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(2)	3.91	26.45	16.84	4.36	27.39	17.96
Modelo AR(30)	3.64	25.65	16.35	4.21	26.18	17.27
Modelo ARX(48, 6) para u_{meal}	3.61	26.61	16.76	4.24	27.01	17.73
Modelo ARX(48, 4) para u_{bolo}	3.63	27.17	17.04	4.26	27.58	18.01

Cuadro 1: Resumen del RMSE para los distintos modelos

3. Modelos ARX

3.1. Descripción

El siguiente grupo de algoritmos a utilizar son los modelos autorregresivos con entradas exógenas ARX. Un modelo ARX(n_a, n_b) se define como

$$y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) + \epsilon_t \tag{12}$$

con n_a el orden de la regresión y n_b el orden de la entrada exógena. Con esto, la predicción es

$$\hat{y}_{t+1}(t) = -a_1 y(t-1) - \cdots - -a_{n_a} y(t-n_a) + b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) \tag{13}$$

Las entradas exógenas a analizar serán las de alimentos y bolo de insulina, denominadas u_{meal} y u_{bolo} respectivamente. Los modelos entrenados sólo consideran una de las dos variables, es decir, un sistema SISO, por lo que falta realizar un modelo combinando ambas entradas.

3.2. Entrenamiento

La librería *statsmodels* para *Python* cuenta con un modelo para entrar modelos ARX, pero está optimizada solo para regresiones, y la entrada exógena tolerable tiene orden 1. El resto de las librerías utilizadas cuentan con el mismo problema. Por lo tanto, para poder entrenar los modelos, se utilizó *MATlAB*, ya que cuenta con funciones de estimación de modelos ARX optimizados en sistemas dinámicos.

Dado que la función utilizada por *MATLAB* es rápida (utiliza factorización QR para resolver las ecuaciones lineales de la estimación de mínimos cuadrados), se realizó un barrido tanto para n_a y n_b entre 1 y 60 para el conjunto de entrenamiento como de prueba. Para cada modelo se calculó el RMSE_{t+6}, y se extrajo la combinación que minimizara el conjunto de prueba. Finalmente, se importaron los parámetros de estos modelos a *Python* y se analizaron en detalle.

3.3. Búsqueda de combinaciones óptimas

En las figuras 1 y 2 se puede ver el resultado de la búsqueda para la combinación óptima de n_a y n_b para el RMSE_{t+6}. Notar que para ambas figuras el conjunto de entrenamiento y de prueba tienen un comportamiento similar, donde la tendencia es siempre mejorar al aumentar el orden de n_a y n_b para el entrenamiento, pero el gráfico del conjunto de prueba muestra que esto genera un sobreentrenamiento. El valor mínimo se alcanza para $n_a = 48$ y $n_b = 6$ para u_{meal} y $n_a = 48$ y $n_b = 4$ para u_{bolo}. Adicionalmente es interesante notar que que para $n_a > 1$ y n_b libre, el error se mueve entre 24 y 26 mg/dL, por lo que un modelo más complejo podría generar un beneficio pequeño.

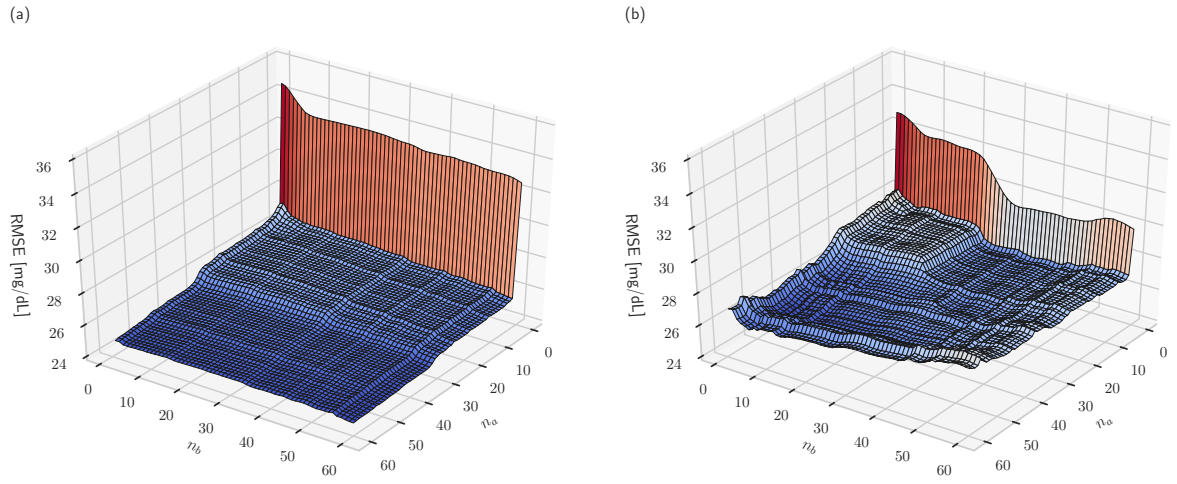

Figura 1: Gráfico de RMSE_{t+6} para entrada u_{meal}. (a) Conjunto de entrenamiento; (b) Conjunto de prueba

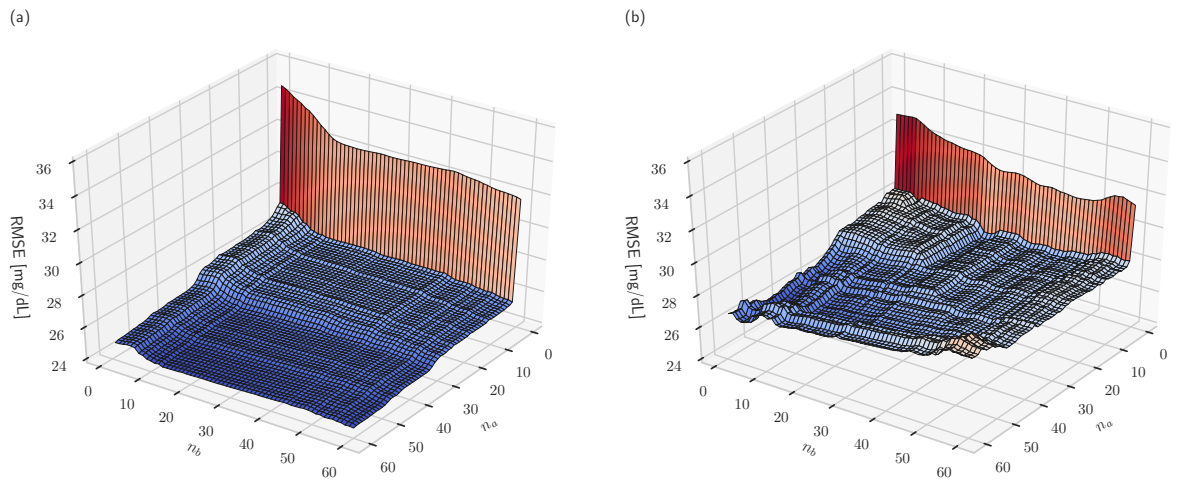

Figura 2: Gráfico de RMSE_{t+6} para entrada u_{bolo}. (a) Conjunto de entrenamiento; (b) Conjunto de prueba

Por lo mismo, se analizarán los dos casos señalados previamente, es decir ARX$(48, 6)$ para u_{meal} y ARX$(48, 4)$ para u_{bolo}.

3.4. Resultados

3.4.1. Modelo ARX$(4, 6)$ para u_{meal}

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. Cada gráfico tiene un indicador de cuando ocurrió un evento de ingesta de comida.

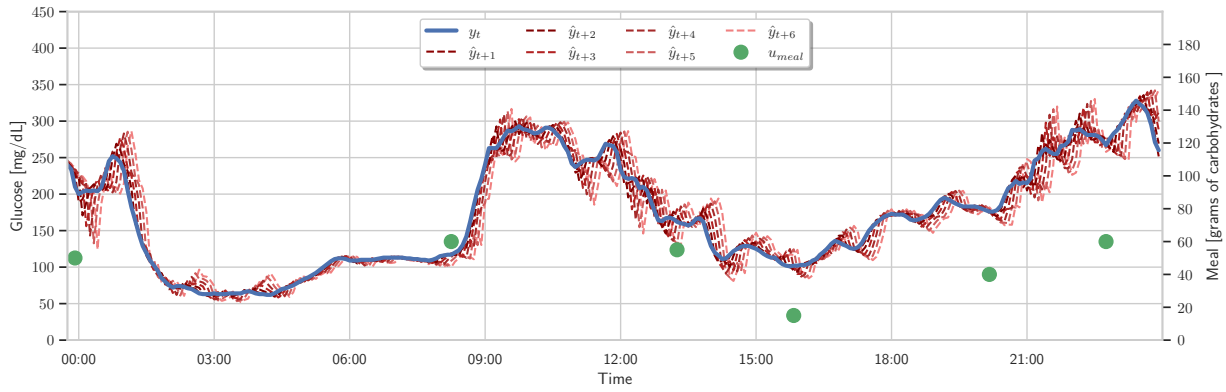

Figura 3: Gráfico de predicción de glucosa para todas las predicciones para ARX$(48, 6)$

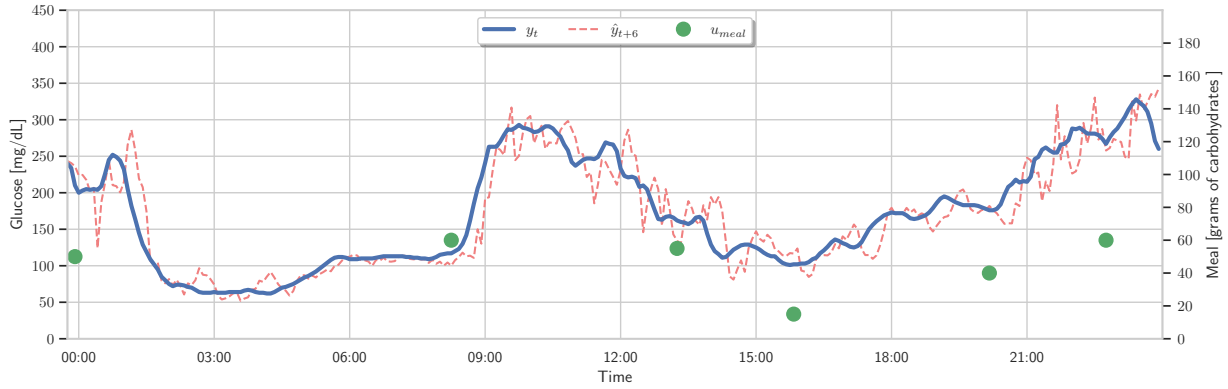

Figura 4: Gráfico de predicción de glucosa para seis pasos adelante para ARX$(48, 6)$

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.

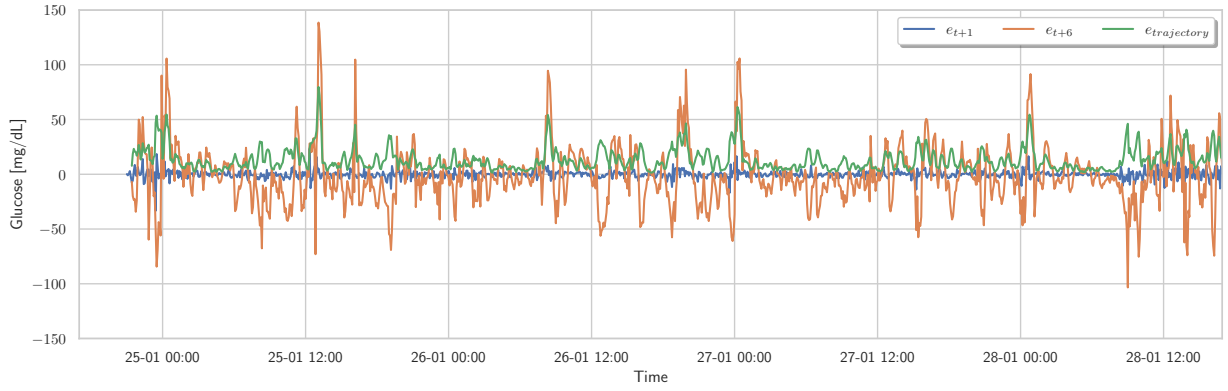

Figura 5: Gráfico del error en función del tiempo para el conjunto de entrenamiento para ARX$(48, 6)$

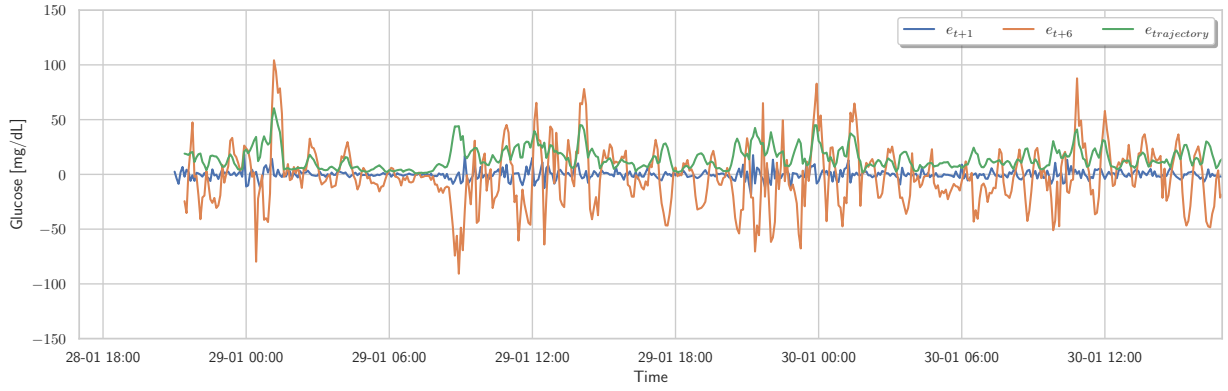

Figura 6: Gráfico del error en función del tiempo para el conjunto de prueba para ARX$(48, 6)$

En la tabla 2 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.

	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	1104	1099	1099	528	523	523
Media	-0.11	-1.67	13.23	-0.14	-2.09	14.69
Desviación estándar	3.61	26.57	10.3	4.24	27.02	9.92
Mínimo	-33.37	-103.35	1.42	-18.78	-90.75	1
25 %	-1.89	-15.88	5.99	-2.18	-18.85	7.16
50 %	-0.11	-3.48	10.1	-0.39	-2.31	12.09
75 %	1.55	10.63	17.18	1.68	12.27	19.76
Máximo	23.29	138.5	79.77	18.24	104.23	60.41

Cuadro 2: Resumen estadístico del error

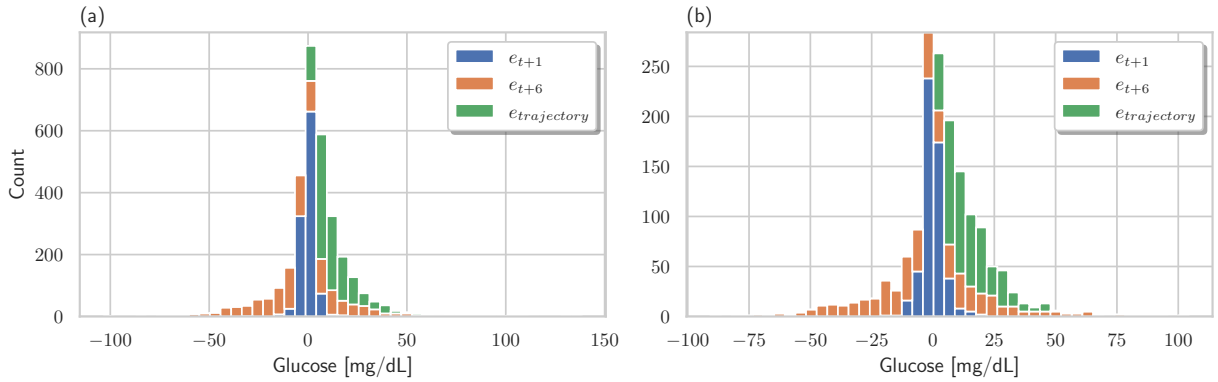

Figura 7: (a) Histograma para el conjunto de entrenamiento; (b) Histograma para el conjunto de prueba

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 $\gamma = N/2 - 1$.

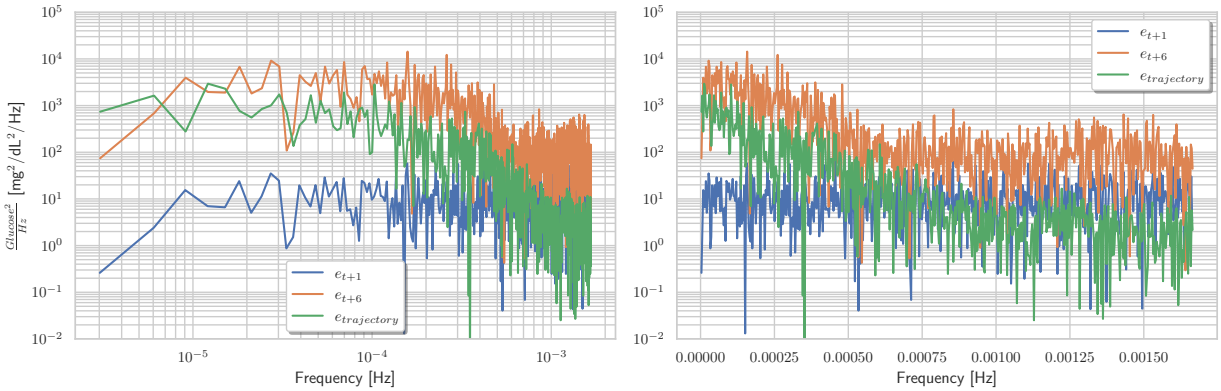

Figura 8: Gráfico del periodograma para el error del conjunto de entrenamiento para ARX$(48, 6)$

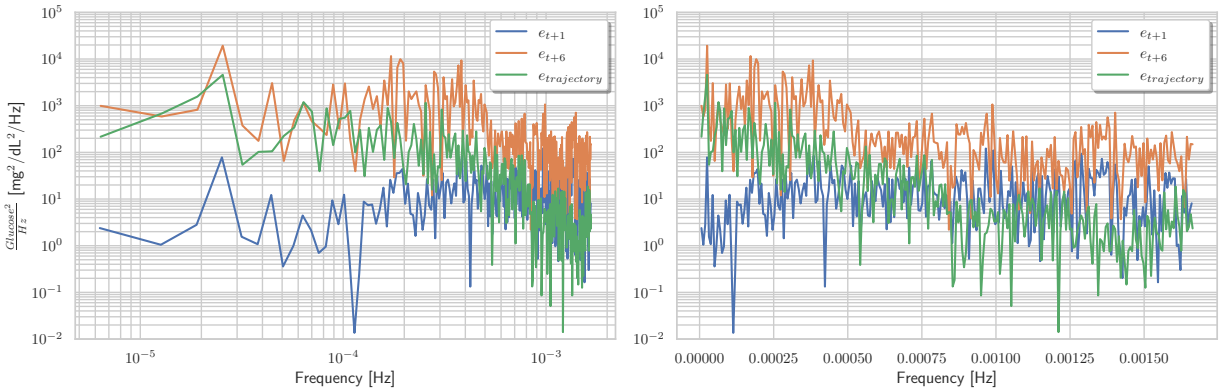

Figura 9: Gráfico del periodograma para el error del conjunto de prueba para ARX$(48, 6)$

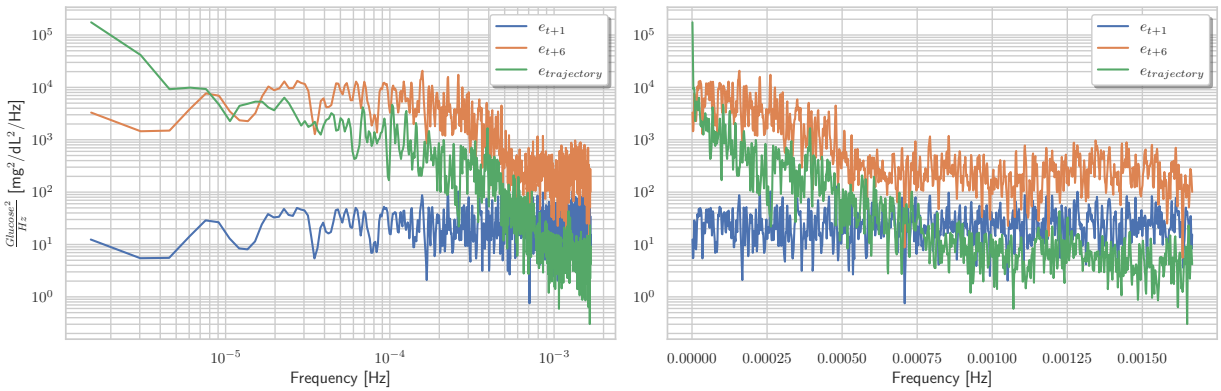

Figura 10: Gráfico de la estimación del espectro para el error del conjunto de entrenamiento para ARX$(48, 6)$

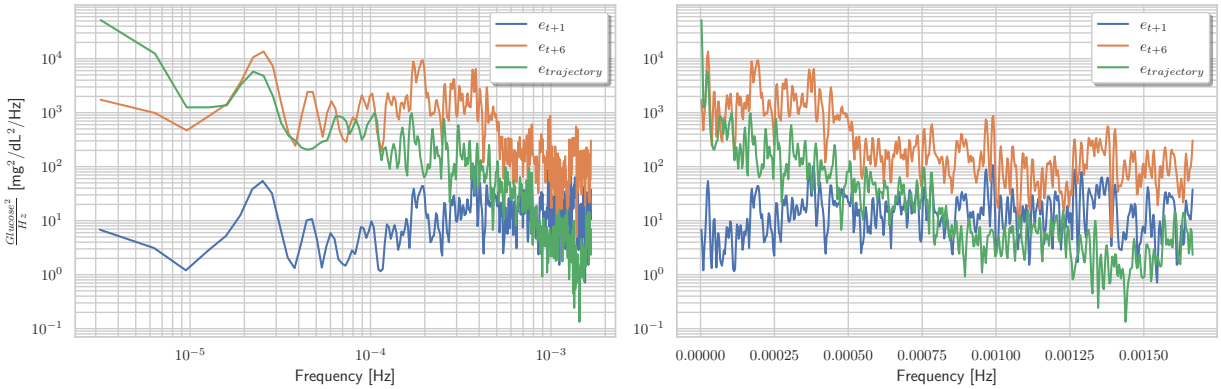

Figura 11: Gráfico de la estimación del espectro para el error del conjunto de prueba para ARX$(48, 6)$

Métricas

Los resultados de las métricas obtenidas bajo este método se muestran en la tabla 3. Hasta el momento, no se ha calculado TG ni ESOD-n, dado que falta estudiar a profundidad su utilidad y si es útil definirlas para la predicción de un paso adelante o la trayectoria total.

	Entrenamiento			Prueba		
	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
RMSE	3.61	26.61	16.76	4.24	27.01	17.73
TG		x			x	
ESOD-n		x			x	

Cuadro 3: Resumen de métricas para ARX$(48, 6)$

3.4.2. Modelo ARX$(4, 4)$ para u_{bolo}

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. Cada gráfico tiene un indicador de cuando ocurrió un evento de inyección de insulina.

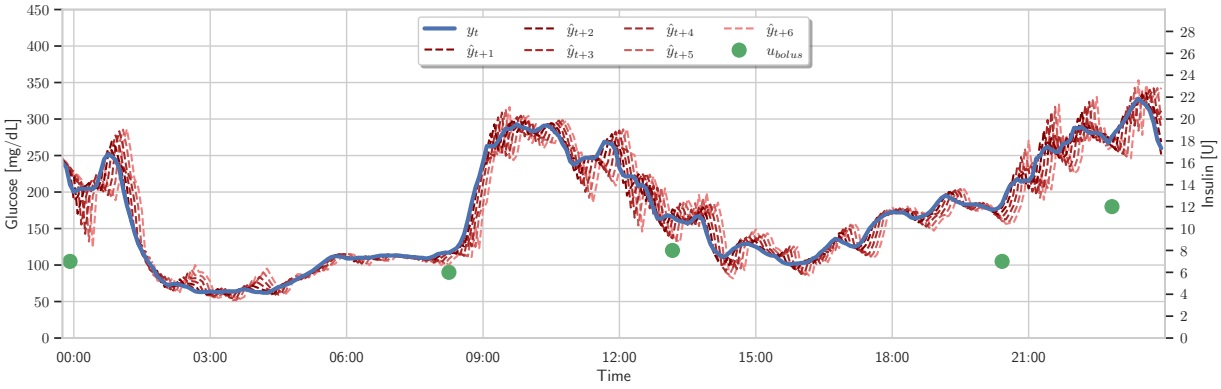

Figura 12: Gráfico de predicción de glucosa para todas las predicciones para ARX$(48, 4)$

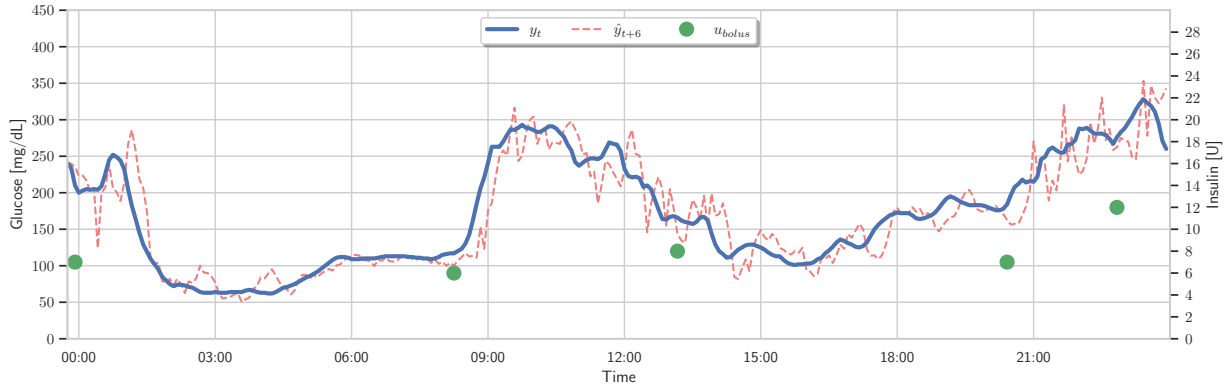

Figura 13: Gráfico de predicción de glucosa para seis pasos adelante para ARX$(48, 4)$

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.

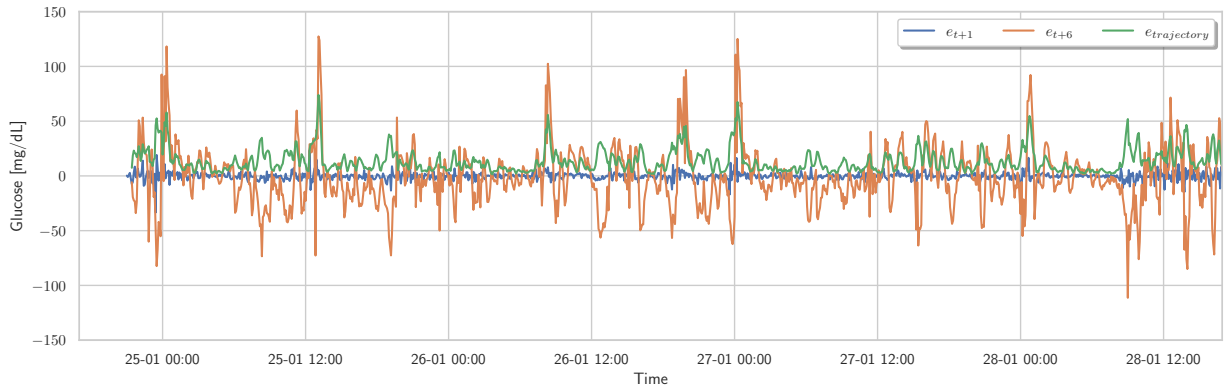

Figura 14: Gráfico del error en función del tiempo para el conjunto de entrenamiento para ARX$(48, 4)$

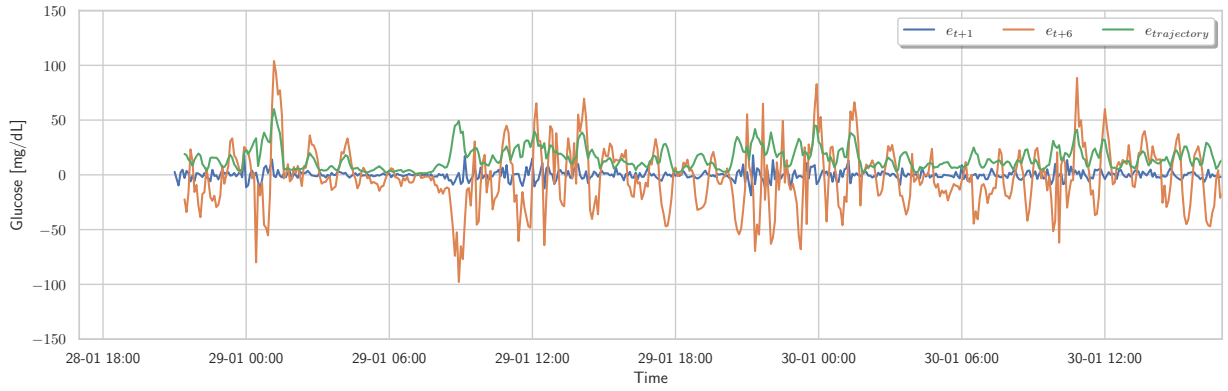

Figura 15: Gráfico del error en función del tiempo para el conjunto de prueba para ARX$(48, 4)$

En la tabla 4 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. En comparación con el las pruebas con entrada de ingestas de comida, este modelo presenta mayor dispersión de los valores, aunque la diferencia es bastante pequeña.

	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	1104	1099	1099	528	523	523
Media	-0.1	-1.73	13.47	-0.14	-2.07	14.98
Desviación estándar	3.63	27.12	10.45	4.26	27.053	10
Mínimo	-33.45	-111.31	1.42	-18.66	-97.86	1.01
25 %	-1.92	-15.93	6.05	-2.31	-18.15	7.41
50 %	-0.04	-3.04	10.3	-0.366	-3.04	12.61
75 %	1.55	11.52	17.72	1.79	13.14	20.04
Máximo	23.34	127.39	73.96	18.15	104.02	60.02

Cuadro 4: Resumen estadístico del error

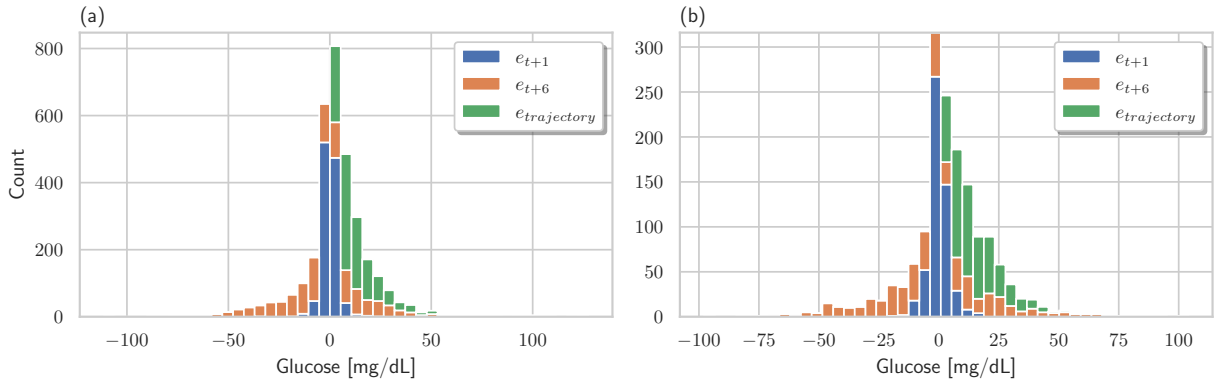

Figura 16: (a) Histograma para el conjunto de entrenamiento; (b) Histograma para el conjunto de prueba

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 $\gamma = N/2 - 1$.

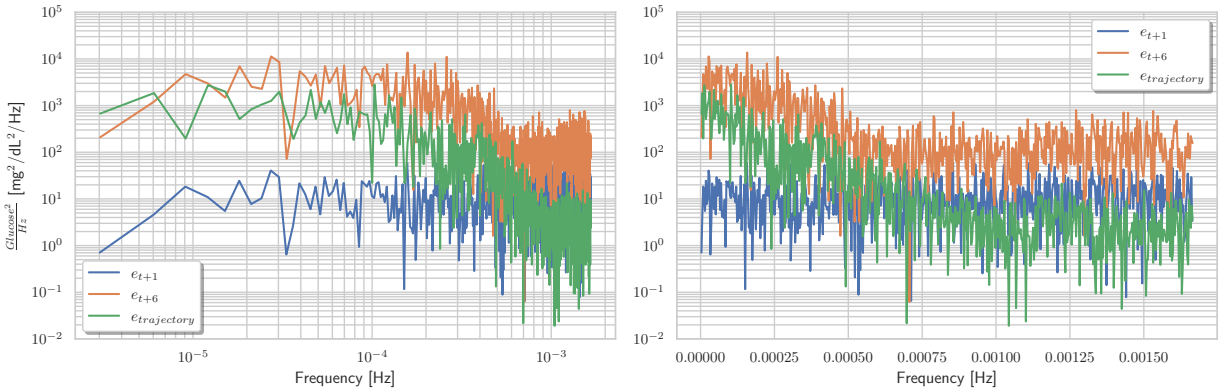

Figura 17: Gráfico del periodograma para el error del conjunto de entrenamiento para ARX$(48, 4)$

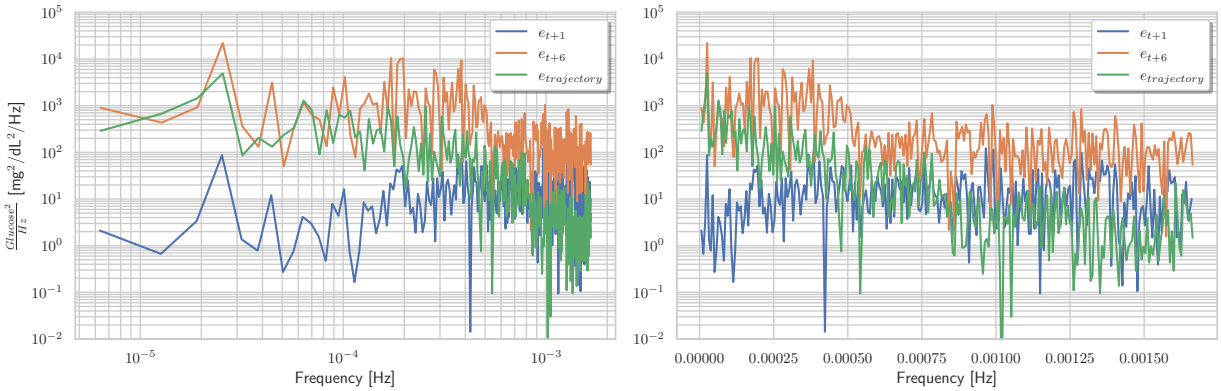

Figura 18: Gráfico del periodograma para el error del conjunto de prueba para ARX$(48, 4)$

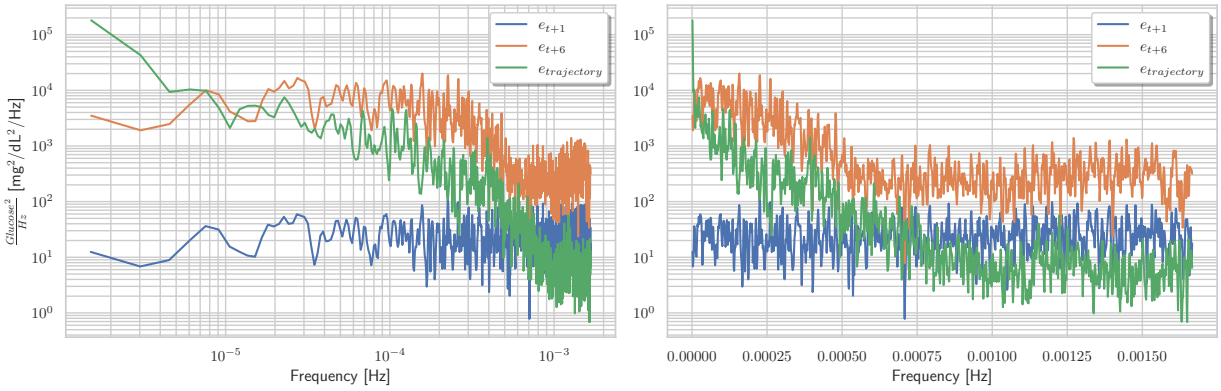

Figura 19: Gráfico de la estimación del espectro para el error del conjunto de entrenamiento para ARX$(48, 4)$

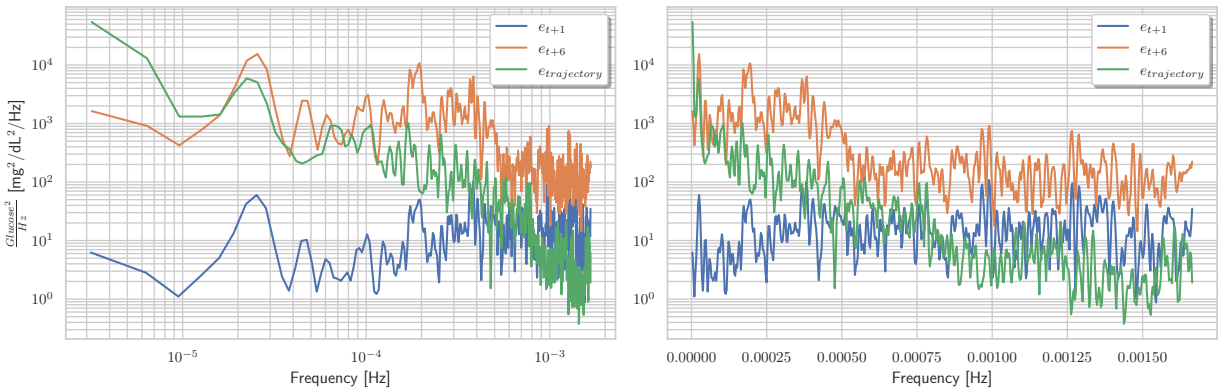

Figura 20: Gráfico de la estimación del espectro para el error del conjunto de prueba para ARX$(48, 4)$

Métricas

Los resultados de las métricas obtenidas bajo este método se muestran en la tabla 5. Hasta el momento, no se ha calculado TG ni ESOD-n, dado que falta estudiar a profundidad su utilidad y si es útil definirlas para la predicción de un paso adelante o la trayectoria total.

	Entrenamiento			Prueba		
	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$	ϵ_{t+1}	ϵ_{t+6}	$\epsilon_{trajectory}$
RMSE	3.63	27.17	17.04	4.26	27.58	18.01
TG		x			x	
ESOD-n		x			x	

Cuadro 5: Resumen de métricas para ARX$(48, 4)$