The Use of MATLAB for Working with Big Data on the Example of Processing Information from the World Network of Magnetic Observatories INTERMAGNET

Anatoly Korobeynikov [1]

[1] *Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of The Russian Academy of Sciences St.-Petersburg Filial (SPbF IZMIRAN), University embankment 5, St.-Petersburg, 191023. RUSSIA*

Abstract

In this paper, using the data from the geomagnetic observatory as an example, the use of MATLAB for working with big data using the Datastore is considered. The source text in MATLAB language is presented. It is shown that a user who is not a highly qualified programmer in the MATLAB environment can work with big data using the considered tools.

Keywords

MATLAB, Big Data, Geomagnetic Observatory, INTERMAGNET, DataStores, Tall Array, GPU

1. Introduction

At present, all over the world, and specifically in Russia, in order to work with big data the developing technology called «Big data» is widely used. It includes a variety of tools, as well as various methods and approaches used in processing big data when solving various problems [1,2].

The MATLAB system is the instrumentarium of this paper. Using it, problems in various subject areas can be solved [2-9].

The concept of big data is not dependent on the amount of data. This is due to the exponential growth of computer resources. For example, currently it is common to measure big data in terabytes, and after a while, it will be measured in petabytes. Proceeding from this, in working with big data, one should understand the data that does not fit into the computer's memory.

The data itself is on the following website: https://www.intermagnet.org/imos/imos-list/imos-details eng.php?iaga_code=SPG.

These data were obtained on the part of the St. Petersburg branch of the Institute of Terrestrial Magnetism and the ionosphere of the Russian Academy of Sciences geomagnetic observatory "Saint-Petersburg" (international IAGA-code SPG), the information from which enters international INTERMAGNET network (INTERMAGNET - International Real-Time Magnetic Observatory Network) [10,11].

2. Problem statement

Due to constant accumulation of data on the geomagnetic field at the geomagnetic observatory, the development of Big Data technology and the emergence of high power tools for working with big data in the MATLAB system, the task was set to master the work with this toolset. Moreover, it is necessary that a researcher who is not a highly qualified specialist in the field of MATLAB system programming can perform the processing and analysis of big data.

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL:
ORCID:

3. Proposed solution.

For working with big data, the following text shows the use of the Datastore mechanism.First of all, the data located on the above site is selected. In this work, data for the period from January 1, 2020 to March 31, 2021 was selected.

These are text files named spg20200101qmin.min - spg20210331qmin.min. 441 files are found. After that, we rewrite these files into the directory that we intend to make Current Folder after starting MATLAB.

In recent versions of MATLAB, there is a very convenient tool called the Live Editor. To start it, you need to go to the LIVE EDITOR tab. After that, we type the commands:

```
clc
clear
Fs=1/60; % Sampling rate 1 min
fileName = 'spg202*.*';  % File names
files_Datastores = …%  Reading data in Datastores
    fileDatastore(fileName,'ReadFcn',@read_spg_file,'FileExtensions','.min');
SPG_Datastore = readall(files_Datastores);
SPG=[]; % For convenience, we transform the data into a timetable
for k=1 : size(SPG_Datastore,1)
    TT=SPG_Datastore{k};
    TT.DATETIME=datetime(TT.DATETIME,'InputFormat','yyyy-MM-dd HH:mm:ss.SSS');
    TT=table2timetable(TT);
    TT = removevars(TT,"DOY"); % Removed column named DOY
    SPG=[SPG;TT];
end
SPG = sortrows(SPG,"DATETIME");
clearvars TT;
stackedplot(SPG); % Displaying a graph of the initial data
```

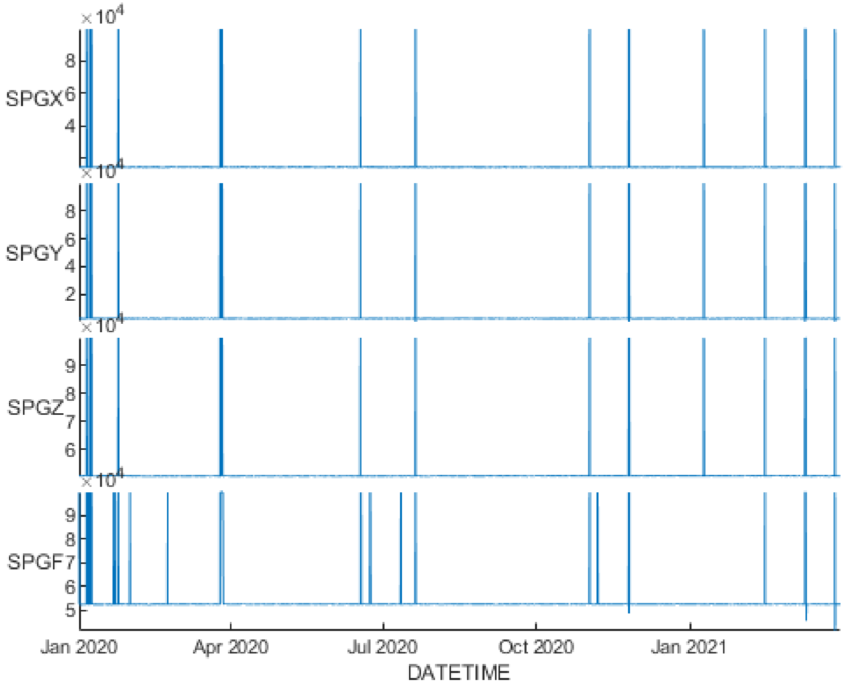

Figure 1: Graphs of initial data

In Figure 1 shows that the initial data have outliers arising from various reasons. For example, when there is a power outage. Let's remove these outliers using the following commands.

```
SPG{find(SPG.SPGX>0.145*summary(SPG).SPGF.Max), …
 ["SPGX","SPGY","SPGZ","SPGF"]}=NaN; % Replacing emissions
[SPG,missingIndices1] = fillmissing(SPG,"linear");
[SPG,outlierIndices] = rmoutliers(SPG,"grubbs","DataVariables","SPGF");
[SPG,missingIndices2] = fillmissing(SPG,"linear",...
"MaxGap",calyears(17),"DataVariables","SPGF");
clearvars missingIndices1 missingIndices2 outlierIndices;
% Gaps replaced by linear interpolation
stackedplot(SPG); % Display the cleaned data graph
```

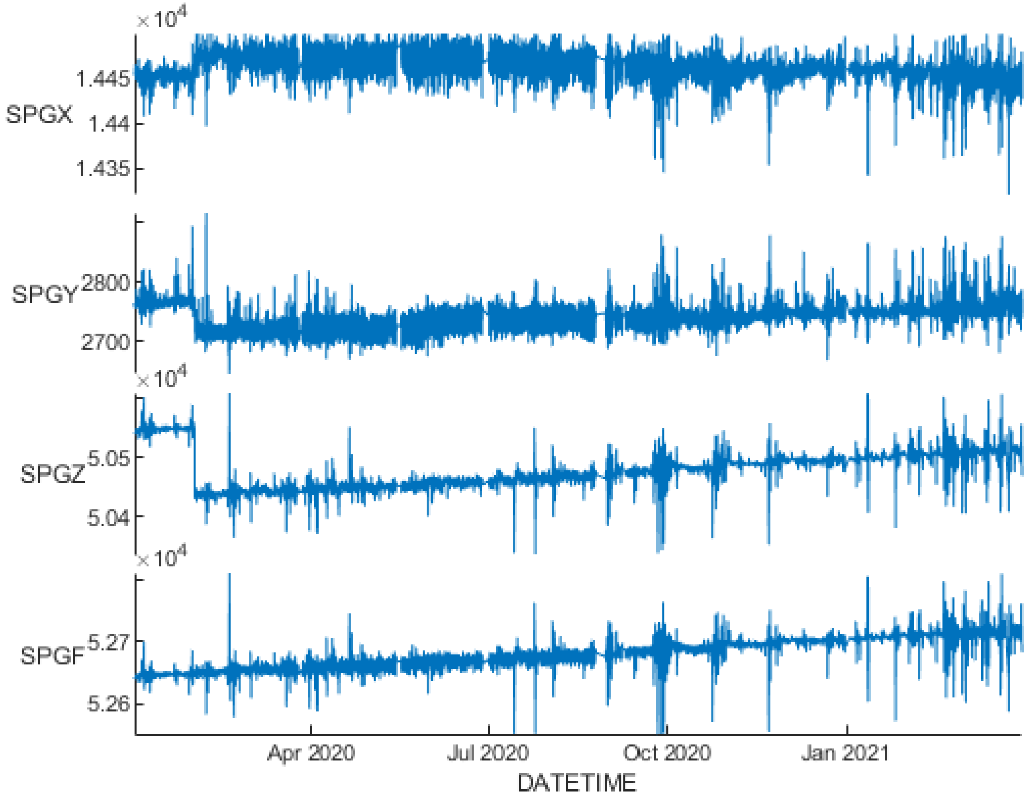

Figure 2: Graphs of cleaned data

In Figure 2 shows that there are no outliers in the data. Further work with data depends on the task set by the researcher. For example, you can find linear trends in the data over the study period. This can be done using the following commands.

```
deg_trend=1; % Trend – straight
X_trend=SPG.SPGX-detrend(SPG.SPGX,deg_trend);
Y_trend=SPG.SPGY-detrend(SPG.SPGY,deg_trend);
Z_trend=SPG.SPGZ-detrend(SPG.SPGZ,deg_trend);
F_trend=SPG.SPGF-detrend(SPG.SPGF,deg_trend);
stackedplot(timetable(SPG.DATETIME,X_trend));
% Sequential display of graphs of X, Y and Z components
```

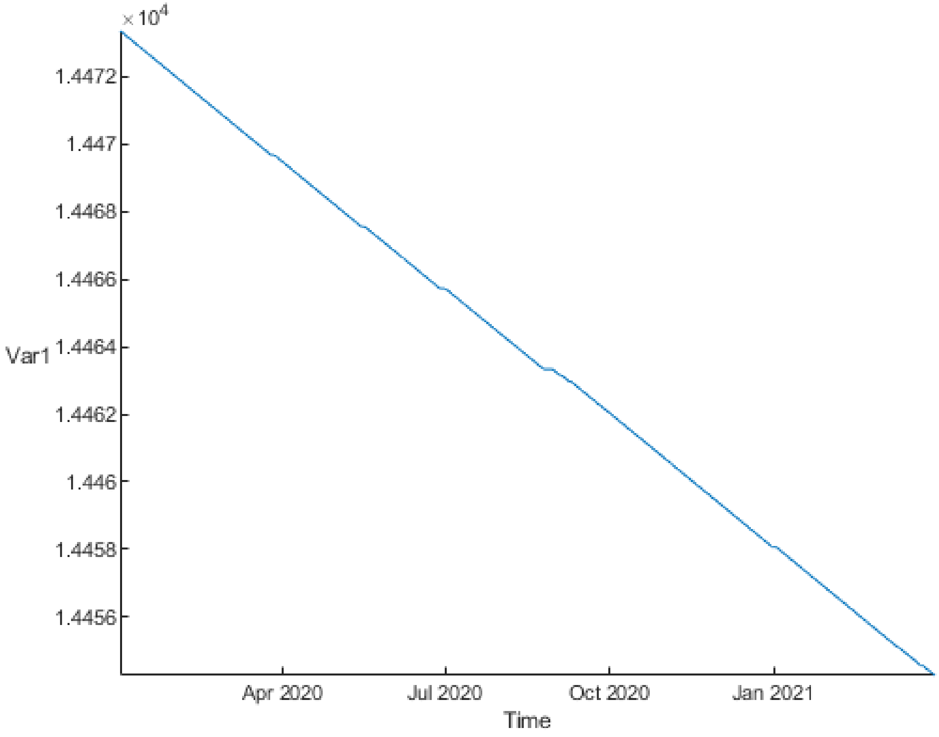

Figure 3: Trend Chart X – Components

stackedplot(timetable(SPG.DATETIME,Y_trend));

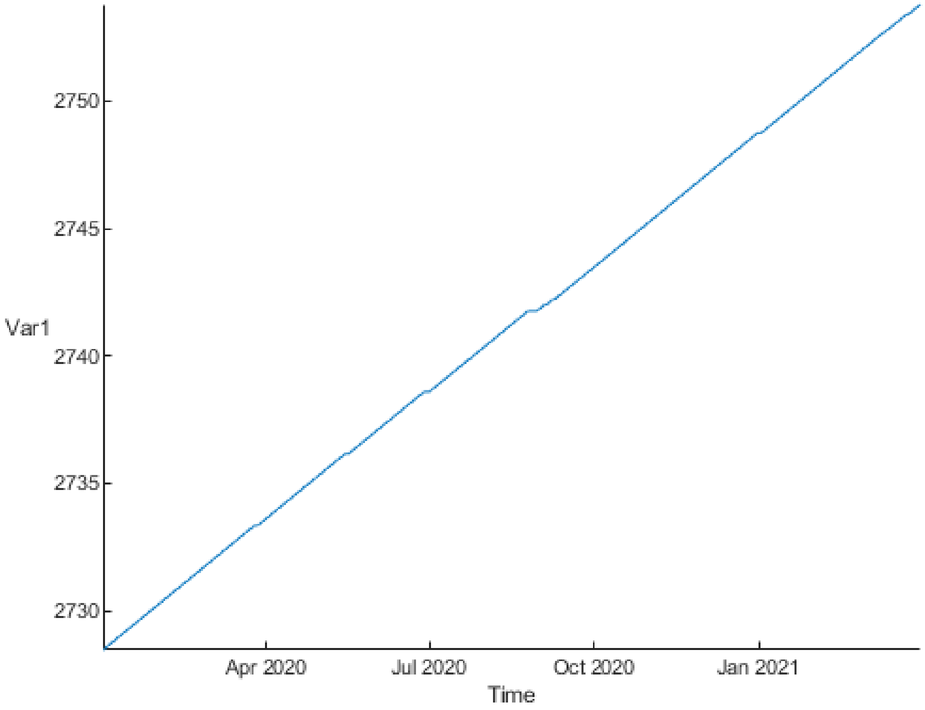

Figure 4: Trend Chart Y – Components

stackedplot(timetable(SPG.DATETIME, Z_trend));

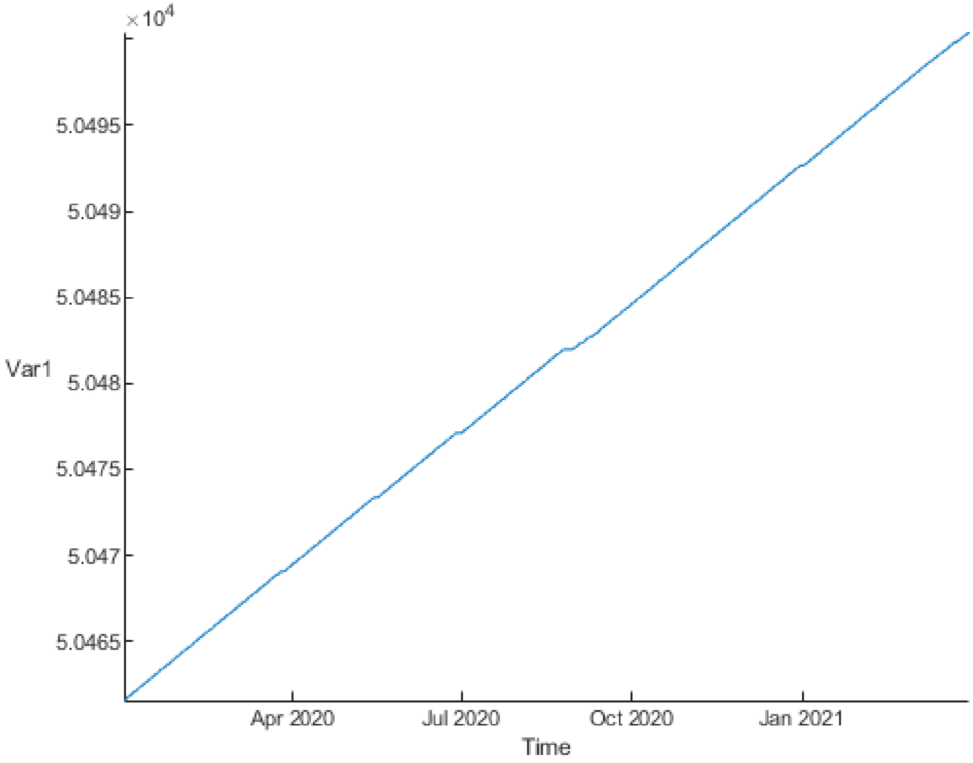

Figure 5: Trend Chart Z - Components

Below is the source code of the file read function.

```
    function spg_table = read_spg_file(filename, startRow, endRow)
    if nargin<=2
      startRow = 29;  endRow = inf;
    end
    formatSpec = '%24s%3f%13f%10f%10f%f%[^\n\r]'; % Data format
    fileID = fopen(filename,'r'); % Opening a text file
    dataArray = textscan(fileID, formatSpec, endRow(1)-startRow(1)+1, 'Delimiter', '', 'WhiteSpace', '',
    'TextType', 'string', 'HeaderLines', startRow(1)-1, 'ReturnOnError', false, 'EndOfLine', '\r\n');
    for block=2:length(startRow)
      frewind(fileID);
      dataArrayBlock = textscan(fileID, formatSpec, endRow(block)-startRow(block)+1, 'Delimiter', '',
      'WhiteSpace', '', 'TextType', 'string', 'HeaderLines', startRow(block)-1, 'ReturnOnError',
      false, 'EndOfLine', '\r\n');
      for col=1:length(dataArray)
         dataArray{col} = [dataArray{col};dataArrayBlock{col}];
      end
  end
  dataArray{1} = strtrim(dataArray{1});
  fclose(fileID); % Closing a text file
% Creating an output variable
spg_table = table(dataArray{1:end-1}, 'VariableNames',
 {'DATETIME','DOY','SPGX','SPGY','SPGZ','SPGF'});
End
```

4. Conclusion

As seen from the presented source text, the MATLAB system provides a fairly clear and convenient toolset for working with big data, which allows to solve various problems.

For example, presented in Fig. 3 - Fig. 5 graphs allow us to conclude that there is a displacement of the geomagnetic pole. For a more accurate change of its location, further work with the data is required.

In addition, it should be noted that in the MATLAB system, there are also tools for working with big data. For example tall array.

To speed up calculations, you can use parallel calculations using "worker" or graphics processes [12-14].

5. References

[1] Makshanov A. V. Big data. Big Data: a textbook for universities / A. V. Makshanov, A. E. Zhuravlev, L. N. Tyndykar. - St. Petersburg: Lan, 2021. - 188 p. ISBN 978-5-8114-6810-2

[2] R. Suganya, S. Rajaram, A. Sheik Abdullah. Big Data in Medical Image Processing//CRC Press, 2018. – 210 page, https://doi.org/10.1201/b22456, ISBN: 9781138557246

[3] Amos Gilat. MATLAB An Introduction with Applications. Fifth Edition// Per. from English Smolentsev N.K. - M.: DMK Press, 2016 .-- 416 p. ISBN 978-1-11862-986-4

[4] A. Y. Grishentsev, A. G. Korobeynikov. Solution model of inverse problem of ionosphere vertical sounding//Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2011. № 2 (72). P. 109-113. https://www.elibrary.ru/item.asp?id=15632791

[5] A. U. Grishentcev, A. G. Korobeynikov. Interoperability tools in distributed geoinformation systems//Journal of radio electronics, № 3, 2015. https://www.elibrary.ru/item.asp?id=23327290

[6] Korobeynikov A. G., Grishentsev A. Yu, Svatkina M. N. - Application of intelligent agents of magnetic measurements for monitoring railway infrastructure objects//Cybernetics and programming. – 2013. – № 3. – P. 9 - 20. DOI: 10.7256/2306-4196.2013.3.8737 URL: https://nbpublish.com/library_read_article.php?id=8737

[7] Gvishiani A.D. Big Data, FAIR Data and Open Data for Systems Analysis//Workshop Big Data and Systems Analysis. International Institute for Applied Systems Analysis (IIASA). Laxenburg, Austria. 24–25 February 2020.

[8] Wei, J., Yang, W. Using Big Data to Establish Mathematical Model Method to Identify the Safety Displacement System of Oil Storage Tank. Chem Technol Fuels Oils 56, 593–600 (2020). https://doi.org/10.1007/s10553-020-01172-0

[9] J. Skeivalas, E. K. Parseliunas D. Slikas , R. Obuchovski, and R. Birvydiene. On Analysis of Seismic Vibrations Data Applying Doppler Effect Expression//ADVANCES IN CIVIL ENGINEERING, 2021, № 8839828. DOI: 10.1155/2021/8839828. https://downloads.hindawi.com/journals/ace/2021/8839828.pdf

[10] Korobeynikov A. G. Processing and analysis of data from the Russian segment of the world network of magnetic observatory INTERMAGNET//International Journal of Humanities and Natural Sciences. 2018. - № 8. - P. 91-98. https://www.elibrary.ru/item.asp?id=35619975

[11] Kudin, D., V., Soloviev, A. A., Sidorov, R., V., Starostenko V. I., Sumaruk Yu. P., Legostaeva O. V. Advanced Production of Quasi-Definitive Magnetic Observatory Data of the INTERMAGNET Standard//GEOMAGNETISM AND AERONOMY, Vol 61, № 1, Page 54-67. DOI: 10.1134/S0016793221010096

[12] R. D. Team, RAPIDS: Collection of Libraries for End to End GPU Data Science, 2018. [Online]. Available: https://rapids:ai

[13] M. Shabaninezhad, M.G. Awan, G. Ramakrishna. MATLAB package for discrete dipole approximation by graphics processing unit: Fast Fourier Transform and Biconjugate Gradient//Journal of Quantitative Spectroscopy and Radiative Transfer, Vol. 262, 2021, ISSN 0022-4073, https://doi.org/10.1016/j.jqsrt.2020.107501.

[14] Tao Zhang, Wang Kan, Xiao-Yang Liu. High performance GPU primitives for graph-tensor learning operations//Journal of Parallel and Distributed Computing, Vol. 148 (2021), Pages 125-137. DOI: 10.1016/j.jpdc.2020.10.011