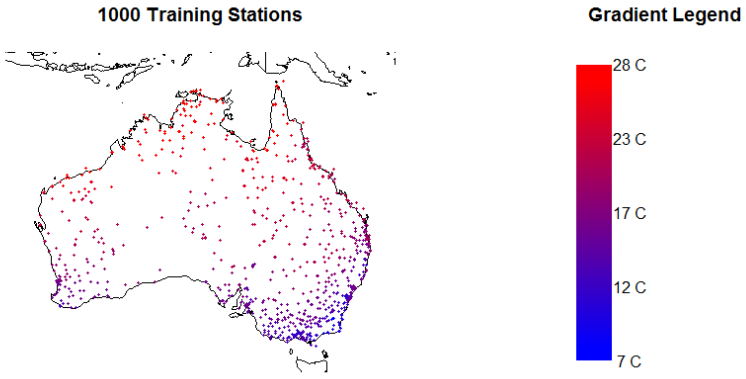
Sam Baugh
NonParametric Inference
Data Analysis Project Writeup

1. Description of the Data

The goal of this project is to interpolate 1980 mean temperatures over the whole of Australia using data collected by a set of 1,588 weather observations stations located throughout the country. This data was obtained from the Australian Bureau of Meteorology climate data online service (url: `http://www.bom.gov.au/jsp/awap/temp/`). At each of the 1588 weather stations, temperature was collected daily at regular intervals. Then the daily mean temperature for each day was calculated by averaging the maximum and minimum recorded temperatures. In the dataset that we use, each station has a single mean temperature value that is the average of the daily mean temperatures of the 365 days of 1980. The Australia Bureau of Meteorology underwent quality control measures on this dataset to exclude errors by confirming extreme values and by comparing the observations of nearby stations.

Australia is advantageous for spatial modeling as the shape is relatively square (in comparison to other continents) which is preferable since temperature observations over water are more sparse than land observations, so we prefer for our dataset to have minimal water boundaries. However, a disadvantage of the Australia dataset is that the data is not uniformly distributed throughout the continent, with a dense collection of observations near the southeastern coast (where over half the population is located) and a sparse distribution of observations in the western interior. As such, we expect our interpolations in the western interior to be less accurate than those near the southeastern coast. We also note that we eliminated the data from Tasmania, as its physical separation from the mainland renders it unnecessary for our purposes.

Since we wish to evaluate the interpolative ability of our models, we generated a fixed random partition of our 1588 observations into a 1000 observation training set and a 588 observation test set. We plot the training data below, overlaid on a map of Australia along with a temperature gradient. Temperatures appear to decrease relatively smoothly from the warm north (latitudes near the equator) to the cool south.



2. Parametric Analysis

The Matern kernel is commonly used to model spatial covariances. The kernel takes the following form:

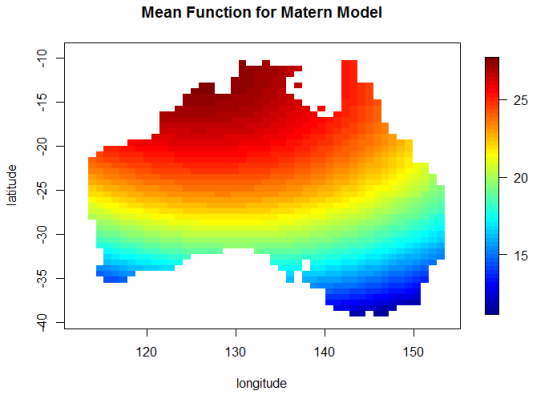
$$K(\mathbf{x},\mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{d(\mathbf{x},\mathbf{y})}{\rho}\right)^{\nu}\mathcal{K}_{\nu}\left(\sqrt{2\nu}\frac{d(\mathbf{x},\mathbf{y})}{\rho}\right)$$

where d is Euclidean distance, Γ is the gamma function, and $\mathcal{K}_{\nu}$ is the modified Bessel function of the second kind. Intuitively, ρ can be interpreted as a range parameter and ν can be interpreted as a 'smoothness' parameter (a Gaussian process with the Matern kernel is $\lceil \nu - 1 \rceil$ times differentiable).

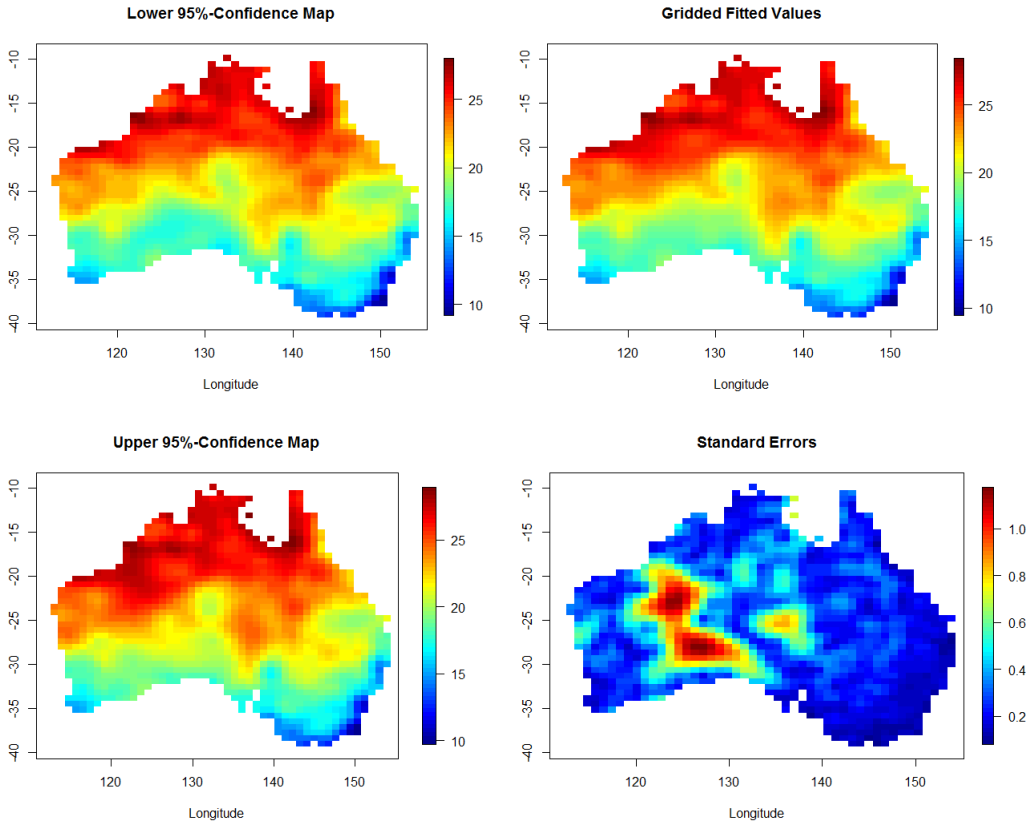
To use this kernel we make the assumption that the underlying process is Gaussian, specifically that any finite set of observations $\mathbf{x} \in \mathbb{R}^n$ has a multivariate normal distribution with mean $\mathbf{m}(\mathbf{x})$ and covariance matrix $\Sigma(\mathbf{x}) = \phi^2 K_{\nu,\rho}(\mathbf{x}) + \sigma^2 I$, where σ^2 is the 'nugget' parameter or variance, ϕ is the 'sill' parameter or measure of spatial variability, and $K(\mathbf{x})$ is the $n \times n$ matrix such that $K_{ij} = K(x_i, x_j)$ for the Matern kernel specified above. Given the 1000 data points of our training data, we use REML (restricted maximum likelihood) in order to estimate the mean function as well as the parameters of the Matern covariance function. We obtain the following Matern parameters:

$$\hat{\phi} = 0.05661127,\ \hat{\rho} = 4.51786655,\ \hat{\sigma} = 1.05287098,\ \hat{\nu} = 1.5$$

and we obtain the following mean function $\hat{\mathbf{m}}(\mathbf{x})$, evaluated on the the 50×50 grid over Australia in order to show the captured trend:

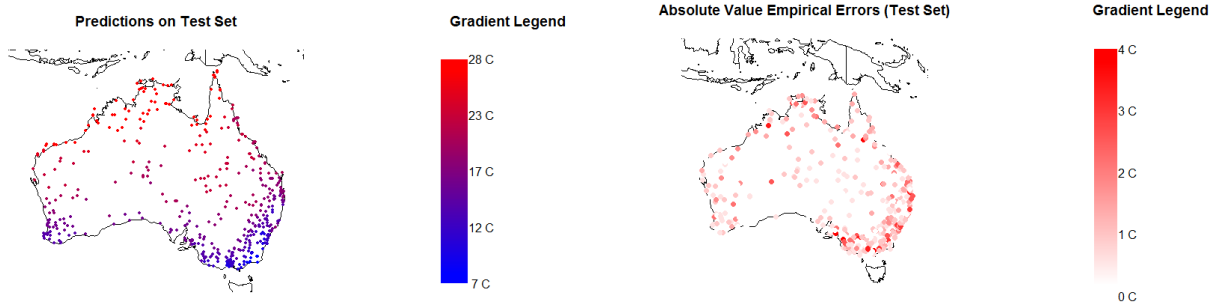


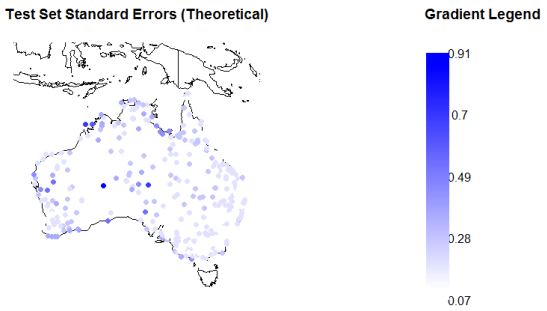
We note that the Matern parameters $\hat{\phi}, \hat{\rho}$, and $\hat{\sigma}$ are estimates of the corresponding MLEs obtained through gradient descent, and the smoothness estimator $\hat{\nu}$ was obtained through experimentation (the MLE of ν is difficult to estimate as there is currently no known algorithm to compute the Matern gradient with respect to ν). Using the techniques of kriging we interpolate onto the 50×50 grid over Australia in order to obtain a temperature heatmap. We also include the lower and upper 95% confidence maps and the theoretical standard errors of our predictions plotted on the grid.



The differences between the confidence maps are subtle and somewhat hard to spot with the naked eye. However, the standard error plot is more insightful, as we can see that the standard prediction errors for locations in the Western interior are much higher than those in other parts of the country, which is as expected given that observations are significantly sparser in that region.

We now evaluate our model by predicting the temperature value for the 588 observations in the test set. We display the following three plots containing our predictions, the absolute empirical error values, and the theoretical standard errors.





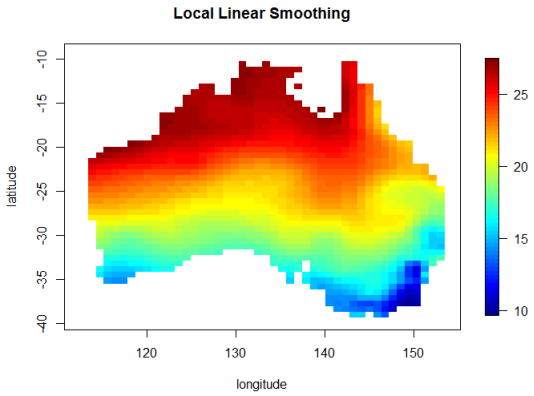
What is interesting is that the errors seem to be fairly large in the dense observation region on the southeastern coast in comparison to the rest of the country, despite the fact that our standard errors for that region are relatively small. We believe that the reason for this is the inevitable boundary bias of estimating near the coast. The mean squared error on the test set 1.715023.

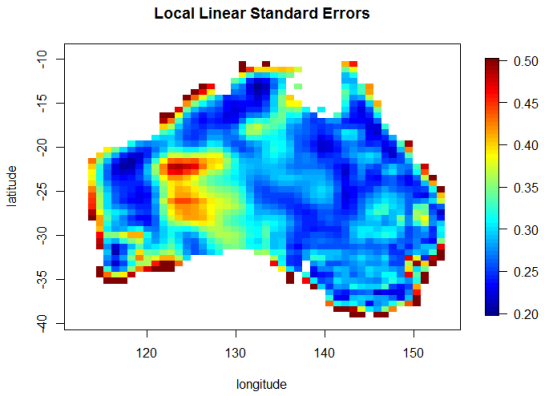
Other parametric models considered for analyzing this data include Gaussian process models with kernels other than the Matern and more complicated Gaussian process models. Non-Gaussian models are more complicated and were not considered. Other kernels considered include the squared exponential kernel, which the Matern kernel converges to as $\nu \to \infty$, and the Rational quadratic kernel. We chose the Matern kernel because it allows for the estimation of smoothness and range parameters. The squared exponential kernel is infinitely smooth, and while our data is fairly smooth we get a better fit using the Matern kernel with smoothness parameter $\nu = 1.5$ than by using squared exponential kernel.

These kernels are all isotropic, meaning their values only depend on the distance between two points. Since the data has a clear north-sound trend, we considered fitting more complicated Gaussian process models with anisotropic components. However, we did not achieve a significantly better fit with any such models, indicating that our estimated mean adequately takes care of the apparent anisotropies.

3. NonParametric Analysis

We propose local linear smoothing as an appropriate nonparametric model. Unlike the parametric model, this does not assume any sort of Gaussian process on the underlying process. However, similar to the parametric model it assumes that there is an underlying mean function which is relatively smooth, and as such that smoothing the data approximates this mean function. Also, the standard errors given by local linear smoothing assumes that the residuals of the observed values minus the means $\epsilon = y - \mathbf{m}(x)$ are independent Gaussians. We present the smoothed fit over the 50×50 grid, as well as the standard errors, below (with a chosen bandwidth parameter of 0.05):

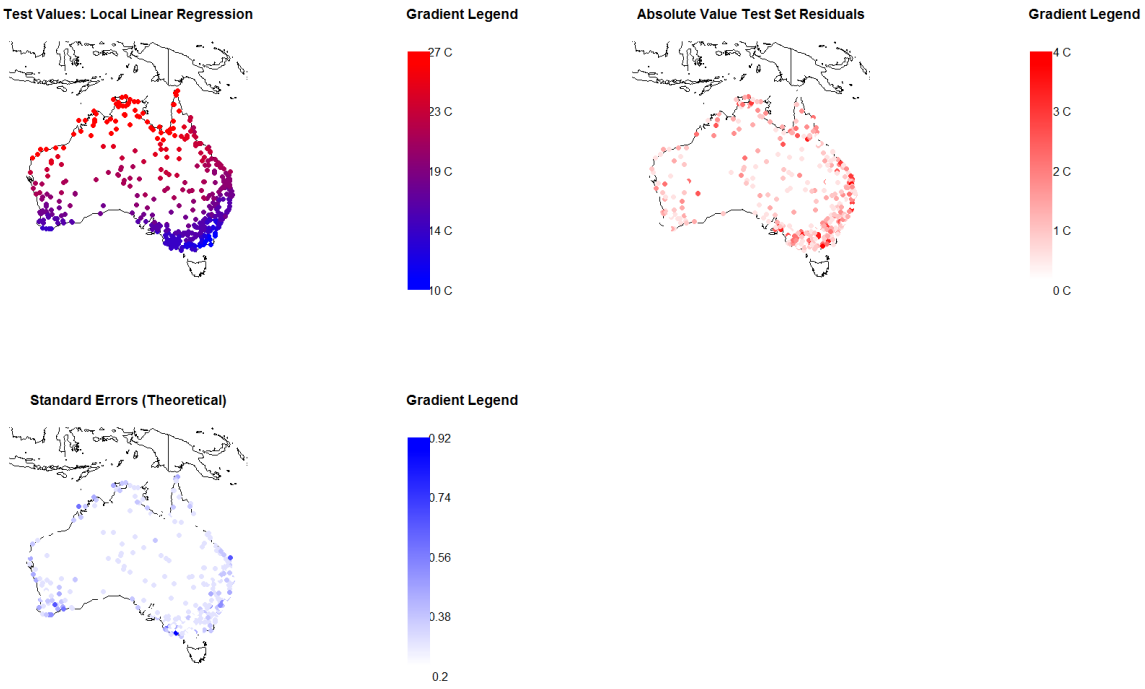




We do not include confidence maps here, because as in the parametric model the standard error plot is much more insightful. We note that we did not choose the bandwidth via cross-validation, as the resulting fit was over-smoothed and did not capture a lot of the details. As such, we opted for a fit with less bias and higher variance than with the estimated optimum bandwidth.

We can see that the Western interior has high standard errors as expected, similar to that of the parametric model fitted in the previous section. However, we notice that the standard errors around the boundaries are much higher than that of the parametric model, which is indicative of the difficulty of accurately smoothing data near the boundary.

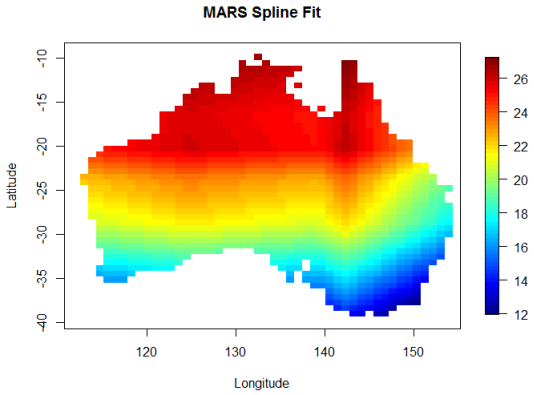
In order to evaluate our model, we predict the values of the test set, and plot the results (the predictions, the absolute value of the errors, and the theoretical standard errors) below.



We can see again that our parametric model has higher standard errors around the Western interior, while our nonparametric model has relatively higher standard errors around the boundary.

We note that our mean squared error is 1.689835 over the test dataset.

We tried fitting other nonparametric models to this data, particularly multivariate adaptive regression splines (called "mars" in R). We did not see any major difference between the mars fit and our local linear fit, as the optimal mars spline is plotted below:



This fit is very similar to the local linear smoothing fit. However, mars standard errors can only be obtained through boostrapping, whereas local linear smoothing estimates have computable, closed form standard errors and as such confidence intervals. Given this, we prefer the local linear fit over the spline fit in order to better quantify the uncertainty in our estimates.

4. Discussion

The mean squared errors of the parametric and nonparametric fits on the test dataset are very similar, with an MSE of 1.715 for the parametric method and an MSE of 1.689 for the nonparametric method. The nonparametric method is slightly better, however we must note that the theoretical standard errors around the boundaries are much higher than those of the parametric method. As such, if we maintain the assumptions of the parametric model hold we conclude that the parametric model is superior as with it we can predict with higher levels of accuracy.

However, we must exercise caution with this preference because the standard errors of the Matern model are deceptive if the model assumptions do not hold. In order to make a more definitive preference, I would propose more rigorously checking the assumptions made by the parametric Matern model. We note that both the nonparametric model and the parametric model assume that the data is distributed as a multivariate Gaussian process, however the local linear smoothing method does make parameterized assumptions regarding the de-trended data whereas the Matern model does. As such, in order to ascertain that the Matern model is appropriate one should test whether the de-trended data actually has an isotropic covariance structure. If the isotropic assumption does not hold then the Matern model is inappropriate, and as such theoretical standard errors are misleading. In such a case, I would suggest choosing the nonparametric model, as while its theoretical standard errors are worse, the uncertainty that they convey would be more accurate than the standard errors of the incorrect parametric model.