

Tarea 1. Análisis de datos.

Daniel Mateo Bonilla Ramírez
Septiembre de 2018

4.6

Muy buen trabajo :)

Introducción

A continuación se llevará a cabo el análisis exploratorio de un subconjunto de datos provenientes de mediciones del campo de velocidad alrededor de un cuerpo en un túnel de viento en el *Laboratoire de Mécanique de Lille*. Tales mediciones se obtuvieron utilizando SPIV ('*Stereo Particle Image Velocimetry*') con la finalidad de estudiar las características del flujo en presencia de un gradiente de presión adverso, y antes de que se forme una zona de separación. Los datos recolectados y la documentación son de acceso público y se encuentran disponibles en la base de datos [TurBase](#).

En este trabajo se tomará un punto en el dominio de medición y se utilizarán técnicas de estadística descriptiva para extraer información sobre el comportamiento general de las tres componentes de velocidad en dicho punto. Dado que se cuenta con información de la velocidad cerca de una superficie sólida, se buscará que el punto elegido permita identificar procesos que ocurren allí.

Origen de los datos

La información con la que se trabajará proviene de mediciones experimentales llevadas a cabo en un túnel de viento con longitud de $21.6m$, ancho $2m$ y alto $1m$ en el *Laboratoire de Mécanique de Lille* con las cuales se pretendía estudiar el comportamiento de la capa límite en presencia de un gradiente de presión adverso (es decir, $\partial P/\partial x > 0$). Tal gradiente de presión fue obtenido ubicando en el piso del túnel un obstáculo cuya geometría garantizaba que no se alcanzaría la separación, esto con el fin de evitar que el flujo fuera completamente tridimensional (Stanislas, Foucaut, y Kostas, 2005). La forma de dicho obstáculo se ilustra en la Figura 1.

Las mediciones se llevaron a cabo en estado permanente con una velocidad media del aire de $10m/s$. Para caracterizar el flujo se elaboraron mediciones en dos secciones transversales al eje del túnel ubicadas a $17.100m$ y $17.900m$ de la entrada; estas distancias corresponden respectivamente al punto más alto del obstáculo y un punto en la transición de aguas abajo. Adicionalmente, se midió el campo de velocidad en la sección longitudinal central tomando ventanas centradas en éstas abscisas, y en tres secciones paralelas a la superficie del obstáculo a $3, 9$ y $25mm$; todas ellas centradas en la sección ubicada a $17.900m$ de la entrada (Stanislas y cols., 2005). En la Figura 2 se ilustra la localización de estos planos de medición.

En las secciones transversales y longitudinales se capturaron 2000 campos de velocidad con una frecuencia de $0.25s$ con pixeles de $0.75 \times 0.75mm$. En las tres secciones paralelas a la superficie del obstáculo se capturaron 1000 campos de velocidad con la misma frecuencia e igual tamaño de celda.

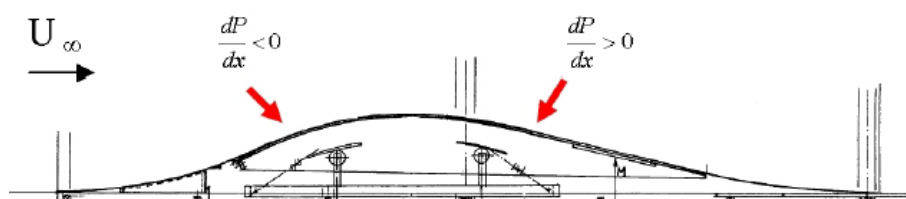


Figura 1. Esquema del obstáculo utilizado para generar el gradiente de presión adverso. Tomado de Stanislas y cols. (2005)

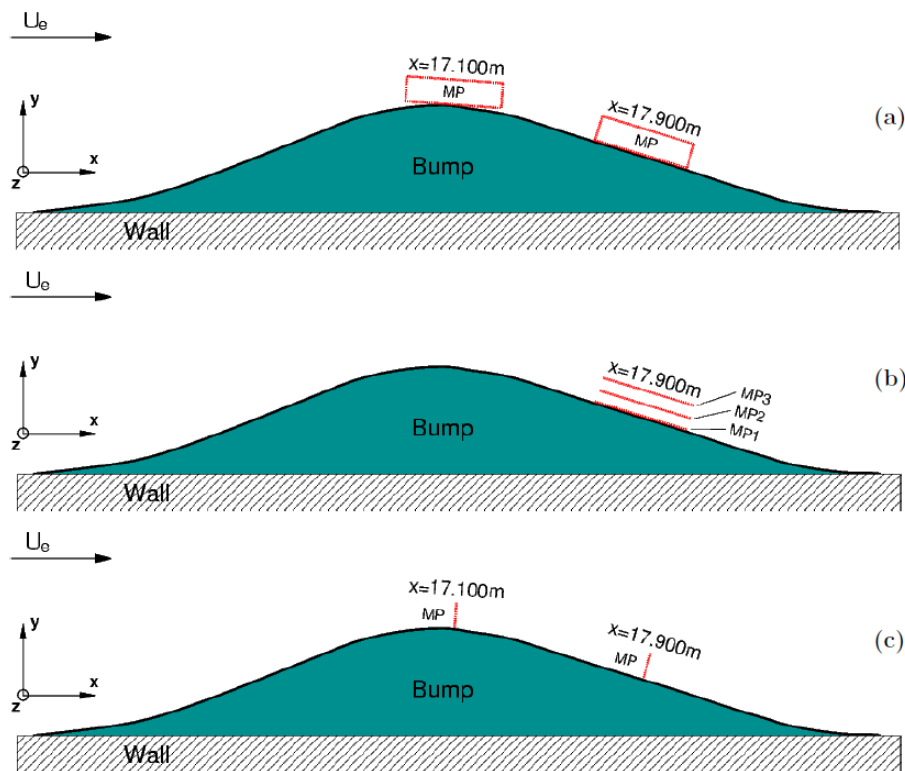


Figura 2. Localización de los planos de medición (MP).
Tomado de Stanislas y cols. (2005)

Elección del punto de estudio

Para los fines de este trabajo se utilizarán los datos de velocidad medidos en la sección longitudinal ubicada a $17.100m$ de la entrada por contar con una mayor extensión (respecto a las mediciones en las secciones paralelas a la superficie) y, dado que si bien por ahora no se llevará a cabo un análisis espacial, es de interés personal observar cómo se genera la transición de un flujo con gradiente de presión negativo a positivo.

Para simplificar el análisis del comportamiento en el tiempo se escoge un punto del dominio. La elección se hizo buscando que estuviera lo más cerca posible a la pared, tuviera una pequeña cantidad de datos faltantes, y permitiera observar lo que ocurre cerca del punto de cambio en el signo del gradiente de presión. Dado que se espera que el cambio de signo en el gradiente de presión se ubique cerca de la sección de máxima altura (en $x = 17.100m$), se buscó que el punto a estudiar estuviese sobre esta sección o cerca de ella.

Al observar los 'flags' reportados con los datos se encontró que alrededor de esta sección, en los puntos medidos a una distancia menor a $2.25mm$ de la pared la cantidad de datos que debían ser removidos, que no se habían medido o que se habían calculado por interpolación era muy alta por lo que se tomó este como el valor de la altura del punto. Así mismo, se encontró que en $x = 17100.75m$ a $2.25mm$ de la pared no habían datos faltantes o que debieran ser removidos, por lo que se escogió esta como la posición idónea.

Características de las series de tiempo

En la Figura 3 se presenta la serie de tiempo de las tres componentes de velocidad en el punto ubicado a $17100.75mm$ de la entrada al túnel de viento y a $2.25mm$ de la superficie del obstáculo. En las series de tiempo ilustradas en la Figura 3 se observa que no parece haber tendencia en

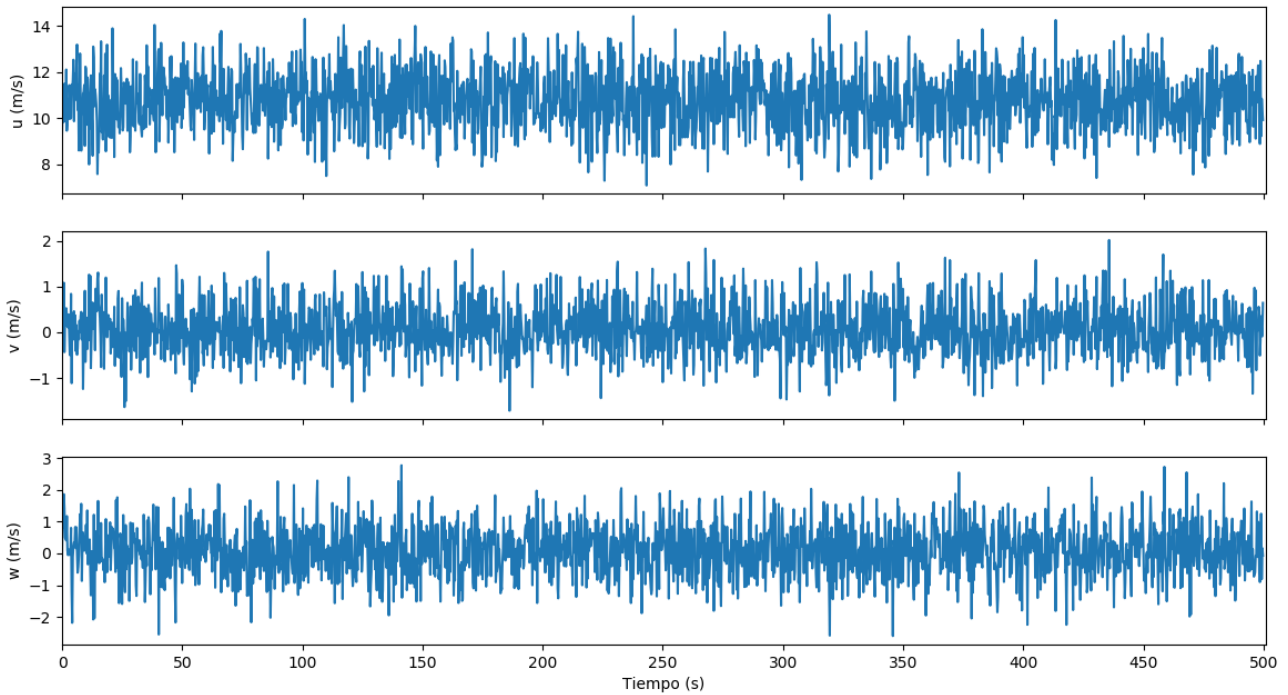


Figura 3. Series de tiempo de las componentes de velocidad en el punto $(17100.75, 2.25)mm$.

ninguna de las componentes de la velocidad, pero sí parecen tener una alta variabilidad. Ambos resultados corresponden con lo esperado: No debería haber tendencia en las series pues esto indicaría no estacionariedad y las mediciones se llevaron a cabo en estado permanente; y como el flujo es turbulento se espera que se presenten importantes fluctuaciones de la velocidad.

Adicionalmente se observa que las componentes en dirección vertical (v) y normal al plano de medición (w) parecen estar centradas en 0, mientras que la velocidad en dirección longitudinal (u) es mucho más grande. Para verificar estas observaciones, en la Tabla 1 se presentan estimaciones de los primeros tres momentos para estas series de datos.

En lo que respecta a la localización, en u se observa la cercanía entre los diferentes estimadores de la localización indicando que la serie se distribuye simétricamente; esto se comprueba al observar los valores para los estimadores de la asimetría (ambos son muy cercanos a 0); además, tal y como se esperaba, en este punto la velocidad media es mayor que la del flujo uniforme debido a la reducción en el tamaño de la sección disponible para el flujo del aire.

En v se corrobora que en efecto la velocidad está centrada cerca de 0, pero tendiendo a ser positiva y, además, los estimadores de la asimetría indican que los datos se extienden un poco más del lado de los valores positivos (aunque la distribución es esencialmente simétrica). Ambos resultados se

Tabla 1. Estimadores de los primeros tres momentos para las series de tiempo.

Medida	Estimador	u	v	w
Localización [m/s]	Media	10.7712	0.0663	0.0943
	Mediana	10.7845	0.0231	0.1057
	Trimedia	10.7832	0.0386	0.1015
Dispersión [m/s]	Desv. Estándar	1.2646	0.5517	0.7879
	IQR	1.8007	0.7091	1.0440
	MAD	0.8995	0.3516	0.5213
Asimetría	Skewness	-0.0224	0.2162	-0.0264
	Yule-Kendall	-0.0029	0.0874	-0.0162

consideran razonables pues en este punto, debido a la inercia de las partículas de fluido, se esperaría que éstas mantuvieran su tendencia a seguir subiendo como lo venían haciendo antes del punto de máxima altura, lo que implica una velocidad con componente vertical positiva; sin embargo, su magnitud no podría ser muy grande pues de otra manera se presentaría un desprendimiento de la capa límite y se sabe que el obstáculo fue diseñado de forma tal que esto no ocurriese con valores de velocidad moderados.

La cercanía entre los estimadores de la localización en w indican nuevamente una distribución bastante simétrica como lo evidencian los estimadores de la asimetría. En esta serie, si bien la velocidad media no es muy alta, sorprende que, ante la ausencia de agentes externos que cambien el momentum de las partículas de fluido en esta dirección, su valor no sea más cercano a 0. El resultado obtenido indica una tendencia al movimiento en dirección transversal; la exploración de una posible causa para esto requiere un análisis espacial en el que se estudie la conservación de la masa y el momentum lineal en la vecindad de este punto.

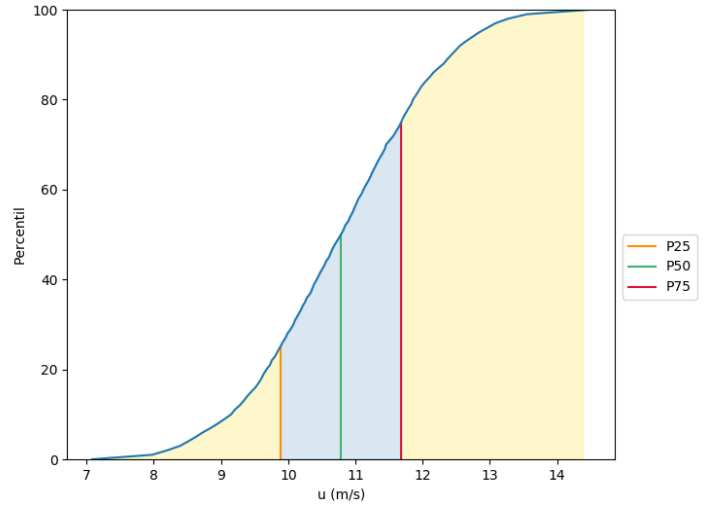
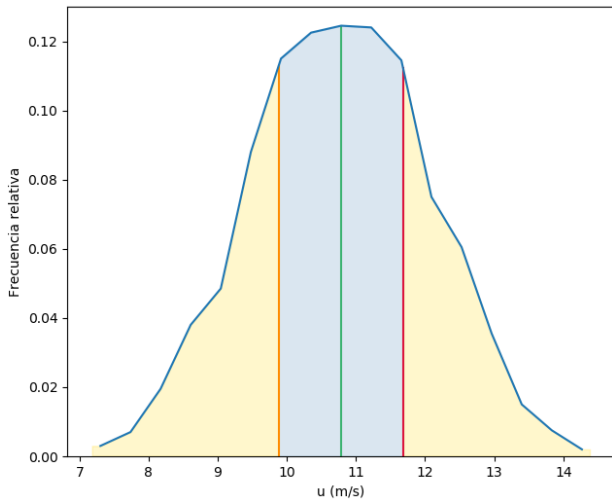
En lo concerniente a la dispersión, se observa que los valores más altos de sus estimadores están en u y los más bajos en v . Este resultado se considera razonable dado que, en promedio, las mayores velocidades están en u , seguidas por las de w y finalmente en v , lo que indica que hay una preferencia por el transporte del momentum tanto en el flujo medio como en el turbulento (asociado con las fluctuaciones) en x , seguida por z y finalmente y . Abandonando por un momento el rigor, esto es equivalente a decir que el flujo es *más turbulento* en x que en z y este a su vez es *más turbulento* que el flujo en dirección y . Es evidente además que, en v y w la dispersión es significativamente mayor que el valor medio; esto quiere decir que en estas direcciones el transporte de momentum es responsabilidad del flujo turbulento, principalmente.

En cuanto a la asimetría, nuevamente se resalta la cercanía de los estimadores a 0 indicando que las series son esencialmente simétricas. Como se esperaba el valor de la asimetría utilizando 'Skewness' es mayor (en valor absoluto) que el estimado por Yule-Kendall debido a la forma de cálculo que amplifica significativamente el efecto de los valores extremos.

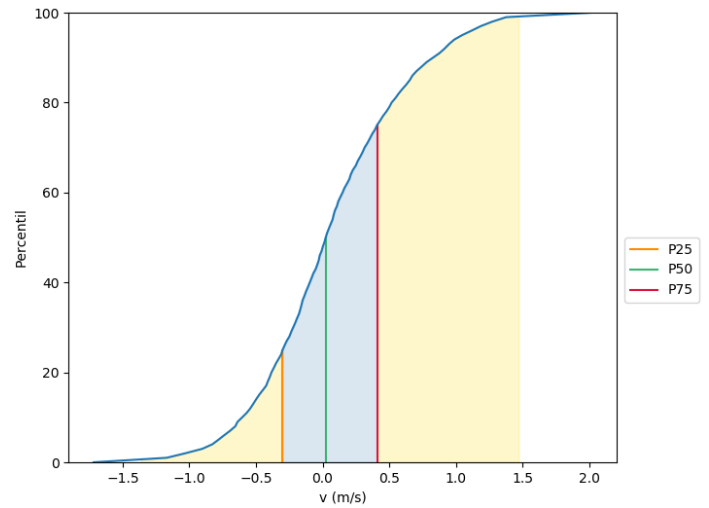
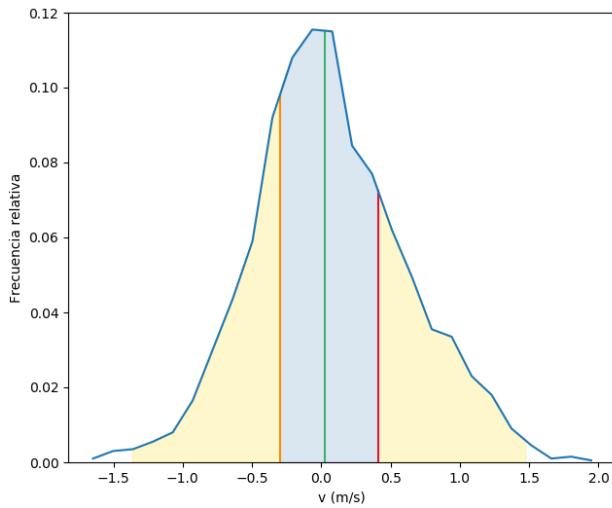
Con el fin de observar la proporción de valores extremos presentes en las series analizadas y visualizar mejor la distribución y especialmente la dispersión de los datos, en la Figura 4 se presentan las distribuciones de frecuencia relativa y frecuencia acumulada para las series de tiempo de las tres componentes de velocidad. En cada una de estas figuras se han elaborado además rectas verticales indicando la localización de los tres cuartiles y se ha sombreado en azul los datos localizados entre el cuartil superior e inferior, y en amarillo los datos localizados a menos de 1.5 veces el rango intercuartil (IQR) del cuartil superior e inferior. Nótese que al graficar los cuartiles y sombrear las regiones descritas, se está tomando como estimador de localización y dispersión de los datos la mediana y rango intercuartil respectivamente; si bien en este caso todas las medidas de localización resultaron muy similares, se escoge la mediana por ser un estimador resistente y robusto que podría ser empleado en otros casos donde los datos valores atípicos extremos fuesen importantes (Wilks, 2006). Para la dispersión, también se busca emplear un estimador robusto y resistente; se ha escogido el IQR sobre MAD (que también tiene estas características) debido a que los datos atípicos se suelen definir con base en el IQR como aquellos que se encuentren a más de 1.5 veces esta distancia desde el cuartil superior e inferior (Devore, 2005; Wilks, 2006).

Se observa que en las tres distribuciones la extensión del rango intercuartil en proporción con la de los datos es aproximadamente igual indicando que las tres series de tiempo tienen dispersiones similares (aunque la dispersión estimada con MAD no es tan parecida). En la pdf para v ilustrada en la Figura 4b se logra apreciar una pequeña asimetría como lo advertían los estimadores previamente calculados.

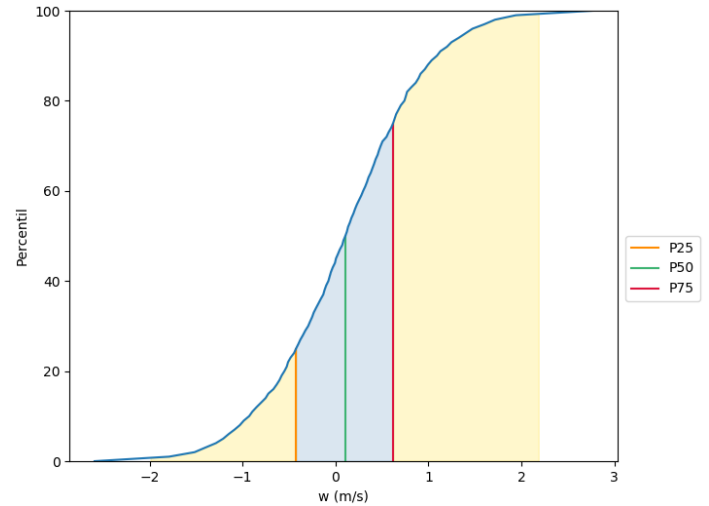
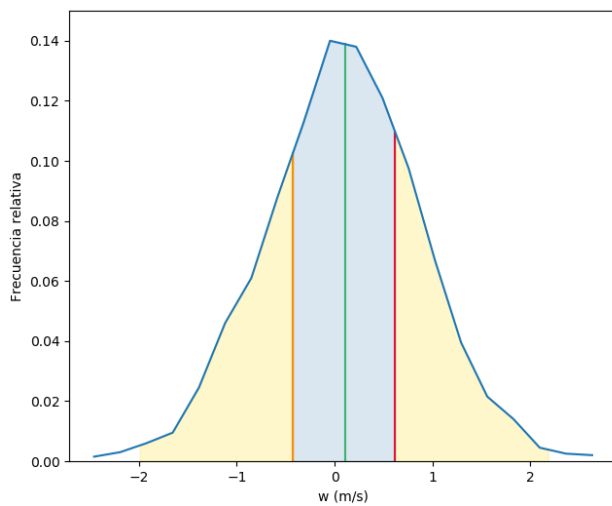
Si se acepta la definición de valores atípicos como aquellos localizados a más de 1.5IQR del cuartil superior e inferior, se observa que u no tiene valores atípicos y en v y w la proporción de ellos es



(a) Estimaciones de la pdf y cdf para la serie de u



(b) Estimaciones de la pdf y cdf para la serie de v



(c) Estimaciones de la pdf y cdf para la serie de w

Figura 4. Distribuciones de densidad de probabilidad y probabilidad acumulada para las series de tiempo. La región sombreada en azul indica los datos en el rango intercuartil y en amarillo los datos a menos de 1.5IQR del cuartil superior o inferior.

muy pequeña comparada con el grueso de los datos.

Estacionariedad de las series

Se busca ahora determinar si las series de tiempo presentan o no cambios en sus características con el tiempo; es decir, si es o no estacionario. En principio, se espera que los resultados de los análisis que se elaboren muestren estacionariedad dado que las mediciones se condujeron en un flujo permanente. Una primera evidencia de la estacionariedad es la Figura 3.

Análisis cualitativo de estacionariedad Todos los análisis presentados son cuantitativos

Para evaluar la estacionariedad de la serie se buscó, por medio de ensayo y error, una ventana de observación con longitud tal que la cantidad de datos permitiera hacer estimaciones representativas de los momentos de la distribución real y se pudieran captar algunos de los procesos del flujo turbulento. Elaborando gráficas de los histogramas y estimadores con diferentes tamaños de ventana, finalmente se escogió una ventana de observación de 2 minutos.

Para cada una de las series se obtuvo la distribución de densidad de probabilidad aproximada en todas las posiciones posibles de la ventana así como sus estimadores de localización, dispersión y asimetría. Los resultados de este procedimiento se ilustran en las figuras 5 a 7.

En la serie de la componente u de la velocidad, la Figura 5a no parece mostrar variaciones significativas en los intervalos de clase alrededor de los cuales se concentran las velocidades, aunque sí muestra leves cambios en la forma en que se concentran en tales intervalos. Esto se traduce en pequeños cambios en la localización y dispersión de los datos como se ilustra en las figuras 5b y 5c respectivamente. Es importante mencionar que, si bien en la Figura 5b la serie parece tener una tendencia decreciente, la diferencia máxima es del orden de $0.3m/s$; justamente debido a que esta diferencia es tan pequeña, se optó por no graficar allí los percentiles 10 y 90 de la serie (estos dificultaban observar esta pequeña variación).

La dispersión de la serie parece también estar cambiando en el tiempo con un rango de variabilidad importante observando cada uno de los estimadores por separado. Si bien tales cambios parecen ser cíclicos, la serie no tiene la extensión suficiente para corroborarlo.

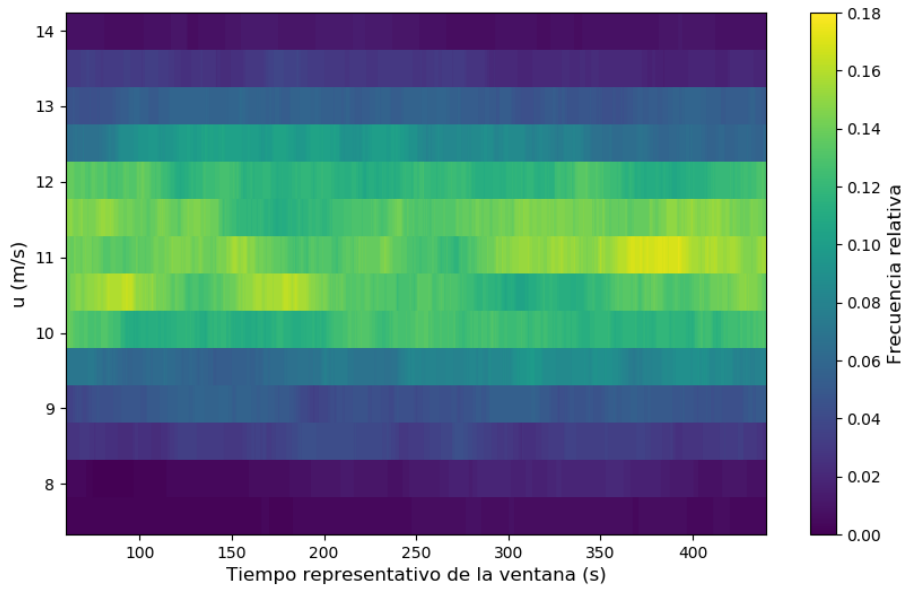
En la Figura 5d llama la atención la amplia zona de la serie donde la estimación de la asimetría por Yule-Kendall tiene signo opuesto a la respectiva estimación con 'Skewness'; en cualquier caso, los valores tienden a ser cercanos a cero y, como se esperaba, mayores (en valor absoluto) las estimaciones por 'Skewness' debido a la forma de cálculo.

En conclusión, gráficamente no parece haber una tendencia *significativa* en los momentos de la serie de u .

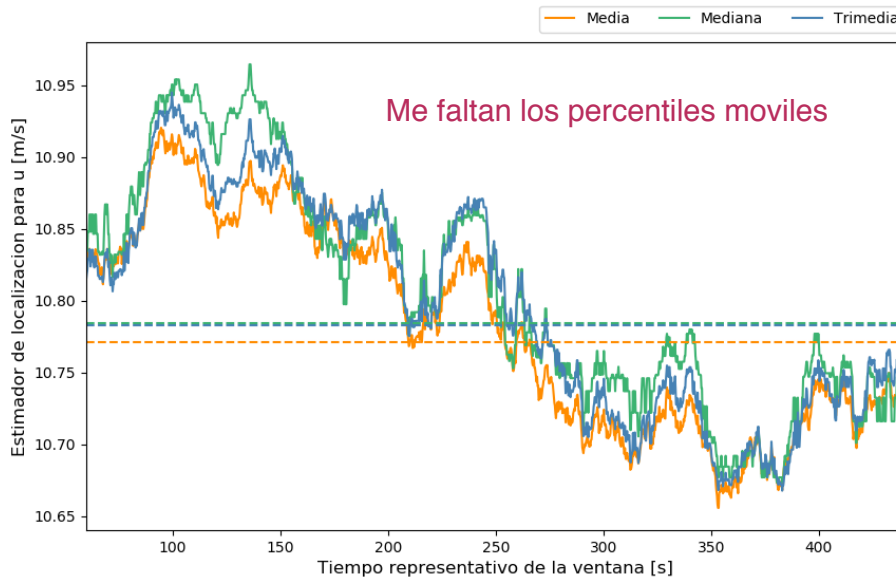
En la serie de la componente vertical de la velocidad, la Figura 6a ilustra una marcada concentración de los datos en el intervalo de clase que incluye al cero y no parece mostrar una variabilidad significativa en la dispersión por lo que se esperaría que no hubiese una tendencia marcada en la localización o la dispersión; esto se observa en las gráficas para la localización y dispersión presentadas en las figuras 6b y 6c respectivamente.

En la localización se observa que la serie podría tener una tendencia cíclica en su media; sin embargo la extensión de la misma impide corroborar esta hipótesis; aún así, de existir tal tendencia, el hecho de que la variación máxima sea menor a $0.1m/s$ indica que esta sería muy leve. Es de notar que así como en los estimadores de localización para u , estos presentan un comportamiento muy similar. Parecen solo estar un poco desfasados.

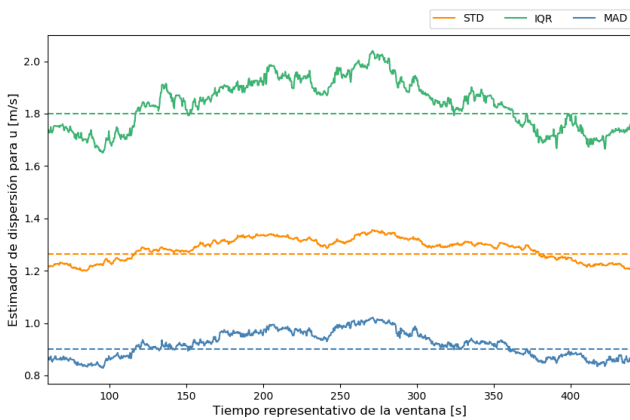
Comparando las escalas de valores en la Figura 6c y 6b se observa que la variabilidad es muy importante en esta dirección; indicando que en y el momentum es transportado principalmente por el flujo turbulento. No parecen haber signos de tendencia en la dispersión de esta serie.



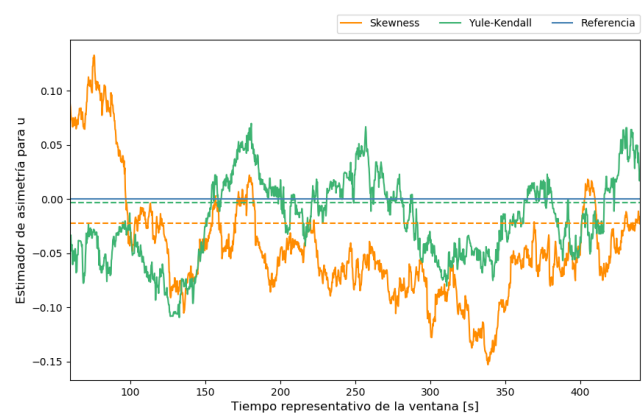
(a) Variación de las frecuencias relativas de u con el tiempo.



(b) Variación de los estimadores de localización de u con el tiempo.

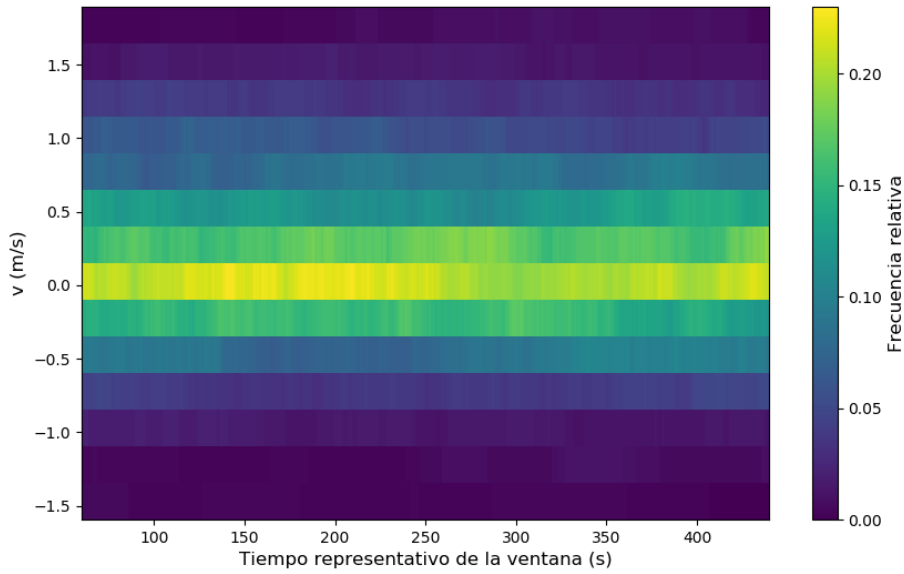


(c) Variación de la dispersión de u con el tiempo.

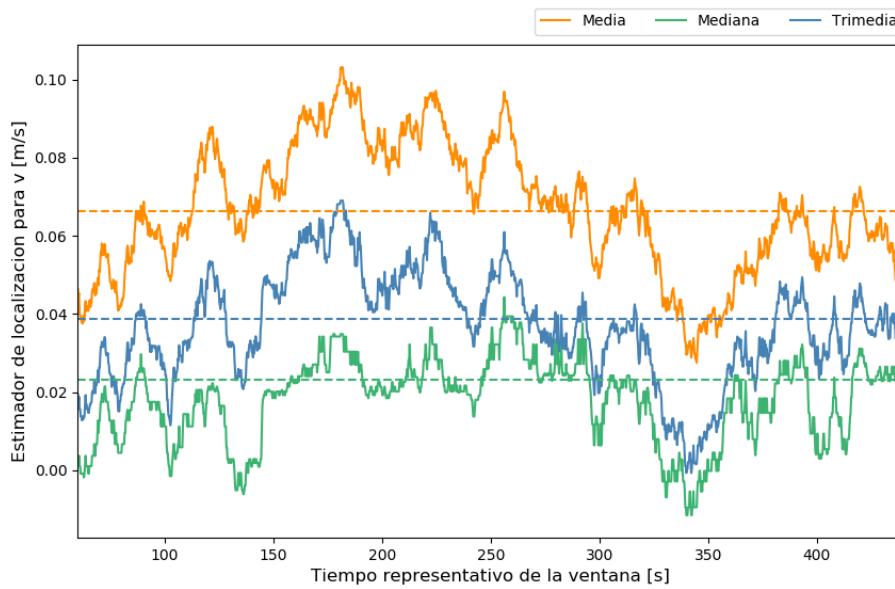


(d) Variación de la asimetría de u con el tiempo.

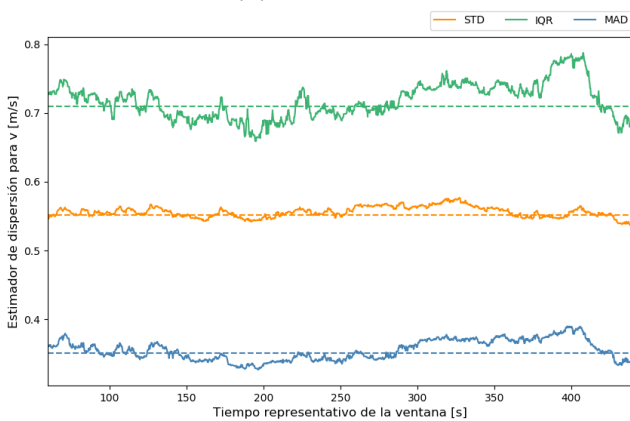
Figura 5. Variación con el tiempo de los estimadores para la distribución de densidad de probabilidad, localización, dispersión y asimetría de la serie u . Las líneas discontinuas indican los respectivos estimadores en la serie completa.



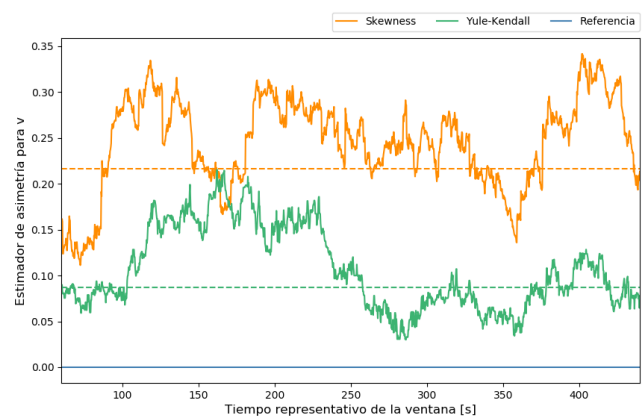
(a) Variación de las frecuencias relativas de v con el tiempo.



(b) Variación de los estimadores de localización de v con el tiempo.

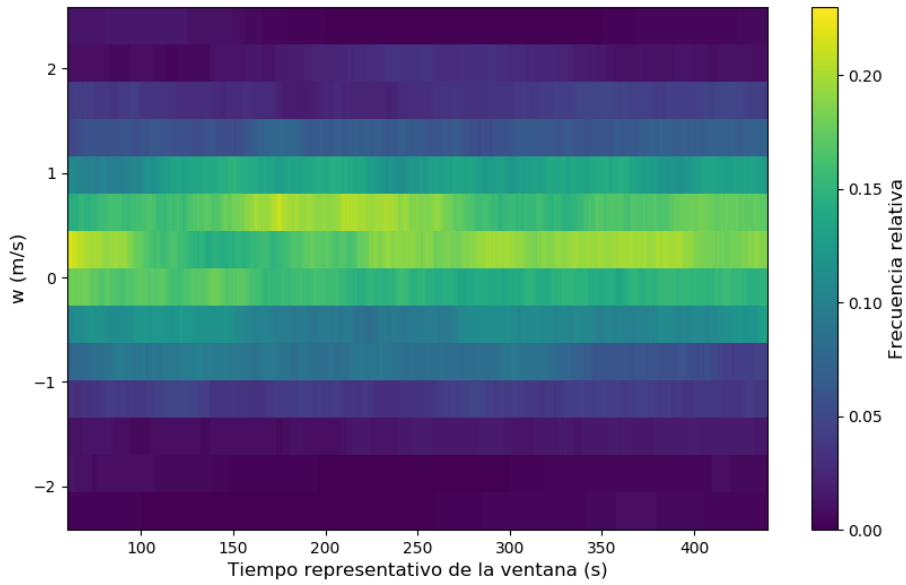


(c) Variación de la dispersión de v con el tiempo.

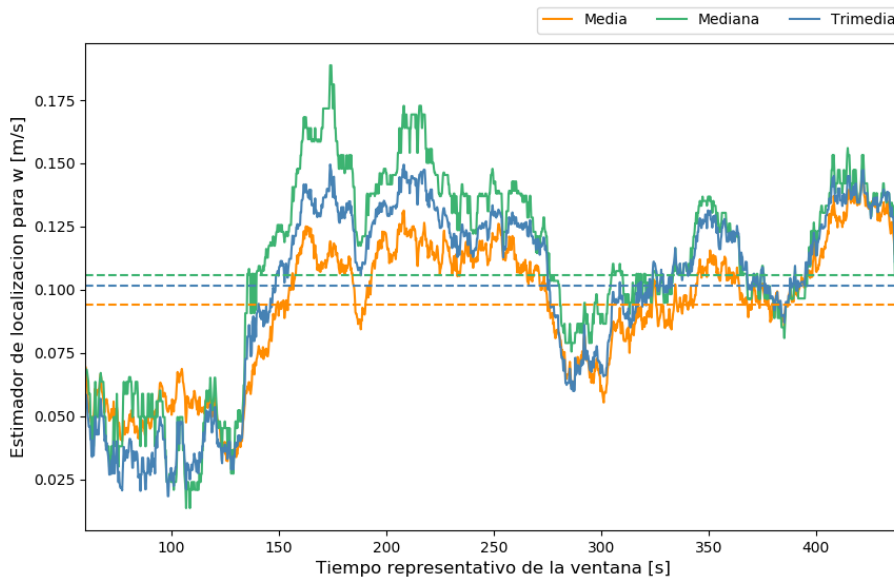


(d) Variación de la asimetría de v con el tiempo.

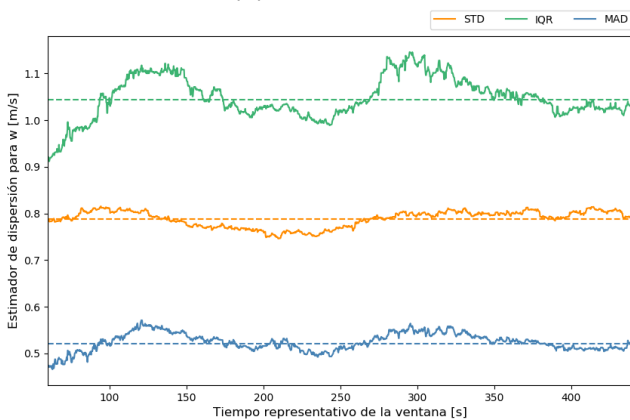
Figura 6. Variación con el tiempo de los estimadores para la distribución de densidad de probabilidad, localización, dispersión y asimetría de la serie v . Las líneas discontinuas indican los respectivos estimadores en la serie completa.



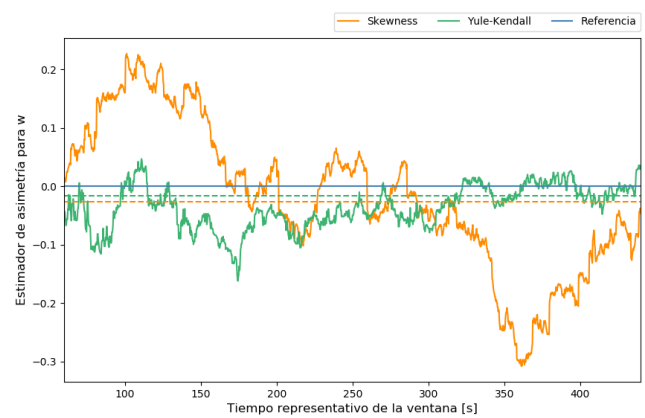
(a) Variación de las frecuencias relativas de w con el tiempo.



(b) Variación de los estimadores de localización de w con el tiempo.



(c) Variación de la dispersión de w con el tiempo.



(d) Variación de la asimetría de w con el tiempo.

Figura 7. Variación con el tiempo de los estimadores para la distribución de densidad de probabilidad, localización, dispersión y asimetría de la serie w . Las líneas discontinuas indican los respectivos estimadores en la serie completa.

En cuanto a la asimetría, se observa que en todas las ventanas de tiempo los estimadores tienen signos positivos indicando que la serie siempre tiene una mayor extensión a su lado positivo; como se mencionó previamente este resultado se esperaba y se considera debido a la inercia de las partículas de aire. Se aprecia nuevamente que las estimaciones de Yule-Kendall son menores en magnitud que las estimaciones empleando 'Skewness'. No se aprecian tendencias en los estimadores de asimetría de la serie.

Nuevamente, la evidencia gráfica no parece mostrar tendencias *significativas* en los momentos de la serie de v .

En la serie de w , la Figura 7a tiene un comportamiento muy similar al descrito en la Figura 5b para u ; en este caso, los datos tienden a concentrarse más en los intervalos de clase entre 0 y $1m/s$ alternando la mayor concentración de velocidades entre estos y sin presentar grandes variaciones en la dispersión de los mismos. Estas observaciones se corroboran en la Figura 7b donde se observa que los estimadores de localización se ubican siempre entre el intervalo de clase cercano a 0 (por encima) y el cercano a 1 (por debajo). La forma en que varían los histogramas no parece mostrar tendencias en la media o dispersión de la serie.

En la Figura 7b llama la atención el salto que ocurre en el estimador de la localización alrededor de los $150s$; sin embargo, antes de este salto y después de él los estimadores de localización tienden a ser constantes hasta transcurridos $300s$ cuando parece haber una pequeña tendencia creciente. Nuevamente, se hace énfasis en que las magnitudes de estas tendencias no parecen ser significativas; en este caso el cambio máximo es de alrededor de $0.15m/s$.

En cuanto a la dispersión, nuevamente se observa que la variabilidad es muy importante en esta dirección indicando que en z el transporte del momentum también se debe principalmente a la turbulencia. Así mismo, los estimadores de dispersión parecen ser bastante uniformes (presenta cambios menores a $0.1m/s$ en todos excepto el IQR) de manera que, junto con lo observado de la Figura 7a, se concluye que la serie no parece tener tendencia en la dispersión.

En la asimetría, ilustrada en la Figura 7d, se observa una importante variabilidad en los estimadores, pero no parece haber una tendencia clara; mientras Yule-Kendall tiende a permanecer bastante uniforme con valores negativos predominantes aunque muy cercanos a 0, mientras que la estimación por 'Skewness' tiene un comportamiento mucho menos regular. Nuevamente es clara la tendencia de Yule-Kendall a indicar valores más pequeños en magnitud que 'Skewness'. Dado que se sabe que Yule-Kendall es un estimador resistente mientras que 'Skewness' no lo es (Wilks, 2006), se considera que Yule-Kendall explica mejor el comportamiento de la asimetría en este caso.

De acuerdo con lo anterior, parece que la serie de velocidades en z no presenta tendencias *significativas* en sus momentos.

Análisis cuantitativo de estacionariedad

Para determinar cuantitativamente si la serie presenta o no correlación con el tiempo se utiliza el test no paramétrico de Mann-Kendall. En él, se evalúa:

H_0 : Los datos son obtenidos aleatoriamente de una distribución; es decir, los datos no tienen tendencia.

vs

H_a : Los datos presentan tendencia creciente o decreciente.

Dado el planteamiento de H_a , definido un nivel de confiabilidad de la prueba $1 - \alpha$, la hipótesis nula se puede rechazar si el estadístico de prueba z_{MK} satisface que:

$$|z_{MK}| \geq z_{1-\alpha/2}$$

Tabla 2. Resultados de la prueba de Mann-Kendall con $\alpha = 1\%$ para la media.

Serie	z_{MK}	$z_{1-\alpha/2}$	Resultado
u	-37.274	-2.576	Rechaza H_0
v	-11.181	-2.576	Rechaza H_0
w	21.534	-2.576	Rechaza H_0

Análisis de tendencia de los
otros estadísticos?
Percentiles?

En la Tabla 2 se presentan los resultados del test de Mann-Kendall sobre la media móvil de cada una de las series.

Se concluye que de acuerdo con el test de Mann-Kendall que las tres series de tiempo presentan tendencia en la media con una confiabilidad del 99%. Más aún, las series de u y v presentan tendencias decrecientes mientras que la serie para w presenta tendencia creciente.

El resultado es sorprendente desde el punto de vista que las tendencias observadas en la media no eran muy marcadas y va en contra de lo que se esperaba a partir del conocimiento del fenómeno físico. La presencia estadística de las tendencias se atribuye a la ausencia de empates en las series: Debido a la gran cantidad de decimales que tenían los datos empleados en este análisis se sabía que sería muy difícil encontrar dos velocidades exactamente iguales lo que demuestra que para el test de Mann-Kendall no importa la magnitud de los cambios mientras estos existan y sean sistemáticos. Es evidente entonces su potencia para detectar tendencias.

En este caso, el valor de prueba es $z_{1-\alpha/2}$ debido a que la hipótesis alternativa planteada implicaba que el test sería de dos colas. Si en lugar de esta, la hipótesis alternativa fuese "Los datos presentan tendencia creciente", el test sería de una cola y el estadístico de prueba se compararía con $z_{1-\alpha}$ (el proceso sería análogo si H_a fuese "Los datos presentan tendencia decreciente", solo que ahora se compararía si $z_{MK} \leq -z_{1-\alpha}$) (Stanislas y cols., 2005).

Referencias

- Devore, J. L. (2005). *Probabilidad y Estadística para ingeniería y ciencias*. doi: 10.1017/CBO9781107415324.004
- Gilbert, R. O. (1987). *Statistical methods for environmental pollution monitoring*.
- Stanislas, M., Foucaut, J. M., y Kostas, J. (2005). Investigation of near wall turbulence structure of an APG TBL using double SPIV.
- Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (Vol. 59). Descargado de <http://cdsweb.cern.ch/record/992087/> {%}5Cn<http://www.tandfonline.com/doi/abs/10.1198/jasa.2007.s163>{%}5Cn<http://doi.wiley.com/10.1002/met.16> doi: 10.1002/met.16