J. Space Weather Space Clim. 2022, **12**, 14

https://doi.org/10.1051/swsc/2022009

JSWSC

Available online at:
www.swsc-journal.org

Topical Issue - Ionospheric plasma irregularities and their impact on radio systems

RESEARCH ARTICLE OPEN ACCESS

Semi-supervised classification of lower-ionospheric perturbations using GNSS radio occultation observations from Spire Global's Cubesat Constellation

Giorgio Savastano[1,*], Karl Nordström[2], and Matthew J. Angling[2]

[1] Spire Global, 33 rue Sainte Zithe, 2763 Luxembourg, Luxembourg
[2] Spire Global, Skypark 6, 64 – 72 Finnieston Square, Glasgow, G3 8ET, UK

Received 9 November 2021 / Accepted 24 March 2022

Abstract – This study presents a new methodology to automatically classify perturbations in the lower ionosphere using GNSS radio occultation (RO) observations collected using Spire's constellation of CubeSats. This methodology combines signal processing techniques with semi-supervised machine learning by applying spectral clustering in a metric space of wavelet spectra. A "bottom-up" algorithm was applied to extract E layer information directly from Spire's high-rate (50 Hz) GNSS-RO profiles by subtracting the effect of the F layers. This processing algorithm has been implemented in our ground segment to operationally produce high rate sTEC profiles with a vertical resolution of better than 100 m. The key idea behind the semi-supervised classification is to produce a database of labeled clusters that can be used to classify new unlabeled data by determining which cluster it belongs to. A dataset of more than 12,000 GNSS-RO profiles collected in 2019 containing sTEC perturbations is used to find the initial clusters. This dataset is used to represent the climatology of ionospheric perturbations, such as MSTIDs and sporadic Es. The wavelet power spectrum (WPS) is computed for these profiles, and a metric space is defined using the Earth mover's distance (EMD) between the WPS. A self-tuning spectral clustering algorithm is used to cluster the profiles in this metric space. These clusters are used as a reference database of perturbations to classify new sTEC profiles by finding the cluster of the closest profile of the clustered dataset in the EMD metric space. This new methodology is used to construct an automated system to monitor ionospheric perturbations on a global scale.

Keywords: ionosphere / GNSS-RO / perturbations / classification / cubesats

1 Introduction

The ionosphere remains an important region of the atmosphere for study due to its potential impact on radio systems (Angling et al., 2012) and its coupling to the wider geospace environment (Schunk & Nagy, 2009). The ionosphere is strongly coupled to the magnetosphere and solar wind as well as to the lower atmosphere. In the case of the latter, the ionosphere can be affected by events such as extreme terrestrial weather (Chou et al., 2017), earthquakes and tsunamis (Galvan et al., 2011; Savastano et al., 2017; Astafyeva, 2019), explosions (Jacobson et al., 1988; Huang et al., 2019) and rocket launches (Booker, 1961; Mendillo et al., 1975; Savastano et al., 2019).

This paper focuses on detecting and classifying perturbations in the low altitude ionosphere, i.e., between 80 and 150 km. We can define such perturbations as deviations from the median ionosphere that have been produced by a geophysical or anthropogenic driver. These perturbations can be generated by multiple mechanisms such as plasma instabilities (Fejer & Kelley, 1980), sporadic E (Es) (Haldoupis, 2011), and traveling ionospheric disturbances (TIDs) (Yeh & Liu, 1974).

Sporadic E (Es) layers are thin (0.5–5 km) layers of enhanced electron density occurring at E region heights (90–120 km). Its occurrence is strongly latitude and season-dependent (Whitehead, 1970, 1989); Es appears mainly in the daytime at midlatitudes in the summer hemisphere, while Es rates are generally low in winter. The horizontal extent of Es can range from 10 to 1000 km (Wu, 2005), and layers may last from minutes to hours. At midlatitudes, sporadic E formation is a complex process of interaction between wind shears in the lower thermosphere and long-lived metallic ions originating from meteors (Arras, 2010; Haldoupis, 2011).

*Corresponding author: giorgio.savastano@spire.com

TIDs are a type of ionospheric perturbations consisting of wave-like fluctuations of the electron density in the ionosphere. These disturbances are the ionospheric manifestation of atmospheric gravity waves that naturally occur at many different scales (Georges, 1967; Georges & Hooke, 1970). Based on their phase velocity, wave period, and horizontal wavelength, TIDs are often classified into medium-scale TID (MSTID) and large scale TID (LSTID) (Ogawa et al., 1987; Crowley & Rodrigues, 2012).

Es and TID climatologies have been well characterized through studies using ionosondes and other ground instruments over many years. The first global maps of sporadic E occurrence were produced during the International Geophysical Year in 1957–58 (Leighton et al., 1962). However, such measurement campaigns are limited in their geographic coverage and may also be limited in their vertical resolution. The high-resolution vertical structure of Es has been studied using sounding rockets (Aubry et al., 1966; MacKenzie & Sayers, 1966; Wakabayashi et al., 2005). However, such measurements are even more geographically sparse.

Beyond scientific studies, real-time monitoring of TIDs and Es layers is of great interest to many terrestrial applications, such as radio communications, global navigation satellite system (GNSS) users, space weather data assimilation models, and natural hazards warning systems (Savastano et al., 2017). However, this remains a challenging goal because, on a global scale, there are large parts of the ionosphere that are not surveyed by any ground-based instrumentation. One way to overcome this is to use space-based radio occultation (RO) measurements to sense the ionosphere. Such measurements are made by estimating the total electron content (TEC) along the line of sight between the GNSS transmitter in medium Earth orbit (MEO) and a receiver in low Earth orbit (LEO) (Hajj & Romans, 1998; Hajj et al., 2002). GNSS-RO studies have been used to study the climatology of Es (Wu et al., 2005; Arras, 2010; Tsai et al., 2018). Using high rate data, RO profiles can capture very fine structures in the E region of the ionosphere (Wu, 2018).

The use of GNSS-RO for routine perturbation monitoring presents its own difficulties. For example, the Spire RO CubeSat constellation currently collects in excess of 15k GNSS-RO profiles per day. It is not possible for each of these to be manually examined and classified with the type of perturbation (or none) that it may contain. Therefore, this paper aims to describe how machine learning methods may be used to group similar profiles into clusters and how these may then be used to classify new profiles in real-time.

2 Materials and methods

2.1 Data collection system

Spire operates a rapidly growing constellation of more than 100, 3U CubeSats known as LEMURs (low-Earth multi-use receivers). Each LEMUR carries a software-defined GNSS-RO receiver (STRATOS) that is capable of making open-loop, dual-frequency, multi-constellation (GPS, GLONASS, QZSS, and Galileo) RO measurements. The LEMUR CubeSats fly in a range of LEO orbits and at altitudes between 400 km and 600 km.

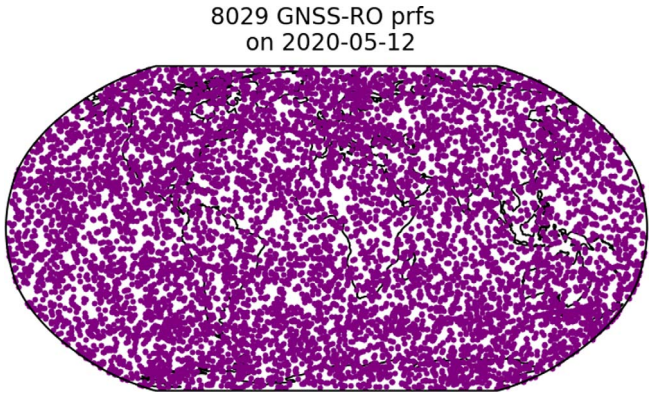

Fig. 1. Coverage map showing 8029 Spire GNSS-RO profiles collected on 12 May 2020.

Presently, Spire processes over 15k GNSS-RO profiles each day, with the expectation of further growth due to new satellite launches and an increased duty cycle of the existing fleet. Figure 1 shows an example of the coverage of Spire GNSS-RO profiles collected on 12 May 2020. A description of the measurement system can be found (Angling et al., 2021).

Space-based GNSS-RO sTEC measurements are collected in an atmospheric limb sounding geometry where the GNSS receiver (Rx) is on an LEO satellite (Angling et al., 2021). As the LEO satellite moves in its orbit, the GNSS satellite (Tx) is seen to rise above or set below the horizon. The ray path from the Tx to Rx is quasi horizontal and can be characterized by the height of its tangent point. Thus, at each time epoch, the sTEC measurement made along each ray path can be associated with the tangent height in order to construct an sTEC profile.

2.2 Dataset

Spire's GNSS-RO level-1b atmPhs files follow the COSMIC (Constellation Observing System for Meteorology, Ionosphere, and Climate) standard file format defined by UCAR (University Corporation for Atmospheric Research). These files contain profiles of signal to noise (SNR), L1 and L2 excess phase measurements sampled at a rate of 50 Hz. In determining the excess phase, the error terms that are common to both frequencies (i.e., errors in the precise orbit determination (POD) and both transmitter and receiver clock noise) have been removed by the standard GNSS-RO processing system. Thus, a single frequency excess phase measurement can be used to estimate relative TEC without forming the L1–L2 linear combination. This is desirable since forming the combination introduces an error term due to the fact the L1 and L2 signals experience different bending and thus travel over slightly different ray paths. Although this is often neglected for space to ground paths, for RO, the path separation can be 100s of meters (Svehla, 2018); this is greater than the vertical resolution that is achieved for the E region measurements and cannot be ignored. Furthermore, the L1 signal has a higher signal-to-noise ratio (SNR) than the L2 signal and therefore provides the best phase measurement among the available GNSS-RO data. For these reasons, we used the bottom-up approach described in (Wu, 2018) in order to estimate high-rate (50 Hz sampling) relative

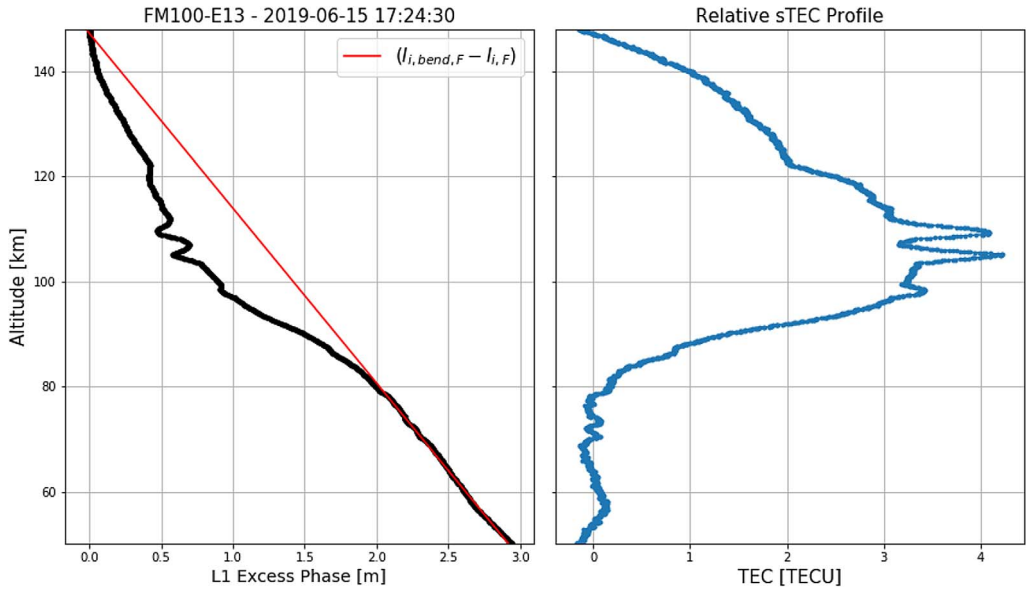

Fig. 2. Bottom-up processing applied to Spire high-rate (50 Hz) excess phase data in order to estimate relative sTEC profiles. The left panel shows the L1 excess phase profile (black) for Spire satellite FM100 from Galileo E13 and the linear contribution of the F region estimated from the profile below 80 km. The right panel shows the relative sTEC profile that is obtained after removal of the linear trend.

sTEC profiles in the E region from L1-only excess phase observations. The main assumption of the algorithm is that the contribution of the F region to the excess phase observations varies linearly with height when the tangent point is below 80 km altitude. The basis for the linear assumption is described in (Wu, 2018) Section 3, and the simulation validation results are given in (Wu, 2018) Appendix C. Figure 2 shows the bottom-up approach applied to high-rate 50 Hz excess phase data in order to estimate relative sTEC profiles. There are many advantages in analyzing the sTEC measurements instead of inverted quantities (such as total electron content). The main one is that traditional retrieval processes (e.g., the Abel inversion) often assume spherical symmetry (Hajj & Romans, 1998), which may not be valid for ionospheric perturbations, and make the retrieved quantities more difficult to interpret.

Limitations of the linear assumption in the presence of strong TEC gradients or ionospheric anisotropy in the F region may lead to unphysical relative TEC profiles (see top right panel of Fig. 3). However, since the proposed semi-supervised classification algorithm is based on cropped wavelet power spectra that do not include very long wavelengths (see Sect. 2.3.1), the classification results are not affected. The main reason for implementing the bottom-up algorithm is to create E-region perturbation profiles more easily physically interpretable. This is an important step for any semi-supervised machine learning algorithm which relies on labels created by scientists during the training phase (see Sect. 3).

GNSS-RO sTEC profiles are nearly instantaneous snapshots of the E region of the ionosphere with a typical measurement time of 1–2 min. This is much shorter than the typical timescales of atmospheric and ionospheric processes that may cause perturbations (Alexander et al., 2008). Therefore, we can treat the ionosphere as fixed in time during the measurement and perform an analysis of vertical structures while ignoring the temporal variability.

The 50 Hz sampling of the relative sTEC profile corresponds to a vertical resolution of around 100 m and thereby allows the detection of fine vertical structures in the E region ionosphere. An added benefit is that only single-frequency observations are needed for this technique.

In order to represent the climatology of ionospheric perturbations, such as MSTIDs (Duly et al., 2013) and sporadic E (Wu et al., 2005), we randomly selected 100 profiles per day for the whole year 2019. We implemented two selection criteria to reduce the number of corrupted profiles in our dataset. The first criterium is that the minimum of the signal to noise ratio (SNR) along the profile must be greater than 15 dB to avoid profiles affected by possible changes in the satellite attitude or by radio-frequency interference (RFI). The second criterium deals with the breakdown of the linear assumption of the bottom-up approach, which is avoided by computing the standard deviation of the sTEC profile below 70 km altitude and disregarding profiles having a standard deviation greater than 1 TECU. In our study, the number of GNSS-RO profiles where the linear assumption broke down was less than 1% of the total.

In order to select only profiles containing ionospheric perturbations, we analyzed the wavelet power spectrum (WPS) of each profile (see details in Sect. 2.3), and we selected only profiles meeting the criteria:

$$\begin{cases} \max(\mathrm{WPS}) > 0.20\,\mathrm{TECU}^2 \\ \mathrm{sum}(\mathrm{WPS}) > 15\,\mathrm{TECU}^2 \end{cases}$$

where the maximum and the sum of the WPS allows the selection of profiles with either localized or extended power intensity in the wavelet spectra.

Figure 3 displays the final dataset of perturbed sTEC profiles used as input for the classification pipeline. The dataset contains more than 12,000 sTEC profiles, with a homogeneous geographic distribution in latitude and longitude. However, it is

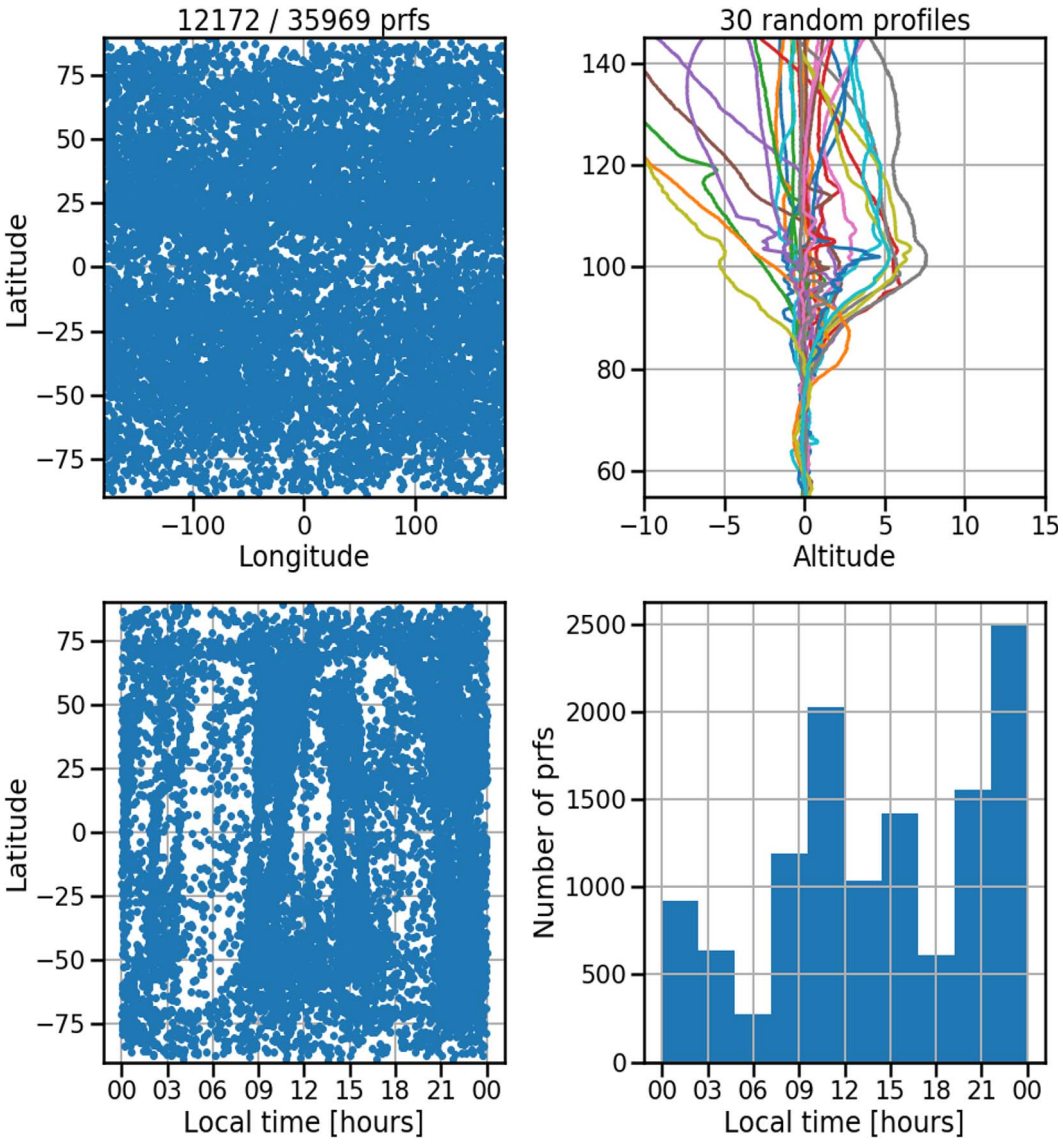

Fig. 3. Dataset of 12,172 perturbed sTEC profiles. Upper left: location of all the sTEC profiles in latitude vs. longitude. Lower left: location of all the sTEC profiles in latitude vs. local time. Upper right: 30 randomly selected profiles. Lower right: local time histograms of sTEC profiles.

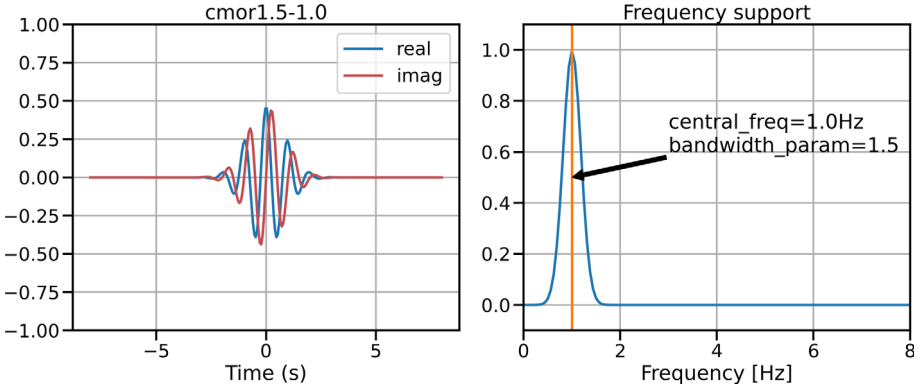

Fig. 4. Complex Morlet wavelet function used in this pipeline and its frequency support.

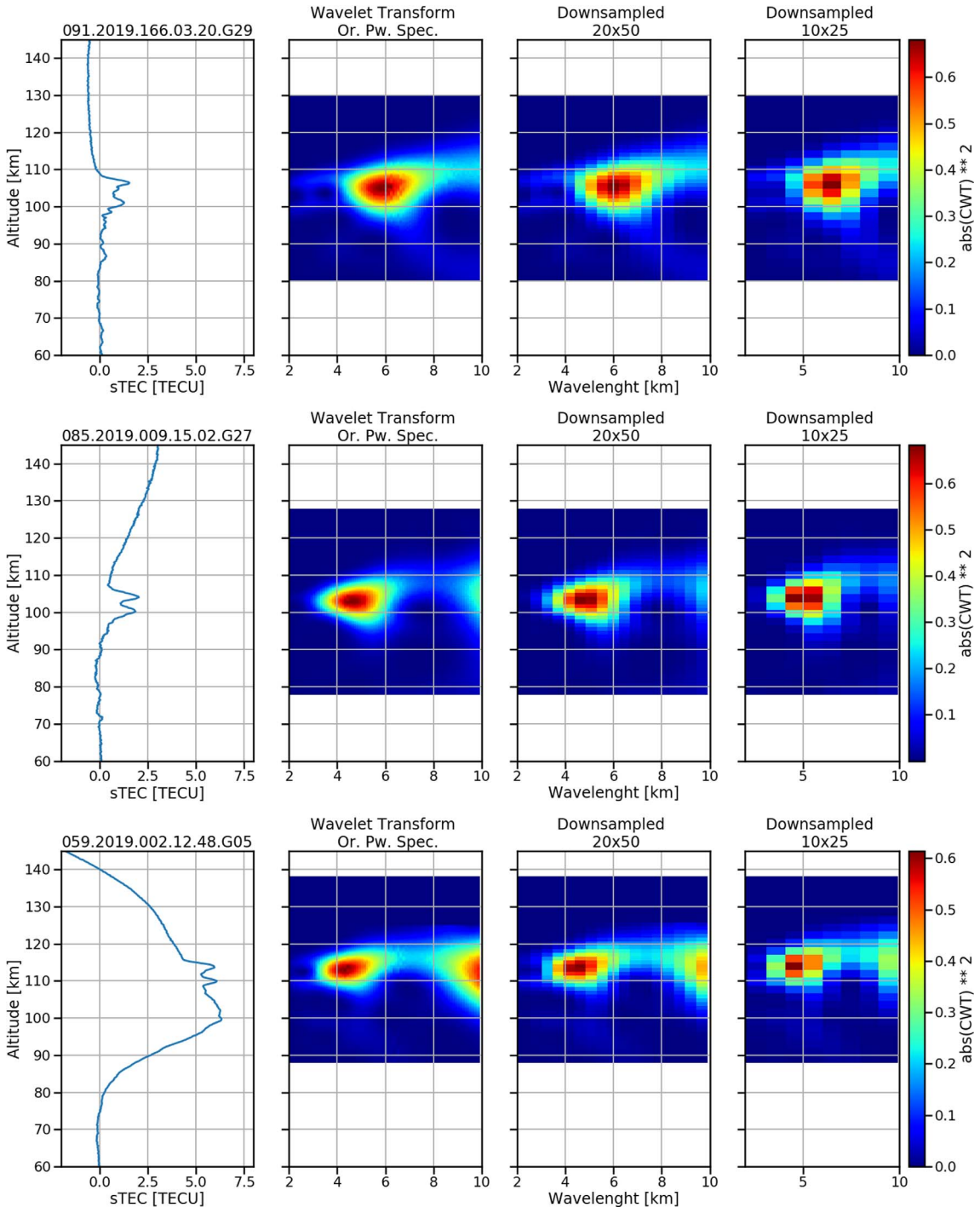

Fig. 5. Centered and downsampled wavelet spectra. (From the left) first column shows three sTEC profiles, second column shows the original wavelet power spectra, third and fourth columns show the downsampled spectra to 20 × 50 and 10 × 25 pixels, respectively.

important to highlight that the profiles do not have a uniform distribution in local time; most of the perturbed profiles have a local time around local noon and after sunset. This is a combination of two effects: the climatology of ionospheric perturbations and Spire's satellites orbits, which are mostly sun-synchronous (i.e., producing observations at a fixed local time).

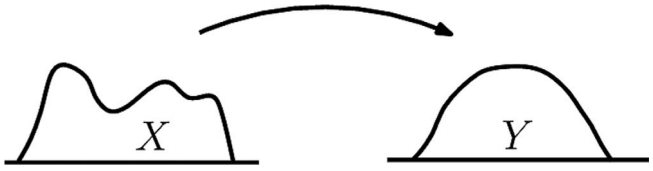

Fig. 6. Concept of the EMD between two probability distributions.

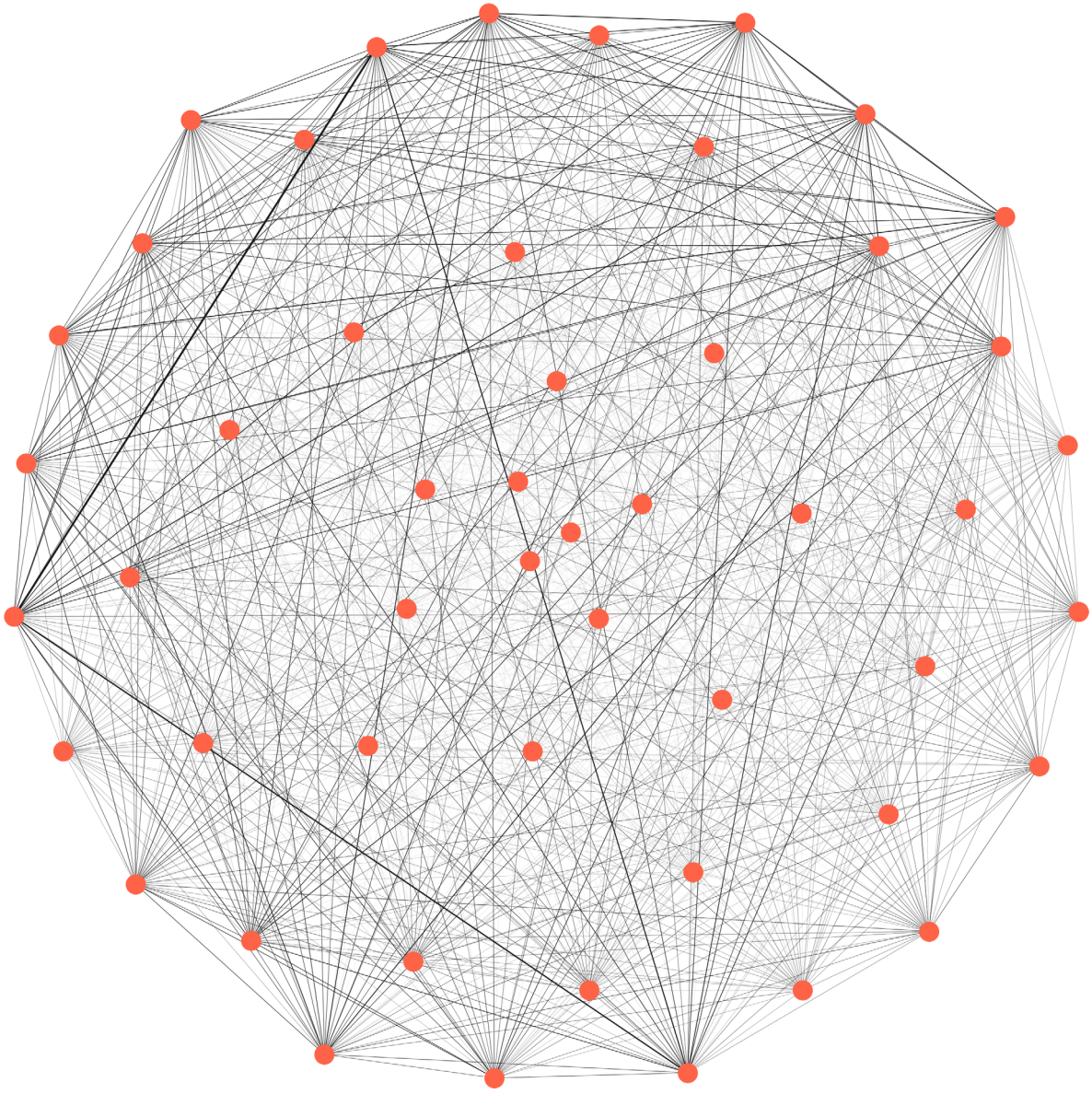

Fig. 7. A partial representation (i.e., 50 nodes) of the fully connected graph obtained from our complete dataset (i.e., 12,172 nodes). Each node (red dot) represents an sTEC profile, and the edges of different widths represent the similarity between nodes, defined by the EMD distance.

2.3 ML-based classification pipeline

In this section, we present the pipeline used to classify perturbations in the lower ionosphere. This pipeline is based on a new methodology that combines signal processing techniques, such as wavelet analysis, with unsupervised machine learning clustering algorithms, such as spectral clustering.

2.3.1 Wavelet transformation

The first step of this pipeline is to resample the sTEC profiles into a regular vertical grid that extends from 50 to 145 km. This step allows the computation of physically meaningful wavelet spectra of the profiles. We use the *scipy.interpolate.interp1d* function, which implements several interpolation algorithms, such as nearest neighbor search (NNS) and spline functions, effectively enabling the downsampling of profiles to a 200 m vertical resolution.

After resampling the profiles, we compute the wavelet spectrum for each profile using the *PyWavelets* python package (Lee et al., 2019). We store each spectrum into a $(n \times n_x \times n_y)$ matrix, where n is the number of profiles and n_x and n_y are the number of columns and rows of the 2-D spectrum. The wavelet spectrum is cropped in wavelength from 1 to 10 km and in altitude from 70 to 125 km in order to avoid edge effects.

The wavelet spectra are computed using a complex Morlet wavelet, which provides a good compromise between scales and wavelength resolution, and it is defined by:

$$\Psi(x) = \frac{1}{\sqrt{\pi f_b}} e^{2i\pi f_c x} e^{x^2 e^{f_b}}$$

where f_b and f_c represent the bandwidth parameter and the wavelet center frequency, respectively. In our pipeline, the wavelet center frequency was set to 1.0 Hz and the bandwidth parameter to 1.5 (Fig. 4).

In order to cluster together perturbations occurring at different altitudes, we center the altitude of the spectrum around the maximum of the spectrum amplitude with an altitude window of ±25 km. In this way, we are unaffected by the absolute altitude values. For sTEC profiles where the maximum spectrum amplitude occurs at a distance <25 km from the edge, we zero-pad to fill the missing values in the wavelet spectrum space.

The next step is downsampling the wavelet power spectra. This step aims to speed up the clustering computations by reducing the number of pixels per spectra without losing important geophysical information. This is achieved by resizing the spectra using the INTER_AREA algorithm of the *OpenCV* python library.

Figure 5 shows the centered and downsampled wavelet power spectra for three sTEC profiles. The first column shows the perturbed sTEC profiles, the second column shows the original wavelet power spectra, and the third and fourth columns display the downsampled spectra to 20 × 50 and 10 × 25 pixels, respectively. In our pipeline, we have used the spectra downsampled to 20 × 50.

2.3.2 Computing similarity between profiles

The earth mover's distance (EMD) is used to define a distance metric in the space of RO wavelet spectra. EMD is a method to compute the distance between two multi-dimensional probability distributions in some feature space D. The concept of using the EMD to measure perceptual similarity between gray-scale images was first explored by (Peleg et al., 1989). More recently, EMD has been utilized for color – or texture-based similarity (Greenspan et al., 2000; Rubner et al., 2000). EMD-based similarity analysis (EMDSA) is now an effective tool used in many pattern recognition (Grauman & Darrell, 2004) applications and has recently been applied to unsupervised learning problems in the physical sciences (Komiske et al., 2019).

The key concept is that the "distance" between two distributions is defined as the minimal amount of work that must be performed to transform one distribution into the other. If we consider the distributions as two different ways of piling up a certain amount of earth (hence "earth mover's distance") over the region D, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of earth moved times the distance by which it is moved (Fig. 6).

The EMD is notoriously expensive to compute for distributions in dimensions higher than one. As we will ultimately want to operationally calculate the distance to all profiles in the reference dataset for each new profile collected by the Spire constellation, it is not feasible to rely on a full EMD calculation for the two-dimensional wavelet spectra. By converting the wavelet spectra into an integer-weight point cloud representation, a proxy for the full EMD can be efficiently computed. In this representation, an EMD calculation can be cast as a transportation problem and solved with a minimum cost flow algorithm (Villani, 2003; Santambrogio, 2009; Korte & Vygen, 2012; Li et al., 2018). The point cloud representation is defined by considering each row along the altitude dimension as a point in an N-dimensional space (where N is given by the wavelet spectrum resolution along the wavelength dimension). Our EMD proxy is then computed using the Euclidean distance between the points to define a point cloud EMD calculation that can be efficiently solved with *scipy.optimize.linear_sum_assignment*.

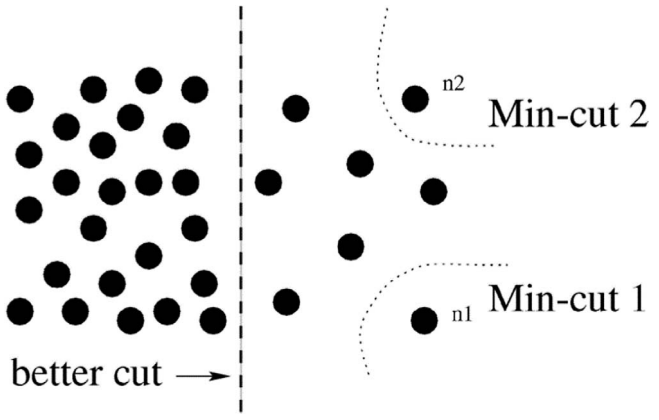

Fig. 8. A case where minimum cut gives a bad partition. Figure from (Shi & Malik, 1997).

Since we have already centered and cropped around the largest perturbation in the wavelet spectrum, the loss of altitude information does not impact the results significantly: on the reference dataset, both the Pearson and Spearman r correlation coefficients between the full EMD calculation and our point cloud EMD proxy is >0.99. This indicates that it is a good proxy and can be calculated more than 3 orders of magnitude faster than when using *pyemd.emd* on a modern laptop. The proxy calculation is inspired by more general approaches to efficiently approximate the EMD in higher dimensions, such as the sliced Wasserstein (Bonneel et al., 2015; Kolouri et al., 2019); however, we rely on our pre-processing pipeline rather than Monte Carlo sampling to reduce the complexity of the calculation without losing necessary information. The final distance matrix has the shape of (n × n), where n is the number of profiles.

2.3.3 Clustering profiles

Spectral clustering is a technique with roots in spectral graph theory (Chung, 1997). The core idea is to perform a spectral analysis of a graph's Laplacian matrix to find clusters in the graph.

We can use a graph $G = (V, E)$ as an abstraction of our dataset, where the nodes V represent the wavelet spectra of the sTEC profiles and the width of the edges E connecting the nodes represents the weight of each edge $w(i; j)$, which in turn is a function of the similarity between nodes i and j. Figure 7 shows an example of a fully connected graph. Clustering a graph means partitioning the set of nodes V into disjoint sets V_1, V_2, . . . , V_n, where the similarity among nodes in a set V_i is high and across different sets V_i, V_j is low.

For example, a graph $G = (V, E)$ can be partitioned into two disjoint sets A and B, $A \cup B = V$, $A \cap B = \emptyset$, by removing those

Table 1. Ionospheric perturbations classes.

Classes	Sub-classes		
1. Sporadic E (Es)	1a. Large	1b. Medium	1c. Small
2. Traveling ionospheric disturbances (TIDs)	2a. Large	2b. Medium	2c. Small
3. Ionosphere irregularities	3a. Large	3b. Medium	3c. Small
4. Smooth profiles	4a. Large	4b. Medium	4c. Small

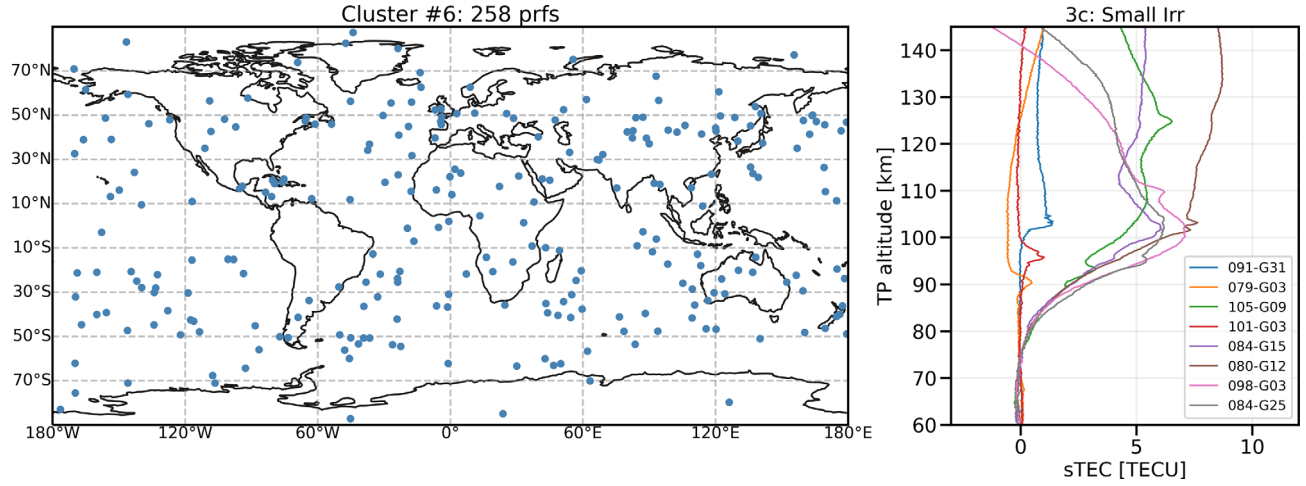

Fig. 9. Ionospheric perturbations cluster example. The left map shows the location of the 258 sTEC profiles within the cluster. On the right, 8 randomly selected profiles from the cluster. All the profiles show similar small-scale sTEC ionospheric perturbations.

edges that connect the two clusters. The degree of dissimilarity between clusters can be computed as the total weight of the edges that have been removed. In graph theoretic language, this is called the *cut* and is defined as:

$$\mathrm{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v).$$

The optimal bi-partitioning of a graph is the one that minimizes this cut value. Although the number of potential partitions increases exponentially with the number of nodes, finding the minimum cut of a graph is a well-studied problem, and there exist efficient algorithms for solving it (Wu & Leahy, 1993).

The minimum cut criterium favors cutting small sets of isolated nodes in the graph. This is not surprising since the *cut* value increases with the number of edges crossing the cut. Figure 8 illustrates this case; assuming the edge weights are inversely proportional to the distance between any two nodes, we see that the cut that partitions out node n_1 or n_2 will have a very small value. In fact, any cut that partitions out individual nodes on the right half will have a smaller cut value than the cut that partitions the nodes into the left and right halves.

To avoid this problem of clustering small sets of isolated nodes in the graph, a disassociation measure known as *normalized cut* (*N*cut) was defined in (Shi & Malik, 1997) as:

$$N\mathrm{cut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(A, V)} + \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(B, V)}$$

where assoc(A, V) is the total connection from nodes in A to all nodes in the graph, and assoc(B, V) is similarly defined. With this definition of the disassociation between the groups, the *cut* that partitions out small, isolated points will no longer have a small *N*cut value since it will almost certainly be a large percentage of the total connection from that small set to all other nodes. For example, in the case illustrated in Figure 8, we see that the cut_1 value across node n_1 will be 100% of the total connection from that node.

The graph's Laplacian matrix L can be written as:

$$L = I^T I = D - W$$

where I is the incidence matrix of the graph, D is the degree matrix, and W is the adjacency matrix. Many methods, such as the normalized cuts algorithm (Shi & Malik, 1997), can be used to find k clusters in the graph using the first k eigenvectors of the graph's Laplacian matrix (Von Luxburg, 2007).

When working with data that is not directly available in the form of a graph, the first step is to construct an affinity matrix $A \in R^{n \times n}$, where n is the number of nodes (i.e., profiles in our case). An affinity matrix is just like an adjacency matrix, except the entries express how similar nodes are to each other. We can use the EMD distance matrix as an affinity matrix, in which a value of zero means identical elements, and high values mean very dissimilar elements. This EMD distance matrix can be transformed into an affinity matrix that is well suited for the *N*cut algorithm by applying a Gaussian kernel (Ng et al., 2002):

$$\begin{cases} A_{ij} = \exp\left(-\frac{d^2(p_i, p_j)}{\sigma^2}\right) & \text{for } i \neq j \\ A_{ij} = 0 & \text{for } i = j \end{cases}$$

where the scaling parameter σ is a free parameter representing the width of the Gaussian kernel and controls how rapidly the

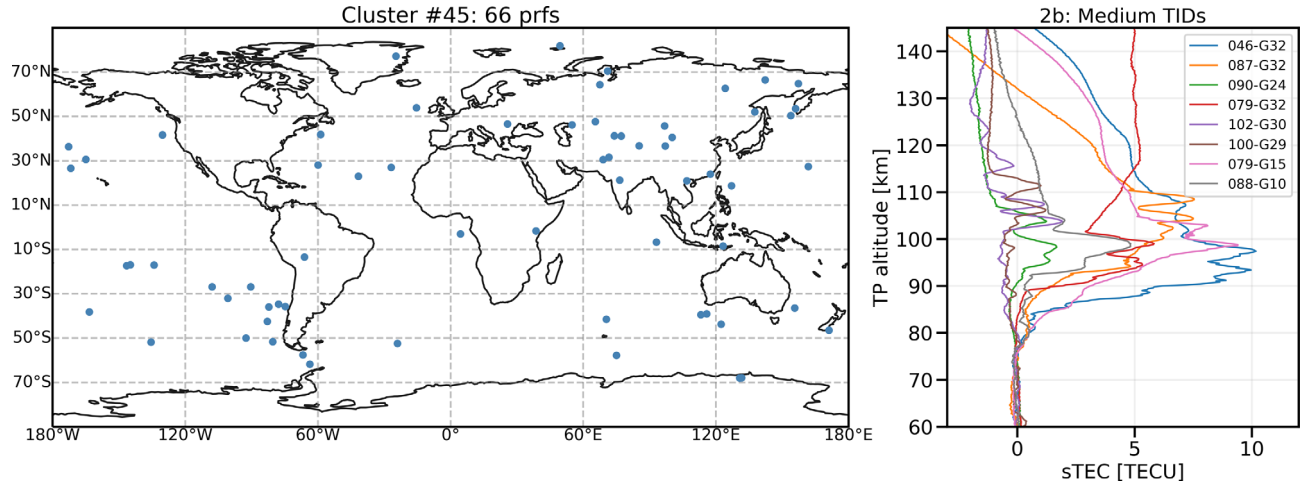

Fig. 10. Ionospheric perturbations example cluster. The left map shows the location of the 66 sTEC profiles within the cluster. On the right, 8 randomly selected profiles from the cluster. All the profiles show similar medium-scale TIDs structures.

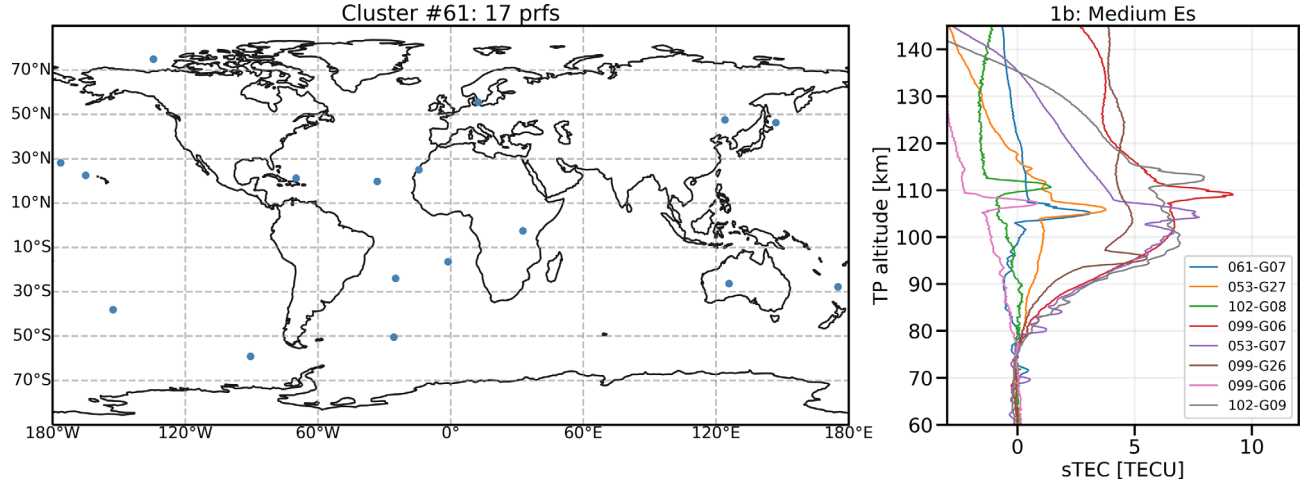

Fig. 11. Ionospheric perturbations example cluster. The left map shows the location of the 17 sTEC profiles within the cluster. On the right, 8 randomly selected profiles from the cluster. All the profiles show similar Es structures.

similarity matrix A falls off with the distance between the points p_i and p_j. The term $d(p_i, p_j)$ represents the distance function. As previously described, we compute the distance matrix by calculating a proxy for the pairwise earth mover's distance (EMD) between all the elements of the matrix of downsampled wavelet spectra obtained from the perturbed sTEC profiles.

Instead of selecting a single arbitrary scaling parameter σ, we compute a local scaling parameter σ_i for each point p_i (Zelnik-manor & Perona, 2005). The Gaussian kernel can now be expressed as:

$$\exp\left(-\frac{d^2(p_i, p_j)}{\sigma_i \sigma_j}\right).$$

Using a local scaling parameter for each point allows self-tuning of the point-to-point distances according to the local statistics of the neighbourhood surrounding point i. A simple choice for σ_i is:

$$\sigma_i = d(p_i, p_b)$$

where p_b is the b'th neighbor of point p_i. In our pipeline, we used the standard value of $b = 7$ (Zelnik-manor & Perona, 2005).

Once we have computed the adjacency matrix, we can apply the *k-ways normalized cuts* method to find the best k clusters (Shi & Malik, 1997) using the implementation in *sklearn.cluster.SpectralCluster*. The optimum number of clusters for a specific dataset can be estimated with the eigengap heuristic algorithm, which finds the number k such that all eigenvalues $\lambda_1, \ldots, \lambda_k$ are very small, but λ_{k+1} is relatively large (Von Luxburg, 2007).

3 Results

In this section, we present the results from our ML-based classification pipeline, which found 145 clusters in our dataset. The next step of our semi-supervised machine learning pipeline

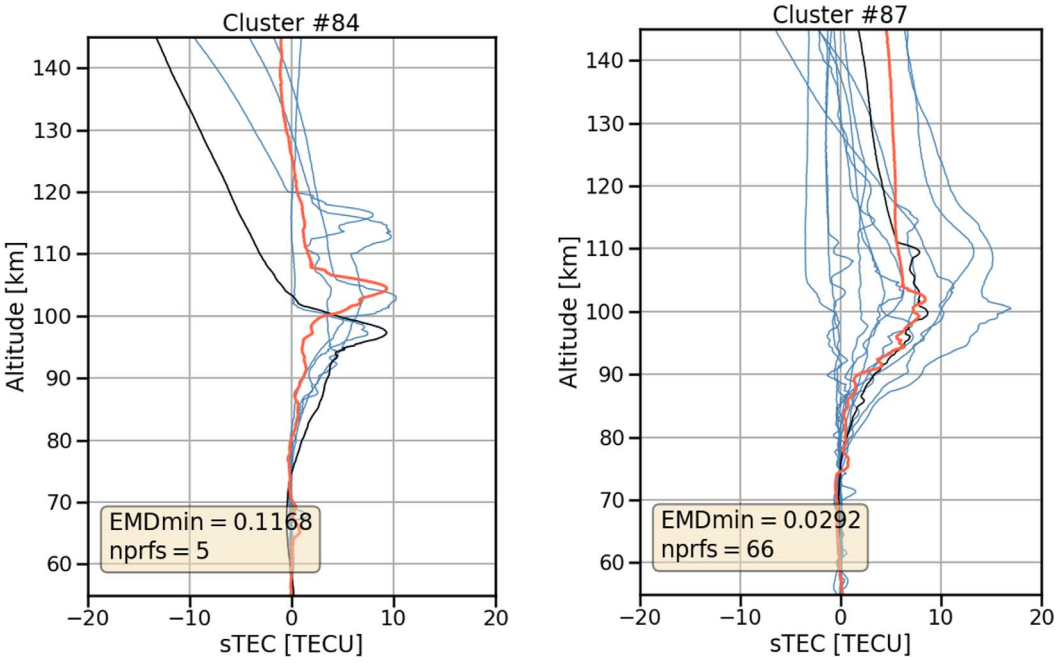

Fig. 12. Classification of a new sTEC profile (in red) as Es (left panel) and TID (right panel). The black profiles represent the closest profile with minimum EMD distance. Ten profiles from the same cluster are displayed in blue. The minimum EMD value and the total number of profiles in that cluster are displayed in the yellow box.

is to label and merge these clusters into a smaller number of classes of interest. For our application, we identify four main classes (Table 1).

Some of the clusters that belong to the database of ionospheric perturbations are presented in the following figures. Each entry of the database also contains metadata that describes the type of ionospheric perturbation in that cluster.

Figure 9 shows an example cluster that contains 258 reference profiles with small perturbations. The map on the left shows the geographic location of the sTEC profiles. The right-hand panel shows 8 profiles randomly selected from the cluster. All the profiles from this cluster contain similar small-scale sTEC ionospheric irregularities, therefore the cluster was assigned to class 3c.

Figure 10 shows another example cluster that contains 66 profiles. All the profiles from this cluster exhibit medium-scale wave-like ionospheric perturbations, therefore the cluster was assigned to class 2b. The final example in Figure 11 shows sporadic E structures. There are 17 profiles in this cluster that are assigned to class 1b.

Once the reference database of ionospheric perturbations has been constructed, we can use it to classify new profiles. First, we compute the wavelet spectrum of the new profile, then we compute the EMD distance between the new spectra and all the spectra in the reference database, and lastly, we assign the new profile to the same class as that of the closest profile (i.e., where the EMD to the reference data set is minimized).

Figure 12 shows the classification of two new profiles (in red). The black profiles represent the closest profile in the EMD space. Ten profiles of the same cluster are represented in blue. The EMD minimum value and the total number of profiles of the clusters are reported in the yellow box.

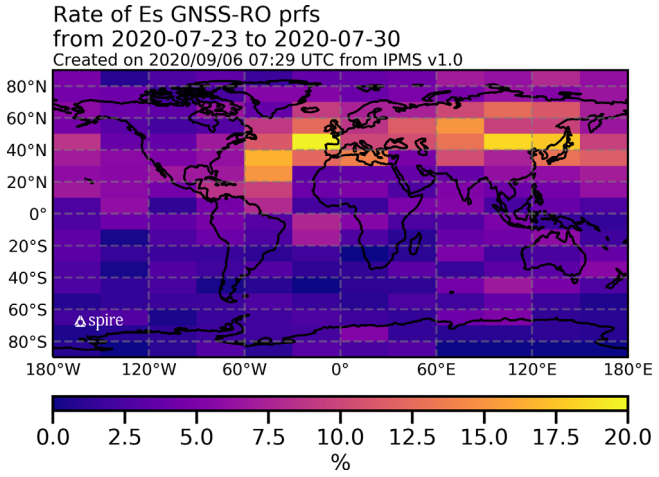

Fig. 13. Relative density of Es from Spire profiles to total number of profiles gathered globally from 2020/07/23 to 2020/07/30. The Es perturbations are picked up at fairly high rates at mid-latitude in the northern hemisphere (summertime), consistent with sporadic E climatological models.

After deploying the automated classification system, it is possible to monitor the rate at which Es and TID events are detected. Figures 13 and 14 show the final output of Spire's ionosphere perturbations monitoring system (IPMS) based on the high-rate GNSS-RO profiles. The rates of Es and TID events show a clear hemispheric asymmetry, which agrees well with the climatology of these perturbations. In particular, we can clearly see that the highest rate of TIDs occurred near the

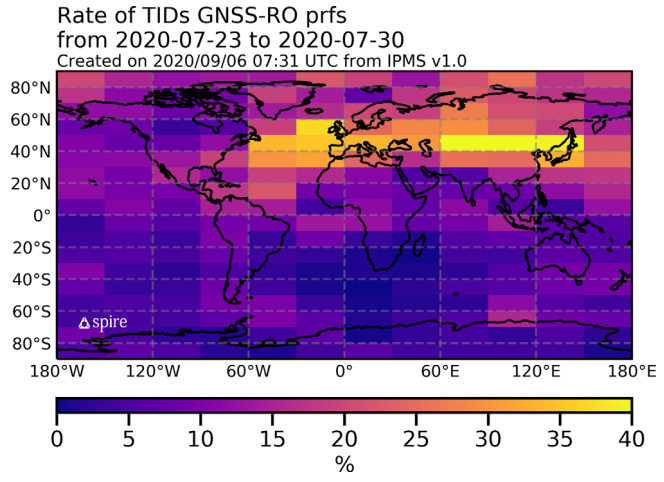

Fig. 14. Relative density of TIDs from Spire profiles to total number of profiles gathered globally from 2020/07/23 to 2020/07/30. The TIDs are picked up at high rates (i.e., more than 50%) at mid-latitude in the northern hemisphere (summertime), especially in the Chinese and Japanese sectors, consistent with TID climatological studies.

Japanese region. These results agree well with the literature on summertime nighttime TIDs in Japan (Saito et al., 2001, 2002; Otsuka et al., 2007).

4 Conclusions

In this study, we extracted valuable information on the lower ionosphere from Spire's constellation of CubeSats. In particular, Spire Global's Lower Earth Orbit CubeSats produce high-rate (50-Hz) sTEC profiles derived using the GNSS-RO technique, which can be used to detect ionospheric features such as sporadic E and TIDs. We designed and implemented an innovative semi-supervised ML-based classification algorithm that combines wavelet signal processing and EMD-based similarity clustering. In this approach, the nonlinearity and nonstationarity of GNSS-RO data can be dealt with more rigorously than in methods using the traditional paradigm of constant frequency and amplitude (e.g., Fourier analysis).

Our results show the possibility of classifying different ionospheric perturbations, which then provides the opportunity for an automated system to monitor ionospheric perturbations on a global scale. This system is currently working operationally and may be of interest to many terrestrial applications, such as radio communications, global navigation satellite system (GNSS) users, space weather data assimilation models, and natural hazard warning systems.

In the future, we may explore the implementation of the ML-based algorithm onboard Spire CubeSats in order to autonomously detect anomalies and prioritize the downlink of RO profiles.

Acknowledgements. This work was funded in part by the Government of Luxembourg through the ESA contract 4000125716/18/NL/MH/mg in the Luxembourg National Space Programme (LuxIMPULSE). The editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References

Alexander P, de la Torre A, Llamedo P. 2008. Interpretation of gravity wave signatures in GPS radio occultations. *J Geophys Res Atmos* **113(D16)**: D16117. https://doi.org/10.1029/2007JD009390.

Angling MJ, Cannon PS, Bradley P. 2012. Ionospheric propagation. In: *Propagation of radiowaves*, 3rd edn., Barclay LW (Ed.), Institution of Engineering and Technology, London, UK, pp. 199–233. https://doi.org/10.1049/PBEW056E_ch12.

Angling MJ, Nogués-Correig O, Nguyen V, Vetra-Carvalho S, Bocquet F-X, et al. 2021. Sensing the ionosphere with the Spire radio occultation constellation. *J Space Weather Space Clim* **11**: 56. https://doi.org/10.1051/swsc/2021040.

Arras C. 2010. A global survey of sporadic E layers based on GPS radio occultations by CHAMP, GRACE and FORMOSAT–3/COSMIC [Application/pdf], *Doctoral dissertation*, Deutsches GeoForschungsZentrum GFZ, Potsdam, 119 p. 32 MB. https://doi.org/10.2312/GFZ.B103-10097.

Astafyeva E. 2019. Ionospheric detection of natural hazards. *Rev Geophys* **57(4)**: 1265–1288. https://doi.org/10.1029/2019RG000668.

Aubry M, Blanc M, Clauvel R, Taieb C, Bowen PJ, et al. 1966. Some rocket results on sporadic *E*. *Radio Sci* **1(2)**: 170–177. https://doi.org/10.1002/rds196612170.

Bonneel N, Rabin J, Peyré G, Pfister H. 2015. Sliced and Radon Wasserstein barycenters of measures. *J Math Imag Vis* **51(1)**: 22–45. https://doi.org/10.1007/s10851-014-0506-3.

Booker HG. 1961. A local reduction of F-region ionization due to missile transit. *J Geophys Res* **66(4)**: 1073–1079. https://doi.org/10.1029/JZ066i004p01073.

Chou MY, Lin CCH, Yue J, Tsai HF, Sun YY, Liu JY, Chen CH. 2017. Concentric traveling ionosphere disturbances triggered by Super Typhoon Meranti (2016). *Geophys Res Lett* **44(3)**: 1219–1226. https://doi.org/10.1002/2016GL072205.

Chung F. 1997. Spectral graph theory. In: *CBMS Regional Conference Series in Mathematics, 92*, American Mathematical Society, Providence, Rhode Island, USA. http://www.ams.org/books/cbms/092.

Crowley G, Rodrigues FS. 2012. Characteristics of traveling ionospheric disturbances observed by the TIDDBIT sounder. *Radio Sci* **47(4)**: RS0L22. https://doi.org/10.1029/2011RS004959.

Duly TM, Chapagain NP, Makela JJ. 2013. Climatology of nighttime medium-scale traveling ionospheric disturbances (MSTIDs) in the Central Pacific and South American sectors. *Ann Geophys* **31(12)**: 2229–2237. https://doi.org/10.5194/angeo-31-2229-2013.

Fejer BG, Kelley MC. 1980. Ionospheric irregularities. *Rev Geophys* **18(2)**: 401. https://doi.org/10.1029/RG018i002p00401.

Galvan DA, Komjathy A, Hickey MP, Mannucci AJ. 2011. The 2009 Samoa and 2010 Chile tsunamis as observed in the ionosphere using GPS total electron content: Tsunami signatures observed in TEC. *J Geophys Res: Space Phys* **116(A6)**: A06318. https://doi.org/10.1029/2010JA016204.

Georges TM. 1967. Evidence for the influence of atmospheric waves on ionospheric motions. *J Geophys Res* **72(1)**: 422. https://doi.org/10.1029/JZ072i001p00422.

Georges TM, Hooke WH. 1970. Wave-induced fluctuations in ionospheric electron content: A model indicating some observational biases. *J Geophys Res* **75(31)**: 6295–6308. https://doi.org/10.1029/JA075i031p06295.

Grauman K, Darrell T. 2004. Fast contour matching using approximate Earth mover's distance. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Washington, DC, USA, 27 June – 2 July 2004. https://doi.org/10.1109/cvpr.2004.1315035.

Greenspan H, Dvir G, Rubner Y. 2000. Region correspondence for image matching via EMD flow. In: *Proceedings – IEEE Workshop on Content-Based Access of Image and Video Libraries, CBAIVL 2000*, Hilton Head, SC, USA, 12 June. IEEE, pp. 27–31. https://doi.org/10.1109/IVL.2000.853835.

Hajj GA, Romans LJ. 1998. Ionospheric electron density profiles obtained with the Global Positioning System: Results from the GPS/MET experiment. *Radio Sci* **33(1)**: 175–190. https://doi.org/10.1029/97RS03183.

Hajj GA, Kursinski ER, Romans LJ, Bertiger WI, Leroy SS. 2002. A technical description at atmospheric sounding by GPS occultation. *J Atmos Sol-Terr Phys* **64(4)**: 451–469. https://doi.org/10.1016/S1364-6826(01)00114-6.

Haldoupis C. 2011. A tutorial review on sporadic E layers. In: *Aeronomy of the Earth's Atmosphere and Ionosphere*, Abdu M, Pancheva D (Eds.). IAGA Special Sopron Book Series, Vol. **2**, Springer, Dordrecht, pp. 381–394. https://doi.org/10.1007/978-94-007-0326-1_29.

Huang CY, Helmboldt JF, Park J, Pedersen TR, Willemann R. 2019. Ionospheric detection of explosive events. *Rev Geophys* **57(1)**: 78–105. https://doi.org/10.1029/2017RG000594.

Jacobson AR, Carlos RC, Blanc E. 1988. Observations of ionospheric disturbances following a 5-kt chemical explosion: 1. Persistent oscillation in the lower thermosphere after shock passage. *Radio Sci* **23(5)**: 820–830. https://doi.org/10.1029/RS023i005p00820.

Kolouri S, Nadjahi K, Simsekli U, Badeau R, Rohde G. 2019. Generalized sliced Wasserstein distances. In: *Advances in neural information processing systems*, Vol. **32**, Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (Eds.), Curran Associates Inc., Red Hook, NY, USA. Retrieved from https://proceedings.neurips.cc/paper/2019/file/f0935e4cd5920aa6c7c996a5ee53a70f-Paper.pdf.

Komiske PT, Metodiev EM, Thaler J. 2019. Metric space of collider events. *Phys Rev Lett* **123(4)**: 041801. https://doi.org/10.1103/physrevlett.123.041801.

Korte BH, Vygen J. 2012. *Combinatorial optimization: Theory and algorithms*, Springer-Verlag, New York, NY. https://doi.org/10.1007/978-3-642-24488-9.

Lee G, Gommers R, Waselewski F, Wohlfahrt K, O'Leary A. 2019. PyWavelets: A Python package for wavelet analysis. *J Open Source Softw* **4(36)**: 1237. https://doi.org/10.21105/joss.01237.

Leighton HI, Shapley AH, Smith EK. 1962. The occurrence of sporadic E during the IGY. In: *Ionospheric sporadic*, Pergamon, pp. 166–177. https://doi.org/10.1016/b978-0-08-009744-2.50018-7.

Li W, Ryu EK, Osher S, Yin W, Gangbo W. 2018. A parallel method for Earth mover's distance. *J Sci Comput* **75(1)**: 182–197. https://doi.org/10.1007/s10915-017-0529-1.

MacKenzie EC, Sayers J. 1966. A radio frequency electron density probe for rocket investigation of the ionosphere. *Planet Space Sci* **14(8)**: 731–740. https://doi.org/10.1016/0032-0633(66)90103-6.

Mendillo M, Hawkins GS, Klobuchar JA. 1975. A sudden vanishing of the ionospheric F region due to the launch of Skylab. *J Geophys Res* **80(16)**: 2217–2228. https://doi.org/10.1029/ja080i016p02217.

Ng A, Jordan M, Weiss Y. 2001. On spectral clustering: Analysis and an algorithm. In: *Advances in neural information processing systems*, Vol. **14**, Dietterich T, Becker S, Ghahramani Z (Eds.), MIT Press, Cambridge, MA, USA. https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.

Ogawa T, Igarashi K, Aikya K, Maeno H. 1987. NNSS satellite observations of medium-scale traveling ionospheric disturbances at southern high-latitudes. *J Geomagn Geoelect* **39(12)**: 709–721. https://doi.org/10.5636/jgg.39.709.

Otsuka Y, Onoma F, Shiokawa K, Ogawa T, Yamamoto M, Fukao S. 2007. Simultaneous observations of nighttime medium-scale traveling ionospheric disturbances and e region field-aligned irregularities at midlatitude. *J Geophys Res: Space Phys* **112(6)**: A06317. https://doi.org/10.1029/2005JA011548.

Peleg S, Werman M, Rom H. 1989. A unified approach to the change of resolution: Space and gray-level. *IEEE Trans Pattern Anal Mach Intell* **11(7)**: 739–742. https://doi.org/10.1109/34.192468.

Rubner Y, Tomasi C, Guibas LJ. 2000. Earth mover's distance as a metric for image retrieval. *Int J Comput Vis* **40(2)**: 99–121. https://doi.org/10.1023/A:1026543900054.

Saito A, Nishimura M, Yamamoto M, Fukao S, Kubota M, et al. 2001. Traveling ionospheric distrubances detected in the FRONT campaign. *Geophys Res Lett* **28(4)**: 689–692. https://doi.org/10.1029/2000GL011884.

Saito A, Nishimura M, Yamamoto M, Fukao S, Tsugawa T, et al. 2002. Observations of traveling ionospheric disturbances and 3-m scale irregularities in the nighttime F-region ionosphere with the MU radar and a GPS network. *Earth Planets Space* **54(1)**: 31–44. https://doi.org/10.1186/BF03352419.

Santambrogio F. 2009. Absolute continuity and summability of transport densities: Simpler proofs and new estimates. *Calc Var Partial Differ Equ* **36(3)**: 343–354. https://doi.org/10.1007/s00526-009-0231-8.

Savastano G, Komjathy A, Verkhoglyadova O, Mazzoni A, Crespi M, Wei Y, Mannucci AJ. 2017. Real-time detection of tsunami ionospheric disturbances with a stand-alone GNSS receiver: A preliminary feasibility demonstration. *Sci Rep* **7(1)**: 46607. https://doi.org/10.1038/srep46607.

Savastano G, Komjathy A, Shume E, Vergados P, Ravanelli M, et al. 2019. Advantages of geostationary satellites for ionospheric anomaly studies: Ionospheric plasma depletion following a rocket launch. *Remote Sens* **11(14)**: 1734. https://doi.org/10.3390/rs11141734.

Schunk R, Nagy A. 2009. *Ionospheres*, Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/cbo9780511635342.

Shi J, Malik J. 1997. Normalized cuts and image segmentation. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 17–19 June 1997, pp. 731–737. https://doi.org/10.1109/cvpr.1997.609407.

Svehla D. 2018. *Geometrical theory of satellite orbits and gravity field*, Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-76873-1.

Tsai LC, Su SY, Liu CH, Schuh H, Wickert J, Alizadeh MM. 2018. Global morphology of ionospheric sporadic E layer from the FormoSat-3/COSMIC GPS radio occultation experiment. *GPS Solut* **22(4)**: 118. https://doi.org/10.1007/s10291-018-0782-2.

Villani C. 2003. *Topics in optimal transportation*, Vol. **58**, American Mathematical Society, Providence, Rhode Island. https://doi.org/10.1090/gsm/058.

Von Luxburg U. 2007. A tutorial on spectral clustering. *Statist Comput* **17(4)**: 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Wakabayashi M, Ono T, Mori H, Bernhardt PA. 2005. Electron density and plasma waves in mid-latitude sporadic-E layer observed during the SEEK-2 campaign. *Ann Geophys* **23(7)**: 2335–2345. https://doi.org/10.5194/angeo-23-2335-2005.

Whitehead JD. 1970. Production and prediction of sporadic E. *Rev Geophys* **8(1)**: 65. https://doi.org/10.1029/RG008i001p00065.

Whitehead JD. 1989. Recent work on mid-latitude and equatorial sporadic-E. *J Atmos Terr Phys* **51(5)**: 401–424. https://doi.org/10.1016/0021-9169(89)90122-0.

Wu DL. 2005. Sporadic E morphology from GPS-CHAMP radio occultation. *J Geophys Res* **110(A1)**: A01306. https://doi.org/10.1029/2004JA010701.

Wu DL. 2018. New global electron density observations from GPS-RO in the D- and E-Region ionosphere. *J Atmos Sol-Terr Phys* **171**: 36–59. https://doi.org/10.1016/j.jastp.2017.07.013.

Wu Z, Leahy R. 1993. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Trans Pattern Anal Mach Intell* **15(11)**: 1101–1113. https://doi.org/10.1109/34.244673.

Wu DL, Ao CO, Hajj GA, De La Torre Juarez M, Mannucci AJ. 2005. Sporadic E morphology from GPS-CHAMP radio occultation. *J Geophys Res: Space Phys* **110(A1)**: A01306. https://doi.org/10.1029/2004JA010701.

Yeh KC, Liu CH. 1974. Acoustic-gravity waves in the upper atmosphere. *Rev Geophys* **12(2)**: 193–216. https://doi.org/10.1029/RG012i002p00193.

Zelnik-manor L, Perona P. 2005. Self-Tuning Spectral Clustering. In: *Advances in neural information processing systems*, Vol. **17**, Saul L, Weiss Y, Bottou L (Eds.), MIT Press, Cambridge, MA, USA. Retrieved from https://proceedings.neurips.cc/paper/2004/file/40173ea48d9567f1f393b20c855bb40b-Paper.pdf.

Cite this article as: Savastano G, Nordström K & Angling MJ 2022. Semi-supervised classification of lower-ionospheric perturbations using GNSS radio occultation observations from Spire Global's Cubesat Constellation. *J. Space Weather Space Clim.* **12**, 14. https://doi.org/10.1051/swsc/2022009.