

Automatic detection of the thermal electron density from the WHISPER experiment onboard CLUSTER-II mission with neural networks

Nicolas Gilet, E. de Leon, R. Gallé, X. Vallières, Jean-Louis Rauch, Klet Jegou, Luca Bucciantini, V. Savreux, Pierrette Décréau, P. Henri

▶ **To cite this version:**

Nicolas Gilet, E. de Leon, R. Gallé, X. Vallières, Jean-Louis Rauch, et al.. Automatic detection of the thermal electron density from the WHISPER experiment onboard CLUSTER-II mission with neural networks. Journal of Geophysical Research Space Physics, 2021, 126 (3), pp.e2020JA028901. 10.1029/2020JA028901 . insu-03141096

HAL Id: insu-03141096
https://insu.hal.science/insu-03141096

Submitted on 15 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Check for updates

Automatic detection of the thermal electron density from the WHISPER experiment onboard CLUSTER-II mission with neural networks

N. Gilet [1], E. De Leon[1], R. Gallé[1], X. Vallières[1], J-L. Rauch[1], K. Jegou[1], L. Bucciantini[1], V. Savreux[1], P. Décréau[1] and P. Henri[1,2]

[1]Laboratoire de Physique et Chimie de l'Environnement et de l'Espace (LPC2E), CNRS, Université d'Orléans, Orléans, France

[2]Laboratoire Lagrange, OCA, UCA, CNRS, Nice, France

Key Points:

- We have applied self-learning methods to predict the key plasma regions crossed by the CLUSTER-II spacecraft in the Earth magnetosphere using the WHISPER instrument
- The extraction of the thermal electron density from WHISPER active (sounding mode) and natural (passive mode) electric field spectra is automatically done in the free solar wind, in the magnetosheath region and in the plasmasphere
- Such automatic procedure could be used for future data processing of electric field experiments onboard space missions (for instance AM^{2}P onboard BepiColombo or MIME onboard JUICE)

Corresponding author: Nicolas GILET, nicolas.gilet@cnrs-orleans.fr

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2020JA028901.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Abstract

The WHISPER (Waves of HIgh frequency and Sounder for Probing Electron density by Relaxation) instrument has been monitoring the bulk properties of the plasma environment around Earth for more than twenty years. Onboard the 3-D Earth magnetospheric CLUSTER-II mission, this experiment delivers active and natural electric field spectra, in a frequency interval ranging respectively from 3.5 to 82 kHz, and from 2 to 80 kHz. The thermal electron density, a key parameter of scientific interest and major driver for the calibration of particles instrument, is derived from spectra.

Until recently, the extraction of the thermal electron density required a manual intervention. To automate this process, self-learning algorithms based on Multilayer Neural Networks have been implemented. The evaluation of the thermal electron density from WHISPER spectra depends on the plasma region encountered by the spacecraft. First, a fully-connected neural network has been implemented to predict the plasma region, using only the active spectra measured by the WHISPER instrument. Secondly, a specific neural network has been implemented to predict the thermal electron density for each plasma region. The model reaches up to 98% prediction accuracy for some plasma regimes. Two thermal electron density prediction models were trained, a first one to process data from the free solar wind and magnetosheath regions, and a second one for the plasmasphere region. The prediction accuracy can reach up to 95% in the free solar wind and magnetosheath regimes, and 75% in the plasmasphere.

1 Introduction

The CLUSTER-II mission has been successfully monitoring the plasma bulk properties of the Earth magnetosphere for more than twenty years (Escoubet et al., 1997). Launched in July and August 2000, this ESA mission is the first to study the ionized medium in near-Earth space in three dimensions, thanks to a tetrahedral configuration of four identical spacecraft (Escoubet et al., 2015). Each spacecraft follows polar orbits around the Earth, with a ~57 h period. The elliptic orbits have been changed throughout the operations, from a closest perigee at 250 km to a farthest apogee at 120,000 km, which, together with the local time drift of the orbit plane throughout a year allows the spacecraft to cross the different plasma regions of the Earth magnetosphere. The payload is identical for each satellite, consisting of 11 experiments dedicated to the measurements of the surrounding plasma, including five instruments from the Wave Experiment Consortium (WEC, Lefeuvre et al. (1993)), among which the WHISPER (Waves of HIgh frequency and Sounder for Probing Electron density by Relaxation; (Décréau et al., 1997))

This article is protected by copyright. All rights reserved.

Accepted Article

instrument. The WHISPER instrument makes use of the two far double sphere electric antennas either in active (sounding) mode or in passive (natural) mode to characterize the plasma bulk properties such as the thermal electron density, the magnetic field amplitude as well as to provide useful information about the electron velocity distribution function (evdf) (Trotignon et al., 2001). In sounding mode, an active spectrum is built on-board by local excitation of the surrounding plasma at a frequency sweeping the 3.5-82 kHz range. In passive mode, the transmitter is switched off and the (electric) antennas are left at floating potential. The instrument becomes a natural wave receiver and spectrum analyser, delivering spectrograms over the 2-80 kHz frequency range (Décréau et al., 1997). The plasma bulk properties can be extracted, under certain conditions, from the active and natural electric field spectrograms measured by the WHISPER experiments. In particular, the thermal electron density can be deduced from the characteristics of natural waves in natural mode and from the plasma resonances triggered in active mode (Trotignon et al., 2003). The thermal electron density is one of the key datasets provided by the WHISPER experiment which is used for scientific interest (Canu et al., 2001; Darrouzet et al., 2008; El-Lemdani Mazouz et al., 2009; Kougblénou et al., 2011; Sandhu et al., 2016) and is the main driver of the calibration of the particles experiments (Johnstone et al., 1997; Trotignon et al., 2010). However, the extraction of the electron density requires a careful analysis of active and/or natural spectra and can be tricky in some plasma regions (Trotignon et al., 2001). Indeed, the location of plasma resonances in the electric field spectrograms, related to the plasma parameters, strongly depends on the plasma conditions such as the magnetic field or the evdf (Gilet et al., 2017). The presence and frequency position of plasma resonances are documented in theoretical works (Bernstein, 1958). They are also revealed by numerical modeling of the experimental responses (Chasseriaux et al., 1972; Wattieaux et al., 2019). However, a study of the resonances patterns of both active and natural spectra measured by the experiment is required (Trotignon et al., 1986, 2001). In the case of the WHISPER instrument, the thermal electron density is extracted via a semi-automatic pipeline, and sometimes fully manually, e.g. for cross-calibration purposes or detailed scientific analysis (Trotignon et al., 2010).

Fully-automatic algorithms, based on Machine Learning and Deep Learning methods, are developing quickly and some are already implemented in order to simplify data processing of space experiments. For instance, such algorithms have been implemented to detect space weather events (e.g. space weather forecasting, Camporeale et al. (2018)), to extract plasma parameters from space experiments (Zhelavskaya et al., 2016) or to predict the plasma regions

This article is protected by copyright. All rights reserved.

Accepted Article

in near-Earth regions (Nguyen et al., 2019; Breuillard et al., 2020). In order to automate the extraction of the thermal electron density from the electric field spectra measured by the WHISPER experiment, we propose to implement neural network methods (LeCun et al., 2015; Goodfellow et al., 2016). As explained above, the detection of the thermal electron density strongly depends on the plasma regime, mainly driven by the magnetic field strength (Trotignon et al., 2010). Therefore, the implemented pipeline proceeds in two steps. The first one is to determine the plasma region to separate (i) regions where the electron plasma frequency is the only resonance in the frequency bands (i.e. electron cyclotron frequency is neglected), typically corresponding to the solar wind and the magnetosheath, and (ii) regions where the electron plasma frequency is of the order of the electron cyclotron frequency, typically corresponding to the plasmasphere. To this purpose, we have implemented a fully-connected (dense) neural network classifying WHISPER spectra into three classes: free solar wind, magnetosheath and other plasma regions, including the plasmasphere, the tail and the cusp. The prediction accuracy can reach up to 98% for some plasma regimes. The second step is to automatically extract the thermal electron density from WHISPER electric field spectra using neural networks specifically implemented for each plasma region or group of plasma regions. In the free solar wind and the magnetosheath regions, we chose a recurrent (GRU) neural network architecture (Cho et al., 2014) to predict the electron plasma frequency from which the thermal electron density is derived. The prediction accuracy reaches up to 95% with a tolerance of one frequency bin (i.e. 162.8 Hz, the difference between two scanned frequencies). For the plasmasphere region, the upper hybrid frequency, which is more easily detected in active spectra than the plasma frequency, is predicted, then the thermal electron density is derived using the electron cyclotron frequency (obtained from the magnetic field measurement). The corresponding self-learning model is a fully-connected (dense) neural network. The prediction accuracy reaches up to 75% with a tolerance of one frequency bin.

A pipeline based on the self-learning algorithms detailed in this study is currently under development, to deliver the thermal electron density from the free solar wind and the magnetosheath region to the Cluster Science Archive (Laakso et al., 2010). The efficiency of the automatic detection of the thermal electron density shows that such automated methods could be used to extract the plasma bulk properties from future experiments such as mutual impedance experiments onboard BepiColombo (PWI/AM^{2}P) (Trotignon et al., 2006) and JUICE (RPWI/MIME) missions (Grasset et al., 2013).

This article is protected by copyright. All rights reserved.

Accepted Article

This paper is organized as follows. First, the WHISPER experiment and its corresponding dataset are described in Sect 2. Secondly, we give a high-level description of the neural networks in Sect. 3. Thirdly, we detail the methods used to predict the plasma region encountered by the spacecraft from WHISPER spectra and we explain the results given by the best neural network in Sect. 4. In Sect. 5, we describe the predictions of thermal electron density applied in different plasma regimes. In Sect 6, we discuss the applicability of such methods in the data processing of the WHISPER instrument and future electric field experiments such as PWI/AM^{2}P and the RPWI/MIME onboard respectively BepiColombo and JUICE missions. Finally, we conclude our study in Sect 7.

2 Instrumentation and data

In this section, we first summarize the main characteristics of the WHISPER instrument (Sect 2.1). Then, we introduce its key datasets mainly composed of active and natural electric field spectra and of the thermal electron density (Sect 2.2). Finally, we give an overview of the semi-automatic algorithms currently used to extract the thermal electron density from the WHISPER spectra (Sect 2.3).

2.1 WHISPER instrument

As part of the Wave Experiment Consortium (WEC, Lefeuvre et al. (1993)), the Waves of HIgh frequency and Sounder for Probing Electron density by Relaxation (WHISPER) instrument is a relaxation sounder that monitors the *in situ* space plasma bulk properties such as the thermal electron density (Décréau et al., 1997). Such instruments have been previously deployed in several space missions which operated in various ionized environments such as near Earth (GEOS-1 & 2, ISEE-1, Viking) or in the interplanetary medium (Ulysses) (Décréau et al., 1978; Harvey et al., 1979; Bahnsen et al., 1986; Stone et al., 1992). The WHISPER principle is based on electric coupling between pairs of electric dipole antennas embedded in the surrounding plasma. This specific experiment consists of a pair of receivers and transmitters associated with parts of two WEC instruments: the sensors of the Electric Field and Wave (EFW, Gustafsson et al. (1997)) experiment and the data processing capability of the Digital Wave Processing (DWP, Woolliscroft et al. (1997)) experiment. Figure 1 shows the WHISPER experiment configuration, which consists of four orthogonal wire booms carrying spherical sensors (8 cm in diameter) at the tips and deployed around the spacecraft. The electric antennas have sphere-to-sphere separations of 88 m. Further information on the WHISPER electron-

This article is protected by copyright. All rights reserved.

ics (operational features and performances) are described in detail in Décréau et al. (1997). Note that the WHISPER instrument is identical on each of the four CLUSTER-II spacecraft.

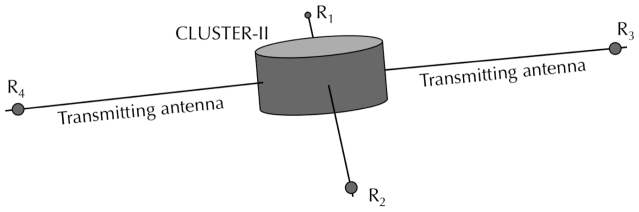


Figure 1. Illustration of the WHISPER relaxation sounder configuration, identical for the four CLUSTER-II spacecraft, using the EFW electric field antennas as transmitters and two of the four spherical booms R_i (8 cm in diameter) at the end of each antenna as receivers. The electric antennas have sphere-to-sphere separations of 88 m.

Two operational modes are used alternatively.

First, in sounding (active) mode, transmitters emit a sinusoidal signal from one antenna pair, at a given time t and frequency f, during a short time duration (~1 ms). Each train pulse covers a frequency band of 976.6 Hz centered on the given frequency f. Next, the signal is received on the EFW probes of a receiving pair (R_1R_2 or R_3R_4) at $t+\delta t$. Using onboard Fourier analysis of the instrument response around the transmitted frequency and by sweeping the transmitted frequency over the 3.5-82 kHz frequency range, a 162.8 Hz frequency resolution (512 bins) E-field spectrum is constructed on-board.

Secondly, in passive (natural) mode, transmitters are switched off. The WHISPER experiment becomes a simple electric receiver and perfoms a Fourier analysis to acquire a natural spectrum in the 2 to 80 kHz frequency range, with a frequency resolution of 162.8 Hz, corresponding to 512 bins (Décréau et al., 1997).

The sounding mode is operated alternately with the natural wave mode, with all four spacecraft following the same timeline. The WHISPER instrument is highly configurable and a number of operational parameters can be set by telecommand from ground. The resulting time resolution, depending on these parameters, varies from 0.3 to 3.4 s in natural mode and is generally of 1.5 s in active mode. The typical operational pattern includes 3 s of active mode, followed by 49 s of natural mode.

The relaxation sounder is based on the resonance principle of the propagation of plasma eigenmodes (Krall & Trivelpiece, 1973). Basically, the plasma characteristic frequencies of

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

the surrounding medium are triggered when these frequencies are in the emitted frequency range. When the transmitted pulse frequency is close to some plasma characteristic frequencies, very intense echoes are received. These stimulated signals are called plasma resonances. The WHISPER experiment operates close to the plasma characteristic frequencies which are typically found in the radio frequency (RF) range in Earth magnetospheric plasma. This effect allows the extraction of characteristic frequencies, which are directly related to the plasma bulk properties. In natural mode, only the natural electric emissions are monitored. Plasma bulk properties can thus be inferred from the analysis of active and natural spectra, among which:

The electron plasma frequency, noted f_{pe}, which is directly related to the thermal electron density as follows:

$$f_{pe} = \frac{1}{2\pi}\sqrt{\frac{n_e e^2}{\varepsilon_0 m_e}}, \tag{1}$$

where n_e is the thermal electron density, e is the electric charge, ε_0 is the vacuum permittivity and m_e the electron mass. The formula can be simplified as follows: $f_{pe}[kHz] \sim 9\sqrt{n_e[cm^{-3}]}$.

The electron cyclotron frequency and its harmonics, noted nf_{ce}, reads:

$$f_{ce} = \frac{1}{2\pi}\frac{eB_0}{m_e}, \tag{2}$$

where B_0 is the magnetic field amplitude. The formula can be simplified as follows: f_{ce} [kHz] $\sim 0.028\ B_0$ [nT]. Note that the electron cyclotron frequency can be extracted using the measurement of the magnetic field amplitude given by the magnetometer onboard the CLUSTER-II spacecraft (Balogh et al., 1997).

The upper hybrid frequency f_{uh}, related to f_{ce} and f_{pe}, and defined as follows:

$$f_{uh} = \sqrt{f_{pe}^2 + f_{ce}^2}, \tag{3}$$

Bernstein frequencies: in magnetized plasma, Bernstein frequencies, noted f_{qn}, can occasionally be measured by the WHISPER experiment. They satisfy the following conditions: $f_{qn} > 2f_{ce}$ and $nf_{ce} < f_{qn} < (n+1)f_{ce}$ (Bernstein, 1958). The frequency location of Bernstein modes has been tabulated in the Hamelin diagram for a plasma at thermal equilibrium (Hamelin, 1978; Trotignon et al., 2001). A shift between the tabulated frequency and the observed f_{qn} can reveal the presence of a non-maxwellian evdf, e.g. two-electron-temperature plasma (Belmont, 1981). Note that the lower hybrid frequency can be observed in specific cases (e.g. deep plasmasphere, Kougblénou et al. (2011)). This characteristic frequency is mainly driven by the motion coupling of ions and electrons which allows to constrain the plasma ion composition.

This article is protected by copyright. All rights reserved.

Accepted Article

2.2 WHISPER dataset

Both active/natural electric field spectra and the derived thermal electron density compose the key science datasets of the WHISPER experiment. The active and natural spectrograms show typical signatures encountered by the spacecraft in the different key plasma regions of the Earth magnetosphere. The characteristic signature of waves indicates the nature of the plasma region and, combined with the spacecraft position, reveals the different magnetospheric boundaries and regions. Figure 2 illustrates an example of several plasma region crossings by C3 (SAMBA) spacecraft from 18 Feb 2007 at 09:43 UT to 19 Feb 2007 at 09:53 UT associated with the frequency/time WHISPER active and natural spectra (second and third panels), the variation of the electron cyclotron frequency f_{ce} derived from the FGM measurement (fourth panel) and the variation of the electron plasma frequency f_{pe} extracted by *ad hoc* methods (fifth panel). The satellite evolved from the free solar wind at large distances from Earth (~17.0 R_e), then crossed the bow shock at ~16.3 R_e around 17:23 UT on 18 Feb 2007 and penetrated the magnetosheath until around 00:06 UT on 19 Feb 2007 at ~10.7 R_e. This plasma region crossing is well identified by a strong increase of the electron plasma frequency f_{pe} (here from ~20 kHz to ~30 kHz). Then, the spacecraft entered the inner magnetosphere region in the cusp region (around 00:10 UT). Note that f_{pe} was not extracted using *ad hoc* methods due to a low signal-to-noise ratio. Finally, C3 spacecraft crossed the plasmasphere (characterized by higher f_{ce} values) a few minutes later.

Figure 3 shows typical active (black curves) and natural (red dotted curves) electric field individual spectra measured by the WHISPER experiment in the key plasma regions with their corresponding plasma resonances extracted by *ad hoc* methods (see Sect. 2.3). The spectra have been expressed in dB and normalized between 0 and 1. First and second panels show the active and natural spectra measured *in situ* respectively in the free solar wind and in the magnetosheath regime. In this regime, the electron temperature and the magnetic field amplitude are low, leading to the excitation of only one resonance usually triggered at, or close to, the electron plasma frequency f_{pe} (shown by blue vertical dotted bars), from which the thermal electron density is derived. For instance, in the free solar wind (resp. magnetosheath regime), f_{pe} ~ 21.1 kHz, (resp. f_{pe} ~ 48.2 kHz) corresponding to n_e ~ 5.51 cm^{-3} (resp. n_e ~ 28.76 cm^{-3}). Third panel shows the spectra measured by WHISPER in the plasmasphere regime. In this region, the electron cyclotron frequency f_{ce} (~ 5.5 kHz) and its harmonics (green) can be measured by the WHISPER instrument as well as Bernstein resonances (magenta) and the upper hybrid frequency f_{uh} (~ 29.4 kHz) (yellow). f_{pe} is then derived from f_{uh} and f_{ce} (~ 28.9 kHz).

This article is protected by copyright. All rights reserved.

The WHISPER dataset used for the self-learning algorithms presented in this paper consist of active/natural electric field spectra which have been processed by *ad hoc* methods to extract the thermal electron density. We briefly discuss this pipeline in the following section.

2.3 *Ad hoc* methods

In order to provide scientific analysis support and dissemination of high-level data, several semi-automatic algorithms have been developped to extract the thermal electron density (Trotignon et al., 2010). They are based on a manual analysis of the plasma region (a resonance pattern) based on WHISPER active and natural spectra, EFW spacecraft potential measurements and FGM magentic field measurements, coupled with an automatic extraction of spectral features (resonance, low cutoff, ...) from WHISPER active and/or natural spectra. The algorithms are specific to (i) the plasma region and (ii) the operational mode (active and natural). In the free solar wind and the magnetosheath, the extraction of the electron plasma frequency is usually performed by (i) identifying the frequency position at maximum amplitude for active spectra or, (ii) identifying the low frequency cut-off for natural spectra. When natural signatures are unclear, a proxy can also be derived, obtained from the cross-calibration of the EFW spacecraft potential and the electron plasma frequency obtained from WHISPER active spectra analysis. In the plasmasphere, the thermal electron density is extracted by (i) the location of the upper hybrid frequency or by (ii) manual analysis. For each plasma regime, the extraction process includes a manual selection of the frequency band which contains the desired characteristic frequencies, with the risk of a bad selection.

This study aims to a fully-automatic electron density production on the easiest regions to analyze (i.e. solar wind and magnetosheath regimes) in order to deliver data faster and reduce manpower and subjectivity due to the manual operations. For this purpose, we have implemented automatic routines based on neural networks. First, the region determination is the key to apply the appropriate algorithm for density calculation. Therefore, we have implemented a neural network which predicts the plasma regime. Secondly, we have applied a specific neural network to determine the thermal electron density for the solar wind/magnetosheath region (for production tool) and for the plasmasphere region (for specific events). We detail the main stages of neural network process in the next section.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

3 Multilayer neural networks

In this section, we provide a general description of neural networks of direct interest to this study. Such methods are derived from a simplified model of the neural system based on the dynamical electrochemical properties of biological neurons and their inter-connections (Mcculloch & Pitts, 1943).

Rosenblatt (1957) was the first to implement such a biological architecture, called Perceptron, to solve simple linear problems dealing with two classes of signals. This model was composed of a single layer of artificial neurons which is fully-connected to the input and the output layers. Each modern neural network is based on this model (Lecun & Bengio, 1995; Goodfellow et al., 2016). We provide a functional description of an artificial neuron in Sect 3.1. Then, we give the basic aspects of the architecture of the neural network based on a stack of artificial neuron layers (Sect. 3.2). Finally, we introduce the implementation of such tools in Sect 3.3.

3.1 Artificial neuron

An artificial neuron is an information-processing unit that produces an output y from an input signal x (Haykin, 1999). The neuron is the basic unit component of any neural network method. The input signal is typically expressed by a vector:

$$x = (x_1, \ldots, x_m) \in \mathbb{R}^m$$

The signal x of dimension m is connected to the neuron k by a set of synapses characterized by the synaptic weight, also expressed by a vector of dimension m:

$$w_k = (w_{k,1}, \ldots, w_{k,m}) \in \mathbb{R}^m$$

The signal x is multiplied by the synaptic weights which is represented in mathematical terms by a scalar product between the vectors x and w_k, called u_k:

$$u_k = \sum_{j=1}^{m} w_{k,j} x_j$$

Note that a bias, called b_k, can be added to u_k. This bias allows the scalar product u_k to deviate from 0, that can lead to a better updates of the synaptic weights during the learning phase. Then, an operation to u_k is performed by an activation function called φ. This function allows to (i) limit the amplitude range of the output signal and to (ii) introduce a non-linearity leading to a more complex transformation of the input data. The popular activation functions used on artificial neurons are (i) the hyperbolic tangent function, that is differentiable, continuous

This article is protected by copyright. All rights reserved.

Accepted Article

and delivering output values ranging from -1 to 1, (ii) the *ReLu* function ($ReLu(x) = max(x, 0)$) which is not differentiable at 0 but fast to compute, and (iii) the sigmoïd function defined as $\varphi(x) = 1/(1 + \exp(-x))$ for which the output value ranges from 0 to 1 and the derivative is non-null which allows the neural network to make some progress at every step. The output y_k of the artificial neuron is then expressed by:

$$y_k = \varphi(u_k + b_k)$$

Figure 4 summarizes the architecture of such artificial neuron with an m-dimensional signal input and m synaptic weights. Note that artificial neurons can be more elaborated, such as for recurrent neural network (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) from which the output y_k depends on the previous state of the neuron, or for convolutional network (LeCun et al., 2015) where each neuron performs a filtering operation (i.e. discrete convolution) on a specific part of the data. We discuss the architecture of the artificial neural network composed of several neuron layers in the following section.

3.2 Artificial Neural Networks

The main idea of neural networks is to extract linear combinations of input data as derived features, and then model the output as a nonlinear function of these features. These supervised methods are mainly used for classification and/or regression tasks (Goodfellow et al., 2016). As illustrated in Figure 5, a neural network is composed of several layers of artificial neurons that are interconnected with neurons from the previous and the following layer. The classical layers from literature (LeCun et al., 2015) are:

Input layer: This layer is composed of the content of the input dataset. In term of mathematical representation, each component of the dataset, called X_i, is expressed as a vector of the size of the data (here m) and the entire dataset composed of n data, is expressed by a matrix $\mathbf{X}$:

$$\mathbf{X} = (X_1, \ldots, X_m) \in \mathbb{R}^{m,n}$$

Hidden layers: The intermediate layers are called hidden layers and are composed of a set of artificial neurons (see Sect 3.1). These layers allow the neural network to learn complex tasks by extracting progressively more meaningful features from the input dataset.

This article is protected by copyright. All rights reserved.

Accepted Article

Output layer: Neural networks produce a predicted output for a given input X_i defined by a vector called y_i^* as follows:

$$y_i^* = (y_1, \ldots, y_N)$$

The output y_i^* can be expressed as a function depending on the input X_i and the free parameters (i.e. synaptic weights + additional parameters) as follows:

$$y_i^* = f(X_i; \mathbf{W})$$

where $\mathbf{W}$ is the whole synaptic weights $w_{i,j}$. f can be seen as a combination of functions of the following form:

$$f(x) = \mathbf{f}^{(d)}(\ldots \mathbf{f}^{(i)}(\ldots(\mathbf{f}^{(2)}(\mathbf{f}^{(1)}(x)))\ldots)\ldots)$$

where d is the number of hidden layers and $f^{(i)} = (f_1^{(i)}, f_2^{(i)}, \ldots, f_w^{(i)})$, where $f_w^{(i)}$ is the activation function $\varphi_{w,i}$ associated to the i^{th}-artificial neuron on the w^{th} hidden layer. Usually, for a single classification task, the sigmoïd function is chosen as the activation function of the last layer. Then, the predicted coordinates y_i of the output y_i^* are expressed between 0 and 1 and their sum equals to 1. Therefore, the output y_i^* can be seen as a probability of belonging to a class. Moreover, the corresponding classifier of the input data X_i is defined as $C(X_i) = argmax(y_i^*)$.

Neural networks are defined in a way that each neuron is structured and connected to each other. This architecture (i.e. structure) strongly depends on the problem to solve. In litterature, there exist several categories of neural network architectures. In a first category, the connection between artificial neurons can flow only in one direction from the input layer to the output layer. Note that the neurons can be fully or partially interconnected. This architecture, called multilayer feedforward network, is the simplest neural network architecture. The hidden layers are seen as feature extractions of the input dataset. In a second category, the recurrent neural network differs from the feedforward neural network in that neurons can have backward pointing connections that can be understood as a form of memory. This architecture is useful when the input data are related in time. Such methods are commonly used for the data processing of temporal series. In this study, we have implemented both architectures.

To predict the class of the input data, the neural network needs to be trained, i.e. the synaptic weights $\mathbf{W}$ need to be updated. The training step of the neural network can be seen as an optimisation problem. First, the free parameters $\mathbf{W}$ are randomly initialized and a computation of current y^* is realized. Then, the accuracy of the prediction is computed by measuring the difference between the current estimated class C_i and the true training class $C_{i,true}$. For that, we compute an objective function (i.e. a loss function), such as the categorical crossentropy

This article is protected by copyright. All rights reserved.

Accepted Article

function is defined as follows:

$$J(\theta) = \sum_{i}^{m} y_{i,true} \log(f(X_i; \mathbf{W}))$$

Then, we need to update the free parameters of the neural network to model the prediction on the training outputs. For that, the loss function J is minimized by a gradient descent method (called optimizer) using the backpropagation technique (Rumelhart et al., 1986). This technique uses the fact that the derivative of J with respect to the synaptic weights can be expressed as the prediction error. Note that the derivative of J is only compared to a limited part of the synaptic weights, chosen randomly, which is the main idea of the Stochastic Gradient Descent (SGD). Therefore, we choose new free parameters that reduce the loss function:

$$\mathbf{W}_{new} = \mathbf{W}_{old} - \epsilon \nabla_{\mathbf{W}} J(X, \mathbf{W}_{old})$$

where ϵ is the step of the descent which is commonly called the learning rate. Note that some optimizers add a more complex step or/and descent direction in order to speed up the convergence of the gradient descent (see ADAM method (Kingma & Ba, 2014), Adagrad method (Duchi et al., 2011) or RMSprop method (Tieleman & Hinton, 2012)). The update is performed iteratively for a fixed number of times called epochs.

Usually, the input and the output data are split into three different sub-datasets: (i) a training dataset, (ii) a validation dataset and, (iii) a test dataset. The training dataset is fed to the neural network which tunes the free parameters to make correct predictions on the dataset itself. Therefore, this dataset need to be representative of the global problem. The validation and the test dataset are used to know how well the neural network will generalize to new cases. If the model performs well on the training data, but it does not generalize, we say that the model overfits the training data, i.e. learning features are too specific to the training dataset. This overfitting can be reduced by using a regularization technique called dropout (Srivastava et al., 2014). Basically, this technique ensures that every neuron has a probability p of being temporarily ignored during one training step. The hyperparameter p is called the dropout rate.

To evaluate the performance on the different datasets, we compute an accuracy score such as, for a classification task, the ratio of correct predictions, which is defined as:

$$acc = \frac{1}{N} \sum_{i}^{N} \mathbb{1}_{C_i = C_{i,true}}$$

where $C_{i,true}$ is the true class of the data X_i. We have used this accuracy score in the following to evaluate the performance of the tested self-learning methods.

Note that the accuracy score does not take into account the class distribution (some classes could appear much more frequently than others) and does not give information about the num-

This article is protected by copyright. All rights reserved.

ber of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Neural network performances can be described by the following metrics: (i) the accuracy of the positive precisions among all predicted positives, called precision and computed as TP/(TP+FP), and (ii) the ratio of positive instances, called recall and computed as TP/(TP+FN), that are correctly detected by the method. As FP and FN are quite similar in this study, the performance is simply illustrated by the accuracy score.

3.3 Implementation and ressources

Nowadays, neural network methods are largely and actively used in space science, especially in space weather applications for the data processing and the automatic detection of space weather events (Camporeale et al., 2018). Zhelavskaya et al. (2016) applied neural networks to automatically determine the electron density from an electric field experiment onboard Van Allen Probes mission (Mauk et al., 2013). For that, they used the measured electric field spectra and some geomagnetic conditions such as magnetic field line or the geometric index. For this study, we chose to apply two separated neural networks to (i) determine the plasma region encountered by the spacecrafts (Sect. 4) and to (ii) extract the thermal electron density (Sect. 5), using only the WHISPER spectra, therefore ignoring external conditions.

Several other automatic methods have been tested to process the WHISPER dataset. It must be noted that a large number of sub-operational modes exist, leading to different frequency bandwidth (hence several frequency steps). Unsupervised methods (i.e. unlabeled algorithms) such as KMEANS (Arthur & Vassilvitskii, 2007) or DBSCAN (Ester et al., 1996) are sensitive to these modes, i.e. the results are driven by operational mode and not plasma regimes. Neural networks have been retained because they can overcome this issue. Moreover, they do not require feature extraction (i.e. data representation) to improve the accuracy of the prediction, in contrary to the majority of self-learning algorithms. Indeed, the hidden layers of the neural networks (see Sect. 3.2) allow the extraction of different features from the data (which are unknown and possibly difficult to interpret).

Neural networks have been implemented using Python libraries *Scikit-Learn* (Pedregosa et al., 2011) and *Keras* (Chollet et al., 2015). They have been run on a workstation featured with an AMD Threadripper 2990 WX (32 cores at 3.4 GHz) and a GPU card Nvidia Geforce 1660 GTX (6 Go DDR5 with 1406 CUDA cores).

Hereafter, we detail the neural networks used to predict the plasma regions (resp. the thermal electron density) in Section 4 (resp. in Sect. 5).

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

4 Detection of the plasma regimes

In this section, we describe the implementation of several neural networks to predict the plasma region into three classes: solar wind, magnetosheath and other plasma regions (plasmasphere, tail, cusp, ...). First, we detail the input data and the output (i.e. labels) of the neural networks (Sect. 4.1). Secondly, we compare the performance of the neural networks (Sect. 4.2). Finally, we apply the best neural network to more WHISPER data (Sect. 4.3).

4.1 Data and labels

To predict the plasma region crossed by CLUSTER-II at a specific time, we have used the WHISPER data format defined for the Cluster Science Archive (Trotignon et al., 2010). Each active spectrum contains 480 frequency bins covering the 3.5-81.5 kHz band. Note that for the sub-operational modes measured at different frequency bandwidth (see Sect. 2.1), the missing values are replaced by 0. Each natural spectrum contains 470 frequency bins covering the 3.5-79.9 kHz band. Moreover, all spectra have been expressed in decibels ($20log_{10}$) and individually normalized between 0 (amplitude min.) and 1 (amplitude max.). Therefore, each input data can be expressed as:

$$X_i = (act_1, \ldots, act_{480}, nat_1, \ldots, nat_{470})$$

where act_i (resp. nat_i) is the i^{th} value of the active spectrum (resp. natural spectrum).

The spectra X_i are classified as three different plasma regimes: (i) free solar wind, (ii) magnetosheath region and, (iii) a class containing various ionized medium (plasmasphere, tail, cusp, ...). Therefore the output y^* can be seen as a vector of three dimensions containing the estimated probability for each class. The plasma region C_i is chosen as $argmax(y^*)$.

In order to train the neural network and validate its ouputs, the region labels have been identified manually for a substantial set of data randomly chosen during the mission lifetime. We ensured that the training and the test datasets were sufficiently representative of the entirety of WHISPER spectra by verifying that selected data distributions in terms of (i) value of thermal electron density and (ii) the maximum amplitude in dB were consistent/representative with the overall same distributions. Finally, the total number of spectra are the following: 113,829 spectra from the free solar wind, 118,632 spectra from the magnetosheath and 113,551 spectra from other plasma regions and from an additionnal specific dataset composed of 40,762 spectra mesured in the plasmasphere. From this dataset, we randomly took 67% of the spectra from each plasma region to build the training (95%) and the validation dataset (5%). The

This article is protected by copyright. All rights reserved.

testing dataset is composed of the remaining 33%. In order to classify WHISPER spectra according to the plasma region encountered by the CLUSTER-II spacecraft, we have tested several neural networks and we have optimized the neural network characteristics (number of hidden layers, number of neurons, learning rate, ...) using GridSearch method (Géron, 2019). A comparison between some architectures is presented in the following section.

4.2 Comparison of several architectures

We have compared two neural networks architectures: (i) a fully-connected (DENSE) neural network (see Sect. 3.2) and (ii) a recurrent (GRU) neural network (Cho et al., 2014). Each architecture has been tested with three different input datasets contening respectively: (i) only the active spectra (A), (ii) the active and natural spectra (A+N) and, (iii) only the natural spectra (N).

To compare the performance of these models, we have first computed the variation of the accuracy on each input dataset during the training task. Results are shown in Figure 6. First, each model converged after four iterations (i.e. epochs). Secondly, we see that the neural networks which only learnt with the natural spectra (green and yellow curves) have the poorest accuracy (from 0.84 to 0.91) whatever the architecture (dense or recurrent). This observation is consistent with the fact that the plasma resonances are not actively excited in passive mode (see Fig. 3). Therefore, the plasma signatures are generally less visible in the natural spectra. Finally, the best accuracy is obtained with the fully-connected (DENSE) architecture with the active and natural spectrum combination (purple curve). The accuracy goes up to 97.5% at the last epoch. We see that the recurrent (GRU) architecture gives a lower accuracy than the fully-connected network, whatever the input data.

Secondly, we have applied these neural networks to the test dataset composed by 24,324 spectra measured in the free solar wind, 25,562 spectra measured in the magnetosheath region and 51,406 in other plasma regions (including the plasmasphere). We have computed the accuracy and the output probability given by these models. These parameters are presented in Table 1. On one hand, we can see that the dense neural network trained with only the active spectrum (DENSE A) gives the best accuracy for the solar wind and the magnetosheath (resp. 97.9 and 96.4%). However, the recurrent neural network trained with only the active spectra (GRU A) gives the best accuracy for the other plasma regions (98.5%). On the other hand, we have computed (i) the mean probability associated to the predicted plasma region Ci given by the neural network (fourth column), (ii) the mean probability associated to the predicted

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Table 1. Accuracy and mean probability of several neural networks implemented to predict the plasma region (Solar Wind, Magnetosheath and Other): GRU with active + natural spectrum in input (GRU A+N), GRU with only the active spectrum (GRU A), GRU with only natural spectrum, fully-connected NN with active + natural (DENSE A+N), fully-connected NN active (DENSE A) and fully-connected with only natural spectrum (DENSE N).

Model	Solar Wind	Magnetosheath	Other	Mean Proba.		
				Total	Correct Pred.	Wrong Pred.
GRU A+N	94.9	93.0	94.2	0.95	0.96	0.77
GRU A	95.6	92.8	98.5	0.97	0.98	0.75
GRU N	75.8	76.6	87.2	0.89	0.92	0.77
DENSE A+N	94.3	94.6	95.0	0.96	0.97	0.77
DENSE A	97.9	96.4	96.9	0.95	0.96	0.66
DENSE N	75.2	83.5	87.8	0.92	0.94	0.81

plasma region (i.e. maximum probability for each spectrum) when the model predicted the correct class and, (iii) the mean probability associated to the predicted plasma region (i.e. maximum probability for each spectrum) when the model predicted a wrong class.

We see that GRU A model has the highest mean probability for correct prediction (0.98) and DENSE A model has the lowest mean probability for wrong predictions (0.66). Therefore, if we choose to impose a threshold on the probability to delete the wrong predictions, more correct predictions will be kept with both models. In the following, we chose to apply the DENSE A model to predict the plasma regime.

Table 2 summarizes the characteristics of this neural network, which is used to predict the plasma regime. The training part needs about 1 hours for 200,000 active spectra and the prediction (i.e. testing) part takes around 5 seconds for 50,000 active spectra. We have applied the model to more WHISPER spectra. We discuss the results in the following section.

This article is protected by copyright. All rights reserved.

Accepted Article

Table 2. Characteristics of the best neural network model used to predict the plasma regime.

Model	**Plasma Regime**
Input	Active spectrum
# of spectra	305,480
Pre-processing	normalized dB
Architecture	Dense
# hidden layers	3
# neurons	1024+1024+1024
Activation function	ReLU/Sigmoïd
Dropout	30%
Loss function	Categorical Crossentropy
Optimizer	Adam
Metrics	Accuracy
Number of parameters	3,076,099
Output	3 classes (SW,MS,Other)

This article is protected by copyright. All rights reserved.

Accepted Article

4.3 Application of the best model

We have applied the model presented in Table 2 to WHISPER measurements presented in Sect 2.2 from 18 Feb 2007 at 09:43 UT to 19 Feb 2007 at 09:54 UT when the C3 spacecraft crossed several plasma boundaries. Figure 7 (middle panel) shows the corresponding classification obtained from the neural network: solar wind (yellow), magnetosheath region (red) and other (blue). We see that the plasma region boundaries are quite well defined. For instance, we note that the model predicted a solar wind region from the beginning to 17 Feb 2007 at ~17 h corresponding to the bow shock crossing (see Fig. 3). Then, the model predicted a magnetosheath region until 19 Feb 2007 at ~00:00 corresponding to the penetration of the satellite into the magnetopause. Finally, the model predicted, as expected, a magnetosphere regime. Resulting predictions are consistent with the shape of the spectra. First, only one strong resonance is observed on the active spectra (upper panel). Then, the magnetosheath region corresponds to a lower maximum amplitude leading to a low signal-to-noise ratio. Finally, the plasmasphere is well defined by the presence of several plasma resonances, especially the harmonics of the electron cyclotron frequency, which depend on the magnetic field amplitude.

Note that the magnetic field amplitude could provide a direct way to distinguish the plasma regimes. However, we only used the WHISPER dataset in order to set up an electron density pipeline independent from the other CLUSTER-II instruments, that is mainly motivated by operational considerations.

We observe that wrong predictions can be observed inside the boundaries. For instance, some spectra are labeled as magnetosheath whereas they seem to belong to the plasmasphere or the tail. In particular, in the free solar wind, the model sometimes predicted a magnetosheath regime. This can be due to (i) instrumental artefact such as a low signal-to-noise ratio or the presence of interferences potentially affecting the noise level or (ii) the fact that the prediction is only based on WHISPER observations. Indeed, WHISPER spectra acquired in the free solar wind and in the magnetosheath regions usually exhibit the same features, leading to an ambiguous prediction. Similar observations have been done with unsupervised and supervised methods (e.g. decision tree). In the other plasma regimes, including the plasmasphere, the wrong predictions are mainly due to (i) a high amplitude of the magnetic field (leading to an electron cyclotron frequency above the WHISPER frequency range) and, (ii) a poor signal-to-noise ratio.

Let us analyse the prediction of the plasma region compared to external parameters such as the magnetic field amplitude and the spacecraft position. Figure 8 shows the predicted plasma

This article is protected by copyright. All rights reserved.

region for one orbit (from 4 Jan 2013 at 00:00 UT to 6 Jan 2013 at 4:00 UT) with respect to the electron cyclotron frequency (expressed in kHz in logarithmic scale) and the X_{gse} coordinate (Russell, 1971), normalized by the Earth radius. As expected, the solar wind region is associated with a large distance from Earth (higher than 10 Earth radii) on the Sun side and a low electron cyclotron frequency (i.e. low magnetic field). Moreover, the magnetosheath region is associated with a large Earth distance range (from 2.5 to 10 Earth radii) and intermediate electron cyclotron frequencies (from 0.1 to 1 kHz). The other plasma regions are predicted close to Earth, especially the plasmasphere regime, where the plasma region is driven by a high magnetic field amplitude (i.e. high electron cyclotron frequency). Therefore, we can conclude that the predicted plasma regions are mainly driven by the spacecraft location and the magnetic field, despite the fact that the prediction is only implemented with WHISPER spectra.

In both case studies (Fig. 7 and 8), an interference or less likely other signatures (e.g. lower hybrid frequency) can be incorrectly interpreted as the unique plasma resonance leading to a classification in the magnetosheath or solar wind regime. In order to correct these wrong predictions and avoid such isolated detections, plasma regions time/space continuity could be taken into account. To this end, we have implemented another step in the predicted plasma region process: the *k*-nearest neighbors algorithm (Cover & Hart, 1967; Géron, 2019). This method is generally applied for data classification tasks using pre-labeled data as input. For each unlabeled data (or test data), the method identifies the *k*-nearest neighbors taken from the labeled data and determines the most contributing class. If we consider a uniform contribution between the *k*-nearest neighbors, the most contributing class is then equivalent to the most frequent class among the *k*-nearest neighbors. The resulting class is then assigned to the test data.

We have applied this method to the spatial domain around the Earth. In this context, the input data are the points along the spacecraft orbit (defined by their GSE coordinates) and previously labeled by the neural network. For each grid point defining a volume around the Earth (corresponding to an unlabeled test data), the algorithm determines the k-nearest neighbors and the resulting class. Finally, we have decided to take advantage of this large-scale labelling process to correct the region associated with each point of the orbit by imposing the label (i.e. plasma region) of the closest volume point.

Figure 9 shows the results for one complete orbit of C1 spacecraft in the XY_{gse} frame, normalized by Earth radius, from 4 Jan 2013 at 00:00 UT to 6 Jan 2013 at 4:00 UT. The plasma

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

regions predicted by the neural network are given by the color points (yellow: solar wind, red: magnetosheath and blue: other) along the orbit. The background color represents the plasma regions predicted by the *k*-nearest neighbors algorithm from the neural network predictions as inputs. We see that wrong detections made by the neural network can be corrected by the large-scale classification given by the k-nearest neighbors algorithm, as the plasma region boundaries are now well defined. Note that using this method with a high value for k, the plasma region is less sensitive to wrong predictions (not shown here) but cannot account for fast multiple boundary crossings (e.g. when skimming the bow shock).

As explained in Sect. 2.2, the detection of the thermal electron density depends on the plasma regime. The automatic prediction of the plasma region is then used as an input of the extraction of the thermal electron density. We present the automatic model and the results of the prediction of the thermal electron density in the following section.

5 Prediction of the thermal electron density

In this study, we chose to define two separated neural networks which apply in different plasma regimes. First, in the solar wind/magnetosheath regime, a self-learning algorithm is implemented in order to fully automate the extraction of the thermal electron density. We describe the method and the results in Sect. 5.1. Secondly, we focused on the plasmasphere region from which the extraction of the thermal electron density is more intricate, and *ad hoc* methods require a manual intervention. We present the neural network and the given predictions in Sect. 5.2.

5.1 In the solar wind/magnetosheath regime

5.1.1 Data and labels

As explained in Sect. 2.2, the two plasma regions considered now (solar wind and magnetosheath, respectively) are characterized by the presence of a main resonance on spectra located at, or close to, the electron plasma frequency. Therefore, the thermal electron density can be directly derived from the predicted electron plasma frequency, and is obtained in the 0.15 to 82.37 cm^{-3} range (corresponding to the 3.5-82 kHz frequency range). Note that *ad hoc* methods can provide an electron plasma frequency which could be not aligned on the emitted frequency bins of the active mode. In that case, the electron plasma frequency has been translated onto the nearest active frequency bin. Therefore, the frequencies have been rescaled

This article is protected by copyright. All rights reserved.

Accepted Article

on the emitted frequency bins of the active mode. Thus, the output of the model can be seen as a vector of 480 classes corresponding to the emitted frequencies defined in the 3.5-82 kHz bandwidth with a 162.8 Hz resolution. For the input data, we have used the active spectra and the time-closest natural spectra measured by WHISPER where the thermal electron density has been extracted from *ad hoc* methods adapted to the solar wind and the magnetosheath regimes.

Note that we have chosen not to consider a class corresponding to the case where *ad hoc* methods were not able to extract the electron plasma frequency, that means that spectra without associated plasma frequency were discarded from the analysis. The prediction of such models are sensitive to the presence of this class (not shown here), in particular because the extraction of the electron plasma frequency from *ad hoc* methods is mainly based on natural spectra and then assigned to the closest active spectrum in time. Therefore, two similar active spectra will not necessarily lead to the same plasma frequency estimations. Actually, the training part can be strongly affected by such a case, when the model learning is based on active spectra.

The labels (i.e. the electron plasma frequencies) have been identified manually for a substential set of data selected during the entire mission lifetime. We ensured that the data is representative of the entire WHISPER dataset, in particular with respect to the distribution of the electron plasma frequency for each plasma regime. The total numbers of spectra are the following: 73,709 spectra for the solar wind and 77,458 spectra for the magnetosheath regime. From this dataset, we randomly took 67% of the spectra from each plasma region to build the training and the validation dataset. The testing dataset is composed of the remaining 33%.

5.1.2 Comparison of several architectures

Like in Sect. 4, we have tested several neural network architectures to predict the thermal electron density in the solar wind and in the magnetosheath regime. We have implemented two architectures: (i) a fully-connected (DENSE) neural network and, (ii) a recurrent (GRU) neural network. Each architecture has been trained and tested with the same dataset composed of three sub-datasets with (i) only active spectra (A), (ii) with active and natural spectra (A+N) and, (iii) only natural spectra (N).

Figure 10 shows the evolution of the accuracy of the two neural network architectures separately trained with the three sub-datasets. First, we see that the two architectures tested with only the natural spectra give the worst accuracy (between 0.1 to 0.2), probably due to lower signal-to-noise ratio. As explained in Sect. 4, in passive mode, the signal-to-noise ratio is lower

This article is protected by copyright. All rights reserved.

Accepted Article

than the sounding (active) mode. Therefore, the determination of the plasma frequency is more tricky for this operational mode. Secondly, the best accuracy is obtained by the recurrent neural network (GRU) tested with only the active spectra. The accuracy reaches up to 0.7 at the last epoch whereas the accuracy of the other neural networks do not exceed 0.6. Thus, we chose to apply the GRU architecture with only the active spectrum to predict the electron plasma frequency in the free solar wind and in the magnetosheath regime. Table 3 summarizes the main parameters of the retained neural network. The training part needs about 2 hours for 200,000 active spectra and the prediction (i.e. testing) part takes around 40 seconds for 50,000 active spectra. We have applied the model to more WHISPER spectra. We discuss the results of this model on the WHISPER dataset in the following section.

5.1.3 Results

In this section, we comment on the results of the thermal electron density automated detection in the solar wind/magnetosheath region using the best neural network chosen in the previous section.

First, we have tested the recurrent neural network presented in section 5.1 on 24,323 (resp. 25,561) active spectra acquired in the free solar wind (resp. in the magnetosheath regime). We have computed the difference between the predicted electron plasma frequency ($f_{pe,pred}$) and the electron plasma frequency given by *ad hoc* methods ($f_{pe,adhoc}$), considered as ground truth. Figure 11 shows the percentage of correct predictions within a tolerance interval (expressed in kHz) for different neural network architectures. We can see that the recurrent (GRU) neural network with active spectra as input (represented by the red line) gives the best absolute accuracy (70%). By increasing the tolerance interval by one frequency bin (± 0.182 Hz) for each predicted electron plasma frequency, we see that recurrent and dense models are give quite similar accuracy (95%). This large increase of the precision is mainly due to the fact that the electron plasma frequency, extracted by *ad hoc* methods, is given as a real value and the predicted electron plasma frequency is expressed into the 480 frequency bins of the active mode. Moreover, the detection of the electron plasma frequency by *ad hoc* methods, sometimes, can be done with only the natural spectrum. Therefore, in the active spectrum, the predicted plasma frequency can be different by one or several bins compared to the plasma frequency given by *ad hoc* methods.

Secondly, we have run the neural network on active spectra measured by WHISPER on C1 on 13 Feb 2012 between 00:02 and 13:21 UT. At this time, the spacecraft were in the free

This article is protected by copyright. All rights reserved.

Accepted Article

Table 3. Summary of the neural network used to determine the thermal electron density in the Solar Wind and Magnetosheath region

Model	**SW/MS**
Input	Active spectrum
# of spectra	151,167
Pre-processing	normalized dB
Architecture	Recurrent (GRU)
# hidden layers	3
# neurons	1024+1024+1024
Activation function	ReLU/Sigmoïd
Dropout	30%
Loss function	Categorical Crossentropy
Optimizer	Adam
Metrics	Accuracy
Number of parameters	17,215,491
Output	f_{pe} on 480 freq. bins (3.5 to 81.5 kHz)

This article is protected by copyright. All rights reserved.

Accepted Article

solar wind region characterized by a single plasma signature observed on active spectra. Figure 12 shows the prediction of the electron plasma frequency f_{pe}, expressed in frequency bins (red points) and the electron plasma frequency given by *ad hoc* methods (black points).

For this case study, we have computed the relative error made by the self-learning algorithm compared to the thermal electron density n_e derived from the electron plasma frequency f_{pe}, extracted using *ad hoc* methods. The distribution of the electron density relative error is shown in Figure 13 (upper panel). We see that the relative error is low. Notably, the mean of the relative error is around 0.01. The bottom panel shows the scatter plot of the predicted thermal electron density n_e versus the density extracted by *ad hoc* methods. We see that the predictions follow the electron density given by *ad hoc* methods. However, there exist wrong predictions for which the model gave $n_e \sim$ 2-5cm^{-3} and *ad hoc* methods, $n_e \sim$12 cm^{-3}. Taking into account the probability y_i corresponding to the predicted class given by the model (colorbar), we see that the probability is low (from 0.2 to 0.4). Therefore, it could be possible to delete these wrong predictions by imposing a reasonable threshold on the probability y_i.

We also predicted the thermal electron density in the solar wind and the magnetosphere regions for a larger amount of data. We have applied the self-learning method to 29,735 WHISPER data measured onboard C1 in 2012. Figure 14 shows three scatter plots of the thermal electron density showing the predicted electron density versus the density detected by *ad hoc* methods. The first panel shows all the data. The middle panel shows the data where the probability is higher than 0.2. The amount of data has decreased to 28,094. The right panel shows the data where the probability is higher than 0.5. The amount of data has decreased to 21,386. We see that the number of wrong predictions also decreases. Therefore, if we want to automatically extract the thermal electron density, it is possible to impose a probability threshold in order to minimize the wrong predictions.

5.2 In the plasmasphere regime

5.2.1 Data and labels

The prediction of the thermal electron density in the plasmasphere is more tricky than in the solar wind and the magnetosheath regions. In active spectra, the resonance corresponding to the electron plasma frequency is less visible, when it is, than the resonance located at the upper hybrid frequency (see Figure 3). Therefore, we have chosen to identify the upper hybrid frequency in a first step, and then, knowing the electron cyclotron frequency from the

This article is protected by copyright. All rights reserved.

Accepted Article

measurement of the magnetic field amplitude of FGM, to derive the electron plasma frequency. This automatic pipeline is shown in figure 15.

We have used active spectra and the closest natural spectrum measured by WHISPER for which the thermal electron density has been extracted using *ad hoc* methods. We only kept spectra (i) that are not labeled in the solar wind and magnetosheath region and (ii) spectra where the electron cyclotron frequency (derived from the magnetometer) is higher than 1.5 kHz. Then, we deleted 16% of the dataset ($\sim$ 58,000 spectra) by applying thresholds on several characteristics of active spectra. In particular, for each active spectrum, we have computed the number of resonances that are not located close to the electron cyclotron frequency or to its harmonics. A high value of this number (i.e. higher than 5) means that the key plasma resonances, especially the resonance associated with the upper hybrid frequency, if measured, are not well-described in the spectrum. Moreover, we have modified the threshold to be less restrictive using the mean amplitude at low and high frequencies in order to limit the impact of the noise at such frequencies. A combination of the two descriptors allows to delete different kinds of active spectra which are considered as not suitable for detection of the electron plasma frequency or the upper hybrid frequency.

On the selected active spectra, a top-hat filter (Serra, 1983) has been applied in order to amplify the plasma resonances. Moreover, the spectra have been normalized in dB between 0 and 1. The output of the model can be seen as a vector of 480 classes corresponding to the emitted frequencies defined on the 3.5-81.5 kHz bandwidth with a precision of 162.8 Hz. Note that the electron plasma frequency can be lower than 3.5 kHz. 360,607 spectra measured in the plasmasphere have been used. A careful study of this dataset has been performed in order to adopt a well representative dataset of the WHISPER instrument in the plasmasphere. In particular, we ensured that in both datasets, the distribution of the upper hybrid frequency, compared to the electron cyclotron frequency and its harmonics or Bernstein waves, was similar to the entire measurements of the WHISPER experiment. From the entire dataset, we randomly took 66.7% of the spectra to build the training and the validation datasets. The testing dataset is composed of the remaining 33.3%.

Table 4 summarizes the characteristics of the best neural network used in the plasmasphere to predict the upper hybrid frequency. Note that the model uses a hybrid activation function composed of an Exponential Linear Unit (ELU) and a hyperbolic tangent function (Clevert et al., 2016; Manessi & Rozza, 2018). We apply the Nesterov Accelerated Gradient method (Nesterov, 1983). This method measures the gradient of the cost function not at the local po-

This article is protected by copyright. All rights reserved.

sition but slightly ahead in the direction of the momentum vector that points generally in the right direction. Note that the low dropout rate (5%), considered as a hyperparameter of the model, has been obtained using a grid search method (Bergstra & Bengio, 2012). We applied this neural network to the WHISPER dataset. The results are shown in the following section.

5.2.2 Results

Figure 16 shows an almost perfect case of active spectrogram in the plasmasphere measured on C1 (RUMBA) on 5 August 2002 between 17:00 and 20:40 UT. It shows several plasma resonances corresponding to the electron cyclotron frequency f_{ce} and its harmonics, the upper hybrid frequency f_{uh}, and Bernstein frequencies. The corresponding frequency bins of the electron plasma frequency f_{pe}, when obtained with *ad hoc* methods, are shown in the bottom panel as black points, while those predicted by the automatic pipeline are shown in red.

We have computed the relative error of the predicted thermal electron density derived from the predicted f_{pe} for this case study. The distribution of the relative error is shown in Figure 17 (upper panel). The mean relative error is around 0.09, thus larger than in the solar wind/magnetosheath regime. We have also computed the scatter plot of the predicted thermal electron density versus the density given by *ad hoc* methods (Fig. 17, lower panel). We can observe that the model seems sometimes perturbed by other resonances, as illustrated by vertical lines formed by the points on the scatter plot for this case study. The probability y_i of each prediction is represented by the colorbar. In contrast with the solar wind or the magnetosheath region results, the model gives a lower probability of correct predictions (around 0.2). This low probability is mainly due to the fact that the active spectra shown in this case study are of exceptional quality, with all of the key plasma resonances well-described. This kind of spectra is rare and thus poorly represented in the training dataset, then leading to a low probability value for the prediction. For most of spectra measured in the plasmasphere, the probability of correct predictions is quite similar to the probability given in the solar wind/magnetosheath regime.

6 Future data processing

In this section, we first describe the applicability of self-learning methods for the data processing of WHISPER spectra (section 6.1). Secondly, we discuss the applicability of such automatic data processing for the future active electric field experiments onboard spacecraft such as BepiColombo/Mio or JUICE (section 6.2).

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Table 4. Characteristics of the best neural network used to determine the upper hybrid frequency in the plasmasphere regime. The hybrid activation function is a combination of an Exponential Linear Unit (ELU) and a hyperbolic tangent function.

Model	**Plasmasphere**
Input	Active + Natural spectrum
# of spectra	360,607
Pre-processing	normalized dB + additionnal processing (top hat)
Architecture	Dense
# hidden layers	2
# neurons	512+480
Activation function	Hybrid function*
Dropout	5%
Loss function	Categorical Crossentropy
Optimizer	Nesterov
Metrics	Accuracy
Number of parameters	964,032
Output	f_{uh} on 480 freq. bins (3.5 to 81.5 kHz)

This article is protected by copyright. All rights reserved.

Accepted Article

6.1 WHISPER data processing

In this study, we have applied self-learning algorithms to predict the thermal electron density on specific data or events. We could apply these methods to the complete WHISPER dataset, present and future, and replace *ad hoc* methods (see Sect. 2.3), currently in use, by the self-learning methods shown in this study. However, the high cadence of WHISPER spectra measurements requires an optimization of the manual intervention, especially for deleting wrong detections. In order to minimize this intervention, we have implemented a new pipeline that evaluates the self-learning algorithms in the solar wind and the magnetosheath regime. This tool, named FP JEDAI (Judging Electron Density by Artificial Intelligence), allows to plot, visualize and validate the predictions made by the neural networks. It is developed with the main objective of reducing the tedious and time-consuming electron density manual determination or *ad hoc* methods selection. FP JEDAI combines two models described in this study: the region classification model and the f_{pe} determination model only applied in the solar wind and the magnetosheath regimes.

First, the model used in the tool predicts the plasma region for each spectrum. Using the time stamp and the probabilities provided by the model, large intervals are automatically selected. This step allows to automatically delete the spurious predictions, in particular the predictions of dynamic plasma regions that varies a lot in a short time period. Secondly, the electron plasma frequency is extracted from the selected spectra. Combined with a probability threshold, a contrast formula, that gives information about the resonance-to-noise ratio, is computed in order to select only the best predictions. The interface allows manual selection of the electron plasma frequency in order to delete the bad predictions. Finally, the thermal electron density is computed and saved in a file.

FP JEDAI could allow to deliver density files with human intervention reduced up to 10 times for WHISPER active spectra (compared to the *ad hoc* methods) and only limited to a validation step (as opposed as a preprocessing step and a validation step for the currently used *ad hoc* methods). For instance, the processing of one month of data for the four spacecraft requires approximatively 4h of manual intervention whereas the *ad hoc* methods required around 40h. In order to limit ambiguous predictions, we imposed a severe threshold (around 0.4) on the probability associated to each density estimate. This can lead to a 20%-loss in the number of electron densities given by neural networks as compared to *ad hoc* methods.

This article is protected by copyright. All rights reserved.

Accepted Article

6.2 Future electric field experiments

Self-learning algorithms could be used for future space missions such as BepiColombo (Benkhoff et al., 2010) or JUICE (Grasset et al., 2013) which include electric field experiments such as mutual impedance experiments (Trotignon et al., 2006) or quasi-thermal noise spectroscopy (Moncuquet et al., 2006).

For instance, the Active Measurement of Mercury's Plasma (AM^2P, (Trotignon et al., 2006)) experiment is one element of the Plasma Wave Investigation (PWI, Kasaba et al. (2020)) consortium onboard the Mercury Magnetospheric Orbiter (Mio/MMO) of the BepiColombo mission, successfully launched in October 2018. From March 2026, the mutual impedance experiment aims to determine the thermal electron density in the Hermean magnetosphere for plasma frequencies in the 0.7-120 kHz band. The Mio/MMO spacecraft will fly along an elliptic polar orbit of $400 \times 11,824$ km with a 9.3 h period (Benkhoff et al., 2010), so that the spacecraft will cross several plasma regions (Yagi et al., 2017; Kasaba et al., 2020) such as the free solar wind, where the electron cyclotron frequency is expected to be much smaller than the electron plasma frequency, and the Hermean plasmasphere, where the electron cyclotron frequency and its harmonics will be much larger than the electron plasma frequency. Therefore, the electric field spectrogram should present similarities, in terms of frequency position of plasma signatures, with the Earth magnetospheric observations made by the WHISPER instrument. Prediction of the plasma region will also be highly desirable to automatically extract the plasma bulk properties from the AM^2P measurements. In contrast with the WHISPER experiment which has been operating for more than twenty years, the mutual impedance experiment should monitor the Hermean plasma environment for two and a half terrestrial years at most (including a planned one terrestrial year extension). Moreover, the AM^2P measurement cadence will be lower than the one of WHISPER instrument, with active measurements planned to be performed every 160–200 s in the Hermean magnetosphere and every 320–400 s in the free solar wind. Therefore, the amount of AM^2P active spectra will be substantially lower than those of the WHISPER dataset. In order to characterize the capability to implement a neural network to predict the plasma regime, we have studied the influence of the amount of spectra needed for an efficient prediction of the plasma regime.

Figure 18 shows the accuracy of the neural network on a test dataset with respect to the number of active spectra in the training dataset from 10 to 10,000 spectra. The corresponding time period to measure such number of spectra with the AM^2P instrument is given by the colored verticals bars from 1 day (blue) to 4 months (red). The training dataset is built with

This article is protected by copyright. All rights reserved.

Accepted Article

spectra randomly chosen. In order to be independent of this random part, we have run the model several times with different training datasets. The overall accuracy score is then given with an error bar (shown in vertical black lines) for each case. We see that the accuracy goes up from 0.5 to 0.95, meaning that an automatic pipeline to predict the plasma region will be efficient after 4 months of operations. Note that this estimation is based on the assumption that all spectra are exploitable, especially in the solar wind and the magnetosheath regions where a clear single signature is expected to be observed. Therefore, the number of spectra actually needed in the training dataset could be higher in real operational conditions.

However, in order to increase the size of the training dataset, it could be possible to model the electric field experiments. Indeed, modeling of the experimental response has been performed in several plasma conditions such as in interplanetary plasma or planetary plasma (Béghin, 1995; Gilet et al., 2017; Wattieaux et al., 2019; Gilet et al., 2019). Recent studies show that modeling of the experiment taking into account the effect of a vacuum sheath around the experiment and the spacecraft, are close to *in situ* observations (Wattieaux et al., 2019). Moreover, thanks to a comparison between the observations and the dataset of modeled spectra, it is possible to characterize the surrounding plasma. Wattieaux et al. (2020) provided a characterization of two-electron-temperature-plasma in the ionosphere of 67P/Churyumov-Gerasimenko using a combination of the observations made by the mutual impedance probe (RPC-MIP, Trotignon et al. (2007)) and the modeling of the instrument response. Therefore, it is possible to generate a dataset of modeled electric field spectra in order to train a model to predict the plasma region and, eventually, the thermal electron density. A complementary alternative could be to make use of modelling of the plasma regions in the context of the JUICE mission using ionospheric modeling of Ganymede (Leclercq et al., 2016; Leblanc et al., 2017; Carnielli et al., 2019) from which the plasma bulk properties can be extracted to model the instrumental responses of mutual impedance experiments (Gilet et al., 2017; Wattieaux et al., 2019).

7 Conclusion

In this study, we have implemented several automated pipelines based on neural network methods to extract the thermal electron density from the electric field spectra measured by WHISPER instrument onboard the four CLUSTER-II spacecraft. We have seen that the determination of the thermal electron density mainly depends on the plasma regime, which is driven by the magnetic field amplitude and the solar activity. In the electric field spectrogram, the ef-

This article is protected by copyright. All rights reserved.

fect of the magnetic field leads to the presence of multiple plasma resonances, thus complicating the determination of the electron plasma frequency.

Therefore, we first predicted the plasma region encountered by the WHISPER instrument from the active and natural spectra. We have used a fully-connected (dense) neural network composed of three hidden layers with additional dropout. We have shown that the model reaches an accuracy score close to 97% for the three plasma regimes. Then, we have implemented two neural networks in the solar wind/magnetosheath region (i.e. low magnetic field) and in the plasmasphere (i.e. high magnetic field).

In the low magnetic field regime, the main resonance is close to the electron plasma frequency, that is directly related to the thermal electron density. Therefore, we chose to predict the electron plasma frequency, defined in the frequency range of the WHISPER instrument. We have used a recurrent neural network based on Gated Recurrent Unit (GRU) built with three hidden layers of 1024 neurons with additional dropout (30%). In the free solar wind and in the magnetosheath regimes, the prediction accuracy score reaches up to 95%. A fully-automatic pipeline is currently under development (FP JEDAI) is order to produce datasets of thermal electron density using outputs of the neural networks (prediction and probability).

For the plasmaspheric region (in practice, for electron cyclotron frequencies higher than 1.5 kHz), the spectral signature at the upper hybrid frequency, related to the electron cyclotron frequency and the electron plasma frequency, is much more simple to detect than the signature at the electron plasma frequency. Therefore, we first chose to predict the upper hybrid frequency. Then, with the cyclotron frequency computed from the magnetic field amplitude measured by the magnetometer, it is possible to automatically determine the thermal electron density. We have shown that active spectra need to be pre-processed in order to increase the accuracy of the upper hybrid frequency prediction. In the plasmasphere regime, the prediction accuracy score reaches up to 75%. For the other plasma regions such as the cusp or the tail where the signal-to-noise ratio is low, the extraction of the thermal electron density is more tricky. A specific automatic method is currently under development.

Finally, we have shown that automatic methods could be applied for future data processing of electric field experiments such as mutual impedance experiments onboard BepiColombo (Hermean environment) and JUICE (jovian environment) missions with constraints related to the size of the dataset. The efficiency of the automatic process strongly depends on the quality and the representativeness of the training dataset. We have shown that in the context of BepiColombo, a self-learning predictive model of PWI/AM^{2}P instrumental response in the Her-

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

mean environment should only be effective after 4 months of operations around Mercury. However, modeling of the instrumental responses (Gilet et al., 2017; Wattieaux et al., 2019) could be used to feed the training dataset and shorten this period.

Acknowledgments

The CLUSTER-II WHISPER electric field spectra and the thermal electron densities used in this study are publicly available to the world-wide science community on the ESA Cluster Science Archive (CSA, Laakso et al. (2010)) at `https://csa.esac.esa.int`. Some of the results presented in this paper have been obtained through the use of a high-performance workstation funded by the French Alternative Energies and Atomic Energy Commission (CEA). The workstation has been designed and maintained by Matthieu Garnung and Sebastien Celestin at LPC2E. We acknowledge CNES fundings through APR.

References

Arthur, D., & Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Soda '07: Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Bahnsen, A., Jespersen, M., Ungstrup, E., Pottelette, R., Malingre, M., Decreau, P. M. E., ... Pedersen, B. M. (1986). First VIKING results: High frequency waves. *NASA STI/Recon Technical Report N*, *87*.

Balogh, A., Dunlop, M. W., Cowley, S. W. H., Southwood, D. J., Thomlinson, J. G., Glassmeier, K. H., ... Kivelson, M. G. (1997, January). The Cluster Magnetic Field Investigation. *Space Science Review*, *79*, 65-91. doi: 10.1023/A:1004970907748

Béghin, C. (1995, March). Series expansion of electrostatic potential radiated by a point source in isotropic Maxwellian plasma. *Radio Science*, *30*, 307-322. doi: 10.1029/94RS03167

Belmont, G. (1981, November). Characteristic frequencies of a non-maxwellian plasma: A method for localizing the exact frequencies of magnetospheric intense natural waves near $;_{pe}$. *Planetary and Space Science*, *29*(11), 1251-1266. doi: 10.1016/0032-0633(81)90130-6

Benkhoff, J., van Casteren, J., Hayakawa, H., Fujimoto, M., Laakso, H., Novara, M., ... Ziethe, R. (2010, January). BepiColombo: Comprehensive exploration of Mercury:

This article is protected by copyright. All rights reserved.

Mission overview and science goals. *Planetary and Space Science*, *58*, 2-20. doi: 10.1016/j.pss.2009.09.020

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*, 281-305. Retrieved from http://dblp.uni-trier.de/db/journals/jmlr/jmlr13.html#BergstraB12

Bernstein, I. B. (1958, January). Waves in a Plasma in a Magnetic Field. *Physical Review*, *109*(1), 10-21. doi: 10.1103/PhysRev.109.10

Breuillard, H., Dupuis, R., Retino, A., Le Contel, O., Amaya, J., & Lapenta, G. (2020, September). Automatic classification of plasma regions in near-Earth space with supervised machine learning: application to Magnetospheric Multi Scale 2016-2019 observation. *Frontiers in Astronomy and Space Sciences*, *7*, 55. doi: 10.3389/fspas.2020.00055

Camporeale, E., Wing, S., & Johnson, J. (2018). *Machine learning techniques for space weather*. Elsevier Science. Retrieved from https://books.google.fr/books?id=zaRBDwAAQBAJ

Canu, P., Décréau, P. M. E., Trotignon, J. G., Rauch, J. L., Seran, H. C., Fergeau, P., ... Yearby, K. (2001, Oct). Identification of natural plasma emissions observed close to the plasmapause by the Cluster-Whisper relaxation sounder. *Annales Geophysicae*, *19*(10), 1697-1709. doi: 10.5194/angeo-19-1697-2001

Carnielli, G., Galand, M., Leblanc, F., Leclercq, L., Modolo, R., Beth, A., ... Jia, X. (2019). First 3d test particle model of ganymede's ionosphere. *Icarus*, *330*, 42 - 59. doi: https://doi.org/10.1016/j.icarus.2019.04.016

Chasseriaux, J. M., R. Debrie, R. D., & C. Renard, C. R. (1972, October). Electron density and temperature measurements in the lower ionosphere as deduced from the warm plasma theory of the h.f. quadrupole probe †. *Journal of Plasma Physics*, *8*, 231-253. doi: 10.1017/S0022377800007108

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D14-1179 doi: 10.3115/v1/D14-1179

Chollet, F., et al. (2015). *Keras.* https://github.com/fchollet/keras. GitHub.

This article is protected by copyright. All rights reserved.

Accepted Article

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, *abs/1511.07289*.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. , *13*, 21- 27. Retrieved from http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1053964

Darrouzet, F., de Keyser, J., Décréau, P. M. E., El Lemdani-Mazouz, F., & Vallières, X. (2008, Aug). Statistical analysis of plasmaspheric plumes with Cluster/WHISPER observations. *Annales Geophysicae*, *26*(8), 2403-2417. doi: 10.5194/angeo-26-2403-2008

Décréau, P. M. E., Béghin, C., & Parrot, M. (1978, November). Electron density and temperature, as measured by the mutual impedance experiment on board GEOS-1. *Space Science Reviews*, *22*, 581-595. doi: 10.1007/BF00223942

Décréau, P. M. E., Fergeau, P., Krannosels'kikh, V., Leveque, M., Martin, P., Randriamboarison, O., ... Mogensen, P. B. (1997, Jan). Whisper, a Resonance Sounder and Wave Analyser: Performances and Perspectives for the Cluster Mission. *Space Science Review*, *79*, 157-193. doi: 10.1023/A:1004931326404

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning*, *12*, 2121–2159.

El-Lemdani Mazouz, F., Rauch, J., Décréau, P., Trotignon, J., Vallières, X., Darrouzet, F., ... Suraud, X. (2009). Wave emissions at half electron gyroharmonics in the equatorial plasmasphere region: Cluster observations and statistics. *Advances in Space Research*, *43*(2), 253 - 264. Retrieved from http://www.sciencedirect.com/science/article/pii/S0273117708003517 doi: https://doi.org/10.1016/j.asr.2008.06.007

Escoubet, C. P., Masson, A., Laakso, H., & Goldstein, M. L. (2015, Oct). Recent highlights from Cluster, the first 3-D magnetospheric mission. *Annales Geophysicae*, *33*(10), 1221-1235. doi: 10.5194/angeo-33-1221-2015

Escoubet, C. P., Schmidt, R., & Goldstein, M. L. (1997, Jan). Cluster - Science and Mission Overview. *Space Science Reviews*, *79*, 11-32. doi: 10.1023/A:1004923124586

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd international conference on knowledge discovery and* (p. 226-231).

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media. Retrieved

This article is protected by copyright. All rights reserved.

Accepted Article

from https://books.google.fr/books?id=HHetDwAAQBAJ

Gilet, N., Henri, P., Wattieaux, G., Cilibrasi, M., & Béghin, C. (2017). Electrostatic potential radiated by a pulsating charge in a two-electron temperature plasma. *Radio Science*, *52*(11), 1432–1448. Retrieved from http://dx.doi.org/10.1002/2017RS006294 (2017RS006294) doi: 10.1002/2017RS006294

Gilet, N., Henri, P., Wattieaux, G., Myllys, M., Randriamboarison, O., Béghin, C., & Rauch, J.-L. (2019, April). Mutual impedance probe in collisionless unmagnetized plasmas with suprathermal electrons - Application to BepiColombo. *Frontiers in Astronomy and Space Sciences*, *6*, 16. doi: 10.3389/fspas.2019.00016

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA, USA: MIT Press. (http://www.deeplearningbook.org)

Grasset, O., Dougherty, M. K., Coustenis, A., Bunce, E. J., Erd, C., Titov, D., ... Van Hoolst, T. (2013, Apr). JUpiter ICy moons Explorer (JUICE): An ESA mission to orbit Ganymede and to characterise the Jupiter system. *Planetary and Space Science*, *78*, 1-21. doi: 10.1016/j.pss.2012.12.002

Gustafsson, G., Bostrom, R., Holback, B., Holmgren, G., Lundgren, A., Stasiewicz, K., ... Wygant, J. (1997, Jan). The Electric Field and Wave Experiment for the Cluster Mission. *Space Science Reviews*, *79*, 137-156. doi: 10.1023/A:1004975108657

Hamelin, M. (1978). *Contributionà l'étude des ondes électrostatiques et électromagnétiques au voisinage de la fréquence hybride basse dans le plasma ionosphérique* (Theses, Université d'Orléans). Retrieved from https://tel.archives-ouvertes.fr/tel-00752196

Harvey, C. C., Etcheto, J., & Mangeney, A. (1979, March). Early Results from the ISEE Electron Density Experiment [Article published in the Special issues: Advances in Magnetospheric Physics with GEOS- 1 and ISEE - 1 and 2. (Proceedings of the 13th ESLAB Symposium, innsbruck, 1978. - pp. 3-133)]. *Space Science Reviews*, *23*(1), 39-58. doi: 10.1007/BF00174110

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Prentice Hall.

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. Retrieved from https://doi.org/10.1162\%2Fneco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

Johnstone, A. D., Alsop, C., Burge, S., Carter, P. J., Coates, A. J., Coker, A. J., ... Woodliffe, R. D. (1997, January). Peace: a Plasma Electron and Current Experi-

This article is protected by copyright. All rights reserved.

Accepted Article

ment. *Space Science Reviews*, *79*, 351-398. doi: 10.1023/A:1004938001388

Kasaba, Y., Kojima, H., Moncuquet, M., Wahlund, J.-E., Yagitani, S., Sahraoui, F., ... Usui, H. (2020, June). Plasma Wave Investigation (PWI) Aboard BepiColombo Mio on the Trip to the First Measurement of Electric Fields, Electromagnetic Waves, and Radio Waves Around Mercury. *Space Science Reviews*, *216*(4), 65. doi: 10.1007/s11214-020-00692-9

Kingma, D., & Ba, J. (2014, 12). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kougblénou, S., Lointier, G., Décréau, P. M. E., Trotignon, J. G., Rauch, J. L., Vallières, X., ... Pickett, J. (2011, Nov). Lower hybrid resonances stimulated by the four CLUSTER relaxation sounders deep inside the plasmasphere: observations and inferred plasma characteristics. *Annales Geophysicae*, *29*(11), 2003-2018. doi: 10.5194/angeo-29-2003-2011

Krall, N. A., & Trivelpiece, A. W. (1973). *Principles of plasma physics*. McGraw-Hill Book Company.

Laakso, H., Perry, C., McCaffrey, S., Herment, D., Allen, A. J., Harvey, C. C., ... Turner, R. (2010, Jan). Cluster Active Archive: Overview. *Astrophysics and Space Science Proceedings*, *11*, 3-37. doi: 10.1007/978-90-481-3499-1_1

Leblanc, F., Oza, A. V., Leclercq, L., Schmidt, C., Cassidy, T., Modolo, R., ... Johnson, R. E. (2017, September). On the orbital variability of Ganymede's atmosphere. *Icarus*, *293*, 185-198. doi: 10.1016/j.icarus.2017.04.025

Leclercq, L., Modolo, R., Leblanc, F., Hess, S., & Mancini, M. (2016, March). 3D magnetospheric parallel hybrid multi-grid method applied to planet-plasma interactions. *Journal of Computational Physics*, *309*, 295-313. doi: 10.1016/j.jcp.2016.01.005

Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech and time series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 255–258). The MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved from `https://doi.org/10.1038/nature14539` doi: 10.1038/nature14539

Lefeuvre, F., Roux, A., Delaporte, B., Dunford, C., Woolliscroft, L. J. C., Davies, P. N. H., ... Gough, M. P. (1993, March). *The Wave Experiment Consortium.* Cluster: Mission, Payload, and Supporting Activities.

This article is protected by copyright. All rights reserved.

Accepted Article

Manessi, F., & Rozza, A. (2018). Learning combinations of activation functions. *CoRR*, *abs/1801.09403*. Retrieved from http://arxiv.org/abs/1801.09403

Mauk, B. H., Fox, N. J., Kanekal, S. G., Kessel, R. L., Sibeck, D. G., & Ukhorskiy, A. (2013, November). Science Objectives and Rationale for the Radiation Belt Storm Probes Mission. *Space Science Reviews*, *179*(1-4), 3-27. doi: 10.1007/s11214-012-9908-y

Mcculloch, W., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 127–147.

Moncuquet, M., Matsumoto, H., Bougeret, J.-L., Blomberg, L. G., Issautier, K., Kasaba, Y., ... Zarka, P. (2006, January). The radio waves and thermal electrostatic noise spectroscopy (SORBET) experiment on BEPICOLOMBO/MMO/PWI: Scientific objectives and performance. *Advances in Space Research*, *38*, 680-685. doi: 10.1016/j.asr.2006.01.020

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate o($1/k^2$). *Dokl. Akad. Nauk SSSR*, *269*, 543-547. Retrieved from https://ci.nii.ac.jp/naid/10029946121/en/

Nguyen, G., Aunai, N., Michotte de Welle, B., Jeandet, A., & Fontaine, D. (2019). Automatic detection of the earth bow shock and magnetopause from in-situ data with machine learning. *Annales Geophysicae Discussions*, *2019*, 1–22. Retrieved from https://angeo.copernicus.org/preprints/angeo-2019-149/ doi: 10.5194/angeo-2019-149

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton project para*. Cornell Aeronautical Laboratory. Retrieved from https://books.google.fr/books?id=P_XGPgAACAAJ

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, p. 318-362). MIT Press.

Russell, C. T. (1971, January). Geophysical coordinate transformations. *Cosmic Electrodynamics*, *2*, 184-196.

Sandhu, J. K., Yeoman, T. K., Fear, R. C., & Dand ouras, I. (2016, Nov). A statistical study

This article is protected by copyright. All rights reserved.

Accepted Article

of magnetospheric electron density using the Cluster spacecraft. *Journal of Geophysical Research (Space Physics)*, *121*(11), 11,042-11,062. doi: 10.1002/2016JA023397

Serra, J. (1983). *Image analysis and mathematical morphology*. USA: Academic Press, Inc.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929-1958. Retrieved from `http://jmlr.org/papers/v15/srivastava14a.html`

Stone, R. G., Bougeret, J. L., Caldwell, J., Canu, P., de Conchy, Y., Cornilleau-Wehrlin, N., ... Goldstein, M. L. (1992, January). The Unified Radio and Plasma wave investigation. *Astronomy & Astrophysics*, *92*, 291-316.

Tieleman, T., & Hinton, G. (2012). *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude.* COURSERA: Neural Networks for Machine Learning.

Trotignon, J. G., Béghin, C., Lagoutte, D., Michau, J. L., Matsumoto, H., Kojima, H., ... Pottelette, R. (2006, January). Active measurement of the thermal electron density and temperature on the Mercury Magnetospheric Orbiter of the BepiColombo mission. *Advances in Space Research*, *38*, 686-692. doi: 10.1016/j.asr.2006.03.031

Trotignon, J. G., Décréau, P. M. E., Rauch, J. L., Randriamboarison, O., Krasnoselskikh, V., Canu, P., ... Fergeau, P. (2001, Oct). How to determine the thermal electron density and the magnetic field strength from the Cluster/Whisper observations around the Earth. *Annales Geophysicae*, *19*(10), 1711-1720. doi: 10.5194/angeo-19-1711-2001

Trotignon, J. G., Décréau, P. M. E., Rauch, J. L., Vallières, X., Rochel, A., Kougblénou, S., ... Masson, A. (2010, January). The WHISPER Relaxation Sounder and the CLUSTER Active Archive. *Astrophysics and Space Science Proceedings*, *11*, 185-208. doi: 10.1007/978-90-481-3499-1_12

Trotignon, J. G., Etcheto, J., & Thouvenin, J. P. (1986, Apr). Automatic determination of the electron density measured by the relaxation sounder on board ISEE 1. *Journal of Geophysical Research*, *91*(A4), 4302-4320. doi: 10.1029/JA091iA04p04302

Trotignon, J. G., Michau, J. L., Lagoutte, D., Chabassière, M., Chalumeau, G., Colin, F., ... Zamora, P. (2007, February). RPC-MIP: the Mutual Impedance Probe of the Rosetta Plasma Consortium. *Space Science Reviews*, *128*, 713-728. doi: 10.1007/s11214-006-9005-1

Trotignon, J. G., Rauch, J. L., Décréu, P. M. E., Canu, P., & Lemaire, J. (2003, Jan). Ac-

This article is protected by copyright. All rights reserved.

Accepted Article

tive and passive plasma wave investigations in the earth's environment: The cluster/whisper experiment. *Advances in Space Research*, *31*(5), 1449-1454. doi: 10.1016/S0273-1177(02)00959-6

Wattieaux, G., Gilet, N., Henri, P., Vallières, X., & Bucciantini, L. (2019, September). RPC-MIP observations at comet 67P/Churyumov-Gerasimenko explained by a model including a sheath and two populations of electrons. *Astronomy & Astrophysics*, *630*, A41. doi: https://doi.org/10.1051/0004-6361/201834872

Wattieaux, G., Henri, P., Gilet, N., Vallières, X., & Deca, J. (2020). Plasma characterization at comet 67p between 2 and 4 a.u. from the sun with the rpc-mip instrument. *Astronomy and Astrophysics*.

Woolliscroft, L. J. C., Alleyne, H. S. C., Dunford, C. M., Sumner, A., Thompson, J. A., Walker, S. N., ... Gough, M. P. (1997, January). The Digital Wave-Processing Experiment on Cluster. *Space Science Review*, *79*, 209-231. doi: 10.1023/A:1004914211866

Yagi, M., Seki, K., Matsumoto, Y., Delcourt, D. C., & Leblanc, F. (2017, November). Global Structure and Sodium Ion Dynamics in Mercury's Magnetosphere With the Offset Dipole. *Journal of Geophysical Research (Space Physics)*, *122*(11), 10,990-11,002. doi: 10.1002/2017JA024082

Zhelavskaya, I. S., Spasojevic, M., Shprits, Y. Y., & Kurth, W. S. (2016, May). Automated determination of electron density from electric field measurements on the Van Allen Probes spacecraft. *Journal of Geophysical Research (Space Physics)*, *121*(5), 4611-4625. doi: 10.1002/2015JA022132

This article is protected by copyright. All rights reserved.

Accepted Article

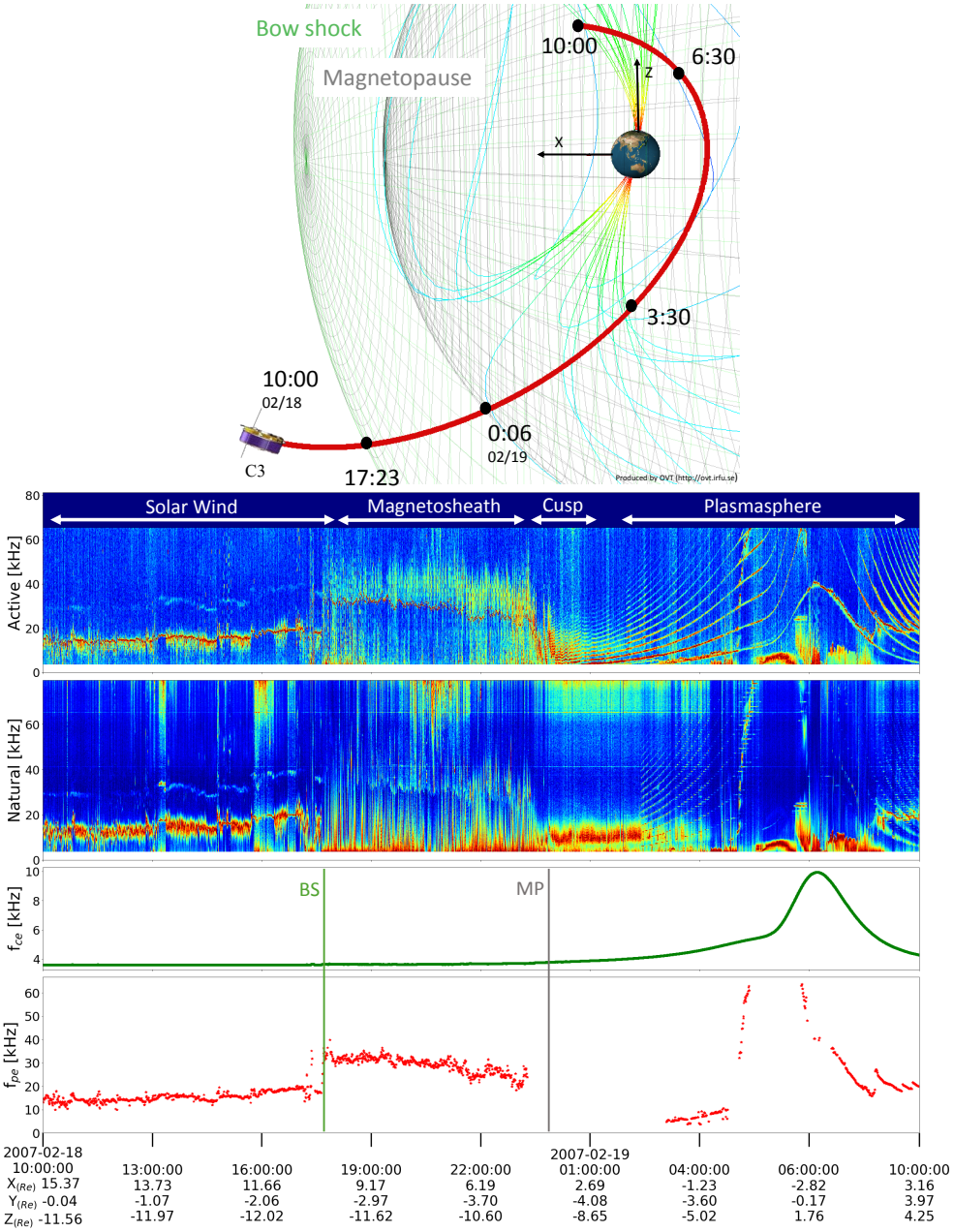


Figure 2. Overview of the WHISPER dataset with parameters of interest measured from 18 Feb 2007 at 09:43 UT to 19 Feb 2007 at 09:54 UT on the C3 (SAMBA) spacecraft. *Upper panel*: orbit of the C3 spacecraft (red line) and field lines in the X-Y GSE coordinate system, normalized to the Earth radius R_e, produced by the Orbit Visualization Tool (OVT) (available at http://ovt.irfu.se/). C3 crossed the bow shock and the magnetopause successively at 17:23 and 00:06 UT. *Second and third panel*: Frequency/time active and natural electric field spectrograms measured by WHISPER, expressed in dB normalized between 0 and 1 (red: maximum amplitude, blue: minimum amplitude). *Fourth panel*: Variation of the electron cyclotron frequency f_{ce}, expressed in kHz, derived by the measurement of the magnetic field amplitude by FGM. *Fifth panel*: Variation of the electron plasma frequency f_{pe}, expressed in kHz, extracted from the WHISPER spectra by *ad hoc* methods (see Sect. 2.3).

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

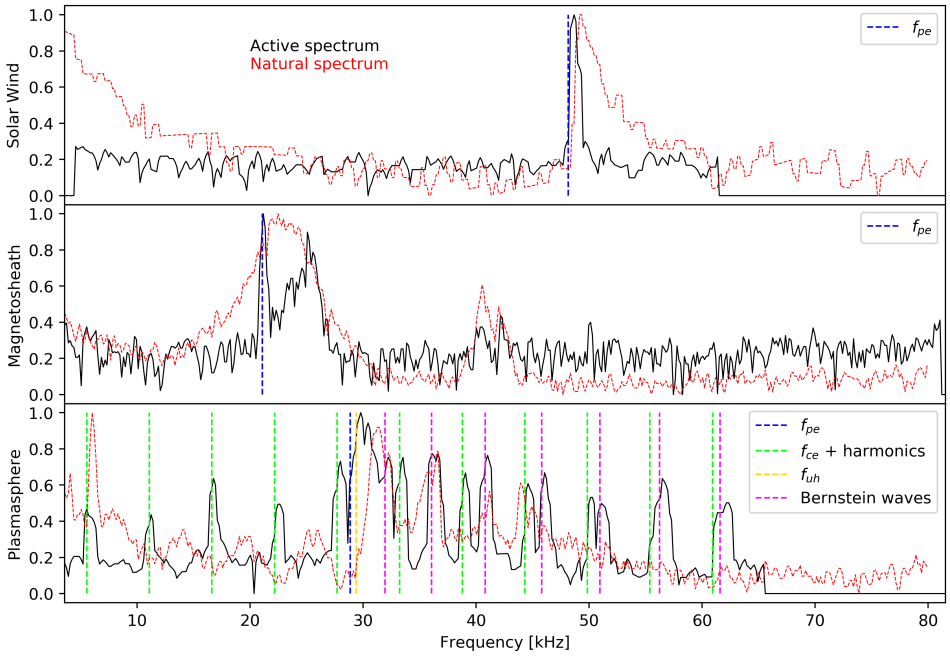


Figure 3. Typical WHISPER individual active (black curve) and natural (red dotted curve) electric field spectra measured in several plasma regimes: in solar wind (first panel), in magnetosheath (second panel) and in the plasmasphere (third panel). The frequency position of the plasma resonances are given by vertical dotted line: the electron plasma frequency f_{pe} (blue), the electron cyclotron frequency f_{ce} and its harmonics (green), the upper hybrid frequency f_{uh} (yellow) and Bernstein frequencies f_{qn} (magenta). All spectra have been expressed in dB and each of them normalized in amplitude between 0 and 1.

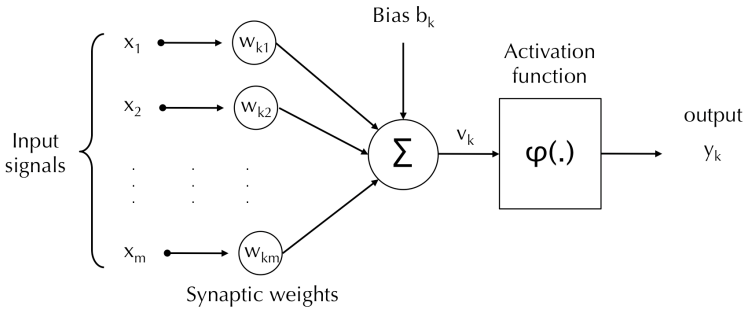


Figure 4. Architecture of an artificial neuron. The neuron takes m input values x_i, multiplies each input by a synaptic weight, $w_{k,i}$, and sums them along with a bias, b_k. An activation function ϕ is then applied to the output which gives the final output y_k of the neuron.

This article is protected by copyright. All rights reserved.

Accepted Article

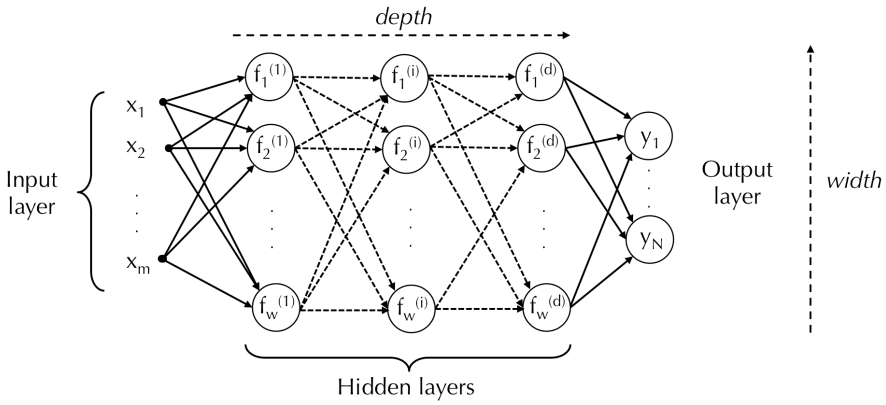


Figure 5. Illustration of the architecture of a multilayer feedforward (dense) neural network composed by several layers: an input layer X, d hidden layers and an output layer y. The neural network is also defined by a combination of activation function $f_i^{(j)}$. The depth of the neural network is defined by the number of hidden layers. The width is defined by the number of neurons on each hidden layer.

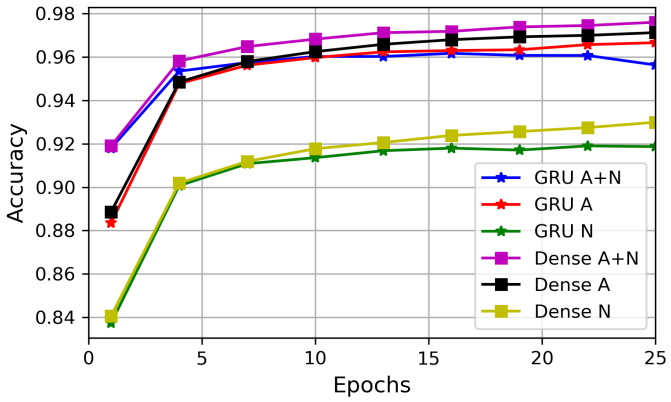


Figure 6. Accuracy of several neural networtk architectures used to predict the plasma region on the training dataset: (i) fully-connected (DENSE) neural network (squared color line) and, (ii) recurrent (GRU) neural network (asterisk color line), in the training dataset composed of (i) only active spectra (A), (ii) active and natural spectra (A+N) and, (iii) only the natural spectra (N).

This article is protected by copyright. All rights reserved.

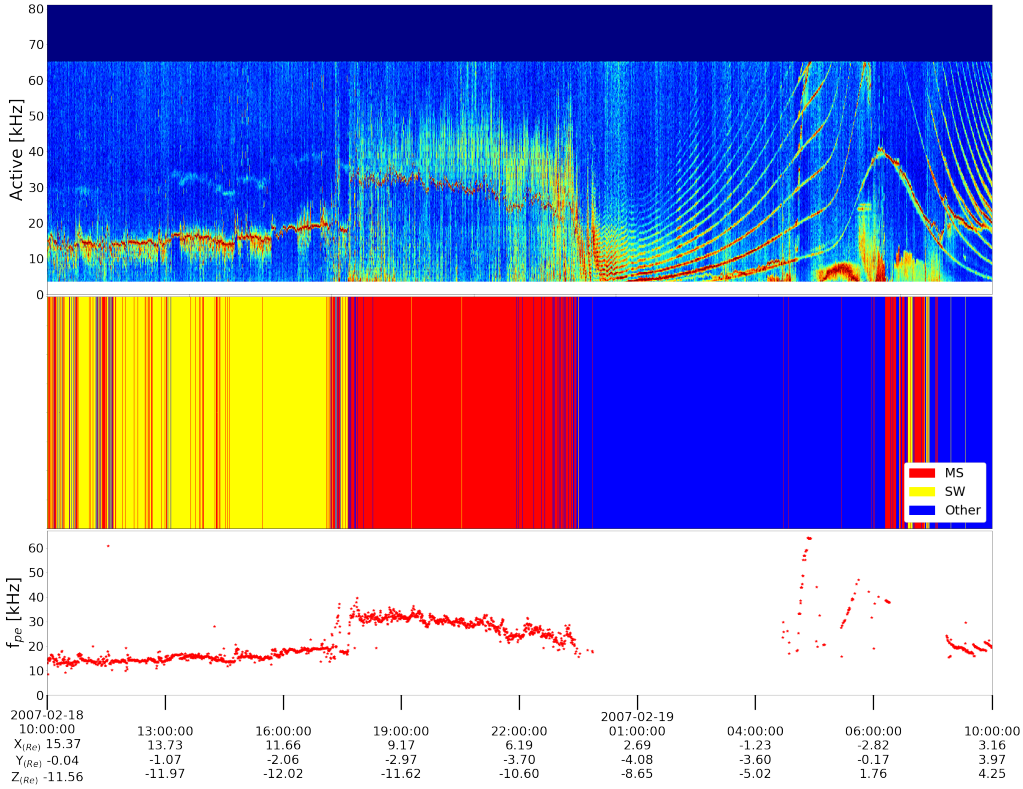


Figure 7. Prediction of the plasma region and electron plasma frequency f_{pe} for the case shown in Fig 2 for the WHISPER measurement on C3 (SAMBA) spacecraft from 18 Feb 2007 at 9:43 UT to 19 Feb 2007 at 09:54 UT. *First panel*: active WHISPER spectrogram expressed in dB. *Second panel*: Predicted plasma regime: Solar wind (SW, yellow), Magnetosheath (MS, red) and Other (blue). *Third panel*: Predicted electron plasma frequency f_{pe} expressed in kHz.

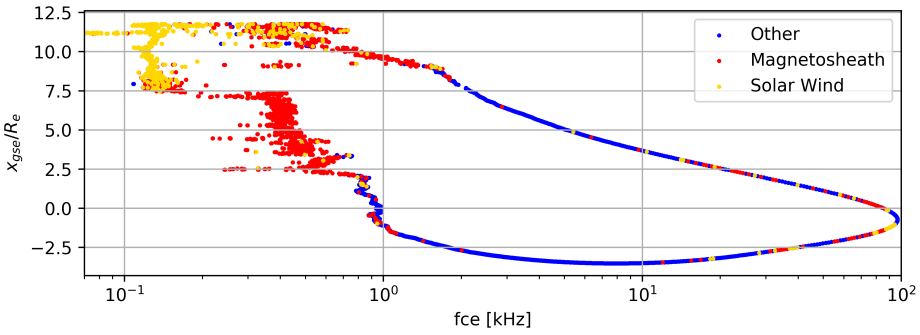


Figure 8. Prediction of the plasma region from 4 Jan 2013 at 00:00 UT to 6 Jan 2013 at 4:00 UT compared to the X_{gse}, normalized by Earth radius and the electron cyclotron frequency f_{ce}, expressed in kHz in logarithmic scale. The solar wind region is indicated by yellow points, the magnetosheath region by red points and the other regions by blue points.

This article is protected by copyright. All rights reserved.

Accepted Article

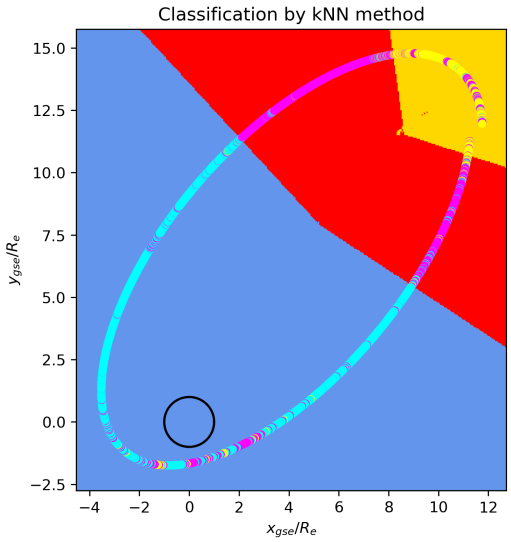


Figure 9. Classification of the plasma region on one complete orbit described by C1 (RUMBA) from 4 Jan 2013 at 00:00 UT to 6 Jan 2013 at 4:00 UT by *k*-nearest neighbors algorithm using the plasma regions predicted by the neural network model, in the XY_{gse} coordinates (normalized by Earth radius). The color background shows the predicted plasma regimes: free solar wind (yellow), magnetosheath (red) and other (blue).

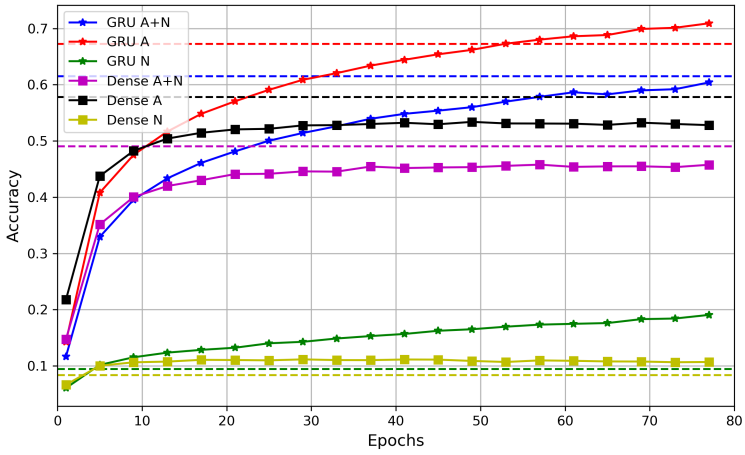


Figure 10. Accuracy of several neural network architectures to predict the thermal electron density on the training dataset: (i) fully-connected (DENSE) neural network (squared color line) and, (ii) recurrent (GRU) neural network (asterisk color line), in the training dataset composed of (i) only active spectra (A), (ii) active and natural spectra (A+N) and, (iii) only natural spectra (N) measured in the solar wind or the magnetosheath regime. Accuracies on the test dataset, at 80 epochs, are given by the horizontal dotted line with the same color code.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

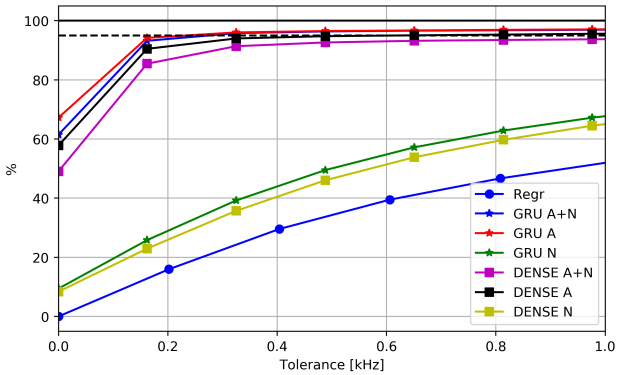


Figure 11. Percentage of well predicted electron plasma frequencies compared to a tolerance threshold, expressed in kHz, for several neural network architectures: recurrent network (GRU), dense (DENSE), or by a regression method. The black vertical dotted line shown the 95% accuracy threshold.

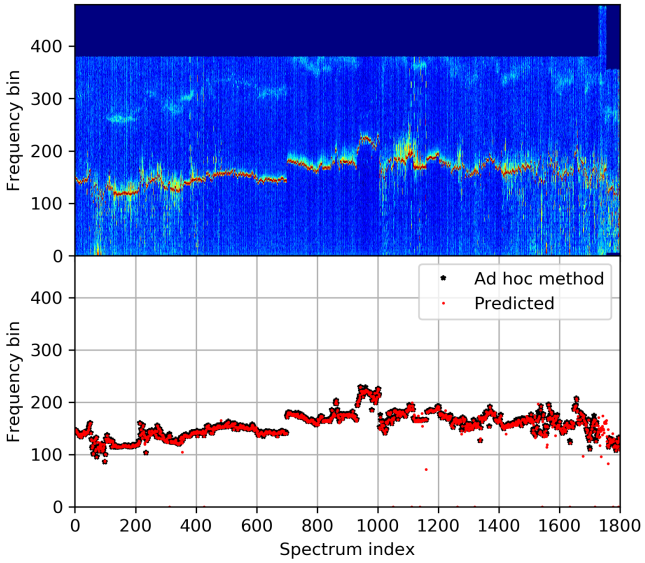


Figure 12. *Upper panel*: Active WHISPER spectrogram measured on C1 (RUMBA) on 13 Feb 2012 between 00:02 UT and 13:21 UT. *Lower panel*: Electron plasma frequency f_{pe}, expressed in frequency bins, given by *ad hoc* methods (black points) and predicted by neural network (red points). Note that the size of red points has been decreased for the sake of clarity.

This article is protected by copyright. All rights reserved.

Accepted Article

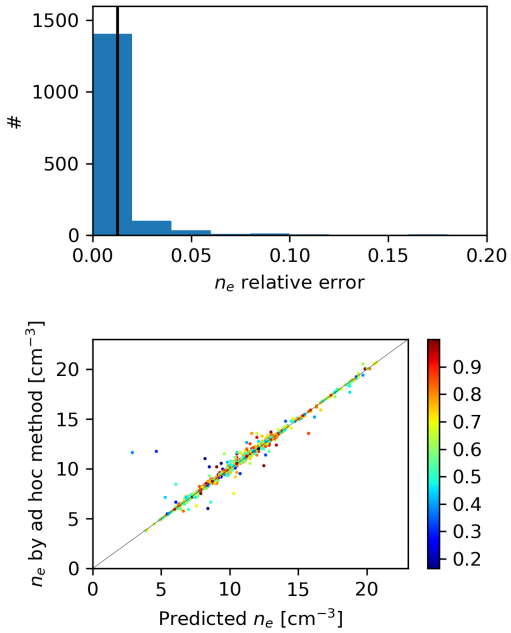


Figure 13. *Upper panel*: Relative error distribution of the predicted thermal electron density on the data shown in Figure 12. The mean relative error is given by the black vertical line. *Lower panel*: Scatter plot of the predicted thermal electron density versus the density given by *ad hoc* methods, expressed in cm^{-3}. The class probability is given by the colorbar (blue: min, red: max).

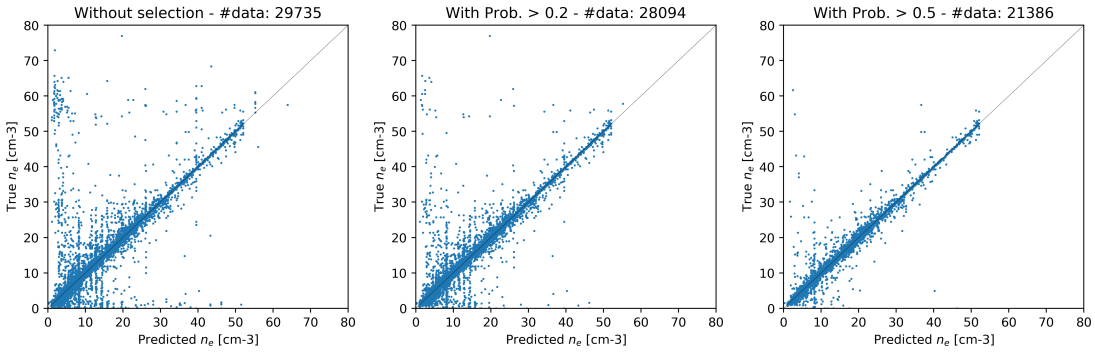


Figure 14. Scatter plots of the thermal electron density showing the predicted density versus the density detected by *ad hoc* methods. Both density are expressed in cm^{-3}. The data corresponds to the WHISPER measurements made onboard C1 spacecraft in 2012 (i) without threshold on the probability (*left panel*), (ii) with a threshold of 0.2 (*middle panel*) and, (iii) with a threshold of 0.5 (*right panel*).

This article is protected by copyright. All rights reserved.

Accepted Article

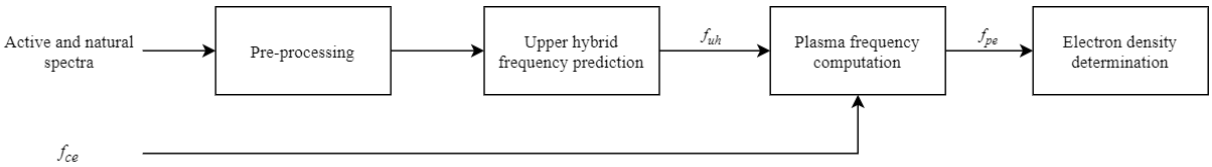


Figure 15. Illustration of the automatic pipeline used to predict the thermal electron density in the plasmasphere regime.

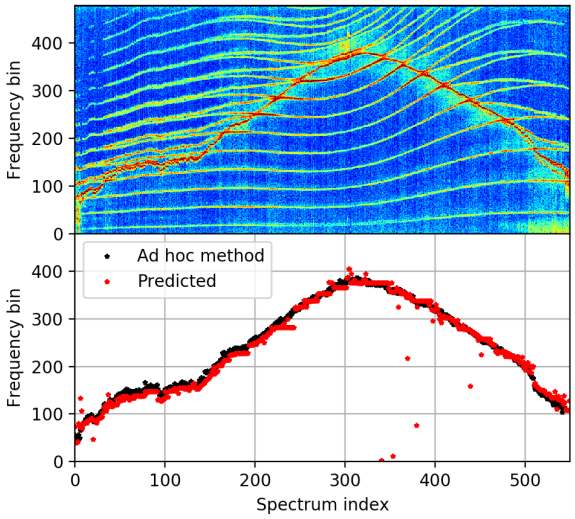


Figure 16. *Upper panel*: Active WHISPER spectrogram measured on C1 on 5 Aug 2002 between 17:00 and 20:40 UT. *Lower panel*: Electron plasma frequency f_{pe}, expressed in frequency bins, given by *ad hoc* methods (black points) and predicted by the neural network (red points).

This article is protected by copyright. All rights reserved.

Accepted Article

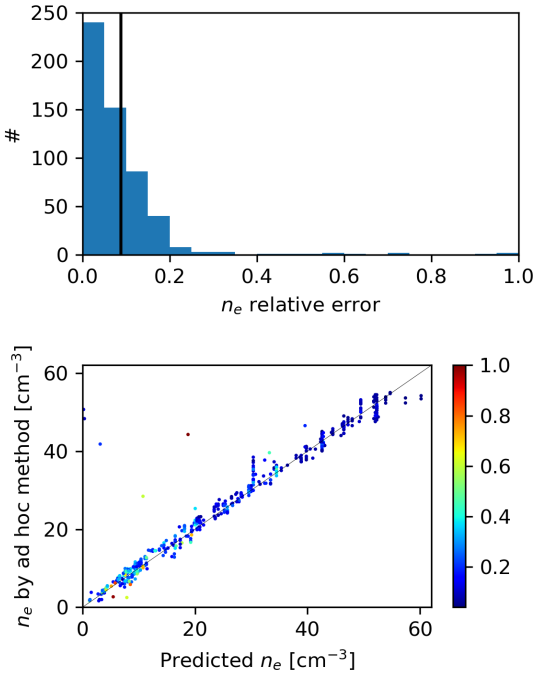


Figure 17. *Upper panel*: Relative error distribution of the predicted thermal electron density on the data shown in Figure 16. The mean relative error is given by the black vertical line. *Lower panel*: Scatter plot of the predicted thermal electron density versus the density given by *ad hoc* methods, expressed in cm^{-3}. The class probability is given by the colorbar (blue: 0, red: 1).

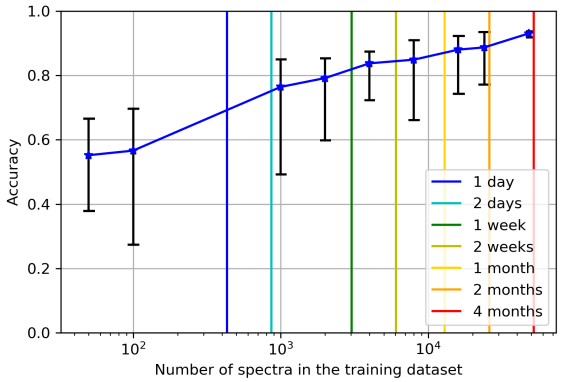


Figure 18. Accuracy score of an automatic plasma region model compared to the amount of spectra in the training dataset in the AM^{2}P experiment context. The error bar is given in black vertical bar. The time period corresponding to the AM^{2}P measurement rate is given by verticals bars from 1 day (blue vertical bar) to 4 months (red vertical bar).

This article is protected by copyright. All rights reserved.