frontiers | Frontiers in Astronomy and Space Sciences

TYPE Perspective
PUBLISHED 27 September 2022
DOI 10.3389/fspas.2022.1008888

OPEN ACCESS

EDITED BY
Noora Partamies,
The University Centre in Svalbard, Norway

REVIEWED BY
Luke Barnard,
University of Reading, United Kingdom
Nithin Sivadas,
National Aeronautics and Space Administration, United States

*CORRESPONDENCE
Elena A. Kronberg,
kronberg@geophysik.uni-muenchen.de

SPECIALTY SECTION
This article was submitted to Space Physics,
a section of the journal
Frontiers in Astronomy and Space Sciences

RECEIVED 01 August 2022
ACCEPTED 09 September 2022
PUBLISHED 27 September 2022

CITATION
Kronberg EA (2022), Data analysis in space physics: My experience and lessons learned.
Front. Astron. Space Sci. 9:1008888.
doi: 10.3389/fspas.2022.1008888

Data analysis in space physics: My experience and lessons learned

Elena A. Kronberg*

Department of Earth and Environmental Sciences, Ludwig Maximilian University of Munich, Munich, Germany

The specific area of investigation in this perspective is data analysis in space physics. This paper is intended to be useful for those who start working with observations in space physics, especially with a focus on charged particle measurements. I forward lessons I learned regarding the data analysis such as calibration, statistics and machine learning. I also list practices which I find important in research in general. An outlook on possible future directions in space physics is given.

KEYWORDS

space physics, charged particle observations, data analysis, statistics, machine learning

1 Introduction

A wide spectrum of methods of data analysis in space physics are well presented in, e.g., Paschmann and Daly (1998) and Paschmann and Daly (2008). However, there are common mistakes that are not described in the literature. Because of time pressure and "result orientation", people sometimes do not bother enough to familiarize themselves with the data and the calibrations involved. It is also common to put undue trust in data. In Section 2.1, I describe the lessons I learned about the handling of data and best practices for communicating with data providers. Large data sets can provide global pictures of physical processes in the space environment. Statistical methods in space physics and also their misuse can be found in, e.g., Reiff (1990). However, the processing of large data sets can easily lead to erroneous results if not done carefully. In Section 2.2, common mistakes in the processing of the large data sets are described. Machine learning techniques are popular for dealing with long observation series. Their application in space sciences is highlighted, e.g., in Bortnik and Camporeale (2021). In Section 2.3, common mistakes in the application of these methods are pointed out. In Section 3, general lessons for space physicists, which I learned during my career, are discussed. In Section 4, I will give an outlook on directions in space physics.

2 Lessons: The data analysis

2.1 Get to know your data

I always recommend to students:

- Lesson 1: Use the data carefully, read the metadata, read User Guides and Calibration Reports, contact PIs and Co-investigators about the data

Data are often taken from a data archive. After my graduation, I became a member of the RAPID [the Research with Adaptive Particle Imaging Detector (Wilken et al., 1997)] team at the Max Planck Institute for Solar System Research led by Dr. Patrick Daly, the principal investigator (PI) of this instrument on the Cluster mission (Escoubet et al., 2001) by the European Space Agency. My job was to work on calibration and preparation of the RAPID data for the Cluster Science Archive (CSA, https://www.cosmos.esa.int/web/csa), writing the User Guide (Daly and Kronberg, 2022), Calibration report (Kronberg et al., 2022) and Interface Control Document [ICD, (Daly et al., 2021)]. Before this project, during my doctorate, I was using data from the Galileo mission to analyze the dynamics of the Jovian magnetosphere without considering that the data could have errors and biases. However, working on data calibrations and processing I understood how much data is altered before it is archived. I learned about the problems and corrections of particle data, how many iterations are needed to obtain "ideal" values, and how much work is involved in the preparation of scientific datasets. For example, to convert raw counts measured by the Cluster/RAPID into electron differential fluxes, one has to apply geometry factors, take the time-dependent efficiency of the detectors into account, shift the initial energy threshold of the lowest channel, and remove the solar contamination and the pedestal noise, see more details in Kronberg et al. (2022). I realized the importance of archiving the data and maintaining of data archives such as CSA. I advocate for this work to be recognized by the scientific community by using DOIs for data sets and related documents such as User Guides, Calibration reports and ICDs.

The highlighting characteristic of the CSA is that the data quality is controlled by a dedicated team. The observations from different instruments, spacecrafts and missions are cross-calibrated and, therefore, the data is complemented and improved. The calibrations which were applied on the data are well described in the Calibration Reports and User Guides of the corresponding instruments. The calibrations, are therefore, not a black box and a user can, in principle, take the raw data and apply a calibration procedure to receive a scientific product. Another advantage of the CSA is that the archiving team works closely with scientists. Because of this interaction, the archive offers many useful scientific products and convenient interfaces, e.g., for plotting.

I show several examples of my own work in which the lesson above was crucial to avoid wrong results.

In my work on the origin of energetic ion events measured upstream of the Earth's bow shock by STEREO, Cluster, and Geotail missions (Kronberg et al., 2011), I worked on explaining upstream events observed far away ($> 70\ R_E$) from the Earth. For this I have combined observations from the above-mentioned space missions. I used particle measurements by STEREO which were given to me in the form of an ASCII table without any metadata. Measurements by the Cluster/RAPID instrument are delivered to the archive in keV units for the particle intensities. Being naive, I thought that the same is true for the STEREO data. It was quite striking that the energetic particle intensities measured by STEREO were very strong compare to those measured near the Earth by the Cluster and Geotail. We even had an explanation for such an interesting observation. Luckily, before submitting the manuscript an expert in STEREO data has noted that instead of keV, the intensity units are MeV. This spoiled our initial interpretation of the data (which by the way was very exciting), we needed to rework the interpretation quite a lot but we avoided submitting an incorrect study.

Here is another instructive example. The Van Allen Probe mission has discovered a temporal third radiation belt which was observed for more than 4 weeks (Baker et al., 2013). Generally the data in radiation belts observed by the RAPID instrument were considered to be rather useless due to background contamination. A warning about this issue has been stated in the RAPID User Guide. Still, a manuscript using Cluster/RAPID observations was submitted to the Nature journal, about the discovery of a third radiation belt which is persistent on long time scales, for several months and during several years. This could have been a great discovery. The reviewers have commented that the manuscript can be published if the RAPID experts confirm that this belt is not a contamination of the observations. The RAPID team was already working on simulations of the RAPID/Imaging Electron Spectrometer (IES) in the radiation belt environment. The detector was bombarded with particles at an energy spectrum corresponding those in the radiation belts. Our results have shown that the "third radiation belt" is indeed a contamination (Kronberg et al., 2016). Unfortunately, the manuscript was not published in Nature but we avoided the publication of wrong results. Since then, a novel cleaning technique for background contamination, also described in the RAPID Calibration report, has helped to make the RAPID data in the radiation belts useful for science. This allowed, for example, an extensive statistical study of radiation belts (Smirnov et al., 2019) and the deduction of information on particle anisotropy for the calculation of the wave power of chorus waves (Breuillard et al., 2015). We also created a guide on how to calculate adiabatic invariants using the Cluster/RAPID data (Smirnov et al., 2020a) and the LSTAR product for the CSA.

Eventually, the Galileo/Energetic Particle Detector (EPD, (Williams et al., 1992)) ion observations which I used for my doctorate, never doubting their accuracy, were corrected for radiation background contamination. It did not affect the results of my thesis. However, in my recent study of the ion acceleration in plasmoids (Kronberg et al., 2019), I excluded the formerly included helium observations because after the correction we did not have a sufficient amount of reliable helium data. Thanks to EPD experts!

- Lesson 2: Question "gold standards"

The data are not "static", meaning they may change after many years if a better calibration technique is found. Moreover, calibrations are a form of measurement interpretation. They can be subjective. This can affect older studies. This can also affect "gold

standards in observations". For example, the charged particle observations by Combined Release and Radiation Effects Satellite (CRRES) launched in 1990 were considered as a "gold standard". The charged particle observations by Polar and LANL satellites were cross-calibrated with those from the CRRES. I was working on the cross-calibration of protons observed by Cluster between the two instruments: the RAPID and the Cluster Ion Spectrometry [CIS, (Rème et al., 2001)]. The cross-calibrations were relatively good (Kronberg et al., 2010) [they were redone later for both instruments but still having relatively good agreement, see Kronberg et al. (2022)]. However, comparing the RAPID proton observations with those from the Polar mission we found a difference of about one order of magnitude. We were not happy to see this, because the data from the Polar mission were well aligned with the "gold standard". However, the agreement of the measurements by the CIS and the RAPID instruments and later the agreement found with the measurements from the Van Allen Probes (new "gold standard" in the radiation belts) and observations from the Arase mission gave us confidence in our data.

2.2 Statistics

- Lesson 3: A value of zero is also a measurement, do not remove it without a reason

It is generally advisable to plot the data to check the type of the distribution, analyze outliers and clean the data before doing statistics. It often happens in particle observations that zeroes are ignored because they are not suitable for logarithmic plots. Also, the absence of an observation (commonly indicated by "fill values" in the data) is often not distinguished from values of zero. In plots, both are then shown as data gaps. Please remember that a value of zero is also a valid measurement, meaning that there was no particle entering the detector at a specific energy at this time. Slip-ups in post processing are less likely if NaNs (not a number, defined in IEEE 754) are used for missing values.

- Lesson 4: Be careful with interpolations

Another mistake which I often observe is interpolation of data between large data gaps. Such inappropriate interpolations often remain undiscovered in the data. Please make sure that the interpolation is reasonable. For example, the spacecraft should not cross several different plasma regimes during a data gap. I recommend avoiding interpolations or using them only for short data gaps.

- Lesson 5: Be careful with possible solar cycle related biases in statistics

You should use as much data as possible. Different phases of the solar cycle (which is 22 years!) may lead to quite different statistical results depending on the phase the sampling was done. In space observations it is often difficult to avoid biases related to the solar cycle, but you need to be aware of it.

- Lesson 6: Please calculate the uncertainty of your results

I always tell my students: please add error bars. I often see a lack of error bars in manuscripts which I review, and conclusions are made just based on a visible trend or the difference of the color in a spectrogram. It is especially dangerous if a spectrogram is made using a rainbow color map (Borland et al., 2007). The differences often appear less prominent when using perceptually uniform color maps. You should always question the uncertainty of the results and separate signal from noise. Even simple, random uncertainties can create a statistical or systematic bias (Sivadas and Sibeck, 2022). Remember that measurements have (systematic) error. We usually measure only a subset of a population, leading to sampling errors. This is very obvious but often ignored. Also calibrations of the data introduce errors but this is usually not taken into account in most studies and data sets.

In charged particle measurements, individual intensity measurements may have different uncertainties, depending on how many counts were accumulated during the time interval used to derive the intensity. In proper data archives, such as CSA, an uncertainty is provided for each measurement. This is especially important for the estimation of the spectral slope in particle distributions.

There are many methods used by statisticians for problem of separation of signal from noise and making conclusions under uncertainties, see, e.g., Wasserstein et al. (2019). Conclusions in space physics have to be made by taking into account all known uncertainties.

2.3 Machine learning

Applying machine learning techniques to observations in space physics for derivation of prediction and forecasting models can be very useful. My students found the book by Geron (2019) quite useful.

- Lesson 7: Be careful with splitting time series

One common mistake is to apply the Scikit-Learn (Pedregosa et al., 2011) train_test_split procedure on time series and getting excellent predictions that occur because a model just interpolates between adjacent times.

- Lesson 8: Make sure there is no overfitting

One easily gets excited about an excellent performance of the model on training data. However, this is often a sign of overfitting.

Namely, there is a large discrepancy in the performance on the training data and the (unseen) validation data (Ghojogh and Crowley, 2019). In this case the model just remembers the training data. In the ideal case the gap between training and validation errors should be small (Goodfellow et al., 2017).

- Lesson 9: Be careful with the interpretation of feature importances

One should be careful with interpretation of the importance of features (for predictors such as solar wind parameters) for understanding underlying physics. Machine learning models combine individual features to get the best output result and this combination can vary from model to model and also be different from considering one input and one output variable in isolation. Please also consider uncertainties of the importances.

- Lesson 10: Archive your codes, data sets and models

It is great to archive the codes, the data and the models on, for example, zenodo or make them available through GitHub, so that other scientists can build up on it in future studies.

3 Discussion: General lessons for space physicists

- Lesson 11: Do not try to accommodate the data with the expected physical picture: physics is complicated and there can be various reasons for why the data does not fit.

For example, one can expect that the mass loading from the moon Io in the Jovian magnetotail leads to a pressure increase in the magnetodisk (I searched for a long time for such signatures in the data during my PhD). However, it can be that the disk just becomes larger and the plasma pressure equilibrium does not change.

- Lesson 12: Use as many observational points as possible

The physical picture may become more complex and bring more questions but it also helps to make a global picture of a phenomena. For example, in Kronberg et al. (2017a) we used observations from 14 satellites to monitor a substorm event. This gave us an opportunity to simultaneously observe phenomena which are usually studied separately such as current sheet flapping, magnetic field dipolarization, signatures of reconnection in the near-Earth tail, dispersionless and dispersed injections and their propagation, electron acceleration by ultra low frequency waves etc.

- Lesson 13: Do not give up if you believe in your research after your manuscript is rejected: it will become better.

A couple of my now well cited papers were initially rejected. However, it can happen that one has to give up on a manuscript because one realizes that the approach was wrong.

- Lesson 14: Find a mentor

It is great to have a mentor who can give directions and set up goals. For different aspects of a scientific career one may need different mentors. Also one can learn a lot from younger people.

- Lesson 15: Be a part of such a team as at an International Space Science Institute (ISSI) in Bern

One of the best places to conceive scientific ideas is the International Space Science Institute (ISSI) in Bern, Switzerland, which allows to gather teams of experts and make them collaborate closely in an informal way for about 1 week several times. About one third of my first author papers were conceived in this place.

4 Outlook

In summary I outline several directions which in my opinion should be developed in space physics:

- Lesson 16: Combination of data and models

Communication between observers and modelers is difficult, although a lot of effort has been made in this direction. Models should be verified by observations. Observations, on the other hand, can be better understood if they are related to physical models.

- Lesson 17: Combine different energies and species

It is common to separate, for example, in the inner magnetosphere the regions by the energy of electrons: plasmasphere (less than ~3 eV), warm plasma cloak [~10 eV–3 keV, (Chappell et al., 2008)], ring current (~3–100 keV) and the radiation belts (above 50 keV). However, efforts are still needed to understand how the particle populations move along these energy scales. For example, the dynamics of cold ions and electrons is still not well understood (Delzanno et al., 2021). Assessing just bulk energies without considering cold and energetic parts may be misleading (Kronberg et al., 2017c). Measurements of heavy ions are still far from ideal and their influence on the magnetospheric dynamics is not well understood (Kronberg et al., 2014).

- Lesson 18: Look in 3D

In space physics it is common to map the data to the equatorial plane in GSE/GSM coordinates. This is fine. But a lot of new physics

is also hidden if one looks at the magnetosphere in 3D. Examples are mysterious north-south hemispheric asymmetries and diamagnetic cusps (from the particle observations point of view).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work is supported by the German Research Foundation (DFG) under number KR 4375/2-1 within SPP "Dynamic Earth."

Acknowledgments

I am thankful to Songyan Li, Aljona Blöcker and Patrick W. Daly for helpful advices.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baker, D. N., Kanekal, S. G., Hoxie, V. C., Henderson, M. G., Li, X., Spence, H. E., et al. (2013). A long-lived relativistic electron storage ring embedded in earth's outer van allen belt. *Science* 340, 186–190. doi:10.1126/science.1233518

Borland, D., Russell, R. T., and Ii, T. (2007). Rainbow color map (still) considered harmful. *IEEE Comput. Graph. Appl.* 27, 14–17. doi:10.1109/mcg.2007.323435

Bortnik, J., and Camporeale, E. (2021). Ten ways to apply machine learning in Earth and space sciences. *EOS* 102. doi:10.1029/2021EO160257

Breuillard, H., Agapitov, O., Artemyev, A., Kronberg, E. A., Haaland, S. E., Daly, P. W., et al. (2015). Field-aligned chorus wave spectral power in Earth's outer radiation belt. *Ann. Geophys.* 33, 583–597. doi:10.5194/angeo-33-583-2015

Chappell, C. R., Huddleston, M. M., Moore, T. E., Giles, B. L., and Delcourt, D. C. (2008). Observations of the warm plasma cloak and an explanation of its formation in the magnetosphere. *J. Geophys. Res. (Space Phys.* 113, A09206. doi:10.1029/2007JA012945

Daly, P. W., and Kronberg, E. A. (2022). *User guide to the RAPID measurements in the Cluster Science Archive (CSA)*. Paris: Tech. Rep. CAA–EST–UG–RAP.

Daly, P. W., Mühlbachler, S., and Kronberg, E. A. (2021). *Cluster Science Archive: Interface Control document for RAPID*. Paris: Tech. Rep. CAA–EST–ICD–RAP.

Delzanno, G. L., Borovsky, J. E., Henderson, M. G., Resendiz Lira, P. A., Roytershteyn, V., and Welling, D. T. (2021). The impact of cold electrons and cold ions in magnetospheric physics. *J. Atmos. Solar-Terrestrial Phys.* 220, 105599. doi:10.1016/j.jastp.2021.105599

Escoubet, C. P., Fehringer, M., and Goldstein, M. (2001). The cluster mission–introduction. *Ann. Geophys.* 19, 1197–1200. doi:10.5194/angeo-19-1197-2001

Geron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow*. Sebastopol: O'Reilly Media, Inc.

Ghojogh, B., and Crowley, M. (2019). *The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial*. Ithaca: arXiv e-prints, arXiv:1905.12787.

Goodfellow, I., Bengio, Y., and Courville, A. (2017). *Machine learning basics*. Ithaca: MIT Press, 98–165.

Kronberg, E. A., Ashour-Abdalla, M., Dandouras, I., Delcourt, D. C., Grigorenko, E. E., Kistler, L. M., et al. (2014). Circulation of heavy ions and their dynamical Effects in the magnetosphere: Recent observations and models. *Space Sci. Rev.* 184, 173–235. doi:10.1007/s11214-014-0104-0

Kronberg, E. A., Bučík, R., Haaland, S., Klecker, B., Keika, K., Desai, M. I., et al. (2011). On the origin of the energetic ion events measured upstream of the Earth's bow shock by STEREO, Cluster, and Geotail. *J. Geophys. Res.* 116, A02210. doi:10.1029/2010JA015561

Kronberg, E. A., Daly, P. W., Dandouras, I., Haaland, S., and Georgescu, E. (2010). Generation and validation of ion energy spectra based on cluster RAPID and CIS measurements. *Clust. Act. Archive, Stud. Earth's Space Plasma Environ.*, 301–306. doi:10.1007/978-90-481-3499-1_20

Kronberg, E. A., Daly, P. W., and Vilenius, E. (2022). *Calibration report of the RAPID measurements in the Cluster Science Archive (CSA)*. Paris: Tech. Rep. CAA–EST–CR–RAP.

Kronberg, E. A., Grigorenko, E. E., Malykhin, A., Kozak, L., Petrenko, B., Vogt, M. F., et al. (2019). Acceleration of ions in jovian plasmoids: Does turbulence play a role? *J. Geophys. Res. (Space Phys.* 124, 5056–5069. doi:10.1029/2019JA026553

Kronberg, E. A., Grigorenko, E. E., Turner, D. L., Daly, P. W., Khotyaintsev, Y., and Kozak, L. (2017a). Comparing and contrasting dispersionless injections at geosynchronous orbit during a substorm event. *J. Geophys. Res.* 122, 3055–3072. doi:10.1002/2016JA023551

Kronberg, E. A., Li, K., Grigorenko, E. E., Maggiolo, R., Haaland, S., Daly, P. W., et al. (2017c). Dawn-dusk asymmetries in the near-earth plasma sheet: Ion observations. *Dawn-Dusk Asymmetries Planet. Plasma Environments,Geophysical Monogr. Ser.* 230, 243–253. doi:10.1002/9781119216346.ch19

Kronberg, E. A., Rashev, M. V., Daly, P. W., Shprits, Y. Y., Turner, D. L., Drozdov, A., et al. (2016). Contamination in electron observations of the silicon detector on board Cluster/RAPID/IES instrument in Earth's radiation belts and ring current. *Space weather.* 14, 449–462. doi:10.1002/2016SW001369

Paschmann, G., and Daly, P. W. (1998). *Analysis methods for multi-spacecraft data. ISSI scientific reports series SR-001*. ESA/ISSI, Vol. 1. ISBN 1608-280X, 1998. *ISSI Scientific Reports Series* 1.

Paschmann, G., and Daly, P. W. (2008). *Multi-spacecraft analysis methods revisited*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Reiff, P. H. (1990). The use and misuse of statistics in space physics. *J. Geomagnetism Geoelectr.* 42, 1145–1174. doi:10.5636/jgg.42.1145

Rème, H., Aoustin, C., Bosqued, J. M., Dandouras, I., Lavraud, B., Sauvaud, J. A., et al. (2001). First multispacecraft ion measurements in and near the Earth's magnetosphere with the identical Cluster ion spectrometry (CIS) experiment. *Ann. Geophys.* 19, 1303–1354. doi:10.5194/angeo-19-1303-2001

Sivadas, N., and Sibeck, D. G. (2022). Regression bias in using solar wind measurements. *Front. Astronomy Space Sci.* 9, 924976. doi:10.3389/fspas.2022.924976

Smirnov, A. G., Kronberg, E. A., Daly, P. W., Aseev, N. A., Shprits, Y. Y., and Kellerman, A. C. (2020a). Adiabatic invariants calculations for cluster mission: A long-term product for radiation belts studies. *J. Geophys. Res. (Space Phys.* 125, e27576. doi:10.1029/2019JA027576

Smirnov, A. G., Kronberg, E. A., Latallerie, F., Daly, P. W., Aseev, N., Shprits, Y. Y., et al. (2019). Electron intensity measurements by the cluster/RAPID/IES instrument in Earth's radiation belts and ring current. *Space weather.* 17, 553–566. doi:10.1029/2018SW001989

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p0.05. *Am. Statistician* 73, 1–19. doi:10.1080/00031305.2019.1583913 <

Wilken, B., Axford, W. I., Daglis, I., Daly, P., Guttler, W., Ip, W. H., et al. (1997). Rapid - the imaging energetic particle spectrometer on Cluster. *Space Sci. Rev.* 79, 399–473. doi:10.1023/A:1004994202296 10.1007/978-94-011-5666-0_14

Williams, D. J., McEntire, R. W., Jaskulek, S., and Wilken, B. (1992). The Galileo energetic particles detector. *Space Sci. Rev.* 60, 385–412. doi:10.1007/978-94-011-2512-3_16