Draft version February 11, 2023
Typeset using LaTeX **twocolumn** style in AASTeX631

Correlation between High Fluctuations in Temperature and Mortality for Various Classes of Ages in Rome

Fontana Aleksandar

Abstract

Climate change is one of the most important topics in our society and it is also important to understand its impact on our lives. In this study, I examine the correlation between high fluctuations in temperature and mortality for different classes of ages in Rome. I have used two different data sets: the first, which refers to mortality for different classes of ages (≤ 64, $65-74$, $75-84$ and ≥ 85 years); and the second, which refers to the mean temperature, and I study the period from January 1^{st} 2011 to December 31^{st} 2019, with a sampling period of 1 day.

I found a linear correlation between temperature and mortality for all age groups with lags of 0, 3, 4 and 5 days. Both hot and cold temperatures are correlated with mortality, and I found that February, March and April are the periods usually more correlated.

Keywords: Correlation — Rome — Temperature — Climate change — Global warming — Mortality — Fast Fourier Transform — Continuous wavelet analysis — Empirical Mode Decomposition — Pearson correlation coefficient — Maximal information coefficient

1. INTRODUCTION

Climate change is one of the most important topics in our society. Since the 1800s, as humans, we have been the main contributor to this change, primarily due to the burning of fossil fuels such as oil and gas 1. The consequences of climate change include different things, where warmer temperatures are only one of this; there is also water scarcity, severe fires, rising sea levels, melting polar ice, catastrophic storms, and others Hardy (2003) Denman (2008). We are trying to reduce the impact of our activities on climate change, but at the moment, we are doing it very slowly. It is seen as a problem that we will face in the future, but some effects are also visible now. Climate action requires significant financial investments from governments and companies. But climate action is much more expensive. As consumers and citizens of developed nations, it is also our job to push nations to take the right steps in the right direction.

In this study I examine the correlation between high fluctuations in temperature and mortality for different class of ages. Already many studies have shown that hot and cold temperatures had effects on increased deaths Guo et al. (2012), but none of them have shown it in Italy.

2. DATASET

In this study we have used two different datasets: the first one, which refers to mortality in Rome, has been downloaded by the "Istituto Nazionale di Statistica" through their website; and the second one, which refers to the temperature, has been downloaded by the weather station " Roma Ciampino" through the website "IlMeteo.it". During the period from January 1^{st} 2011 to December 31^{st} 2019 I obtained a time series of mortality for every different class of ages in Rome, with a sampling period of 1 day. Over the same period, I also obtained a time series of the mean temperature in Rome, with the same sampling period. I present a slice of the datasets that I am using in this experience in Fig. 2 2.

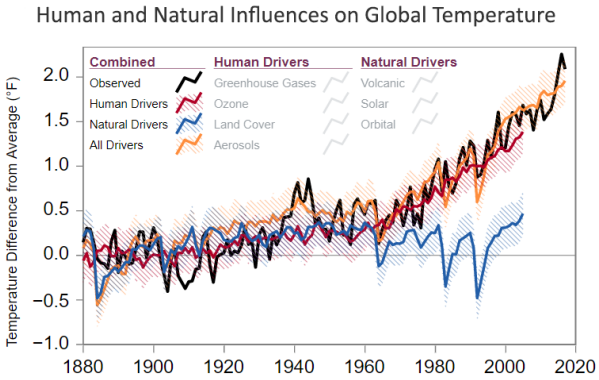


Figure 1. Credit: epa (2022)

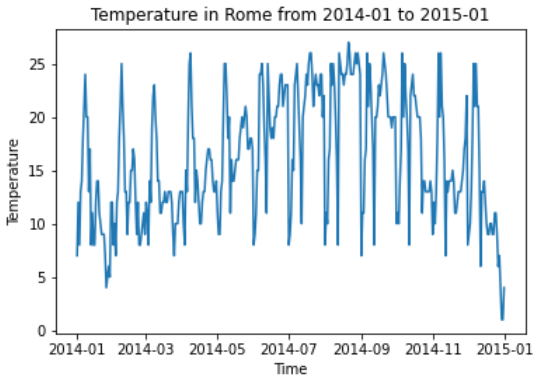


Figure 2. This is a slice of the daily mean temperature in Rome

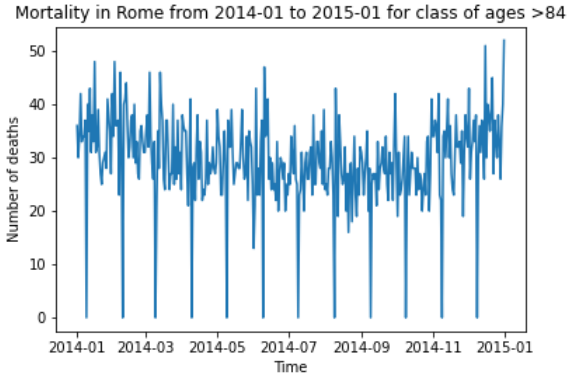


Figure 3. This is a slice of the daily number of deaths in Rome

The raw data obtained are then organized in a Pandas Dataframe; I resample the class of ages into only four (≤ 64, $65-74$, $75-84$ and ≥ 85 years).

3. METHODS

To carry out the intended analysis I need to detrend the datasets.

3.1. *Detrend with Empirical Mode Decomposition (EMD)*

The Empirical Mode Decomposition is a method introduced by Huang et al. (1998) for analyzing non-linear and non-stationary data. The decomposition is based on the local characteristic time scale of the time series, so the main characteristic of EMD is that the base is empirical and highly adaptive. Every dataset can be decomposed into multiple Intrinsic Mode Functions (IMF). By the nature of the decomposition, the last IMF refers to the trend of the dataset, so the last IMF is removed from the original dataset. In this way, the data set was detrendized without removing any oscillations. The result of such decomposition is presented in an energy-frequency-time distribution, the Hilbert spectrum. The main advantage of this method is that it is no longer needed to perform the convolution of the signal with some kind of assumed harmonic base functions but instead the introduction of IMFs allows to extract the actual instantaneous phase and frequency.

3.2. *Singular analysis of the two datasets*

Before to study the correlation of the two datasets it is also interesting to study their properties.

3.2.1. *Fast Fourier Transform*

The first approach is to use a fast Fourier transform and study the periodofran obtained. Due to the fact that the periodogram obtained by FFT of a time series is a biased estimator Scargle (1982), I refine the estimation using the routine described in Chatfield (2003) without using the Prewhitening phase. This phase consists of a detrendization that I have already done.

1. *Windowing*: To contain boundary effects due to the truncation of the signal, a Hamming window function is applied to the time series.

2. The FFT is computed.

3.2.2. *Wavelet analysis*

To carry out additional information from the study, I used wavelet analysis. The main difference between this analysis and the FFT one is that it can also be applied to time series that contain non-stationary power at many frequencies Daubechies (1990). In this analysis we convolve the time series with different scales of the same wavelet function. The wavelet function, called Mother Wavelet, is a function with zero mean and is localized in both time and frequency spaces. It can be orthogonal or not, in the first case it led to a discrete wavelet transform, and the second one led to a continuous wavelet transform. For the estimation of the scalogram we are interested only in the continuous wavelet transform and we will use the Morlet Wavelet.

Due to possible noise, I exclude from the analysis every oscillation with a period shorter than $\sqrt{2}$ the sampling period of the dataset. The Cone of Influence (COI) is the region of the wavelet spectrum where boundary effects become significant. For the Morlet Wavelet this is $\sqrt{2}$ the sampling period of the signal. I plot the spectrogram, which is the power spectrum of the continuous wavelet transform.

3.3. *Correlation*

3.3.1. *Linear correlation with Pearson correlation coefficient*

The Pearson correlation coefficient is a well-known measure used in statistics Cohen et al. (2009), it is also known as Pearson's r, and it is a measure of linear correlation between two sets of data.

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where: *cov* is the covariance; σ_X is the standard deviation of X; σ_Y is the standard deviation of Y; X and Y are the two datasets.

So, it is essentially a normalized measurement of covariance from -1 to +1, where +1 refers to a perfect linear correlation, but also -1 Fig. 4.

Due to the fact that Pearson's r is different from 0 only if there is a linear correlation, I need to linearize the possible correlation. To do this, I consider only the absolute value of the temperature oscillation because I expect a correlation to high mortality from high and low temperatures. If used Pearson's r to complete datasets it gives only a single value, it is useful to apply it to moving windows, to see if there are some region major correlated.

Another important aspect is that Pearson's r can be applied to datasets shifted from one to the other.

3.3.2. *Non-linear correlation with Maximal information coefficient*

The Maximal information coefficient (MIC) is a heuristic measure of linear and non-linear correlation Zhang et al. (2014). It uses binning as a means to apply mutual information on continuous variables. The rationale behind MIC is that the bins for both variables should be chosen in such a way that the mutual information between the variables is maximal. So, in a perfect scenario:

$$H(X_b) = H(Y_b) = H(X_b, Y_b)$$

where H is the entropy information.

Because MIC maximizes the entropy, the bins would have the same size; each bin of X will roughly correspond to a bin in Y. Because X and Y are real numbers, it is usually possible to create exactly a bin for each value in X and Y and thus maximize the entropy. To solve this problem, usually the number of bins for X (n_X) and the number of bins for Y (n_Y) need to follow this relation:

$$n_X * n_Y \leq (N)^{0.6}$$

where N is the size of the complete data sample.

In this case the result value is normalized to [0,1], where 1 corresponds to a perfect correlation. In Fig. 4 there is the comparison between Pearson's r and MIC.

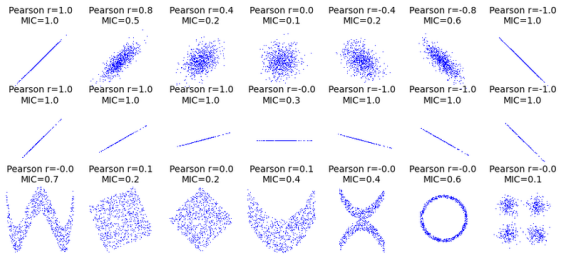


Figure 4. In this image there is a comparison between Pearson's r and Maximal information coefficient

4. RESULTS & DISCUSSIONS

Before discussing the results it is important to note that the temperature dataset is not complete. Due to the fact that only a few values are missing (7), I use a simple interpolation method to fill the gaps; this will not change the results of the study.

I now present the results obtained with the different techniques presented in the previous section. In this section are presented only the images that refer to the group of ages ¿84, beacuse is the one it is the most interesting one. The images that refer to the other groups of ages are in the Appendix.

4.1. *Empirical Mode Decomposition and detrend*

We can now proceed with the empirical mode decomposition of the signals in order to retrieve the IMFs and the trend of the signals Fig. 5.

Once I obtained the trend and the IMFs of the signal I plot the Hilbert spectrum of the two time series. Then I recompose the signals without the trend, as mentioned in the previous section I will continue the study on the time series detrendized

4.2. *Fourier Transform & Periodogram*

First of all I present the power spectrum obtained by computing the FFT of the signals and plotting the results. In Fig. 6 I plot the power spectrum of the mean temperature and it is possible to detect 5 peaks that refer respectively to oscillations of 9, 99, 117, 216 and 324 days. Probably the second and the third refer to the same type of oscillation; however, the last refer to an annual oscillation.

In Fig. 7 I plot the Power spectrum of the mortality (class of ages ≤ 64) and there is only an important peak at 9 days.

As previously anticipated, now I estimate the periodogram of the time series implemented in Python in two different functions: **Welch** and **Periodogram**.

In the periodogram of the mean temperature (Fig. 8) it is also possible to detect the 5 principal peaks of the

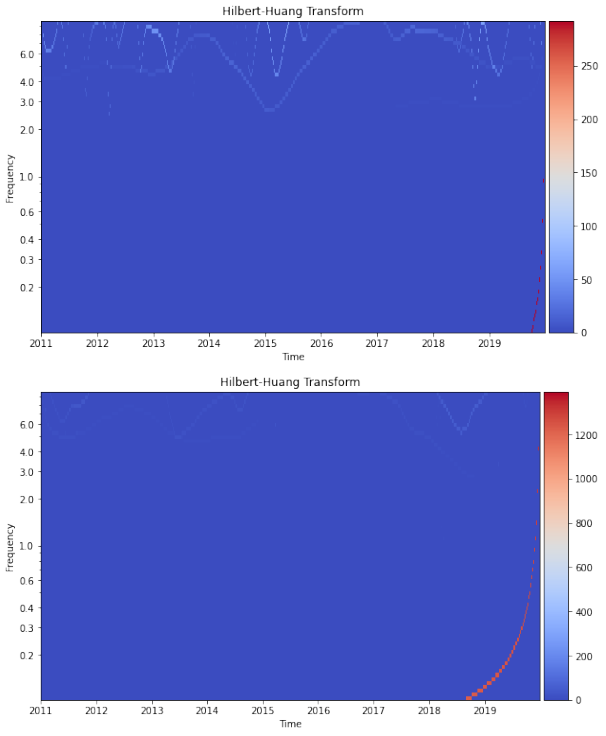


Figure 5. On top the hilbert huang transform of the mean temperature; on bottom the hilbert huang transform of the mortality for group of ages > 84

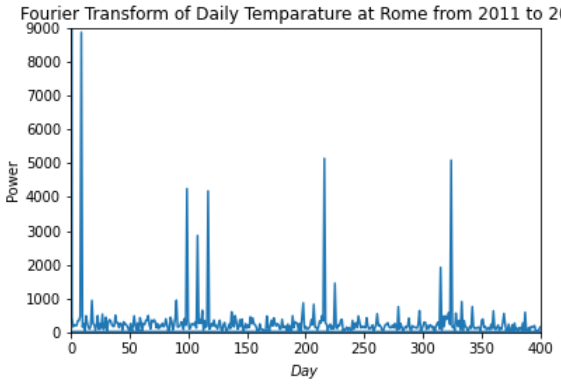


Figure 6. Fourier Transform of the daily temperature at a sampling frequency of 1 day.

Fourier transform. There is also a peak in the mortality periodogram (Fig. 9).

4.3. *Continuous Wavelet Transform & Scalogram*

Then I study the dataset with the continuous wavelet analysis. The datasets used are stationary, so the results presented in the scalogram (Fig. 10 and Fig. 11) are pretty much similar to the Fourier analysis; but there are also some small variations in intensity over time to the peaks that are not visible in the Fuorier analysis.

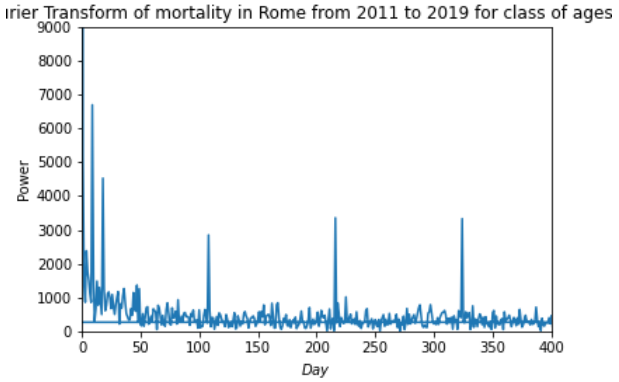


Figure 7. Fourier Transform of the mortaly for class of ages > 84 at a sampling frequency of 1 day.

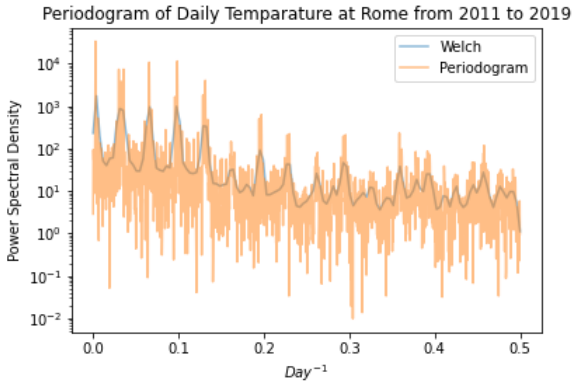


Figure 8. Periodogram obtained with both the Welch function and the Periodogram function in Python. The first peaks corresponds to the one's seen in the Fourier analisys

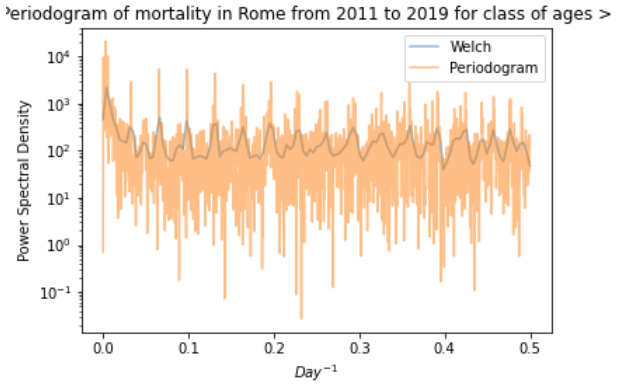


Figure 9. Periodogram obtained with both the Welch function and the Periodogram function in Python. The peaks reflect the ones seen in the Periodogram of Daily temperature

Obviusly the stationarity of the datasets is not strong.

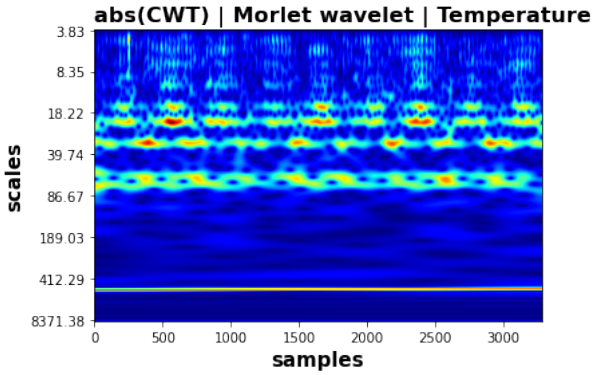


Figure 10. Scalogram of the temperature, the first peaks founds corresponds to the one's obtained in the Fourier study

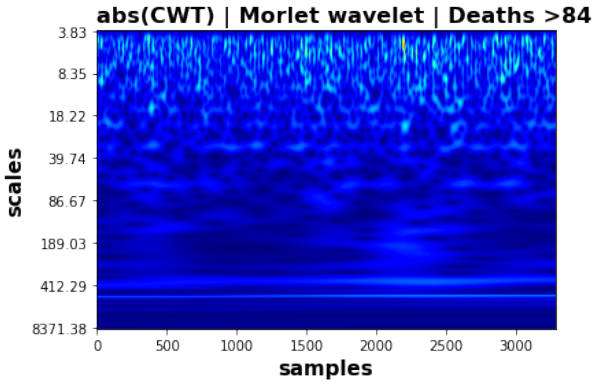


Figure 11. Scalogram of the deaths > 84, in this case the peaks are not completely visible

4.4. *Linear correlation & Time Lagged Correlation*

As said in the prevoius section, I linearize the possible correlation by studying it between $|\delta(T)|$, where $\delta(T)$ is the temperature variantion from the mean value, and $\delta(D^j)$, that is the variation from the average number of deaths for the class of ages j.

Obviously, there can be some time-lagged correlation between the two datasets, so I initially use the Pearson's r to understand the most probable time lag (Fig. 12). As expected, it is clear that the correlation between the two datasets is higher if I study the fourth class of ages , furthermore for all the classes of ages there is a peak to an offset of 4 days. But there is also a Pearson's r ≥ 0.10 for an offset of 3 and 5 days, so I define $\delta T_i^\star$ as:

$$\delta T_i^\star = mean(\delta T_{i-3}, \delta T_{i-4}, \delta T_{i-5})$$

where i is the current date studied. Then, with a moving window of 30 days, I plot the Pearson's correlation between the datasets Fig. 13. As shown in the figure, there is a higher correlation between mortality and $T_i^\star$ than between T_{i-4}. In Appendix is possible to see in much more detail the region with Pearson's r > 0.6.

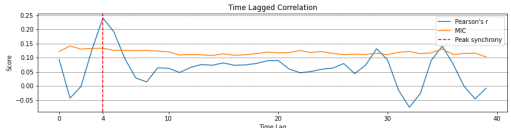


Figure 12. In this image are plot the score value of Maximal information coefficient and Pearson's r for different offset. There is a peak value for r at 4, instead for MIC there is not a peak, the value is always around 0.10

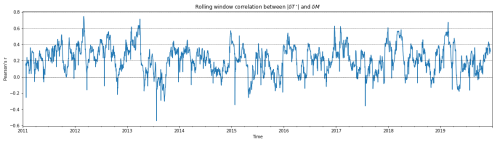


Figure 13. In the image is plot the correlation between $|\delta T_i^\star|$, define as $|mean(\delta T_{i-3}, \delta T_{i-4}, \delta T_{i-5})|$ and the number of deaths for the class of ages > 84, with a moving window of 30 days. So, for every point in the plot, Pearson r is evaluate considering also 15 days before and 14 days after.

4.5. *MIC correlation*

I can now proceed with the study of the MIC correlation between $\delta(T)$ and $\delta(D^j)$ and plot it Fig. 12.

The values given by MIC are pretty similar; this behavior is not so strange because, as shown in Fig. 4, the MIC method has some trouble if the values are spread around the correlation line. Therefore, in this case, it is much better to use a simple linear correlation.

5. CONCLUSION

I have performed a detailed analysis of mortality in Rome for four different classes of ages (≤ 64, $65-74$, $75-84$ and ≥ 85 years) and also a detailed analysis of the mean temperature at Rome for the period from January 1^{st} 2011 to December 31^{st} 2019.

Then I find a linear correlation between the temperature and the classes of ages > 74, ≤ 84 and > 84, with Person's r ~ 0.2 and periods of time when this correlation is ≥ 0.6. So as expected, older people are more susceptible to climate change.

Obviously, this analysis can be refined using a dataset that contains also the reasons for the deaths. Because I expect a stronger correlation for certain types of death, it can also be extended to the Covid period to see if there is a change in the correlations. With a more refined dataset I can also study the correlation taking into account other factors such as air pollution.

REFERENCES

2022, United States Environmental Protection Agency, https://www.epa.gov/climatechange-science

Chatfield, C. 2003, The Analysis Of Time Series - An Introduction, 6th edn., Chapman & Hall/CRC Texts in Statistical Science (Chapman and Hall/CRC)

Cohen, I., Huang, Y., Chen, J., et al. 2009, Noise reduction in speech processing, 1

Daubechies, I. 1990, IEEE Transactions on Information Theory, 36, 961, doi: 10.1109/18.57199

Denman, K. L. 2008, Marine ecology progress series, 364, 219

Guo, Y., Punnasiri, K., & Tong, S. 2012, Environmental health, 11, 1

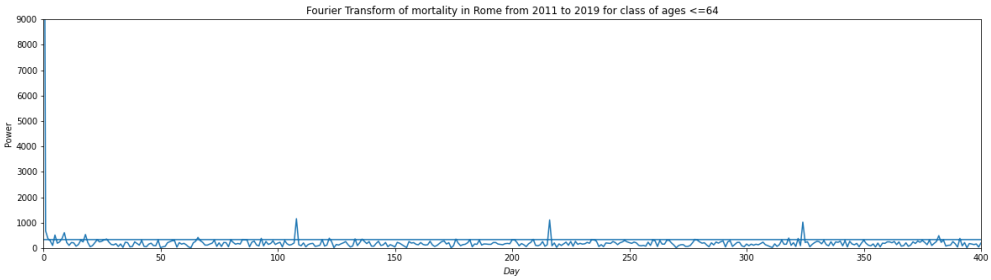
Hardy, J. T. 2003, Climate change: causes, effects, and solutions (John Wiley & Sons)

Huang, N. E., Shen, Z., Long, S. R., et al. 1998, Proceedings of the Royal Society of London Series A, 454, 903, doi: 10.1098/rspa.1998.0193

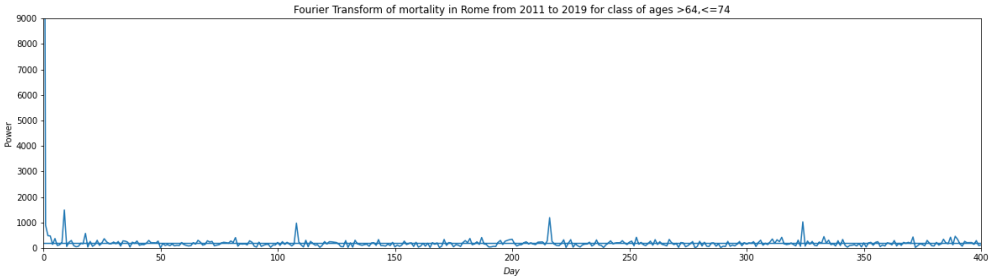
Scargle, J. D. 1982, ApJ, 263, 835, doi: 10.1086/160554

Zhang, Y., Jia, S., Huang, H., Qiu, J., & Zhou, C. 2014, Scientific reports, 4, 1

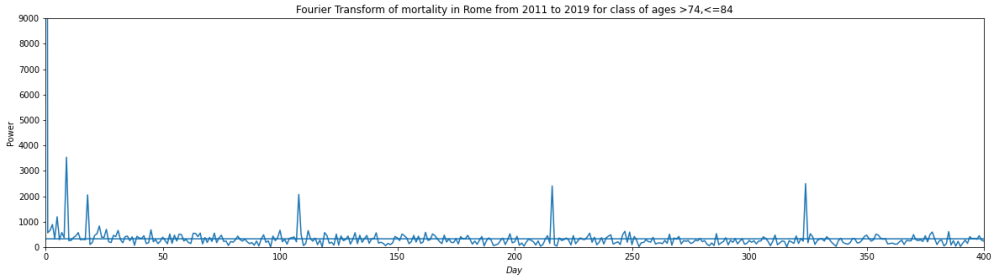
APPENDIX



Fourier Transform of the mortaly for class of ages ≥ 64 at a sampling frequency of 1 day. The peaks are not so high

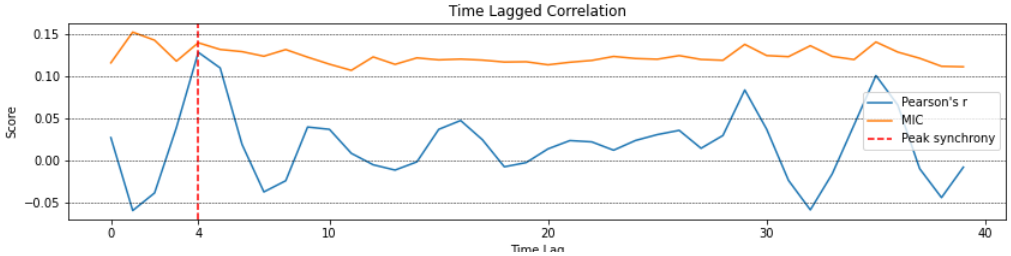


Fourier Transform of the mortaly for class of ages > 64, ≤ 74 at a sampling frequency of 1 day. We can see the first

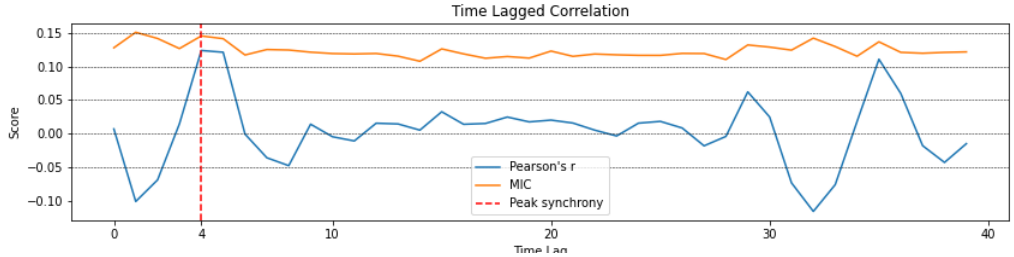


peaks near 0.

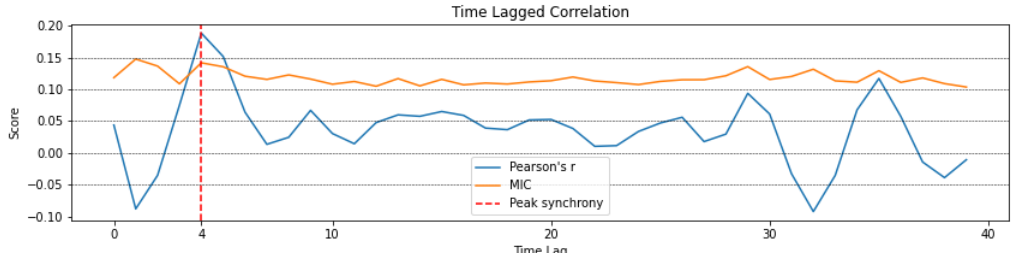
Fourier Transform of the mortaly for class of ages > 74, ≤ 84 at a sampling frequency of 1 day. In this case the peaks are higher, the results are much more similar to the case > 84.



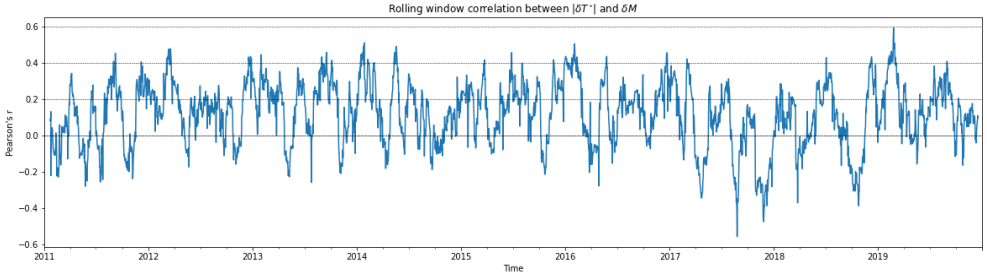
In this image are plot the score value of Maximal information coefficient and Pearson's r for different offset. Case $\geq$ 64, we can say that there is not a correlation between the two datasets.



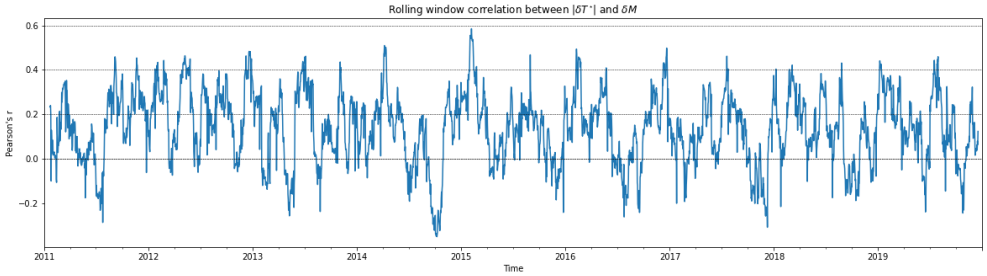
In this image are plot the score value of Maximal information coefficient and Pearson's r for different offset. Case $>$ 64, $\leq$ 74, also in this case there is not a correlation.



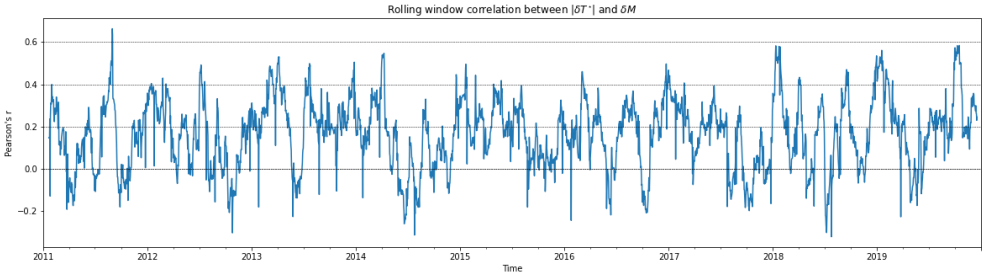
In this image are plot the score value of Maximal information coefficient and Pearson's r for different offset. Case $>$ 74, $\leq$ 84, finally we can see a correlation with lag 4 days.



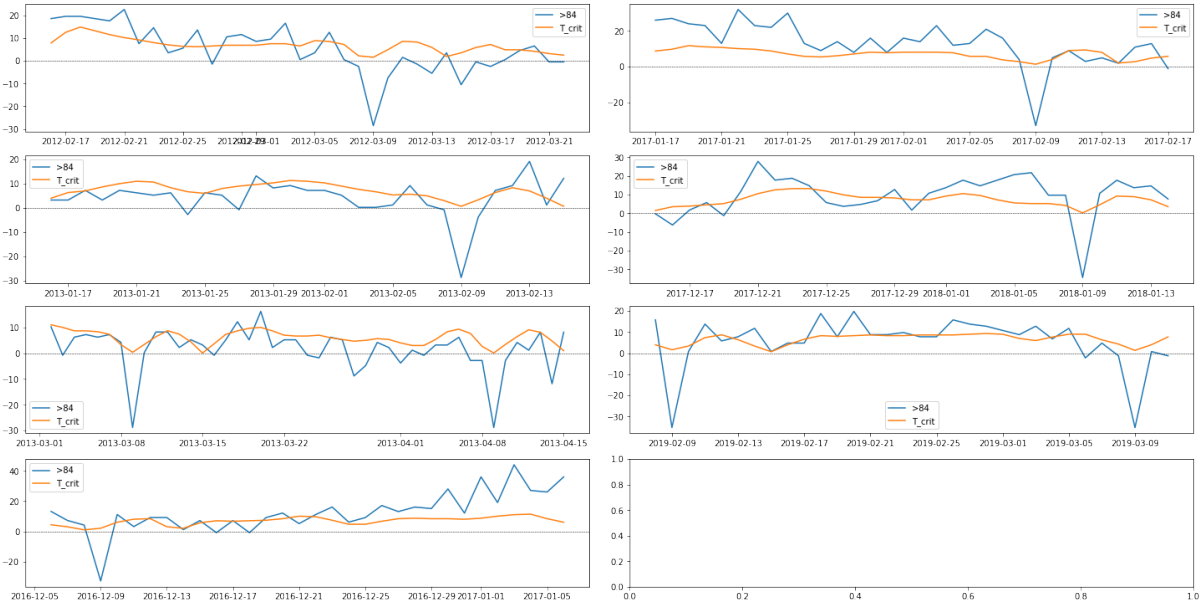
In the image is plot the correlation between $|\delta T_i^{\star}|$ and the number of deaths for the class of ages $\geq$ 64, with a moving window of 30 days. Obviously some periods can have a much higher correlation than others, but I do not think that the results tell us something important, because that could happen by accident.



In the image is plot the correlation between $|\delta T_i^{\star}|$ and the number of deaths for the class of ages $>$ 64, $\leq$ 74, with a moving window of 30 days. In this case we see a higher correlation than the previous.



In the image is plot the correlation between $|\delta T_i^\star|$ and the number of deaths for the class of ages > 74, ≤ 84, with a moving window of 30 days. Finally we see a correlation between the datasets, some periods have also r$\simeq 0.6$



This are the periods with r> 0.6, for the case of > 84. We can see that February, March and April seems to be the periods with higher correlation.