Measurement 208 (2023) 112426

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

A Machine Learning hourly analysis on the relation the Ionosphere and Schumann Resonance Frequency☆

Carlos Cano-Domingo [a,*], Ruxandra Stoean [b,c], Gonzalo Joya [d], Nuria Novas [a], Manuel Fernandez-Ros [a], Jose Antonio Gazquez [a]

[a] Ceia3, Engineering Department, University of Almeria, Spain
[b] Romanian Institute of Science and Technology, Romania
[c] University of Craiova, Romania
[d] Departamento Tecnología Electrónica, University of Malaga, Spain

ARTICLE INFO

Keywords:
Schumann Resonance
Extreme Low Frequency
Machine Learning
Ionosphere
Explanatory models
Electro-magnetic signal analysis

ABSTRACT

The Schumann Resonances arise from the constructive interference of dozens of near-simultaneous lightning strikes every second, mostly located in the tropics. Characterizing the Schumann Resonance signal variation is a complex task due to the number of variables affecting the electromagnetic composition of the ionosphere and the Earth. We describe a novel approach for investigating the behavior of this variation by focusing on specific hours of the day. This study further explores this preliminary influence by means of a machine learning framework composed of six conceptually different algorithms. Fourteen external variables, related to the ionosphere condition, are considered as the predictors for the monthly Schumann Resonance frequency variation along five years of real data, for each of the first six modes and separated by the hour of the day. The results provide a clear evidence of the importance of selecting a particular hour to observe the influence of the Ionosphere parameters on the Schumann Resonance frequency variation.

1. Introduction

SR constitutes electromagnetic signals that propagate along the earth–ionosphere cavity in the *Extremely Low Frequency* (ELF) band [1]. SR signals have been deeply studied by their frequency spectrum. Frequency central modes are well known to be around 7.8 Hz, 14 Hz, 20 Hz, 26 Hz, 33 Hz 39 Hz for the first six modes of SR, Price [2], see Fig. 1.

The electromagnetic cavity properties are primarily influenced by the Earth surface and the lower ionosphere electromagnetic condition. Due to the lack of significant changes in the Earth surface, changes in the lower ionosphere are well-documented as one of the most important sources of modifications in the SR spectrum. From a theoretical and simulated point of view, many authors have tried to characterize the spectrum in the steady condition. However, the conductive profile of the ionosphere has a high dependency on multiple conditions. The Source–Observer distance has been undoubtedly established as the key part of changes in the variation of the SR signal [11]. As an example, in [12], the authors established a strong correlation between their frequency variation experimental data and the Source–Observer distance. It is also interesting to observe the different relationships between the distance of the three main thunderstorm centers and the variation of each SR frequency mode. The variation within the SR spectrum is largely attributed to the intensity and distribution of these global lighting activities, both the intensity and frequency variation. It is also that the solar effect is considered one of the most critical factor for changes in the lower ionosphere conductivity, a periodic diurnal and seasonal pattern. As a consequence, the regular pattern can be observed in the SR frequency spectrum, Tatsis et al. [13].

☆ The authors thank the Andalusian Institute of Geophysics. The Ministry of Economics and Competitiveness of Spain financed this work, under Project TEC2014-60132-P, in part by Innovation, Science and Enterprise, Andalusian Regional Government through the Electronics, Communications, and Telemedicine TIC019 Research Group of the University of Almeria, Spain and in part by the European Union FEDER Program and CIAMBITAL Group. by I+D+I Project UAL18-TIC-A025-A, the University of Almeria, and the European Regional Development Fund (FEDER). R. Stoean was supported by grants of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, Romania, project number 178PCE/2021, PN-III-P4-ID-PCE-2020-0788 and project number 408PED/2020, PN-III-P2-2.1-PED-2019-2227, within PNCDI III.

* Corresponding author.
E-mail addresses: carcandom@ual.es (C. Cano-Domingo), rstoean@inf.ucv.ro (R. Stoean), gjoya@uma.es (G. Joya), nnovas@ual.es (N. Novas), mfernandez@ual.es (M. Fernandez-Ros), jgazquez@ual.es (J.A. Gazquez).

https://doi.org/10.1016/j.measurement.2022.112426
Received 22 May 2022; Received in revised form 18 October 2022; Accepted 30 December 2022
Available online 6 January 2023

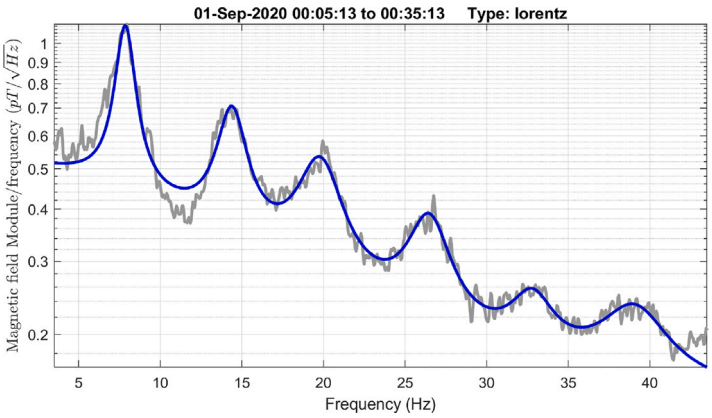


Fig. 1. Example of a $\mathcal{H}_{NS}$ spectrum of the Sierra de Filabres observatory. Blue line: Lorentzian fit. Gray Line: Raw signal.

Table 1
Ionosphere variables - short description and references.

Variable name	Group	Abrv	Unit	Brief explanation	Ref
Total Electron Content	Ionosphere	TEC	1×10^{-16} m^{-2}	Total Electron Content of the whole Ionosphere	Reinisch and Galkin [3]
Geomagnetic Index Ap	Earth	Ap	Ap index	The Ap-index is the earliest occurring maximum 24 h value	Matzka et al. [4]
Geomagnetic Index Kp	Earth	Kp	Kp index	The K-index of the 3-hourly range in magnetic activity.	Matzka et al. [4]
Global temperature	Earth	Temp	Degree	Global temperature anomaly	NASA GISS [5] and Lenssen et al. [6]
Adjusted solar flux	Solar	SolarFlux	1×10^{-22} W m^{-2} Hz^{-1}	Adjusted electromagnetic power received by the sun	Tapping [7]
Lightning activity - USA	Earth	LightningUS	Lightning strikes per day	Lightning activity measured in US	Cecil et al. [8]
Ionosphere sporadic E layer	Ionosphere	hEs	km	Minimum virtual height of the sporadic E layer	Reinisch and Galkin [3]
Ionosphere E layer	Ionosphere	hE	km	Minimum virtual height of the E layer	Reinisch and Galkin [3]
Ionosphere F1 layer	Ionosphere	hF1	km	Minimum virtual height of the F1 layer	Reinisch and Galkin [3]
Ionosphere F2 layer	Ionosphere	hF2	km	Minimum virtual height of the F2 layer	Reinisch and Galkin [3]
Simulated TEC in D layer	Ionosphere	hD	1×10^{-16} m^{-2}	TEC in the height layer at the observatory location	Bilitza [9]
Total number of sunspots	Solar	Sunspot	Events per day	A historical measure of the sun power	Clette and Lefèvre [10]
Northern hemisphere sunspots	Solar	SunspotN	Events per day	A historical measure of the sun power - North hemisphere	Clette and Lefèvre [10]
southern hemisphere sunspots	Solar	SunspotS	Events per day	A historical measure of the sun power - south hemisphere	Clette and Lefèvre [10]

SR has received increased attention in the last five years due to the growing number of ELF sensor stations. Many studies have aimed to research the hypothetical link between these electromagnetic ELF signals and other natural phenomena. The use of SR towards forecasting earthquake or predicting biomedical indicators is currently under study. The relation between the ionosphere and Schumann Resonance has not been fully established from either an analytical point of view or a simulation model. Furthermore, these previous works focused on a specific aspect of this relationship and not the general prototype, which would be an extensive model of far different contributions to a highly complex problem, Tritakis et al. [14].

The lightning activity is recognized as being the most crucial source of the SR wave energy. The seasonal variation of SR has been studied centered in their relation with changes of activity and location of three major thunderstorms. In previous studies, the reached conclusion was that, focusing on the hours with the maximum activity of these thunderstorm centers, it is possible to see a similar pattern between the SR frequency peak and the variation of the level of activity of this thunderstorm, in terms of lightning discharges rate, Soler-Ortiz et al. [15].

There is a vast amount of literature on the theoretical analysis of the SR signal. Various approaches have been proposed to study the relationship between the electromagnetic signal and the cavity properties from an analytical point of view, with a considerable advance. As early as in [16], the authors proposed a model based on a two-scale height for characterizing the conductivity of the ionosphere in a mathematical model. In their comparison of mathematical approaches, Galuk et al. [17] shows a review of three solutions and concludes that the heuristic knee model does not grant a realistic conductivity profile of the atmosphere. In a recent study of Prácser et al. [18], assumptions seem to be well-founded to explain the impact of the Day–Night Asymmetries in the SR spectrum. Unfortunately, these approaches try to describe the SR steady condition without providing insights about SR variations over a long period.

On the other hand, simulation studies have been performed using highly complex electromagnetic systems with substantial results. In [19], they developed a software framework to reproduce the general behavior of the SR signal. Other approaches were focused on modeling the 3-D electromagnetic wave propagation as in [20,21], using an electromagnetic software package. This result shows a significant milestone to properly characterize the averaged behavior of the SR signal. There

Table 2
First most important variables using Shapley method for the ML model with and without adding the Source–Observer distance.

SR mode	Without distance		With distance		
1	Ap	hF2	AM	Ap	hF2
	Sunspot south	Sunspot total	AS	Sunspot south	Sunspot total
2	hEs	hF2	AM	hEs	hF2
	hD	Lightning USA	AS	Sunspot total	Lightning USA
3	hEs	hF2	AM	hEs	hF2
	Sunspot south	Lightning USA	AS	hD	Lightning USA
4	hE	hF2	AM	hE	hF2
	hEs	Lightning USA	AS	hEs	Lightning USA
5	hEs	hF2	AM	hEs	hF2
	Lightning USA	Sunspot south	AS	Lightning USA	Sunspot south
6	hE	hF2	AM	hE	hF2
	hEs	Sunspot south	AS	hEs	Sunspot south

is still considerable uncertainty with regards to the variation of the ELF signal, and simulation fails to address this condition due to the extreme complexity of the 3D Earth–ionosphere Electro-Magnetic model.

A growing body of literature has investigated the SR signal using experimental data, with two different approaches.

- **Characterize SR variation**: some recent studies have focused on showing the variation of SR frequency and intensity of the first mode in different locations around the globe. In [22], the authors analyze the long-term variation of the SR in a UK observatory, focusing on the first resonant mode. A longer time span is exposed in [23], where near 20 years of data comparison are presented. The result shows a significant difference between the Arctic and Antarctic SR observatory. In [13], the authors highlight the diurnal and seasonal differences using the data gathered in the Northwest of Greece. They also added valuable information about the correlation with lightning activity.
- **Relation with other phenomena**: recent studies have focused on exploring the usage of SR for earthquake forecasting. In [24], the authors investigate the relation of SR with a large earthquake in Mexico, focusing on the first three SR modes with a window of 15 days before and after the earthquake. Hayakawa et al. [25] also deepens the theoretical and experimental analysis of the relationship between SR anomalies and two offshore earthquakes in 2021. In [26], the authors explore the use of ML to forecast earthquakes based on SR signals. Recent studies about the correlation with local lightning discharges have been performed in [27] with a visible outcome. In [28], the predicted relation with the solar cycle is pointed out. It also shows a great agreement between certain SR intensity records and the long-term variation of solar fluxes. Finally, it draws our attention to study the relation between SR and Heart Rate in [29]. They reported a preliminary study about the possible effects of the electromagnetic wave in the ELF band and changes in the population heart rate.

These methods show a potential initial relationship between a specific SR tendency and a unique event in a particular location of time and space. However, the primary defect in these research works is that they do not exploit the analysis using the general SR variation as a whole over an extensive period and related with possible variables that can affect the propagation condition. Furthermore, the use is limited in general to the first SR mode, while, in our experience, all first six modes provide valuable information. This group has presented two studies in line with this purpose. The first one is to automatically segment and extract the individual transient ELF events, Domingo et al. [30], and their main features, with the aim of using an artificial computing model to explore their relation with other phenomena. The second one targets a correlation study about the relation of the ionosphere variables with the SR frequency [In press].

This work is based on the previous findings that there are distinct hours of the day when there is a particular influence between 14 of the ionosphere variables and the SR frequency variation.

The aim of this study is to demonstrate that the relationship between the lower ionosphere and the SR frequency variation is significantly different for each of the hours of the day. With this in mind, we collect the monthly average over a 5-year period for 14 given external variables (Table 1). These 14 variables have been chosen due to a previous work we have carried out in [31] in which the 14 variables are used and explained. The main reason is to select around five variables related to the state of the ionosphere, around five about the state of the solar effect, to take into account the well-documented solar influence in the SR, and around five about the state of the earth. This input will be in correspondence with the SR frequency at a given hour of a month and in a certain SR mode. Hence, 144 Combined SR models are obtained, one for each of the 6 SR modes and every of the 24 h. We subsequently appoint a framework of six traditional ML algorithms to develop different regression methods for each of the recorded vectors, using the 14 external variables as predictors. For the sake of simplicity, we will name these external variables as ionospheric variables in the following. These ionospheric variables are classified into three groups as can be seen in the mentioned Table 2: Solar data, Ionospheric data and Earth data. We have the assumption that there is a physical link between the SR frequency variation and the ionospheric variables for each Hour of the day and each SR mode, following Eq. (1). But, there are no analytical solutions to model it. Thus, this work is based on the premise that the ML methods are able to discover and capture this relationship as well as the influence of each ionospheric variable in the previous mentioned equation.

$$SR_{Hour,Mode} = f(hD, hF2, hF1, \ldots, SunspotN) \tag{1}$$

The ML techniques used in this research can be divided into three categories, concerning the ML approach. Each of these categories focuses on detecting and inferring different types of variation in their prediction. Finally, the ultimate aim of this research is to group the six ML methods obtaining each particular $SR_{Hour,Mode}$. Each obtained group is called the Combined $SR_{Hour,Mode}$ Model. Thus, we obtain 144 Combined SR Models, which are capable of observing any type of variation between different hours and modes in the SR signal. Each Combined SR Model is evaluated, focusing on the importance of each predictive variable and its accuracy.

In the pre-experimental phase, a time-series forecast approach was also considered. However, the relation between the SR frequency variation and retarded versions of the 14 predictor variables were not relevant. Specifically, multiple delay versions of all predictors were used as independent inputs for the preliminary ML model. In conclusion, the results do not fulfill the requirement of this research. For this reason, the time dimension has not be considered in this research. A complete explanation will be exposed in Section 3. Also, although data scarcity could be considered as a problem for the ML application, due to the monthly average procedure, it is possible to see that with the collected amount of records the results are stable and robust. To contrast the assumption that SR frequency variation is far more

related to ionospheric variable than the SR intensity variation we have performed a case of study in Section 4.8. While we have demonstrated that certain variables have a significant dependence on the frequency value, it is additionally shown that the same methodology applied to the intensity data does not reveal any particular relationship.

This article is divided into six sections. Section 1 gives a brief overview of the hypothesis and aim of this study. Section 2 describes the data used in the study, i.e the ionospheric variables, and also the SR experimental data. The methodology is outlined in Section 3, along with a brief explanation of each of the six ML algorithms used. In Section 4, the results of our experiments are outlined, along with a detailed discussion of the most relevant outcomes. Finally, the conclusions are drawn in the final section.

2. Data

As will be covered in Section 3, the problem is formulated as a regression task, where the independent and dependent variables are constituted in the following manner. First, the vector is composed of 14 components, where each one is the monthly average of the daily value of each of the 14 ionosphere variables, which are the predictors in the ML framework. One vector is considered for each month for the interval of study (Jan 2016–Dec 2020). Therefore, there are 60 vectors (12 months × 5 year) are taken as the input of the ML methods. The output of the ML methods is represented by the monthly average of the SR mode frequency for each of the hour of the day. The output is thus a vector of 60 positions. Consequently, since the SR frequency is reported over 24 h in 6 modes, 144 Combined SR models will be constructed in order to discover the relationship between the external variables and the SR frequency.

2.1. External data

We have summarized all the information about the 14 external variables used in this work in Table 1. They are gathered from the same period, i.e. 2016 to 2020. In conclusion, there are 60 input vectors (12 months × 5 years) of 14 predictive values. For the sake of consistency, we have named the simulated content of the D layer as hD. Because the important result for this study is the use of it as an explanatory variable

2.2. Schumann resonance data

The SR experimental data to be used as ground truth has been obtained using the Sierra de los Filabres ELF observatory developed by the research group TIC019 at the University of Almería. The data is arranged in 30 min long segments. Each of them is processed using the average periodogram technique, Parra et al. [32]. The frequency and intensity of the central peaks are extracted using a six Lorentzian fit algorithm. A detailed explanation of the observatory can be found in [33], and the data analysis in [34]. An example of a SR 30 min register can be seen in Fig. 1. Although the two orthogonal fields $\mathcal{H}_{NS,EW}$ are measured in our observatory, the substantial differences between the two fields make it almost impossible to develop a ML approach for both channels. As a consequence, we have focused this research on $\mathcal{H}_{NS}$. The differences between sensors are related to a particular interference that we have very close to our ELF observatory which only affects substantially the $\mathcal{H}_{EW}$ sensor.

In this study, the frequency peak value of the first 6 modes is separated into 24 different time series. These are averaged by month to reduce the variability of the natural phenomena along with the perturbance measures due to the low value of the electromagnetic signal. Other time period has been considered, however, due to the winter months difficulties strongly affecting the consistency of the results, some register has been removed due to the high part of noise.

To sum up, the final composition of the SR data are 144 (24 h × 6 modes time series), each composed of 60 values, one for each month during the 5 years from 2016 to 2020. Averaging the value per month is a necessary step in order to reduce the variability.

It is also important to remark that we evaluated the records through a comparison with the SR intensity variation (Section 4.8), in which case there is clearly no dependence between them and the ionosphere variables.

3. Methodology

The regression formulation consists of samples of 14 ionosphere records predicting the SR frequency vectors. The length of the measurement is adjusted to 60 points. All the ionosphere variables are normalized to a mean of 0 and a standard deviation of 1.

The methodology for this regression problem is based on the used of 6 different ML algorithms to identify the relation between the ionosphere variables, used as predictors, and the SR signal separated by hour and by mode. Each of the 144 SR signals (6 modes × 24 h) is used as the dependent variable, whose variation will be forecasted using the 14 ionospheric parameter variation series as independent variables in the ML regression tasks, after being learned from the ground truth (Fig. 2).

The relevance of each ionosphere variable will be assessed independently by every used ML approach and a combined view will be subsequently expressed. The combination of the decisions of multiple ML techniques is important [35,36], as each method addresses the problem from a different perspective; as such, the final framework will encompass a multi-faceted model.

For this point on, the SR vectors will be referred to as the dependent variable while the ionosphere variables will be presented as the independent ones. An overview of the methodology can be seen in Fig. 2. As it was commented in the introduction, it is clear from the literature that the Source–Observer distance is the most critical factor for the SR frequency variation. However, the data available about the location of the three thunderstorm centers is based on a previous study [37], using an analytical model. This model only depends on the month of the year, so the variance between years cannot be taken into account. As it was mentioned before, the study is focused on discovering the influence of different variables and the SR frequency variation. For this reason, we have decided to not include this analytical variable due to the lack of enough experimental data. However, as part of the validation methods, a subsection has been introduced.

3.1. ML algorithms

In order to identify and to adapt different types of variations and relations between the dependent and the independent variables, three different groups of ML algorithms have been used. An example of each ML model prediction can be seen in Fig. 3.

1. Analytical methods focus on finding simple relationships from the regression formulation: RIDGE linear regression, Friedman et al. [38], and MARS, Friedman [39].
2. Black-box methods are complex methods with strong non-linear behavior, which can predict the output with outstanding performance. However, these methods are not explanatory.: ANN, Lantz [40], and SVM, Karatzoglou et al. [41].
3. Ensemble methods specialize in iteratively improving model accuracy: RF, Breiman [42], and GBM, Ridgeway [43].

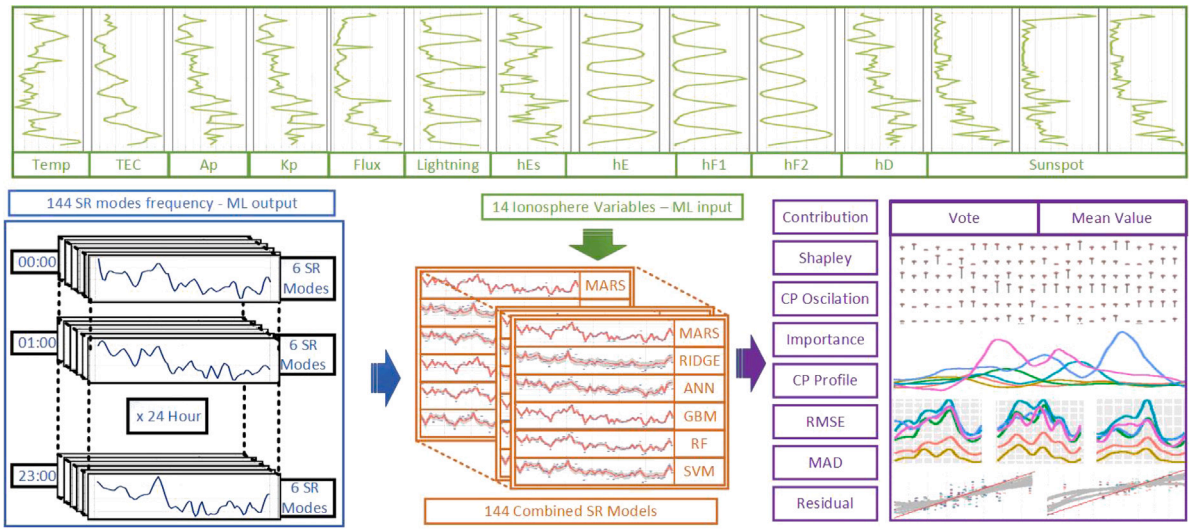


Fig. 2. Overflow of the methodology: A framework of 6 ML methods are appointed to investigate the dependence of the SR frequency variation, assessed in 24 h and 6 modes as ground truth, on 14 ionospheric variables. Several measures of variable relevance, the error and residuals are computed from the combined SR model.

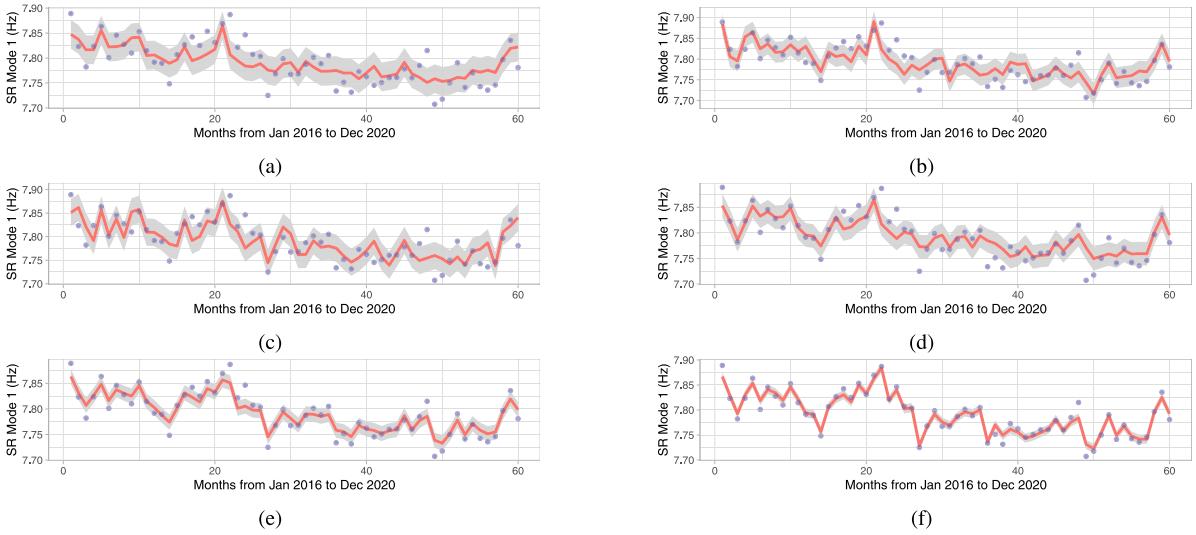


Fig. 3. Example of a run of every selected ML method to adapt to the data from the frequency values of SR 1st at 00:00 h from the 60 months. The red line denotes the predicted values, the blue points stand for the original values and the shadow area points to the resulting RMSE. a: RIDGE, b: MARS, c: ANN, d: SVM, e: RF, f: GBM.

3.1.1. *RIDGE linear regression (see Fig. 3(a))*

Ridge regression is a particular regularization method where all the attributes have to be considered in the tuning process. These methods enlarged the classical approach of linear regression adding constraints or regularization of the estimated coefficients. This additional step allows the model to reduce the variance and adapt it to the given variables.

In RIDGE, linear regression is performed by adding a new term to the minimization of the cost function, primarily given by the sum of squared residuals (SSE), as shown in Eq. (2). The tuning parameter γ exploits the option of penalizing the increment of the β_i coefficients for the attributes.

$$minimize\left(SSE + \gamma \sum_{i=1}^{k} \beta_i^2\right) \tag{2}$$

3.1.2. *MARS (see Fig. 3(b))*

This ML method is an algorithm focused on modeling the behavior of high dimensional data, having a non-linear relationship between them and the prediction. The algorithm creates a set of piece-wise linear functions to adapt to the non-linear data. The number of models is automatically calculated based on the data. The process involves a recursive partitioning algorithm for capturing high order iterations.

In Eq. (3), the MARS function $\hat{f}(x)$ is composed of the addition of the polynomial functions weighted by a coefficient c_i. In the case of this study, $B(x)$ is a two degree polynomial for each of the segments.

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x) \tag{3}$$

3.1.3. *ANN (See Fig. 3(c))*

This method is inspired by the brain structure, with a vast amount of interconnections and a very flexible composition. The crucial part of the ANN is the process of learning the features, which is done by readjusting the weight values in each constituent neuron. The ANN can internalize complex model relationships using the input samples and their output.

Each activation neuron follows Eq. (4), where b_k is its corresponding bias value, while $x_{k,i} * w_{k,i}$ is the learned weight applied to its input. In this methodology, two hidden layers are used with a logistic formulation as a activation function for the hidden layer and a ReLu

function for the output neuron.

$$output_k = f\left(b_k + \sum_{i=1}^{N_k} x_{k,i} * w_{k,i}\right) \quad (4)$$

3.1.4. SVM (see Fig. 3(d))

The main idea of this ML method is to find a hyper-plane that represents the best separation among the data points. In addition, the separating hyper-plane must have the highest distance to the points which lie closest to itself, i.e. the support vectors. The method is also known as maximum margin classifier. SVM are very close to ANN methods due to the definition of the kernel functions, which also allow non-linear separations.

3.1.5. RF (see Fig. 3(e))

A generalization of the classical decision tree approach is exploit with this ML method. In RF, multiple new sets are created by sampling the original data collection, where each of them has a different combination of features. The results of the decision tree models applied are then aggregated. The RF model allows the creation of a very generalized model, in which the length of the data set is not crucial, due to the randomization process of creating the subsets.

3.1.6. GBM (see Fig. 3(f))

Another approach to improve the decision tree algorithm different from that of RF is performed in GBM, by following a sequential assemble learning model. It is based on the development sequence of a decision tree in which the prediction of the next step is always more accurate than the previous one. The main idea is to overcome the error in the previous learner prediction. As such, each tree predicts the error of the previous one — thereby gradually boosting the performance.

3.2. ML metrics

In order to compare the values of the different ML methods, a normalization was performed using DALEX [44]. The individual value of a variable metric is divided by the sum of that metric in all variables for a given ML method.

3.2.1. Contribution

This metric evaluates how much the prediction changes when adding more variables to the model for a specific test case. It starts from the intercept and predicts the dependent variable, then adds a variable and checks the difference between the predicted value with and without that variable. The process continues until all the variables have been included in the model. The main drawback of this method is that the first attributes added to the model present more importance if there is dependency among the variables, because all the variability has been considered in the previous predictor. The contribution metric has a strong dependency on the order.

3.2.2. Shapley

In order to remove the order dependency, multiple orderings are exploited. The results for each variable are averaged and assigned to each function individually. The principal disadvantage of this method is that, if more than one variable is highly dependent on another, this metric could greatly underestimate their relevance.

3.2.3. Ceteris-Paribus Profiles

It is possible to see how the predicted value changes if only one predictor is considered for variation, while the others remain at their average value. *Ceteris-Paribus* (CP) Profiles are focused on a specific predictor in order to see how the output changes when a small change in only this variable is considered for a given test case.

3.2.4. Ceteris-Paribus Oscillation

This metric aggregates Ceteris-paribus Profiles considering all the predictor variables. The idea behind this metric is that it is possible to compare the influence of different variables, by evaluating the magnitude of a change in the predicted value when only one variable is changed. The possible values of the variable are taken in an interval centered around its mean.

3.2.5. Importance

Considering values outside the data set is performed by the use of this metric. The hypothesis of this metric is that it is possible to know the importance of a given variable to the model prediction, by checking the increment of the error when this particular variable is removed from the model. It is possible to compare the influence of two variables, by comparing which one of the two increases the residual error more, if eliminated from the model.

3.2.6. Performance

Performance is a common metric to understand the accuracy of a given model. The two most used approaches for evaluating the goodness of fit for regression methods are used in this research, i.e. *Root Mean Squared Error* (RMSE) and *median absolute-deviation* (MAD).

- **RMSE** (Eq. (5)): the most used goodness-of-fit metric that sums the total amount of difference between the predicted and the original value. The sum is squared to have the same scale as the prediction data. This metric is differential, which is crucial for the optimization problem.
- **MAD** (Eq. (6)): this metric allows to measure the goodness of fit when the data presents a significant amount of outliers. The lack of mathematical properties for optimization is the major drawback.

$$RMSE = \sqrt{\sum_{i}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (5)$$

$$MAD = median(|\hat{y}_1 - y_1|, |\hat{y}_2 - y_2|, \ldots, |\hat{y}_n - y_n|) \quad (6)$$

3.2.7. Residual analysis

The study of the differences between the predicted values and the original ones corresponds a very important measure to evaluate the performance of a prediction. In this research we have also been concerned with the visualization of the distribution of residuals in relation with the magnitude of the training data. The purpose is to show that the residuals are larger or smaller when the real value is farther from the mean.

3.3. Combined SR model

The six ML methods of each Hour and SR mode are composed in a Combined SR model, which constitutes the final step to extract representative metrics about the relationship of the SR frequency variation and each of the ionosphere variables. This Combined SR model is constituted by the following two approaches. Two different techniques are used to aggregate the importance.

3.3.1. Ranking Vote

For each ML model, the most five relevant variable are taken. This number was chosen as a compromised between the common three ranking and half of the ionospheric variables used. The Ranking Vote value is the count of all the variables that are selected in more than five ML rankings. This approach allows us to identify whether in a particular hour the dependent variable can be related to a specific ionospheric variable or not. If the vote value is 0, there are no clear link between any of the ionosphere variables and the dependent variable. On the

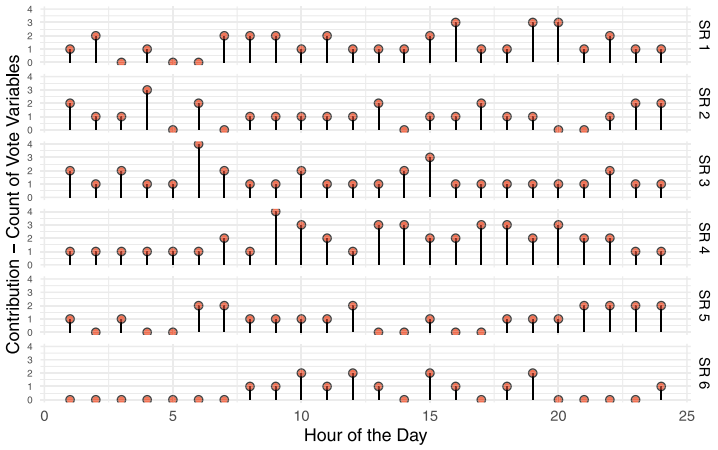


Fig. 4. Result of the vote for the **contribution** metric for the six first SR modes. The *y*-axis represents the number of variables that are among the five most important for at least 5 out of the 6 ML methods.

other hand, if the value is 2 or greater, the dependent variable cannot be linked with a single ionospheric variable only. However, when the vote value is 1, we can establish a relationship among the evolution of the dependent variable with a particular ionosphere variable.

3.3.2. Mean value

The Combined SR mean value result is the average of the outcomes of the six ML methods. This metrics allows to considered all the possible relationship that each of the methods can identify.

4. Results and discussion

Following the application of the discussed methodology, the obtained results will be presented and commented with respect to the chosen metrics. For an objective assessment of the outputs, the Leave-1-out cross-validation was conducted on the current data set.

4.1. Contribution

As it was mentioned in Section 3.2.1, the contribution metric shows the added importance of the independent variables when a test case is presented. The main drawback is the lack of consistency when multiple independent variables are related.

In Fig. 4, the result of the contribution vote is shown. It can be seen that in the SR mode 6th the vote results are almost 0 in the interval [19:00–6:00]. It is consistent with the fact that this SR mode is hardly distinguishable from the background noise, as it can be seen in Fig. 1. This level of noise triggers that the ML approaches cannot agree on which variables are the most important. In contrast, the 4th SR mode shows an interval [22:00–7:00] in which, almost every hour, there is just one relevant variable for most of ML methods. The 1 value has a remarkable connotation when it comes to seeing the modeled relationship. When the ML methods agree on just one ionospheric variable, it can be possible to establish a clear predominant relationship between this ionospheric variable and the SR mode in a specific hour.

It is also interesting to notice the decreasing tendency of the 4th SR mode [8:00–12:00], which provides additional insight into the validity of our method. In the 1st SR mode, it can be seen that the value falls to 0 three times during the first hours of the day. It means that none of the variables are relevant for this hour. However, it is always possible to see a value of one or more for the rest of the days. The 3rd SR mode shows a particular different behavior from the rest of the SR modes, with around just one variable in the vote in the interval [15:00–4:00]. In the 5th SR mode, the vote value falls to zero at seven different hours, which could mean that there is not a relevant ionospheric variable with enough importance to make the ML methods agree on the same. This figure also provides additional information if a vertical view is observed. We can see that at 7:00, all the SR modes have a value of 1 except the 1st with two variables. The same happens with the 17:00 h, but in this hour, the 4th SR mode disagrees with a value of 3. These two hours show a particular effect that could imply that the ionospheric variable contribution is very prominent at this time, and it is easy to distinguish the relation with each particular model.

The 24 h averaged most relevant ionospheric variable for each mode can be seen in Fig. 5.

It is important to remark that only two ionosphere variables monopolize the most relevant ionospheric predictors in contribution: hF2 and Lightning. This result is consistent with the figure before. For the first mode, the contribution value is below 10% until 9:00, from that point is always higher than 20%. Therefore, the value falls to 0 in some moments within the first hours of the day. The 2nd SR modes contribution is highly concentrated on two peaks at 2:00 and 17:00; these two shapes can also be observed around the same hour in the vote plot. The 3rd and 4th modes showed a high value within all the 24 h, which in the vote means that there is no contribution value below 1. The 3rd SR mode also has a very high peak in the interval [5:00–12:00] which in the vote result can be seen as a trough among these hours with a value of 1 in the middle. This ionosphere variable hoards the contribution at these hours. In accordance with the vote result, the 5th SR mode is higher than 15% for the interval [5:00–11:00] and [17:00–22:00]. This pattern is also clear in the vote when the values drop to 0 outside this interval. It is important to notice that even the most relevant predictor, considering the average value for its contribution, for the 6th SR has a minimal contribution for all the hours, reaching a maximum of 15% around 11:00. As a consequence, the ML algorithms do not have any critical variable in common during most of the hours; just when that ionosphere variable is around the maximum, we can see an agreement among the methods applied.

4.2. Shapley

In Section 3.2.2, the Shapley method was introduced. The main difference with regard to the previous method is that Shapley uses multiple random permutations in the order of predictors to obtain an averaged metric, whose mean does not depend on the specific order for getting the contribution value. There are several points of similarity between the vote for Contribution in Fig. 4 and the Shapley one. The main difference is regarding the 3rd SR mode, where the values for the Shapley vote are around 1, with a high peak at 15:00. It can be appreciated that the 6th SR mode presents small peaks at 4:00, 10:00, 15:00, and 19:00.

The four most relevant predictors for each of the SR modes are shown in Fig. 6. It can be seen that the hF2 variable plays an important

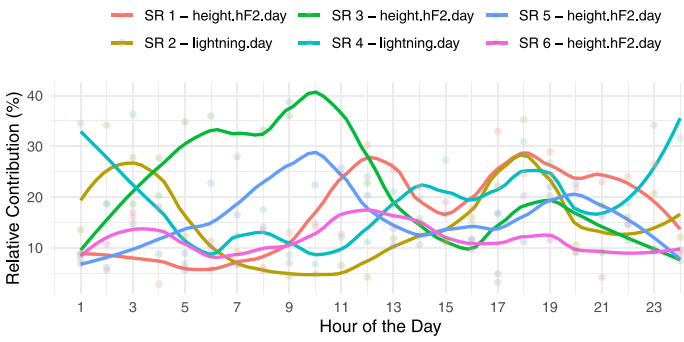


Fig. 5. Highest normalized contribution variable for each of the six SR modes for the 24 h.

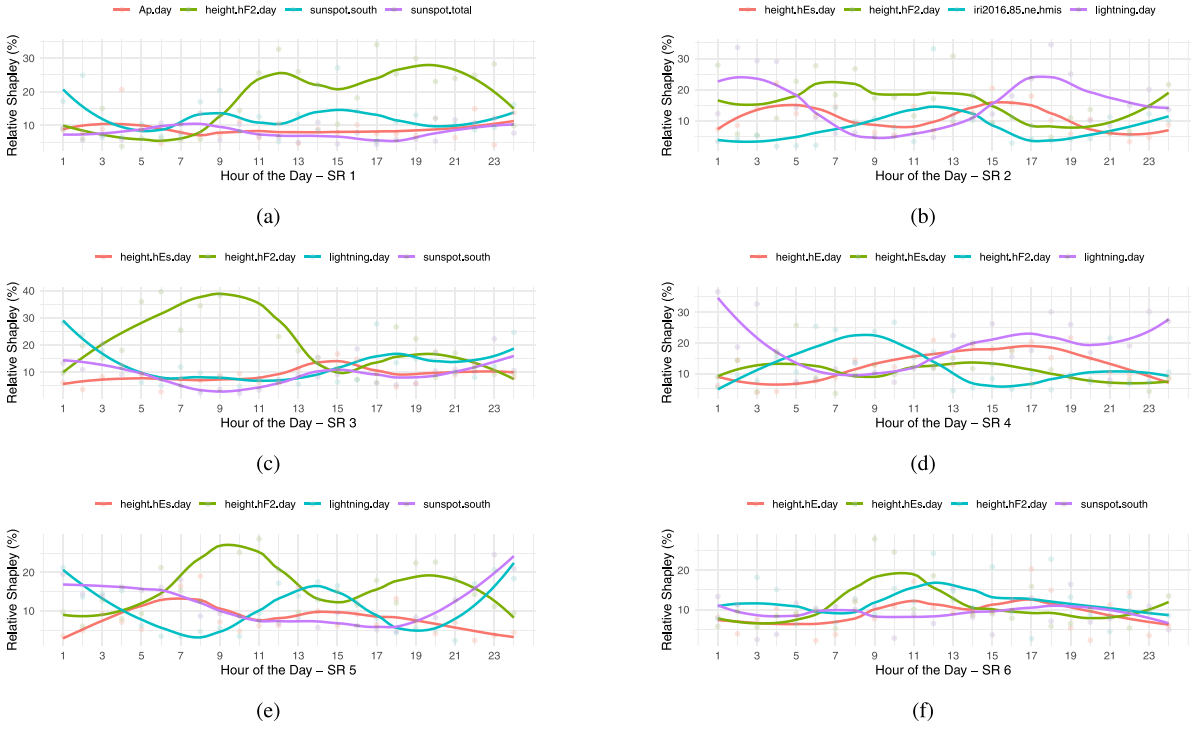


Fig. 6. The normalized **Shapley** result for the first six SR modes for the 24 h. a: SR mode 1, b: SR mode 2, c: SR mode 3, d: SR mode 4, e: SR mode 5, f: SR mode 6.

role for all the modes. Therefore, hF2 has the maximum Shapley value for the odd modes, 1st, 3rd and 5th. For the even modes the hF2 is also important but the highest Shapley value is the US lightning for the 2nd and 4th and hEs for the 6th mode. This noticed different behavior for the groups of even and odd SR modes is also found in the literature exposed in Section 1.

A special observation that needs to be however highlighted is that the shape of the hF2 bears stronger similarities among the two groups. The 1st mode, Fig. 6(a), shows another important variable, SunspotS with values around 12% for almost all 24 h but two peaks at 8:00 and 14:00. Noticeably, the number of sunspots seen from the southern hemisphere is present in other modes but the highest values are reached for the 1st. Therefore, the 1st mode is the only one which has two sunspot related variables in the four most important. This leads to a preliminary result that the highest solar influence is present in the first mode.

The 2nd mode, Fig. 6(b), shows that the Shapley value is shared among the four most important variable. hF2 and Lightning-US have a complementary behavior, one is always around 20% when the other drops to 10%.

Surprisingly, hD appears just in this mode, but with a substantial importance between 7:00 and 13:00.

The common pattern between the 2nd mode and the hD can be produced by the fact the hD is the closest layer to the earth surface and, contrary to the 1st mode, it is not affected by the solar influence at the same scale. hF2 in the 3rd mode in Fig. 6(c) shows the highest Shapley value among all the modes, 40% almost between 5:00 and 10:00. The interval hours are almost complementary to the 1st SR mode. It is important to remark that the peak is 50% higher than in the 1st mode. When the hF2 falls, the Lightning-US and hEs increase their value up to 20% in the interval from 17:00 to 19:00. The significant value of the Lightning-US from 22:00 to 2:00, around 30%, is also noticeable. Surprisingly, the pattern of the 5th mode reveals great similarities with the 6th mode, Fig. 6(e). These two modes do not only have the same four important variables but also the patterns show similarities in the four variables, with recognizable differences. Broadly speaking, the study of the differences among the two modes at specific hours can narrow a possible predictor of some ionospheric variable, i.e. hF2 at 5am.

Lightning-US activity has the most Shapley value in the 4th mode, Fig. 6(d). The highest values are obtained during noon, up to more than 30%, and big values from 11:00 to 4:00. It matches with the hours in which the hF2 influence is the lowest. Interestingly, the hE plays an important role of more than 15% from 8:00 to 19:00. The

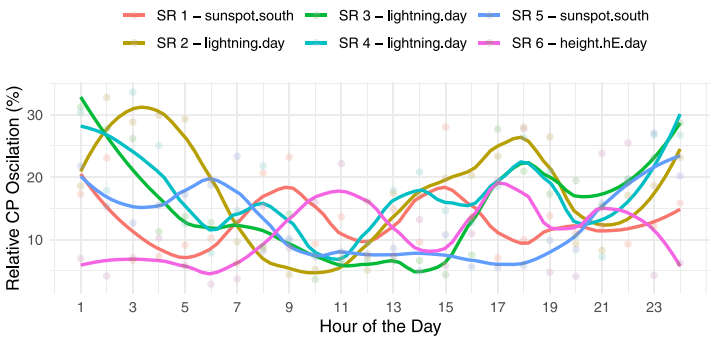
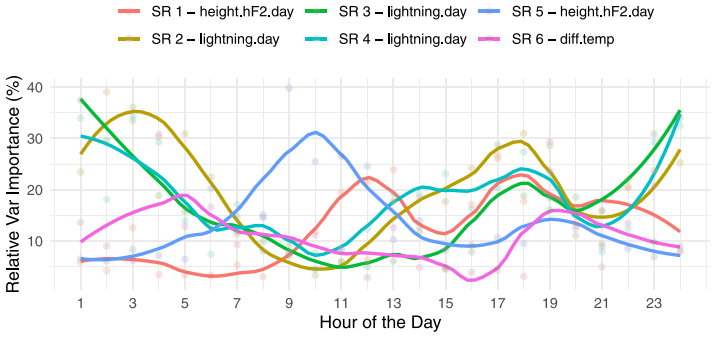
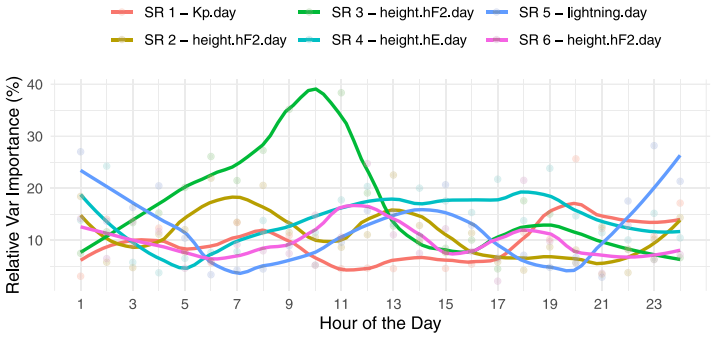


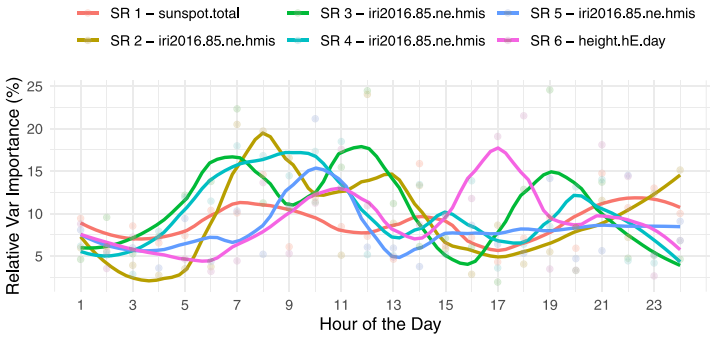
Fig. 7. Most important normalized CP **Oscillation** variable for each of the six SR modes for the 24 h.



(a)



(b)



(c)

Fig. 8. The highest prediction regarding normalized **Variable importance** for each of the six SR modes for the 24 h. a: **1st** most important Variable, b: **2nd** most important Variable, c: **3rd** most important Variable,.

influence of the hE is only present in this mode and in the 6th mode, Fig. 6(f), however, with a completely different pattern. In this mode, the presence of the hEs is strongly remarkable. The highest Shapley values for this mode, 25%, are obtained by this ionospheric variable and also these values are concentrated in the narrow interval from 7:00 to 10:00, which can bear an interesting link. Given the above, it can

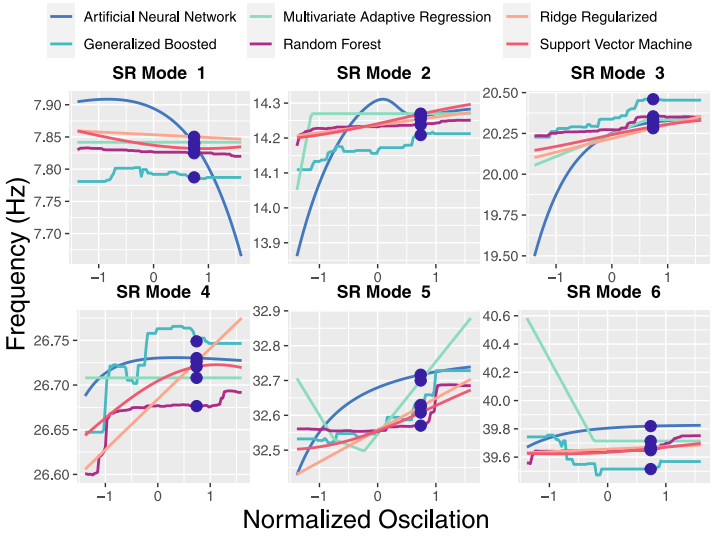


Fig. 9. CP Profile for the Height of the ionosphere (oscillation of hF2) at 8:00 for the six SR modes.

be concluded that the main influence is given by the hF2, but with a common pattern between odd and even modes.

The possibilities of narrowing the hours of interest have been exposed, and also the usage of contrasting two similar modes for explaining ionospheric variables.

4.3. Ceteris-Paribus Oscillation

The result of the CP Oscillation shows the impact on the predicted output when an oscillation is performed on a single predictor while the rest of the independent variables remain constant. The CP Oscillation vote is slightly different from the one shown in Fig. 4. The most significant difference is located in the 6th SR mode, where there is always around one relevant predictor in the interval [0:00–12:00 UTC]. The difference from the previous metrics is straightforward, because the values of the predictor are allowed to oscillate outside the original data limit.

The most important variable of each SR mode concerning the CP Oscillation can be seen in Fig. 7. Contrary to the last two metrics, the most common highest CP Oscillation value is not the hF2 but the Lighting-US, present in the 2nd, 3rd and 4th mode. It is also important to note that this variable presents the highest values, more than 20%. It can be seen that these three modes present a slightly similar pattern, very high from 21:00 to 4:00 and also a peak between 13:00 and 18:00. The values are considerably higher for the 2nd mode. The highest value for 1st and 4th modes is SunspotS. There are clear differences between the two. For the 1st mode, it shows an oscillating behavior with 3 peaks centered at 0:00, 8:00 and 14:00. On the other hand, the 5th mode presents a wide peak in the interval from 21:00 to 7:00 with a value of 15% and a wide trough for the rest of the hours. Opposite to the rest of the modes, the 6th mode exposed the highest dependency in terms of CP oscillation with the hE variable. The behavior of this variable shows three strong peaks at 10:00, 16:00 and 20:00 and an almost 0 value from 0:00 to 5:00. This metric shows relevance when not only the values in the dataset are used. It can be seen that the behavior is sometimes consistent with the previous variable, especially in the 2nd mode. Lightning-US results provide a preliminary indication to focus on the 2nd, 3rd and 4th modes for studying the lightning activity at 1:00 and 17:00.

4.4. Variable importance

Variable importance shows the increase of the residual error when a given predictor is excluded from the model training.

To focus on the most relevant features, the three most important values, considering the 24 h, for each mode can be seen in Fig. 8.

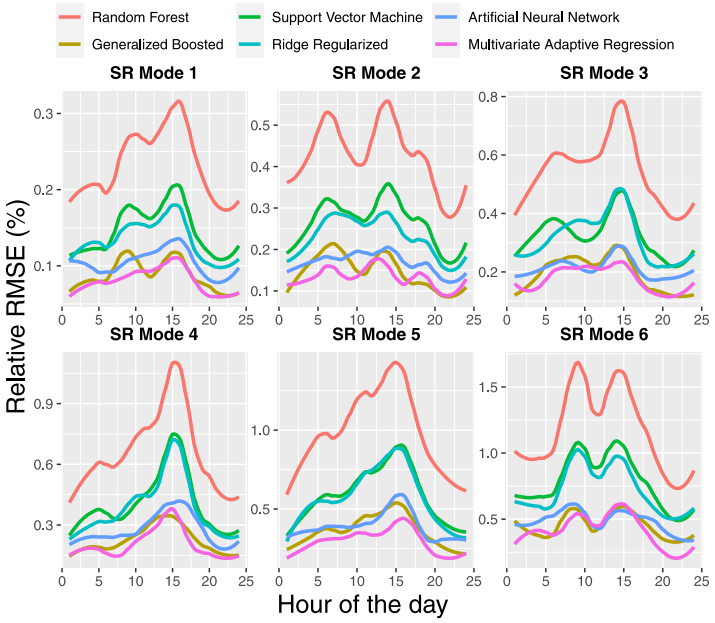
The importance of the hF2 variable can be noticed in the fact that five of the six modes have it among the three most important variables. It is also possible to observe that hF2 has the highest importance for the odd modes. Particularly interesting is that the 3rd and 5th modes have the peaks value at the same hour 10:00 with almost 30% and 40% respectively. Although hF2 is among the most important variables for the 2nd and 6th, considering 24 h average, neither of the two modes shows a substantial relevance at any hour, being less than 17%.

Lightning-US is present in 4 SR modes and the most important for three modes, 2nd, 3rd and 4th. Surprisingly, the pattern is almost identical for these three modes. It shows a strong importance during the interval from 22:00 to 6:00, with a prominent peak at 0:00, and then a considerable peak around 18:00. It is also important to remark that the values are appreciably higher for the 2nd and 4th modes. It is interesting that hD is the third most important ionospheric variable for four of the six modes. Although the lines are significantly different, all share a high value between 6:00 and 13:00 and a minimum low value at 1:00. It is also important to remark that the 1st mode is the only one which has Kp and Sunspot as an important variable. The 6th mode also behaves very differently from the rest. As the most important variable it has the Temperature, with a substantial high value at 5:00 and 19:00, around 20% and hE as the third most important.

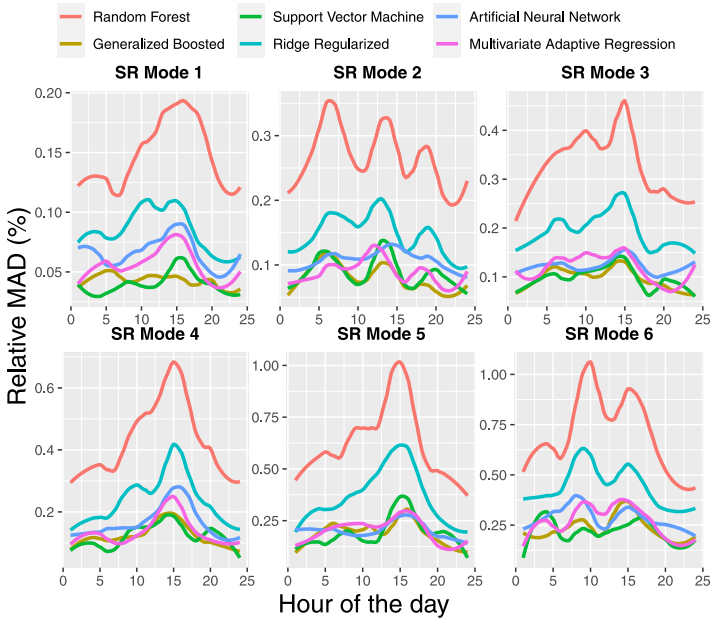
In conclusion, this result provides additional insight about the relation between the odd modes, with substantial similarities among them. It is also possible to notice the particular behavior of the first mode, with just one variable in common with the rest, i.e. hF2, while the others are related to the Solar influence (Kp and Sunspot Total).

4.5. Ceteris-Paribus Profile

The CP profile explains how the ML model changes when a particular predictor is slightly modified while the rest remain constant. To show that every ML behaves in a similar trend, it has been chosen to study the hF2 predictor at 8:00 in Fig. 9. It can be seen that although each ML model shows a somewhat different pattern, the slope and direction of each mode is very similar. This metric exposes that each ML can learn different types of relationships between the dependent variable and the predictors. However, the trends are very similar no matter what ML is chosen. The pattern outlined in the mentioned figure is present almost every hour and shared among all the predictors.



(a)



(b)

Fig. 10. Goodness-of-fit evolution for the six SR modes and the six ML algorithms under study along 24 h. a: **RMSE**, b: **MAD**.

4.6. Performance metrics

To show the relative importance of the error, these metrics are divided by the mean value for each mode and shown as a percentage. In order to make a comparison between modes possible, a relative value is performed. The performance metric values are divided by the mean value SR frequency of each mode and multiplied by 100. The RMSE for the first six SR modes can be seen in Fig. 10(a) for the six ML methods. It is possible to see that the behavior is broadly similar among all the ML methods and SR modes, with a considerable higher value for all at 15:00 and a common minimum at 20:00. There are also several points with higher values that are common for the ML algorithms. It is important to remark that the GBM model presents the lowest RMSE for all the SR modes. Concerning the MAD metrics in Fig. 10(b), the result is in line with the previously mentioned. The best performance is reached with the GBM model for all SR modes. The hourly evolution of this metric is not as consistent among methods as for the RMSE, with more differences between SR modes and hours. The y-axis shows that the error is relatively similar for all the modes concerning the mean value of the SR mode, with a maximum value of 0.7% RMSE and 0.4% MAD.

4.7. Residual analysis

In Fig. 11, the original outcome can be seen against the predicted value. We have chosen two cases:

- **17:00 - Worst case**
- **00:00 - Best case**

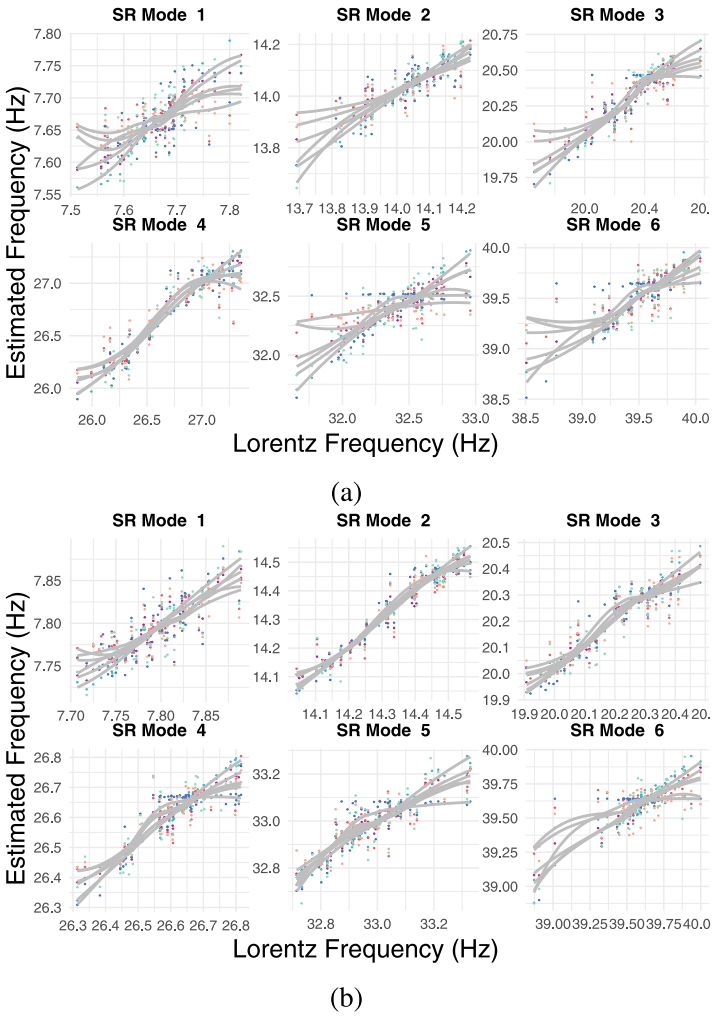


Fig. 11. Comparison between the prediction and the original outcomes for all the ML algorithms and the six modes. x-axis: Lorentzian values, y-axis: estimated value. Gray line: Smooth line for each ML algorithm. a: At **15:00**, b: At **00:00**.

It is possible to discern that the values around the mean frequency are consistent between the real value for each mode and those predicted by the ML algorithms. However, when it comes to the values of the outliers, a higher difference is obtained. In the worst case Fig. 11(a), it can be seen that the 1st SR mode shows a high error for the values outside the mean. Surprisingly, the 4th SR mode exhibits a very similar pattern to the $x = y$ line.

On the other hand, the second plot also points to the prediction at 00:00, in order to expose a good prediction case (Fig. 11(b)). In this scenario, it is possible to see a high output correlation in every mode and for every ML algorithm. It is observable that the values for the 1st mode are more dispersed than for any other mode, so the prediction is a very complex task even in the best case. Nevertheless, the result trends shown in this section can be considered as evidence that the ionosphere parameters can be taken into account to predict the SR frequency variations.

4.8. Case study: Frequency vs. Intensity

In this subsection, we want to bring confirmation that the heuristic-hypothesis that it is possible to use a ML approach for modeling the monthly average SR mode frequency variation for each of the 24 h using the monthly average of the ionospheric variables values. We have the assumption that there is a significant link between the ionosphere variables and SR mode frequency variation. But, we have also presumed that this link is not relevant for the SR mode intensity variation. With the aim of assessing our results we have applied the same methodology to the SR mode intensity variation. As it was exposed in Section 1, the condition of the ionosphere modifies the conductivity of the cavity, which leads to a change in the propagation properties of the ionosphere, while the intensity is much more sensitive to the local condition around the ELF observatory. We have applied the same methodology to the intensity values of the first six SR modes. The data treatment has been equal to the one for the frequency values. The intensity values correspond to magnetic field spectral density of the SR peaks, originally in $10^5 \times pT/\sqrt{\text{Hz}}$. In Fig. 12(a) the RMSE of the intensity variable can be seen. Following the frequency approach, the RMSE values are divided by the mean SR intensity of each mode and multiplied by 100. It is clear that These values are around 10% and 20%. There are firmly higher than those reached for the frequency which is around 40 more times. A special mention has to be made to the fact that although the predicted intensity values are in the same range as the real ones, there is not a good enough predictor using the ionospheric variables. A similar plot to the one shown for the Frequency, but using Intensity values, can be seen in Fig. 12(b). It is noticeable that the values are only similar when the model predicts a value around the mean, which is not significantly better than using the mean value as the predictor. As we have presumed, the ML methods are not able to capture the relationship between the ionospheric variables and SR mode intensity variation. This contrast test provides additional insight to the previously mentioned hypothesis that the result of our methodology applied to SR frequency variation is valid. However, as expected, there is no substantial evidence to exploit the same for intensity prediction.

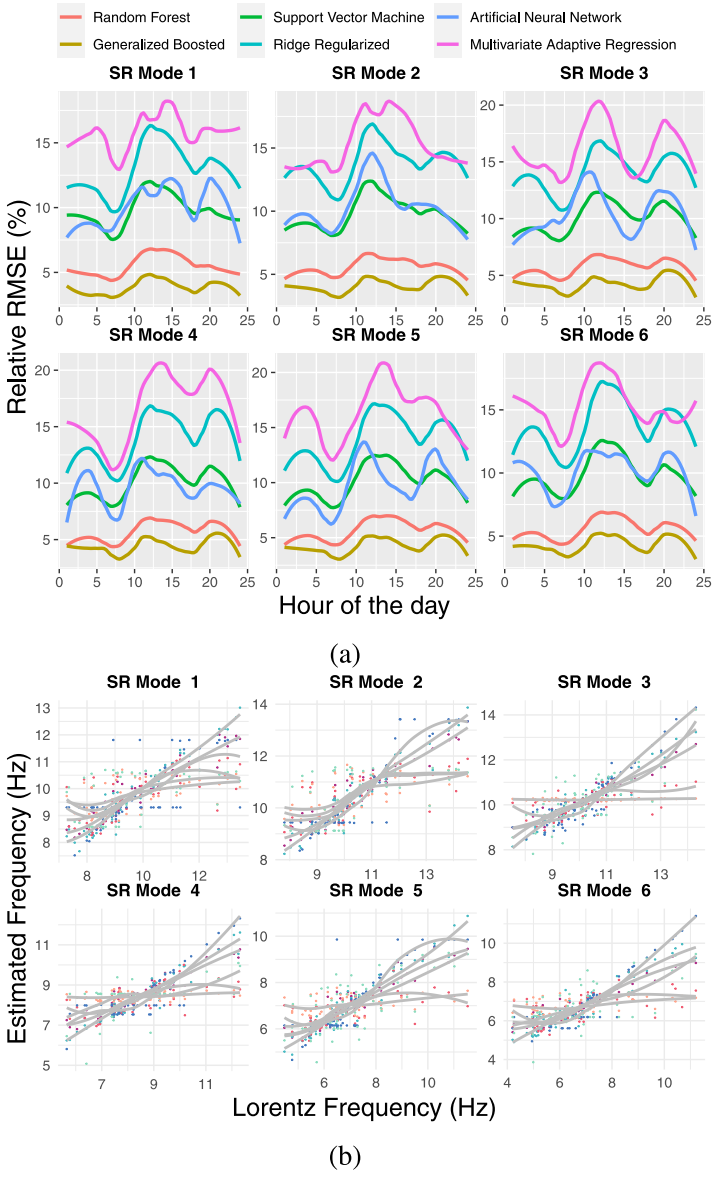


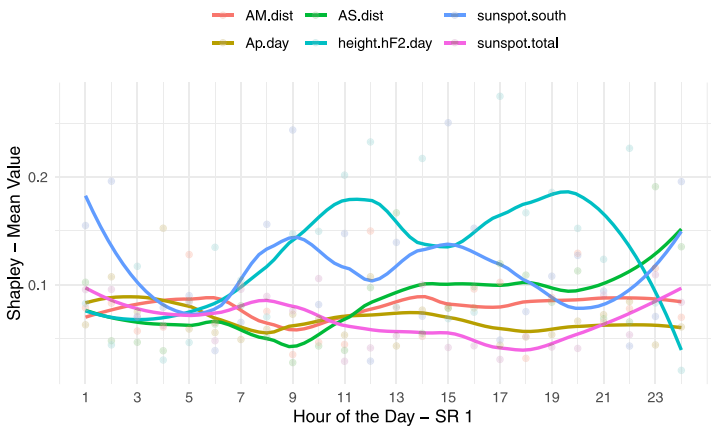
Fig. 12. Result for the Intensity SR variation along 24 h with the ML algorithms for the six first SR modes. a: **Performance RMSE** x-axis: Hour of the day. y-axis: Relative RMSE to mean SR frequency, b: **Residual Plot** Comparison between the prediction and the original outcomes. x-axis: Lorentzian values y-axis: Estimated value. Gray line: Smooth line for each ML algorithm.

4.9. Case study: Distance to the three thunderstorm centers

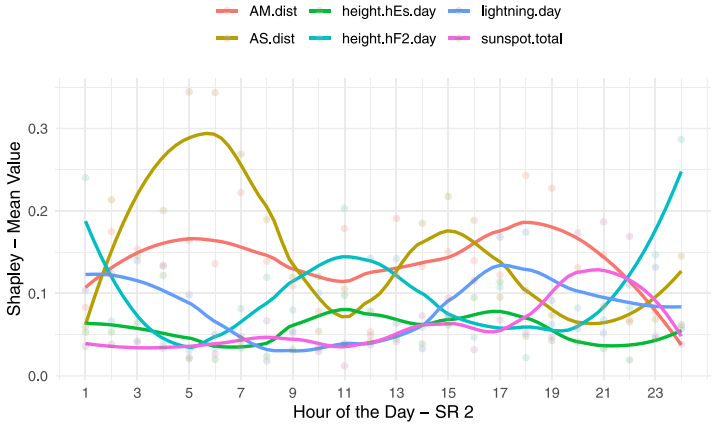
It is clear that the most important variable is the distance between the lighting activity and the ELF observatory. The lightning activity is heavily concentrated in three points around the world. Due to the fact that each one is located on a different continent, it has been established in the literature the name of the continent for each one, i.e. the Asian, African and American thunderstorm center. To evaluate our ML model, we have added these three distances, using the analytical solution [37], repeating the same values for every year. This validation method has two main outcomes:

- The ML model recognizes the importance of the Source–Observer distance for the majority of the SR modes and hours.
- The result without adding the three distance variables has maintain the accuracy because their presence does not alter the relative values of the rest of variables.

We have chosen the Shapley metric because it is the most representative for the importance of the variables. In Table 2 the most important variables for each set can be seen. In the right column there is the result for the set with the distance, and it is clear that the American and Asian thunderstorm centers have an enormous importance for all the six SR modes. This result is in line with the theoretical analysis. The sensor we have used for this study is North–South oriented and these two thunderstorm centers are parallel to the sensor. In the same way, although the African thunderstorm is very close to our observatory and also is the most intensive one, the sensor is not sensitive enough due to the orthogonal relative orientation between this center and our observatory. The other variables, without taking into account the distance ones, are presented among the most important almost identically in both cases, with the exception of the 2nd and 3rd SR modes. In this two cases, the results show that the height of the D layer is not among the most important for the variable set with distance variables and the sunspot of the south hemisphere is not considered in the 3rd mode, when the distance variables are not included. In order to compare the differences between the ML model using the original set of variables and the one with the distance included, two figures have been added (Figs. 13(a) and 13(b)). Comparing the tendency of these two modes



(a)



(b)

Fig. 13. The normalized **Shapley** result for the first two SR modes for the 24 h adding the Source–Observer distance to the ML model. a: SR mode 1, b: SR mode 2.

with that shown in the Shapley part (Figs. 6(a) and 6(b)) it is clear that the model does not change its behavior on the whole, when distance variables are added. However, a few differences can be observed related to the second mode and in relationship with the hF2 variables. These small differences can be expected since the system has more variables to split the importance. It is also important to note that some collinearity can be anticipated between the distance variability and other variables that are affected by changes in the ionosphere

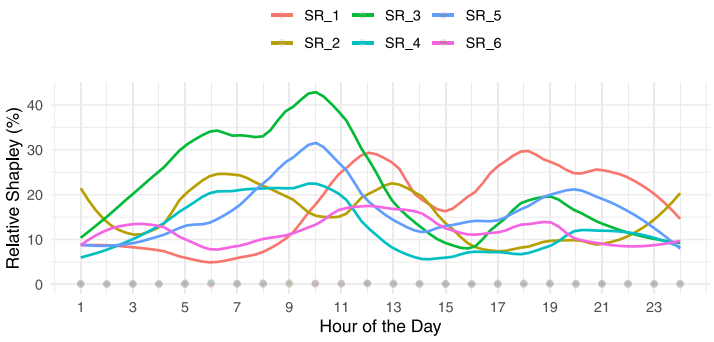
In Fig. 14, the detailed tendency of the hF2 can be seen for both cases: top row shows the ML model without adding the three distances between our ELF observatory and the thunderstorm centers, while bottom row includes these three variables. It can be seen that the pattern is very similar between these two, with the exception of the 2nd SR mode. In this SR mode, the difference is very clear in the first hours of the day, which is an effect very interesting to further explore. Unfortunately, studying this effect is out of the scope of present research.

To sum up, the outcomes of this validation are very satisfactory, the ML model is able to recognize the importance of the distance variables and also it does not change the tendency of the other variables. It is also important to remark that this relation is the best-documented one, and this strong correlation does not provide new insight about the SR variability.

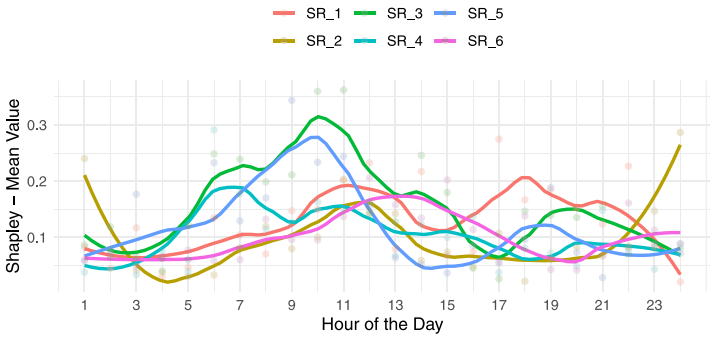
4.10. Discussion

As hypothesized, our experiments prove that the dependence between the ionospheric variables and the SR frequency variation exists, and this dependence is highly determined by the hour of the day in which the relationship is studied.

We are currently exploring the option of adding longer time series from other observatories. These could be used to test the methodology as well as also considering other different parameters. This addition could lead to a more accurate system, for example, a full solar cycle (i.e. 11–12 years) of data might show that the correlation with sunspots is not as strong, though it might reveal the opposite. However, the use of a longer period of time to measure the SR is out of scope of this research, mainly because our observatory has been recording continuously from 2016, so no data from before can be taken into account. Nonetheless, we are actively looking for collaborating with other observatories and gather together different SR data to develop a more reliable ML model, which could be used even for detecting changes in the ionosphere, or in the lightning activity based on SR measurements. We have used a variety of methods to evaluate the order and importance of the ionospheric variable to the ML interpretation, with all showing a common conclusion: we can observe that the SR modes behave completely different in their relation with the ionospheric variables among modes but also with respect to the hour of the day. As an example, the 1st SR mode is highly dependent on the hF2 for the 10:00 h, while at 14:00, the dependence is much more evident with the number of sunspots in the Southern hemisphere based on the Shapley metric. These studies provide further insight into the hypothesis proposed by our team in previous studies, Soler-Ortiz et al. [15],Cano-Domingo et al. [34].



(a)



(b)

Fig. 14. The normalized **Shapley** result for hF2 variable for the 24 h for the first six SR modes. a: Using 14 external variables, not adding Source–Observer distances, b: Using 14 external variables, not adding Source–Observer distances.

It is also noticeable that we have checked the adequacy of the number of data points through a comparison using the same methodology over the SR intensity value, with a radical difference among the accuracy for the dependent variable: the frequency dependence clearly superior, with a magnitude of around 40 times.

5. Conclusion

We studied the capability of 14 ionospheric variables to condition the SR frequency variation for each mode. In summary, the paper:

- Exposed the importance of considering the ionospheric variables to determine the SR frequency value.
- Provided additional insights about the dependence of the SR frequency variation on the hour of the day.
- Examined the prediction accuracy of the framework by considering the SR frequency versus the intensity variation in comparison as the dependent variable.
- Showed the capacity of the ionospheric variables to estimate the mean value of SR frequency value.
- Compared and combined 6 different methods of traditional ML to model the behavior of the SR frequency variation to adapt to different types of relationship.
- Contributed to further the knowledge of the common pattern among the odd modes and the same for the even modes in terms of SR frequency variation.
- Our result suggests that the importance of some ionosphere variables is significantly higher over others, such as the height of the F2 layer or the lightning activity of the USA.
- The validation methods reveal the importance of the Source–Observer distance as a predictor of the SR frequency variation for all the six first SR modes, and points a new line of research based on the relative position of the observatory for different hours.

As future steps, the use of data from different observatories could provide additional insight in order to support the dependencies presented in this paper. Another line of study is to further research this variation when a smaller average time is utilized.

CRediT authorship contribution statement

Carlos Cano-Domingo: Conceptualization of this study, Methodology, Software, Data preparation, Writing. **Ruxandra Stoean:** Co-ordination, Software, Data preparation, Writing, Review. **Gonzalo Joya:** Methodology, Editing, Review. **Nuria Novas:** Co-ordination, Editing, Review. **Manuel Fernandez-Ros:** Hardware development, Review. **Jose Antonio Gazquez:** Hardware development, Review.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jose Antonio Gazquez Parra, Noria Novas, Carlos Cano and Manuel Ros reports financial support was provided by The Ministry of Economics and Competitiveness of Spain. Ruxandra Stoean reports financial support was provided by Executive Agency for Higher Education, Research, Development and Innovation Funding Romania. Carlos Cano Domingo reports travel was provided by Erasmus. Jose Antonio Gazquez Parra, Noria Novas, Carlos Cano and Manuel Ros reports was provided by European Union.

Data availability

The authors do not have permission to share data.

References

[1] W.O. Schumann, "Uber die strahlungslosen Eigenschwingungen einer leitenden Kugel, die von einer Luftschicht und einer Ionosphärenhülle umgeben ist, Z. Naturforsch. - Sect. A J. Phys. Sci. (ISSN: 18657109) 7 (2) (1952) 149–154, http://dx.doi.org/10.1515/zna-1952-0202.

[2] Colin Price, ELF electromagnetic waves from lightning: The schumann resonances, Atmosphere (ISSN: 20734433) 7 (9) (2016) http://dx.doi.org/10.3390/atmos7090116.

[3] Bodo W. Reinisch, Ivan A. Galkin, Global ionospheric radio observatory (GIRO), Earth, Plan. Space (ISSN: 18805981) 63 (4) (2011) 377–381, http://dx.doi.org/10.5047/eps.2011.03.001.

[4] J. Matzka, C. Stolle, Y. Yamazaki, O. Bronkalla, A. Morschhauser, The geomagnetic kp index and derived indices of geomagnetic activity, Space Weather (ISSN: 15427390) 19 (5) (2021) e2020SW002641, http://dx.doi.org/10.1029/2020SW002641.

[5] NASA GISS, 2019: GISS surface temperature analysis (GISTEMP), version 4, 2019, URL https://data.giss.nasa.gov/gistemp/.

[6] Nathan J.L. Lenssen, Gavin A. Schmidt, James E. Hansen, Matthew J. Menne, Avraham Persin, Reto Ruedy, Daniel Zyss, Improvements in the GISTEMP uncertainty model, J. Geophys. Res.: Atmos. (ISSN: 21698996) 124 (12) (2019) 6307–6326, http://dx.doi.org/10.1029/2018JD029522.

[7] K.F. Tapping, The 10.7 cm solar radio flux (f10.7), Space Weather (ISSN: 15427390) 11 (7) (2013) 394–406, http://dx.doi.org/10.1002/swe.20064.

[8] Daniel J. Cecil, Dennis E. Buechler, Richard J. Blakeslee, Gridded lightning climatology from TRMM-LIS and OTD: Dataset description, Atmos. Res. (ISSN: 01698095) 135 (2014) 404–414, http://dx.doi.org/10.1016/J.ATMOSRES.2012.06.028.

[9] Dieter Bilitza, IRI the international standard for the ionosphere, Adv. Radio Sci. (ISSN: 16849973) 16 (2018) 1–11, http://dx.doi.org/10.5194/ars-16-1-2018.

[10] Frédéric Clette, Laure Lefèvre, The new sunspot number: Assembling all corrections, Sol. Phys. (ISSN: 1573093X) 291 (9–10) (2016) 2629–2651, http://dx.doi.org/10.1007/s11207-016-1014-y.

[11] G. Satori, Monitoring Schumann resonances–ll . Daily and seasonal frequency, Science 58 (13) (1996) 1483–1488.

[12] Gabriella Sátori, Bertalan Zieger, El Niño related meridional oscillation of global lightning activity, Geophys. Res. Lett. (ISSN: 00948276) 26 (1999) 1365–1368, http://dx.doi.org/10.1029/1999GL900264.

[13] G. Tatsis, V. Christofilakis, S.K. Chronopoulos, G. Baldoumas, A. Sakkas, A.K. Paschalidou, P. Kassomenos, I. Petrou, P. Kostarakis, C. Repapis, V. Tritakis, Study of the variations in the Schumann resonances parameters measured in a southern Mediterranean environment, Sci. Total Environ. (ISSN: 18791026) 715 (2020) 136926, http://dx.doi.org/10.1016/j.scitotenv.2020.136926.

[14] Vasilis Tritakis, Ioannis Contopoulos, Janusz Mlynarczyk, Vasilis Christofilakis, Giorgos Tatsis, Christos Repapis, How effective and prerequisite are electromagnetic extremely low frequency (ELF) recordings in the schumann resonances band to function as seismic activity precursors, Atmosphere (ISSN: 20734433) 13 (2) (2022) http://dx.doi.org/10.3390/atmos13020185.

[15] Manuel Soler-Ortiz, Manuel Fernández-Ros, Nuria Novas Castellano, Jose A. Gazquez Parra, A new way of analyzing the schumann resonances: A statistical approach, IEEE Trans. Instrum. Meas. (ISSN: 15579662) 70 (2021) http://dx.doi.org/10.1109/TIM.2021.3073435.

[16] D.D. Sentman, Approximate schumann resonance parameters for a two-scale-height ionosphere, J. Atmos. Terr. Phys. (ISSN: 00219169) 52 (1) (1990) 35–46, http://dx.doi.org/10.1016/0021-9169(90)90113-2.

[17] Yu P. Galuk, A.P. Nickolaenko, M. Hayakawa, Knee model: Comparison between heuristic and rigorous solutions for the Schumann resonance problem, J. Atmos. Sol.-Terr. Phys. (ISSN: 13646826) 135 (2015) 85–91, http://dx.doi.org/10.1016/j.jastp.2015.10.008.

[18] Erno Prácser, Tamas Bozóki, Gabriella Sátori, Janos Takatsy, Earle Williams, Anirban Guha, Two approaches for modeling ELF wave propagation in the earth-ionosphere cavity with day-night asymmetry, IEEE Trans. Antennas and Propagation (ISSN: 15582221) 69 (7) (2020) 4093–4099, http://dx.doi.org/10.1109/TAP.2020.3044669.

[19] Tamás Bozóki, Ern Prácser, Gabriella Sátori, Gergely Dálya, Kornél Kapás, János Takátsy, Modeling Schumann resonances with schupy, J. Atmos. Sol.-Terr. Phys. (ISSN: 13646826) 196 (2019) 105144, http://dx.doi.org/10.1016/j.jastp.2019.105144.

[20] Christian Kwisanga, Coenrad J. Fourie, 3-D modeling of electromagnetic wave propagation in the uniform earth-ionosphere cavity using a commercial FDTD software package, IEEE Trans. Antennas and Propagation (ISSN: 0018926X) 65 (6) (2017) 3275–3278, http://dx.doi.org/10.1109/TAP.2017.2695532.

[21] Jamesina J. Simpson, Allen Taflove, Three-dimensional FDTD modeling of impulsive ELF propagation about the earth-sphere, IEEE Trans. Antennas and Propagation (ISSN: 0018926X) 52 (2) (2004) 443–451, http://dx.doi.org/10.1109/TAP.2004.823953.

[22] Andrea Pizzuti, Alec Bennett, Martin Füllekrug, Long-term observations of schumann resonances at portishead (UK), Atmosphere (ISSN: 20734433) 13 (1) (2022) 38, http://dx.doi.org/10.3390/atmos13010038.

[23] A.V. Koloskov, A.P. Nickolaenko, Yu M. Yampolsky, Chris Hall, O.V. Budanov, Variations of global thunderstorm activity derived from the long-term Schumann resonance monitoring in the Antarctic and in the Arctic, J. Atmos. Sol.-Terr. Phys. (ISSN: 13646826) 201 (February) (2020) 105231, http://dx.doi.org/10.1016/j.jastp.2020.105231.

[24] Pablo Sierra Figueredo, Blanca Mendoza Ortega, Marni Pazos, Daniel Rodríguez Osorio, Ernesto Andrade Mascote, Víctor M. Mendoza, René Garduño, Schumann Resonance anomalies possibly associated with large earthquakes in Mexico, Indian J. Phys. (ISSN: 09749845) 95 (10) (2021) 1959–1966, http://dx.doi.org/10.1007/s12648-020-01865-6.

[25] M. Hayakawa, J. Izutsu, A. Yu Schekotov, A.P. Nickolaenko, Yu P. Galuk, I.G. Kudintseva, Anomalies of Schumann resonances as observed near Nagoya associated with two huge (M-7) Tohoku offshore earthquakes in 2021, J. Atmos. Sol.-Terr. Phys. (ISSN: 13646826) 225 (August) (2021) 105761, http://dx.doi.org/10.1016/j.jastp.2021.105761.

[26] K. Florios, I. Contopoulos, G. Tatsis, V. Christofilakis, S. Chronopoulos, C. Repapis, Vasilis Tritakis, Possible earthquake forecasting in a narrow space-time-magnitude window, Earth Sci. Inform. (ISSN: 18650481) 14 (1) (2021) 349–364, http://dx.doi.org/10.1007/s12145-020-00535-9.

[27] G. Tatsis, A. Sakkas, V. Christofilakis, G. Baldoumas, S.K. Chronopoulos, A.K. Paschalidou, P. Kassomenos, I. Petrou, P. Kostarakis, C. Repapis, V. Tritakis, Correlation of local lightning activity with extra low frequency detector for Schumann Resonance measurements, Sci. Total Environ. (ISSN: 18791026) 787 (2021) 147671, http://dx.doi.org/10.1016/j.scitotenv.2021.147671.

[28] Tamás Bozóki, Gabriella Sátori, Earle Williams, Irina Mironova, Péter Steinbach, Emma C. Bland, Alexander Koloskov, Yuri M. Yampolski, Oleg V. Budanov, Mariusz Neska, Ashwini K. Sinha, Rahul Rawat, Mitsuteru Sato, Ciaran D. Beggan, Sergio Toledo-Redondo, Yakun Liu, Robert Boldi, Solar cycle-modulated deformation of the earth–Ionosphere cavity, Front. Earth Sci. (ISSN: 22966463) 9 (August) (2021) 689127, http://dx.doi.org/10.3389/feart.2021.689127.

[29] Inga Timofejeva, Rollin McCraty, Mike Atkinson, Abdullah A. Alabdulgader, Alfonsas Vainoras, Mantas Landauskas, Vaiva Šiaučinait, Minvydas Ragulskis, Global study of human heart rhythm synchronization with the earth's time varying magnetic field, Appl. Sci. (Switzerland) (ISSN: 20763417) 11 (7) (2021) 2935, http://dx.doi.org/10.3390/app11072935.

[30] Carlos Cano Domingo, Nuria Novas Castellano, Manuel Fernandez Ros, Jose Antonio Gazquez-Parra, Segmentation and characteristic extraction for Schumann Resonance transient events, Measurement (ISSN: 0263-2241) 194 (2022) 110957, http://dx.doi.org/10.1016/j.measurement.2022.110957, URL https://www.sciencedirect.com/science/article/pii/S0263224122002329.

[31] Carlos Cano Domingo, Nuria Novas Castellano, Ruxandra Stoean, Manuel Fernandez-Ros, Jose A. Gazquez Parra, Schumann resonance modes and ionosphere parameters: An annual variability comparison, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, http://dx.doi.org/10.1109/TIM.2022.3194912.

[32] J.A.G. Parra, M.F. Ros, N.N. Castellano, R.M.G. Salvador, Techniques for schumann resonance measurements: A comparison of four amplifiers with a noise floor estimate, IEEE Trans. Instrum. Meas. (ISSN: 15579662) 64 (10) (2015) 2759–2768, http://dx.doi.org/10.1109/TIM.2015.2420376.

[33] J.A. Gazquez, R.M. Garcia, N.N. Castellano, M. Fernandez-Ros, A.J. Perea-Moreno, F. Manzano-Agugliaro, Applied engineering using Schumann Resonance for earthquakes monitoring, Appl. Sci. (Switzerland) (ISSN: 20763417) 7 (11) (2017) 1113, http://dx.doi.org/10.3390/app7111113.

[34] C. Cano-Domingo, M. Fernandez-Ros, N. Novas, J.A. Gazquez, Diurnal and seasonal results of the Schumann Resonance Observatory in Sierra de Filabres, Spain, IEEE Trans. Antennas and Propagation (ISSN: 15582221) 69 (10) (2021) 6680–6690, http://dx.doi.org/10.1109/TAP.2021.3069537.

[35] Ruxandra Stoean, Catalin Stoean, Adrian Sandita, Daniela Ciobanu, Cristian Mesina, Ensemble of classifiers for length of stay prediction in colorectal cancer, in: Ignacio Rojas, Gonzalo Joya, Andreu Catala (Eds.), Advances in Computational Intelligence, Springer International Publishing, Cham, ISBN: 978-3-319-19258-1, 2015, pp. 444–457.

[36] Mohamed Mohandes, Mohamed Deriche, Salihu Aliyu, Classifiers combination techniques: A comprehensive review, IEEE Access 6 (2018) 19626–19639, http://dx.doi.org/10.1109/ACCESS.2018.2813079.

[37] A.P. Nickolaenko, G. Sátori, B. Zieger, L.M. Rabinowicz, I.G. Kudintseva, Parameters of global thunderstorm activity deduced from the long-term Schumann resonance records, J. Atmos. Sol.-Terr. Phys. (ISSN: 13646826) 60 (1998) 387–399, http://dx.doi.org/10.1016/S1364-6826(97)00121-1.

[38] Jerome Friedman, Trevor Hastie, Robert Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22, URL https://www.jstatsoft.org/v33/i01/.

[39] Jerome H. Friedman, Multivariate adaptive regression splines, Ann. Statist. 19 (1) (1991) 1–67, http://dx.doi.org/10.1214/aos/1176347963.

[40] Brett Lantz, Machine Learning with R: Expert Techniques for Predictive Modeling, Packt publishing ltd, 2019.
[41] Alexandros Karatzoglou, Alexandros Smola, Kurt Hornik, Achim Zeileis, Kernlab - an S4 package for kernel methods in R, J. Stat. Softw. 11 (9) (2004) 1–20, http://dx.doi.org/10.18637/jss.v011.i09, URL https://www.jstatsoft.org/index.php/jss/article/view/v011i09.
[42] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[43] Greg Ridgeway, Generalized Boosted Models: A guide to the gbm package, Update 1 (1) (2007) 2007.
[44] Przemyslaw Biecek, DALEX: Explainers for complex predictive models in R, J. Mach. Learn. Res. 19 (1) (2018) 3245–3249, URL https://jmlr.org/papers/v19/18-416.html.