Don't Delete Carriage Returns Above. The Title of Your Paper Goes Here Using the Paper Title Style

**Team 08**

**Dmitrii Kharkovskii /////**matric///

**Naheed Anjum Arafat /////**matric///

**Liu Dake /////**matric///

**Wu Lifu /////**matric///

**Daren Tan Zong Loong** A0111855J

Affiliations and Addresses Go Here Using the Affiliation and Address Style  
publications15@aaai.org (to delete?)

Abstract

This paper presents a real world application which relies on the Gaussian process (GP) models for probabilistic non-linear regression and prediction. Our motivational example makes use of a 1 degree by 1 degree HadSST2 dataset to predict the sea surface temperature (SST) anomaly for any valid coordinate on the map which can then be interpreted to other valuable information such as weather predictions. By providing SST anomaly values for specific coordinates, it should allow experts to plan and make better decisions based on a distribution of SST anomaly values.

1. Introduction

Sea surface temperature (SST) anomaly is a crucial oceanographic variable that are provided by continuous observations from ships, vessels and buoys. These SST anomalies are good indications of El Niño and La Niña climate cycle. For more specific locations, it can influence weather patterns across regions.

At the equator, one degree of longitude and latitude both cover about 111 kilometers, or just about 70 miles. By being able to pin-point a coordinate and produce a good measure of SST anomaly at the specific coordinate, it gives a better representation of the anomaly situation within the one degree grid. However, deploying buoys and vessels in order to cover areas within a 1 degree grid is inefficient.  
  
This brings us to our motivating application, which aims to use the Gaussian Process (GP) to provide a good distribution of SST anomaly values within the one degree grid for water-vehicles to be able to know where to move within the one degree grids.

Prediction of SST Anomalies

Our problem can be modeled by a Gaussian Process, the input is the locations (including the locations which have measurement in the HadSST2 dataset and the locations whose SST anomalies we are going to predict) and the output is the SST anomalies.

From the HadSST2 dataset, we have observation of some output, with which we can use the Bayesian model to predict the mean and uncertainty of the SST anomalies (variance) of the locations that are directly not observed. The non-parametric nature of GP model also gives us enough flexibility not to worry about whether our model can fit the data since we are giving prediction as a distribution over all such models (parameters) rather than just one model parameter.

Qualitative Advantages of GP Models

Simple Linear Regression (available in Excel) is not appropriate for modeling relationships between latitudes, longitudes and sea-surface temperature as the values of SST anomalies depends on more than one variable (requires both latitude and longitude). It is easy to comprehend from the following plots that the relationship between latitude, longitude and SST anomaly values is not linear.

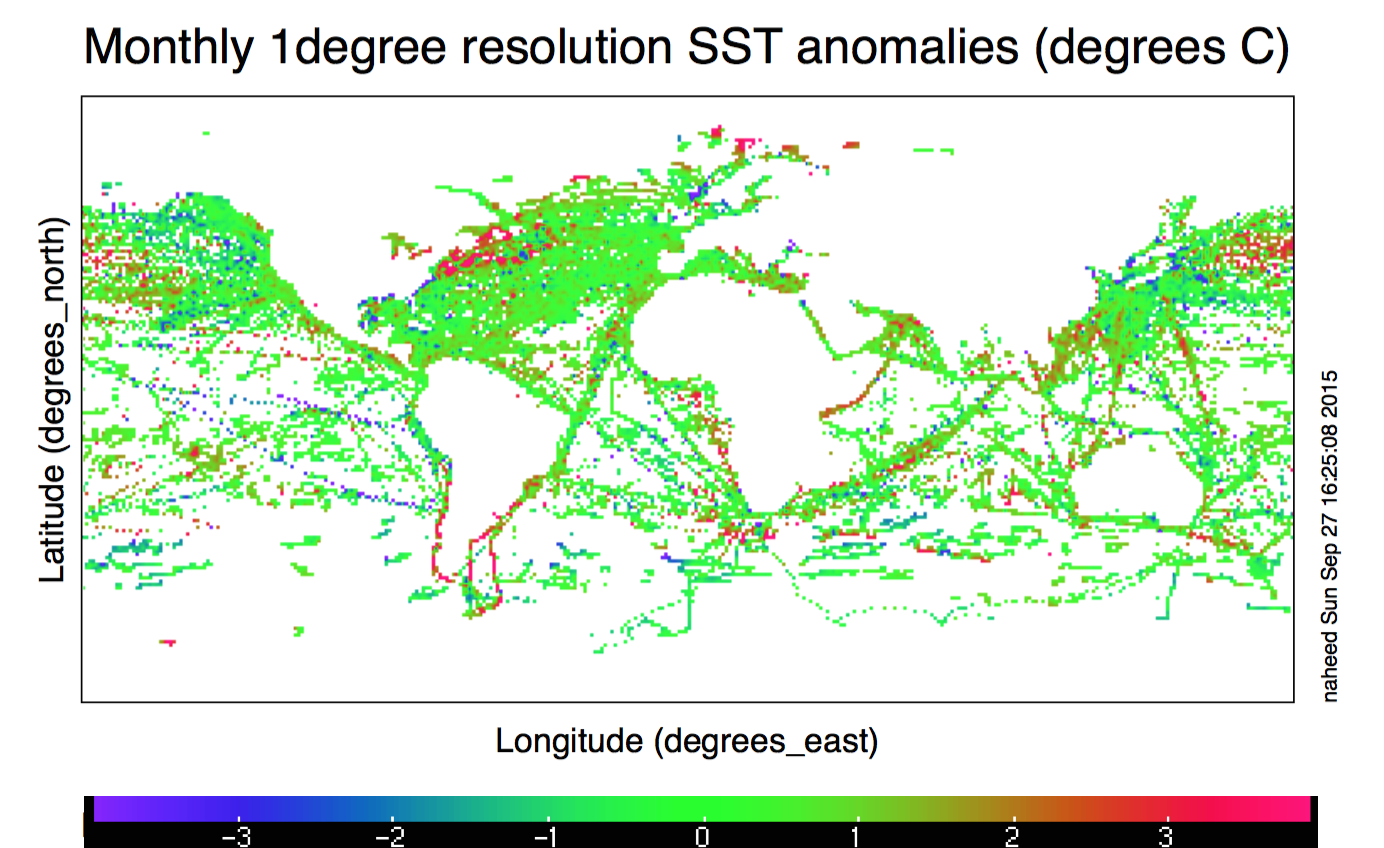


Figure 1. Monthly 1degree resolution of SST anomalies (degrees C)

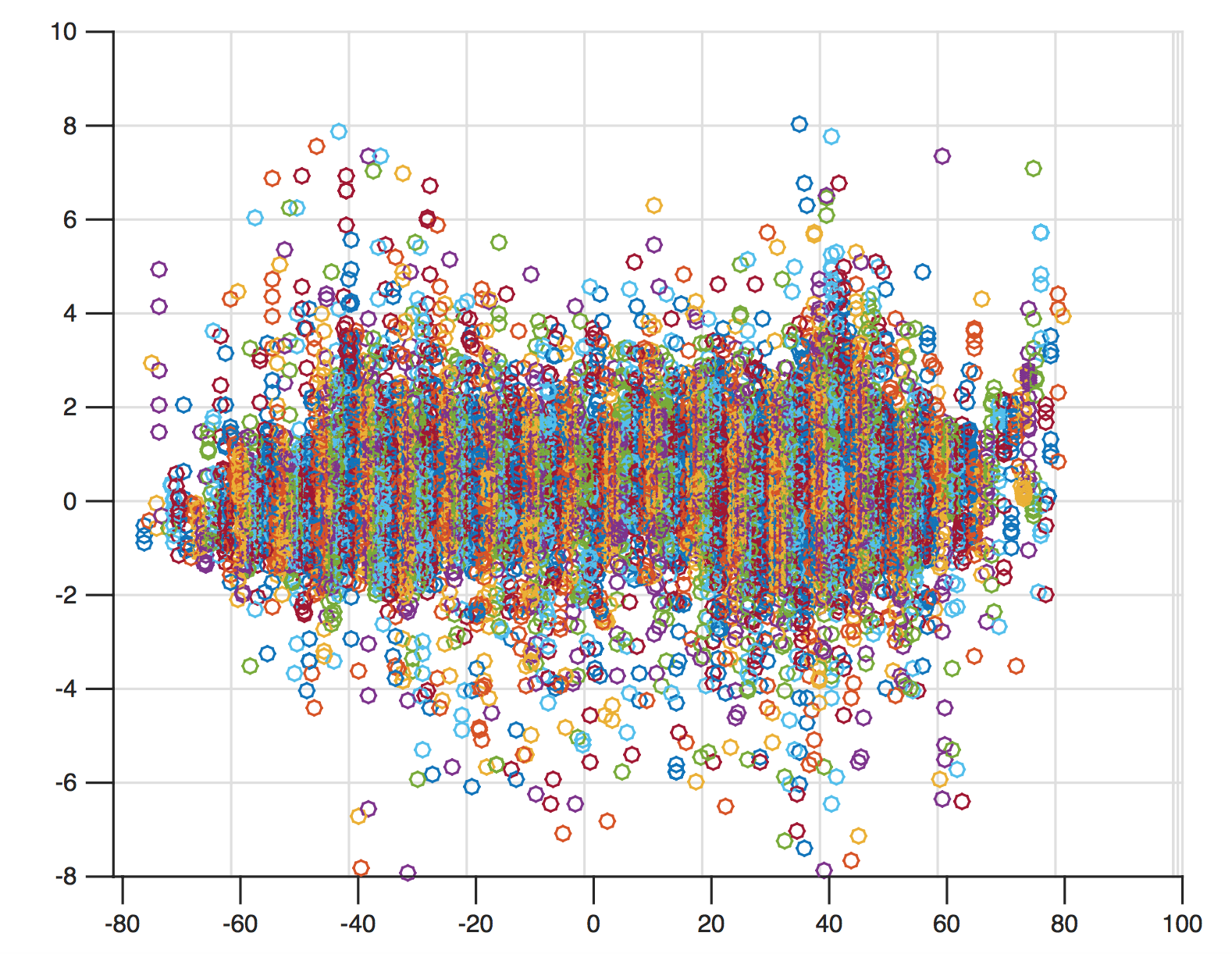


Figure 2. Plot of SST values across latitudes

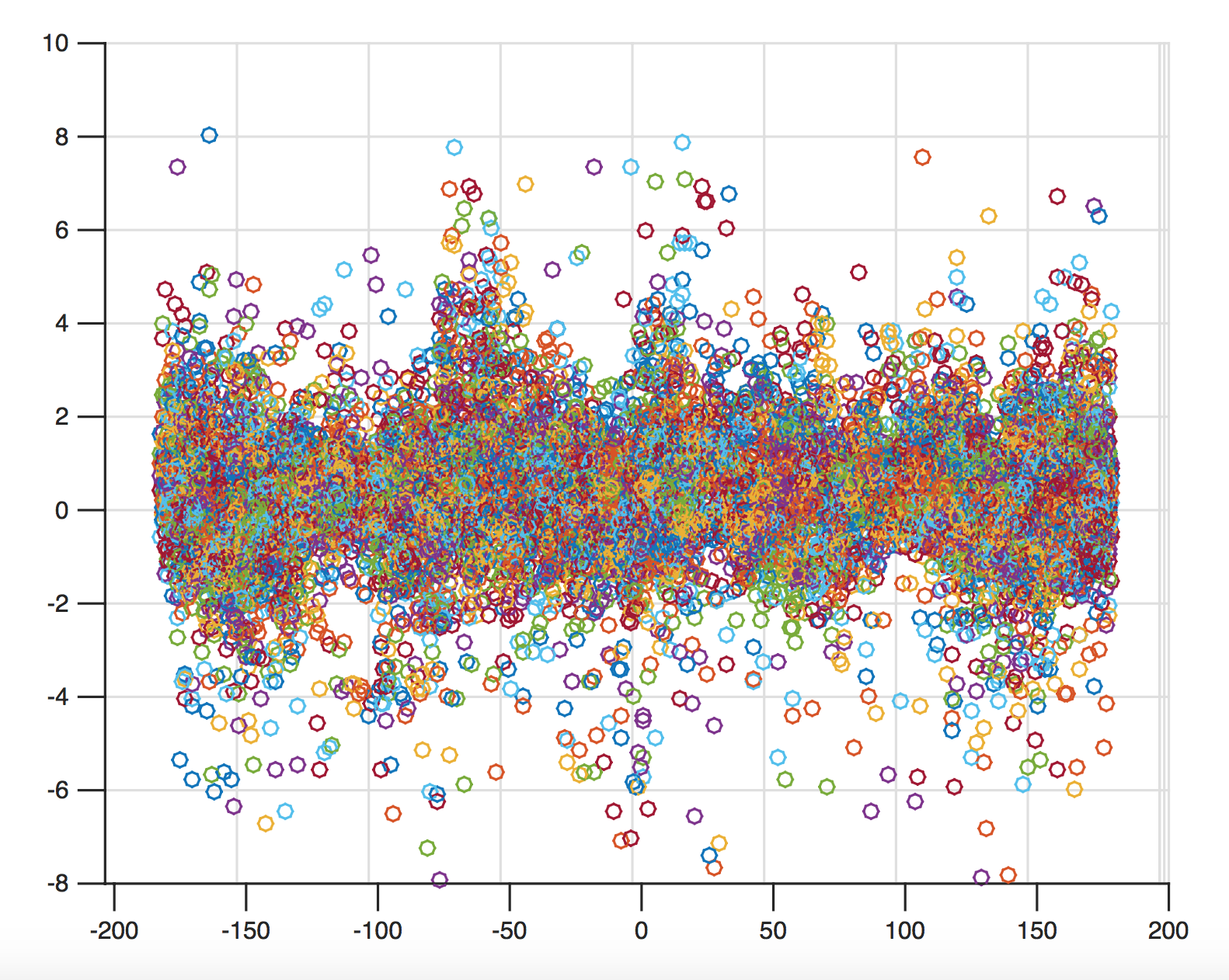


Figure 3. Plot of SST values across longitudes

From those plots, we observe that most of the regions have mean SST anomaly value (MSstA) of 0. However adjacency of blue, green and red regions (Figure 1) signifies that the relation is not linear. As a matter of fact, It is difficult to make any kind of assumptions about the relation between SST and the model parameters (i.e. whether it is linear or non-linear) as well as between the SST and the coordinates. Since the SST anomaly values varies across the coordinates in an unpredictable manner which is expected from any natural phenomena. Therefore, a non-parametric model would be a better choice for this particular application.

However, the SST anomaly values are spatially correlated, as observed from large spatially adjacent region of green dots and trails of red and blue dots (Figure 1).  Gaussian Process (GP) is particularly suitable for this kind of non-parametric regression modeling where spatial correlations among the independent variable (regressor) are known. Since we are interested to predict the MSstA for any region with known predictive uncertainty (derived from GP regression), which later can be used for planning a navigation route for a water vehicle.

**Data Requirements**

For our application, we will be relying on the 1 degree version of the HadSST2 dataset. This dataset has been refined in a way that it is time homogenous and has already passed the basic ICOADS quality checks that filter out invalid land locations or duplicates. The HadSST2 is a large data set, (approximately 65,000 data points) and is used to map a global field of SST anomalies monthly.

Our application also requires a recent dataset (i.e. using specific dataset from the same year/month) with uniformly distributed data points in order to project the SST anomalies distribution within each 1 degree grid.

**Interpretation of Data & Decision Making**

The output of our GP would be a more continuous and complete map for the ocean area. With the output of the GP, human experts would be able to see the gradient trend of ocean temperature anomaly, and gain information on how other co-related climate phenomenon (e.g. direction of ocean current, sunshine intensity at different latitude) reflected/affects ocean temperature anomaly.

Moreover, with the whole timeline of series change of data, users would be able to predict the future appearance of ocean temperature anomaly in anyplace.

**2. Technical Approach**

GP models are very popular in environmental sensing and spatial analysis because they can take into account the spatial correlation between the points of the input data which exactly fit our real world application - we assume that since the sea temperature changes continuously, the values at points, which are close to each other, should be more correlated. So using GP model will allow us to use the advantage of spatial correlation.

Another important aspect of the GP model is that it allows us to deal with noisy observations, which might also be useful. Since we do not know if the process of obtaining the HadSST2 dataset filters out noisy measurements, we should assume that the measurements are noisy.

**Additional Insights**

Since there are missing values in the dataset along with some unnecessary land co-ordinates which do not constitute our dataset, we had to trim those values which resulted in approximately 15,000 data points before running a Full-rank GP on this dataset.

**Ideas for Performance Improvements**

One proposed method of improving time efficiency is to divide the full datasets into smaller subsets that are justifiably uncorrelated (i.e. a point in pacific ocean is by no means correlated with a point in Atlantic ocean) and use a full-rank GP on each subset to get the covariance function parameters.

However, the parameters derived this way might be sub-optimal. Therefore, optimal policy for segmenting the regions into sub regions would be very crucial for the accuracy of the prediction.

For prediction, when a user requests prediction for a certain point, we will calculate which subset it belongs to and use only that covariance function to predict the value of the point. This proposed modification should not affect the results substantially given the regions are segmented optimally and provide an improved time efficiency.

**3. Experimental Evaluation**

In this section, we will demonstrate the various GP models that have been implemented to evaluate its performance in terms of accuracy, time efficiency and how each GP model performs with respect to our proposed application.

As mentioned above, we will be using the monthly data provided in the 1 degree version of the Hadley Centre SST data set (HadSST2)for both training and testing purposes.

**Experimental Setup**

////////////////

////////////////

**Model Comparisons & Analysis**

////////////////

////////////////

**4. Conclusion**

References

Lisa.stat.vt.edu, (2015). Parametric versus Semi/nonparametric Regression Models | LISA: the statistical collaboration laboratory of Virginia Tech. [online] Available at: http://www.lisa.stat.vt.edu/?q=node/7517

[Accessed 29 Sep. 2015]

Rayner, N.A., P.Brohan, D.E.Parker, C.K.Folland, J.J.Kennedy, M.Vanicek, T.Ansell and S.F.B.Tett (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set. Journal of Climate. 19(3) pp. 446-469 [Accessed 6 Oct. 2015]

**Team Contributions**

Dmitrii Kharkovskii -

Naheed Anjum Arafat -

Liu Dake -

Wu Lifu -

Daren Tan Zong Loong -

Margins (TO DELETE)

Papers **must** be printed in two-column format on 8-1⁄2 x 11 inch paper. The margins if you use 8-1/2 x 11 inch paper must be **exactly** as follows:

• Top margin: 3⁄4 of an inch

• Left margin: 3⁄4 inch

• Right margin: 3⁄4 inch

• Bottom margin: 1-1⁄4 inches

The actual image size for your paper (minus margins) is 7

x 9 inches (17.78 x 22.86 cm). Papers that deviate from these measurements will not be published.

Fonts

All fonts must be embedded for your paper to be published. No exceptions.