Investigating Coastal Images

Charlotte Wickham

11th December 2006

Contents

1 Introduction and Data Description

The data consist of images of Pauanui Beach, located on the Coromandel Peninsula in New Zealand. The images are collected hourly during daylight from the 1st of Jan 2004 to the 31st of December 2005. Two types of images are collected: raw images and averaged images. The averaged images are the combination of images collected over 10 minutes (600 images). Examples of the images are shown in Figure 1. In total there are 7826 raw images and 7749 averaged images. These images are collected by the National Institute of Water and Atmospheric Research (NIWA) as part of their Cam-Era project monitoring 8 of New Zealand's beaches. http://www.niwascience.co.nz/services/cam-era/about.

Figure 1: Example images. Both images are taken Jan 1st 2004 at noon. The image on the left is a raw image; the image on the right is an averaged image.

This report provides a primarily exploratory investigation of some of the properties of these images. Some of the objectives that are addressed are: the seasonal behaviour of the images, the identification of the stormy days, and the different patterns of variability across the image.

2 Data Exploration

One way to think of the image data is as each pixel being a individual time series. Here we consider the time series of the pixels' intensities. The intensity is a measure of how bright the pixel is on a black to white scale. The intensity of a completely white pixel is 1 and a completely black pixel is 0. Hence a greyscale representation of an image is a direct representation of the intensity of each pixel.

We start by considering only the summary statistics of each pixels individual time series. Consider looking at the mean intensity a pixel has over the entire two years as well as the variance in this intensity. To simplify the size of this task the images are first subsetted to only include the noon images and then reduced to greyscale and 143 $\times$ 190 pixels (and the remainder of the analyses are conducted on this reduced set). Figure 2 illustrates both of these statistics by ploting the value for each pixel in the pixels location in the image. The first image shows the mean intensity for each pixel. The second shows the variance of

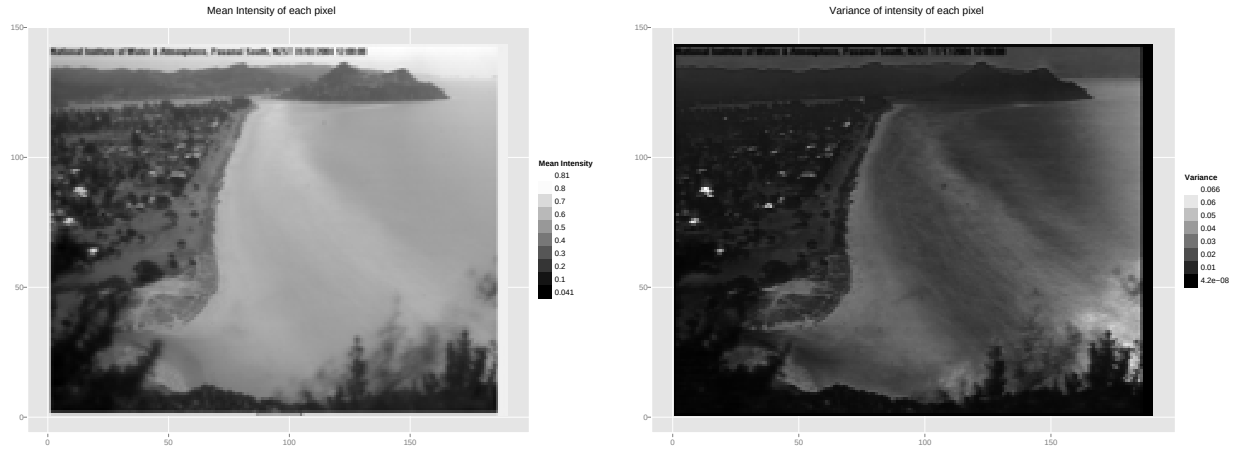


Figure 2: Mean and variance of pixel intensity over time by pixel.

each pixel over time. The average image reveals no surprises. It simply smoothes out any variability (i.e the wave motion, tidal level, weather patterns etc.). The variance image also confirms some expectations. The sea area is one of the more variable areas. The few very

variable areas on land correspond to houses with white roofs. On bright days these tend to be very reflective and have intensities close to one (one of the few areas that ever does). On dull days (or when the sun is not reflecting off them into the camera) they have lower values. It is also of interest that there is a very variable area on the bottom right. This is probably caused by a combination of foliage movement and breaking waves (there are a lot of rocks in the water in this region).

Another way to gain an overview of the data is to formulate a statistic that captures the entire image and then look at the time series of this statistic of the two years. We can do this by looking at the mean intensity of each image as well the variance in the pixel intensity over each image. Plots of the mean and the variance of the pixel intensity are shown in Figure 3. The thicker line is a result of a loess smoother applied to the series. The plots show some unusual and unexpected behaviour.

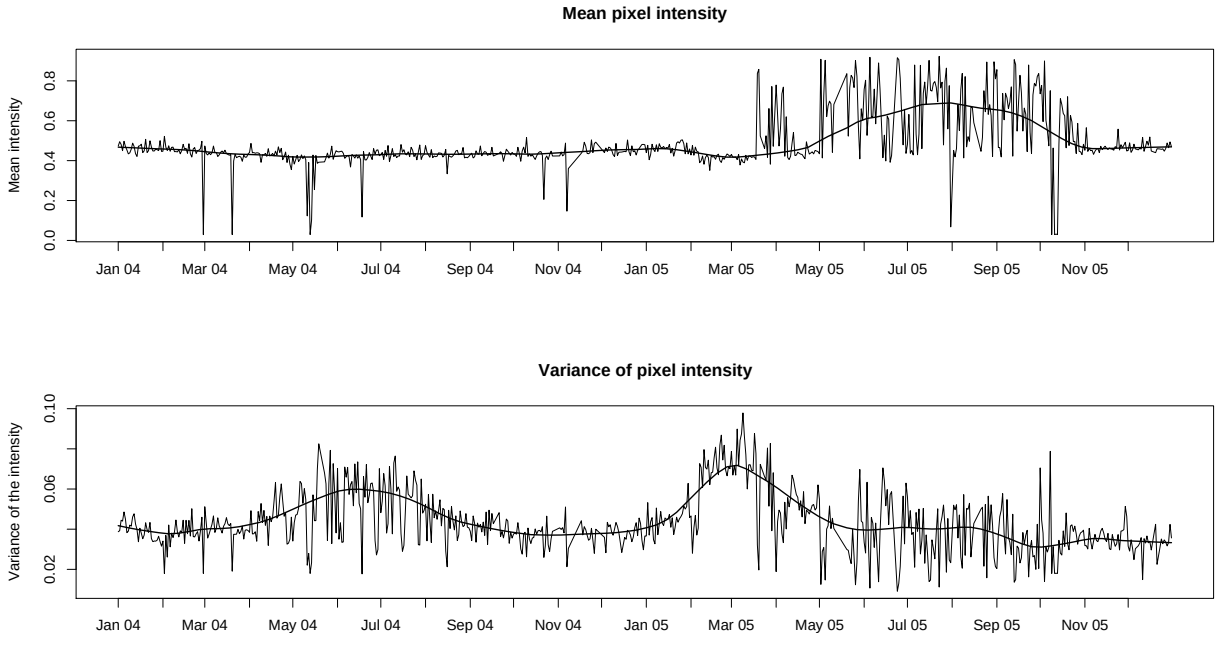


Figure 3: The mean and variance of the pixel intensities over time.

First, considering the mean pixel intensity we can see that in general until about April 2005 the values are around 0.4 with very little variation. The exception being the 14 images with average intensity less than 0.3. These images were found to be blank images and are dropped from any further analysis. The strange behaviour occurs at about April 2005. The mean intensity varies wildly with some very high values. These images were investigated and two probably connected phenomenon were discovered. The first occurs between noon on the 6th and noon on the 7th of February 2005. The images for these times are shown in Figure 4. We can see the camera has been knocked out of position and now primarily looks at a tree

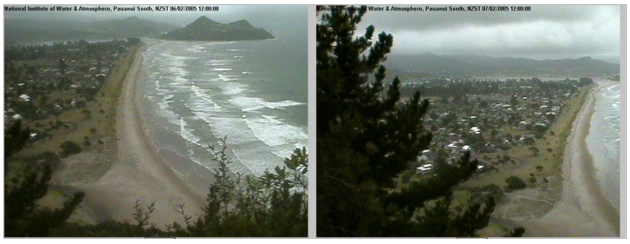


Figure 4: February 6th 2005, February 7th 2005

(this event will be referred to as the "kick" in later analyses). The camera is not returned to its original orientation until the 1st of April 2005. However, the strange behaviour of the images does not cease. Some example of the images after the 1st of April are shown in Figure 5. It is clear from these images that the camera is damaged or malfunctioning.

Figure 5: April 1st, April 8th, April 13th and May 5th 2005.

Now, considering the variance in the intensity of the pixels, we can see this strange behaviour is reflected here too. There is also evidence of some seasonality. The variance of

the pixel intensity appears to be higher during the winter months (remember we are in the Southern Hemisphere) as well as much more variable. Of course with only one year of usable data it is hard to verify this happens every year.

It is slightly surprising to see such constancy in the mean intensity. However, when thinking about the image taking process it is explainable. The camera performs some natural preprocessing by choosing the exposure at which to take the picture. It is easy to imagine the the camera aims for some set average intensity. This explains why the mean intensity is roughly constant across seasons even though we would expect there to be more light during the summer months.

It seems harder to explain the higher variance in pixel intensity in the winter months. One would expect more "dull" days in winter where we would expect the images to contain less variation in intensity.

The exploratory analysis raises a few questions that will be covered in the next few sections. A quick exploration in the next section attempts to identify stormy days. Then the following section investigates the mean and variance series in a little more detail in an attempt to formulate an algorithm for the detection of camera malfunction. The final section attempts to look at all the individual pixel series and reveal some commonality between them.

3 Stormy Days

Consider just the images collected before February 7th. The mean and variance in pixel intensity for each image in this period is plotted in Figure 6. Intuitively we might expect stormy days to be quite uniformly grey both metaphorically and quite literally in terms of image intensity. This would correspond to a very low variance in intensity across the image. Examining the plots in Figure 6 we can see there are some quite distinct troughs in the variance. Some of these correspond to blank images (very low mean intensities) but

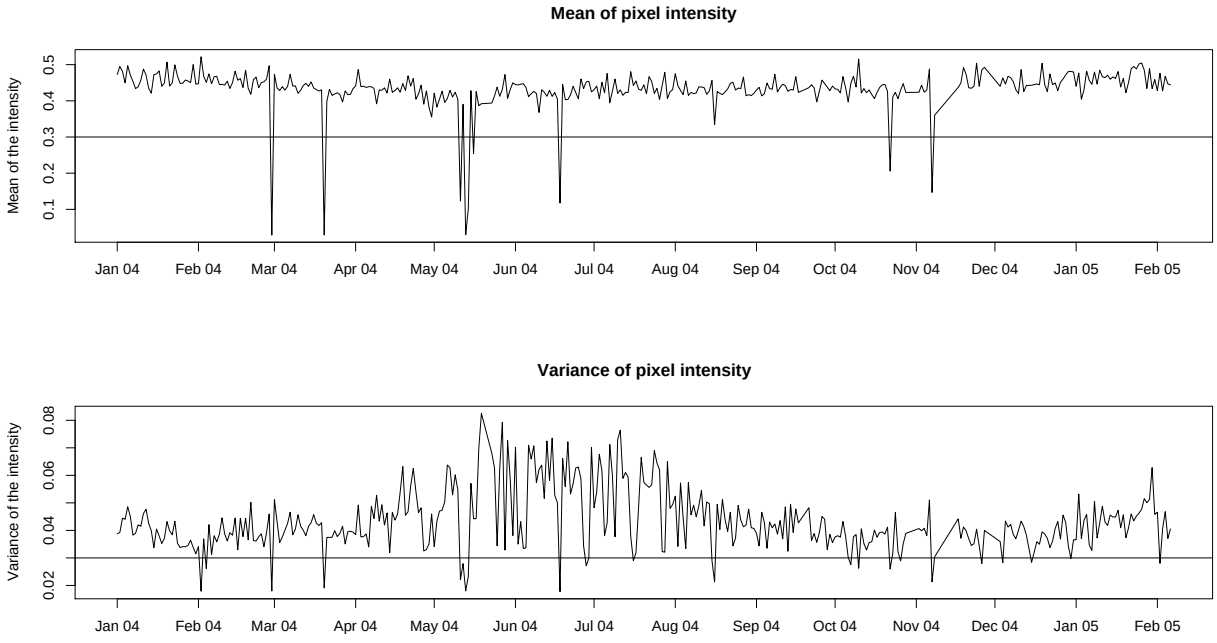


Figure 6: Cutoff values for stormy days.

we might expect the others to be stormy days. We pick the cutoff values of 0.3 in mean intensity and 0.03 in the variance of intensity. A day is classified as stormy if it has a mean intensity greater than 0.3 and a variance in intensity less than 0.03. Using these values 16 days are classified as stormy. The images from these days are shown in Figure 7. We can see this very crude (and slightly arbitrary) cutoff actually works quite well. Fifteen out of the sixteen images appear to be taken during bad weather. The one image that stands out is the one of static (August 16th 2004). So, this method shows good positive predictive value but without classifying every day by hand as either stormy or not stormy it is hard to evaluate its negative predictive value. However, it certainly seems possible that these simple measures could be enough to identify a stormy day (with maybe the addition of a simple measure of spatial correlation to eliminate static images).

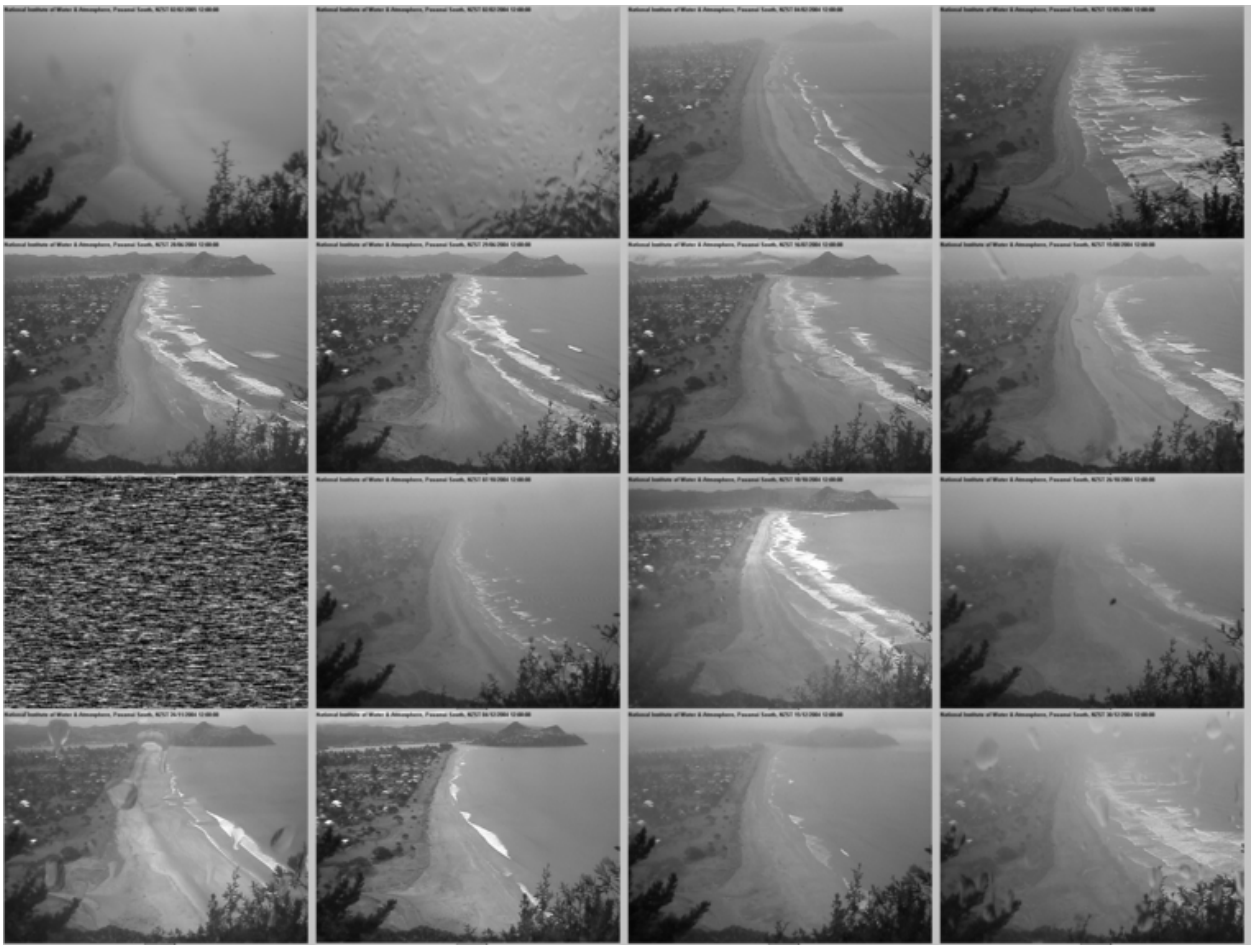


Figure 7: Images for 16 identified stormy days.

4 Malfunction Alert System

This section attempts to find an algorithm to detect the start of the camera malfunction. The nest subsection seeks to fill in observations for the missing images so that the mean and variance series can be more thoroughly analysed using traditional time series tools in the second subsection. The third subsection presents the algorithm and its results on this series.

4.1 Interpolating Missing Values

In the following section the series of the mean and variance of pixel intensity will be investigated more thoroughly. For ease of computation it would be nice be interpolate the values for the missing and blank images. In total over the two year period there are 50 days without images.

The simplest approach to fill in these values would be to replace them with the mean of the series. Since, in particular, the variance series shows some trend it was decided it would be wiser to fill in these values with smoothed values. The loess smoother plotted in Figure 3 is used to fill in the missing values. Figure 8 shows the two series along with points indicating the interpolated values. These series are now used in the following section to

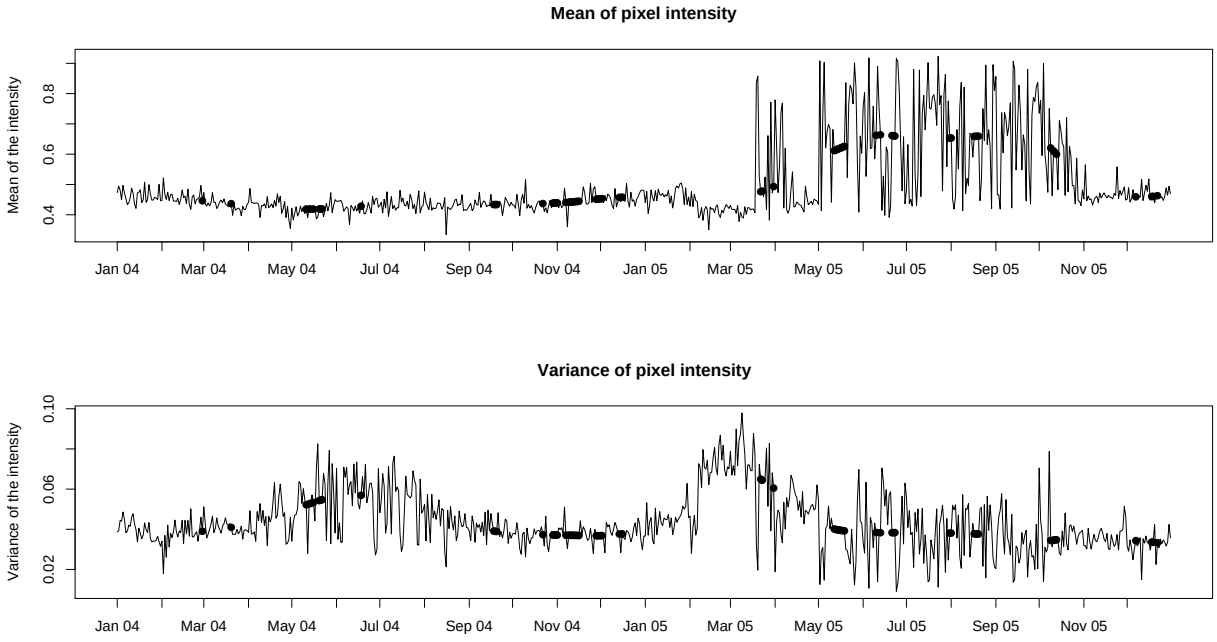


Figure 8: Mean and variance of pixel intensity with missing values filled in.

further examine the behaviour of these series.

4.2 Behaviour pre and post "kick"

With the intent of developing an malfunction detection algorithm we first investigate the properties of the series pre and post "kick". The series of the mean intensity is considered in detail through this section but investigation of the variance of the pixel intensity showed very similar results and the final results of the variance series are presented at the end of this section.

First the estimated autocorrelation function and the partial autocorrelation function for both pre and post "kick" series are shown in Figure 9. Remarkably both series show

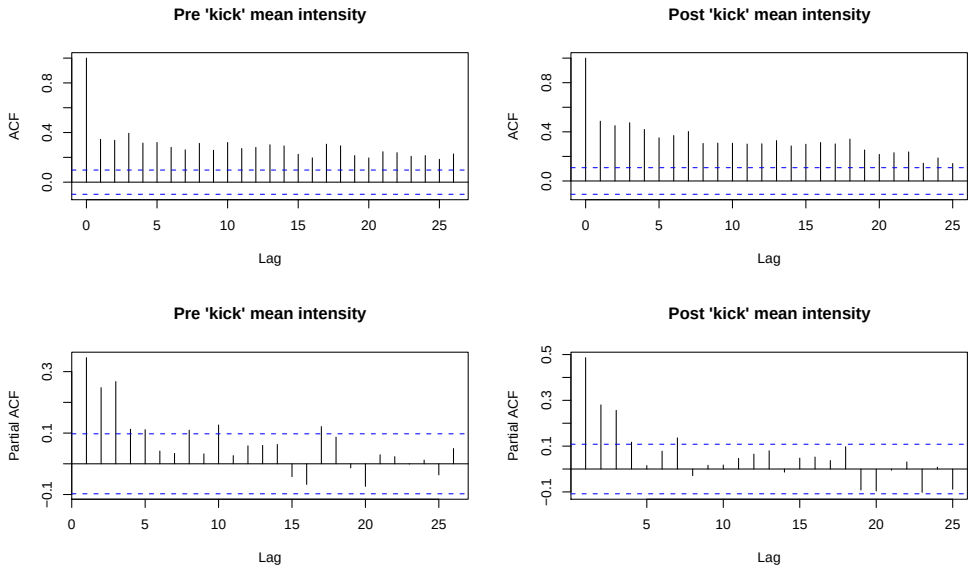


Figure 9: The estimated autocorrelation and partial autocorrelation functions for the pre "kick" and "post" kick periods.

very similar behaviour on these measures despite their apparent differences when plotted against time. They both show a slow decrease in the autocorrelation function and a sharp decrease after the first three lags in the partial autocorrelation function. The only real difference between the two series is the magnitude of the correlations: the post "kick" series having slightly larger correlations. The slow decrease in the autocorrelation function and

Table 1: Results of fitting an MA(1) model to the differenced series

	Pre "kick"	*Post "kick"*
Estimate of MA coefficient	-0.8947	-0.8028
Standard error of estimate	0.0247	0.0393
Estimate of σ^2	0.00047	0.01387

the evidence of a trend in the time series suggests taking first differences of the two series (Shumway and Stoffer [2000] Section 2.8). The autocorrelation function and partial autocorrelation function for the series after taking first differences are shown in Figure 10. The

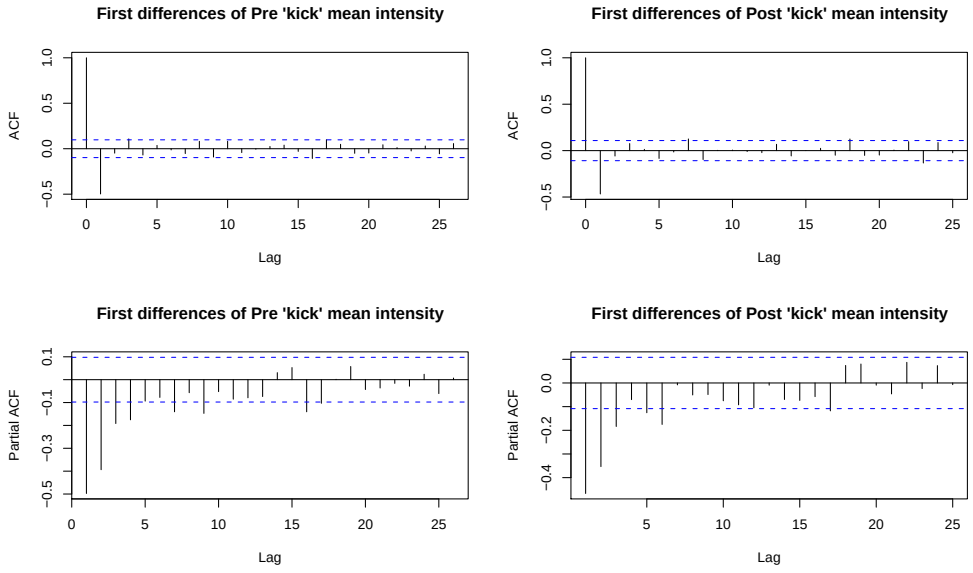


Figure 10: The estimated autocorrelation and partial autocorrelation functions of the first order differences of the pre "kick" and "post" kick periods.

two series are still similar. The autocorrelation function drops after lag 1 and the partial autocorrelation function slowly decreases. This suggests both differenced series would be modeled quite well as moving average series of order 1 (MA(1)) or equivalently the original series modeled as an ARIMA(0,1,1) model. This model is fit to both series. The results are shown in Table 1 The MA coefficients for both series are quite similar. The estimates for σ^2 however are hugely different. This gives some suggestion that this might be a good way

to distinguish between the behaviour of a series of normal images and a series of corrupt images. The next section suggests a procedure for doing this.

Diagnostic plots for the pre "kick" series are shown in Figure 11. The standardized residuals show no obvious trends. The estimate of the autocorrelation function shows no evidence of correlation between the residual. The Ljung-Box statistic gives no evidence against the residuals being white noise. Therefore, we judge the model fit to be good. The same diagnostics were plotted for the post "kick" series and no evidence of lack of fit was discovered.

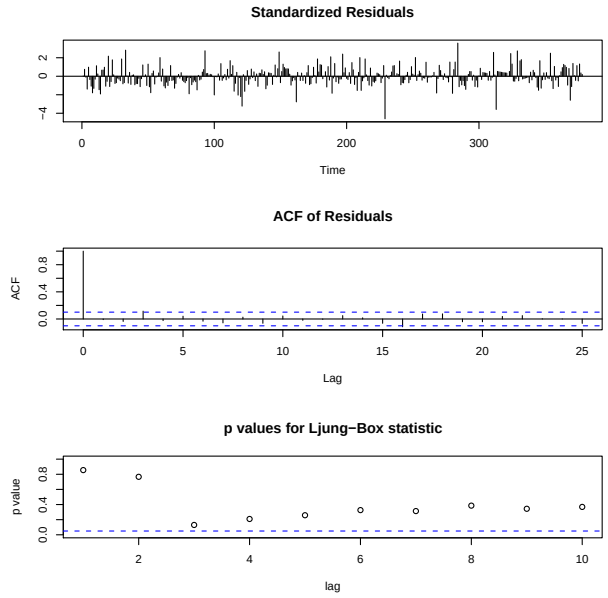


Figure 11: Diagnostic plots for the ARIMA(0,1,1) fit to the mean pixel intensity pre "kick".

4.3 Dynamic fitting approach

Since both the pre and post "kick" series seem well modeled by an ARIMA(0,1,1) we consider a dynamic fitting approach. We suggest fitting an ARIMA(0,1,1) to the last 50 observations and examining the fitted MA coefficient and the estimate of σ^2. This procedure is followed for the entire series and the estimates are plotted in Figure 12. The estimates show some

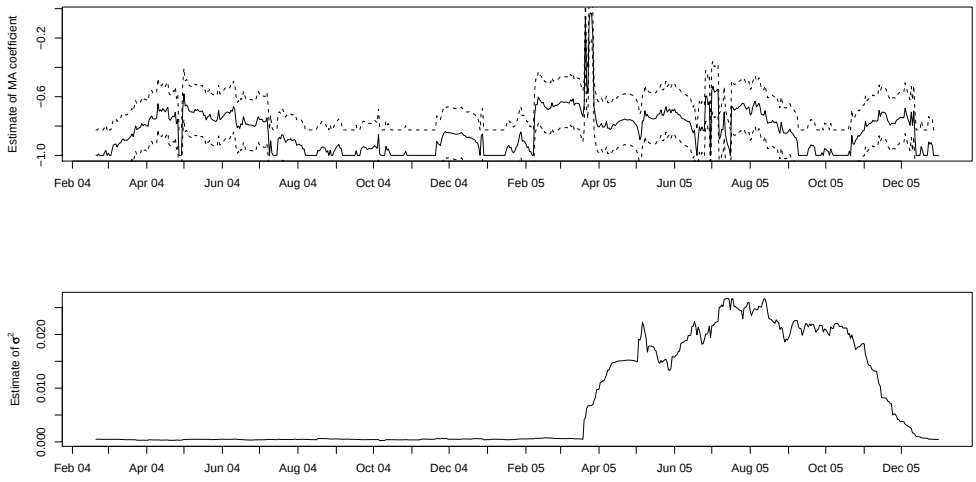


Figure 12: The estimated MA coefficient and σ^2 over time for the mean series.

desirable properties. The estimate of the MA coefficient jumps at the time the camera changes orientation and falls when it is returned to its original position. The estimate of σ^2 jumps when the images start to be corrupted. Unfortunately, since we are fitting over 50 values these jumps lag about one month behind the actual events.

An ARIMA(0,1,1) also models the variance in the pixel intensity well and the same procedure is attempted with this series. The results are shown in Figure 13. There is no clear indicator of the events we are searching for so this series does not seem suitable for malfunction detection.

(Note: I realised that perhaps a much simpler way of doing this would be to come up with a measure of spatial correlation in the image. A "good" image would be expected to have large areas of pixels that have intensities that are highly correlated to their neighbours. And in fact, this pattern of spatial correlation would not be expected to change that much over time (particularly in areas of open water or land). Whereas a corrupt image would not have the same degree of spatial correlation. This would also have the advantage that the jump would be instantaneous. I didn't have enough time to try implementing this.)

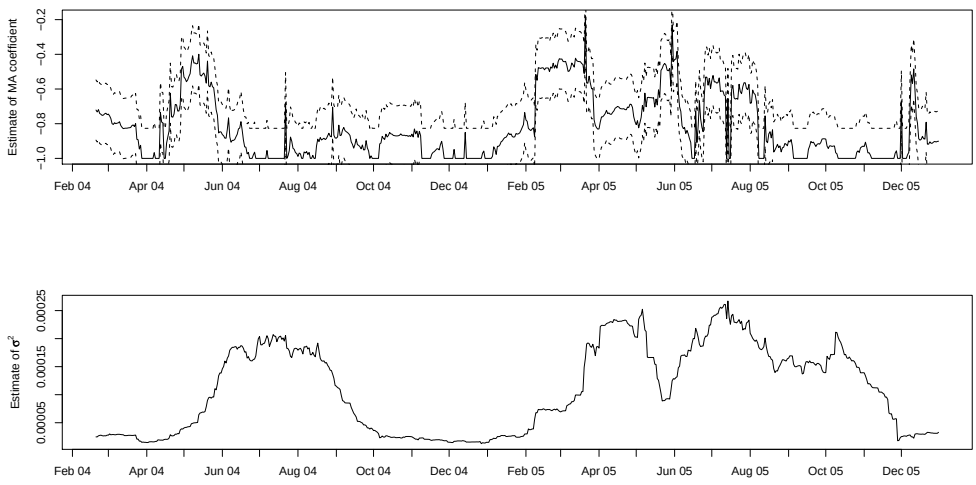


Figure 13: The estimated MA coefficient and σ^2 over time for the variance series.

5 Modeling Variability

This section attempts to explore the different types of variation amongst the pixels. Consider each pixel as having a time series (over only the useable images) of intensity. We want to answer the question: which pixels show similar patterns in intensity over time? Since we are primarily interested in the patterns in these series rather than their differences in mean intensity we first standardise each series to have mean zero and variance one. We will attempt to explore the variation in these series using functional principal component analysis. A brief description of this method is given and then the method is applied to the pixel intensity series.

5.1 Functional Principal Components (FPCA)

In standard multivariate principal components analysis (PCA) we seek to uncover directions in the observation space on which the variability in the data is maximised. These directions are the principal components and are defined by a weight vector. The inner product of an observation and the weight vector gives a value known as the score on the principal

component.

In FPCA we consider the observation space to be a function space and the principal components are defined by weight functions. We can get scores for each pixel by integrating their intensity function against the weight function (an inner product in the functional world). Ramsay and Silverman [1997] gives a more in depth discussion of the method. With the pixel data we consider the time series of the intensity as a function. We could simply interpolate the observed intensities to approximate this function but it turns out to be more desirable to preform some smoothing first. This is discussed in the next section.

5.2 Applying FPCA to the pixel intensities

The individual pixel intensity series are quite noisy so we reduce this (and the dimensionality of the problem) by first approximating each series using a Fourier basis. We consider two bases: one of size 150 and one of size 50. An example of the approximations using these two bases is shown in Figure 14 where a individual pixel's series is plotted along with the Fourier approximations. Both appear to do quite well although obviously the basis of size 50 results in a much smoother approximation. We can now think of each series as a point in 150 and 50 dimensional space respectively. Functional principal components analysis was preformed on both approximations. The results were compared and the analysis based on a Fourier basis of size 50 was much more interpretable (the approximation based on 150 terms still contained too much noise). The results for that fit are now presented.

5.3 Results

Figure 15 plots the variance explained by each component. The "elbow" in this plot appears to be around the third principal component. We therefore assume that components four and above are probably capturing mostly noise. We will now examine the first three components

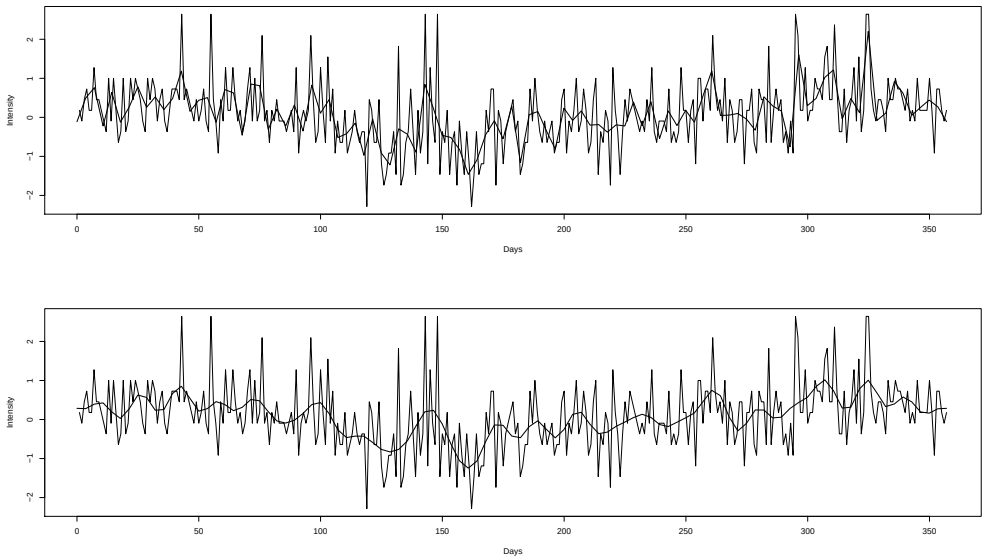


Figure 14: Example of the first pixel's raw series (thin lines) and its projection onto the Fourier basis with 150 terms (thick line on top plot) and its projection onto a Fourier basis with 50 terms (thick line on bottom plot).

in more detail.

Figure 16 summarises the weight function of each principal component. The solid line represents the mean of the series. The "+" represent the curve if a multiple of the principal component is added to the mean and the "-" represent the curve if a multiple of the principal component is subtracted from the mean curve.

The first principal component appears to extract the overall trend of the pixel intensity. We can see that an addition of this component results in a peak in the summer months and a trough in the winter. A subtraction of this component results in a peak in the winter month and a slight trough in the summer months. So, we expect pixels scoring highly on this component to have series with central troughs and those that score lowly to have central peaks.

The second component is harder to discern. It appears that a subtraction of this component from the mean results in more variability (the peaks are higher and the troughs

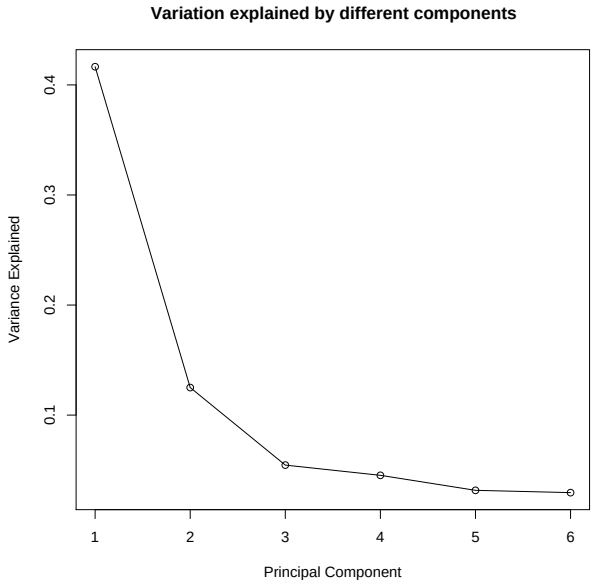


Figure 15: Screeplot of principal components

lower).

The third component seems to identify differences between the first and second halves of the series. Adding this component to the mean we see that it results in a higher curve until about day 170 and then it results in a curve lower than the mean. Subtraction of this component causes the opposite effect. Pixels that score highly on this component would be expected to have higher than average intensities in the autumn and lower than average in the spring.

The Figures 17, 18 and 19 (pages 20, 21 and 22) help us understand what each principal component is capturing. The first plot on each page plots the score of each pixel. This tells us what part of the image scored highly on this component and which parts score lowly. The second plot shows the time series of the pixel intensity for a sample of 100 points from the image. These points are chosen using the regular grid to ensure a representative sample from the image. These time series plots help us determine what type of pattern in the time series the principal component is identifying.

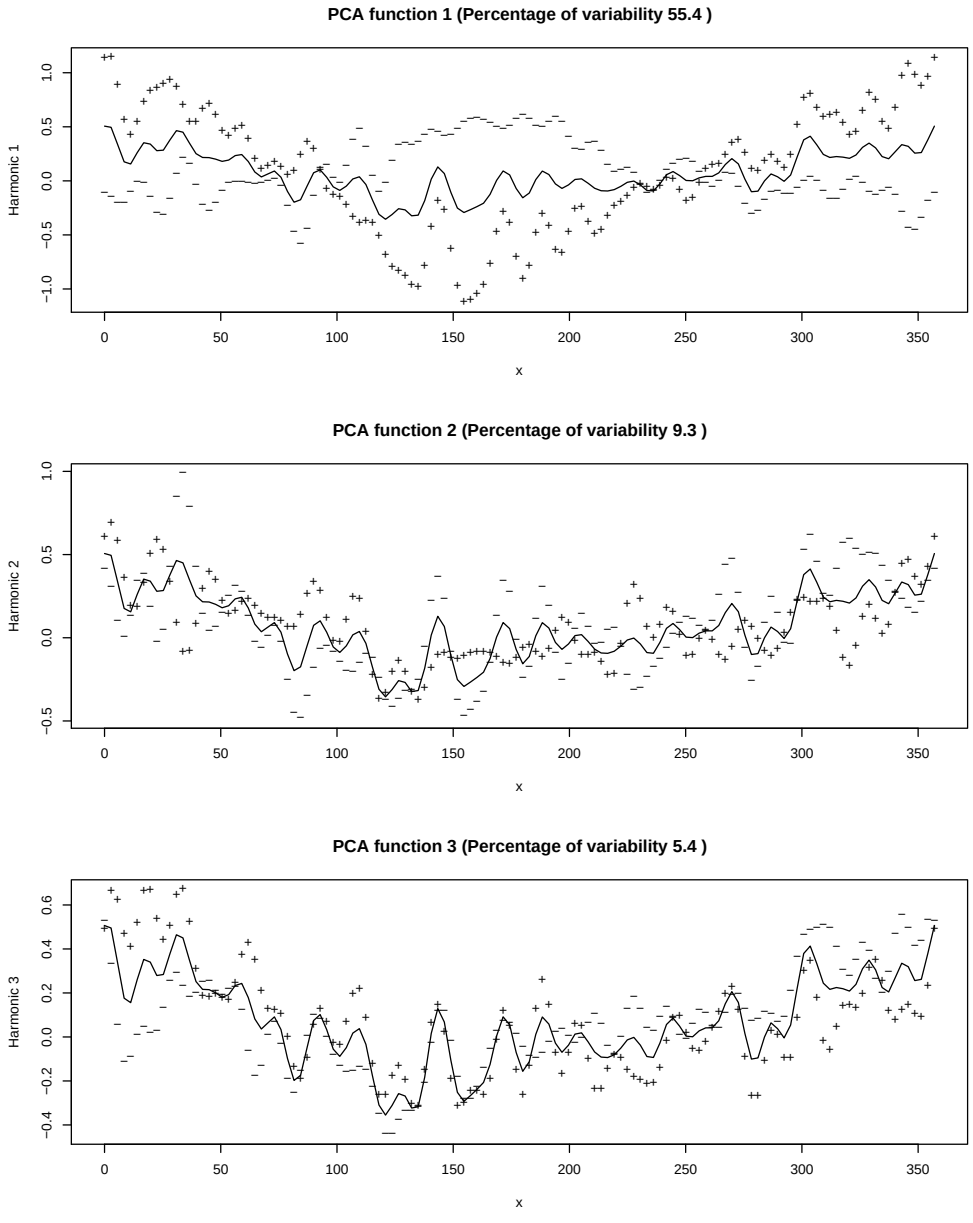


Figure 16: Summary of principal components from fit.

The first principal component seems to split the image into two: the sea which scores low on this component, and the land which scores high. From, the time series plots and the previous summary plots this tells us that the pixels on land tend to have a dip in winter and the sea areas have a peak in winter. Perhaps this can be explained by the lower position of the sun in the sky in winter which may result in more reflection off the water but lower light levels in general.

The areas identified with high scores on the second principal component are the shoreline and much of the ocean and sky. Any vegetation seems to score very low on this component. The low scoring areas do seem a lot more variable (looking at the bottom panel of the time series plots).

The wave break area scores high on the third principal component. Much of the developed land also scores high. The beach itself scores low on this component. The pixels plotted in the time series that score high (top panel) appear to have a slight downward trend.

6 Conclusion

The discovery of the corrupt images in the data provided some surprise and further challenges.

Stormy days were found to be well described by having low pixel intensity variance (but not being blank).

An algorithm that dynamically fits MA(1) models to the last 50 observations of the mean pixel intensity was found to show dramatic shifts in the estimates when the camera started malfunctioning. This provides a starting point for a detection of malfunction system.

The functional principal components analysis provided some interesting insight into the patterns of variability across the image. It is interesting how well the first component divided the sea from the land. One of the most interesting findings from this section was the peak

Figure 17: Functional Principal Component One

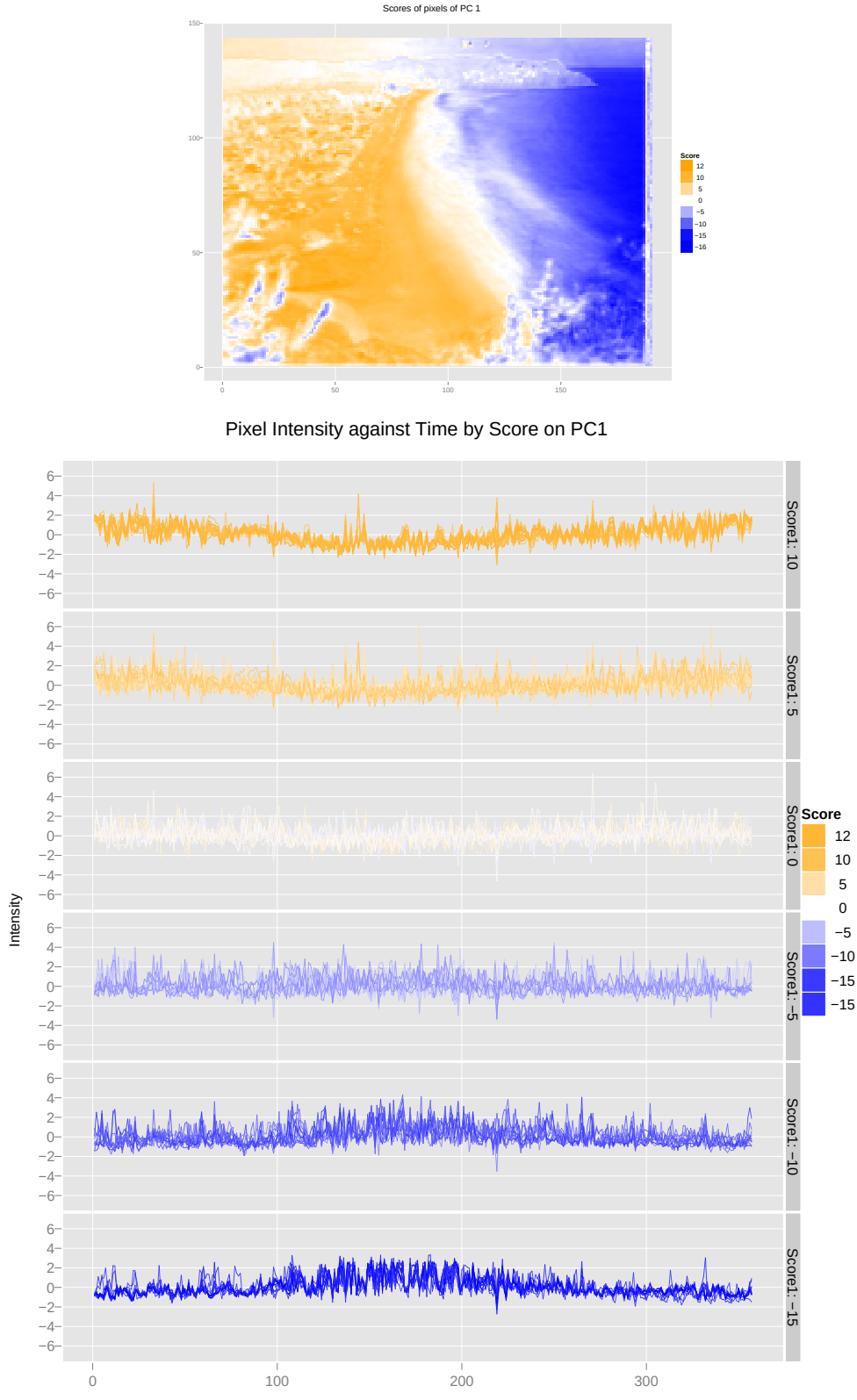
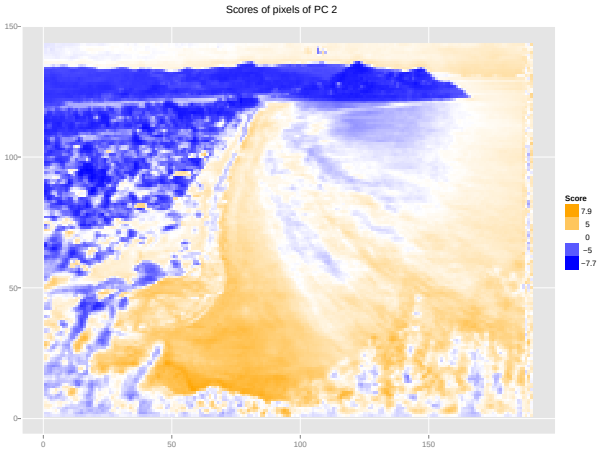


Figure 18: Functional Principal Component Two



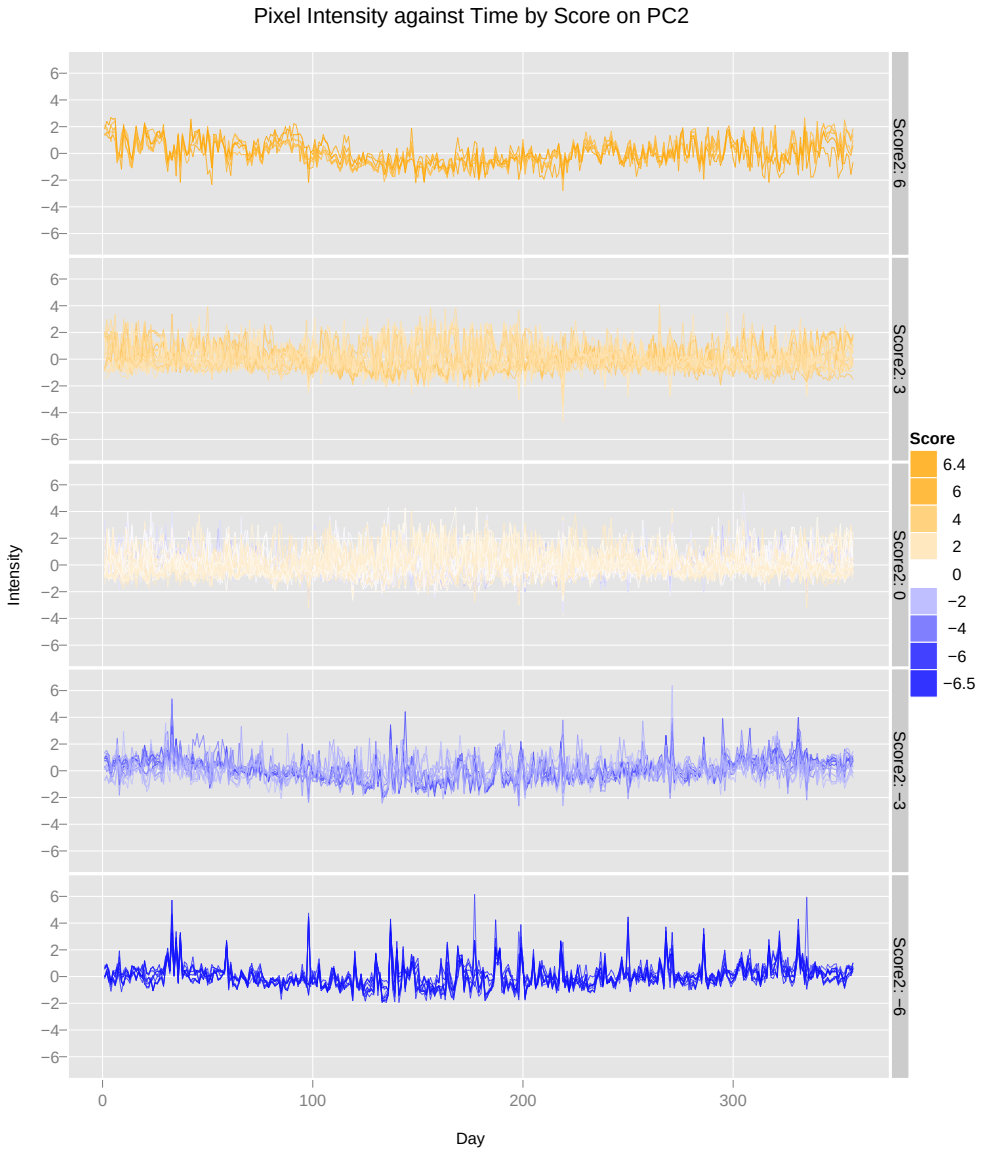
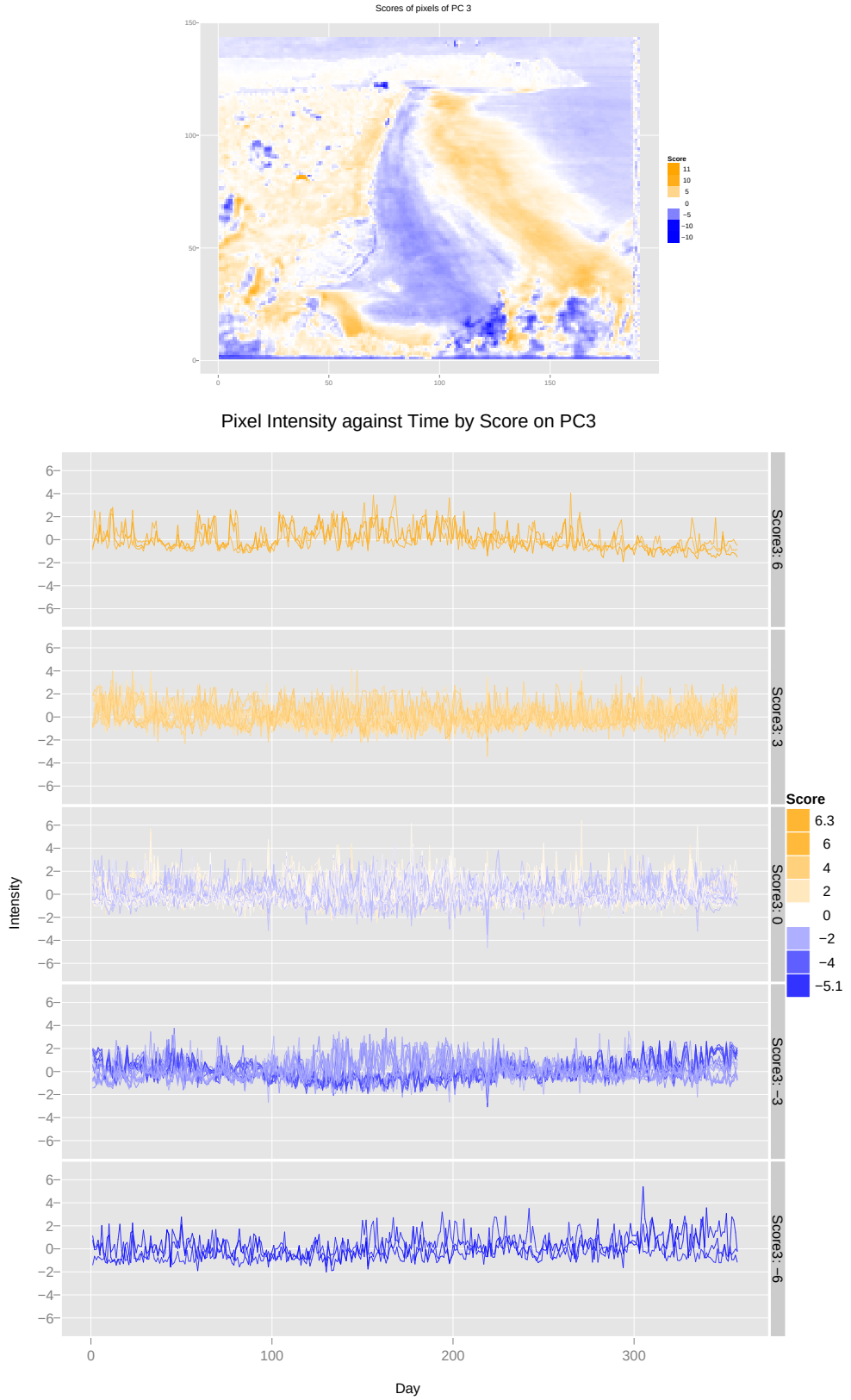


Figure 19: Functional Principal Component Three



in winter of intensities across the sea and the dip in those on land.

6.1 Ideas for future work

The results of the last section went some way to explain the different types of variation in the individual pixel time series. It might also be interesting to perform a similar analysis but perform the PCA on the spectra of the individual series. This may highlight the areas of the image that vary periodically at different frequencies.

All the analyses in this report were performed on greyscale images. It might be interesting to extend these analysis to deal jointly with the red, green and blue channels. Perhaps the proportion of blue in an image is a good measure of the amount of sea and sky.

It would be interesting to collect more data to more fully explain the images. For example weather records and high tide records for the period observed. Then perhaps steps could be made to find relationships between the images and these other records.

This report only scratches the surface of the possibilities of exploring the information contained in these images.

7 Some Technical Details

The image manipulation (resizing, converting to greyscale, cropping, and moving between formats) was all implemented in a command line tool called Imagemagick. The images were analysed in R (R Development Core Team [2006]) using the pixmap package (Bivand et al. [2006]) to read in the images. The tools in the package fda (Ramsay and Wickham [2006]) were used for the functional principal components analysis and the package ggplot (Wickham [2006]) for the figures in the corresponding section.

References

Roger Bivand, Friedrich Leisch, and Martin Mchler. *pixmap: Bitmap Images ("Pixel Maps")*, 2006. R package version 0.4-5.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Ramsay and B. Silverman. *Applied Functional Data Analysis*. Springer-Verlag, 1997.

J. O. Ramsay and Hadley Wickham. *fda: Functional Data Analysis*, 2006. URL `http://www.functionaldata.org`. R package version 1.1.6.

Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.

Hadley Wickham. *ggplot: An implementation of the Grammar of Graphics in R*, 2006. R package version 0.4.0.