Procesos gaussianos: metodología y aplicaciones
Taller de Tesis I - Entrega 3
Maestría en Explotación de Datos y Descubrimiento del Conocimiento – UBA 2023 cuat. 1

G. Sebastián Pedersen

sebasped@gmail.com

Última actualización: 5 de junio de 2023

Resumen

En este trabajo se estudian a los procesos gaussianos y su aplicabilidad para predecir series de tiempo, en particular series de tiempo de producción de hidrocarburos. Los procesos gaussianos son un método no paramétro probabilístico de regresión. Es este trabajo por un lado se indaga metodológicamente a los procesos gaussianos, explicitándose sus virtudes y limitaciones teóricas (modelado por gaussianas, incertidumbre natural, etc.). Y por otro lado se aplican los procesos gaussianos para predecir la producción mensual de hidrocarburos en la Argentina, encontrándose las dificultades típicas del caso (diseño del kernel, elección y optimización de los parámetros, etc.). Los resultados de este trabajo, si bien con las dificultades antes mencionadas, por lo menos hacen que la técnica merezca más tiempo de estudio y experimentación, pues a primer análisis y experimentación predice dentro de límites razonables (o, si no es el caso hay indicios suficientes para suponer que podría hacerlo).

Índice

1. Introducción

En este trabajo se estudian a los procesos gaussianos tanto desde un punto de vista teórico, asi como desde un punto de vista práctico mediante una aplicación en concreto. Los procesos gaussianos son un método no paramétro probabilístico de regresión. Los objetivos son dos. Por un lado hay un objetivo metodológico sobre los procesos gaussianos que involucra estudiarlos, entenderlos y ahondar en ellos. Por otro lado se investiga las bondades de los procesos gaussianos para realizar predicciones sobre series temporales, tomando como ejemplo práctico datos de producción mensual de petróleo y gas desde 2006 publicada por la Secretaría de Energía de la Nación. Los datos son abiertos y accesibles desde:

`https://datos.gob.ar/dataset/energia-produccion-petroleo-gas-por-pozo-capitulo-iv`

Como referencia general al problema de forescasting (predicción sobre series temporales) se puede consultar [5]. Como referencias generales a procesos gaussianos se pueden consultar [1], [2], [3] y [4]. Como referencia particular de procesos gaussianos aplicados a la predicción de producción de petróleo y gas se puede consultar [6].

2. Marco teórico: procesos gaussianos

2.1. Trabajos previos

- Como trabajos previos específicos: hacer referencia a [6] y por las dudas buscar otros.
- Como referencias relevantes generales: buscar las partes correspondientes de [1], [3], [4], [2].

2.2. Qué es un proceso gaussiano ($\mathcal{GP}$)

2.2.1. Idea y relación con la regresión clásica

El problema de regresión clásica busca encontrar una $f(x) = y$ que mejor ajuste a un conjunto de datos $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^2$. La f se propone parametrizada por una cantidad fija de parámetros (sería la elección del modelo) y el "mejor ajuste" viene dado por encontrar los parámetros que minimizan una función de costo o función objetivo (típicamente el error cuadrático) entre la f y los datos $\mathcal{D}$. Por ejemplo en el problema de regresión lineal clásico la cantidad de parámetros está dada por la cantidad de covariables de los datos a utilizar (incluyendo un posible feature engineering), y la f se propone como una función lineal en los parámetros (podría ser no lineal en los datos).

Los modelos que utilizan una cantidad fija de parámetros se llaman *paramétricos*. En contraste a éstos existe los modelos *no paramétricos* cuya cantidad de parámetros no es fija (por ejemplo KNN o SVM). Por supuesto siempre hablando de parámetros del modelo a aprender. También existen los hiperparámetros del modelo que sí pueden ser fijos, independientemente de si el modelo es paramétrico o no. Por ejemplo en KNN un hiperparámetro del modelo sería el radio elegido para considerar a un punto vecino de otro.

Los procesos gaussianos ($\mathcal{GP}$) son un método no paramétrico probabilístico de regresión. Este modelo busca encontrar ya no una única f que mejor ajusta a los datos, sino una distribución sobre funciones $P(f)$ que mejor ajusten a los datos, donde $f : \mathcal{X} \to \mathbb{R}$ es una función del espacio de entrada $\mathcal{X}$ (espacio en donde vivirán los datos $\mathcal{D}$). Como su nombre ya lo indica en los $\mathcal{GP}$ la distribución de las f será gaussiana. Una de los principales benificios de los $\mathcal{GP}$ es que se obtiene una cuantificación sobre la incertidumbre del ajuste, dada naturalmente por la distribución $P(f)$.

2.2.2. Definición formal de proceso gaussiano ($\mathcal{GP}$)

Dado que el espacio de entrada $\mathcal{X}$ de la $f : \mathcal{X} \to \mathbb{R}$ bien puede ser infinito, se puede pensar a la f como un vector infinito dimensional. Como en la práctica solamente se pueden trabajar con finitos valores, podemos pensar que dados $\{x_i\}_{i=1}^n$ la f es un vector evaluando en cada x_i, es decir $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))$, y en donde la $f(\mathbf{x})$ tendrá distribución gaussiana. Más precisamente:

Definición (proceso gaussiano): un proceso gaussiano ($\mathcal{GP}$) es una colección de variables aleatorias tal que cualquier subconjunto finito tiene distribución (conjunta) gaussiana.

Para entender esta definición, apliquémosla al problema de regresión formulado anteriormente en 2.2.1. Será entonces $P(f)$ un $\mathcal{GP}$ si para cualquier subconjunto $\{x_i\}_{i=1}^n \subset \mathcal{X} = \mathbb{R}$ la distribución $P(f(\mathbf{x}))$ de $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))$ es una gaussiana multivariada. Las variables aleatorias representan, en este caso, los valores $f(x_i)$ para cada x_i.

Como cualquier gaussiana multivariada queda completamente caracterizada por su media y su covarianza, para un $\mathcal{GP}$ definimos sus funciones media y covarianza como:

$$m(x) = \mathbb{E}\{f(x)\} \tag{1}$$
$$K(x_1, x_2) = \mathbb{E}\{[f(x_1) - m(x_1)]\,[f(x_2) - m(x_2)]\} \tag{2}$$

De esta forma podemos escribir al $\mathcal{GP}$ y a su marginal como:

$$f \sim \mathcal{GP}(m, K) \tag{3}$$
$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})) \tag{4}$$

Como esta marginalización a una gaussiana debe cumplirse para cualquier subconjunto finito de variables aleatorias, y el conjunto $\mathcal{X}$ original bien puede ser infinito, se pide que el $\mathcal{GP}$ cumpla con esta propiedad de marginalización (llamada propiedad de consistencia). Notemos que esta propiedad se cumple para la función de covarianza definida en la ecuación 2.

Elecciones de funciones de media y covarianza: es normal asumir que $m = 0$ pues siempre se pueden referenciar los datos al promedio. Para la función de covarianza K, también llamada kernel, hay que pedir que sea semi definida positiva (para cada subconjunto finito). La elección del kernel en general dependerá del problema en particular que se desee estudiar. Una elección popular es el kernel *exponencial cuadrático* (square exponential), también conocido como radial basis function (RBF) o kernel gaussiano:

$$K_{SE}(x_1, x_2) = \sigma^2 \exp\left(-\frac{(x_1 - x_2)^2}{2l^2}\right) \tag{5}$$

donde σ y l son parámetros interpretables.

Observación: destacamos que los parámetros de los kernels son relevantes a la hora de predecir, y que en general se estiman a partir de los datos de entrenamiento.

Comentario para mi: kernels hay muchos y su elección está relacionada con la naturaleza de los datos. Además sus parámetros se entrenan por máxima verosimilitud o MCMC (y deben ser entrenados para que el $\mathcal{GP}$ prediga mejor. Se puede ejemplificar la influencia de los parámetros en la predicción). Queda en el tintero comentar estas cosas.

2.3. Muestreo de un $\mathcal{GP}$

2.3.1. Muestreo *prior*

Mirando la definición de la sección 2.2.2 podemos entonces pensar que un $\mathcal{GP}$ define un *prior* sobre funciones, y por lo tanto ya podemos realizar muestreos para cada subconjunto finito $\{x_i\}_{i=1}^n \subset \mathcal{X} = \mathbb{R}$, previa elección de la forma funcional de la covarianza K, pues tendremos que:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

donde $\mathbf{x} = (x_1, \ldots, x_n)$, $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))$ y $K(x_i, x_j)$ viene dado por K_{SE} de la ecuación 5. Asumiendo media igual a cero, tomando $\sigma = l = 1$ para los parámetros del K_{SE}, y muestreando tres veces para 100 valores de x equiespaciados entre 0 y 10, obtenemos 3 muestras del correspondiente *prior* del $\mathcal{GP}$ como se puede ver en la figura 1. El intervalo de confianza corresponde a moverse 1,96 devíos estándares para cada valor $f(x_i)$.

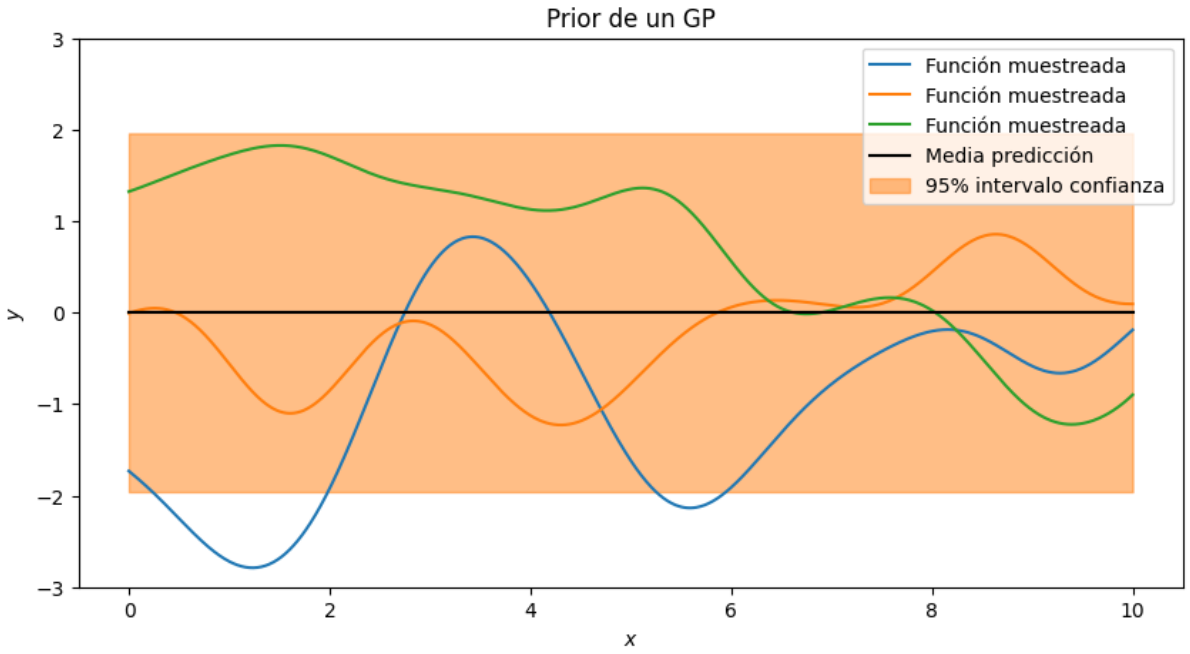

Figura 1. Prior de un $\mathcal{GP}$. A modo ilustrativo se grafican 3 muestras del prior del $\mathcal{GP}$.

2.3.2. Predicción *posterior* sin ruido

Supongamos que tenemos observaciones $(X, f(X))$ donde $X = (x_1, \ldots, x_n)$ y queremos realizar una predicción sobre un conjunto de valores X^* de tamaño n^* (X y X^* pueden o no compartir valores). De la definición de $\mathcal{GP}$ se desprende inmediatamente que la distribución conjunta es:

$$\begin{bmatrix} f(X) \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(X^*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix} \right) \tag{6}$$

Como para realizar la predicción necesitamos la distribución de $f(X^*)$ (condicional a los datos conocidos, por supuesto), nos valeremos del siguiente resultado:

Proposición: dado un prior $\mathcal{GP}$, la posterior también es un $\mathcal{GP}$ y se puede condicionar sobre los datos conocidos para obtener:

$$f(X^*)|f(X), X \sim \mathcal{N}(m_{X^*|X}, \Sigma_{X^*|X}) \tag{7}$$

Donde la media y covarianza son:

$$m_{X^*|X} = m(X^*) + K(X^*,X)K^{-1}(X,X)(f(X) - m(X)) \tag{8}$$

$$\Sigma_{X^*|X} = K(X^*,X^*) - K(X^*,X)K^{-1}(X,X)K(X,X^*) \tag{9}$$

Observemos que tanto la media como la covarianza de la distribución de $f(X^*)$ dependen tanto de las observaciones conocidas (que juegan el papel de datos de entrenamiento) como de los valores de X^* en donde deseamos realizar las predicciones $f(X^*)$.

Veamos un ejemplo de esto en acción. Supongamos nuevamente media cero y kernel SE con $\sigma = l = 1$. Tomemos a X como 6 datos entre 0 y 10, y las $f(X)$ dadas por $f(x) = x\sin(x)$. Tomemos ahora para predecir a X^* como 100 datos equiespaciados entre 0 y 10. Realizamos 3 muestreos de la $f(X^*)|f(X), X$ que se pueden ver en la figura 2.

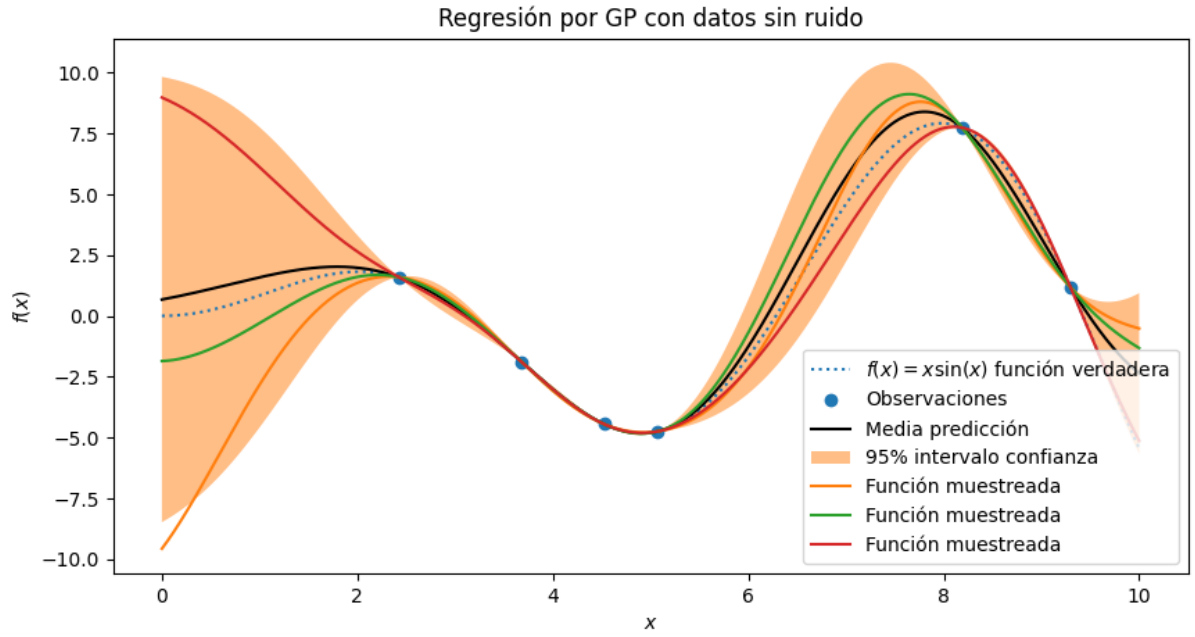

Figura 2. Regresión por $\mathcal{GP}$ con datos sin ruido. A modo ilustrativo se grafican 3 muestras del posterior del $\mathcal{GP}$.

Observemos que la predicción puede realizarse tanto interpolando como extrapolando a los datos conocidos. Además el intervalo de confianza es más angosto cuando los datos conocidos están menos distantes.

2.3.3. Predicción *posterior* con ruido

En aplicaciones reales es normal tener mediciones con ruido. Supongamos que tenemos observaciones $(X, f(X))$ donde $X = (x_1, \ldots, x_n)$, pero ahora nuestras observaciones tiene ruido $y_i = f(x_i) + \eta$, donde suponemos que $\eta \sim \mathcal{N}(0, \sigma_n^2)$ y que el ruido es i.i.d. Por lo tanto ahora nuestro conjunto de datos es de la forma (X, Y) donde $Y = f(X) + \eta$.

Es fácil ver que este ruido equivale, en nuestro modelo, a agregar un término diagonal a la covarianza:

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Nuevamente queremos predecir sobre un conjunto X^*, y en este caso la conjunta queda como:

$$\begin{bmatrix} Y \\ f(X^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(X^*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbb{I} & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \tag{10}$$

Al igual que en el caso sin ruido tenemos un resultado sobre la distribución condicional que nos permite realizar predicciones:

Proposición:

$$f(X^*)|Y, X \sim \mathcal{N}(m_{X^*|X}, \Sigma_{X^*|X}) \tag{11}$$

Donde la media y covarianza son:

$$m_{X^*|X} = m(X^*) + K(X^*, X)\left[K(X,X) + \sigma_n^2 \mathbb{I}\right]^{-1}(Y - m(X)) \tag{12}$$

$$\Sigma_{X^*|X} = K(X^*, X^*) - K(X^*, X)\left[K(X,X) + \sigma_n^2 \mathbb{I}\right]^{-1} K(X, X^*) \tag{13}$$

En la figura 3 se puede ver este caso en acción, tomando los mismos parámetros y datos que para el ejemplo sin ruido, y tomando un ruido gaussiano de varianza 1.

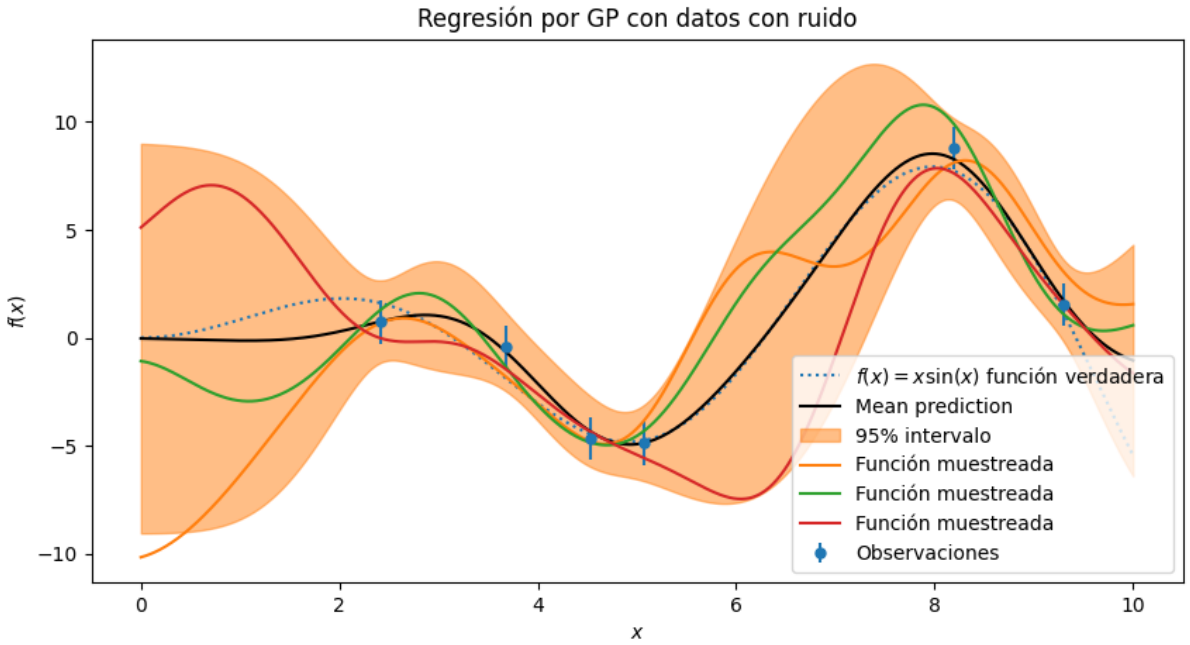

Figura 3. Regresión por $\mathcal{GP}$ con datos *con* ruido. A modo ilustrativo se grafican 3 muestras del posterior del $\mathcal{GP}$.

Si comparamos la figura 3 con la figura 2 observamos que ahora que los datos tienen ruido, la media de la gaussiana ya no necesariamente interpola a las observaciones (léase no interpola la esperanza de las observaciones).

3. Metodología

3.1. Datos: descripción, prepocesamiento y análisis exploratorio

Como mencionamos en la sección 1 los datos son de producción de petróleo y gas mensual desde 2006 publicada por la Secretaría de Energía de la Nación, abiertos y accesibles desde:

`https://datos.gob.ar/dataset/energia-produccion-petroleo-gas-por-pozo-capitulo-iv`

Estos datos tienen información mensual discriminada por pozo, a saber: producción de petróleo, gas y agua; inyección de agua, gas, CO_2 u otro; yacimiento, cuenca, concesión y provincia a la que pertenece; además de datos sobre la empresa operadora del pozo. En total son 971.882 datos.

Como primera medida se procedió a explorar la calidad de los datos, detectándose casos en donde una o más covariables tenían datos faltantes y esto derivaba en una mala carga del dato pues el faltante siempre ocurría en las covariables finales y, observando los demás datos de buena calidad, se podía deducir que había ocurro un "corrimiento" hacia las covariables primeras. Es decir, por poner un ejemplo, si el dato constaba de $x_1, \ldots, x_9$ y x_9 tenía dato faltante, mirando datos de buena calidad, se podía deducir que el faltante en realidad estaba entrea las $x_1, \ldots, x_8$ y no en la x_9. Este tipo de datos se descartaron, siendo su cantidad no significativa sobre el total, por la imposibilidad en muchos casos de identificar adecuadamente a cuál o cuáles covariables pertenecía el faltante.

Como uno de los objetivos de este trabajo será realizar predicciones sobre series temporales de la producción de petróleo y gas mensual, pero a primer acercamiento sin discriminar por pozo (ni por cuenca, provincia, etc.) entonces se procedió a sumar para cada mes la producción de todos los pozos, obteniéndose la producción mensual total de petróleo y gas. Cabe aclarar que un pozo puede, y en general lo hace, producir al mismo tiempo tanto petróleo como gas.

Como los procesos gaussianos que se desean usar en este trabajo son alimentados por series de datos ordenados, por ejemplo por el tiempo en el caso de producción de petróleo y gas, como ya se ejemplificó

en la sección 2.3, posterior a la obtención de la producción total por mes se procedió a inspeccionar los datos para corroborar que los mismos eran adecuados para usarse con la técnica antes mencionada. Las figuras 4 y 5 muestran la producción mensual desde Enero de 2006 hasta Marzo de 2023 de petróleo (en m^3) y de gas (en miles de m^3), comprobando que efectivamente estos datos sirven para alimentar a los procesos gaussianos descriptos en la sección 2.2 y ejemplificados en la sección 2.3.

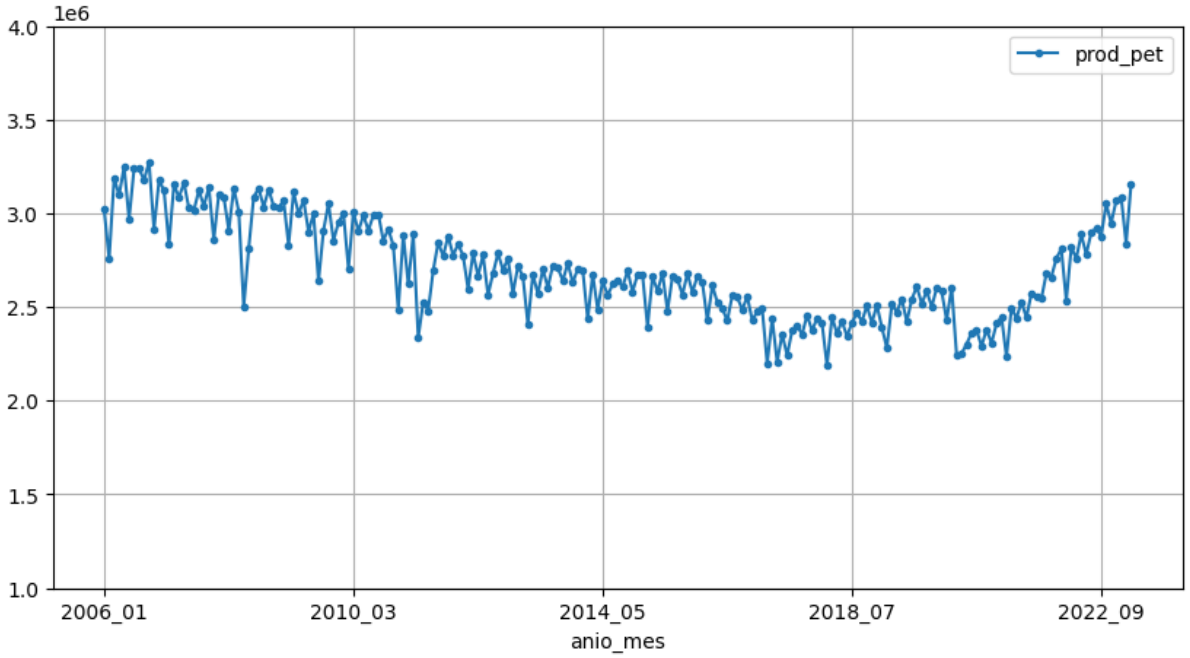

Figura 4. Producción total mensual de petróleo en m^3 en Argentina, desde Enero-2006 hasta Marzo-2023 según la Secretaría de Energía de la Nación.

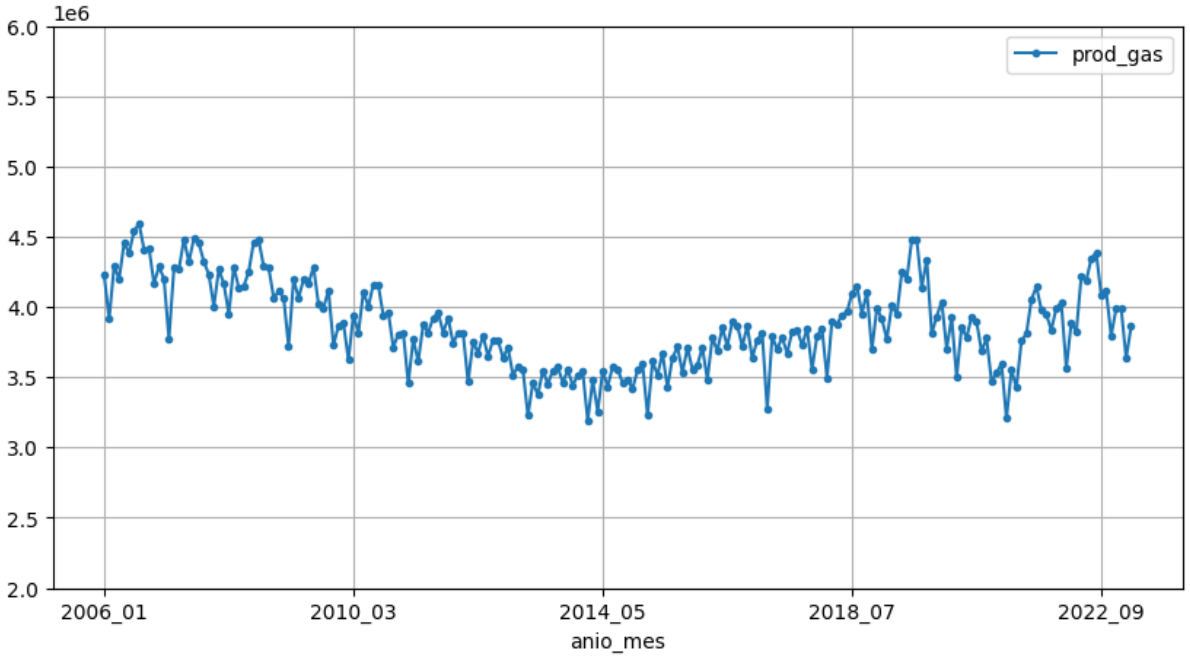

Figura 5. Producción total mensual de gas en miles de m^3 en Argentina, desde Enero-2006 hasta Marzo-2023 según la Secretaría de Energía de la Nación.

3.2. Técnicas de análisis y modelado

- Contar brevemente de nuevo el modelado de los datos, tanto de entrenamiento como de testeo, por una gaussiana multivariada, y contar brevemente de nuevo la técnica para predecir con $\mathcal{GP}$s descripta en 2.3.
- Hablar sobre la medida de incertidumbre dada naturalmente por el $\mathcal{GP}$: la diagonal de la covarianza.
- Describir el o los diseños del kernel para esta serie de tiempo particular: término de tendencia a largo plazo, término de estacionalidad, término de irregularidades a corto plazo, ¿término de ruido?, ¿otros?. Hay un ejemplo interesante de esto en `https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_co2.html#design-the-proper-kernel`.

  También se puede ver [7] y `https://www.cs.toronto.edu/~duvenaud/cookbook/`.

  Cada término es un kernel, y la suma de kernels es un kernel.
- ¿Acá o en Marco Teórico? Describir la influencia de los parámetros de cada kernel. Hablar de la optimización de los parámetros a partir de los datos (hay una explicación linda de esto en [2]). Optimización de los parámetros: máxima verosimilitud y/o MCMC.

4. Resultados y discusión

- Poner el resultado mínimo previo utilizando un kernel RBF.
- Si da el tiempo: probar con distintos kernels (leáse suma de) y con distintas ventanas de tiempo hacia atrás para predecir. Intuición: la ventana determina en cierta medida el kernel. Implicancias de la intuición: posible rediseño del kernel cada vez que se desea predecir.

  Se puede leer del cap. 2 de [7]. Una explicación más informal de lo mismo está en `https://www.cs.toronto.edu/~duvenaud/cookbook/`.
- Limitación: modelado gaussiano, diseño del kernel y optimización de sus parámetros.
- Limitación: los $\mathcal{GP}$ con poco eficientes en altas dimensiones. En algún lado leí que ya se complica si los features sobrepasan un par de decenas.
- Mejora: usar más features y no solamente el tiempo. Investigar sobre esto: por ahora es una inquietud más bien en el aire.
- Mejora: usar procesos gaussianos multi output. Intuición de esto: hacer uso de la correlación entre la producción de petróleo y de gas.
- Mejora: discriminar por zonas. Intuición de esto: cada zona tiene sus propios tiempos y volúmenes de producción.
- Mejora: discriminar por antigüedad de pozo y/o yacimiento. Intuición de esto: el volumen de producción está correlacionado a la antigüedad.

5. Conclusiones

6. Ideas que todavía no puse en ningún lado

- Relacionar los modelos aditivos de series de tiempo clásicos: tendencia largo plazo + estacionalidad1 + estacionalidad2 + ... + irregularidades + ruido, con la aditividad de kernels. Intuición de esto: cada kernel modela un término del clásico.

  Idea general de esto: el $\mathcal{GP}$ es el análogo probabilístico al modelo clásico.

Referencias

[1] Robert B. Gramacy, *Gaussian process modeling, design and optimization for the applied sciences*,
https://bookdown.org/rbg/surrogates/
2023-03-05.

[2] Felipe Tobar, *Aprendizaje de máquinas. Cap. 8: procesos gaussianos*,
https://raw.githubusercontent.com/GAMES-UChile/Curso-Aprendizaje-de-Maquinas/master/notas_de_clase.pdf
2021.

[3] Carl E. Rasmussen; Christopher K. I. Williams, *Gaussian Processes for Machine Learning*,
http://gaussianprocess.org/gpml/
MIT Press, 2006.

[4] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective. Ch. 15: Gaussian processes*,
http://noiselab.ucsd.edu/ECE228/Murphy_Machine_Learning.pdf
MIT Press, 2012.

[5] Rob J. Hyndman; George Athanasopoulos, *Forecasting: Principles and Practice*,
https://otexts.com/fpp3/
3ra. ed., 2021.

[6] Osaro, Etinosa; Okorie, Vivian; Alornyo, Sonia, *Exploring the Usefulness of Gaussian Process Regression for the Prediction of Oil, Water and Gas Production Rates*,
Abril 2023, 14. 10.35248/2157-7463.23.14.506.
https://www.researchgate.net/profile/Etinosa-Osaro/publication/370099861_Exploring_the_Usefulness_of_Gaussian_Process_Regression_for_the_Prediction_of_Oil_Water_and_Gas_Production_Rates/links/643ed71a39aa471a5248f03a/Exploring-the-Usefulness-of-Gaussian-Process-Regression-for-the-Prediction-of-Oil-Water-and-Gas-pdf

[7] David K. Duvenaud, *Automatic Model Construction with Gaussian Processes*,
https://www.cs.toronto.edu/~duvenaud/thesis.pdf
PhD thesis, 2014, University of Cambridge.