muy estable a lo largo del tiempo. La desviación estándar es muy baja, lo que refleja una temperatura casi constante. Este comportamiento es característico de ciudades cercanas al ecuador, donde no existen estaciones climáticas marcadas.

Melbourne presenta una temperatura media templada a cálida. Su variabilidad térmica es intermedia: mayor que en Quito, pero menor que en Oslo.

En resumen, Quito destaca por su clima sumamente estable y predecible. En cambio, Oslo presenta la mayor variación térmica anual, reflejo de un clima con estaciones muy marcadas. Por último, Melbourne muestra un clima más equilibrado, con estaciones presentes pero sin variaciones tan marcadas.

Factor de Correlación Cruzada

Definición: Medida estadística utilizada para determinar la relación entre dos fuentes. La magnitud y el signo del resultado de la correlación cruzada indican la fuerza y la dirección de la relación entre ellas. Un valor alto (positivo o negativo) sugiere una fuerte relación, mientras que un valor cercano a cero indica poca o ninguna relación.

1. Oslo - Quito:

Correlación Baja con signo (-)

El gráfico de dispersión muestra una nube de puntos sin tendencia clara, distribuida de manera horizontal.

2. Oslo - Melbourne:

Correlación Moderada-Alta con signo (-)

El gráfico de dispersión se asemeja a una línea diagonal hacia abajo.

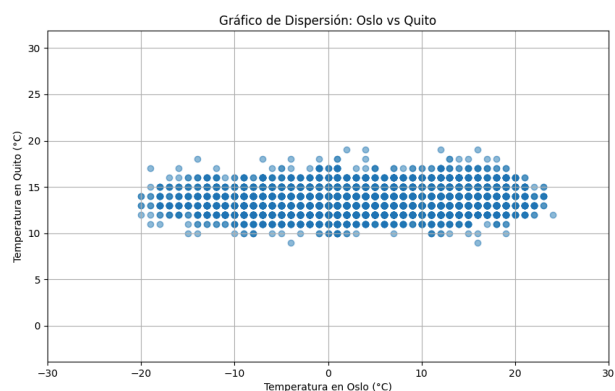
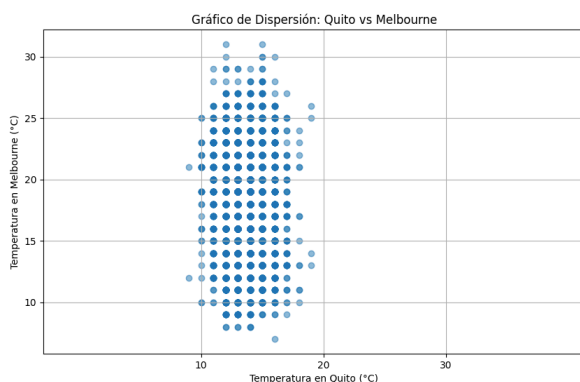
3. Quito - Melbourne:

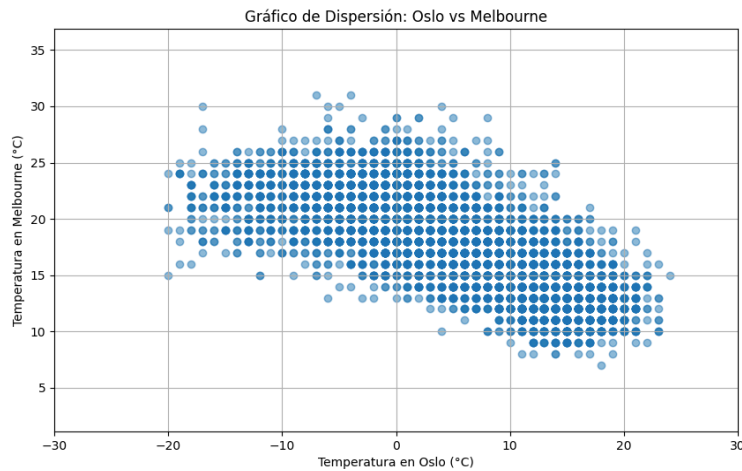
Correlación Baja con signo (+)

El gráfico de dispersión se asemeja a una línea vertical.

1,3: No hay una correlación significativa entre los datos.

2: Existe una correlación fuerte entre los datos. La tendencia es que los días en que la temperatura en Oslo es más elevada, la temperatura en Melbourne suele ser más baja y viceversa. Esta relación inversa resulta coherente considerando que ambas ciudades se encuentran en hemisferios opuestos, por lo que las estaciones del año están invertidas entre sí. El gráfico de dispersión presentado refuerza esta observación, donde se visualiza la inclinación descendente que refleja esta tendencia inversa entre ambas variables.





Parte 2: Modelado de las fuente Markovianas

A partir de las secuencias de símbolos frío (si $t > 11^{\circ}\text{C}$), templado (si $11 \leq t < 19^{\circ}\text{C}$) y cálido (si $t \geq 19^{\circ}\text{C}$), se construyó la matriz de transición para cada ciudad, lo que permite analizar cómo evoluciona el clima día a día. Luego, mediante simulaciones de Monte Carlo, se estimaron la distribución estacionaria de los estados y el tiempo medio de primera recurrencia para cada símbolo. Además, se evaluó el efecto de distintos umbrales de convergencia ϵ sobre la precisión y estabilidad de los resultados, complementando el análisis con gráficos de convergencia.

Matriz de transición

✓	Fx	Tx	Cx
Fy	0.959	0.099	0.000
Ty	0.041	0.865	0.296
Cy	0.000	0.035	0.704

En Oslo, el clima muestra una marcada estabilidad dentro de cada estado. Si hoy hace frío, hay un 95,9 % de probabilidad de que mañana también lo sea, lo que indica una alta persistencia del estado frío. Si el día es templado, existe un 86,5 % de que se mantenga así, aunque también hay un 9,9 % de posibilidad de volver al frío y un 3,5 % de avanzar hacia un estado cálido. Cuando el día es cálido, la

probabilidad de mantenerse en ese estado es del 70,4 %, mientras que el 29,6 % restante representa una transición hacia templado. En resumen, el clima en Oslo tiende a ser bastante estable dentro de los extremos frío y cálido, mientras que el estado templado es más inestable. Además, es improbable un cambio directo de frío a cálido, característico de sus estaciones.

✓	Fx	Tx	Cx
Fy	0.054	0.006	0.000
Ty	0.946	0.993	1.000
Cy	0.000	0.001	0.000

En Quito, la dominancia del estado templado es evidente. Si el día comienza frío, hay un 94,6 % de probabilidad de que al día siguiente esté templado. Si hoy está templado, es casi seguro que seguirá igual (99,3 %). En términos prácticos, el clima de Quito puede considerarse puramente templado en

casi todos los días del año, lo que confirma su notable estabilidad térmica.

✓	Fx	Tx	Cx
Fy	0.486	0.032	0.000
Ty	0.514	0.836	0.147
Cy	0.000	0.132	0.853

Melbourne, en cambio, presenta un comportamiento más variable. Si hoy hace frío, hay un 47 % de probabilidad de que siga así y un 49,5 % de que pase a templado, lo que indica una alta probabilidad de cambio. Si el día está templado, existe un 83,6 % de que se mantenga así, pero también un 13,1 % de que avance hacia cálido. Cuando el clima es cálido, hay un 85,3 % de posibilidad de permanecer en ese estado y un 13,1 % de probabilidad de bajar a templado. Esto refleja que el estado cálido es relativamente estable, pero también que las temperaturas pueden cambiar de un día a otro con mayor facilidad que en Oslo o Quito.

Probabilidad estacionaria [$p(F)^*$, $p(T)^*$, $p(C)^*$]

Vector estacionario de Oslo: [0.683 0.283 0.035]

Vector estacionario de Quito: [0.006 0.993 0.001]

Vector estacionario de Melbourne: [0.032 0.506 0.462]

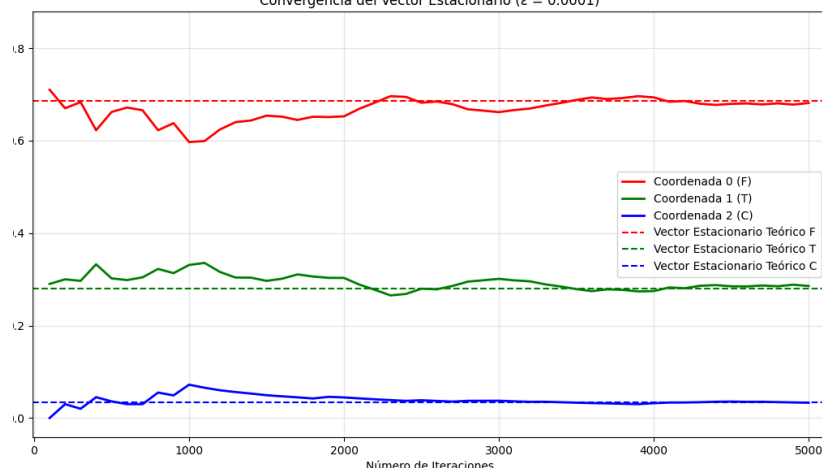
Se aplicó un método de Monte Carlo para calcular el vector estacionario de cada ciudad a partir de su matriz de transición. Para eso, se generaron muchos estados simulados hasta que las frecuencias de aparición se estabilizaron. El resultado muestra la probabilidad a largo plazo de que la temperatura esté en estado frío, templado o cálido (en ese orden).

-Oslo tiene un clima mayormente frío, con una menor proporción de días templados y casi ningún día cálido.

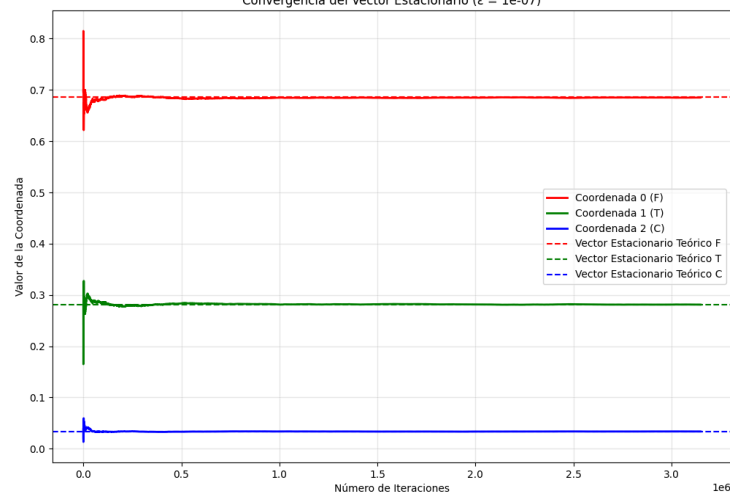
-Quito mantiene una temperatura templada durante casi todo el año, lo cual es consistente con su ubicación ecuatorial. Tiene un clima muy estable, sin grandes fluctuaciones térmicas.

-Melbourne muestra un clima variable, con una distribución relativamente equilibrada entre estados templados y cálidos.

Convergencia del Vector Estacionario ($\epsilon = 0.0001$)



Convergencia del Vector Estacionario ($\epsilon = 1e-07$)



Los gráficos muestran que las coordenadas convergen a los vectores estacionarios teóricos. El umbral ϵ afecta la velocidad: $\epsilon = 0,001$ converge en aprox. 5000 iteraciones, mientras que $\epsilon = 1e-07$ requiere aprox. 300 mil, sin cambiar significativamente los valores finales. Un ϵ más alto ahorra tiempo, pero reduce precisión; un ϵ más bajo mejora precisión a costa de más iteraciones.

Tiempo medio de 1ra recurrencia

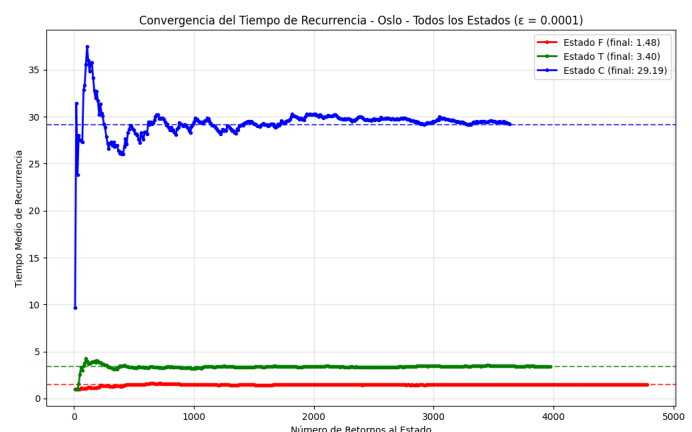
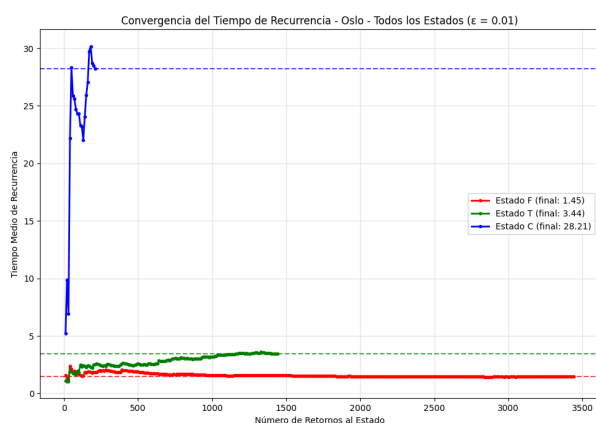
	F	T	C
Oslo	1.464	3.588	30.992
Quito	156.987	1.008	1463.018
Melbourne	30.997	1.999	2.179

Nota: cada “paso” representa un día ya que simulamos temperaturas diarias.

Oslo muestra un comportamiento relativamente equilibrado entre los estados frío y templado. El estado frío tiene una recurrencia muy rápida (aprox 1,46 días), lo que indica que aparece con mucha frecuencia. El estado templado tarda algo más en reaparecer (aprox 3,59 días), pero sigue siendo común. En cambio, el estado cálido es claramente el menos frecuente, con un tiempo medio de recurrencia de (aprox. 31 días), lo que sugiere que los días cálidos son pocos, probablemente estén concentrados en verano.

Quito presenta un patrón completamente distinto. El estado templado tiene un tiempo de recurrencia de aprox. 1 día, lo que confirma que es el estado dominante. Por el contrario, tanto frío como cálido tienen tiempos de recurrencia muy altos (aprox. 157 y 1463 días respectivamente), lo que indica que son extremadamente raros y se comportan como excepciones puntuales.

Melbourne, en cambio, tiene tiempos de recurrencia bastante equilibrados para templado (aprox. 2 días) y cálido (aprox. 2,18 días), lo que refleja que ambos estados se alternan regularmente. El estado frío tiene una recurrencia más espaciada (aprox. 31 días), por lo que aparece menos seguido, pero tampoco es tan raro como en Quito.



Tanto el tiempo medio de recurrencia como la convergencia del vector estacionario muestran que el umbral ϵ influye en la velocidad de convergencia y el número de iteraciones requeridas, sin cambiar notablemente los valores finales. Un ϵ más alto (0,01) acelera la estabilización, mientras que un ϵ más bajo (0,0001) ofrece mayor precisión a costa de más tiempo, manteniendo una dinámica comparable en ambos análisis.

Parte 3: Entropía, Huffman y Compresión

	H1	Hcond
Oslo	1,052	0,389
Quito	0,064	0,063
Melbourne	1,168	0,695

Oslo: Entropía media. Notable disminución en Hcond, saber el estado actual ayuda mucho a predecir el siguiente.

Quito: Bajísima entropía, el clima es muy predecible, casi no hay incertidumbre.

Melbourne: Mayor entropía total y condicional, más incertidumbre y menor previsibilidad.

Análisis de códigos de Huffman

Quito

$$C([F, T, C]) = [01, 1, 00]$$

$$C([FF, FT, TF, TT, TC, CT]) = [01010, 00, 011, 1, 01011, 0100]$$

$$< L_1 >: 0.064 \leq 1.007 < 1.064 \quad < L_2 >: 0.064 \leq 1.007 < 1.064$$

En orden 1 se observa que la longitud promedio está muy lejos de la entropía, lo que indica una codificación poco eficiente a pesar de que los datos son extremadamente predecibles. Sin embargo, al extender la fuente a orden 2, la longitud promedio se reduce considerablemente y se acerca mucho más al límite inferior de Shannon. Esto permite, gracias a la frecuencia considerable del símbolo 'TT' (98%) aprovechar mejor las repeticiones constantes en los datos, logrando una codificación mucho más eficiente. Esto se ve reflejado en la compresión¹ donde se logra la tasa más alta de orden 1 (87,41%) y orden 2 (93,59%).

Conclusión: Sistema altamente redundante y predecible. Muy bueno para compresión y predicción. El clima es casi invariablemente templado.

Oslo

$$C([F, T, C]) = [1, 01, 00]$$

$$C([FF, FT, TF, TT, TC, CT, CC]) = [1, 0010, 0011, 01, 00000, 00001, 0001]$$

$$< L_1 >: 1.052 \leq 1.314 < 2.052 \quad < L_2 >: 0.721 \leq 0.780 < 1.221$$

¹ Se tomó como tamaño original la cantidad total de días multiplicado por 8 bits, el tamaño de un char.

Podemos observar por el Teorema de Shannon que la diferencia entre la longitud y la entropía no es considerable pero aún hay margen para mejorar la codificación. Al extender la fuente a orden 2, la longitud promedio desciende notablemente y se acerca mucho más a la cota inferior, lo cual tiene sentido en una ciudad con estaciones marcadas. Por ejemplo si hace frío probablemente sea invierno, donde los días son mayormente fríos, pocas veces templados y nunca cálidos. Esta mejora es coherente con la baja entropía condicional observada. Debido a estos factores, se logra una gran tasa de compresión (83,57%), la cual mejora mucho al tener en cuenta el contexto (90,17 %).

Conclusión: Aunque hay algo de variabilidad, el clima sigue siendo bastante estructurado y compresible. Las transiciones pueden predecirse fácilmente (baja Hcond).

Melbourne

$$C([F, T, C]) = [00, 1, 01]$$

$$C([FF, FT, TF, TT, TC, CT, CC]) = [101010, 10100, 101011, 0, 100, 1011, 11]$$

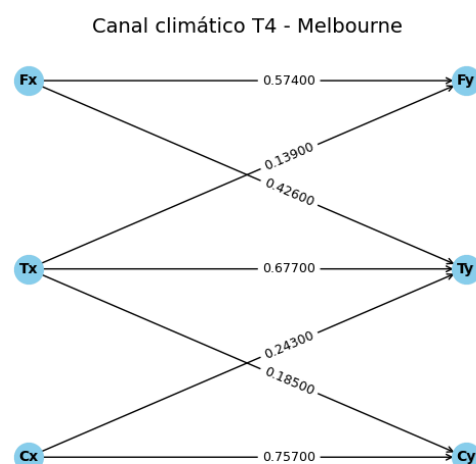
$$< L_1 >: 1.168 \leq 1.491 < 2.168 \quad < L_2 >: 0.931 \leq 0.975 < 1.431$$

La distancia entre la entropía y la longitud media es mayor que en Oslo o Quito, tanto en orden 1 como en orden 2. Esto indica que, si bien la codificación es válida, no se puede comprimir tanto porque los datos son más variados y los símbolos no tienen tanta relación entre sí. Al extender la fuente a orden 2, la eficiencia mejora, pero la longitud media sigue alejada del mínimo teórico. Esto se debe a que hay menos repeticiones claras en los datos, lo que hace más difícil aprovechar patrones. En resumen, Melbourne muestra que cuando los datos son más impredecibles, usar pares de símbolos sigue ayudando, pero la mejora es más limitada. Esto hace que el chiste “Melbourne, donde podés experimentar las cuatro estaciones antes del almuerzo” tenga sentido. La dificultad de predecir el próximo estado se ve claramente reflejada en la compresión, donde, aunque se logren tasas altas (Orden 1: 81% ; Orden 2: 87%), son las más bajas entre las ciudades.

Parte 4: Canal Melbourne ruidoso

La señal original proveniente de Melbourne fue transmitida por el satélite SpaceHeat-42, pero lo que llegó a la base terrestre difiere de la fuente original, mostrando evidencia de distorsión. Para evaluar la calidad de esta transmisión, se construyó la matriz de transición T4, donde se observa cómo cada símbolo de entrada (Fx, Tx, Cx) es transformado en una salida (Fy, Ty, Cy), y en base a eso, se analizan el ruido y la información mutua.

↙	F _x	T _x	C _x
F _y	0.574	0.139	0.000
T _y	0.426	0.677	0.243
C _y	0.000	0.185	0.757



La matriz evidencia una alta frecuencia de errores, sobre todo entre los estados templado (T) y frío (F), que se confunden entre sí de forma significativa. Por ejemplo, un símbolo Fx tiene una alta probabilidad de ser recibido como Ty. En contraste, el estado cálido (C) muestra una mayor estabilidad, manteniéndose más fiel a su valor original en la transmisión.

El valor de ruido obtenido fue 1.023, lo cual refleja que existe una cantidad considerable de incertidumbre entre lo que se transmite y lo que efectivamente se recibe. Es decir, el canal no conserva de forma precisa la información climática original, y en muchos casos, deducir el símbolo transmitido a partir del recibido se vuelve poco confiable.

Además, la información mutua calculada fue 0.319, un valor relativamente bajo, que indica que solo una pequeña parte del contenido original logra transmitirse con claridad. En otras palabras, el canal T4 transmite poca información útil, y gran parte del mensaje original se ve afectado por distorsión.

En conclusión, el canal T4 no puede considerarse confiable para transmitir datos climáticos, ya que los errores frecuentes y la poca información efectiva que transporta lo convierten en una herramienta ineficiente, más cercana a una falla de comunicación que a una innovación tecnológica real.

Conclusión

Este trabajo permitió comprender la importancia de las herramientas computacionales en el estudio de la Teoría de la Información, ya que hacen posible un análisis más detallado de los datos y facilitan la visualización de patrones y tendencias que de otro modo pasarían desapercibidos. Además, se evidenció cómo la distribución de los datos influye de manera significativa en su transmisión.

A través de técnicas de análisis y representación de datos, fue posible identificar comportamientos estacionales, comparar características climáticas entre distintas ciudades y simular escenarios estocásticos de forma más precisa.