Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería Eléctrica

Reporte de Modelos

Nawel Carimán Fuenzalida

17 de Septiembre del 2020

---

1. Metodología

1.1. Procesamiento de datos

Base de datos

La base de datos utilizada cuenta con datos entre el 2020-01-24 17:00:00 al 2020-01-30 16:55:00, es decir, 6 días de información. La variable a predecir será la de medición continua de glucosa *sensor_glucose*. Este dato se encuentra en una tabla SQL junto a otras variables entregadas por el monitor continuo. Los datos fueron ordenados para tener una tasa de muestreo T_s de 5 minutos. Se utilizará la variable $y(t)$ para denotar la variable *sensor_glucose* medida en el tiempo t.

Pre procesamiento

La variable del sensor continuo de glucosa $y(t)$ cuenta con una pequeña pérdida de datos (5 valores), por lo que se realizó una interpolación lineal para manejar este problema. No se utilizó un mecanismo más sofisticado ya que la cantidad de información perdida es poca a comparación de la totalidad de los datos.

Datos de entrenamiento y prueba

El conjunto de datos se dividió en entrenamiento y prueba, con 4 y 2 días respectivamente, es decir, 66,7 % de entrenamiento y 33,3 % de prueba.

1.2. Entrenamiento, predicción y desempeño

Para cada algoritmo, se busca generar un modelo predictivo de un paso adelante (5 minutos). Luego, de modo recursivo, se obtendrá una trayectoria de predicción de seis pasos adelante (30 minutos) para finalmente evaluar el desempeño bajo distintos indicadores.

Denotaremos a la predicción como $\hat{y}_{t+k}(t), k = 1, \ldots, 6$ como la predicción de k pasos adelante al tiempo t de y.

Entrenamiento y predicción

Cada algoritmo de predicción será detallado en cada sección, donde se mostrará un gráfico de la variable $y(t)$ y las predicciones a futuro.

Desempeño

Los indicadores de desempeño se indican a continuación:

- Error de predicción: Para cada modelo, se calcularán tres tipos de errores:

  1. Error de un paso adelante: Este se define como

  $$\epsilon_{t+1}(t) = \hat{y}_{t+1}(t) - y(t) \tag{1}$$

  2. Error de seis paso adelante: Este se define como

  $$\epsilon_{t+6}(t) = \hat{y}_{t+6}(t) - y(t) \tag{2}$$

  3. Error de trayectoria: Este se define como

  $$\epsilon_{trajectory}(t) = \left[\frac{1}{6}\sum_{k=1}^{6}\left(\hat{y}_{t+k}(t) - y(t)\right)^2\right]^{1/2} \tag{3}$$

  Notar que el error de trayectoria cuenta una cota inferior (valor mínimo posible es cero) a diferencia de los demás errores. Luego, se presentará como resultado un resumen estadístico e histograma del error, gráficos en función del tiempo y un análisis en frecuencia, donde se mostrará el periodograma definido como

  $$Y_N(k) = \left|\frac{1}{\sqrt{N}}\sum_{t=1}^{N} y(t) e^{\frac{2\pi k i t}{N}}\right|^2 \tag{4}$$

  para $k = 1, \ldots, N$. También se mostrará una estimación del espectro, definida como

  $$\hat{\Phi}_y^N(\omega) = \sum_{\tau=-\gamma}^{\gamma} w_\gamma(\tau)\hat{R}_y^N(\tau)e^{-i\tau\omega} \tag{5}$$

  con $w_\gamma(\tau)$ una función ventana y $\hat{R}_y^N(\tau)$ la función de autocorrelación definida como

  $$\hat{R}_y^N(\tau) = \frac{1}{N} = \sum_{t=\tau}^{N} u(t)u(t-\tau) \tag{6}$$

  El valor de γ suele estar limitado a $\gamma = \pm N/2 - 1$, valor que se utilizará generalmente a menos que se indique lo contrario.

- Error cuadrático medio (RMSE): Este se define como:

  $$RMSE_i = \left[\frac{1}{N}\sum_{k=1}^{N}\epsilon_i(k)^2\right]^{1/2} \tag{7}$$

donde N es el número total de puntos y $\epsilon_i(k)$ son los tres errores descritos previamente, obteniendo tres errores cuadráticos medios; uno para un paso adelante, uno para seis pasos adelante y uno para la trayectoria.

- Ganancia temporal (TG): Esta se define como:

$$delay = \arg\min_{i\epsilon[0,L]} \left\{ \frac{1}{N-L} \sum_{k=1}^{N-L} \left(\hat{y}_{t+6}(k+i) - y(k)\right)^2 \right\} \tag{8}$$

$$TG = (L - delay) \cdot \Delta t \tag{9}$$

con Δt correspondiente al tiempo de muestreo y L el horizonte de predicción.

- Energía normalizada de la diferencia de segundo orden (ESOD-n): Esta se define como:

$$ESOD_n = \frac{ESOD(\hat{y}_{t+6})}{ESOD(y)} \tag{10}$$

$$= \frac{\sum_{k=3}^{N} \left(\hat{y}_{t+6}(k) - 2\hat{y}_{t+6}(k-1) + \hat{y}_{t+6}(k-2)\right)^2}{\sum_{k=3}^{N} \left(y(k) - 2y(k-1) + y(k-2)\right)^2} \tag{11}$$

2. Resultados - Resumen

En la tabla 1 se resume el RMSE para los distintos modelos entrenados hasta el momento.

	Entrenamiento			Prueba		
	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
Modelo de persistencia	6.95	34.88	23.53	6.98	33.17	22.66
Modelo AR($n_a = 2$)	3.91	26.45	16.84	4.36	27.39	17.96
Modelo AR($n_a = 30$)	3.64	25.65	16.35	4.21	26.18	17.27

Cuadro 1: Resumen del RMSE para los distintos modelos

3. Modelos AR

3.1. Descripción

En esta sección se utilizará el modelo autorregresivo AR de orden n_a, AR(n_a). Este se define como

$$y(t) = C + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) + \epsilon_(t) \tag{12}$$

Y con esto, la predicción es

$$\hat{y}_{t+1}(t) = C + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) \tag{13}$$

La librería utilizada fue *Statsmodels* de *Python*, donde el modelo de regresión cuenta con una constante C que no es ponderada por ningún valor previo. Otros programas o librerías no cuentan como esta constante.

3.2. Entrenamiento

Como se mencionó previamente, se utilizó la librería *Statsmodels* para el entrenamiento. Particularmente se utiliza *AutoReg*, clase que usa el criterio de máxima verosimilitud condicional.

3.3. Búsqueda de ordenes óptimos

Para definir los modelos a analizar, se para el orden de la regresión n_a entre 0 y 60, es decir, con un pasado máximo de 5 horas. Con esto se calculó el error del conjunto de entrenamiento y de prueba, donde los resultados se ven en los gráficos 1 y 2. De estas figuras notamos que para ordenes de $n_a > 1$ no existen una gran disminución del error. Además notamos que para n_a cercano a 30 pareciera existir un mínimo para el conjunto de pruebas.

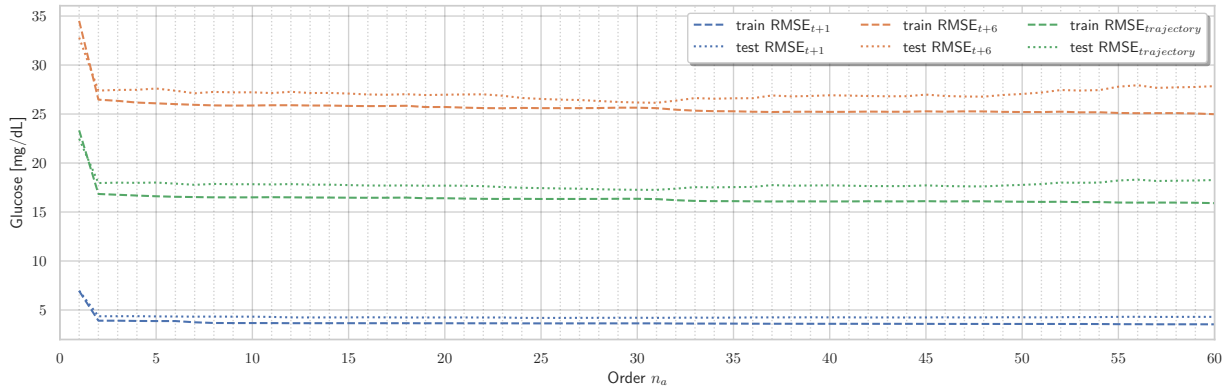

Figura 1: Gráfico del error de modelos AR para distintos ordenes de n_a

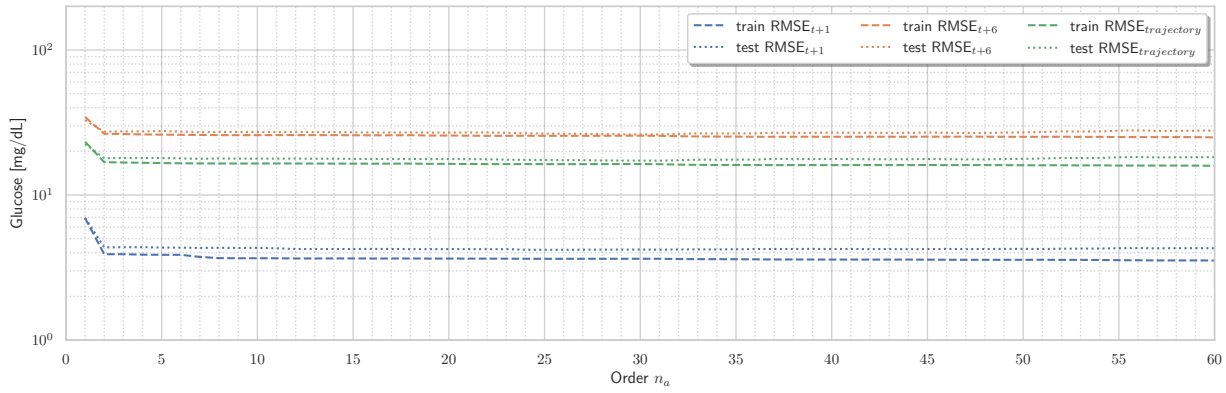

Figura 2: Gráfico del error de modelos AR para distintos ordenes de n_a en escala logarítmica

Luego, se buscó los valores que minimizaran los distintos conjuntos de entrenamiento, donde se obtuvo que con $n_a = 59$ se minimiza el error de entrenamiento (para los tres tipos de error), mientras que para el conjunto de prueba, $n_a = 25$ minimiza el error de 1 paso adelante y $n_a = 30$ el error de trayectoria y de 6 pasos adelante.

Por otro lado, la función de autocorrelación parcial (PACF) señala que un buen orden para la regresión es $n_a = 2$, mientras que en [paper de benchmark] se utilizan ordenes de 3, 6, 9 y 12. En la tabla 2 se resumen los resultados específicos para estos modelos.

	Entrenamiento			Prueba		
n_a	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
2	3.91	26.47	16.84	4.36	27.39	18
3	3.91	26.34	16.76	4.36	27.46	18
6	3.87	26.01	16.55	4.34	27.38	17.9
9	3.67	25.86	16.49	4.32	27.21	17.83
12	3.65	25.9	16.45	4.25	27.27	17.86
25	3.63	25.6	16.33	4.19	26.54	17.44
30	3.64	25.65	16.35	4.21	26.18	17.27
59	3.54	25.06	15.95	4.3	27.76	18.22

Cuadro 2: Resumen del RMSE para distintos n_a

En base a lo anterior, se analizó en profundidad los modelos para n_a de 2 y 30, ya que el primero es el modelo más sencillo que tiene un desempeño aceptable, mientras que el segundo es el que minimiza el los errores de 6 pasos adelante y de trayectoria para el conjunto de prueba.

3.4. Resultados

3.4.1. Modelo AR($n_a = 2$)

En las figuras 3 y 4 se muestra un gráfico para la predicción de la glucosa para todas las predicciones y para la de seis pasos adelante respectivamente.

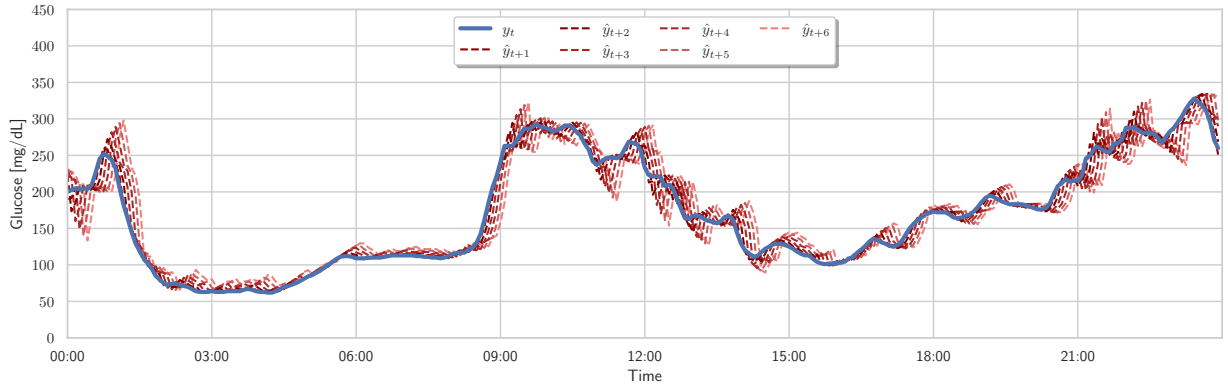

Figura 3: Gráfico de predicción de glucosa para todas las predicciones para $n_a = 2$

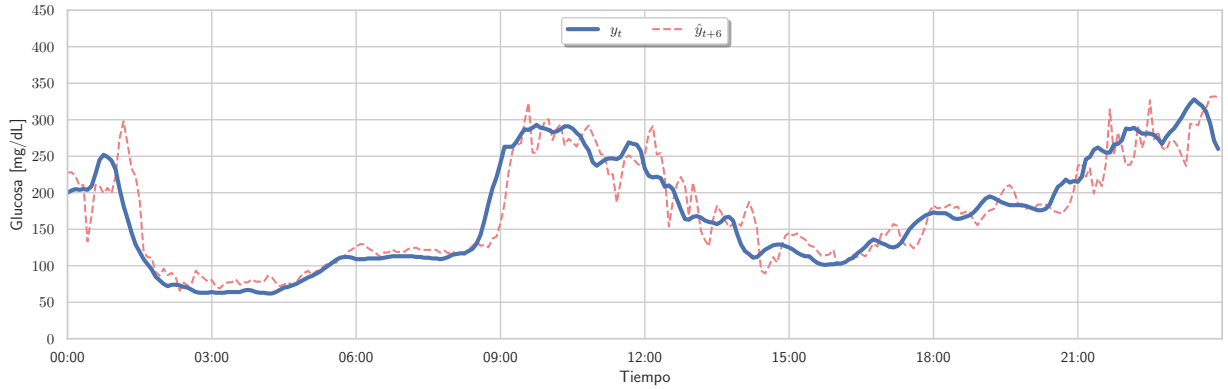

Figura 4: Gráfico de predicción de glucosa para seis pasos adelante para $n_a = 2$

Error de predicción

En las figuras 5 y 6 se puede ver los distintos errores en función del tiempo para el conjunto de entrenamiento como de prueba. Estos gráficos a simple vista no reflejan mucha diferencia con los obtenidos en el gráfico de persistencia.

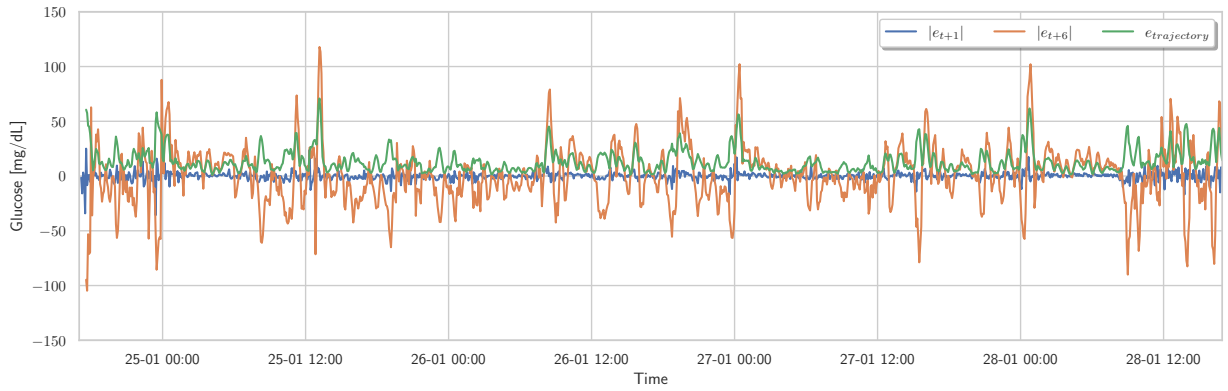

Figura 5: Gráfico del error en función del tiempo para el conjunto de entrenamiento para $n_a = 2$

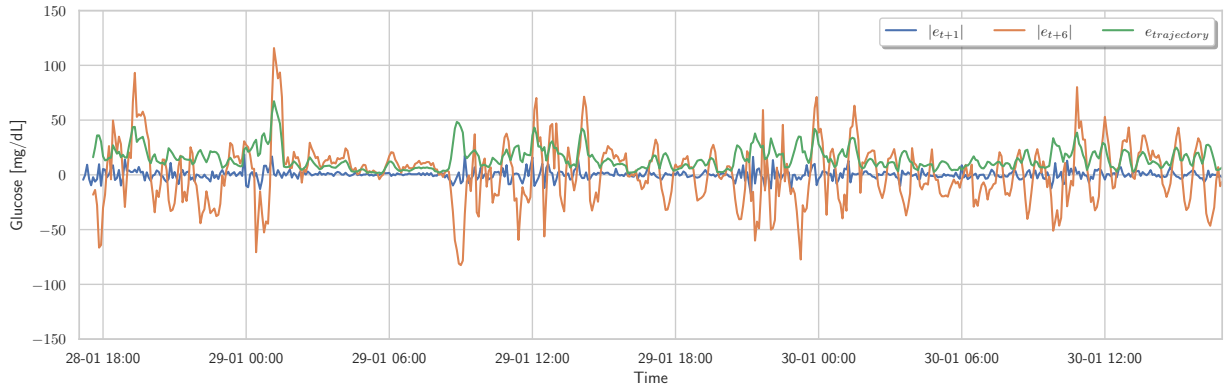

Figura 6: Gráfico del error en función del tiempo para el conjunto de prueba para $n_a = 2$

En la tabla 3 se puede ver el resumen estadístico de los errores para el conjunto de entrenamiento como de prueba, mientras que en la figura 7 se muestran los histogramas para cada conjunto. Estos gráficos si reflejan una disminución de la dispersión del error en relación al modelo de persistencia.

	Entrenamiento			Prueba		
	$\epsilon_{t+1}(t)$	$\epsilon_{t+6}(t)$	$\epsilon_{trajectory}(t)$	$\epsilon_{t+1}(t)$	$\epsilon_{t+6}(t)$	$\epsilon_{trajectory}(t)$
Número de datos	1150	1145	1145	574	569	569
Media	0	0.14	13.28	0.09	1.5	15
Desviación estándar	3.91	26.48	10.36	4.36	27.37	9.96
Mínimo	-35.76	-104.76	1.04	-18.44	-82.46	1.18
25 %	-1.74	-14.9	6.28	-2.04	-16.66	7.75
50 %	0.1	0.47	10.56	0.1	3.44	12.5
75 %	1.64	13.36	16.39	2.02	16.81	19.61
Máximo	25.02	117.86	70.83	18.72	115.81	67.26

Cuadro 3: Resumen estadístico del error para $n_a = 2$

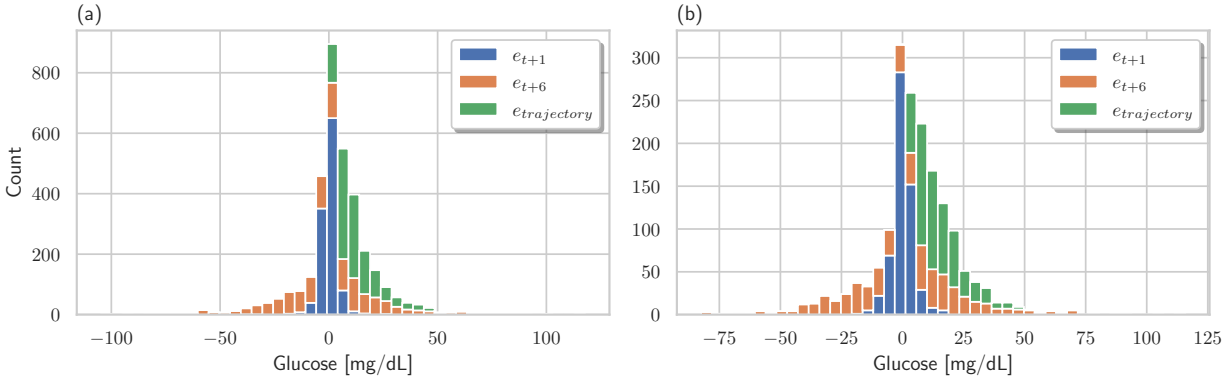

Figura 7: (a) Histograma para el conjunto de entrenamiento; (b) Histograma para el conjunto de prueba

Análisis en frecuencia

En las figuras 8 y 9 se muestra el periodograma para cada conjunto, mientras que en las figuras 10 y 11 se muestra una estimación del espectro para una ventana hanning con $\gamma = N/2 - 1$.

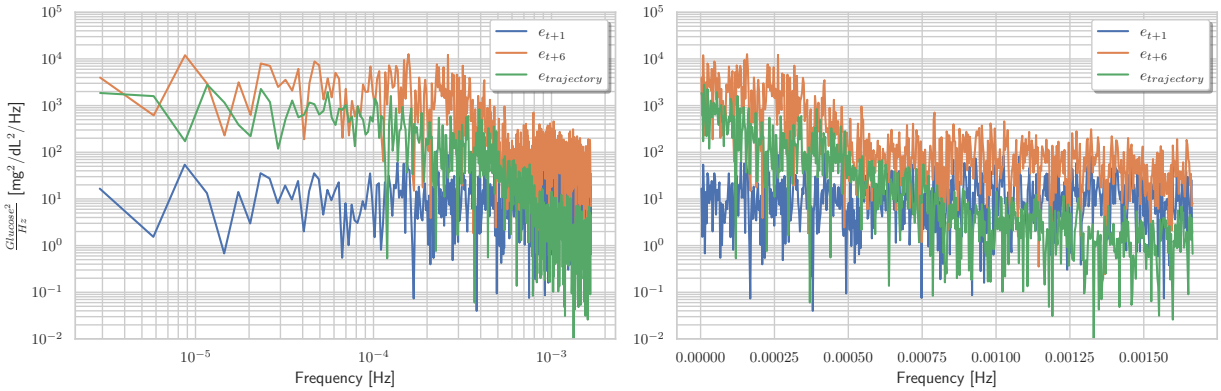

Figura 8: Gráfico del periodograma para el error del conjunto de entrenamiento para $n_a = 2$

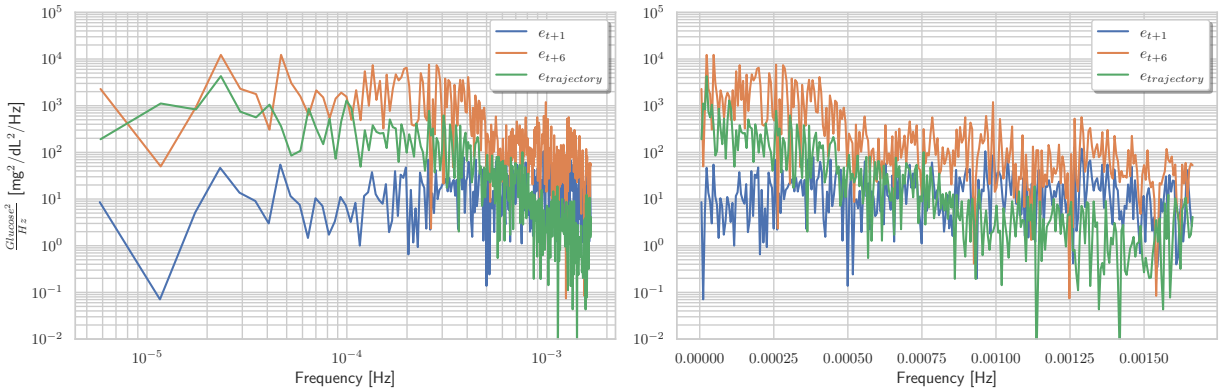

Figura 9: Gráfico del periodograma para el error del conjunto de prueba para $n_a = 2$

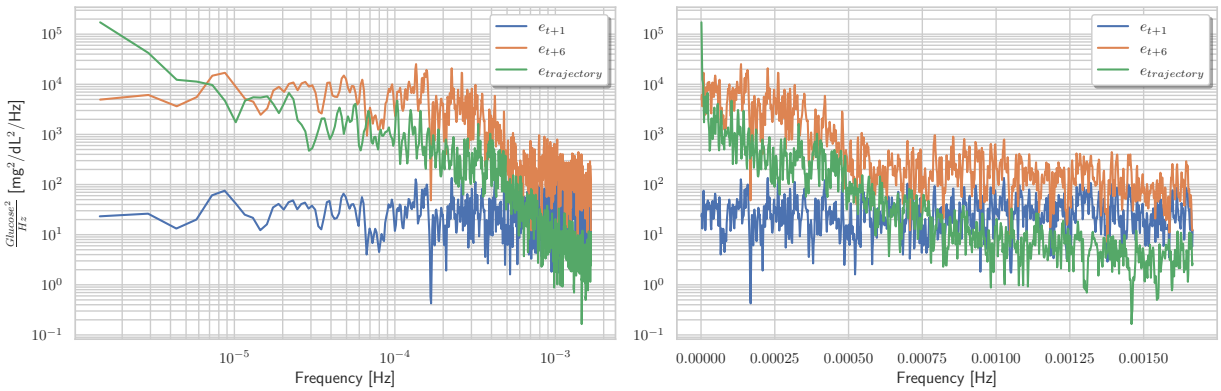

Figura 10: Gráfico de la estimación del espectro para el error del conjunto de entrenamiento para $n_a = 2$

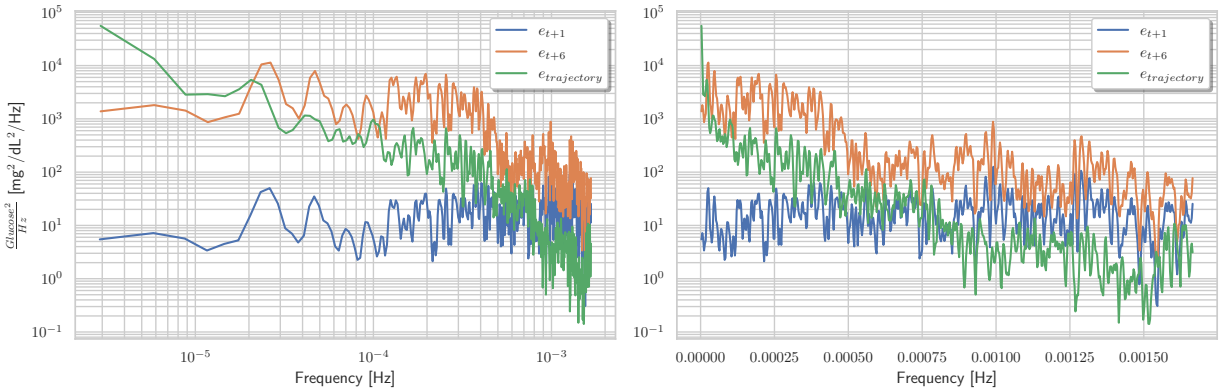

Figura 11: Gráfico de la estimación del espectro para el error del conjunto de prueba para $n_a = 2$

Métricas

Los resultados de las métricas obtenidas bajo este método se muestran en la tabla 4.

	Entrenamiento			Prueba		
	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
RMSE	3.91	26.45	16.84	4.36	27.39	17.96
TG		x			x	
ESOD-n		x			x	

Cuadro 4: Resumen de métricas

3.4.2. Modelo AR($n_a = 30$)

En las figuras 12 y 13 se muestra un gráfico para la predicción de la glucosa para todas las predicciones y para la de seis pasos adelante respectivamente.

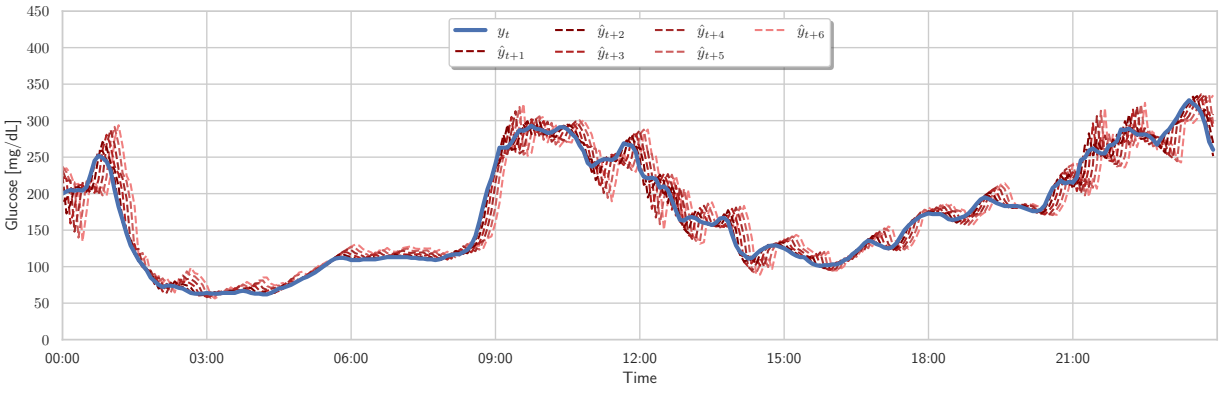

Figura 12: Gráfico de predicción de glucosa para todas las predicciones para $n_a = 30$

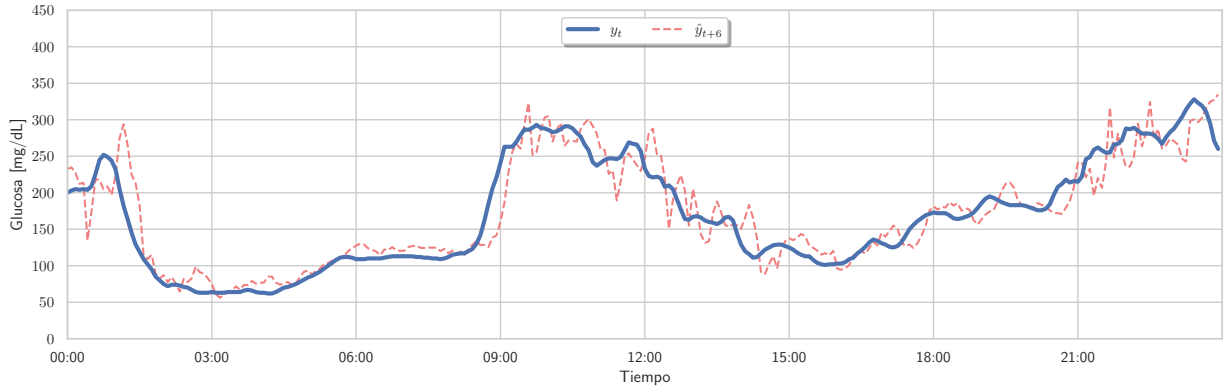

Figura 13: Gráfico de predicción de glucosa para seis pasos adelante para $n_a = 30$

Error de predicción

En las figuras 14 y 15 se puede ver los distintos errores en función del tiempo para el conjunto de entrenamiento como de prueba. En este caso, no hay mucha diferencia con el modelo AR($n_a = 2$).

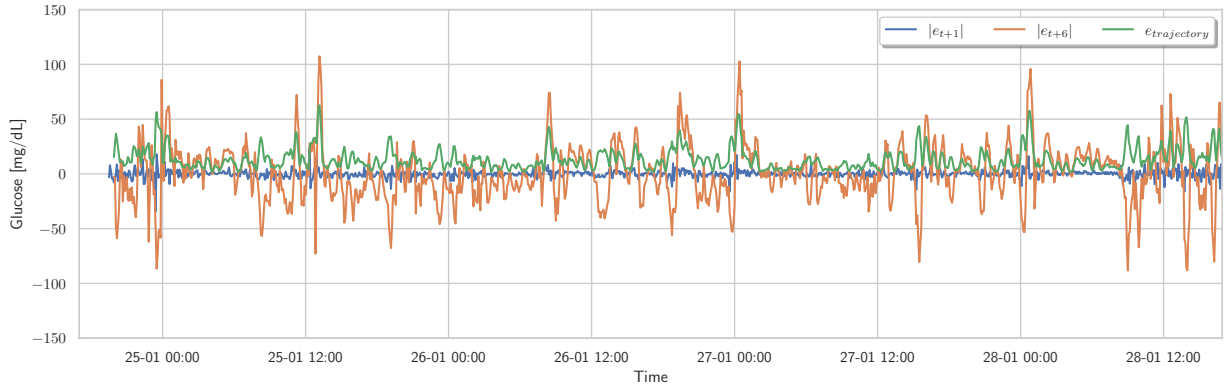

Figura 14: Gráfico del error en función del tiempo para el conjunto de entrenamiento para $n_a = 30$

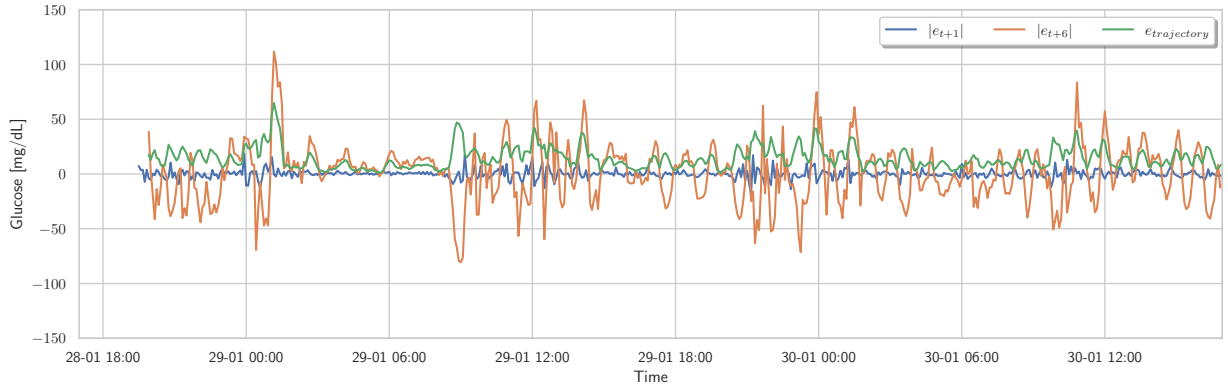

Figura 15: Gráfico del error en función del tiempo para el conjunto de prueba para $n_a = 30$

En la tabla 5 se puede ver el resumen estadístico de los errores para el conjunto de entrenamiento como de prueba, mientras que en la figura 16 se muestran los histogramas para cada conjunto. Nuevamente no hay mucha diferencia con el modelo AR($n_a = 2$).

	Entrenamiento			Prueba		
	$\epsilon_{t+1}(t)$	$\epsilon_{t+6}(t)$	$\epsilon_{trajectory}(t)$	$\epsilon_{t+1}(t)$	$\epsilon_{t+6}(t)$	$\epsilon_{trajectory}(t)$
Número de datos	1122	1117	1117	546	541	541
Media	0	-0.05	13.13	0.02	0.21	14.4
Desviación estándar	3.64	25.66	9.75	4.21	26.2	9.54
Mínimo	-34.27	-88.3	1.6	-18.32	-80.58	1.25
25 %	-1.76	-14.81	6.3	-1.95	-16.67	7.46
50 %	0.1	-0.3	10.59	-0.02	2.11	11.96
75 %	1.65	13.68	16.61	1.83	14.67	19.13
Máximo	23.35	107.47	62.63	18.7	111.77	64.9

Cuadro 5: Resumen estadístico del error para $n_a = 30$

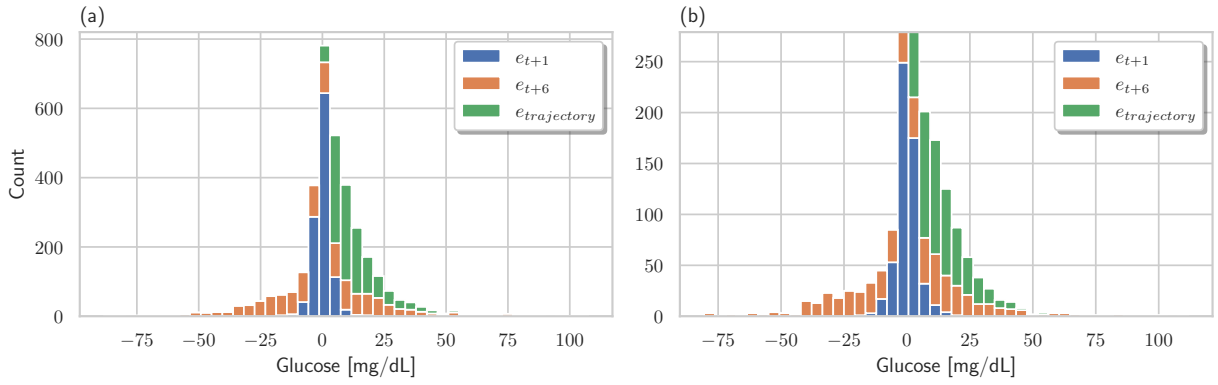

Figura 16: (a) Histograma para el conjunto de entrenamiento; (b) Histograma para el conjunto de prueba

Análisis en frecuencia

En las figuras 17 y 18 se muestra el periodograma para cada conjunto, mientras que en las figuras 19 y 20 se muestra una estimación del espectro para una ventana hanning con $\gamma = N/2 - 1$. Estos no muestran diferencia visual con $AR(n_a = 2)$.

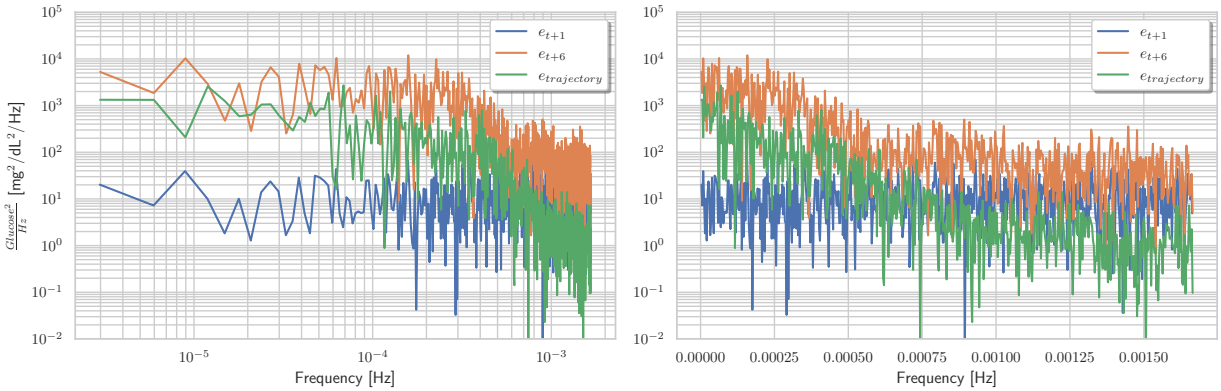

Figura 17: Gráfico del periodograma para el error del conjunto de entrenamiento para $n_a = 30$

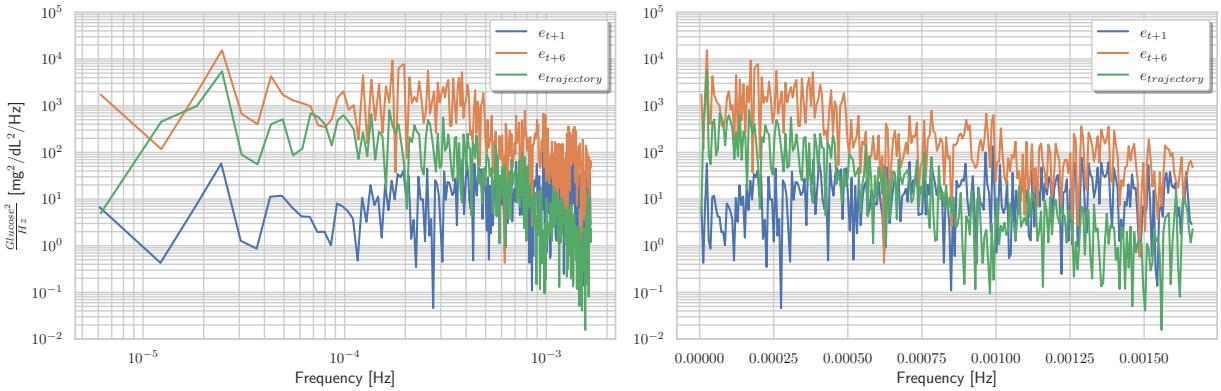

Figura 18: Gráfico del periodograma para el error del conjunto de prueba para $n_a = 30$

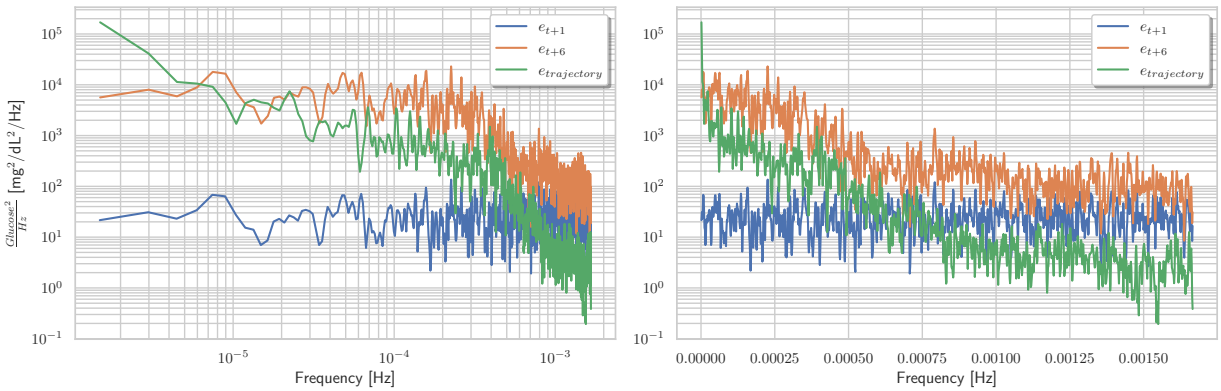

Figura 19: Gráfico de la estimación del espectro para el error del conjunto de entrenamiento para $n_a = 30$

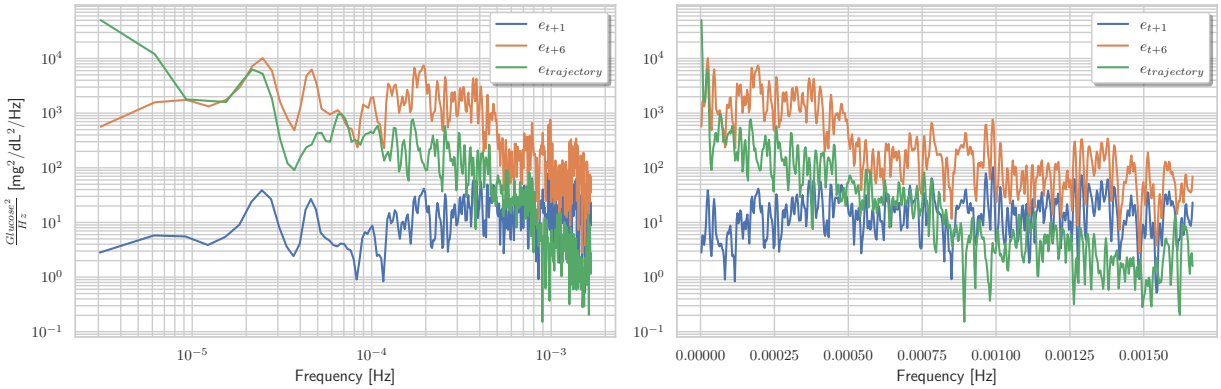

Figura 20: Gráfico de la estimación del espectro para el error del conjunto de prueba para $n_a = 30$

Métricas

Los resultados de las métricas obtenidas bajo este método se muestran en la tabla 6.

	Entrenamiento			Prueba		
	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
RMSE	3.64	25.65	16.35	4.21	26.18	17.27
TG		x			x	
ESOD-n		x			x	

Cuadro 6: Resumen de métricas