Brazilian Journal of Geophysics, **40** (2022), 4, 523–531
Brazilian Geophysical Society
ISSN 2764-8044
DOI 10.22564/brjg.v40i4.2187

A NEW APPROACH FOR LOW-LATITUDE IONOSPHERIC SCINTILLATION PREDICTION

Pedro Alexandre dos Santos [1*], Stephan Stephany [1], and Eurico Rodrigues de Paula [2]

[1]Instituto Nacional de Pesquisas Espaciais – INPE, Department COPDT, São José dos Campos, SP, Brazil
[2]Instituto Nacional de Pesquisas Espaciais – INPE, Department DIHPA, São José dos Campos, SP, Brazil

*Corresponding author: retiarus@gmail.com

ABSTRACT. The prediction of ionospheric scintillation is a current research topic. Data-oriented models have been proposed since there is not a mathematical model able to simulate the complex ionospheric mechanisms for the prediction of scintillation. Data-oriented models employ machine learning algorithms that are trained with known, post-mortem data, and then perform predictions from new data. An ensemble method based on sampling boosting applied for a set of decision trees is employed for the prediction of scintillation in a single low-latitude location in Brazil. The prediction was performed considering two classes, occurrence or absence of scintillation. The method uses as input temporal series of the scintillation index S_4, the total electron content (TEC), and some geomagnetic and solar indexes for that location. The data encompass the summer months of the previous solar cycle years (2010-2018) and were split into mutually-exclusive training and validation/test sets. The prediction performance was promising, showing a potential to be developed for operational use. Data limitations related to time series extension, or also to a balanced distribution of samples/instances covering both classes, since scintillation occurrences are usually scarce, can be further developed as new data are available.

Keywords: S_4 index prediction, machine learning, gradient boosting, GNSS station network

INTRODUCTION

Many socioeconomic activities rely on GNSS (Global Navigation Satellite System) satellite constellations, including the navigation of aircraft, land vehicles, and fluvial/maritime vessels. The Earth's ionosphere presents density irregularities, which cause fluctuation in phase/amplitude of the radio-frequency signals that traverse it. This effect is called ionospheric scintillation and may affect GNSS and/or telecommunication signals. Scintillation is associated with complex electromagnetic phenomena in the ionosphere, which are even described by some mathematical models but still require more resolution. Scintillation is a current research topic due to its relevance. Since low-latitude scintillation is frequent in Brazil, researchers at INPE, the National Institute for Space Research, have been studying ionospheric scintillation and plasma bubbles for more than 50 years (de Paula et al., 2021) and participating in related international cooperation programs and projects with other research institutes and universities worldwide. In addition, INPE hosts the EMBRACE program (Brazilian Studies and Monitoring of Space Weather)[1].

The absence of accurate mathematical models to simulate the ionosphere, and thus predict the occurrence of scintillation, gave rise to data-oriented models, which are based on machine learning. One such model "learns", using known post-mortem data that may include time series of scintillation, total electron content (TEC), and other related geomagnetic and solar variables and indexes. This work employed scintillation and TEC data from 2010 to 2018, covering almost the entire solar cycle 24.

The scintillation data were acquired from the GNSS networks LISN (Low-latitude Ionospheric Sensor Network, Peru), CIGALA-CALIBRA (Concept for Ionospheric Scintillation Mitigation for Professional GNSS in Latin America - Countering GNSS

[1]https://www2.inpe.br/climaespacial/portal/pt/

High Accuracy Applications Limitations due to Ionospheric Disturbances in Brazil), and ICEA (Institute for Airspace Control, Brazil). The S_4 data provided by these networks were then employed to generate S_4 maps by means of inverse distance weighted interpolation (Vani, 2018), which interpolates for each grid point of the map the S_4 values of all available GNSS stations, but weighting each value by the inverse of the distance station-grid point.

The TEC data were generated using the RINEX files provided by the RMBC/IBGE (Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS, Brazilian Institute for Geography and Statistics) using the method adopted by EMBRACE/INPE (Otsuka et al., 2002; de Sousa do Carmo, 2018), which generates a TEC map combining the TEC values of the available stations for the considered time interval.

These maps have some data limitations related to the quality, periodicity, and eventual out-of-service status of particular GNSS stations. Another limitation is that, for the early years of the considered solar cycle, the number of GNSS stations of these networks was lower than now, affecting the monitoring of scintillation. The number of stations has increased over the years.

This work denotes as raw data the S_4 and TEC values extracted from the corresponding maps since these values undergo a specific pre-processing in order to have a more suitable representation for the machine learning algorithm. In the case of this work, the time series of these maps allowed the time series generation of the S_4 and TEC values for a single location with low magnetic latitude, the city of São José dos Campos (SJC), Brazil.

The proposed data-oriented model aims to perform short-term predictions (30 minutes ahead) of the S_4 index during the night over SJC. The prediction performance of the model was evaluated by some specific metrics, showing a potential for future operational use at EMBRACE/INPE.

Ionospheric Scintillation

Nowadays, the Global Navigation Satellite System (GNSS) is commonly employed in many activities, even in critical ones, such as aerial navigation. Besides GPS, other GNSS's were proposed and implemented (or are being implemented) around the world, such as the Russian GLONASS, the Chinese COMPASS, and the European GALILEO. However, GNSS radio-frequency signals may be disturbed by the ionosphere dynamic, mainly by undergoing phase and amplitude fluctuations along the line of sight between GNSS satellites and receivers on ground stations or on-board vehicles, aircraft, or vessels. This phenomenon is called ionospheric scintillation, being measured in amplitude by the S_4 index, the normalized standard deviation of the signal amplitude I acquired at a rate of 50 Hz:

$$S_4^2 = \frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2}. \quad (1)$$

The Earth's ionosphere presents many phenomena directly or indirectly related to the Sun dynamics. The Sun emits radiation and particles that affect the Space Weather between Sun and Earth, including its electric and magnetic fields. In their turn, Earth's electric and magnetic fields are also affected, as well as the ionosphere density of ions and electrons, generating non-homogeneous and non-isotropic structures of different scales of size, ranging from meters to hundreds of kilometers. Such structures represent ionospheric irregularities, being the most important the ionospheric bubbles, regions with depletion of ions and electrons. The boundaries of these irregularities may present high-density gradients that may cause scintillation.

In addition, there is a very known plasma transport process from the magnetic Equator to low latitudes, in both Earth hemispheres, called the Equatorial Ionization Anomaly (EIA), which causes the rise of equatorial plasma, and eventually, the rise of bubbles that migrate towards the magnetic South and East, and may suffer expansion or contraction in size. Bubbles are formed after sunset and generally end up at the beginning of dawn. The occurrences of ionosphere bubbles start to intensify from October or November of each year, presenting peaks along the Summer, and start to decrease from March on.

The general morphology of bubble occurrence and associated mechanisms is well understood, but bubble occurrences present high day-to-day variability, making their prediction difficult (Abdu, 2019). This poses a challenge for its prediction since this work proposes the use of time series that may embed such variability hampering the learning phase of the proposed algorithm.

Prediction of Ionospheric Scintillation

Currently, notwithstanding the existence of mathematical models able to simulate the forming and evolution of ionospheric bubbles, they would require higher temporal and spatial resolutions in order to predict scintillation (Yokoyama, 2017). There are some empirical models that employ or not historical data, but they are inaccurate for the prediction of scintillation (Wernik et al., 2007; Retterer, 2010; Béniguel and Hamel, 2011). Scintillation may impact different socioeconomic services, and although being the occurrence of scintillation restricted to some regions and periods of time, it can critically affect aerial navigation and landing/take-off procedures, which increasingly rely on GNSS services. The prediction of the occurrence of scintillation is actually the prediction of the scintillation index S_4 in a point of the globe, making the abstraction of the existence of a two-dimensional S_4 field. Such index is actually a

measure of the amplitude uniformity of the radio-frequency signal emitted by a GNSS satellite and reaching a ground receiver (there is also another measure considering the phase of the signal). For instance, worldwide institutions acquire raw data from networks of GNSS stations, to derive S_4 values, and the resulting set of these values at discrete points of the globe is then interpolated, providing regional or global S_4 maps that can be thought of as a two-dimensional field.

In the absence of a suitable mathematical model, a common approach is to employ a data-oriented model that is based on a machine learning algorithm. Similar approaches have been employed in different areas such as weather forecast or hydrology prediction and rely on large amounts of data and on the use of supercomputers (Camporeale, 2019). A machine learning algorithm has training, validation, and test phases using mutually exclusive sets of historical data. In the training phase, the algorithm "learns" using known data, adjusting learning parameters (for instance, weights and biases, in the case of a neural network). In the validation phase, other parameters inherent in the algorithm configuration, called hyper-parameters, are optimized and, in the test phase, the resulting data-oriented model is evaluated.

A particular work (McGranaghan et al., 2018) was the first to propose the prediction of scintillation in high magnetic latitudes but relies on the persistence paradigm, i.e. assuming that scintillation may last for some hours, suggesting that the ionosphere has a kind of non-linear memory. However, many works, before and after that one, only proposed data-oriented models, as some works involving the authors (Rezende et al., 2010; Lima et al., 2014, 2015). In particular, a fresh search on the Google Scholar website revealed only a single work in recent years related to scintillation prediction in low magnetic latitudes, as proposed here.

The referred work proposes scintillation prediction using a data-oriented model based on gradient boosting (Zhao et al., 2021). It employs S_4 data of the 2012-2020 period from Brazilian GNSS stations, and also the virtual height of the ionospheric F-layer base provided by specific radars called ionosondes. The proposed prediction is made for the considered GNSS station considering two classes, occurrence or not of scintillation, respectively being below or above the threshold of $S_4 = 0.5$, in the period 2012-2020. The prediction is daily (actually for each night, since scintillation occurs mostly after sunset), yielding a single maximum S_4 value for each night. The training data use S_4 values averaged every 5 minutes, but only the maximum of these values is considered for each night.

The prediction results of that work are promising, but there is a single S_4 value predicted for the entire night, not taking into account its temporal variability. In addition, the machine learning algorithm was trained using only a single maximum value for each night, not considering that, at the same time, the GNSS station could lock different satellites in different azimuthal angles that may yield S_4 values in a range of values from absent/weak, moderate or strong scintillation. Therefore, besides the limitation of a single-value prediction for each night, there is a strong bias to make the prediction of only high S_4 values that are more likely associated with the occurrence of strong scintillation in any azimuthal direction around the GNSS station. The approach is interesting, but the resulting prediction performance given by the F_1 score, a standard metric explained in the next section, was below 85 % in all cases, showing room for improvement.

DATA AND METHODOLOGY

Machine learning algorithms that perform prediction or classification are the same, only focusing on different problems. The prediction/classification performance relies on a suitable amount of good-quality data, which must undergo selection and preprocessing in order to be suitable for the considered algorithm. For instance, raw data may be discretized or normalized. A predictor/classifier algorithm is then applied, "learning" from known data in the training phase, thus generating a specific data-oriented model. Known data refer to input data composed of predictive features plus the expected numerical or categorical output data. This work employs two target classes, occurrence, and absence of scintillation. The next phase is the validation, the tuning of the hyper-parameters of the algorithm, in order to improve the model performance. Finally, the test phase uses another "unknown" set of data (i.e., not including the target class). An analysis of the prediction performance is then made by some selected metrics. These steps, although applied to prediction, are part of a more general Data Science process called Knowledge Discovery in Databases (KDD), which is performed iteratively in order to optimize each step.

Prediction metrics

The standard way to evaluate the prediction/classification performance is by a confusion matrix that summarizes correct and incorrect predictions/classifications for a number of instances to be predicted/classified. In this case, there are only two classes (occurrence/absence of scintillation), being the matrix 2×2. The diagonal elements contain the count of correct predictions for the occurrence class (true positives TP and true negatives TN), while the remaining elements are the counts of mispredictions for the same class (false positives FP and false negatives FN). Some prediction/classification metrics can be derived from the confusion matrix, all having values in the range of 0.0 to 1.0, being the unity their optimal value. These metrics are defined below considering instances of the scintillation occurrence

class:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(TP + FN)}. \tag{5}$$

Usually, it is intended to predict the majority of the scintillation occurrence instances (TP), while minimizing the misprediction occurrences (FN). As a consequence, the number of mispredicted non-occurrences (FP) is increased, and some trade-off must the chosen in such adjustment in order to minimize the number of FNs without increasing too much the number of FPs. The F_1 score is a metric that better optimizes these numbers, being adopted in this work.

Data pre-processing

The data employed in this work cover the 2010-2018 period, but only the summer months, which usually present more scintillation events. The data include time series of the S_4 and TEC indexes, both over SJC.

This work proposes a local S_4 prediction for SJC using local data, and thus no information about the location is employed. S_4 and TEC data of nearby GNSS stations were implicitly considered in the generation of the corresponding maps.

Besides the time series of the S_4 and TEC indexes, other time series were employed to account for the geomagnetic activity and the level of solar activity. The time series of geomagnetic indexes include the AE (Auroral Electrojet), Dst, Ap, and Sym-H/Sym-D, while the time series of solar indexes include the F10.7 (Solar flux), SN (daily sunspot number), IMF (Interplanetary Magnetic Field) and its components B_z and IMF B_y, solar wind velocity, and pressure, respectively, V_{sw} and P_{sw}. In addition, the time series of the virtual height of the F-layer base ($h'F$) was also employed.

The sampling rates of S_4, TEC, and other variables and indexes are higher, but the corresponding time series are downsampled to 30-minute by their maximum, in order to reduce the complexity of the problem and processing needs. The remaining data have a poorer temporal resolution. Predictions refer to finding out if there will be absence or occurrence of scintillation at a given time over SJC using the S_4 threshold value of 0.2 to distinguish between these classes. Many pre-processing steps were required, as described below.

- Generation of time series of S_4 maps from GNSS raw data similarly as performed in the GNSS network CIGALA-CALIBRA/INCT (Vani, 2018), but reducing the S_4 values corresponding to the set of neighboring IPPs (Ionosphere Piercing Point) at each grid point using the maximum value, instead of the average, and using a 10-minute integration interval;
- Generation of time series of TEC using the method adopted by EMBRACE/INPE (Otsuka et al., 2002; de Sousa do Carmo, 2018), also with 10-minute resolution;
- Removal of eventual negative TEC values from the TEC maps, since they have no physical meaning; these negative values may be generated as a flaw of the method that estimates the absolute TEC value from the relative one; each negative value is replaced by the average of the lower non-negative 5% quantile;
- Extraction of the TEC and S_4 values using the corresponding values of the grid point that is nearest to SJC, obtaining the time series of these variables;
- Re-sampling of the S_4 and TEC and the other time series for 30-minute resolution;
- Filling in of the missing values of the S_4 and TEC time series using the last valid value that precedes each missing value;
- Normalization of the S_4 and TEC time series values to the 0.0 to 1.0 interval.
- Use of the libraries Tsfresh (Christ et al., 2016, 2018) and Tsfel (Barandas et al., 2020) for the extraction of predictive features from fixed-length windows of the considered time series, such as maximum value, average, median, Fourier transform coefficients, etc.;
- Calculation of the time derivatives of each time series and estimation of non-linear correlations between different time series or their derivatives;
- Calculation of the average and the climatology deviation is given, for each value of the series, by the difference between the value and the average of precedent values for a fixed-length window;
- Filtering of the predictive features composed of the original features plus the extra features obtained in the former steps; features with variance lower than 0.1 are filtered out since they are possibly meaningless in the learning process;
- Standardization of all predictive features in order to yield zero mean and unitary variance for each one;

- Use of the SMOTE function of the Imblearn library (Lemaître et al., 2017) to generate synthetic instances from the set of original ones, in order to obtain a balanced distribution of instances (occurrence/absence of scintillation) in the training set.

The resulting data from these pre-processing steps are a feature vector for each instant of time, according to the 30-minute resolution. This vector is composed of all the considered predictive features and may be considered as an instance of the input data. Every time a prediction is devised, the feature vector for that instant of time applies.

Data partitioning and validation schemes

Data partitioning generally splits the set of predictive features (in this case, time series of the different indexes) into mutually-exclusive training and validation sets. The first is used to generate a model, this is, it is applied in combination with an optimization algorithm to fit a function to the set, such that the function gives the best possible result given a loss function. The set is used to validate the model performance; in some cases, it can be used to select (tuning) a class of functions (different types of models) from a list of possible functions.

Data partitioning schemes are heuristics that define how to divide the data among the datasets. In this work, a particular double-nested partition scheme was employed, composed of an external partition level to generate a testing and a pseudo-training dataset, and an internal partition level that splits the external pseudo-training dataset into internal training and validation datasets. Therefore, such a nested partition scheme generates training, validation, and test sets. Both the internal and external partitioning levels used here employ one of the following schemes: Time Series Cross Validation with Gaps (GapKFold) or Time Series Cross Validation Gap Walk Forward (GapWalkForward). When using GapKFold, both internal and external schemes are GapKFold; the same is valid for the GapWalkForward.

- Time Series Cross Validation with Gaps (GapKFold): the set of predictive features is divided into k mutually exclusive subsets, each one sorted in time and separated by 30-day gaps, being (k-1) subsets used for training and the remaining one for validation; this scheme is iteratively repeated by replacing the validation set by one of the training sets, rendering k different models (named as k-Cross Validation); the resulting prediction is given by a polling scheme using these k models (here, k was adopted as 5);

- Time Series Cross Validation Gap Walk Forward (GapWalkForward): similar to *GapKFold*, but generates datasets ordered in time, in a way that all data in the training dataset precede the data in the validation dataset. K-different models can be generated: in the first interaction, an ordered subset from all samples sets starts in the first sample; this is the training set. For each interaction, the sample number in the training set will increase from the same amount while keeping the validation set with a fixed size. In the last interaction, all the available samples will be covered when considered the training and validation subsets together (here, k was also adopted as 5).

Gradient Boosting Ensemble algorithms

Two machine learning algorithms of this work are state-of-the-art ensemble methods, i.e. based on a finite set of classifiers/predictors. Two Gradient Boosting Ensemble algorithms are used in the training and validation of the same data resulting in two different models. Test results are given by a polling scheme using these two models. The algorithms are the XGBoost (Chen and Guestrin, 2016) and the CatBoost (Prokhorenkova et al., 2018; Dorogush et al., 2018). However, a specific stopping criterion (early stopping) is applied for each model along the training in order to avoid over-fitting. At every iteration of the training, the prediction performance based on the validation subset is evaluated. Since such performance degrades for too many iterations, an optimal number of training iterations can be defined for each model.

It is worth noting that predictions are made for 6 prediction times, from 30 to 180 minutes of antecedence (spaced by 30 minutes) and, therefore, there are actually 6 models, each one resulting from the combination of the models derived from XGBoost and CatBoost for a given prediction time, using the GapKFold partitioning scheme, and another 6 models, using the GapWalkForward scheme. These 12 models are evaluated separately resulting in 12 prediction results at each instant of time. Each one implies in training, validating, and testing the related subsets, which comprise the number of instances with 30-minute resolution.

Tables 1 and 2 show the number of instances for training, validation and test, respectively for the GapKFold and the GapWalkForward schemes, considering 30-minute predictions. The training and validation subsets of these tables present the number of instances after performing the balancing of instances according to the classes (occurrence/absence of scintillation). In these tables, there are 5 subsets for each validation level, but in the table related to the GapKFold scheme, only first two subsets are shown for the outer level. In the table related to the GapWalkForward scheme, only the first (smaller) and the last (larger) subsets of the outer level are shown. In both, for each outer level, all corresponding 5 subsets of the inner level are shown.

Table 1: Number of instances for the training, validation, and testing subsets using the GapKFold partitioning for the 30-minute prediction, showing subsets (I) and (II) of the outer level cross-validation.

Subset	I					II				
	1	2	3	4	5	1	2	3	4	5
Training	17406	16216	16450	17006	18700	15508	14288	14564	15158	17242
Validation	3256	3256	3255	3255	3255	2987	2987	2987	2986	2986
Test	4406	4406	4406	4406	4406	4406	4406	4406	4406	4406

Table 2: Number of instances for the training, validation, and testing subsets using the GapWalkForward partitioning for the 30-minute prediction, showing the smaller (I) and the larger (V) subsets of the outer level cross-validation.

Subset	I					V				
	1	2	4	4	5	1	2	3	4	5
Training	974	2144	3322	4514	5706	5836	10282	15172	19350	23566
Validation	596	596	596	596	596	3043	3043	3043	3043	3043
Test	3671	3671	3671	3671	3671	3671	3671	3671	3671	3671

RESULTS

This section summarizes the prediction results for the different models, prediction times (30-180 minutes), and data partitioning schemes, using the prediction metrics described before in section Prediction Metrics. It is worth noting that such metrics presented were obtained from the counts of TP, TN, FP, and FN elements for the test subset of each case. As expected, longer prediction times degrade the prediction performance, but for 30 and 60-minute antecedence, the results show the F_1 score above 90 % for the GapKFold scheme, except for the last subset (V) of the GapKFold. The same tendency was observed for the V subset of the GapWalkForward scheme, but the results were poorer.

The same behavior is observed for all prediction times and both partitioning schemes and can be explained as follows. Regardless of the constant-size subsets of the GapKFold scheme, and the "cumulative" size of the GapWalkForward scheme, both V subsets refer to more recent instances; in this case, from the period 2017-2018 (solar minima). The corresponding training and validation subsets precede the V subset, being of the period 2011-2017 (solar maxima). Therefore, these poor results for the V subset are due to the use of data from different phases of the solar cycle.

In general, considering subsets I to IV, the performance of the GapWalkForward scheme was lower than the GapKFold one. The reason is that the GapWalkForward scheme preserves the temporal ordering of the training, validation, and testing sets, and thus the training employs only "past data" in relation to validation and testing sets. On the other hand, the GapKFold training may include samples of "future data" in relation to the validation and test sets. This is a kind of leakage from one set to another and generally improves the prediction performance since in the training phase the model may learn information about the "future".

A second problem may arise using the GapWalkForward scheme, which generates 5 models using increasing sizes of the training and validations sets, since, in the generation of the earlier models, the size of the training and validation sets may be too small precluding the learning process.

However, despite its lower prediction performance, the GapWalkForward scheme would be better for operational purposes, since it does not rely on "future data".

Tables 3 and 5 present results for 30 and 60-minute predictions, respectively for the GapKFold and the GapWalkForward schemes, showing accuracy, precision, and F_1 score for each case. Tables 4 and 6 present results for 30-60-90-120-180-minute predictions for both schemes but showing only the F_1 score.

CONCLUSIONS

The monitoring and prediction of ionospheric scintillation occurrence are current research topics and have importance for GNSS warning systems, mainly those concerned with aerial navigation, and aircraft take-off/landing procedures. Currently, there is only a sole recent work proposing scintillation prediction at low latitudes (Zhao et al., 2021), as already mentioned. Similarly to the proposed approach, it yields

Table 3: Prediction results for 30 min and 60 min antecedence using the GapKFold scheme.

Test subset	30 min			60 min		
	Accuracy	Precision	F_1 **score**	Accuracy	Precision	F_1 **score**
I	0.99	0.97	0.97	0.98	0.95	0.95
II	0.95	0.92	0.93	0.94	0.92	0.91
III	0.93	0.93	0.92	0.92	0.92	0.91
IV	0.95	0.90	0.91	0.94	0.88	0.89
V	0.93	0.72	0.79	0.90	0.57	0.60

Table 4: Prediction results for 30 min to 180 min antecedence using the GapKFold scheme.

Test subset	30 min	60 min	90 min	120 min	150 min	180 min
	F_1	F_1	F_1	F_1	F_1	F_1
I	0.97	0.95	0.95	0.94	0.95	0.93
II	0.93	0.91	0.92	0.92	0.89	0.90
III	0.92	0.91	0.88	0.87	0.88	0.90
IV	0.91	0.89	0.86	0.84	0.84	0.86
V	0.79	0.60	0.48	0.47	0.48	0.49

Table 5: Prediction results for 30 min and 60 min antecedence using the GapWalkForward scheme.

Test subset	30 min			60 min		
	Accuracy	Precision	F_1 **score**	Accuracy	Precision	F_1 **score**
I	0.75	0.61	0.61	0.69	0.50	0.42
II	0.90	0.84	0.86	0.87	0.82	0.83
III	0.88	0.85	0.86	0.83	0.82	0.81
IV	0.93	0.82	0.83	0.92	0.78	0.79
V	0.92	0.68	0.74	0.87	0.52	0.51

Table 6: Prediction results for 30 min to 180 min antecedence using the GapWalkForward scheme.

Test subset	30 min	60 min	90 min	120 min	150 min	180 min
	F_1	F_1	F_1	F_1	F_1	F_1
I	0.61	0.42	0.54	0.42	0.43	0.43
II	0.86	0.83	0.79	0.82	0.74	0.81
III	0.86	0.81	0.78	0.79	0.78	0.78
IV	0.83	0.79	0.75	0.74	0.73	0.72
V	0.74	0.51	0.47	0.47	0.47	0.47

a categorical value (occurrence/absence class), but a single value for all the nights. This may show that such prediction is difficult and a state-of-the-art research topic.

The approach proposed here is very complex but rendered promising prediction results. It was preceded by many simpler approaches with poorer prediction performance, thus requiring a higher complexity in terms of data pre-processing, data partitioning and validation, different algorithm combinations, corresponding prediction models, etc. Processing demands are high for a personal computer but would easily be tackled by any state-of-the-art multiprocessed server. In addition, the proposed approach can be further improved and evaluated, as detailed below for future work. The goal is to reach a prediction performance high enough to be effectively implemented for operational use in the EMBRACE/INPE program and eventually be adapted for use in other countries located at low magnetic latitudes.

Future work

The future work includes the assessment of the proposed approach using currently available data, providing feedback to refine the employed machine learning algorithm by optimizing its hyper-parameters, and enlarging the training data set as more data are made available or even exploring different preprocessing options for the data. It also devises the extension of the presented approach of scintillation local prediction to other Brazilian GNSS stations, eventually generating a training data set using data of all the considered stations but testing separately for each one. Other improvements are also devised:

- Use of more recent ionospheric data for the period 2019-2022, which present better quality, being provided by the 4 networks of GNSS stations available in Brazil (LISN, CIGALA-CALIBRA/INCT, ICEA, RBMC/IBGE), which are being updated with Septentrio[2] receivers, while TEC data are provided by the RBMC/IBGE;

- Use of new S_4 maps to be generated by a new methodology being recently proposed by the author's research group to extract more accurate values for SJC;

- Extension of the proposed prediction for other locations in Brazil that have GNSS stations; this would be performed in two steps: (i) performing training and validation using multilocality data, and evaluating the resulting prediction performance, and (ii) enhancing the predictive data by including location-specific magnetic coordinates.

REFERENCES

Abdu, M. A., 2019, Day-to-day and short-term variabilities in the equatorial plasma bubble/spread F irregularity seeding and development: Progress in Earth and Planetary Science, **6**, 11, doi: 10.1186/s40645-019-0258-1.

Barandas, M., D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, 2020, TSFEL: Time series feature extraction library: SoftwareX, **11**, 100456, doi: 10.1016/j.softx.2020.100456.

Béniguel, Y., and P. Hamel, 2011, A global ionosphere scintillation propagation model for equatorial regions: J. Space Weather Space Clim., **1**, A04, doi: 10.1051/swsc/2011004.

Camporeale, E., 2019, The challenge of machine learning in space weather: Nowcasting and Forecasting: Space Weather, **17**, 1166–1207, doi: 10.1029/2018SW002061.

Chen, T., and C. Guestrin, 2016, XGBoost: A scalable tree boosting system: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 785–794. doi: 10.1145/2939672.2939785.

Christ, M., N. Braun, J. Neuffer, and A. W. Kempa-Liehr, 2018, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – a Python package): Neurocomputing, **307**, 72–77, doi: 10.1016/j.neucom.2018.03.067.

Christ, M., A. Kempa-Liehr, and M. Feindt, 2016, Distributed and parallel time series feature extraction for industrial big data applications: Computing Research Repository - CoRR arXiv:1610.07717, doi: 10.48550/arXiv.1610.07717.

de Paula, E. R., A. O. Moraes, M. A. Abdu, J. H. A. Sobral, I. S. Batista, E. A. Kherani, H. Takahashi, and E. Costa, 2021, Long term studies on scintillation and plasma bubbles, *in* 100 Years of the International Union of Radio Science: URSI Press, 29, 557–562.

de Sousa do Carmo, C., 2018, Estudo de diferentes técnicas para o cálculo do conteúdo eletrônico total absoluto na ionosfera equatorial e de baixas latitudes: Master dissertation (Postgraduate Course in Space Geophysics / Solar-Terrestrial Environmental Science), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, SP, Brazil. (English title translation: Study of different techniques for the calculation of the total electronic content in equatorial and low latitude ionosphere. Available at http://urlib.net/8JMKD3MGP3W34P/3QJNJHE).

Dorogush, A. V., V. Ershov, and A. Gulin, 2018, CatBoost: gradient boosting with categorical features support: Computing Research Repository - CoRR arXiv:1810.11363, doi: 10.48550/arXiv.1810.11363.

Lemaître, G., F. Nogueira, and C. K. Aridas,

[2] https://www.septentrio.com/en/products/gnss-receivers

2017, Imbalanced-learn: A Python toolbox to Tackle the Curse of Imbalanced datasets in Machine Learning: Journal of Machine Learning Research arXiv:1609.06570, **18**, 1–5, doi: 10.48550/arXiv.1609.06570.

Lima, G. R. T. d., S. Stephany, E. R. de Paula, I. S. Batista, and M. A. Abdu, 2015, Prediction of the level of ionospheric scintillation at equatorial latitudes in Brazil using a neural network: Space Weather, **13**, 446–457, doi: 10.1002/2015SW001182.

Lima, G. R. T. d., S. Stephany, E. R. de Paula, I. S. Batista, M. A. Abdu, L. F. C. Rezende, M. G. S. Aquino, and A. P. S. Dutra, 2014, Correlation analysis between the occurrence of ionospheric scintillation at the magnetic equator and at the southern peak of the equatorial ionization anomaly: Space Weather, **12**, 406–416, doi: 10.1002/2014SW001041.

McGranaghan, R. M., A. J. Mannucci, B. Wilson, C. A. Mattmann, and R. Chadwick, 2018, New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning: Space Weather, **16**, 1817–1846, doi: 10.1029/2018SW002018.

Otsuka, Y., T. Ogawa, A. Saito, T. Tsugawa, S. Fukao, and S. Miyazaki, 2002, New technique for mapping of total electron content using GPS network in Japan: Earth and Planetary Science Letters, **54**, 63–70, doi: 10.1186/BF03352422.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2018, Catboost: Unbiased boosting with categorical features: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., 6639–6649.

Retterer, J. M., 2010, Forecasting low-latitude radio scintillation with 3-D ionospheric plume models: 2. Scintillation calculation: Journal of Geophysical Research: Space Physics, **115**, doi: 10.1029/2008JA013840.

Rezende, L. F. C., E. R. de Paula, S. Stephany, I. J. Kantor, M. T. A. H. Muella, P. M. de Siqueira, and K. S. Correa, 2010, Survey and prediction of the ionospheric scintillation using data mining techniques: Space Weather, **8**, S06D09, doi: 10.1029/2009SW000532.

Vani, B. C., 2018, Investigações sobre modelagem, mitigação e predição de cintilação ionosférica na região brasileira: Ph.D. thesis, Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Campus Presidente Prudente, Presidente Prudente, SP, Brazil. (Available at http://hdl.handle.net/11449/153701).

Wernik, A. W., L. Alfonsi, and M. Materassi, 2007, Scintillation modeling using in situ data: Radio Science, **42**, RS1002, doi: 10.1029/2006RS003512.

Yokoyama, T., 2017, A review on the numerical simulation of equatorial plasma bubbles toward scintillation evaluation and forecasting: Progress in Earth and Planetary Science, **4**, 37, doi: 10.1186/s40645-017-0153-6.

Zhao, X., G. Li, H. Xie, L. Hu, W. Sun, S. Yang, Y. Li, B. Ning, and H. Takahashi, 2021, The prediction of day-to-day occurrence of low latitude ionospheric strong scintillation using gradient boosting algorithm: Space Weather, **19**, e2021SW002884, doi: 10.1029/2021SW002884.

dos Santos, P. A.: designed, implemented, tested and evaluated the machine learning models; **Stephany, S.**: supervised the implementations and tests, wrote and revised the manuscript; **de Paula, E. R.**: provided expertise in ionospheric physics, evaluated the test results and revised the manuscript.

Received on May 26, 2022 / Accepted on March 30, 2023