Data science for Space Weather services in Argentina

Jorge H. Namour[(1)(2)], Ticiano Torres Peralta[(1)(2)], María G. Molina[(1)(2)(3)], Alejandro N. Gómez[(1)(2)], Esteban Paz[(1)(2)], Guadalupe Delorme[(1)(2)], Darío Dall'ara[(1)(2)], Guillermo Amaya[(1)(2)], and Benjamín Cortés[(1)(2)].

[1] Tucumán Space Weather Center (TSWC), FACET-UNT, Av. Independencia 1800, Tucumán, Argentina.

[2] Laboratorio de Computación Científica, FACET-UNT, , Av. Independencia 1800, Tucumán, Argentina.

[3] CONICET

jnamour@herrera.unt.edu.ar

Abstract. Space Weather services rely heavily on the data. Challenges include multiple data sources, multiple formats (not always structured data), raw data (direct from the instruments), different data resolutions (in time and in space), poor metadata, data missing (instrument failure, connectivity issues, etc.), bad calibrated data, among many other issues. Bearing in mind the above considerations, we present in this work the main data pipeline design and implementation details for the Tucumán Space Weather Center - TSWC (https://spaceweather.facet.unt.edu.ar/), Universidad Nacional de Tucumán in Argentina as a new web-based system for Space Weather services.

1 Introduction

The Tucumán Space Weather Center–TSWC currently offers two types of products: publicly available and tailored for registered clients space weather products. Currently, the system uses both external data sources (e.g. from international databases and instrumentation networks) and local data sources from instruments. The upper atmosphere monitoring is performed using local instruments such as an AIS-INGV ionospheric sounder, continuous HF Doppler radar system, 2 GNSS receivers, and many more. In addition data sets from national instrument networks (e.g. RAMSAC) are also included.

Each data source provides data with different time and spatial scales and different levels of preprocessing (from images to JSON formatted data).

In the following sections, we discuss the overall architecture of the system and we present details on the different stages of the data pipeline. We also describe details on the implementation and further research steps planned.

2 System architecture

The overall data pipeline for the TSWC includes the following stages: data acquisition, preprocessing, persistent storage, processing and, visualization. In the next sub-sections, we explain in detail each stage.

2.1 Data acquisition

As mentioned before, multiple and heterogeneous data sources feed the TSWC system. We deal mostly with level 0 data [6]. This means raw data or data with low preprocessing in-situ in the instrument PC (local data source). Data arrival to the system is done in two ways: through FTP and by tailored APIs (e.g. NOAA). We also gather data from publicly available space weather databases (external data source) such as RAMSAC [7] (around 800 MB) in Argentina. This structured data is preprocessed (e.g. re-sampled, calibrated, etc) and stored in our database. The total amount of data gathered per day is around 1 TB.

2.2 Data preprocessing

At this stage we tackle several issues such as A) Multiple data formats: We unify using a unique format (e.g. JSON for structured data or time series) when needed. B) Handling missing data: If a null value has a physical meaning (e.g. F1 layer is not visible during nighttime, thus the null value of f0F1 represents this situation) the null is kept. On the other hand, sometimes a null value represents an instrument failure and an alert should be raised to the operator. Other cases are considered as well. C) Temporal data resolution management: often we compared and display time series with different resolutions, downsampling of interpolation is used depending on the case. D) An special case is the calibrated TEC derived from GNSS raw data (arriving compressed with 2 techniques: ZIP and Hatanaka compression), also a calibration process [1],[7] is done before obtaining the time series to be stored. E) Damaged files, we filtered and discarded such files. Modules from the preprocessing stage are:

- Parsers: To ensure data readability by filtering, combining instruments or techniques. Ionosonde data, for example, arrives in plain text files, we parsed and transformed them to a unique format (Tucuman and Bahia Blanca ionosonde files have different structures). The output is a JSON format.

- Cleaners: Mainly in charge of re-sampling when two-time series have different resolutions. For the Machine Learning (ML) based modeling for ionospheric forecasting, Kp index and TEC values have a different resolution (in this case we performed a K nearest neighbors interpolation for Kp). While in other cases downsampling of a high-resolution parameter is enough. Also, missing tuples issues are considered in this module. The strategies for missing data and corrupted files described above are also implemented in this module.

- TEC Calibration: This module implements the decompression and calibration for TEC explained in D). After this preprocessing the 800 MB of RINEX and IONEX files are reduced to around 20 MB final TEC product.

At the end of the pre-processing data is in level 1.

2.3 Persistent storage

MongoDB as a NoSQL database management system has been chosen because we deal with no-structured data (ionograms) and structured data from other Data Bases (DBs) and we need a flexible way of storing it [3]. Sometimes data changes in structure and we need to add it fast into our system without changing the core design. For example, initially, ionosonde data were a simple time series, and later the structure changes to add basic hourly statistics (new fields into the collection).

A tailored API serves the DB to manage the queries. Data here is in level 2 format.

The MongoDB is designed to store a collection for each parameter within the domain, e.g. we have a collection for solar wind main parameters corresponding to the interplanetary medium sub-domain. This strategy is replied for all the external data sources. On the other hand, when storing local data each collection corresponds to a type of instrument. As an example, the data from the ionosonde (both in Tucumán and Bahia Blanca) are stored in a single collection.

2.4 Analysis and processing

Additional processing is applied here. We perform baselines calculation for the main upper atmosphere parameters by implementing different statistics from the data stored (e.g. media, median, special averages -based on quiet days and last 27 days). These baselines are used to set thresholds for the alarm system of space

weather events. We have also implemented machine learning analysis for the now-casting of the upper atmosphere parameters. The ML modeling is not in production yet (not deployed on the site). Currently, we are testing 2 models: a) Single station f0F2 nowcasting (3 h. ahead using a window step of 3 hours back) based in LSTM. Preliminary results are promising with MSE <3. Further validation is still needed before production. (b) Using an NNARX scheme we forecast TEC derived from GNSS data 24 hs. We use the Kp index as the exogenous forcing to the system, and TEC time series 24 hs behind. Also here we have a good preliminary performance. This analysis is not currently available in an operative mode.

2.5 Visualization

With respect to the visualization, TSWC presents different options of plots. We implemented this visualization with the HighCharts library [5]. Additionally, we offer image format display for selected instruments (e.g. ionograms). We developed an API between the back-end (in Python) and the front-end (Django)[2] in Flask Framework [4].

3 Conclusions and further work

This is a work in progress and we planned several other software tools such as:

- Operative implementation using short-term prediction of ionospheric parameters using our LSTM model, we are discussing the proper manner to re-train the model as new data arrives, proper metrics, among other important considerations.
- Implementation of an online WebApp for specialists. Among the features implemented there is a dynamic tool to correct automatically scaled ionograms.
- Implementation of a database of calibrated TEC for Argentina with online access services (partially implemented).

The TSWC is constantly evolving with new products, data, and instruments. In this scenario, the Data Science approach gives us the opportunity to enhance our capabilities in a robust manner.

Acknowledgments:

We acknowledge Laboratorio de Telecomunicaciones (FACET- UNT) for the instrumentation deployment and maintenance.
This work is partially supported by the research projects PIUNT-E689 and PICT04447.

References:

1. Azpilicueta, F., Brunini, C. et al.: Calibration errors on experimental slant total electron content (TEC) determined with GPS. J Geod 81, 111–120 (2007). https://doi.org/10.1007/s00190-006-0093-1
2. Django Homepage: https://www.djangoproject.com/
3. Eoin Brazil, Kristina Chodorow, Shannon Bradshaw: MongoDB: The Definitive Guide (3rd edition), O'Reilly ISBN 9781491954461.
4. Flask Homepage, https://palletsprojects.com/p/flask/
5. HighCharts Homepage, https://www.highcharts.com/
6. Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt, and Gerd Heber: Scientific data management in the coming decade, SIGMOD Rec. 34, 4 (2005), 34–41. DOI:https://doi.org/10.1145/1107499.1107503
7. RAMSAC Homepage: https://www.ign.gob.ar/