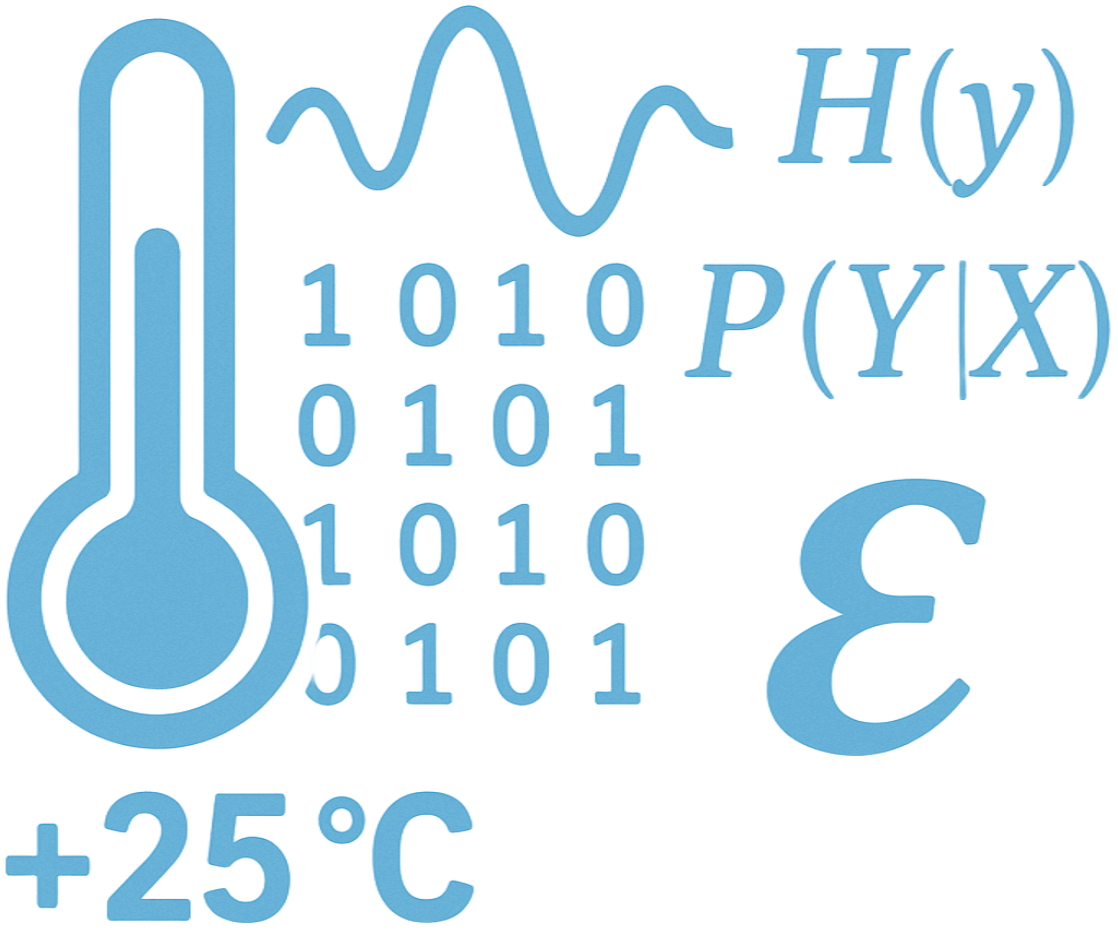
Teoría de la Información - Cursada 2025

Trabajo Práctico Especial



Integrantes:

- Roumec, Iñaki. iroumec@alumnos.exa.unicen.edu.ar.
- Velis, Ulises. uvelis@alumnos.exa.unicen.edu.ar.

Introducción.

La cátedra de la materia ha proporcionado cuatro *datasets* con la finalidad de que se realice el análisis de las temperaturas promedio de tres ciudades con condiciones meteorológicas bastante distintas. Estas tres ciudades son introducidas por la cátedra de la siguiente manera:

- Quito: *donde la temperatura no cambia ni aunque recen diez climas distintos*.
- Melbourne: *donde podés experimentar las cuatro estaciones antes del almuerzo*.
- Oslo: *donde el clima no se decide si quiere ser Siberia o un spa nórdico*.

Brindando un poco de contexto geográfico, Quito es la capital de Ecuador, Melbourne es una ciudad de Australia (fue su capital entre 1901 y 1927) y Oslo es la capital y ciudad más poblada de Noruega. Para poder tener en claro la distancia entre estos países, se invita a observar el siguiente mapa, en el que se encuentran coloreados los tres países a los que pertenecen las ciudades a analizar:

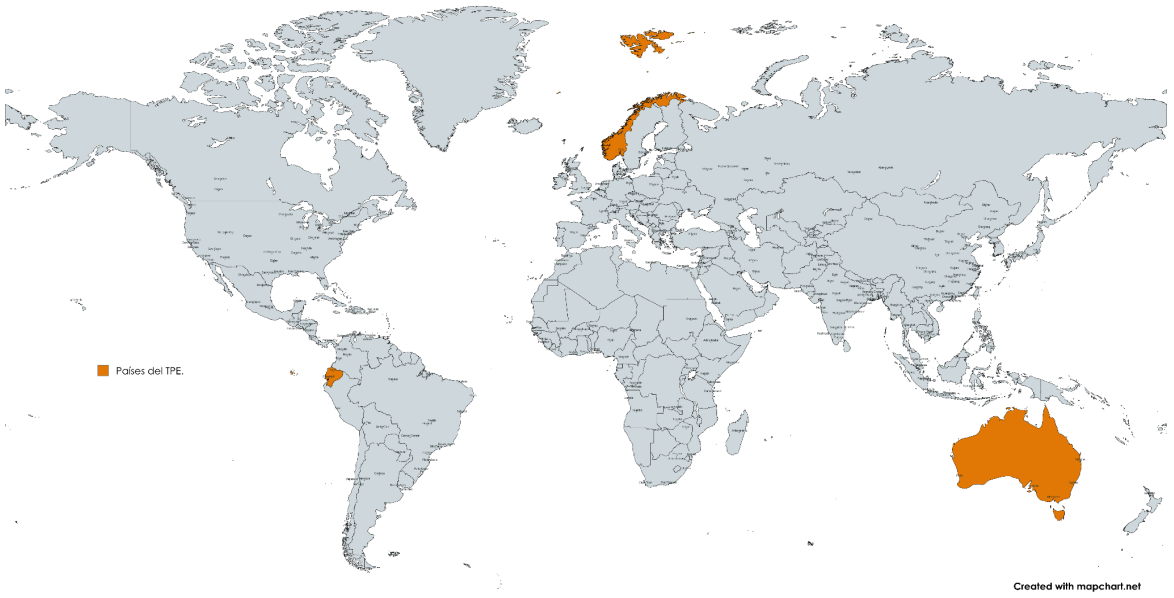


Imagen 1: planisferio en el que se hallan coloreados los países de Ecuador (izquierda), Noruega (medio superior) y Australia (esquina inferior derecha).

A lo largo del informe, se presentarán los resultados más relevantes obtenidos del análisis de los *datasets*, al igual que conclusiones y gráficos que acompañen el desarrollo del informe.

Todo el código escrito queda a disposición en el siguiente repositorio. En él, es posible observar más gráficos, conclusiones y experimentaciones que las presentadas en este informe. Es importante tener en cuenta que algunos resultados, especialmente los de simulación computacional y sus derivados (como la codificación, la tasa de compresión y el Teorema de Shannon), pueden variar levemente entre ejecuciones. Estas variaciones no alteran de ninguna forma las conclusiones planteadas en este documento.

Desarrollo y análisis.

Pasos previos.

Previo a la resolución de los puntos solicitados, se realizó un análisis de los datos a disposición. Durante su observación, se detectaron *outliers* en los *datasets* de Melbourne y Oslo, al igual que en el *dataset* de Melbourne "ruidoso" (este último será recién introducido y utilizado en la última parte). En los dos primeros *datasets* mencionados, el valor de los *outliers* es el mismo: -73. Por otro lado, en el de Melbourne ruidoso, estos *outliers* varían entre -77.0 y -68.0.

Al momento de manejarlos, se evaluaron distintas alternativas, tales como:

- Eliminar todas las entradas del *dataset* en las que aparezca dicha temperatura. Sin embargo, esto implicaría, para mantener consistencia, eliminar dichos días también de los demás *datasets*. Y, al estar realizadas las mediciones por días, perderlas dificultaría la toma de estadísticas entre años, tales como la correlación cruzada.
- Imputar los datos, a partir de sus datos vecinos. En la mayoría de las ocasiones en las que aparece un *outlier*, los valores anteriores siguen un comportamiento predecible, lo que permite estimar adecuadamente el dato que debería ocupar la posición del *outlier*.

Habiendo seleccionado la segunda alternativa como la mejor opción, se realizó una imputación por media móvil de orden 3, reemplazando el *outlier* por el promedio de los tres datos anteriores.

Parte 1: Estadísticas para ingenieros que miran el cielo.

Dadas las señales de temperaturas diarias registradas durante cierto periodo en las tres ciudades anteriormente mencionadas, expresadas como valores enteros, en °C, se pide:

1.1. Calcular la temperatura promedio y la desviación estándar para cada señal S_i y analizar cómo se comportan estadísticamente.

Los resultados obtenidos, redondeados a tres decimales, son los siguientes:

	Quito	Melbourne	Oslo
Promedio	13.604	17.803	4.771
Desvío Estándar	1.302	4.252	8.79

Tabla 1: promedio y desviación estándar de cada dataset.

De lo obtenido, se puede destacar que:

- Quito posee la menor desviación estándar, lo que se traduce en la menor variación, entre las tres ciudades, con respecto a la media, lo cual concuerda con la descripción de la ciudad proporcionada por la cátedra.

- Por otro lado, Oslo posee el promedio de temperatura más bajo (lo que sugiere que es la ciudad más fría de las tres) y la desviación estándar más alta. Esto último indica que es la ciudad, de las analizadas, en la que más dispersas se hallan las temperaturas con respecto a su media.

1.2. Calcular el factor de correlación cruzada entre cada par de señales. Discutir si existen correlaciones significativas.

La **correlación cruzada** es una medida de la similitud entre dos fuentes, considerando un posible desfase temporal entre ellas (al que se llamará *lag*). Se utiliza para identificar si hay una relación entre dos señales. El **coeficiente de correlación cruzada** permite cuantificar esta similitud. Varía entre -1 y +1, donde: -1 indica una correlación negativa perfecta; 0, ninguna correlación; y +1, una correlación positiva perfecta.

Con esto en cuenta, se calculó el factor de correlación para distintos desfases (*lags*). A continuación, se presenta el resultado más destacado: el gráfico del factor de correlación cruzada entre Melbourne y Oslo.

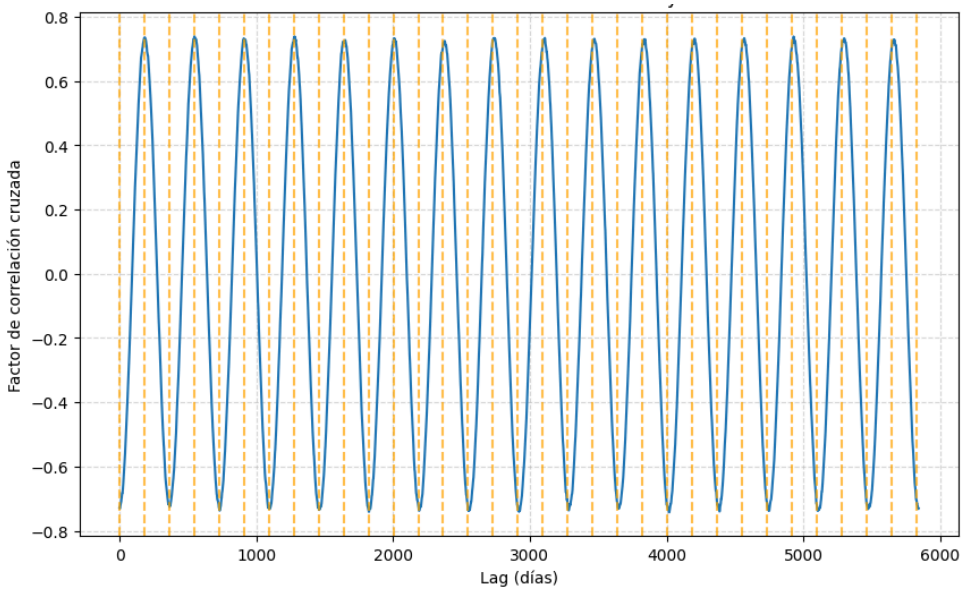


Gráfico 2: factor de correlación cruzada entre Melbourne y Oslo. En el eje x se halla el desfase ("lag"), en días, y en el eje y, el factor obtenido. El lag varía entre 0 y 5843.

A medida que se incrementa el *lag*, los resultados forman una senoidal. En el caso de los pares que incluyen a Quito, el factor de correlación lineal obtenido con suerte supera el valor absoluto 0.1, lo que sugiere que no hay un acople lineal entre Quito y las demás ciudades. Sin embargo, el resultado entre Melbourne y Oslo es diferente, alcanzando el valor absoluto de aproximadamente 0.7, lo que indica un acople estadísticamente relevante entre estas ciudades.

Las marcas anaranjadas son coincidentes con los picos o valles y fueron colocadas cada seis meses, coincidentes con el cambio de estación. Esto se debe a que, en Melbourne, es verano

de diciembre a febrero, mientras que en Oslo, es invierno durante esos mismos meses; y viceversa. Entonces, al realizar un desfase de seis meses, se están comparando las mismas estaciones entre Melbourne y Oslo. Cuando el desfase es múltiplo de un año, se obtiene un valle, el mínimo factor, puesto a que se están comparando estaciones opuestas. Al ser las curvas de comportamiento de las temperaturas promedio de estas dos ciudades similares, es esperable el resultado ante desfases de seis meses.

Analizando lo obtenido, no es posible realizar ninguna afirmación acerca de que "el clima de una ciudad prediga el de la otra", como sugiere la consigna. Al menos no de forma lineal. **El comportamiento observado es normal y esperado dada la naturaleza de las estaciones**.

Parte 2: Una fuente de calor markoviana.

En esta segunda parte del trabajo se pide que, considerando los valores de temperatura t que componen cada señal S_i, se construya una nueva señal T_i compuesta por una secuencia de símbolos discretos F, T o C, definidos según:

- F (frío): si $t < 11\,°C$
- T (templado): si $11 \leq t < 19\,°C$
- C (cálido): si $t \geq 19\,°C$

Con las fuentes discretizadas, se pide, para cada T_i:

2.1. Modelar la fuente con memoria de orden 1 (Markov), obtener la matriz de transición y analizar su comportamiento

A continuación, se presentan las matrices de transición obtenidas, en forma de grafo:

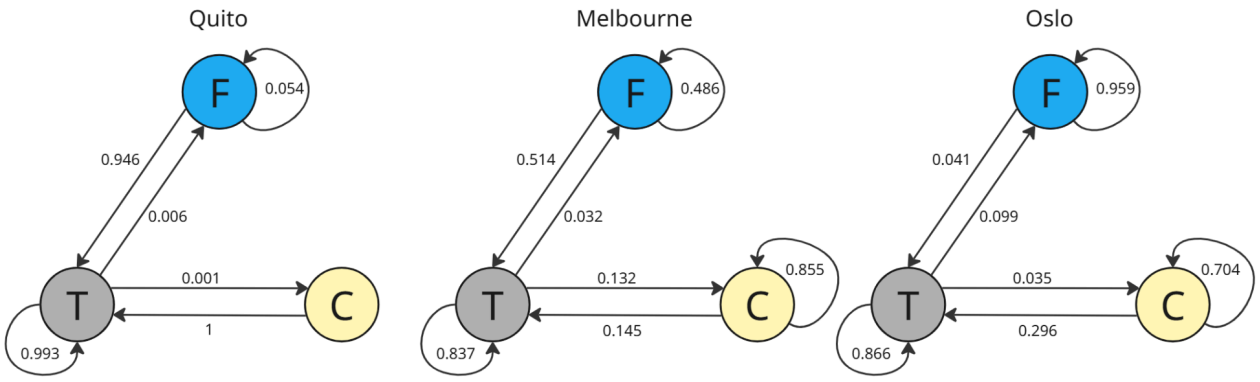


Imagen 1: matrices de transición de cada fuente discretizada.

A partir de ellas, es posible observar que:

- En Quito, independientemente de la temperatura anterior, lo más probable es que la temperatura del siguiente día sea templada. Si el día anterior fue cálido, esto es un hecho.

- En Melbourne, luego de un día frío o templado, lo más probable es encontrarse con otro día templado. Por otro lado, si el día estuvo cálido, lo más probable es que el siguiente día también lo esté.
- En Oslo, dado que un día hubo determinado clima, es muy probable que ese mismo clima se repita al día siguiente.

En todos los casos, entre un día cálido y uno frío, siempre hay, al menos, uno templado de por medio.

2.2. Usar muestreo Monte Carlo para obtener, para cada símbolo su probabilidad estacionaria y su media de primera recurrencia.

Se calcularon los vectores estacionarios para distintos umbrales. A continuación, se presentan los gráficos de convergencia del mayor umbral utilizado (0.005) y del menor (0.00005):

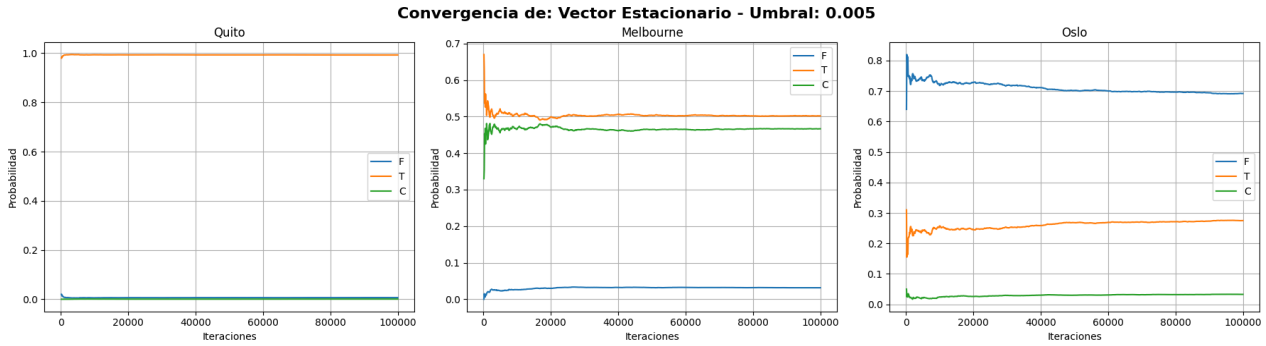


Gráfico 3: gráficos de convergencia del vector estacionario, para cada señal, con un umbral de 0.005.

	F	T	C
Quito	0.006	0.993	0.001
Melbourne	0.032	0.509	0.460
Oslo	0.684	0.282	0.034

Tabla 2: vector estacionario con umbral de 0.005 para cada ciudad.

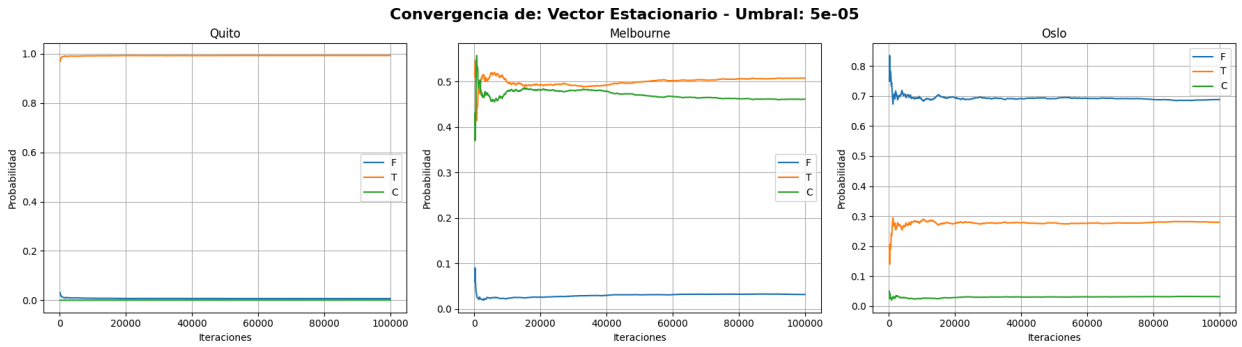


Gráfico 4: gráficos de convergencia del vector estacionario, para cada señal, con un umbral de 0.0000005.

	F	T	C
Quito	0.006	0.993	0.001
Melbourne	0.031	0.509	0.460
Oslo	0.686	0.281	0.033

Tabla 3: vector estacionario con umbral 0.0000005 para cada ciudad.

Los vectores de convergencia varían mínimamente entre umbrales. Con respecto a su convergencia, estos convergen antes del número mínimo de iteraciones establecido, lo que puede notarse observando el eje de las abscisas y en cómo la probabilidad estacionaria de cada uno de los símbolos se acerca a una línea constante.

De los resultados más "precisos" (con el menor umbral definido), es posible analizar que:

- En Quito, es muy raro presenciar un día que no sea templado.
- En Melbourne, la temperatura varía generalmente entre templada y cálida, predominando levemente las temperaturas templadas. Es muy raro hallar un día frío, aunque es más común que en Quito.
- En Oslo, la mayoría de los días son fríos, suele haber templados y casi ninguno cálido.

Con respecto a la media de primera recurrencia, a continuación se presenta el gráfico de convergencia para el umbral más alto utilizado (0.005) y para el menor (0.00005):

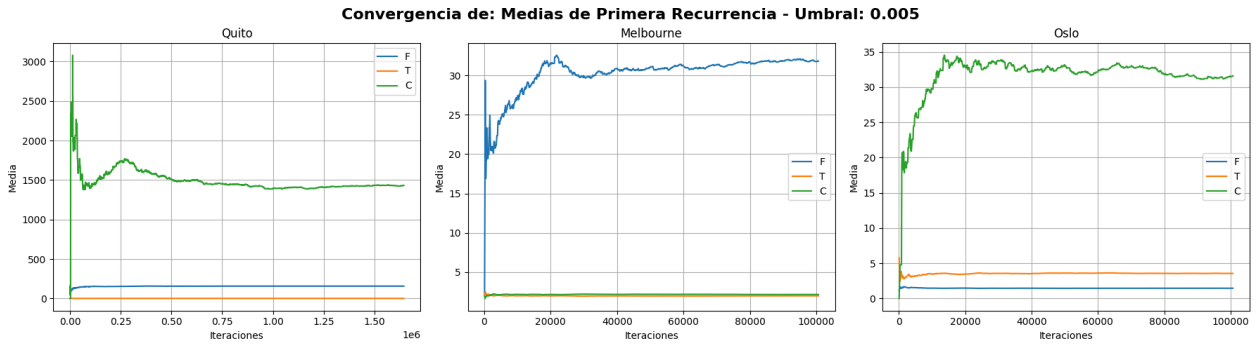


Gráfico 5: convergencia de la media de primera recurrencia, para cada señal, con un umbral de 0.005.

	F	T	C
Quito	155.897	1.007	1462.993
Melbourne	31.797	1.977	2.160
Oslo	1.461	3.547	29.831

Tabla 4: medias de primera recurrencia con umbral 0.005 para cada ciudad.

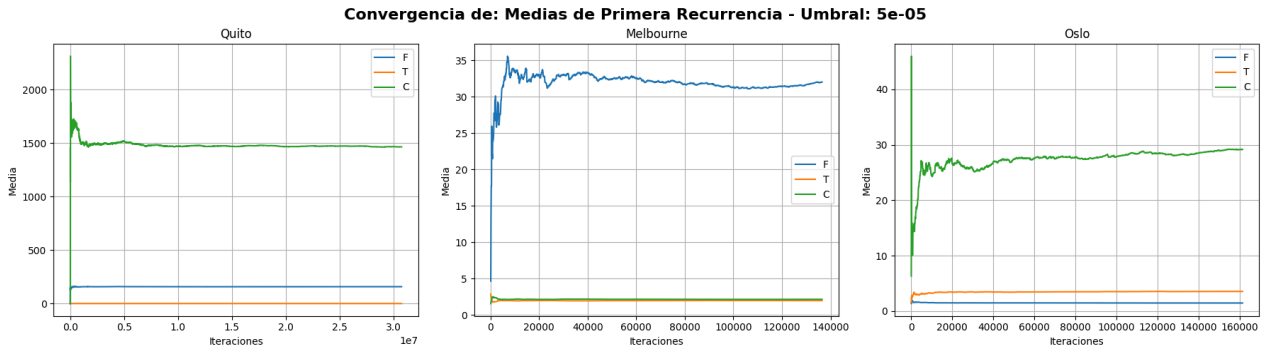


Gráfico 6: convergencia de la media de primera recurrencia, para cada señal, con un umbral de 0.00005.

	F	T	C
Quito	155.002	1.007	1442.447
Melbourne	31.984	1.978	2.159
Oslo	1.457	3.561	30.352

Tabla 5: medias de primera recurrencia con umbral 0.00005 para cada ciudad.

El tiempo medio de primera recurrencia indica, en el contexto del problema, la cantidad de días que, promedio, deben pasar para obtener un clima de características similares. En contraste con los gráficos de convergencia de los vectores estacionarios, es posible observar **una mayor variabilidad antes de alcanzar la convergencia y en los resultados para los distintos umbrales.**

Tanto en el cálculo de los vectores estacionarios como en el cálculo de las medias de primera recurrencia, la similitud entre los valores obtenidos para los distintos umbrales indica que **no es necesario el uso de un umbral chico** para obtener **resultados igualmente aproximados**.

Parte 3: Entropía, Huffman y la batalla por los bits.

3.1. Calcular, para cada fuente T_i, su entropía sin memoria, su entropía con memoria y analizar los resultados.

Los valores obtenidos, redondeados a tres decimales, se resumen en la siguiente tabla:

	Quito	Melbourne	Oslo
H1	0.067	1.173	1.013
Hcond	0.063	0.687	0.38

Tabla 6: entropía sin memoria y con memoria de cada fuente.

La entropía sin memoria se define como la cantidad de preguntas binarias que, en promedio, hay que hacer para, en el contexto del problema, conocer el clima de la ciudad.

- Una entropía baja en Quito se debe a que en la mayoría de los días suele haber el mismo clima y que se requerirían en promedio muy pocas preguntas, o incluso ninguna, para conocerlo.
- Para el caso de Melbourne y Oslo, con una entropía cercana a 1, indica que el clima es variado y que puede ser necesario (contrario a Quito), hacer preguntas para conocer el clima de un día.

Por otro lado, la entropía condicional se define como el mínimo número de preguntas binarias que, en el contexto del problema, en promedio, deben realizarse para conocer el clima de un día dado que se conoce el clima del día anterior. Con esto en cuenta, se analizaron los resultados obtenidos:

- Quito presenta la menor entropía condicional, lo que indica que, conociendo el clima del día anterior, es la ciudad en la que más sencillo es, en promedio, conocer el clima del día siguiente. Esto, nuevamente, es coherente con la descripción de la ciudad, la cual establece que la temperatura no suele cambiar.
- Melbourne, por otro lado, presenta la mayor entropía condicional, lo que se ve reflejado en una mayor incertidumbre para conocer el clima del día dado que se conoce el clima del día de ayer.

- Oslo se mantiene en un punto intermedio, donde el conocimiento del clima del día anterior no es tan útil para saber el clima del día siguiente como en Quito, pero aún así es más útil que en Melbourne.

En las tres ciudades, hay una mayor facilidad para conocer el día siguiente dado que se conoce el anterior. Esto es coherente con la idea de que la entropía con memoria, puesto a que se cuenta con información extra, nunca va a ser mayor a la entropía sin memoria.

3.2. Implementar el algoritmo de Huffman para codificar cada señal T_i y su extensión a orden 2. Aplicar el Teorema de Shannon y analizar resultados.

Las codificaciones de Huffman de orden 1, para cada señal, son las siguientes:

	Quito	Melbourne	Oslo
F	01	10	1
T	1	11	01
C	00	0	00

Tabla 7: codificaciones de Huffman de orden 1 para cada señal.

Y las de orden 2, las que se presentan a continuación:

	Quito	Melbourne	Oslo
FF	01010	101011	1
FT	011	10100	0011
TF	00	101010	0010
TT	1	11	01
TC	01011	1011	00001
CT	0100	100	00000
CC	-	0	0001

Tabla 8: codificaciones de Huffman de orden 2 para cada señal.

Debajo, los resultados obtenidos para el Teorema de Shannon:

Melbourne, orden 1:
$1.173 \leq 1.515 < 2.173$

Quito, orden 1:
$0.067 \leq 1.007 < 1.067$

Oslo, orden 1:
$1.013 \leq 1.297 < 2.013$

Melbourne, orden 2:
$0.930 \leq 0.978 < 1.430$

Quito, orden 2:
$0.065 \leq 0.512 < 0.565$

Oslo, orden 2:
$0.697 \leq 0.765 < 1.197$

Analizando lo anterior, se puede observar que, en los casos de Melbourne y Oslo, cuando la fuente se extiende, la longitud media por símbolo es relativamente cercana al límite inferior del teorema. Sin embargo, no es posible ver este mismo comportamiento en Quito, lo que despertó la pregunta: ¿a qué orden debería extenderse la fuente de Quito para que la diferencia entre el límite inferior del Teorema de Shannon y la longitud media por símbolo sea aproximadamente 0.05?

Para responder la anterior pregunta, se planteó una función que permita realizar el cálculo y se obtuvo el valor de 13 como orden ideal de extensión, con los siguientes valores en el Teorema de Shannon:

$$0.063 \leq 0.108 < 0.140$$

3.3. En cada caso, calcular la longitud total del mensaje codificado (en bits), compararla con la longitud original del archivo y obtener la tasa de compresión.

Se resumen, en la siguiente tabla, los resultados obtenidos:

	Quito	Melbourne	Oslo
Bits - Archivo Original	23758	27281	20449
Bits - Archivo Comprimido a Orden 1	5885	8990	7685
Bits - Archivo Comprimido a Orden 2	2987	5780	4596
Tasa de Compresión - Orden 1	4.037	3.035	2.661
Tasa de Compresión - Orden 2	7.925	4.72	4.449

Tabla 9: comparación entre archivo original y compresiones a orden 1 y 2.

A considerar: se tomó como archivo original el archivo sin discretizar, es decir, con las temperaturas numéricas. Además, se consideró que cada entero se codifica con la cantidad mínima de bits necesaria. En la realidad, el número de bits utilizados no varía y es el mismo para todos, por lo que las tasas de compresión serían incluso más altas. Adicionalmente, se despreciaron los bits de *padding* que el método de Huffman utiliza.

Si bien las tasas de compresión de orden uno son buenas, las de orden dos mejoran significativamente con respecto a las anteriores. Esto sugiere que, **en caso de transmitir datos correspondientes a estas ciudades, se justifica aplicar la lógica adicional para tomar de a pares los símbolos**.

Resulta particularmente llamativa la alta tasa de compresión obtenida para la fuente de Quito de segundo orden, la cual casi duplica la obtenida en la fuente sin extender. Profundizando su análisis, se halló que el símbolo TT, al tener una probabilidad de ocurrencia cercana al 99 %, recibió la menor longitud de código, como es esperable con Huffman. Como resultado, pasó de ocupar 8 bits en el archivo original (dos enteros correspondientes a dos temperaturas templadas), a requerir únicamente 1 bit (la codificación de TT). Al representar TT casi todo el archivo, es normal ver en la tasa de compresión reflejado dicho ratio (8:1).

Parte 4.

Un satélite en órbita transmite la señal S_2 (de Melbourne), pero lo que llega a la base terrestre es S_4 (Melbourne "ruidoso"). Se nos pide, a partir de S_2 y S_4:

4.1. Generar T_4 (de igual manera que se generaron las otras T_i), y construir la matriz de canal comparando T_2 (entrada) y T_4 (salida).

Comparando la entrada y salida, se obtuvo la matriz conjunta. A partir de ella, se calculó la distribución de entrada (probabilidad marginal $P(x)$) y, con ella, se obtuvo la matriz del canal que se presenta a continuación:

↙	F	T	C
F	0.574	0.139	0.0
T	0.426	0.675	0.242
C	0.0	0.186	0.758

Tabla 10: matriz de transición del canal.

4.2. Calcular el ruido del canal, su información mutua y analizar los resultados obtenidos.

El cálculo de los valores solicitados otorgó los siguientes resultados:

- $Ruido = H(Y/X) = 1.023$
- $Información\ Mutua = I(X, Y) = 0.32$

Se define el ruido como el número de preguntas binarias que, en promedio, deben realizarse para conocer el símbolo de salida dado que se conoce el de entrada. Por otro lado, la información mutua se define como el número de preguntas binarias que, en promedio, se ahorran, conociendo que la entrada y la salida están acopladas. También es descrita como la cantidad de información útil (en bits) que el canal transmite en cada salida.

El canal analizado presenta una información mutua muy baja, lo que indica que conocer la entrada aporta poca información útil sobre la salida. Todo esto, acompañado de un ruido relativamente alto (lo que se refleja en una gran incertidumbre acerca de cuál va a ser la salida dada una entrada), hace del canal uno poco confiable para la transmisión de información útil, especialmente si existen alternativas con mejores características.

Conclusiones.

A lo largo del trabajo práctico, se asentaron de forma práctica los conocimientos adquiridos durante la materia, resolviendo, en equipo, las consignas propuestas por la cátedra.

Adicionalmente, se adquirieron y profundizaron conocimientos relacionados al uso de lenguajes de programación que permiten el análisis de datos, como Python y sus librerías de visualización, y el uso de herramientas colaborativas, como Deepnote y GitHub.