Gaussian Processes

Analysis 5 for MIT 6.419x Data Analysis: Modeling and Applications

Evan Rushton

June 2, 2021

1 Problem 2: Identifying long-range correlations

Note that for many of my plots I have (y,x) plotted even though I am using the conventional (x,y) order when referencing points in text.

(5 points): A map with the two points with correlations marked.

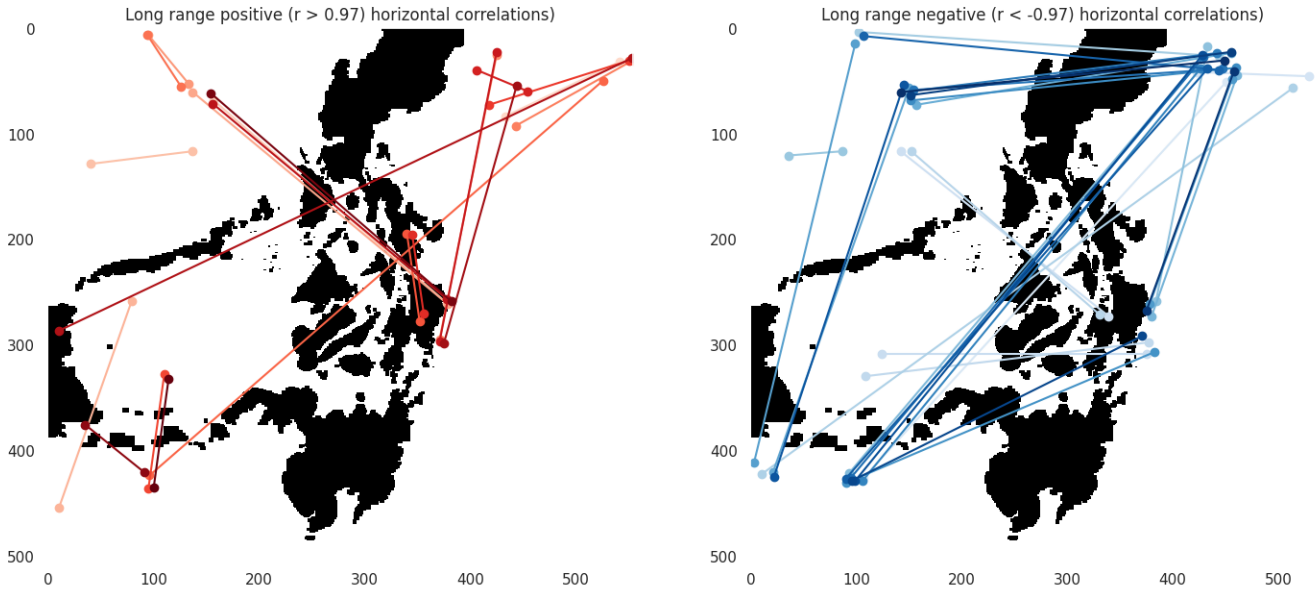

Figure 1: Long-range correlations in oceanic flow in the Philippine Archipelago over 300 hours in Jan 2009

(3 points): Provides an explanation of how the correlation was computed.

Figure 1 shows long range correlations that were identified in horizontal oceanic flow values. Negative correlations are colored blue and positive correlations are colored red (deeper color is stronger correlation). Paired points are joined by line. To find these regions, correlations were found by randomly sampling 10^6 pairs of points and computing the correlation of their 100 time values in the horizontal direction. A threshold of 0.95 for positive and -.95 for negative correlations was set. To ensure long-range correlations a minimum distance was set at 50km. More can be done with the veritcal component of velocity, but for the sake of time I stopped with this preliminary analysis.

(2 points): Provides a convincing commentary on why the two marked locations could be correlated.

Notice that there are some regions where paired points are repeated. POSITIVE REGION 1: (90, 425) with (110, 325) suggests that the horizontal flow of the ocean is the same on those opposing sides of the string of islands by (100, 400). POSITIVE REGION 2: (150, 50) with (375, 250) suggests that the horizontal component by the outer coast near (375, 250) is the same as the inner waters closer to the mainland at (150, 50). It would be good to look at the average magnitudes to see if the correlation is due to generally low values for horizontal flow at these points over time. NEGATIVE REGION 1: (100, 425) with (400, 50) suggests currents traveling in opposing directions. These points may be indicative of particular current cycles and we can see that at these paired points they move in opposing directions. Imagine a vector filed in the top right curling down along the outer coast and another vector field curling inward and down in the bottom left under the islets. These two fields will have opposite horizontal value around the paired points. NEGATIVE REGION 2: (125, 75) with (425, 50) ...

2 Problem 3: Simulating particle movement in flows

(3 points): Provides an explanation of the simulation algorithm, with equations for the evolution of the particle trajectory.

The simulation assumes that the time-varying flow at each location is provided in the data set along 3-hour intervals. A particle in space with a given location has its location $x(t)$ updated by the following update ruie, $x(t+\epsilon) = x(t)+\epsilon v(t)$. For the simplest case, we use $\epsilon = 3$ hours and round each location to the nearest integer to get the vertical and horizontal components of $v(t)$. If the final location is outside of our grid of points, we find a serrogate location defined by the min distance to the grid.

(2 points): Provides a plot of the initial state of the simulation. (3 points): Provides two plots of intermediate states of the simulation. (2 points): Provides a plot of the final state of the simulation.

Figure 2 shows the simulation with 1000 random points. Points starting on land are colored red. The size of the point is proportional to the magnitude of the flow rate and the points get less transparent over time to indicate direction.

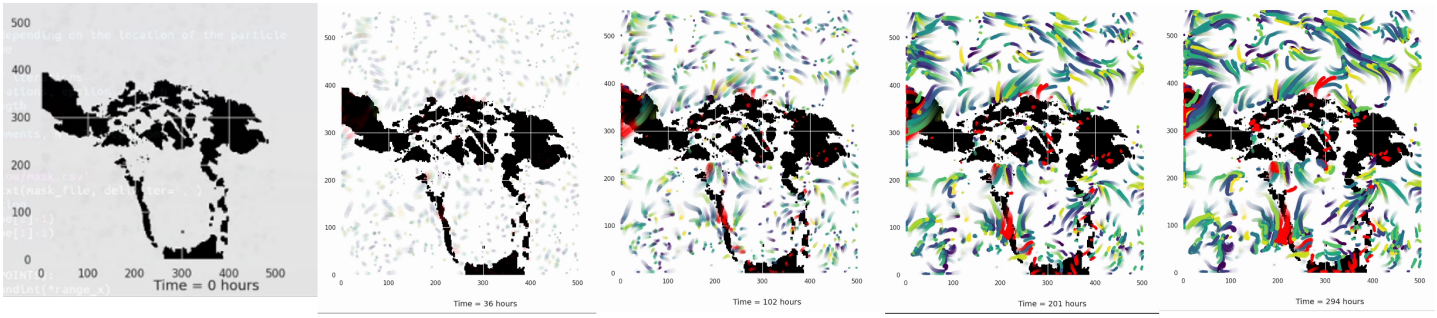

Figure 2: Simulation of 1000 points with $\epsilon = 3$ hours

2.1 A (toy) plane has crashed in the Sulu Sea at T=0.

(3 points): Provides plots showing the state of the simulation at the times: T=48hrs, 72hrs, 120hrs. (Three plots required.)

The simulation is shown in Figure 3 for 1000 points randomly sampled from a guassian distribution with mean $= (100, 350)$ and cov $= \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Given this very narrow variance we can predict with some certainty that the debris will be located in a circle with radius of 15 km centered around $(348\text{km}, 1080\text{km})$.

(3 points): Two or more additional choices of the variances were tried, and three plots of the state of the simulation at the above three times are provided. (Six additional plots required.)

The simulation was run two more times with mean $= (100, 350)$. With cov $= \left[\begin{smallmatrix} 10 & 0 \\ 0 & 10 \end{smallmatrix}\right]$ shown in Figure 4 and then with cov $= \left[\begin{smallmatrix} 100 & 0 \\ 0 & 100 \end{smallmatrix}\right]$ shown in Figure 5. As the variance increases the predicted circle of debris becomes quite large with a radius of 30-45 km for covariance 10 and 70-90 km for covariance 100, still centered around $(348\text{km}, 1080\text{km})$.

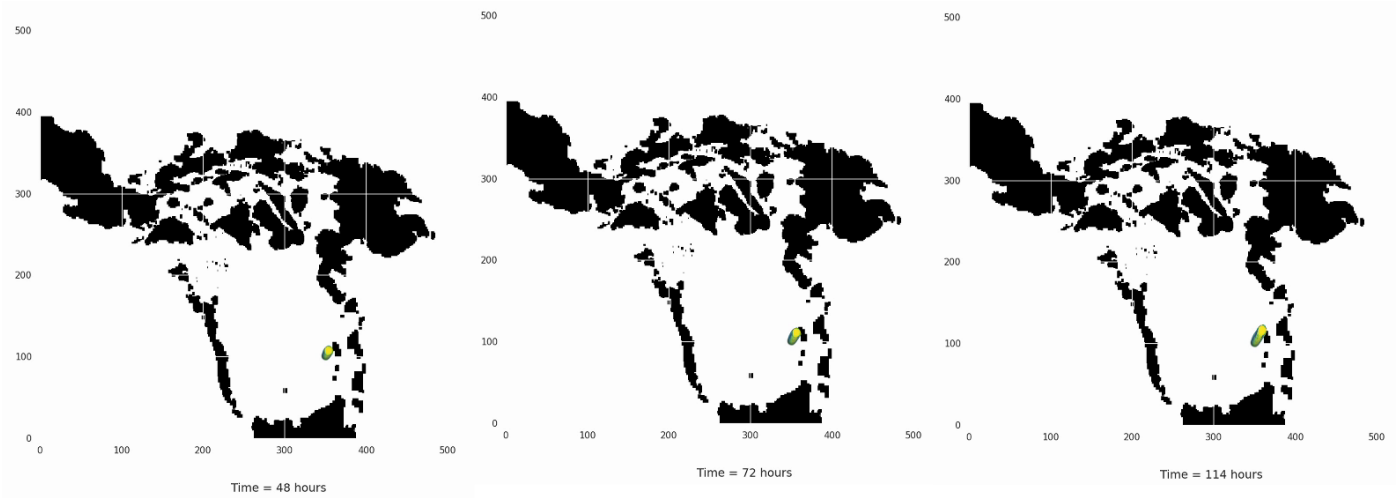

Figure 3: Predicted location of 1000 points with initial location sampled from $\mathcal{N}((100, 350), \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right])$

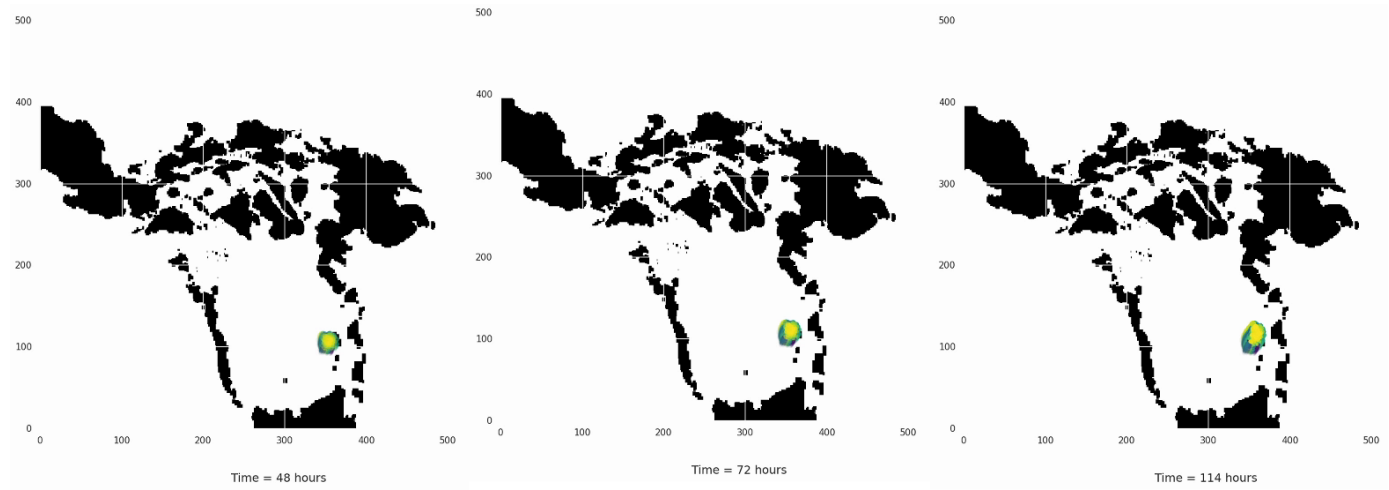

Figure 4: Predicted location of 1000 points with initial location sampled from $\mathcal{N}((100, 350), \left[\begin{smallmatrix} 10 & 0 \\ 0 & 10 \end{smallmatrix}\right])$

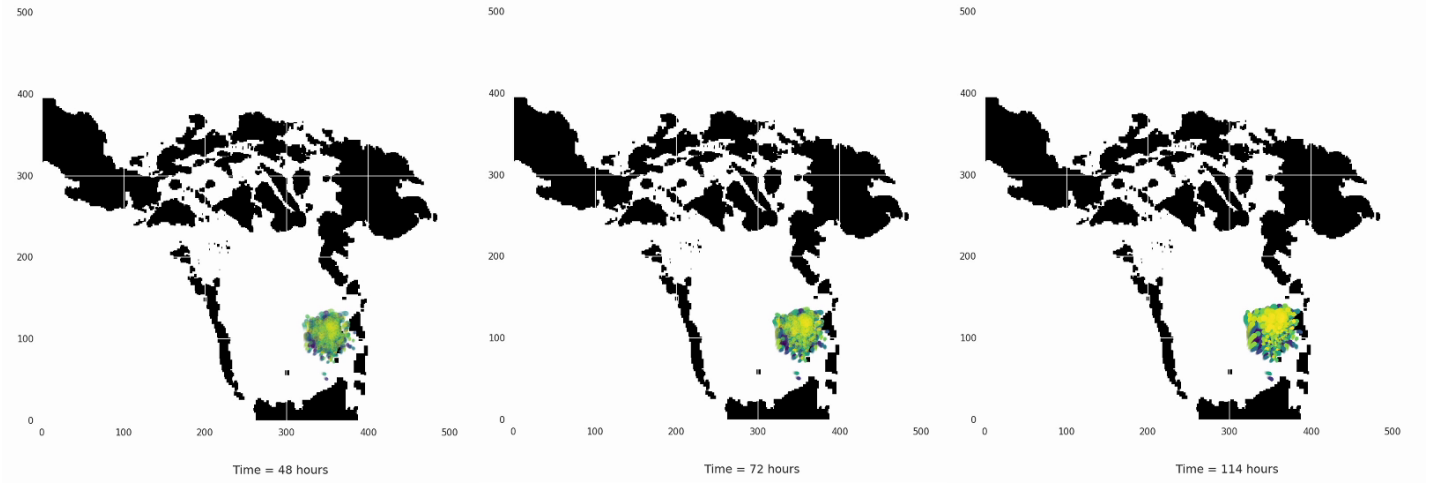

Figure 5: Predicted location of 1000 points with initial location sampled from $\mathcal{N}((100, 350), \left[\begin{smallmatrix} 100 & 0 \\ 0 & 100 \end{smallmatrix}\right])$

(4 points): Comments on where one should concentrate search activities based on the obsered results.

Based on these simulations it makes sens to start the search near the center across all simulations which is around $(348\text{km}, 1080\text{km})$. There is a consistent current in this area that is directed toward the larger island, so it would make sense

to look in an eliptical shape that stretches further in the direction of flow.

3 Part II - Estimating Flows with Gaussian Processes

3.1 Problem 4 [20 pts] Creating a Gaussian process model for the flow.

3.1.1 4a. Find the parameters of the kernel function that best describes the data independently for each direction.

(1 point): States the choice of kernel function and provides a justification for this choice.

The squared exponential kernel function was used because it is commonly used and makes sense given (lat, long) spatial data. I am choosing the location (1200km, 900km) because it is off-shore heading out to sea and looks to be a source that travels in a variety of directions based on the flow field. The histograms for the x- and y- components of velocity at this location are shown in Figure 6 and the mean of these values is used for our guassian process rather than zero since they don't look to be centered about zero. $\mu_x = 0.6476$ and $\mu_y = 0.2288$ at location (1200km, 900km).

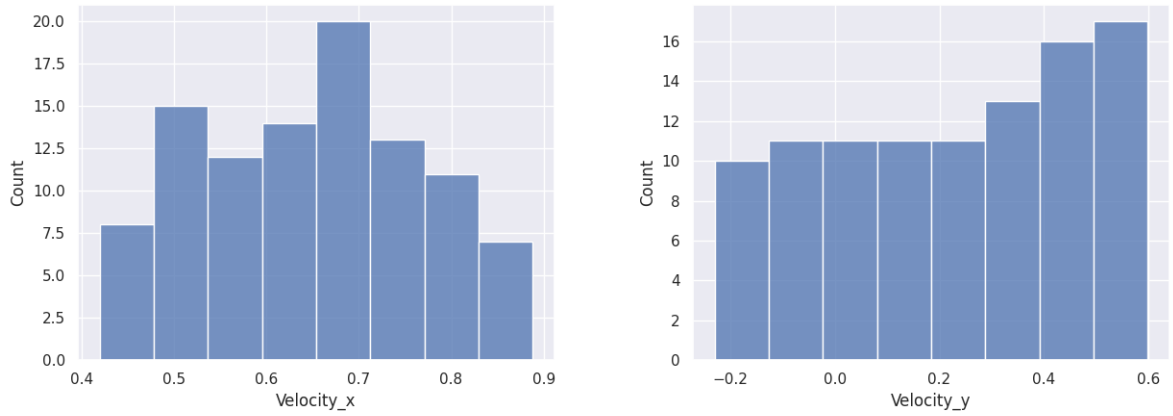

Figure 6: Histograms for the x- and y- velcoity components at point (1200km, 900km over 100 3-day time steps)

(1 point): Identifies the parameters of the kernel function.

The Radial Basis Function kernel has paremeters $\theta = (l, \sigma)$ where l is the lengthscale and σ is the output variance.

(1 point): Explicitly states the search space for each kernel parameter.

The search space for the lengthscale l is 7.2 to 360 hours (.1, 5) and for the output variance σ is -0.5 to 0.5

(1 point): Explicitly states the number of folds (k) for the cross-validation.

I performed 10-fold cross-validation with 10 sets of 90 training points with 10 test points. I ran out of time learning to conduct this analysis. A few more points beyond here.

(3 points): Provides the optimal kernel parameters from the search. (3 points): Provides a plot of the computed cost/performance metric over the search space for the kernel parameters.

The resulting search yielded optimal paramters $\theta = (??, ??)$. The cost/performance metric is plotted over the paramter search space in Figure **??**.

3.1.2 4b. Run the process described in part (4a) for at least three more points in the map, you free to choose more if you wish. What do you observe? Which of your kernel parameters show patterns? Which do not?

(3 points): Provides the optimal kernel values for three new location that are different from the location in Problem 4.a. (Plots do not need to be provided.)

Three other locations were chosen and the optimal kernel values were obtained as displayed in Table 1, (450km, 450km), (450km, 1200km), and (801km, 1200km) for a variety of directions based on the flow field. We use the mean of each input vector for our guassian processes. I ran out of time building this Gaussian Process. I hope to complete this project this week. From here on out my paper is a work in progress, apologies.

Table 1: Seed points chosen for velocity vectors. Note I wasn't able to solve for optimal Kernel parameters yet.

Seed Point	μ_x	μ_y	Parameters
(1200, 900)	0.6476	0.2288	$(l = 1, \sigma = 0.3)$
(450, 450)	0.8243	-0.6249	$(l = 1, \sigma = 0.3)$
(450, 1200)	-0.2697	-0.1950	$(l = 1, \sigma = 0.3)$
(801, 1200)	-0.6492	-0.4166	$(l = 1, \sigma = 0.3)$

(2 points): For each kernel parameter, states if a pattern was observed.

the following pattern was observed...

3.1.3 4c. We have suggested one particular value for τ. Consider other possible values and comment on the effects such parameter has on the estimated parameters and the estimation process's performance. Try at least two values different from that used in Problem 4.a.

(1 point): Provides the optimal kernel values for at least two new choices of τ.

Tau was tested at $\tau = 0.1, 0.01, and 0.0001$. The optimal kernel values are...

(2 points): A plot showing the cost/optimization target is provided for the search space, for each choice of τ.

(2 points): Comments on whether these results differ from those found in Problem 4.a, and on whether results from the choices of τ in the problem differ from each other.

Figure **??** shows the optimization across the search space while varying τ.

3.1.4 4d. Use one library of your choice and compare the obtained results. Did you get the same parameters as in problem 4.a? If not, why are they different?

(2 points): Provides the optimal kernel parameters as found through the software library. (2 points): Provides details on the library used.

This is a cool project, I hope I can complete it. Thanks for taking the time to grade. Be well, Evan Rushton.