MWZ station 10 HHZ exploratory analysis

David Hood

2023-05-23

This draft replaces the May 11th 2023 draft. It is a rewrite to add clarity and explanation about the steps used and reasons, but does not change the conclusions.

Motivation

There are significantly more (56.09% of 100,279) earthquakes at night in the Geonet New Zealand earthquake catalogue. While I have ruled out a number of explanations based on correlations between the catalog and other sources, I have been advised to check the seismograph readings themselves rather than the interpreted values.

A particular, discussion focus was on the STA/LTA calculation for determining likely earthquakes, where STA/LTA is

$$\frac{mean(squaredAmplitudeSamplesFromShortTermWindow)}{mean(squaredAmplitudeSamplesFromLongTermWindow)}$$

With the unusual high signals of earthquake events being more dominant in the short term time window when compared to the long term time window where background noise is more dominant.

If daytime noise is high enough it could be lowering the ratio (via raising the denominator in proportion to the numerator) enough to miss earthquakes give a similar rate of events in the day and night.

If raised noise is not the cause then large signal readings should actually be occurring more frequently at night. This acts as a test of the consistency of the noise explanation by testing what should be a predictable outcome of the noise explanation.

Matawai/ MWZ

Matawai/ MWZ is being looked at in detail, as the first seismograph considered (HAZ) showed a moving skew in the overall distribution through the day (see Appendix 1), and Matawai is near to an Electronic Weather Station that records hourly temperature for examining that phenomena. The two stations share the features of being located near to the East Cape, which has a very high proportion of recorded nighttime earthquakes (58-60% compared to the national average of 56%), the seismographs recording more extreme events at night, and the seismograph showing a daily moving skew in the distribution of recorded entries.

Specific Terms

For this I am using a few specific terms

- *Amplitude Distance* for the size of amplitude movements on the seismograph measured as the absolute value distance from the median of the 10 minute period the reading took place in. Standard calculations

are based on the raw value, but with the offset generated by the skew I wanted to focus of the size of the difference from normal of readings.

- *Extreme events* are amplitude distances so large that they occur at a rate of less than approximately 1 in 10,000 readings (100 in a million) in the data from this seismograph. The level is so high to be clear of all potential human noise at lower readings. Because of binning the data into steps for easy of analysis, this seismograph was technically the slightly tighter criteria of 98.2 in a million for amplitude moves of $10^{4.8}$ or greater.

MWZ Matawai

Location 10, Sensor HHZ

Longitude: 177.528, Latitude: -38.334

Data downloaded as 10 minute intervals from 2010-12-31 to 2021-06-01 (UTC timezone) with 79 days of data unavailable. With 100 samples per second, there are 32,023,172,557 readings in total. Of those readings 15,958,217,096 are in day, 16,064,955,461 at night (where day is on or above the horizon at the centre of the 10 minute interval). As there are 60,000 readings in a 10 minute period where the seismograph did not go off line, that is easily sufficient to establish accurate summary values for each 10 minute period. As there are around 52560 10 minute periods a year for a year of full operation, more than a decade's of readings gives a large number of independent 10 minute samples.

Data processing

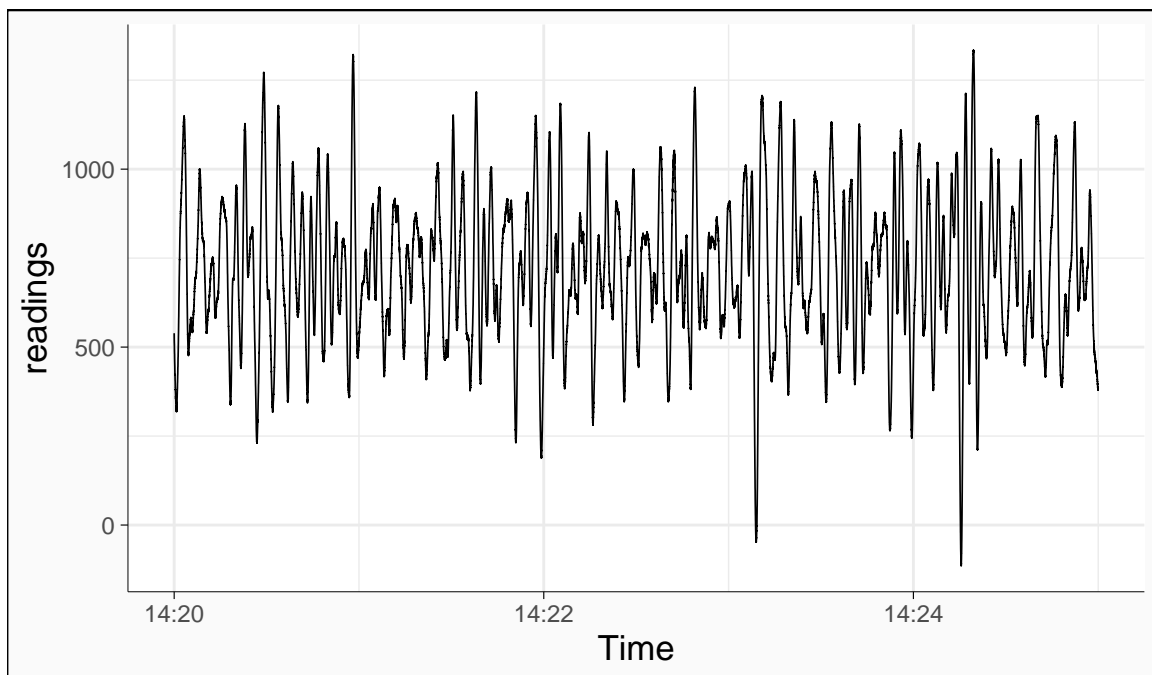A seismograph trace normally displays the amplitude readings over time (figure 01)

Figure 1: MWZ 2010-12-31 14:20 to 14:25 UTC

However, for a given number of entries, you can also express those entries by the frequency at which particular readings occur (figure 02).
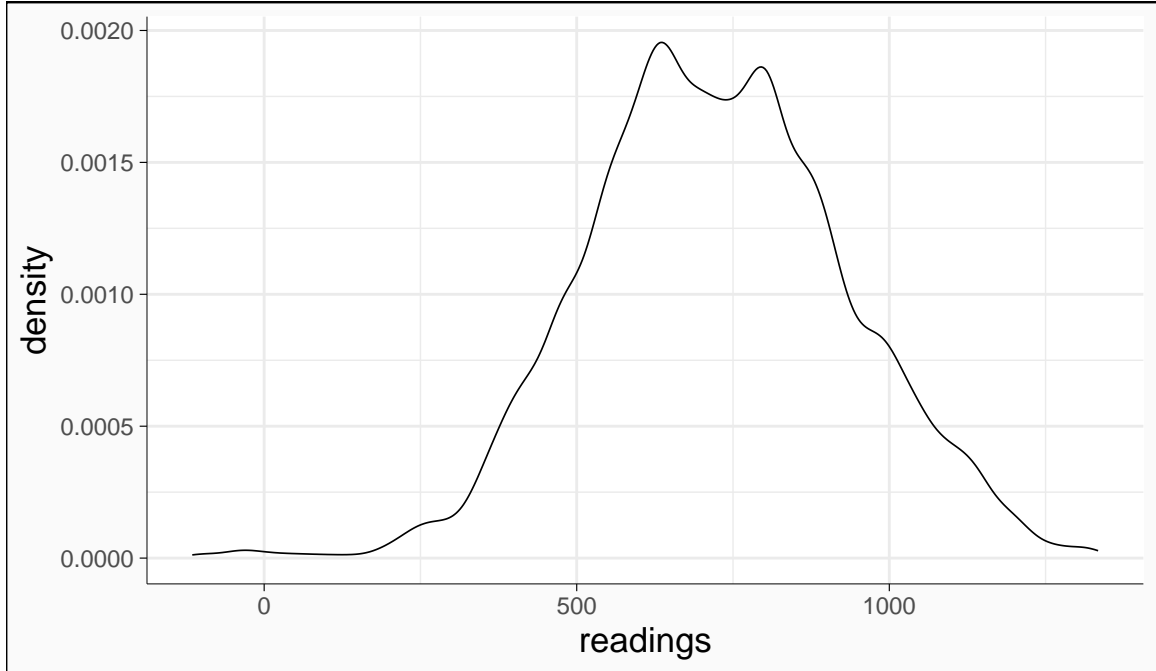
Figure 2: Density plot of previous (fig 1) data

This provides detail on the range of values in the given period. But as the centre of the distribution varies with daily temperate (Appendix 1), I am recentring the distribution on the median of the distribution. This avoids bias when comparing different times caused by the raw readings being centred at different distances from 0 at different times.

It is customary to square the readings to produce a distance from the centre of the amplitude, but I am using the absolute value as I want to minimise the squaring effects on extreme values in aggregation steps.

Among the 32 billion readings from this seismograph, the maximum values approach 7943282 amplitude away from the median. To make the amplitude distances tractable, I am converting to the base 10 log of the amplitude.

To make the 32 billion readings easy to hold and analyse in memory, I am placing the log converted readings into binned 0.1 steps.

As the number of entries in different bins and the number of day and night entries vary, I am making heavy use of average values per 10 minute unit, since increasing the sample size increases the accuracy of the average value (in a Central Limit Theorem square root of the increase in sample size kind if way). The trade off in using mean aggregates heavily is that if it is only a small subset of particular entries of interest, that specific combination is swamped by the overall aggregate.

Seperately, this data has also been linked with the NIWA Motu electronic weather station (Longitude: 177.52941,Latitude: -38.28566) 5.3 km away to compare its hourly temperature readings for investigating skew. 87,832 readings are available on the hour for both the seismograph and thermometer.

Sensor activity

For my original motivating interest, that there are more earthquakes in the catalogue that occur at night, I checked the number of extreme amplitude movements. Readings of that strength occur at an average rate per 10 minute period 24% higher in night than in day (6.52 per 60,000 readings at night vs. 5.26 per 60,000 readings in the day).

For the purposes of influencing the STA/LTA caluclation for earthquakes, as the amplitude strength of this imbalance is far stronger than noise levels it has more effect on the STA than the LTA. This effect is not more noise raising the LTA during the day and effecting the threshold calculation.

Noise

While I avoided the effect of noise on the determination of an earthquake by only looking at extreme events, having this much data can say some things about noise (where noise is the replacement/ masking of quieter movements by higher strength sources of vibration up to the maximum level of noise production, mostly assumed to be during the day).

The aggregate pattern caused by increased noise is expected to be lower readings of quiet (readings central to the distribution), higher levels of readings of noise in the range of the noise production (which replace what would otherwise be lower readings), then unaffected above that. This means I can compared distributions at different times from the same site to gauge the relative effects of noise between those times (figure 3).
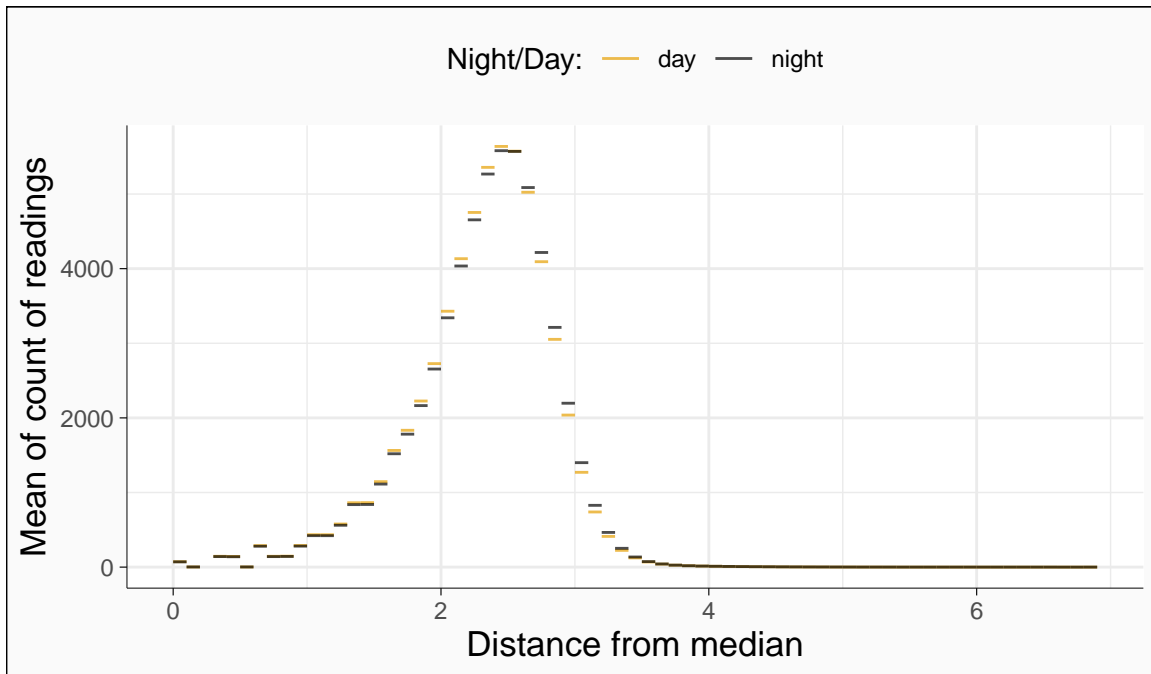

Figure 3: Average count of readings per 10 minutes by distance (log10(Size of amplitude movement from median + 1)) from median

The daytime and night range, and frequency of that range, is fairly similar, and while daytime events do occur at a higher rate up to 2.5 (moves of $10^{2.5}$-$10$2.6), nighttime events occur at a higher rate above that movement bin. Because this change over is close to the peak in the data, this means that, broadly, daytime events are more common among movements in the lower half of the data and nighttime events are more common among movements in the upper half (not just the upper 1/10,000th extreme values).

This is not the expected pattern for noise under the assumption of the same distribution of day and night earthquakes. There, daytime noise is replacing quieter values with larger movements giving a pattern of nighttime dominance in the smallest ranges, followed by an artificial inflation of daytime values in the medium ranges, followed by a similar rate for both in the extreme ranges unaffected by noise.

Because this is not the expected pattern, and the largest amplitude counts on the graphs have 1,000 times the readings of less frequent steps so it is hard to see the detail, I thought it worth exploring the imbalance

at each individual step of amplitude distance relative to those movements in the alternative cycle, in order to better see the relative distances (figure 4).
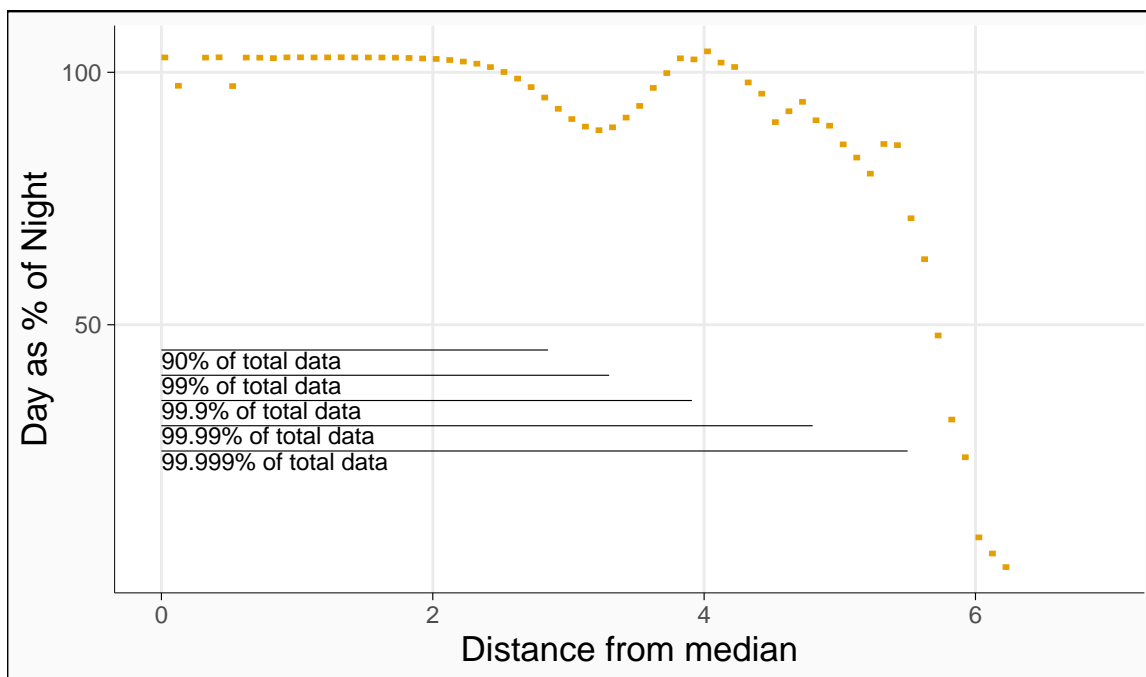

Figure 4: Mean day event rate as percentage of mean night rate by distance (log10(Size of amplitude movement from median + 1)) from median. Events of 6.3 or higher removed due to sparseness

View as the per amplitude bin difference, there is a rise in daytime events (following a period of night dominance at lower levels of activity) peaking at 4. If this is noise, then as the volume of data is reducing the rise is insufficient to account for the swing to night at low levels of activity (nor the continuing trend afterwards). However, if it is noise that un-masked would cause noise entries to otherwise be to the left of their measured position, that would create a steadier gradient in the progression to more extreme events at night.

Appendix 1, Skew and temperature.

When initially analysing data from the HAZ/Te Kaha seismograph, I noticed there was a diurnal and seasonal (but not weekly) move of the centre of the entire distribution. This differs from variation in the earthquake catalogue (which is diurnal, a small component of weekly, and not seasonal), but I thought needed to be investigated.

Checking MWZ/ Matawai it shows the same pattern (figure 5), and had a close by NIWA weather station.

With the suggestion that the skew is caused by temperature, I am plotting the same distribution of centres for those entries that have EWS temperature readings taken at the same time as the seismograph readings.

From that, it is absolutely clear that as the air temperature gets above 20 degrees the seismograph skews, and almost all of the above 20 degrees readings are in the day. So that is the reason, but the question open as to if the skew is influencing results.

Restricting the temperature range to only those entries in the 1 to 20 degree operating range increases the imbalance in extreme events from the nighttime rate being 24% above the day rate to the nighttime rate being 35.6% above the day rate. If removing the values in the skewed region of the data increases the differential, it is unlikely that the skewed region is causing the differential.
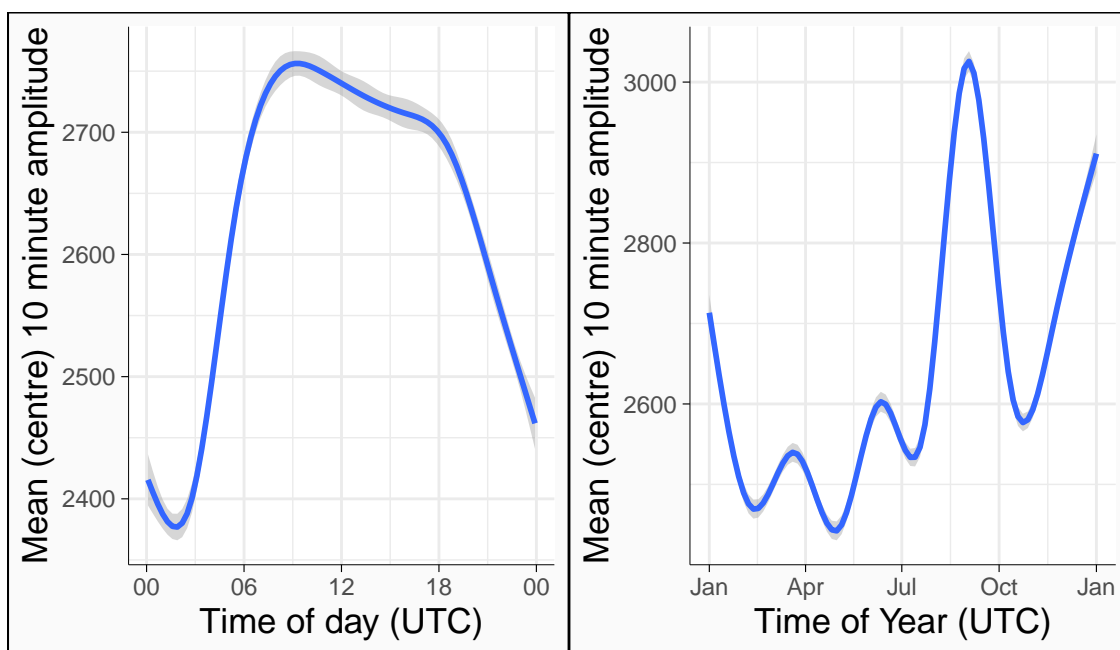
Figure 5: (a) Daily (UTC) and (b) Monthly mean of 10 minutes of readings. GAM line of best fit
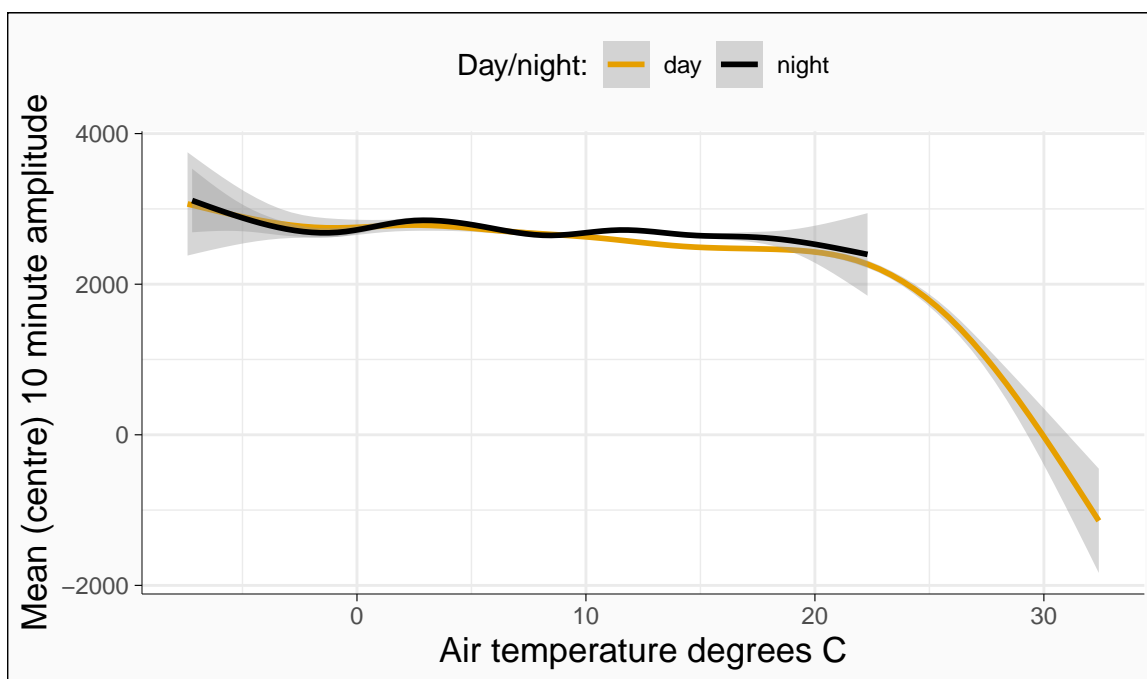

Figure 6: Center (mean) of 10 minute observations by Air temperature.GAM line of best fit

The reason excluding the above 20 degree values has such a dramatic effect is that the majority of daytime large amplitude events occur during the afternoon, and afternoons are the time when most of the above 20 temperatures are (which is also the time when the sun as a gravitational source is travelling through an arc in line with the subduction zone, as is the case with the main time of nighttime extreme events).