Gaussian Process Regression for Ship Dynamics Fitting using AIS Data

Matteo Sartoni

February 21, 2025

1 Problem Introduction

Ship tracking plays a crucial role in maritime surveillance. Cooperative vessels are equipped with an Automatic Identification System (AIS) sensor, which enables them to transmit information such as their ID, latitude, longitude, speed, and heading.
In cases where AIS sensor data is received at a low frequency, one approach to reconstructing the ship's track is to fit the obtained data using Gaussian Process Regression (GPR).

2 Simulation

The objective of this study is to utilize GPR to reconstruct the temporal evolution of the ship's latitude and longitude, using $1/6$ of the data from the selected dataset.

2.1 Dataset and Ship selection

A real and dynamic AIS dataset has been provided in [1]. This dataset contains AIS data for vessels near Piraeus, in Greece.
The selected dataset is the one that comprises data from January 2019. As a first step, the dataset was reordered in chronological order, and only the first vessel with a speed of at least 7 m/s was considered. Fig. 2.1 and Fig. 2.2 show, respectively, the longitude and latitude of the vessel under consideration over a 5-minute simulation period.

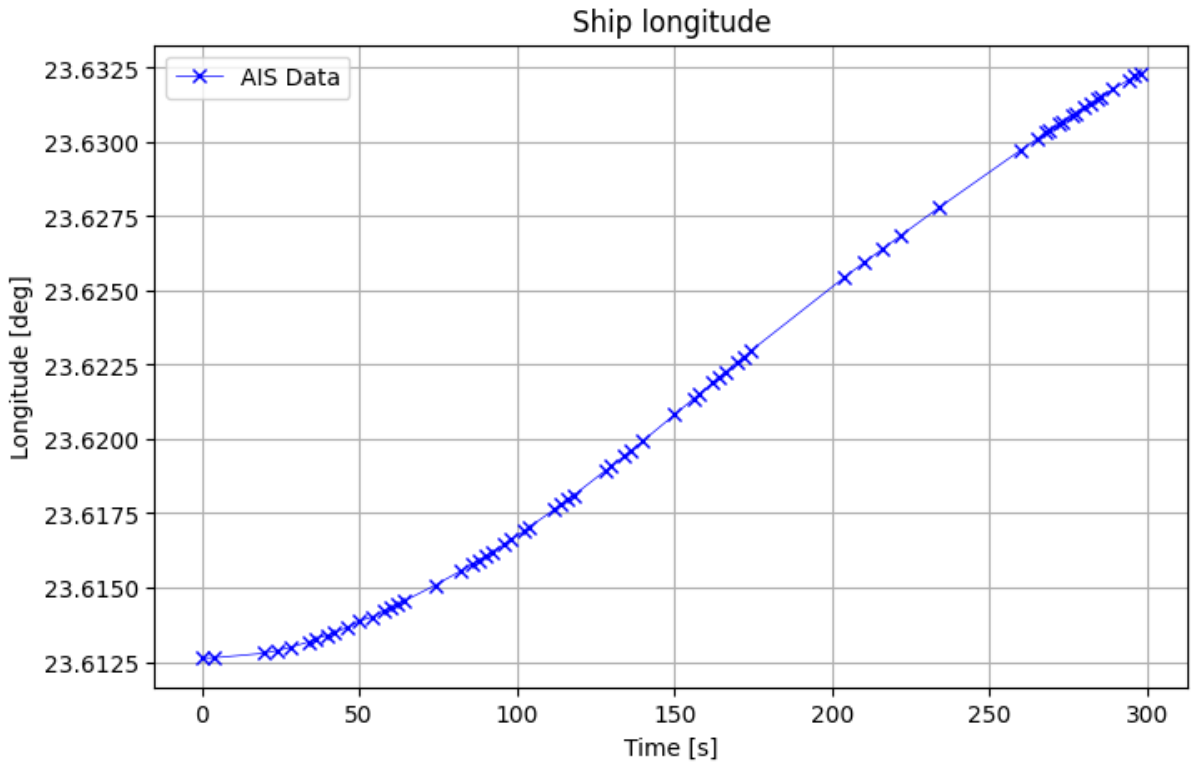

Figure 2.1: Longitude of the considered vessel over 5 minutes.

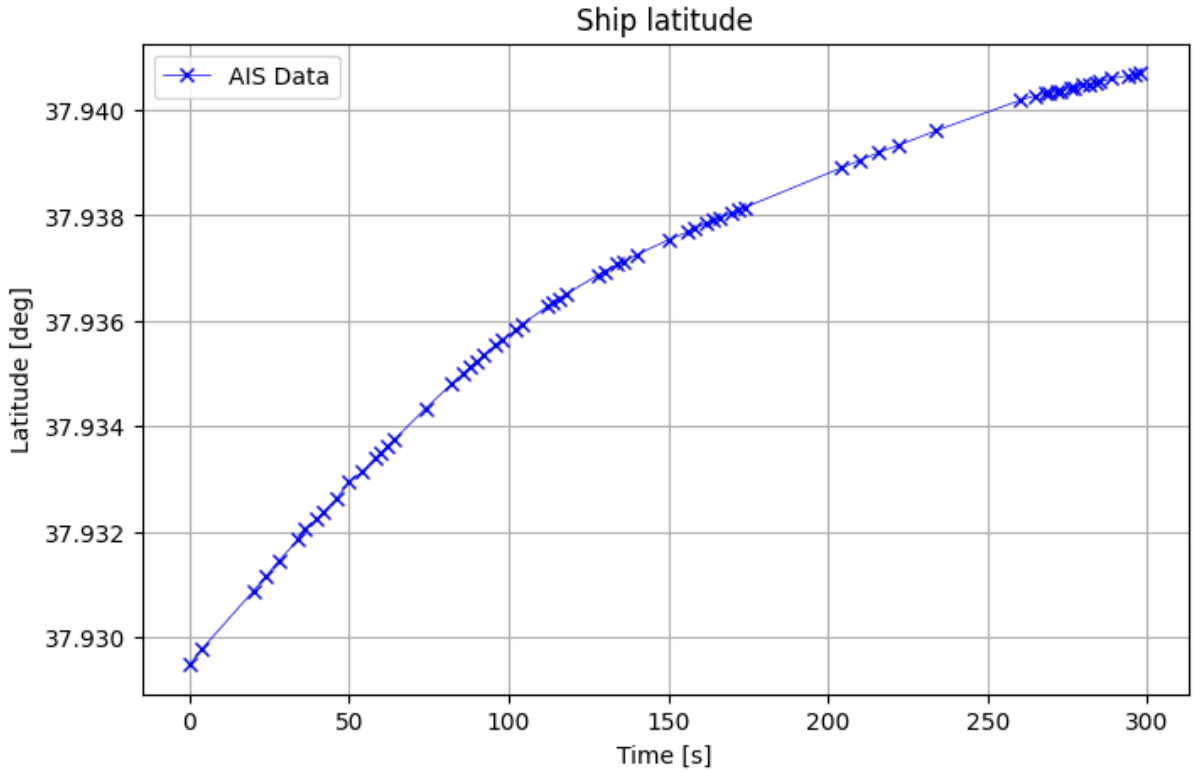

Figure 2.2: Latitude of the considered vessel over 5 minutes.

2.2 Gaussian Process Regression

A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.
A GP is completely defined by its mean function $m(x)$ and covariance function, called also kernel, $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(k), k(x, x')), \tag{2.1}$$

where

$$m(x) = \mathbb{E}[f(x)] \tag{2.2}$$
$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \tag{2.3}$$

The kernel $k(x, x')$ encodes the assumptions about the function's smoothness and other properties.

2.2.1 Gaussian Process Regression with Noisy Data

Consider a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^D$ is the input and $y_i \in \mathbb{R}$ the corresponding output. Let's denote with $X \in \mathbb{R}^{D\times n}$ the input matrix of the training set and with $\mathbf{y} \in \mathbb{R}^n$ the vector containing all the outputs. Therefore we may rewrite the dataset as $\mathcal{D} = \{X, \mathbf{y}\}$
We aim to predict the output $\mathbf{f}^\star \in \mathbb{R}^{n^\star}$ at new inputs $X^\star \in \mathbb{R}^{D\times n^\star}$. The joint distribution of the training output $\mathbf{y}$ and the test output $\mathbf{f}^\star$ is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}^\star \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X, X) + \sigma_n^2 I & K(X, X^\star) \\ K(X^\star, X) & K(X^\star, X^\star) \end{pmatrix}\right), \tag{2.4}$$

where

- $K(X, X) \in \mathbb{R}^{n\times n}$ is the covariance matrix of the training inputs, with $K(X, X)_{[i,j]} = k(x_i, x_j)$.
- $K(X, X^\star) \in \mathbb{R}^{n\times n^\star}$ is the covariance vector between the training inputs and the test input, with $K(X, X^\star)_{[i,j]} = k(x_i, x_j^\star)$.
- $K(X^\star, X^\star) \in \mathbb{R}^{n^\star\times n^\star}$ the covariance of the test input, with $K(X^\star, X^\star)_{[i,j]} = k(x_i^\star, x_j^\star)$..
- σ_n^2 is the variance of the observation noise.

The predictive distribution for $\mathbf{f}^\star$ given the training data is Gaussian

$$\mathbf{f}^\star | X, \mathbf{y}, X^\star \sim \mathcal{N}(\mu^\star, \sigma^{\star 2}), \tag{2.5}$$

with mean and variance

$$\mu^\star = K(X^\star, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \tag{2.6}$$
$$\sigma^{\star 2} = K(X^\star, X^\star) - K(X^\star, X)[K(X, X) + \sigma^{\star 2} I]^{-1}K(X, X^\star). \tag{2.7}$$

2.2.2 Simulation and Results

Kernel selection

The choice of the kernel $k(x, x')$ is crucial in GPR as it defines the properties of the functions we are modeling.
In this simulation, the kernel used is the sum of two kernels, the **Squared Exponential (RBF)** and the **Dot Product** one.
The former is given by

$$k_{se}(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right) \tag{2.8}$$

where σ_f^2 is the signal variance and l is the length's scale, which has been set to 1.
The latter is

$$k_{dp}(x, x') = x \cdot x'. \tag{2.9}$$

This kernel has been used to take into account a general linear trend, with some nonlinear trend incorporated by the RBF kernel.

Result

The training dataset size corresponds to $1/6$ of the AIS data considered, and the values used to train the GPR were randomly selected from the available ones.
Fig. 2.3 and Fig. 2.4 show, respectively, the output of the GPR for longitude and latitude data fitting.

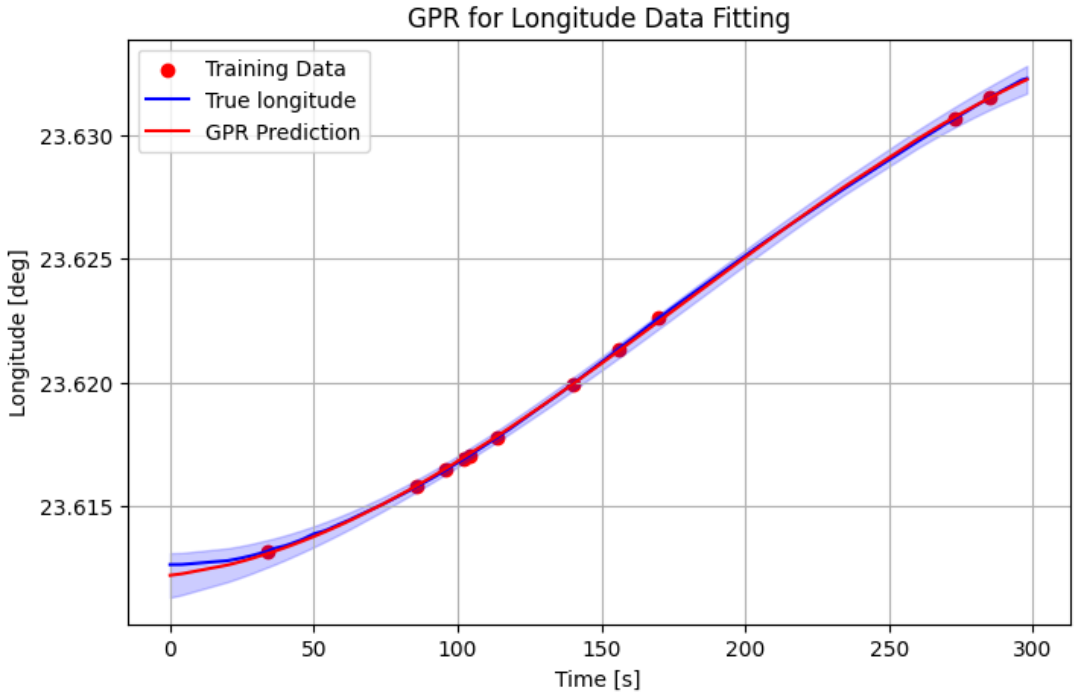

Figure 2.3: Longitude fitted through GPR for the considered vessel over 5 minutes.

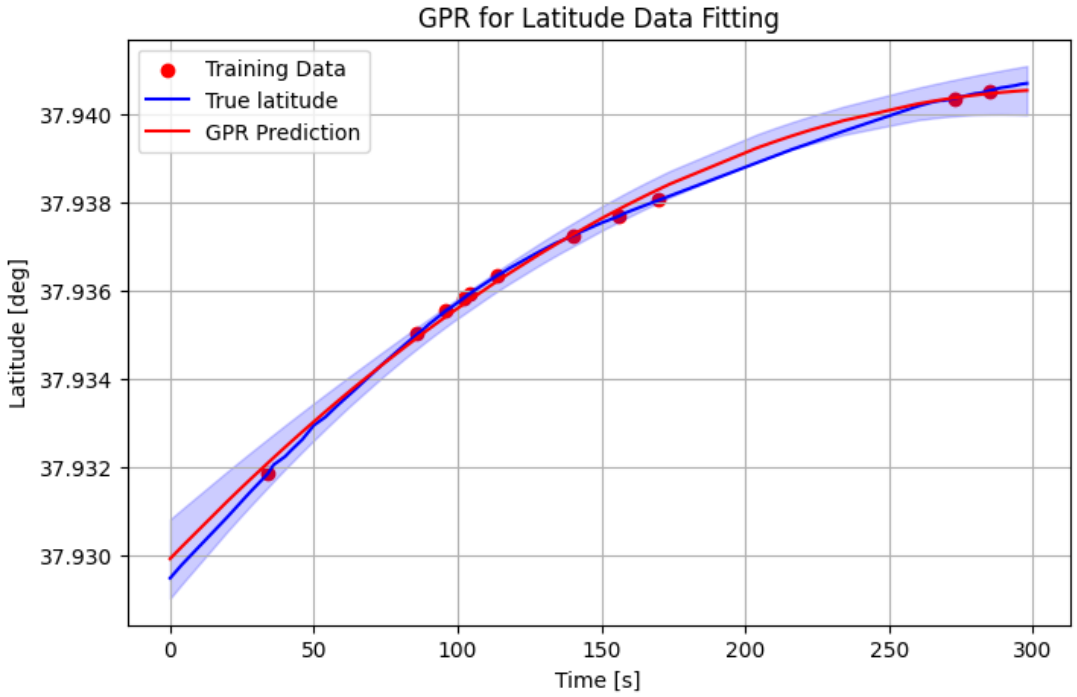

Figure 2.4: Latitude fitted through GPR for the considered vessel over 5 minutes.

Bibliography

[1] Andreas Tritsarolis, Yannis Kontoulis, and Yannis Theodoridis. The piraeus ais dataset for large-scale maritime data analytics. *Data in brief*, 40:107782, 2022.