Visualisation and assisted detection of correlations in solar time-series

A. Falcão, UNINOVA-CA3, Monte da Caparica, Portugal, ajf @uninova.pt
I. Dorotovič, UNINOVA-CA3, Monte da Caparica, Portugal; SÚH Hurbanovo, ivan.dorotovic @suh.sk
Joaquim F. Silva, Dep. Informática, FCT-UNL, Monte da Caparica, Portugal, jfs @fct.unl.pt

Abstract
Finding correlations in multiple sets of time-series can be a difficult task when the amount of data is relatively large. Challenges include working with time-series that cover a long period of time or include a high sampling-rate, but also analysing time-series from a large number of different parameters. In this work we present a tool to assist in the analysis of non-categorical numerical time-series. The objective is to provide an application with a set of easy-to-use functionalities to aid in the detection of correlations in different time-series and the detection of possible periodicities.
The tool provides a user friendly graphical user interface, with plotting and zooming capabilities and intuitive functionality selection. It is supported on a number of operating systems. It features detection of positive and negative correlations between two parameters, or in sets of multiple parameters. It also provides detection of "time-phased correlation", i.e., correlations which may have a time difference (delta) associated, with the tool providing a visual feedback of the correlation values for various deltas. Similarly, periodicities within a particular parameter can also be detected (auto-correlation). Graphical representations of the corresponding 11-year solar cycle and superimposed Gleissberg cycle, for example, can be seen using the tool.

Keywords: Time series, Correlation, Auto-correlation, Periodicity detection

1. INTRODUCTION

Every year, huge amounts of data are generated from scientific programmes and projects. Astronomy-related missions are a significant contributor to this large amount of data. For example, it is estimated that ESO's VLT Survey Telescope will generate in the order of 30 terabytes of data anually. Analysing this extremely large amount of data is beyond the limits of human capability and mechanisms are needed for automatically detecting, or at least helping in the detection of, patterns and relationships within these data.

With that in mind, and considering the Solar Physics community and the challenges faced by analysing an overwhelming amount of data, we proposed to develop a set of tools that would allow the analysis of non-categorical numerical time-series which are common in this domain.

The challenge is in dealing with large time-series; these can be large due to covering a long period of time or due to having high sampling rates. Another consideration is that there also a large amount of different parameters that have to be analysed.

The objective was to provide a graphical user interface for the set of tools with easy to use functionalities allowing a person to visually explore the data, and that could provide assistance in the detection of correlations between parameters as well as hidden periodicities with the time-series.

We will show the tool's capabilities in plotting and navigating the data, support for detection of correlations (both positive and negative correlations) as well as functions for detecting „time-phased" correlations – situations where two parameters may be correlated but with some time difference (delta) between them. This allows detecting possible correlations in parameters associated to some physical phenomenon but due to distances between the sensors the values will be measured at different times, with a certain delay (this can be the case of satellites in opposing orbits where one satellite may measure an increase in a particular

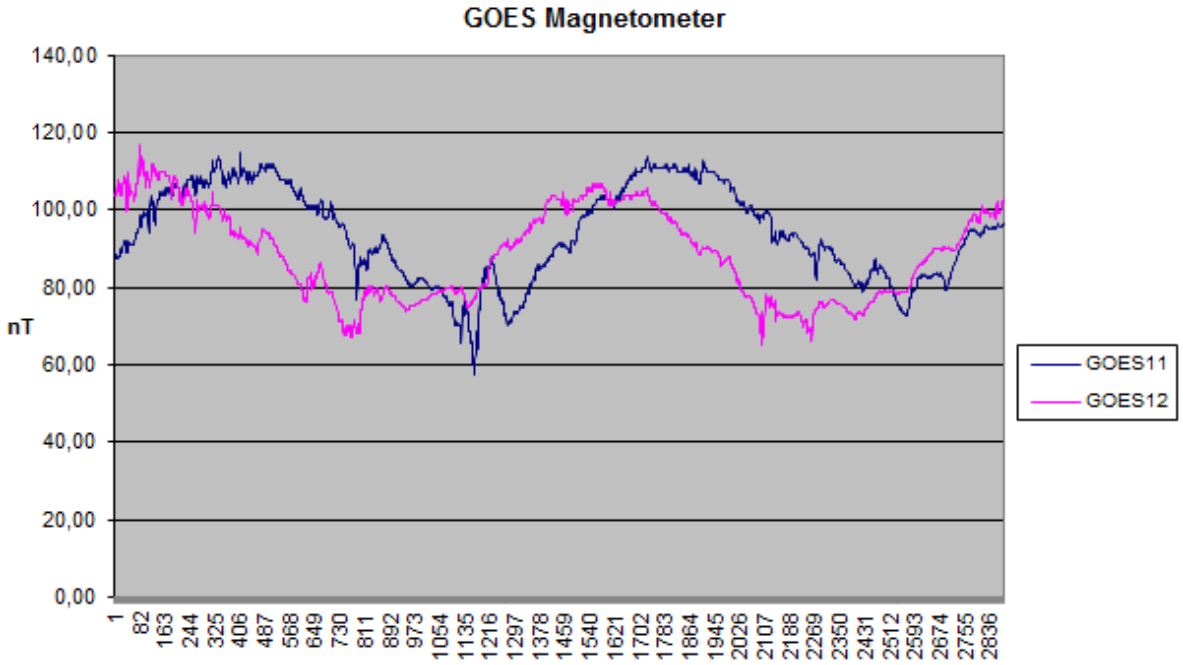
solar parameter, while the other satellite will only measure this same increase after a short time period).

We will also present the use of auto-correlation mechanism as an efficient way to visually find periodicities within a certain time-series parameter.

2. MOTIVATION AND CONTEXT

In this work, we considered non-categorical time-series, i.e., a time-series composed of numerical values from a particular parameter, usually equally spaced over time. The main driver for these developments was to find a way to deal with the very large amounts of different parameters and data volumes with which scientists are faced with nowadays. Manually sifting through this data is a time-consuming and arduous task, which leads to inefficient use of resources.

Consider for example the following plot of two time-series, of magnetometer data obtained from the GOES11 and GOES12 satellites over a particular time-frame, at a sampling rate of 5 minutes (the x axis denotes time samples).



It shows a good example of two correlated time-series, but with an associated temporal deviation between them. With our prototype tool, it is possible to automatically find the deviation value between them (maximizing the correlation value of the two parameters). This data was obtained from SWENET – Space Weather European Network [1].

Another interesting situation is the solar sunspot number time-series (obviously well-known within the Solar Physics community). Retrieving data from the SIDC – Solar Influences Data Analysis Center [2] from the Royal Belgian Observatory it is possible to obtain a relatively long timeseries spanning a few hundred years (specifically since 1749). A plot of this dataset reveals the 11 year solar cycle – but with our tool it is much easier to observe the periodicity within the signal even if it is not so evident in a simple plot.

Due to the work domains of the authors, solar-related parameters were used as a choice of test data. This is also justified by the importance being given to „Space Weather“, which is now the focus of several research and development programmes, including the European Space Agency's SSA – Space Situational Awareness Programme [3]. The study of Space Weather parameters holds great importance due to the impact of the solar influence on Earth. It is a study domain that involves a very large set of different parameters, mostly represented as time-series, and with ongoing and upcoming space missions, the amount of data being generated will be significantly increased. This leads to the necessity of efficient tools for the statistical analysis of these data. A better understanding of the processes involved in the solar activity, will contribute to improved knowledge of the solar cycles and activity, its impact on our planet and effectively aiming for better forecasting models.

3. CORRELATION AND PERIODICITY DETECTION

3.1 Correlation

Current analysis of space weather related time-series may imply a manual manipulation of time series, usually provided in a text format (ASCII file). This leads the researcher to create a graphical visualisation of the time series, either in spreadsheet related tool or a specific plotting environment, and then manually analysing the various dataset plots. The process may not be available in an integrated tool, or the process may

require a command-line execution or similar procedure. This becomes time-consuming and hinders the task of analysing large amounts of different parameters and/or parameters with large amounts of data.

Detecting correlations in these, requires a simple and efficient measure. Several correlation metrics were evaluated, including the Kendall Rank Correlation Coefficient [4], Spearman Correlation [5] and the Pearson Product-Moment Correlation [6, 7].

The Kendall coefficient requires preliminary computation of all pairs of correlations between values and therefore would not scale well for large amount of data. Although the Spearman Correlation considers the difference between the ranks of the various values, does not translate the distance between the actual values. The authors opted instead for the Pearson Correlation.

The Pearson Correlation, which can be represented as:

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

where,

$$cov\ (X,Y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})$$

With N representing the number of elements in the time series, x_i is the i-th value in the series X and $\bar{x}$ is the average. The possible values of the coefficient are between -1, 1, where 1 indicates a perfect positive correlation. Negative correlations are represented by negative numbers.

The Pearson Correlation is extensively used for calculating the correlation between several phenomena, such as location of fluorescent molecules [8], improving the alignment of magnetic resonance imaging [9].

This coefficient turns out to be a very good correlation metric for objective at hand. At handles variables of real values, providing a measure for both positive and negative correlations, with the advantage of being calculated using the time-series direct values (not needing the pre-computation of the series ranks as the Kendall or Spearman correlations). This better reveals any relation between the two time-series, and is computationally more efficient.

3.2 Periodicity Detection

Since we are considering numerical time-series with a continuous domain, Fourier Transforms are a likely method to consider when detecting periodical behaviour. Fourier Transforms provides us with the frequencies and amplitudes present in a signal, which can be denoted as a representation in the frequency domain of the original function. It decomposes a signal into a sum of sines each with its own amplitude and frequency. Although with Fourier Transforms we are able to detect components in the frequency domain, it presents a drawback since it does not provide "when" the component occurs in the signal, i.e., even though we can detect some periodic behaviour, we are not able to visualize where in the signal it occurs.

This issue can be overcome by analysing the signal in short time windows (the *Short Time Fourier* Transformation) which led to the development of *wavelets*, which are adapted to the analysis of non-stationary signals [10,11]. There are several families of wavelets – these are effectively "used to measure" and can be tailored according to the signal being analysed. A wavelet is in essence a mathematical function used to divide a continuous signal into its frequency and time components. In order to select the best wavelet family to use, in [12] the authors present a possible method for choosing a wavelet which has good characteristics in identifying fluctuations in electrical signals. The article demonstrates the difficulty in selecting the right wavelet family for the data being observed.

In order to overcome this difficulty, and the complexity of using wavelets, we opted to use auto-correlation as a way to find periodicities in time-series. By calculating the auto-correlation (correlation of a variable with itself), and successively time-shifting the series, it is possible to visualise any periodic behaviour in the time-series. Although the use of auto-correlation for this is not completely innovative, our implementation is significantly more simplified than other approaches, without the need for successive integration of time intervals.

4. VISUALISATION TOOL

As explained in the previous section, we've used the Pearson Coefficient to calculate the correlation between two variables and as mechanism to visualise periodicities, via auto-correlation.

In this section we will present the visualisation tool that was developed to test and validate these concepts and as an integrated platform to explore Space Weather data. The tool itself was developed using Java [13], in order to provide multi-platform support, being able to run in many supported operating systems. It provides a user-friendly graphical user interface (GUI), which makes navigating the tool easy and simplifies the use of the available functionalities (most functions are available as a right-click sub-menu that pops-up).

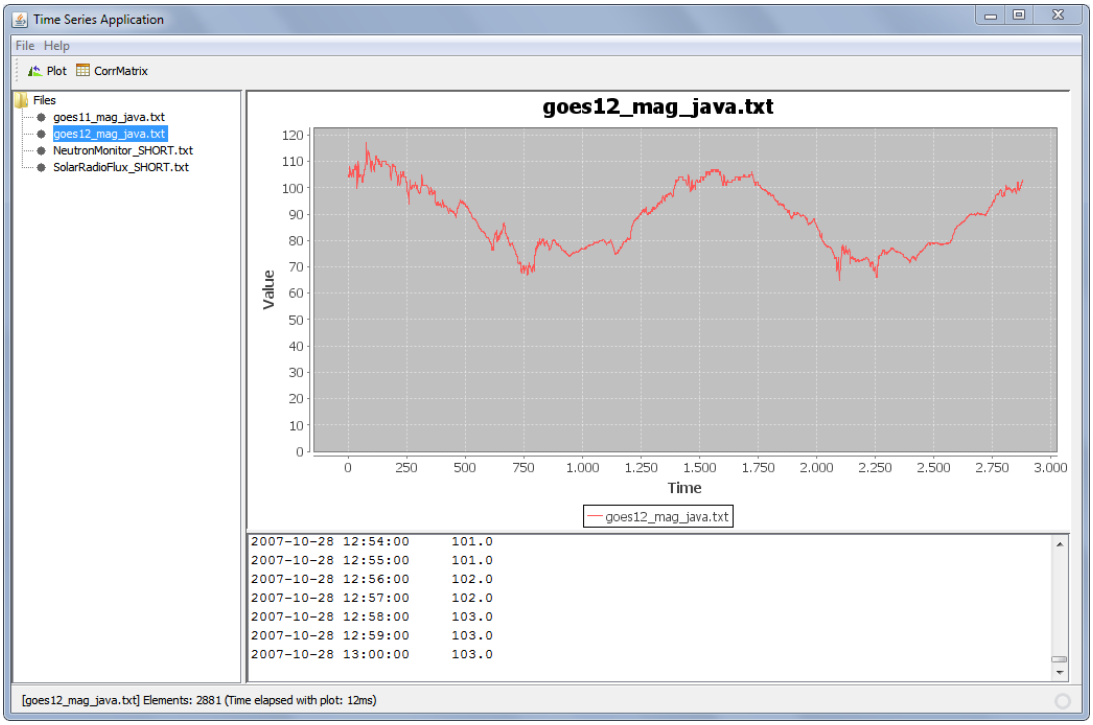
The figure below shows a general view of the tool. It provides three distinct areas:

(i) On the left, a listing of the currently open files (representing time-series)
(ii) A main area showing the graphical representation of the selected parameter
(iii) On the bottom, a console area showing the actual values read from the time-series file.

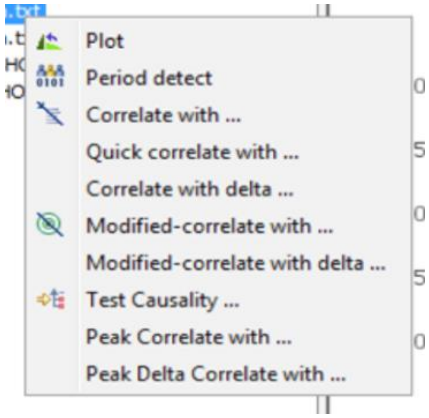
Multiple files can be loaded at once, providing the list of variables on the left hand side tree representation.

The user can then navigate the various parameter files and each time on is selected, the respective graphical representation is shown in the main window.

This window provides zooming and panning capabilities allowing the user to explore in detail large time-series.

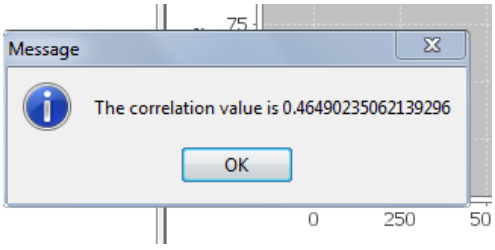


As mentioned, functions are accessed by a right-click sub-menu providing easy navigation:

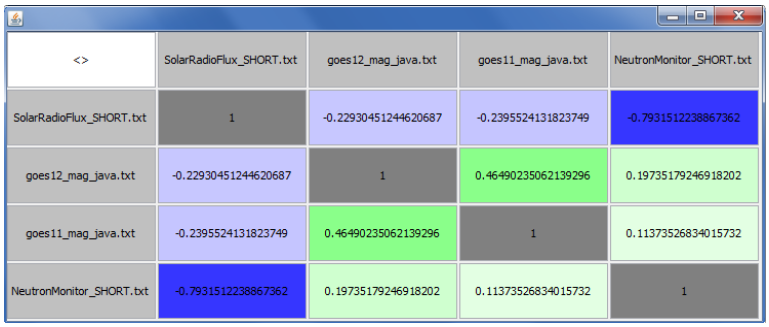


4.1 Correlations

By loading at least two variables, the user can right-click and select "Correlate with" to correlate the two parameters. A simple pop-up window presents the correlation value:



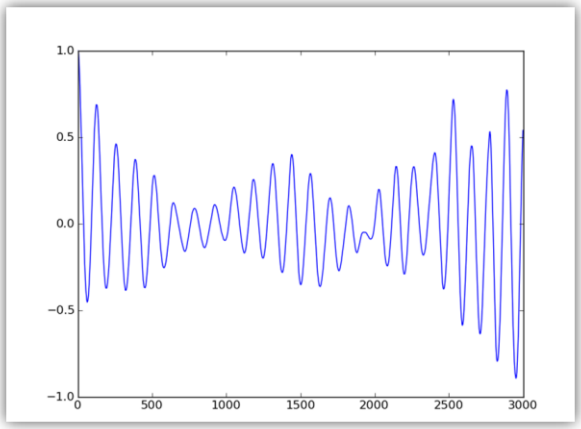
A rather useful feature is to calculate all parameter correlations (two-by-two) – since the method is lightweight, this is not a time-consuming process and is presented relatively fast on a modest computer configuration. A matrix view of all the loaded parameters is shown, with a colour-scale representing the degree of correlation. Green shades represent positive correlations, while blue shades represent negative correlations. More solid colours represent a high correlation value (either positive or negative).



This allows a very rapid visualisation of the correlations between the various parameters that the user is analysing.

4.2 Periodicities

The detection of periodicities in a variable is done by right-clicking a parameter and selecting "Period detect". The auto-correlation is quickly calculated and presented in a window. For example, considering the solar sunspot time series, we can see the in the following image the result of this function.



The x axis represents time samples (in months) and the y axis is the correlation value, ranging from -1 to 1. The number of samples between the peaks (measured using the tool) is around 132 months, i.e., 11 years – clearly representing the 11 year solar cycle. Interestingly, with this visualisation it is also possible to see another lower frequency in the data which is in fact the Wolf-Gleissberg cycle [14, 15]. Although this is not a new discovery, it served as a useful validation of this periodicity detection method.

5. CONCLUSIONS AND REMARKS

Our focus has been on the correlation between parameters, using the Pearson Coefficient. With it we are able to quantify the correlation between parameters (both positive and negative), we can use the same formula to detect periodicities using auto-correlation and provide an easy visualisation associated to computationally lightweight processing.

The visualisation tool itself is useful for the detection of correlations and periodicities within time series. It can be used to analyse correlations between multiple variables. It provides a user friendly graphical user interface and is supported in multiple platforms.

There is still some work to do, as this is currently just a prototype for proof-of-concept demonstration. For future work it is necessary to improve the data loading mechanism, allowing support for even larger datasets (memory offloading techniques and caching), as well as reading parameter metadata if it is available in the file for chart axis captions and units.

REFERENCES

[1] SWENET - Space Weather European Network. [Online] http://www.esa-spaceweather.net/swenet/index.html.

[2] SIDC – Solar Influences Data Analysis Center. [Online] http://sidc.oma.be/.

[3] Space Situational Awareness. [Online] European Space Agency. http://www.esa.int/esaMI/SSA/index.html.

[4] Kendall, M. A New Measure of Rank Correlation. s.l. : Biometrika, 1938. 30, pp. 81-89. doi:10.1093/biomet/30.1-2.81.

[5] Spearman, C. The Proof and Measurement of Association Between Two Things. American Journal of Psychology. 1904, 15, pp. 72-101.

[6] Pearson, Karl. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philos. Trans. Royal Soc. London Ser. A. 1896, 187, pp. 253-318.

[7] Pearson, K. Notes on the History of Correlation. s.l. : Biometrika Trust, Oct. 1920. Vol. 13, 1.

[8] Adler, J., Parmryd, I. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. Cytometry Part A. 2010. 77A, pp. 733–742.

[9] Saad, Ziad S., Glen, Daniel R., Chen,G., Beauchamp, Michael S., Desai, R., Cox, Robert W. A new method for improving functional-to-structural MRI alignment using local Pearson correlation. 2009, Vol. Volume 44, Issue 3, pp. 839-848.

[10] Chan, Kin-pong, Fu, Ada Wai-chee. Efficient Time Series Matching by Wavelets. Sydney, Australia : s.n., 1999. 0-7695-0071-4.

[11] Hubbard, Barbara Burke. The world according to wavelets: the story of a mathematical technique in the making. Natick, MA, USA : A. K. Peters, Ltd., 1996. 1-56881-047-4.

[12] Vega, V., Duarte, C., Ordóñez, G., Kagan, N. Selecting the Best Wavelet Function for Power Quality Disturbances Identification Patterns. Harmonics and Quality of Power. 2008.

[13] Java website: https://www.java.com/en/

[14] Yousef, Shahinaz M. The Solar Wolf-Gleissberg Cycle And Its Influence On The Earth. ICEHM2000. Cairo, Egypt : s.n., 2000. pp. 267-293.

[15] 26. Gleissberg, W. The Eighty-year Sunspot Cycle. 1958, 68, pp. 148-152.